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. 83, at pages 73–74, and is available at the *Magazine*'s website www.maa.org/pubs/mathmag.html. Manuscripts to be submitted should not be concurrently submitted to, accepted for publication by, or published by another journal or publisher.

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

Cover image by Susan Stromquist, following the directions on pages 83–84 for drawing lines on a hyperbolic plane.

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

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

Advertising correspondence should be addressed to

MAA Advertising
1529 Eighteenth St. NW
Washington DC 20036

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

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

Change of address, missing issue inquiries, and other subscription correspondence:

MAA Service Center, *maahq@maa.org*

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

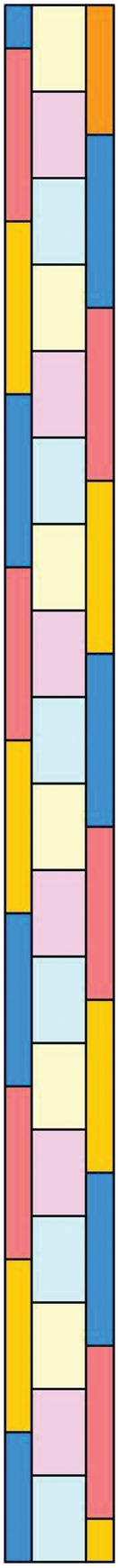
Periodicals postage paid at Washington, D.C. and additional mailing offices.

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

Printed in the United States of America

Vol. 83, No. 2, April 2010

MATHEMATICS MAGAZINE



LETTER FROM THE EDITOR

In the margin we celebrate color and symmetry. How many mathematical stories can you find in this graphic?

What? Is MATHEMATICS MAGAZINE in color? Evidently so, at least for the first few pages of this issue. It is an experiment, and it isn't free. Feedback from readers and prospective authors would be greatly appreciated (mathmag@maa.org). (Full color is now available in the MAA's online publications, including the online version of this MAGAZINE.)

Color enlivens the illustrations in the first article, on how to draw triangles, by Curtis D. Bennett, Blake Mellor, and Patrick D. Shanahan. Perhaps you already know how to draw triangles, but these authors are working in the Thurston model of the hyperbolic plane, and they use their triangles to present a version of Pick's theorem in that environment.

The other articles add color in your imagination. Two separate articles relate group theory to Sudoku puzzles. Carlos Arcos, Gary Brookfield, and Mike Krebs show us how to use symmetry groups to count Sudoku grids, and Jennifer Carmichael, Keith Schloeman, and Michael B. Ward ask when a group multiplication table is itself a Sudoku. Kent Morrison uses multiplication tables in his imaginary casino to rediscover a famous probability distribution. (Where else in the issue are multiplication tables and Sudoku mentioned?)

In the Notes Section, Aaron Melman helps us find eigenvalues and Dimitrios Kodokostas helps us find triangle equalizers. In News and Letters you can find a report of the 38th USAMO.

Sixty years of quickies. This issue's Problems section features Quickie #1000. Quickies are one of the distinctive features of the Problems section that set it apart from other problems columns. We publish two quickies per issue, five issues per volume, and this is Volume 83, so anyone can calculate that the first quickie appeared in Volume. . . Oops. Apparently the rate has varied.

Actually, Quickie #1 appeared in Volume 23, March–April, 1950. Regular problem numbers were restarted at #1 in 1947 when the Magazine took on its present name. You can read more about the history of the MAGAZINE and its Problem section in our April, 2005 issue, or in the MAA book, *The Harmony of the World.* You can see the first quickie on page 148 of this issue.

Walter Stromquist, Editor

CONTENTS

The Mathematical Association of America
1529 Eighteenth Street, NW
Washington, DC 20036

ARTICLES

Drawing a Triangle on the Thurston Model of Hyperbolic Space

CURTIS D. BENNETT
Loyola Marymount University
Los Angeles, CA 90045
cbennett@lmu.edu

BLAKE MELLOR
Loyola Marymount University
Los Angeles, CA 90045
bmellor@lmu.edu

PATRICK D. SHANAHAN
Loyola Marymount University
Los Angeles, CA 90045
pshanahan@lmu.edu

Experiments with a common physical model of the hyperbolic plane presented the authors with surprising difficulties in drawing a large triangle. Understanding these difficulties led to an intriguing exploration of the geometry of the Thurston model of the hyperbolic plane. In this exploration we encountered topics ranging from combinatorics and Pick's Theorem to differential geometry and the Gauss-Bonnet Theorem.

The journey began when one of the authors was teaching a class of non-mathematics majors using Ed Burger and Michael Starbird's popular text *The Heart of Mathematics* [**1**]. In section 4.6, Burger and Starbird describe how to build a model of the hyperbolic plane out of paper by taping together equilateral triangles with 7 triangles around each vertex; FIGURE 1 shows the result. They then ask the following question:

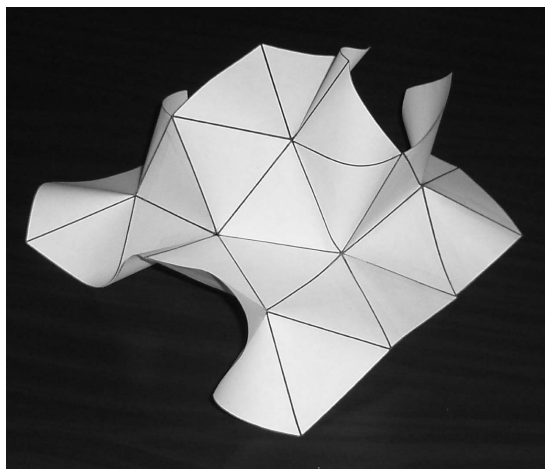


Figure 1 The Thurston model of the hyperbolic plane

Math. Mag. **83** (2010) 83–99. doi:10.4169/002557010X482853.

> Draw a big triangle upon your floppy sheet (the model) spanning several of the pieces by flattening a section on the ground and drawing a straight line, then flattening another section and drawing another straight line, and then completing the triangle in the same way. There is a lot of squashing involved. Now measure the three angles and add them up. What do you get? (section 4.6, problem #18)

This question is unexpectedly difficult to answer and raises interesting questions about the relationship between the model and the hyperbolic plane. For example, what is meant by a "big" triangle? And what is a "straight line"?

The model described by Burger and Starbird was initially suggested by William Thurston as a way for people to get a feeling for hyperbolic space, and has appeared in several books aimed at a general audience, in particular, *The Shape of Space* by Jeffrey Weeks [**5**, p. 151] and *The Heart of Mathematics* [**1**, p. 301]. We encourage readers to construct their own models, both to verify for themselves the results in this paper, and simply because they are very cool toys!

Notice that the Thurston model shown in FIGURE 1 cannot be flattened onto the plane because we are forcing $7\pi/3$ radians to fit around each vertex rather than the 2π radians allowed in the Euclidean plane. However, there are strips of equilateral triangles in the model that *can* be flattened onto the Euclidean plane, as shown in FIGURE 2.

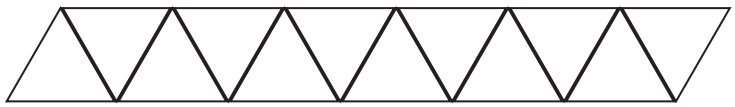


Figure 2 A strip of equilateral triangles in the Euclidean plane

When Burger and Starbird ask us to draw a "big" triangle, it is natural to think in terms of area. However, we can draw a triangle with as much area as we wish within one of these Euclidean strips of triangles, and the result will have an angle sum of π. Since the purpose of the model is to illustrate the *differences* between Euclidean and hyperbolic geometry, this is clearly not what was meant. Instead of looking at area *per se*, we want to draw a triangle containing a large number of the vertices of the model in its interior.

Before we can begin to draw any kind of triangle, big or small, we need to know what we mean by straight lines in the model. Burger and Starbird suggest we should "flatten a section [of the model] on the ground" and draw a straight line on this flattened section. But, then, what of a line that runs along the sides of one of the Euclidean strips shown in FIGURE 2? This certainly seems like a straight line—and yet, since it passes through vertices where the model *cannot* be flattened without folding the model onto itself, they cannot be drawn as Burger and Starbird describe. How should we resolve this? Answering this questions leads to some beautiful mathematics, including the Gauss-Bonnet Theorem relating the area of a hyperbolic triangle to the sum of its angles.

Drawing lines in Thurston models

Before we dive into the nitty gritty of drawing lines and triangles, we need to address to what extent the Thurston model actually models hyperbolic space. It might be better to say that it is an *approximate* model, in the same way that an icosahedron

is an approximate model of the sphere. It is most natural to look at the geometry on the Thurston model induced by its embedding as a surface in $\mathbb{R}^3$; however, this geometry does not strictly satisfy the axioms of hyperbolic geometry (or even incidence geometry!). Alternatively, we can define a map from an actual hyperbolic plane to the Thurston model, and use this map to define the geometry of the model; this results in a different measure of distance, and hence in different lines and polygons. We are often interested in comparing these two perspectives. The "natural" geometry is easier to use in a classroom (as long as we place certain restrictions), so we begin from that point of view by defining *Thurston lines* (these are the lines imagined by Burger and Starbird). In subsequent sections we will define the correspondence between the Thurston model and a standard model of hyperbolic space, the *Poincaré disk model*, and use it to define a different set of lines, the *hyperbolic lines*. By comparing these two notions of lines we will see that this natural geometry, while not the same as the hyperbolic geometry, does provide a useful approximation.

Thurston lines The standard method to define a line in a space is as the shortest path (or *geodesic*) between any two points of the space. In the Thurston model, measuring distance as a surface embedded in $\mathbb{R}^3$, we will call these lines *Thurston lines*. This definition fits in well with the Burger-Starbird problem, as a line on a "flattened section" of the model would be a geodesic. Our definition of Thurston lines will not include all geodesics. The reader is encouraged to think about complications that occur when geodesics lie in sections of the model that cannot be flattened.

We begin by defining some key terms. A *model triangle* will denote one of the Euclidean triangles. Two model triangles are *adjacent* if they share an edge (meaning they have been glued together along an edge). A *model vertex* is a vertex of any model triangle. Intuitively, a Thurston line will have two properties: It never passes through a model vertex, and when it passes through two adjacent triangles, its restriction to the union of the triangles is a Euclidean line segment, as in FIGURE 3. These properties guarantee that a Thurston line lies in a section of the model that can be flattened.

We now formally define a *Thurston line* to be a set of points ℓ such that

1. The restriction of ℓ to any model triangle T is either empty or a line segment of T containing a no vertex of T.
2. If T_1 and T_2 are adjacent triangles sharing edge $\overline{AB}$, with $\ell \cap \overline{AB} = C$, $X_i \neq C$ and $X_i \in \ell \cap T_i$ for $i = 1, 2$, then $\angle X_1CA \cong \angle X_2CB$.

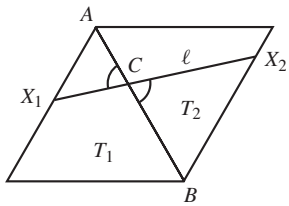


Figure 3 The Thurston line segment $\overline{X_1X_2}$ is the restriction of a Thurston line ℓ to adjacent triangles T_1 and T_2

A *Thurston angle* is now defined naturally as an angle formed by two intersecting Thurston lines. Since the rays of a Thurston angle are subsets of Thurston lines, the vertex of a Thurston angle is *not* a model vertex. Thus, any Thurston angle agrees locally with a Euclidean angle that is inside either a model triangle or two adjacent

model triangles, and we define the measure of a Thurston angle to be its Euclidean measure. Define a *Thurston triangle* as the figure bounded by three Thurston lines.

The Burger-Starbird question can now be rephrased as asking us to draw a Thurston triangle with at least one model vertex in its interior and then find the sum of its angles. Curiously, at most two model vertices can lie in the interior of a Thurston triangle, as we will show.

Drawing large Thurston triangles We now turn to the question of how "big" triangles in our geometry can be, by which we mean how many model vertices they may contain. Suppose first that we have a Thurston triangle in our geometry. That is, we have points A, B, and C such that each of $\overline{AB}$, $\overline{BC}$, and $\overline{AC}$ lies on a piece of the space that can be flattened.

The model triangles partition the interior of $\triangle ABC$ into a collection of complete model triangles and pieces of model triangles. As some of these pieces may be quadrilaterals, we further triangulate the pieces by adding additional edges (but no new vertices). This gives a triangulation of $\triangle ABC$ in which every triangle lies on a flat region of the model, and all the vertices are either model vertices in the interior, or non-model vertices on the boundary. Since all of the triangles in the triangulation are Euclidean, they must each have angle sum of π radians.

To count these triangles, we use Euler's formula for a triangulation of a topological disk: $V - E + F = 1$, where V is the number of vertices, E the number of edges, and F the number of faces in the triangulation. We can write $V = 3 + b + m$, counting the three points A, B, and C, the b additional vertices on the edges $\overline{AB}$, $\overline{BC}$, and $\overline{AC}$, and the m internal model vertices. A standard combinatorial argument shows that the total number of edges in the triangulation is

$$E = \frac{3F + b + 3}{2}.$$

Substituting this into Euler's Formula and solving for F yields

$$F = 1 + b + 2m.$$

Since every triangle has an angle sum of π, the sum of all the angles in the triangulation is $\pi F = \pi(1 + b + 2m)$. On the other hand, the angles around each boundary vertex (excepting A, B, and C) sum to π and the angles around each model vertex sum to $7\pi/3$. So we have two ways of computing the sum of the angles in the triangulation, producing the equation

$$\pi(1 + b + 2m) = \angle A + \angle B + \angle C + \pi b + \frac{7\pi}{3}m.$$

Therefore

$$\angle A + \angle B + \angle C = \pi\left(1 - \frac{m}{3}\right),$$

and we have established the following proposition.

PROPOSITION 1. *Any Thurston triangle $\triangle ABC$ has an angle sum equal to $\pi(1 - m/3)$ radians, where m denotes the number of model vertices in the interior of $\triangle ABC$.*

Since any Thurston triangle must have angles with positive measure, it follows that any Thurston triangle can have at most two model vertices on its interior. A triangle containing two model vertices is shown in FIGURE 4.

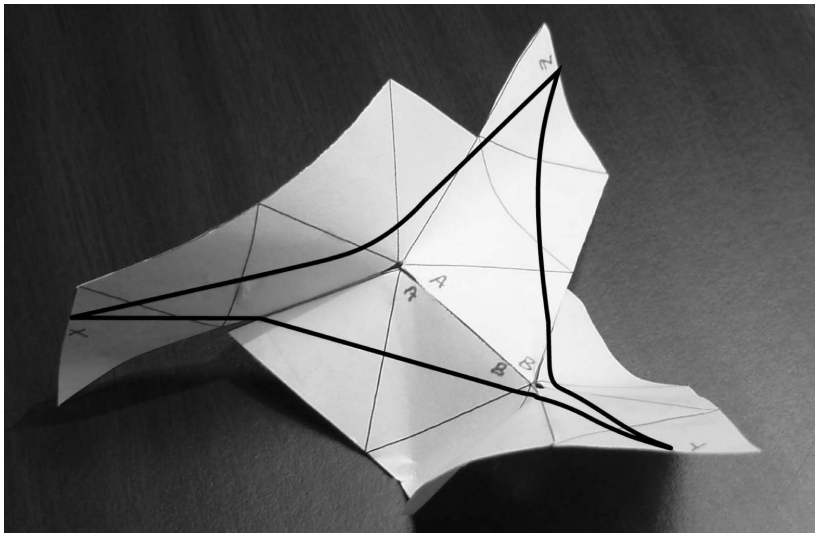


Figure 4 A triangle in the Thurston model enclosing two model vertices

A mapping between the Poincaré and Thurston models Although Thurston lines allow us to get a feel for the curvature of hyperbolic space, they are actually not hyperbolic lines. To define actual hyperbolic lines on the Thurston model, we define a map to the model from one of the standard models of hyperbolic space. We will use the standard Poincaré disk model for the hyperbolic plane, where the geodesics are the diameters of the disk and the circular arcs that are perpendicular to the boundary of the disk. We note that one can tile the Poincaré disk with equilateral triangles so that each angle measures $2\pi/7$, as shown in FIGURE 5. There is a one-to-one correspondence between this tiling and the triangles of the Thurston model. We can use this correspondence to define a bijective mapping from the Poincaré model of hyperbolic space to the Thurston model. The details of this mapping are given in the next two paragraphs for the interested reader, but only the fact of its existence is required for the rest of the paper.

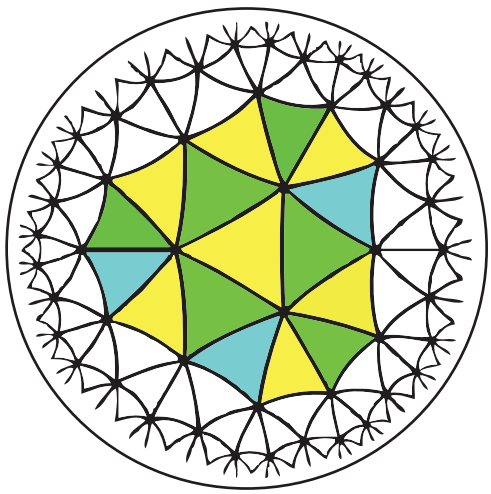


Figure 5 The Poincaré model tiled with equilateral triangles

First, let S be the triangle centered at the origin in the triangulation of the Poincaré disk shown in FIGURE 5. Next, pick a base triangle T in the Thurston model. We define a mapping $f : S \to T$, starting with the Beltrami-Klein disk model of hyperbolic geometry shown in FIGURE 6, where the geodesics are the Euclidean lines in the disk [**2**, pp. 297–301]. We view both the Poincaré model and the Beltrami-Klein model as unit disks in $\mathbb{C}$. Then the function $p(z) = 2z/(1 + |z|^2)$ maps the Poincaré disk to the Beltrami-Klein disk and takes S to a Euclidean equilateral triangle T' centered at the origin of the Beltrami-Klein disk. We map T' to T via a linear rescaling $l(z) = kz$, where k is a positive real constant. It is easy to verify that both p and l are invariant under conjugation by any symmetry of an equilateral triangle. Thus the mapping $f : S \to T$ defined by $f = l \circ p$ is also invariant under these symmetries. Note that the mapping f takes hyperbolic line segments in S to Euclidean line segments in T.

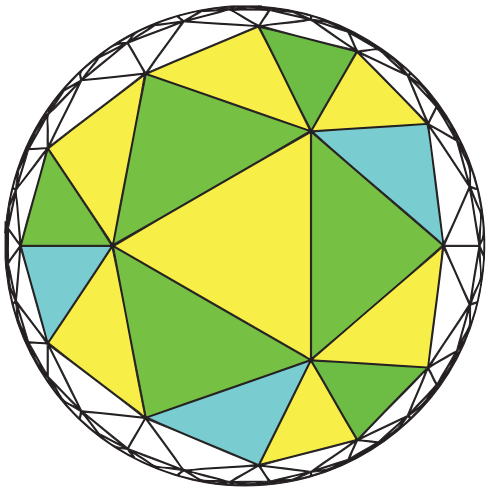


Figure 6 The Beltrami-Klein model tiled with equilateral triangles

We now describe how to extend f to a mapping from the entire Poincaré disk to the Thurston model. Given a triangle S_i of the tiling of the Poincaré disk, there exists an isometry g of the disk such that $g(S_i) = S$. We construct g by choosing a path of triangles from S_i to S in the triangulation, and composing reflections across the sides of the triangles along the path. In the Thurston model, we can inductively reverse this path of triangles and reflections to construct a mapping $\tilde{g}^{-1}$ from T to a unique triangle T_i in the Thurston model. For a point x of S_i, we define $\phi(x) = \tilde{g}^{-1} \circ f \circ g(x)$. To see that this is well defined, observe that if g' were constructed using a different path from S_i to S, then g and g' differ by a symmetry of the equilateral triangle S (and similarly for $\tilde{g}$ and $\tilde{g'}$). Since f is invariant under these symmetries, ϕ is independent of the choice of the path.

We have now defined our mapping ϕ between the models. Under this mapping we have a natural set of lines in the Thurston model, namely the images of hyperbolic lines under ϕ. These lines are only piecewise linear and may pass through model vertices. While, on the face of it, the segments of these hyperbolic lines inside model triangles are Euclidean line segments, the Euclidean distance between two points on the segment is *not* the same as the hyperbolic distance.

However, there is a particular class of these lines that we will call *special hyperbolic lines*, which are also geodesics in the Thurston model. The intersection of a special hyperbolic line with a model triangle is either a side of the triangle or the Euclidean

line segment from a vertex to the midpoint of the opposite side. When these special hyperbolic lines pass through a model vertex, by symmetry there is the same Euclidean angle sum (of $7\pi/6$) on either side. FIGURE 7 shows that these special hyperbolic lines arise naturally in the barycentric subdivision of the tiling of the hyperbolic plane by equilateral triangles. We call the triangles of this subdivision the *barycentric triangles*.

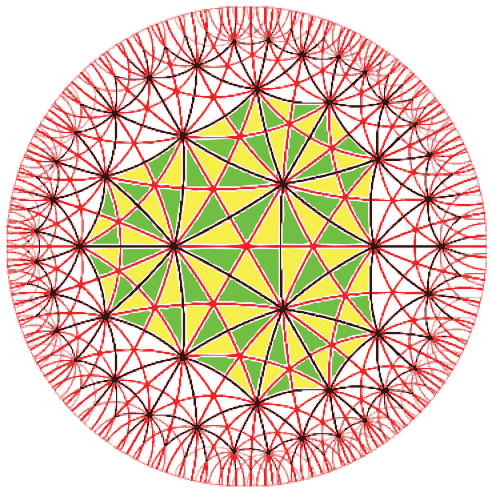


Figure 7 The barycentric subdivision of the Poincaré model, showing the special hyperbolic lines

Drawing large special hyperbolic triangles We have answered Burger and Starbird's question for Thurston triangles, but what if we take a triangle whose edges lie on special hyperbolic lines? Such a triangle is the image of a hyperbolic triangle and, unlike our earlier candidate for a large triangle, can have both internal and boundary model vertices. Triangulate this hyperbolic triangle so that the interior of each small triangle lies inside a model triangle. In this case, again, the angle measure around model vertices in the interior is $7\pi/3$. The angle sum around model vertices on the boundary, however, is only half as much, $7\pi/6$. As in Proposition 1, we discover the following:

PROPOSITION 2. *Any special hyperbolic triangle has angle sum equal to* $\pi(1 - m/3 - n/6)$ *radians, where m denotes the number of model vertices in the interior of the triangle and n denotes the number of model vertices on the edges of the triangle (not including the triangle vertices themselves).*

Since the smallest angle we could realize on a special hyperbolic triangle has measure $\pi/6$, the proposition implies that the largest number of model vertices that could lie on the triangle is 3 (with $m = 0$ and $n = 3$), and this can be realized, as shown in FIGURE 8.

Deflections of hyperbolic lines in the Thurston model

We have defined *hyperbolic lines* in the Thurston model as the images of the geodesics in the Poincaré model; however, aside from the special hyperbolic lines, we have not discussed what these lines look like in our collection of taped-together triangles. As we mentioned before, the image of a hyperbolic line in any model triangle it passes

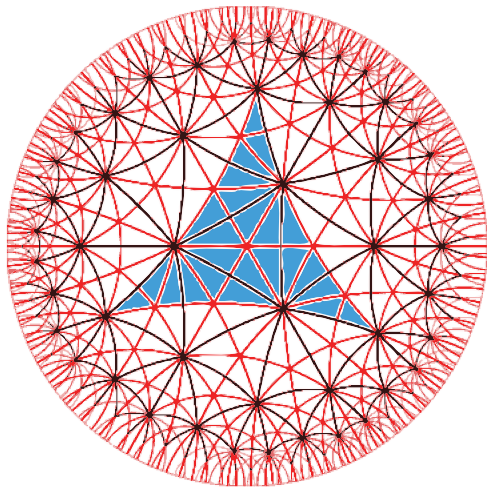


Figure 8 A special hyperbolic triangle in the Poincaré model with three model vertices on the boundary

through is a Euclidean line segment, so the question is how the line bends as it passes between adjacent triangles.

Consider the two equilateral triangles in the Poincaré model on the left in FIGURE 9, together with the hyperbolic line l, and the image of the triangles and the line under ϕ in the Thurston model on the right.

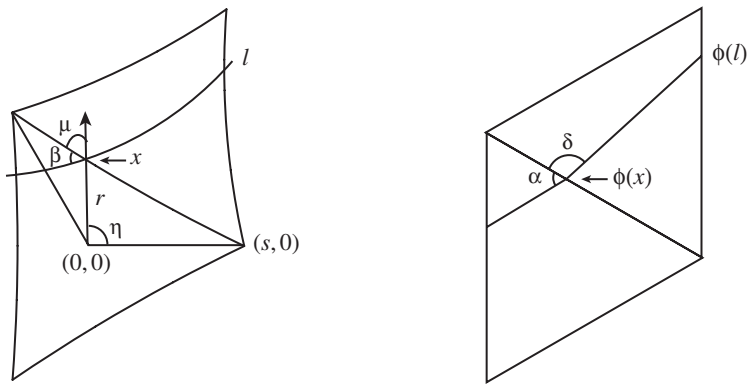


Figure 9 Equilateral triangles and a hyperbolic line in the Poincaré model, and their images in the Thurston model

The angles α and δ in the Thurston model are determined by the angles η and β in the Poincaré model, which together are enough to determine where in the Poincaré disk the hyperbolic line intersects the side of the triangle, as well as the angle of intersection. The formulas for α and δ are quite complicated, involving the derivatives of the mappings ϕ and ϕ^{-1} at the point of intersection. We content ourselves with showing these quantities graphically and leave the (somewhat lengthy) details as an exercise for the reader, with brief answers posted at the MAGAZINE website.

In general, $\alpha + \delta \neq \pi$; we want to measure the *deflection* $\alpha + \delta - \pi$. A graph showing the deflection as a function of the angles η and β appears in FIGURE 10, where η ranges from 0 to $2\pi/3$ and β range from 0 to π.

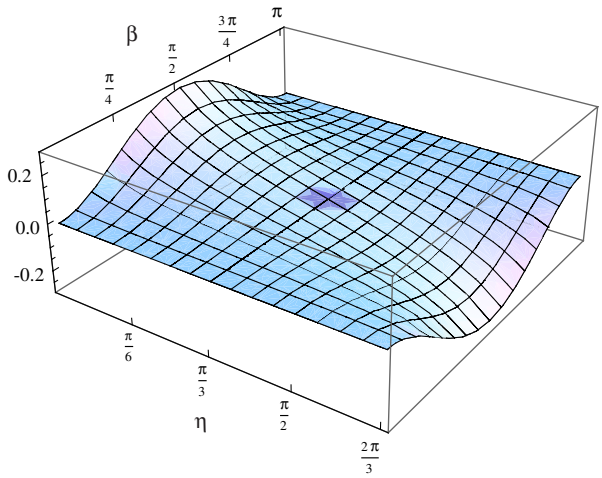


Figure 10 The deflection as a function of the angles η and β

We can now make several interesting observations. First of all, the greatest deflection occurs at $\beta = \pi/2$, when the line is perpendicular to the side of the triangle. FIGURE 11 shows the cross-section of the graph in FIGURE 10 with $\beta = \pi/2$.

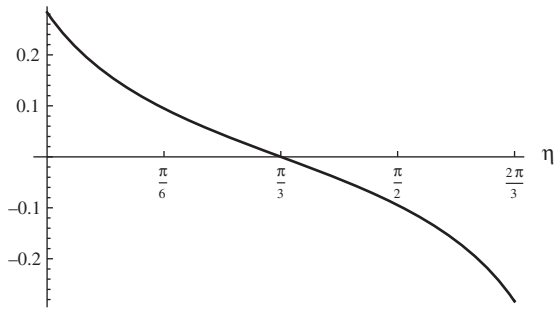


Figure 11 The deflection when $\beta = \pi/2$

The figure shows that, as we approach a vertex, the line is deflected *toward* that vertex, with a maximal deflection that approaches 0.283278 radians (about 16.23 degrees). The amount of the maximal deflection is determined by the equilateral triangle we choose in the Poincaré disk; for the computations that led to FIGURE 11, we chose the triangle centered at the origin with angles measuring $\pi/7$. If we had chosen a larger equilateral triangle (decreasing the angle measures), then this maximal deflection would increase. For example, if the three angle measures were $\pi/8$, the maximal deflection would be 0.469475 radians (about 26.9 degrees). As the angle measures decrease, the number of triangles around each vertex in the corresponding Thurston model increases, and the maximal deflection increases asymptotically toward $\pi/3$. The reason is that, as the number of triangles around each vertex increases, a line passing near one of the vertices will have to pass through more triangles. In the corresponding Thurston model, this means the line will need to be deflected to bend around the vertex. At a deflection of $\pi/3$, a line could be bent into a spiral around a vertex that passes through *all* the triangles around that vertex.

The other interesting observation is that there is *no* deflection when the line passes through the midpoint of a side (when $\eta = \pi/3$). So we see that the Thurston line in

the Thurston model that connects the midpoints along a strip of triangles is also a true hyperbolic line, meaning that it is the image of a hyperbolic line under the mapping from the Poincaré model to the Thurston model.

We can also show that there is no deflection through midpoints directly by symmetry considerations. Consider two adjacent model triangles $\triangle ABC$ and $\triangle BCD$, and let x be the midpoint of the shared edge BC, as shown in FIGURE 12. There is an isometry g of the Poincaré model that takes $\phi^{-1}(\triangle ABC)$ to $\phi^{-1}(\triangle DCB)$ by rotating by π radians around $\phi^{-1}(x)$. Consider a point p on the edge AC and its image $q = \phi g \phi^{-1}(p)$. In the Poincaré model, g preserves lines through $\phi^{-1}(x)$; since it exchanges $\phi^{-1}(p)$ and $\phi^{-1}(q)$, the three points $\phi^{-1}(p)$, $\phi^{-1}(x)$ and $\phi^{-1}(q)$ must lie on the same line in the Poincaré model. The image of this line in the Thurston model is the pair of line segments px and xq. However, since g is an isometry, we know that $|px| = |qx|$, $|Cx| = |Bx|$ and $|pC| = |qB|$, so by Side-Side-Side congruence the triangles $\triangle pxC$ and $\triangle qxB$ are congruent. In particular, $\angle pxC = \angle qxB$, which means that the image of the hyperbolic line is the Euclidean line between p and q. We conclude that there is no deflection through the midpoint x.

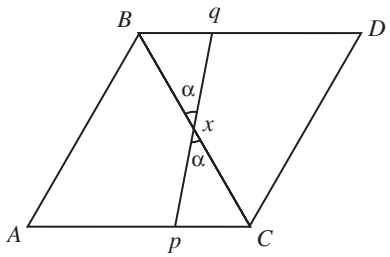


Figure 12 The image $\overline{pq}$ of a hyperbolic line segment through the midpoint x

FIGURE 13 compares hyperbolic lines and Thurston lines in a segment of the Thurston model. In each example, we have drawn both the Thurston line and the hyperbolic line connecting two points in the Thurston model. We can see that when the hyperbolic line is near the midpoints, it is almost straight and very close to the Thurston line; however, when it is farther from the midpoints, the deflections are much greater.

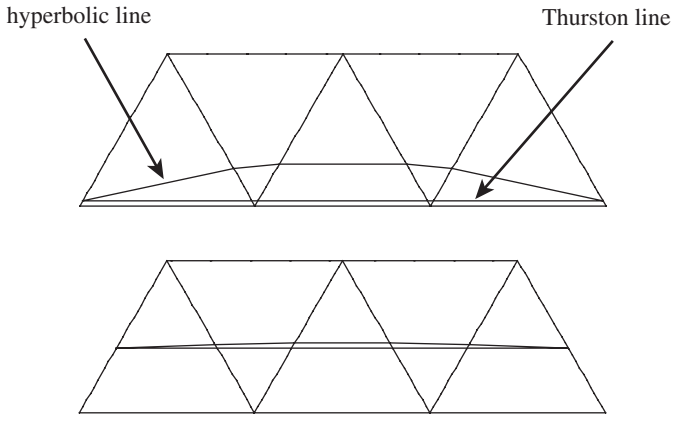


Figure 13 Thurston lines and hyperbolic lines in the Thurston model

Pick's Theorem in Thurston's model

We continue to explore our model by establishing a hyperbolic analog of Pick's Theorem, which gives a simple formula in Euclidean geometry for computing the area of a polygon drawn on a unit square lattice (meaning that the area of one square of the lattice is 1). It has many applications and generalizations [**4, 3**]. Here is the simplest form of Pick's Theorem: If a polygon P is drawn on a square lattice so that all the vertices are lattice points, if there are i vertices inside the polygon, and if there are b vertices on the boundary of the polygon, then the area of the polygon is

$$A(P) = i + b/2 - 1.$$

For example, the area of the polygon in FIGURE 14 is $A = 5 + 7/2 - 1 = 7.5$.

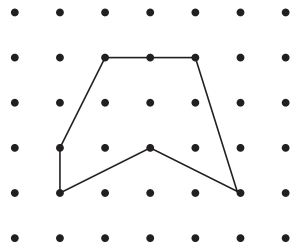


Figure 14 A polygon in a unit square lattice whose area is 7.5 units by Pick's Theorem

The special hyperbolic lines of the Thurston model are the lines corresponding to the barycentric subdivision of our triangulation of hyperbolic space. Notice that all the small triangles formed by this subdivision are congruent, and so they all have the same area α. Now, suppose we have a special hyperbolic figure R in the model, that is, each side of R is made up of special hyperbolic lines and each vertex is either a model vertex, a model center or a model midpoint as in FIGURE 15. We will also assume that R is simply connected and hence a topological disk.

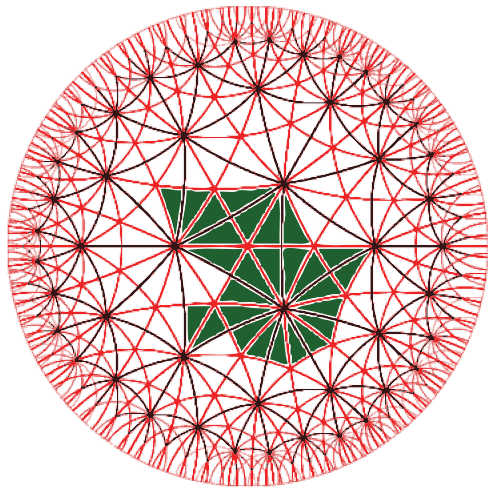


Figure 15 A special hyperbolic figure R (shaded) in the Poincaré model whose area is 27 units by Proposition 3

The area of R is equal to α times the number of barycentric triangles contained in R. So, if we want the area of R, we can count the number of barycentric triangles in R. For this, we once again recall Euler's formula for a tiling of a disk: $V - E + F = 1$. Letting our lattice points be the centers, midpoints, and vertices of the model triangles (so the lattice points are the vertices of the barycentric subdivision), we know that each internal edge lies on exactly two faces, whereas each boundary edge lies on exactly one face. Our faces are all barycentric triangles, so every face is bordered by three edges. Letting V_b be the number of boundary vertices, E_i be the number of internal edges, and E_b be the number of boundary edges, we have $E_b = V_b$ and $2E_i + E_b = 3F$, or $3F = 2E - E_b = 2E - V_b$. Thus $E = (1/2)(3F + V_b)$. Letting V_i denote the number of internal vertices, we have

$$V - E + F = (V_b + V_i) - (1/2)(3F + V_b) + F = 1.$$

Solving for F we obtain

$$F = 2V_i + V_b - 2.$$

But F is the number of barycentric triangles we have in the region R. Consequently, we have proved:

PROPOSITION 3. *Let R be a region bounded by special hyperbolic lines in the Thurston model. Then the area of region R is given by*

$$Area(R) = (2V_i + V_b - 2)\alpha,$$

where α is the area of the barycentric triangle.

In fact, as others have noted, the hardest step in proving Pick's theorem is to show that any minimal triangle has area $1/2$, and the result follows from Euler's formula. In our case, all minimal triangles are congruent, since they are images of the fundamental domain for the group action on the hyperbolic plane, so our result is not too surprising. Notice that in FIGURE 15, the region has 5 internal vertices and 19 boundary vertices, so the area is $(10 + 19 - 2)\alpha = 27\alpha$; and indeed the region contains 27 triangles of the barycentric subdivision.

Suppose we make a slightly different restriction on our region R, namely that all the vertices must be model vertices. One quickly sees that there are only two minimal triangles in this case, the model triangle and a triangle created by bisecting the quadrilateral formed by two adjacent model triangles. By symmetry arguments, both of these triangles have area $\beta = 6\alpha$. As a result, we have:

PROPOSITION 4. *Let R be a region bounded by special hyperbolic lines in the Thurston model, and let V_i denote the number of model vertices inside R and V_b denote the number of model vertices on the boundary of R. Then the area of region R is given by*

$$Area(R) = (2V_i + V_b - 2)\beta,$$

where β is the area of the model triangle.

General Thurston models

Of course, Thurston's model is just one way to model hyperbolic space; there are many others that may allow some constructions to be performed more easily. It turns out that

if we generalize the Thurston model, then we can make flatter models that allow for a wider variety of triangles. We define a *general Thurston model*: Take any regular triangulation of hyperbolic space given by an integer triple (n_1, n_2, n_3), representing the fundamental triangle with angle measures $(2\pi/n_1, 2\pi/n_2, 2\pi/n_3)$. We make one additional requirement that if one of n_1, n_2, n_3 is odd, then the other two are equal. With these conditions, we can create a tiling of hyperbolic space with the property that all angles about any vertex are congruent. Associated to this tiling, we take a Euclidean triangle with angle measures a_1, a_2, and a_3 such that $n_1a_1 = n_2a_2 = n_3a_3$. Then we tape together n_1 vertices of angle measure a_1, n_2 vertices of angle measure a_2, and n_3 angles of measure a_3. In the standard Thurston model, $n_1 = n_2 = n_3 = 7$ and $a_1 = a_2 = a_3 = \pi/3$.

The requirement $n_1a_1 = n_2a_2 = n_3a_3$ means that at each vertex of the model, the excess in angle is the same. This means that the amount the paper must bend in order for us to tape together the triangles is the same at each vertex; we might naively refer to this as the *curvature*. In these general models you lose a little bit of regularity in the sense that the vertices are not evenly spaced out, and it also becomes a little harder to flatten out the space to draw a straight line. On the other hand, you can make the curvature much smaller, allowing you to draw a greater variety of triangles. A simple calculation shows that under these conditions

$$a_i = \frac{n_j n_k}{n_1n_2 + n_2n_3 + n_1n_3}\pi,$$

and that the excess angle glued around a vertex (called the *angle excess*) is

$$\mathrm{E}(n_1, n_2, n_3) = \left(\frac{n_1n_2n_3}{n_1n_2 + n_2n_3 + n_1n_3} - 2\right)\pi.$$

Of course, if $1/n_1 + 1/n_2 + 1/n_3 > 1/2$, we have too little angle around a vertex and our triangle corresponds to a tiling of the sphere, so that $(3, 3, 3)$ produces a tetrahedron, $(4, 4, 4)$ produces an octahedron, and $(5, 5, 5)$ produces an icosahedron. If $1/n_1 + 1/n_2 + 1/n_3 = 1/2$, then our triangle tiles Euclidean space. Thus, for our purposes, we will restrict our attention to the case where $1/n_1 + 1/n_2 + 1/n_3 < 1/2$. Noting that $E(n_1, n_2, n_3)$ is increasing in each n_i, to find the minimal excess we can simply check the smallest possible triples satisfying our conditions, namely $(6, 6, 7)$, $(5, 8, 8)$, $(4, 8, 10)$, and $(4, 6, 14)$. The table below shows the excess for each of these, along with the standard Thurston model $(7, 7, 7)$.

(n_1, n_2, n_3)	$E(n_1, n_2, n_3)$
$(7, 7, 7)$	$\pi/3$
$(6, 6, 7)$	$\pi/10$
$(5, 8, 8)$	$2\pi/9$
$(4, 8, 10)$	$2\pi/19$
$(4, 6, 14)$	$2\pi/41$

Thus, the smallest excess comes with the choice $(4, 6, 14)$, shown in FIGURE 16 (FIGURE 17 gives a schematic you can copy to construct your own model). Although this initially looks quite different from the Thurston model, it actually arises from its barycentric subdivision. That is, this triangle corresponds to the minimal triangle we saw before in the Thurston model!

The arguments that we gave before for the standard Thurston model carry over directly to the new model, thus we have

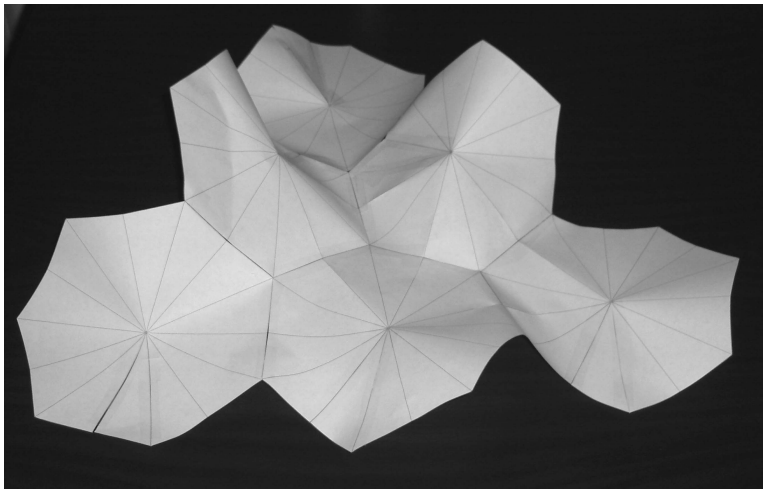


Figure 16 The (4, 6, 14) general Thurston model

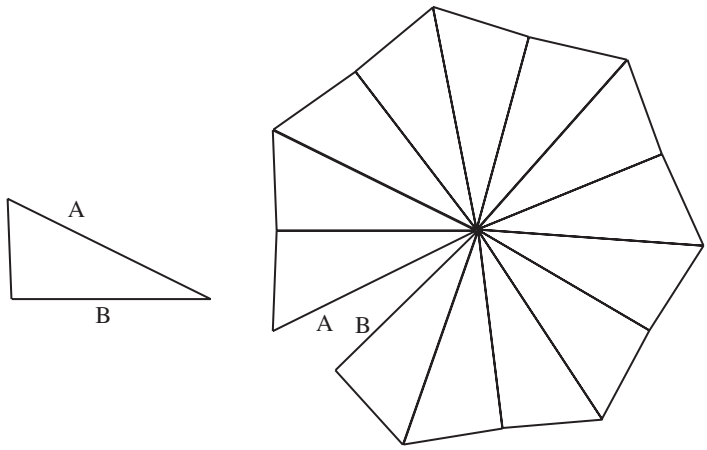


Figure 17 Schematic for the (4, 6, 14) general Thurston model, with the 14th triangle around the vertex to be pasted in along edges A and B

PROPOSITION 5. *In the general Thurston (n_1, n_2, n_3) space, any Thurston triangle $\triangle ABC$ has angle sum equal to $\pi - E(n_1, n_2, n_3)V_i$ where V_i denotes the number of model vertices in the interior of $\triangle ABC$. Moreover, if we take a special hyperbolic triangle with model vertices, then the angle sum is $\pi - (1/2)E(n_1, n_2, n_3)(2V_i + V_b - 3)$ where V_i is the number of model vertices lying in the interior of the triangle and V_b is the number of model vertices on the boundary (including the vertices of the triangle).*

So, in the (4, 6, 14)-model it is possible to draw a Thurston triangle containing as many as 20 model vertices.

We also have an analog of Pick's theorem for the general models:

PROPOSITION 6. *Let R be a region bounded by (n_1, n_2, n_3)-model hyperbolic lines, and again let V_i denote the number of internal model vertices and V_b denote the number of model vertices on the boundary of R. Then the area of region R is given by*

$$Area(R) = (2V_i + V_b - 2)\beta,$$

where β is the area of the model triangle.

Gauss-Bonnet Theorem

We can put Propositions 5 and 6 together to get a special case of the Gauss-Bonnet formula, one of the most important theorems in differential geometry. Recall that the Gauss-Bonnet formula states that the area A of a triangle in a surface of constant curvature κ is given by the formula

$$-\kappa A = (\pi - a_1 - a_2 - a_3),$$

where a_1, a_2, and a_3 denote the measurements of the interior angles of the triangle. We will derive a similar formula relating the area and angle sum of a special hyperbolic triangle in the (n_1, n_2, n_3)-model, whose vertices are all model vertices. This is particularly useful for the (4, 6, 14)-model (or any model where n_1, n_2, n_3 are all distinct), where *all* special hyperbolic triangles have model vertices.

From Proposition 6, the area of a special hyperbolic triangle with model vertices is $A = (2V_i + V_b - 2)\beta$, where β is the area of the model triangle. On the other hand, by Proposition 5, the sum of the angles a_1, a_2, and a_3 of the triangle is given by

$$\begin{aligned} a_1 + a_2 + a_3 &= \pi - \frac{1}{2}E(n_1, n_2, n_3)(2V_i + V_b - 3) \\ &= \pi - \frac{1}{2}E(n_1, n_2, n_3)\left(\frac{A}{\beta} - 1\right) \\ &= \pi - \frac{E(n_1, n_2, n_3)}{2\beta}(A - \beta). \end{aligned}$$

It now follows that we can write the area A in terms of the angles of the special hyperbolic triangle, the area of a model triangle, and the angle excess $E(n_1, n_2, n_3)$. Specifically,

$$A = \beta + \frac{2\beta}{E(n_1, n_2, n_3)}(\pi - a_1 - a_2 - a_3),$$

where a_1, a_2, and a_3 are the measures of the angles of the triangle. Here the curvature of the model is approximated by the angle excess $E(n_1, n_2, n_3)$, which corresponds to our observation that reducing the angle excess results in a flatter model.

We can derive a formula even closer to the Gauss-Bonnet formula by introducing a new variable α_i, defined below (for brevity, we let $E = E(n_1, n_2, n_3)$):

$$a_i = \left(1 + \frac{E}{2\pi}\right)\alpha_i \qquad \text{or} \qquad \alpha_i = \frac{a_i}{1 + \frac{E}{2\pi}}.$$

Then our expression for the area becomes

$$\begin{aligned} A &= \beta + \frac{2\beta}{E}(\pi - a_1 - a_2 - a_3) \\ &= \beta + \frac{2\beta}{E}\left(\pi - \left(1 + \frac{E}{2\pi}\right)(\alpha_1 + \alpha_2 + \alpha_3)\right) \\ &= \beta\left(1 + \frac{2\pi}{E} - \left(\frac{2}{E} + \frac{1}{\pi}\right)(\alpha_1 + \alpha_2 + \alpha_3)\right) \end{aligned}$$

$$= \beta \left(\frac{2}{E} + \frac{1}{\pi} \right) (\pi - \alpha_1 - \alpha_2 - \alpha_3)$$

$$= \beta \left(\frac{2\pi + E}{\pi E} \right) (\pi - \alpha_1 - \alpha_2 - \alpha_3).$$

We have proved the following analogue of the Gauss-Bonnet Theorem:

PROPOSITION 7. *Consider a special hyperbolic triangle with model vertices in the general Thurston (n_1, n_2, n_3)-model, where β is the area of a model triangle. Say that the triangle has area A and angles a_1, a_2, a_3. Then*

$$-\kappa A = (\pi - \alpha_1 - \alpha_2 - \alpha_3),$$

where

$$\alpha_i = \frac{a_i}{1 + \frac{E}{2\pi}}, \qquad \kappa = -\frac{\pi E}{\beta(2\pi + E)} \quad \text{and} \quad E = E(n_1, n_2, n_3).$$

How can we interpret α_i and κ? If we consider the preimage of our special hyperbolic triangle in the Poincaré model of hyperbolic space (as described earlier), then the preimage of an angle a at one of the model vertices (with angle excess $E = E(n_1, n_2, n_3)$) is exactly

$$\alpha = a \frac{2\pi}{2\pi + E} = \frac{a}{1 + \frac{E}{2\pi}}$$

This means that α_i is just the true hyperbolic angle corresponding to the angle a_i at a model vertex.

To understand κ, consider the preimage of a model triangle in the hyperbolic surface of constant curvature -1. This triangle has angles $2\pi/n_1, 2\pi/n_2, 2\pi/n_3$, so by the classical Gauss-Bonnet Theorem, its area γ is

$$\gamma = -\left(\frac{2\pi}{n_1} + \frac{2\pi}{n_2} + \frac{2\pi}{n_3} - \pi \right)$$

$$= \pi - \left(\frac{n_1 n_2 + n_2 n_3 + n_1 n_3}{n_1 n_2 n_3} \right) 2\pi$$

$$= \pi - \frac{\pi}{E + 2\pi} 2\pi$$

$$= \frac{E\pi + 2\pi^2 - 2\pi^2}{E + 2\pi} = \frac{E\pi}{E + 2\pi}.$$

Then $\kappa = -\gamma/\beta$ measures the ratio of the area of the preimage of the model triangle in the surface with constant curvature -1 to the area of the model triangle. As the model triangle gets larger (so the model is flatter), κ will get closer to 0, so κ is a reasonable measure of the curvature of the model. Moreover, as the angle excess E shrinks, κ will also get smaller. This means that the (4, 6, 14)-model, with the smallest angle excess, gives a significantly flatter model, in which it is easier to follow the hyperbolic lines and illustrate the Gauss-Bonnet theorem.

In a college geometry class, we have used these models to introduce students to curvature and the Gauss-Bonnet theorem, without any of the difficult differential geometry required to prove the full Gauss-Bonnet Theorem. Students can be led to discover for themselves one of the greatest theorems of mathematics, starting from no more than paper triangles and tape!

REFERENCES

1. E. Burger and M. Starbird, *The Heart of Mathematics: An Invitation to Effective Thinking*, 2nd ed., Key College Publishing, Emeryville, CA, 2005.
2. M. J. Greenberg, *Euclidean and Non-Euclidean Geometries: Development and History*, W. H. Freeman, New York, 2008.
3. B. Grünbaum and G. C. Shephard, Pick's theorem, *Amer. Math. Monthly* **100** (1993) 150–161. doi:10.2307/2323771
4. I. Niven and H. S. Zuckerman, Lattice points and polygonal area, *Amer. Math. Monthly* **74** (1967) 1195–1200. doi:10.2307/2315660
5. J. Weeks, *The Shape of Space*, 2nd ed., CRC Press, Boca Raton, FL, 2001.

Summary In looking at a common physical model of the hyperbolic plane, the authors encountered surprising difficulties in drawing a large triangle. Understanding these difficulties leads to an intriguing exploration of the geometry of the Thurston model of the hyperbolic plane. In this exploration we encounter topics ranging from combinatorics and Pick's Theorem to differential geometry and the Gauss-Bonnet Theorem.

CURTIS D. BENNETT is a professor of mathematics at Loyola Marymount University in Los Angeles. He received his Ph.D. in mathematics at the University of Chicago in 1990 under the direction of George Glauberman and Mark Ronan. His mathematical research runs from the study of groups and geometries to the scholarship of teaching and learning in mathematics. He was awarded the Deborah and Franklin Tepper Haimo award for Distinguished College or University Teaching of Mathematics in 2010. In his spare time, he likes hiking in the mountains.

BLAKE MELLOR earned his Ph.D. in mathematics from U.C. Berkeley in 1999, under the direction of Robion Kirby. He is currently an associate professor of mathematics at Loyola Marymount University in Los Angeles. His research interests are in knot theory and spatial graphs, which have yet to lead him to a better way to tie his shoelaces. He enjoys martial arts, ballroom dancing, science fiction and playing with his son Eric.

PATRICK D. SHANAHAN is a professor of mathematics at Loyola Marymount University in Los Angeles. He received his Ph.D. in mathematics at U.C. Santa Barbara in 1996, under the supervision of Daryl Cooper. His main area of research is geometric topology, with an emphasis on knot theory. He is also a co-author of the textbook, *A First Course in Complex Analysis with Applications*, currently in its second edition. In his spare time you can find him at the beach surfing with his children Kasey and Cody.

The Multiplication Game

KENT E. MORRISON
California Polytechnic State University
San Luis Obispo, CA 93407
kmorriso@calpoly.edu

The game

You walk into a casino, and just inside the main entrance you see a new game to play—the *Multiplication Game.* You sit at a table opposite the dealer and place your bet. The dealer hits a button and from a slot in the table comes a slip of paper with a number on it that you cannot see. You use a keypad to choose a number of your own—any positive integer you like, with as many digits as you like. Your number is printed on the slip of paper along with the product of the two numbers. The dealer shows you the slip so that you can verify that the product is correct. You win if the first digit of the product is 4 through 9; you lose if it is 1, 2, or 3. The casino pays even odds: for a winning bet of one dollar the casino returns your dollar and one more. Should you stay and play?

It looks tempting. You you have six winning digits and the casino has only three! But being skeptical, you take a few minutes to calculate. You write the multiplication table of the digits from one to nine. Of the 81 products you see that 44 of them begin with 1, 2, or 3, and only 37 begin with 4 through 9. Suddenly, even odds do not seem so attractive! You abandon the game and walk further into the casino.

In the next room you find another table with the same game, but better odds. This table pays \$1.25 for a winning one dollar bet. From your previous count you figure that if the odds favor the casino by 44 : 37, then a fair payout would be 44/37 dollars for a dollar bet; that is almost \$1.19, and this table is offering more. Should you stay and play?

You open your laptop and write a computer program to count the products of the two digit numbers from 10 to 99. (You realize immediately that in this game multiplying by $1, 2, \ldots, 9$ is the same as multiplying by $10, 20, \ldots, 90$, and so you leave out the one digit numbers.) You find that of these 8100 products the casino has 4616 winners and you have 3484. The ratio 4616/3484 is between 1.32 and 1.33, quite a bit more than the \$1.25 being offered. You move on, heading to the back of the casino where you might find the best odds.

Far back in a dark corner you find a high stakes table with the Multiplication Game. This one offers to pay \$1.40 for a winning dollar bet with a minimum bet of \$100. Now you take a little time to think it over. You run your computer program to multiply all the three digit numbers between 100 and 999 and find that 461698 of the products are winners for the casino and 348302 are winners for you. The ratio 461698/348302 is 1.32557 to five decimal places. (The limit of this process, as the number of digits increases, turns out to be about 1.32565.) The odds look good, so you stay to play.

You pick three digit numbers randomly. You win some and you lose some, but after a hundred rounds you find yourself \$450 poorer. Why are you losing? Obviously the casino is not choosing its numbers in the same way you are. If it were, you would be ahead about \$320 by now. You wisely decide it is time to take a break from the table and analyze the game more thoroughly.

Math. Mag. **83** (2010) 100–110. doi:10.4169/002557010X482862.

Applying game theory

The Multiplication Game was first described and analyzed by B. Ravikumar [**10**] as a two-person game in which the players choose n-digit integers for a fixed n. He determined the limit of the optimal strategy as n goes to infinity.

In this article we have modified the game to allow positive integers of any length. Note that it differs from most actual casino games (like blackjack or roulette) in that both you and the casino can play strategically. The outcomes of all possible simultaneous choices of the two players can be represented by an infinite matrix $A = (a_{ij})$ whose rows and columns are indexed by the positive integers. The rows of the matrix correspond to your choices and the columns correspond to the choices of the casino. Looking at the outcome from your point of view, we put 1 into the matrix at the ij location if you win and 0 if the casino wins.

Clearly it is not in either player's interest to choose the same number every time. In the lexicon of game theory this would be called a *pure strategy.* Instead the players must use *mixed strategies,* which are probability distributions on the set of positive integers.

More formally, a mixed strategy is a probability vector $p = (p_1, p_2, \dots)$, where each p_i is non-negative and $\sum_i p_i = 1$. It may also be helpful to think of p as a linear combination of pure strategies, $p = \sum_i p_i \delta_i$, where δ_i is the pure strategy of choosing i with probability 1. Thus, δ_i is the standard basis vector having 1 in the ith location and 0 everywhere else.

Let $f(p, q)$ denote the probability that you win when the you use the mixed strategy p and the casino uses the mixed strategy q. Then

$$f(p, q) = \sum_{i,j} a_{ij} p_i q_j.$$

(This is a doubly infinite sum, as each of i and j can be any positive integer.) You seek to maximize your chance of winning, while the casino seeks to minimize it. That is, you would like to find p so that $f(p, q)$ is as high as possible while the casino chooses q to make it as low as possible.

(We are using $f(p, q)$ to represent a probability, rather than an expected profit, because we want to be flexible about the payoffs. Your average profit per round at the high stakes table is

$$(+140) f(p, q) + (-100)(1 - f(p, q)).$$

The adjustment required for other stakes is apparent.)

This leads to the standard definitions of the *value* of the game to each player. The value of the game to you is

$$v_1 = \sup_p \inf_q f(p, q),$$

and the value of the game to the casino is

$$v_2 = \inf_q \sup_p f(p, q).$$

(To understand these expressions, first note that $\inf_q f(p, q)$ is the worst thing that can happen to you if you choose strategy p. Therefore, v_1 is the best winning probability that you can guarantee for yourself, independent of the casino's choice. Similarly, v_2 is the best result—lowest winning probability for you—that the casino can guarantee for itself.)

Now for any real valued function $f(u, v)$ whose domain is a Cartesian product $U \times V$,

$$\sup_u \inf_v f(u, v) \leq \inf_v \sup_u f(u, v).$$

We leave the proof as an exercise for the reader. It follows that $v_1 \leq v_2$.

For a finite game (one in which the sets of pure strategies for both players are finite) we can apply the Minimax Theorem of von Neumann and Morgenstern. It states that

(a) the values are equal; that is, $v_1 = v_2 = v$; and

(b) there are *optimal strategies* p and q such that $v = \inf_q f(p, q) = \sup_p f(p, q)$.

That means that if you choose p and the casino chooses q, the result v is guaranteed.

Unfortunately, the multiplication game is not a finite game. For infinite games the Minimax Theorem is not true in general; there are examples of infinite games with $v_1 \neq v_2$. In the analysis that follows we will show that the multiplication game is like a finite game, at least to the extent that $v_1 = v_2$.

The analysis

The analysis proceeds in steps.

First we observe that there is some redundancy in the set of pure strategies. For example, the integers 21, 210, 2100, ... all give the same results when chosen by either player. Thus, in the payoff matrix the rows representing these pure strategies are identical, and so we can eliminate all but one of these rows. We can do the same with the columns indexed by these equivalent choices.

Next we observe that the choices don't really need to be integers. A player could choose 2.1 with the same result as choosing 21. We can define a reduced set of pure strategies X as the set of rationals in [1, 10) having a terminating decimal expansion:

$$X = \left\{ \sum_{k=0}^{m} d_k 10^{-k} \mid m \geq 0, d_k \in \{0, 1, \ldots, 9\}, d_0 \neq 0 \right\}.$$

Although the players are no longer choosing integers, we have not changed the game in any essential way, and we now have an efficient description of the infinite sets of pure strategies in which there is no redundancy. The players choose terminating decimals in the interval [1, 10) and the casino wins if the product begins with the digits 1, 2, or 3.

Next we modify the game again, this time possibly in an essential way. We enlarge the sets of pure strategies to include all the real numbers in [1, 10). Although this increases the cardinality of the set of pure strategies from countable to uncountable, the new game is easier to analyze and its solution will be essential in understanding the original game. In a later section we will return to this modification to see whether it affects our results.

So now the casino chooses x and you choose y in [1, 10) with the casino winning if the first digit of xy is 1, 2, or 3. That is, the casino wins if

$$1 \leq xy < 4 \text{ or } 10 \leq xy < 40.$$

The points (x, y) satisfying these inequalities make up the shaded region in FIGURE 1. The white region includes the points for which you win.

Finally, we straighten out the shaded region by using the logarithms of the numbers rather than the numbers themselves. We let the casino choose $a = \log_{10} x$ and

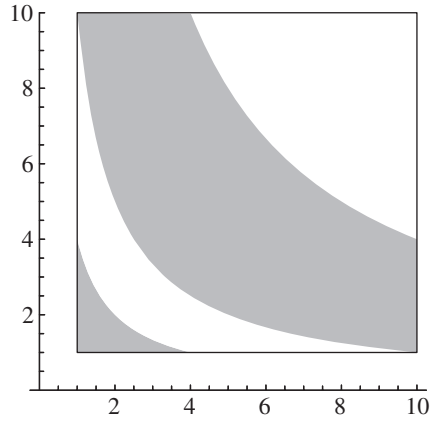


Figure 1 Winning region for the casino

you choose $b = \log_{10} y$. Clearly, choosing a and b is equivalent to choosing x and y. (However, as we will see, mixed strategies can look very different when described in terms of a and b.) Now the casino wins when

$$0 \le a + b < \log_{10} 4 \text{ or } 1 \le a + b < 1 + \log_{10} 4,$$

which is equivalent to

$$(a + b) \bmod 1 \in [0, \log_{10} 4).$$

FIGURE 2 is a complete picture of the game—essentially the payoff matrix. You choose a horizontal line and independently the casino chooses a vertical line. You win if the point of intersection is in the white region.

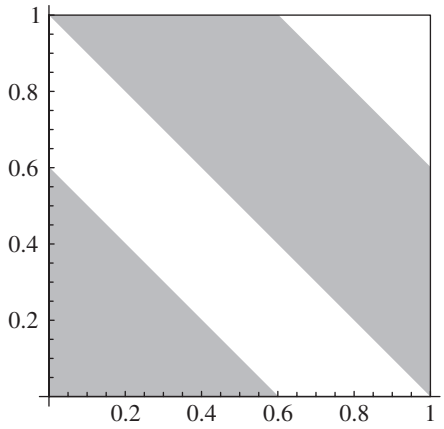


Figure 2 Winning region for the casino

What are mixed strategies now that the players have an uncountable number of choices? They should be probability distributions on the set [0, 1) of pure strategies.

There are many ways to describe probability distributions on an interval. Density functions and cumulative distribution functions (cdf's) come to mind. Most formally, a probability distribution is any measure on the interval, which is a non-negative function

defined on a suitable collection of subsets of $[0, 1)$. The measure must be countably additive and assign the value 1 to the whole interval. The mixed strategies we need can be described in terms of density functions, while pure strategies are discrete distributions concentrated at single points.

Solving the (modified) game

In order to motivate the solution of this game, consider for a moment an analogous finite game. Looking at FIGURE 2 one can see that the same proportion of each horizontal line lies within the shaded region, and the same is true for the vertical lines. A finite game with a similar structure is one in which both players have the same set of pure strategies $\{1, \ldots, n\}$; all entries in the payoff matrix are 1 or 0, and there are the same number of 1's in each row and in each column. The payoff matrix, then, is something like a discrete version of FIGURE 2 with the location of the 1's playing the role of the white region.

For a game with these properties we claim that an optimal strategy for both players is the uniform probability distribution

$$(1/n, \ldots, 1/n).$$

To see this, let p be uniform. No matter which j the column player chooses, the row player wins with probability

$$\sum_i a_{ij} p_i = \sum_i a_{ij}(1/n) = \frac{1}{n} \sum_i a_{ij} = \frac{c}{n},$$

where c is the number of 1's in each row. Thus, $v_1 \geq c/n$. On the other hand, if the column player uses the uniform strategy for q and the row player chooses any row i, then the row player wins with the same probability

$$\sum_j a_{ij} q_j = \sum_j a_{ij}(1/n) = \frac{1}{n} \sum_j a_{ij} = \frac{c}{n},$$

and thus $v_2 \leq c/n$. Since $v_1 \leq v_2$, this proves that $v_1 = v_2$ and from that it follows that the uniform probability distributions are optimal.

With this analogy to guide us, consider what happens when you choose your number uniformly in $[0, 1)$. This means your mixed strategy is the uniform distribution on this interval, which we denote by λ. Assume that the casino uses the pure strategy δ_a. Thus, the outcome is on the vertical line $\{(a, b) | b \in [0, 1)\}$ and the probability that the point (a, b) is in the shaded region is $\log_{10} 4$. That is the casino's probability of winning. Your probability of winning is $1 - \log_{10} 4$. This probability is independent of a, and so $\sup_a f(\delta_a, \lambda) = 1 - \log_{10} 4$. Since you can guarantee that the casino does not win with a probability greater than $\log_{10} 4$, we know that the value of the game to you satisfies $v_1 \geq 1 - \log_{10} 4$.

Now we look at the game from the casino's point of view and reason in the same way. No matter what pure strategy δ_b that you employ, the casino can choose its number uniformly and win with probability $\log_{10} 4$. Thus, the casino can guarantee winning with at least this probability no matter what you do, and so $v_2 \leq 1 - \log_{10} 4$. Since $v_1 \leq v_2$, we see that they are in fact equal, and it follows that the uniform distribution is optimal for both players. These are the conclusions that the MiniMax Theorem would have given us if it had been applicable. Since $\log_{10} 4 \approx 0.60206$, the casino will win just over 60% of the time, which makes the correct odds a bit higher than 3:2. To

make the game fair the casino should pay you a bit more than \$1.50 for your winning one-dollar bet. The exact amount is

$$\frac{\log_{10} 4}{1 - \log_{10} 4} \approx 1.5129,$$

but the casino was only paying \$1.40 in the high stakes game, and so you found yourself losing after playing for a while.

Now we transfer the uniform distribution λ on the logarithms in $[0, 1)$ back to a distribution on $[1, 10)$ that we denote by β. Thus β assigns to the interval between x_1 and x_2 the probability $\log_{10} x_2 - \log_{10} x_1$. This probability distribution β has the density function

$$f(x) = \frac{1}{\ln 10} \frac{1}{x}$$

shown in FIGURE 3. The area between x_1 and x_2 and lying under the graph of f gives the probability that x is between x_1 and x_2.

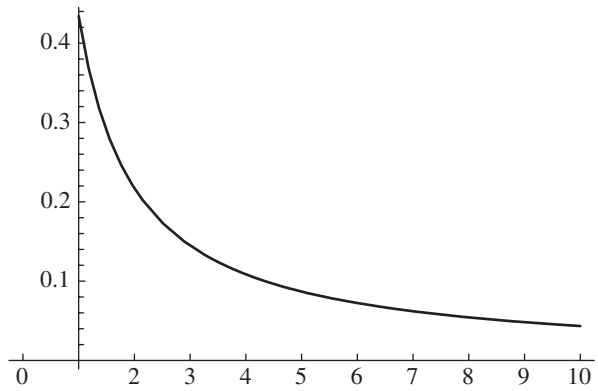


Figure 3 Plot of the Benford density $f(x) = \frac{1}{(\ln 10)x}$

In recent years this logarithmic probability distribution has become known as the *Benford distribution*, named for the physicist Frank Benford who investigated the relative frequency of leading digits of more than 20,000 numbers in several datasets from diverse sources such as populations of cities and the areas of river basins. Although Benford described the phenomenon over 70 years ago [**1**], the first discovery was actually due to the astronomer and mathematician Simon Newcomb who observed the phenomenon more than half a century earlier [**8**]. Newcomb's paper did not spark any further work in the years that followed. Benford, however, was fortunate to have his paper appear just in front of an influential paper in modern physics having Hans Bethe as one of the authors, and so it was widely seen by other scientists. This helped to attach Benford's name to the empirical observations and began what has become a small industry; there are now more than 700 papers in a comprehensive Benford online bibliography [**2**]. (Naming the phenomenon after Frank Benford illustrates *Stigler's Law of Eponymy*, which states that no scientific discovery is named for its discoverer. Appropriately, Stigler attributes his eponymous law to Robert Merton.) Benford's Law describes not just the distribution of the first significant digit but also the distribution of all significant digits. In its general form the law is the logarithmic distribution on the set of real numbers between 1 and 10, and even more generally the number base

can be any positive integer $b \geq 2$, in which case the Benford distribution is supported on interval $[1, b)$. The special cases concerning any particular significant digits can be derived from the continuous distribution. For enlightening accounts of Benford's Law we recommend the articles by Raimi [**9**], Hill [**5, 6**], and Fewster [**3**] or the statistics text by Larsen and Marx [**7**].

What happens if we change the game so that the casino wins if the first digit is prime? Or change the game so that the winner is determined by the value of the second digit of the product? To what extent can we solve the game, i.e., determine the optimal strategies and the value of the game, when the winning conditions are changed?

Let $W \subset [1, 10)$ be the *winning set* for the casino; the casino wins when the product xy lands in W. For the version we have analyzed the winning set W is the interval $[1, 4)$, consisting of the numbers whose first digit is 1, 2, or 3. Now look at the logs of the numbers in W and call that set Z. If the logs are a and b, then the casino wins when $a + b \pmod 1$ is in Z. Now plot the region in the unit square consisting of the points (a, b) for which the casino wins. See FIGURE 4 for the case in which the casino wins when the first digit is prime. In general this region consists of bands between lines having slope -1. Each horizontal line and each vertical line meets the shaded region in the same proportion. The points on the horizontal axis that are in the shaded region are $(a, 0)$ where $a \in Z$. Thus, the proportion of each line that lies in the shaded region is $\lambda(Z)$, the Lebesgue measure of Z, which is the same as $\beta(W)$, the Benford measure of W. Using the same reasoning as before we conclude that if the casino chooses its logarithm uniformly, then it will win with probability $\lambda(Z)$ regardless of what you do, and if you choose your logarithm uniformly you will win with probability equal to $1 - \lambda(Z)$.

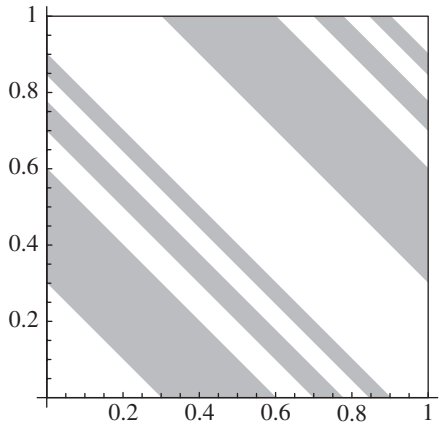


Figure 4 Winning region for the casino for prime first digit

Just how far can we push this approach? Certainly W (or Z) can be any finite union of intervals but we can even allow countable unions of intervals; it does not matter whether they are open, half-open, or closed. But non-measurable sets are too strange to be used for winning sets, because if Z is not Lebesgue measurable, then we cannot make sense of the statement that every horizontal and vertical line meets the set $\{(a, b) \mid a + b \pmod 1 \in Z\}$ in the same proportion. We summarize this discussion in the following theorem.

THEOREM 1. *Let the casino's winning set $W \subset [1, 10)$ be a finite or countable union of intervals. Then an optimal mixed strategy for both you and the casino is*

to choose your numbers from the Benford distribution on $[1, 10)$*, or, equivalently, to choose their logarithms uniformly in* $[0, 1)$*. The probability that the casino wins is* $\beta(W)$*.*

Solving the original game

Now we return to the analysis of the original game in its reduced form in which you and the casino choose numbers in the set X of terminating decimals in $[1, 10)$. We will show that by approximating the Benford distribution we can find mixed strategies that come arbitrarily close to being optimal. That is, we will show that

$$\sup_p \inf_q f(p, q) = \inf_q \sup_p f(p, q) = \log_{10} 4,$$

but we will not actually find strategies that attain this value.

Let X_n be the subset of X whose elements have a terminating expansion with n digits,

$$X_n = \left\{ \sum_{k=0}^{n-1} d_k 10^{-k} \mid d_k \in \{0, 1, \ldots, 9\}, d_0 \neq 0 \right\}.$$

Consider the following strategy. Fix a positive integer n. Generate a random number $a \in [0, 1)$, compute 10^a and then choose $x \in X$ to be the nearest n-digit number less than or equal to 10^a. This defines a probability distribution β_n that is concentrated on the finite set X_n. The probability mass at the point $x \in X_n$ is given by

$$\beta_n\{x\} = \log_{10}(x + 1/10^{n-1}) - \log_{10} x.$$

Let $F(x) = \log_{10} x$ be the cumulative distribution function (cdf) of β and F_n the cdf of β_n. Then F_n has jumps at the points in X_n and is always greater than or equal to F. (See FIGURE 5.) The maximum difference between F_n and F occurs at $x = 1$, where

$$F_n(1) - F(1) = \log_{10}(1 + 1/10^{n-1}) - 0 < \frac{1}{10^{n-1}}. \tag{1}$$

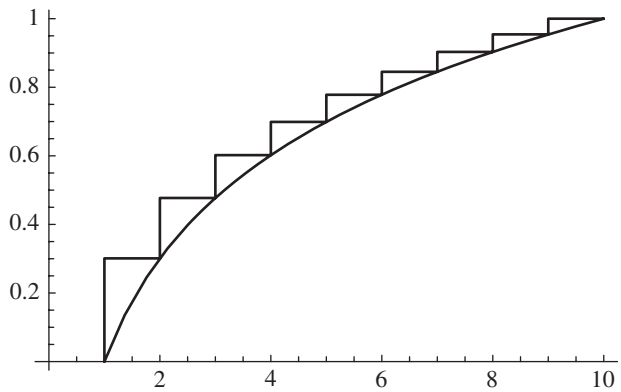


Figure 5 The cdf's of β_1 and β

Now if the casino uses β_n for its mixed strategy and you play $y \in X$ (or, for that matter, any $y \in [1, 10)$), then the casino will win with probability $\beta_n(V_y)$ where

$$V_y = \{x \in X \mid xy \in [1, 4) \cup [10, 40)\}$$

If $\epsilon > 0$, then for n sufficiently large we have $|\beta_n(V_y) - \beta(V_y)| < \epsilon$, and this estimate holds for all y because V_y is either an interval or a union of two intervals in $[1, 10)$ and the measures of V_y can be expressed with the cdf's F_n and F evaluated at the endpoints of the intervals where they differ by at most 10^{n-1} according to (1). Therefore, the casino can guarantee a win with probability at least $\log_{10} 4 - \epsilon$, and so $v_1 \geq \log_{10} 4 - \epsilon$.

Similar reasoning shows you can use β_n for n sufficiently large and guarantee that the casino wins with probability no more than $\log_{10} 4 + \epsilon$. Therefore, $v_2 \leq \log_{10} 4 - \epsilon$. These inequalities hold for all ϵ, and thus $v_1 = v_2 = \log_{10} 4$. It is worth repeating that we have not found an optimal strategy that actually achieves the value but rather a family of strategies that come arbitrarily close to optimal. It is tempting to consider the limit of the β_n, which is β, as an optimal strataegy, but β is not a probability distribution on X. If an optimal strategy μ exists as a probability distribution on X, then it must have the property that $\mu(V_y) = \log_{10} 4$ for all $y \in X$. There is no evident way to produce such a measure even if one exists.

The group game

How far we can generalize the multiplication game? Although there may be other paths to follow, we will assume that there is a binary operation that combines the players' choices to produce the result. Does the operation need to be associative? Commutative? Have an identity? Inverses? And what about the nature of the set on which the operation is defined?

In the original game we use the positive integers with multiplication. The operation is associative, commutative, and has an identity, but does not have inverses. The same is true of the reduced version with the set X, where after multiplying we move the decimal point if necessary to get a number between 1 and 10. However, by extending the set of pure strategies to be all real numbers between 1 and 10, we get inverses, and the result is that we have a group. This group is nicely described as the quotient group $\mathbf{R}_+/\langle 10\rangle$, where $\mathbf{R}_+$ is the multiplicative group of positive real numbers and $\langle 10\rangle$ is the subgroup generated by 10 consisting of all integral powers of 10. The numbers in $[1, 10)$ are unique coset representatives of the subgroup $\langle 10\rangle$. With logarithms we use the set $[0, 1)$ with addition mod 1, which is another way to describe the quotient group $\mathbf{R}/\mathbf{Z}$, where $\mathbf{R}$ is the additive group of real numbers and $\mathbf{Z}$ is the integer subgroup. An isomorphism from $\mathbf{R}/\mathbf{Z}$ to $\mathbf{R}_+/\langle 10\rangle$ is given by $a \mapsto 10^a$. For the $\mathbf{R}/\mathbf{Z}$ game we proved that Lebesgue measure is optimal, and for the $\mathbf{R}_+/\langle 10\rangle$ game it is the Benford measure that is optimal. These measures are special for their respective groups in that they are *invariant*. For Lebesgue measure it means that $\lambda(E) = \lambda(a + E)$ for a subset E of $[0, 1)$ and for $a \in [0, 1)$. For the Benford measure it means that $\beta(E) = \beta(xE)$ for $E \subset [1, 10)$ and $x \in [1, 10)$. (Addition and multiplication must be done in the quotient groups.) Furthermore, λ and β are probability measures, meaning that they are positive measures with total mass equal to one.

There is a class of groups having exactly the properties necessary to generalize Theorem 1, namely the class of *compact topological groups*. Among these groups are the groups we have just described, $\mathbf{R}/\mathbf{Z}$ and $\mathbf{R}_+/\langle 10\rangle$, both of which are abelian and topologically equivalent to circles. Also, every finite group (abelian or not) is a compact topological group with its discrete topology. For infinite non-abelian examples there

are the groups of isometries of $\mathbf{R}^n$ for $n \geq 2$. By contrast, topological groups that are not compact include the real numbers under addition, the non-zero real numbers under multiplication, the invertible $n \times n$ matrices over $\mathbf{R}$ or over $\mathbf{C}$, and any infinite group with the discrete topology (such as the integers under addition).

In general, a topological group is a topological space G together with a continuous group operation $G \times G \to G : (g_1, g_2) \mapsto g_1 g_2$ and a continuous inverse map $G \to G : g \mapsto g^{-1}$. With compactness comes the existence of a unique invariant (under left and right multiplication) probability measure λ known as *Haar measure* [**4**]. For finite groups Haar measure is simply normalized counting measure, whereas for $\mathbf{R}/\mathbf{Z}$ it is Lebesgue measure and for $\mathbf{R}_+/\langle 10\rangle$ it is the Benford measure.

With a compact topological group G we can generalize Theorem 1 as follows. Let $W \subset G$ be a λ-measurable subset. The casino chooses $x \in G$, you choose $y \in G$, and the casino wins if xy is in W. Then for both players an optimal mixed strategy is to use Haar measure λ, and the value of the game, i.e., the probability that the casino wins, is $\lambda(W)$. The proof, the details of which will be omitted, uses the invariance properties of λ to show that when the casino uses λ it does not matter what strategy you use, and when you use λ it does not matter what the casino does.

If any of the hypotheses are relaxed, then we do not have complete solutions to the game. The analysis becomes more difficult, something we have already seen with the original game played on the positive integers or in the equivalent version on the terminating decimals. Two important properties are lacking: inverses and compactness. We can exhibit mixed strategies arbitrarily close to optimal, and thus show that the game has a value, but we cannot exhibit strategies that actually achieve the optimum. Even if we add inverses to the set of terminating decimals by including all rational numbers in $[1, 10)$, the resulting group is not compact. It is countable and infinite and so it cannot carry an invariant probability measure because each element of the group would have the same non-zero mass and so the total mass would be infinite. We doubt that an optimal strategy exists but it is a question we leave unanswered.

Acknowledgment The author would like to thank B. Ravikumar, inventor of the multiplication game, for providing a copy of his paper describing and analyzing the game. The limiting distribution that he found is, of course, the Benford distribution.

REFERENCES

1. F. Benford, The law of anomalous numbers, *Proc. Amer. Phil. Soc.* **78** (1938) 551–572.
2. A. Berger and T. P. Hill, *Benford Online Bibliography*, `http://www.benfordonline.net`, 2009.
3. R. M. Fewster, A simple explanation of Benford's Law, *Amer. Statist.* **63** (2009) 26–32. doi:`10.1198/tast.2009.0005`
4. P. R. Halmos, *Measure Theory*, Springer Verlag, New York, 1974.
5. T. P. Hill, The significant digit phenomenon, *Amer. Math. Monthly* **102** (1995) 322–327. doi:`10.2307/2974952`
6. T. P. Hill, The first-digit phenomenon, *Amer. Sci.* **86** (1998) 358–363.
7. R. J. Larsen and M. L. Marx, *Introduction to Mathematical Statistics and Its Applications*, 4th ed., Prentice Hall, Upper Saddle River, NJ, 2005.
8. S. Newcomb, Note on the frequency of use of the different digits in natural numbers, *Amer. J. Math.* **4** (1881) 39–40. doi:`10.2307/2369148`
9. R. A. Raimi, The peculiar distribution of first digits, *Sci. Amer.* **221**(6) (June 1969) 109–119.
10. B. Ravikumar, A simple multiplication game and its analysis, preprint, 2007.

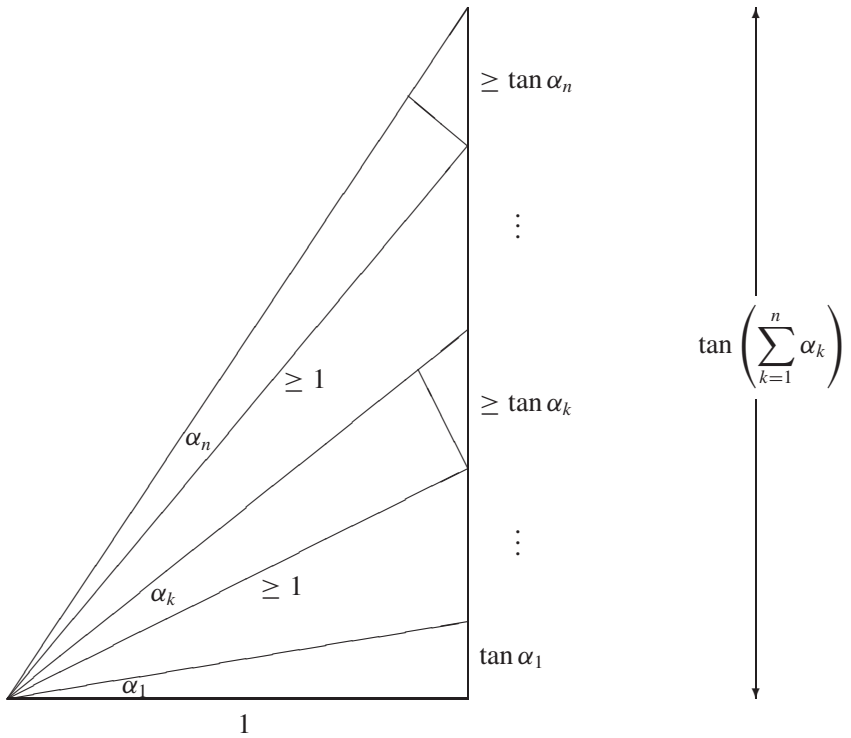
Summary The Multiplication Game is a two-person game in which each player chooses a positive integer without knowledge of the other player's number. The two numbers are then multiplied together and the first digit of the product determines the winner. Rather than analyzing this game directly, we consider a closely related game in which the players choose positive real numbers between 1 and 10, multiply them together, and move the decimal point, if necessary, so that the result is between 1 and 10. The mixed strategies are probability

distributions on this interval, and it is shown that for both players it is optimal to choose their numbers from the Benford distribution. Furthermore, this strategy is optimal for any winning set, and the probability of winning is the Benford measure of the player's winning set. Using these results we prove that the original game in which the players choose integers has a well-defined value and that strategies exist that are arbitrarily close to optimal. Finally, we consider generalizations of the game in which players choose elements from a compact topological group and show that choosing them according to Haar measure is an optimal strategy.

KENT E. MORRISON is professor emeritus at California Polytechnic State University in San Luis Obispo, where he taught for thirty years. He has also taught at Utah State University, Haverford College, and the University of California at Santa Cruz, where he received his Ph.D. and B.A. degrees. Currently he is a visiting researcher at the American Institute of Mathematics in Palo Alto. He has a number of research interests in the areas of algebra, geometry, probability, and combinatorics.

Proof Without Words: A Tangent Inequality

If $\alpha_k \geq 0$ for $k = 1, \ldots, n$ and $\sum_{k=1}^{n} \alpha_k < \pi/2$, then $\tan\left(\sum_{k=1}^{n} \alpha_k\right) \geq \sum_{k=1}^{n} \tan \alpha_k$.



—Rob Pratt
Raleigh, NC 27613-1079
Rob.Pratt@sas.com

Math. Mag. **83** (2010) 110. doi:10.4169/002557010X496407.

Mini-Sudokus and Groups

CARLOS ARCOS
GARY BROOKFIELD
MIKE KREBS
California State University, Los Angeles
Los Angeles, CA 90032-8204
gbrookf@calstatela.edu
mkrebs@calstatela.edu

By now you probably have at least a passing acquaintance with Sudoku, the pencil-and-paper puzzle that has, for the past few years, been displacing advice columns and word jumbles from the back pages of newspapers all over the world.

The rules are simple. One is given a 9×9 grid. Each cell in the grid is to be filled in with one of the digits from 1 to 9. Some of the cells have been filled in already, as in the example below.

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

The puzzler may not fill in the empty cells willy-nilly; he or she must obey the Rule of One, which requires that each row, each column, and each block (the 3×3 subgrids with thick borders) must contain every digit from 1 to 9 exactly once. To simplify our discussion we say that a 9×9 grid completely filled with the digits 1 to 9 such that the Rule of One holds is a *Sudoku*. (So the grid above, then, is *not* a Sudoku according to our definition, because not all of the cells have been filled in. Once all the cells have been filled in, *then* it's a Sudoku.)

The rules of Sudoku suggest many natural mathematical questions: *How do you construct these puzzles? How do you solve these puzzles? How many different Sudokus are there? How many of these are essentially different?* We call two Sudokus *essentially the same*, or *equivalent*, if you can get from one to the other in finitely many steps where a single step might be switching the first two columns, or rotating the grid ninety degrees, or relabeling entries (replacing every 2 with a 7 and every 7 with a 2, for example). We will make this notion of equivalence more precise in the sections that follow.

The answers to the questions above are known. Felgenhauer and Jarvis [**5**] found that there are 6,670,903,752,021,072,936,960 Sudokus. That's a big number. Also, Jarvis and Russell [**10**] found that there are 5,472,730,538 essentially different Sudokus. That's a smaller number. But it's still pretty darn big.

There is no need to limit oneself to 9×9 grids; any grid of size $n^2 \times n^2$ will do. Herzberg and Murty [**8**] use graph-theoretic techniques to provide an asymptotic estimate for the number of $n^2 \times n^2$ Sudokus.

Math. Mag. **83** (2010) 111–122. doi:10.4169/002557010X482871.

To make our discussion accessible to those without a background in graph theory, and to keep things on an order of magnitude that a human can more readily comprehend, we answer these questions about the much simpler, but still interesting, case of 4×4 Sudokus. We call these *mini-Sudokus*. Thus a mini-Sudoku is a 4×4 grid, for example,

1	2	3	4
3	4	1	2
2	1	4	3
4	3	2	1

,

such that the Rule of One holds: Each row, each column, and each block (the 2×2 subgrids with thick borders) contains every digit from 1 to 4 exactly once.

Our main tools come from group theory. In particular, the notion of groups acting on sets will enable us to define precisely what it means for two mini-Sudokus to be essentially the same. Undergraduate math majors will have seen these concepts in a first abstract algebra class and may find that applying newly-learned group theory methods to a familiar, concrete example brings the abstract theory to life.

Various sources discuss the mathematics of Sudoku in general [**3, 4, 7, 8**].

Counting mini-Sudokus

How many mini-Sudokus are there? They can be enumerated in many ways. One method is to consider first the four entries in the upper left 2×2 block. These entries must be 1, 2, 3, and 4, but they can be put in any order. This gives $4! = 24$ ways of filling this block. The reader should confirm that, once this block has been filled, for example,

1	2		
3	4		
		*	
			*

,

then all the other entries are determined by the Rule of One and the choice of the two entries marked $*$. These two entries are arbitrary except that they must be different, so there are $4 \cdot 3 = 12$ ways of choosing them.

Here are the 12 possible mini-Sudokus obtained by filling in the empty cells in the above example:

$\mathtt{A}_1 =$

1	2	3	4
3	4	1	2
2	1	4	3
4	3	2	1

$\mathtt{A}_2 =$

1	2	4	3
3	4	2	1
2	1	3	4
4	3	1	2

$\mathtt{A}_3 =$

1	2	4	3
3	4	2	1
4	3	1	2
2	1	3	4

$\mathtt{A}_4 =$

1	2	3	4
3	4	1	2
4	3	2	1
2	1	4	3

$\mathtt{B}_1 =$

1	2	3	4
3	4	2	1
2	1	4	3
4	3	1	2

$\mathtt{B}_2 =$

1	2	4	3
3	4	1	2
2	1	3	4
4	3	2	1

$\mathtt{B}_3 =$

1	2	4	3
3	4	1	2
4	3	2	1
2	1	3	4

$\mathtt{B}_4 =$

1	2	3	4
3	4	2	1
4	3	1	2
2	1	4	3

$$\mathtt{C}_1 = \begin{array}{|c|c|c|c|}\hline 1&2&3&4\\\hline 3&4&1&2\\\hline 2&3&4&1\\\hline 4&1&2&3\\\hline\end{array} \qquad \mathtt{C}_2 = \begin{array}{|c|c|c|c|}\hline 1&2&4&3\\\hline 3&4&2&1\\\hline 2&3&1&4\\\hline 4&1&3&2\\\hline\end{array} \qquad \mathtt{C}_3 = \begin{array}{|c|c|c|c|}\hline 1&2&4&3\\\hline 3&4&2&1\\\hline 4&1&3&2\\\hline 2&3&1&4\\\hline\end{array} \qquad \mathtt{C}_4 = \begin{array}{|c|c|c|c|}\hline 1&2&3&4\\\hline 3&4&1&2\\\hline 4&1&2&3\\\hline 2&3&4&1\\\hline\end{array}$$

We have labeled these $\mathtt{A}_1, \mathtt{A}_2, \ldots, \mathtt{C}_3, \mathtt{C}_4$ for future reference. Can the reader guess why we have labeled them in this manner?

Now we can calculate the number of mini-Sudokus. There are 24 ways of filling in the upper left 2×2 block, and, once that is done, there are 12 ways of filling in the rest of the grid. This gives a total of $24 \cdot 12 = 288$ different mini-Sudokus. (So, while the number of different 9×9 Sudokus—6,670,903,752,021,072,936,960—is excessively disgusting, the number of different mini-Sudokus is merely two gross!)

Row and column symmetries

Are the 12 mini-Sudokus listed above really that different from one another? After all, interchanging the last two columns of $\mathtt{A}_1$ gives $\mathtt{A}_2$. Similarly, interchanging the bottom two rows of $\mathtt{A}_2$ gives $\mathtt{A}_3$. Indeed, the mini-Sudokus $\mathtt{A}_1$, $\mathtt{A}_2$, $\mathtt{A}_3$, and $\mathtt{A}_4$ differ only by switching columns and/or rows. We would like to say that these mini-Sudokus are essentially the same or, using a more standard nomenclature, that they are equivalent. Similarly, we would like to say that the mini-Sudokus $\mathtt{B}_1$, $\mathtt{B}_2$, $\mathtt{B}_3$, and $\mathtt{B}_4$ are all equivalent (as are $\mathtt{C}_1$, $\mathtt{C}_2$, $\mathtt{C}_3$, and $\mathtt{C}_4$).

But are $\mathtt{A}_1$ and $\mathtt{B}_1$ equivalent? How about $\mathtt{A}_1$ and $\mathtt{C}_1$? Are *all* mini-Sudokus equivalent in some sense?

To answer these questions, we need to be precise about what *equivalent* means. And to do that, we have to understand the set of *mini-Sudoku symmetries*, by which we mean one-to-one onto functions from the set of all mini-Sudokus to itself. We have already mentioned that interchanging the bottom two rows in any given mini-Sudoku always yields another mini-Sudoku. So the operation of interchanging these two rows is a mini-Sudoku symmetry. If we give this symmetry the symbol ρ then $\rho(\mathtt{A}_1) = \mathtt{A}_4$, $\rho(\mathtt{A}_4) = \mathtt{A}_1$, $\rho(\mathtt{A}_2) = \mathtt{A}_3$, and so on. Interchanging the last two columns is also mini-Sudoku symmetry—call it σ.

Composing any two symmetries yields another symmetry. For example, interchanging the bottom two rows, followed by interchanging the last two columns, is also a mini-Sudoku symmetry, which we would write as $\sigma\rho$. The symmetry that leaves all mini-Sudokus unchanged is called the *identity symmetry* and denoted id. Every symmetry γ has an *inverse symmetry* γ^{-1} which undoes whatever the symmetry does, that is, $\gamma\gamma^{-1} = \gamma^{-1}\gamma = \mathrm{id}$. For example, $\rho^{-1} = \rho$ since switching the bottom two rows of a mini-Sudoku twice gives the original mini-Sudoku back. For any three mini-Sudoku symmetries α, β, γ, we have $(\alpha\beta)\gamma = \alpha(\beta\gamma)$, since function composition is associative. In short, the set of mini-Sudoku symmetries is a group.

Now we can explain equivalence. If K is a group of mini-Sudoku symmetries (that is, a subgroup of the set of all mini-Sudoku symmetries), then two mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ are *K-equivalent* if one can be obtained from the other by applying some symmetry in K, that is, $\mathtt{Y} = \gamma(\mathtt{X})$ for some $\gamma \in K$. Since K is a group, this is, in fact, an equivalence relation. The set of all mini-Sudokus that are K-equivalent to $\mathtt{X}$ is called the *K-equivalence class* containing $\mathtt{X}$. Every mini-Sudoku is contained in a unique K-equivalence class. We say that the group K *acts* on the set of mini-Sudokus. Texts by Gallian [**6**] and Rotman [**9**] are good places to learn more about groups acting on sets.

As we have already seen, given a mini-Sudoku, there are some easy ways to make a new mini-Sudoku from it. For example, we could switch the first row with the second

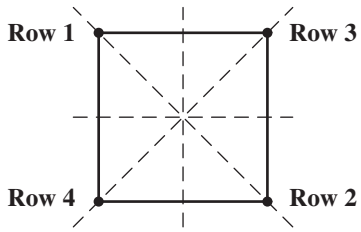
row, and leave the bottom two rows alone. Another example would be to send Row 1 to Row 3, Row 3 to Row 2, Row 2 to Row 4, and Row 4 to Row 1.

Since there are four rows in a mini-Sudoku, we can regard the set of such row symmetries as a subgroup R of the symmetric group S_4. However, not all row permutations are symmetries of mini-Sudokus. For example, the permutation taking Row 1 to Row 2, Row 2 to Row 3, and Row 3 to Row 1, leaving Row 4 unchanged, takes the mini-Sudoku $\mathtt{A}_1$ to

2	1	4	3
1	2	3	4
3	4	1	2
4	3	2	1

which is *not* a mini-Sudoku. Thus R is isomorphic to a proper subgroup of S_4. Which subgroup?

The answer is that R is isomorphic to the dihedral group D_4, the group of symmetries of a square. One way to see this is to draw a square and to label its vertices with the rows of the mini-Sudoku, as below. (Do not confuse this square with the mini-Sudoku grid itself—that comes later!)



The group D_4 consists of 8 symmetries: four rotations, by 0°, 90°, 180°, and 270°, and four reflections through the axes indicated by the dotted lines in the diagram.

Thus, switching the top two rows of a mini-Sudoku corresponds to reflecting the square about the diagonal axis through the vertices labeled Row 3 and Row 4. Rotation of the square by 90° clockwise corresponds to the mini-Sudoku symmetry that sends Row 1 to Row 3, Row 3 to Row 2, Row 2 to Row 4, and Row 4 to Row 1. The reader should check that each symmetry of the square corresponds to a mini-Sudoku row symmetry and vice versa. The isomorphism between R and D_4 is then transparent.

By replacing the word *row* with the word *column* in the above discussion, we get a new group C of mini-Sudoku column symmetries, again isomorphic to D_4. If $\mu \in R$ and $\nu \in C$, then applying μ and then ν to a mini-Sudoku gives the same result as applying first ν then μ. In other words, row symmetries commute with column symmetries. This means that, combining the 8 row symmetries with the 8 column symmetries, we get 64 different symmetries forming a group $R \times C$ isomorphic to $D_4 \times D_4$.

The reader should check that $\mathtt{A}_1$ and $\mathtt{A}_2$ are C-equivalent but not R-equivalent, and that $\mathtt{A}_1$ and $\mathtt{A}_3$ are R-equivalent but not C-equivalent. The mini-Sudokus $\mathtt{A}_1$, $\mathtt{A}_2$, $\mathtt{A}_3$ and $\mathtt{A}_4$ are all in the same $R \times C$-equivalence class. Similar statements hold for $\mathtt{B}_1$, $\mathtt{B}_2$, $\mathtt{B}_3$ and $\mathtt{B}_4$, as well as for $\mathtt{C}_1$, $\mathtt{C}_2$, $\mathtt{C}_3$ and $\mathtt{C}_4$.

Is $\mathtt{A}_1$ $R \times C$-equivalent to $\mathtt{B}_1$ or $\mathtt{C}_1$? That is, is there some combination of row and column symmetries that yields $\mathtt{B}_1$ or $\mathtt{C}_1$ when applied to $\mathtt{A}_1$?

To show that the answer to these questions is no, we associate with each column and row of a mini-Sudoku a partition of the set $\{1, 2, 3, 4\}$—specifically, one of the three

partitions

$$\alpha = \{\{1, 2\}, \{3, 4\}\} \qquad \beta = \{\{1, 3\}, \{2, 4\}\} \qquad \gamma = \{\{1, 4\}, \{2, 3\}\}$$

The notation { } means that order does not matter. For example, $\{\{1, 2\}, \{3, 4\}\}$, $\{\{2, 1\}, \{3, 4\}\}$, $\{\{3, 4\}, \{1, 2\}\}$, and $\{\{4, 3\}, \{2, 1\}\}$ are all different ways of writing α.

The partition associated with a row or a column is fairly obvious—just take the entries and put them in order into $\{\{*, *\}, \{*, *\}\}$ in place of the asterisks. For example, all the row partitions of $\mathtt{A}_1$ are α, and all the column partitions are β. All the row partitions of $\mathtt{B}_1$ are α, but the column partitions are β, β, γ and γ from left to right. The row partitions of $\mathtt{C}_1$ are $\alpha, \alpha, \gamma, \gamma$ from top to bottom, but all the column partitions are β.

For an arbitrary mini-Sudoku $\mathtt{X}$, it is not hard to see that the Rule of One applied to the top two blocks implies that the partitions associated with Row 1 and Row 2 are the same. Of course, the same holds for Row 3 and Row 4, Column 1 and Column 2, and Column 3 and Column 4. So $\mathtt{X}$ is associated with two row partitions and two column partitions. We will record all this information as an ordered pair $[\mathtt{X}]$ of (unordered) pairs of partitions that we call the *partition type* of $\mathtt{X}$. The first entry contains the two row partitions, and the second entry contains the two column partitions. For example,

$$[\mathtt{A}_1] = (\{\alpha, \alpha\}, \{\beta, \beta\}) \qquad [\mathtt{B}_1] = (\{\alpha, \alpha\}, \{\beta, \gamma\}) \qquad [\mathtt{C}_1] = (\{\alpha, \gamma\}, \{\beta, \beta\})$$

Note that $(\{\alpha, \alpha\}, \{\beta, \gamma\})$ and $(\{\alpha, \alpha\}, \{\gamma, \beta\})$ are equal, but $(\{\alpha, \alpha\}, \{\beta, \gamma\})$ is not equal to $(\{\beta, \gamma\}, \{\alpha, \alpha\})$. In particular, $[\mathtt{A}_1] = [\mathtt{A}_2] = [\mathtt{A}_3] = [\mathtt{A}_4]$, $[\mathtt{B}_1] = [\mathtt{B}_2] = [\mathtt{B}_3] = [\mathtt{B}_4]$ and $[\mathtt{C}_1] = [\mathtt{C}_2] = [\mathtt{C}_3] = [\mathtt{C}_4]$.

What makes these partitions useful is how they change under the mini-Sudoku symmetries we have discussed. For example, applying the eight row symmetries in R to the first column of $\mathtt{A}_1$ yields eight different columns:

1	1	3	3	2	4	2	4
3	3	1	1	4	2	4	2
2	4	2	4	1	1	3	3
4	2	4	2	3	3	1	1

However, each of these columns is associated with the same partition, namely β.

Thus column partitions are invariant under the row symmetries, and, similarly, row partitions are invariant under the column symmetries. Of course, the column partitions are simply permuted by column symmetries, and row partitions are permuted by row symmetries. Thus we have the following rule:

RULE 1. *If mini-Sudokus* $\mathtt{X}$ *and* $\mathtt{Y}$ *are* $R \times C$*-equivalent, then their partition types are the same—that is,* $[\mathtt{X}] = [\mathtt{Y}]$.

Since the partition types of $\mathtt{A}_1$, $\mathtt{B}_1$, and $\mathtt{C}_1$ are distinct, no pair of these mini-Sudokus is $R \times C$-equivalent.

The situation we have been considering is typical in mathematics. One has a collection of objects (in our case, mini-Sudokus) and a notion of equivalence, often from a group action (in our case, the group is $R \times C$). The goal is to determine which objects are equivalent. In algebra, the objects might be groups, rings, or fields, and *equivalent* means *isomorphic*. In topology, the objects might be topological spaces or manifolds, and *equivalent* means *homeomorphic*. In linear algebra, the objects might be square matrices, two of which are equivalent if they are similar.

The general strategy for such problems is to attach an *invariant* to each object—something that is the same for equivalent objects. The partition type is an invariant for mini-Sudokus; that is what Rule 1 says.

Other examples of invariants are the order of a group, the elementary divisors of a finite abelian group, the characteristic of a field, the fundamental group of a topological space, the genus of a compact surface, the determinant of a square matrix, and the Jordan canonical form of a square matrix with complex entries. The ideal invariant is easy to compute and completely determines whether or not two objects are equivalent (in which case, we say the invariant is *complete*). The set of elementary divisors is a complete invariant for the set of finite abelian groups. The order of a group is not a complete invariant, however, as nonisomorphic groups can have the same order. The determinant is not a complete invariant for square matrices, but the Jordan canonical form is.

When we defined the notion of *partition type*, we did so carefully, to ensure that it would be an invariant. In particular, it was necessary to define [X] as an ordered pair of *unordered* pairs. For example, let

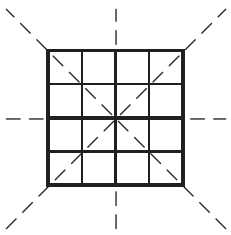
$$\mathtt{D} = \begin{array}{|c|c|c|c|}\hline 3&4&1&2\\\hline 2&1&3&4\\\hline 4&3&2&1\\\hline 1&2&4&3\\\hline\end{array}\,.$$

Notice that D and $\mathtt{B}_1$ are $R \times C$-equivalent; you can obtain one from the other by swapping the two left columns with the two right columns. Rule 1 assures us that they have the same partition type, and indeed we can verify directly that $[\mathtt{D}] = (\{\alpha, \alpha\}, \{\gamma, \beta\}) = [\mathtt{B}_1]$. Had we defined [X] as an ordered pair of *ordered* pairs, however, we would not have had the desired equality $[\mathtt{D}] = [\mathtt{B}_1]$.

Is partition type a complete invariant for mini-Sudokus with respect to $R \times C$-equivalence? In other words, can we find two mini-Sudokus that have the same partition type, but which are not $R \times C$-equivalent? We will see the answer to this question later.

Geometric symmetries

Have we now found all mini-Sudoku symmetries? Definitely not! After all, a mini-Sudoku is itself a square, and it is not hard to see that any symmetry of a square is also a symmetry of mini-Sudokus.



For example, reflecting a mini-Sudoku across its horizontal axis produces a new mini-Sudoku. But this symmetry is just the row symmetry that reverses the order of the rows—it interchanges Row 1 and Row 4, and Row 2 and Row 3. Similarly, reflecting across the vertical axis is a column symmetry.

What about reflections across a diagonal? For concreteness, let τ be the symmetry that reflects a mini-Sudokus across its main diagonal (from top left to bottom right). This symmetry plays an important role in our discussion. As the reader can compute, the partition type of $\tau(\mathtt{A}_1)$ is $[\tau(\mathtt{A}_1)] = (\{\beta, \beta\}, \{\alpha, \alpha\})$, which is enough to show that

this mini-Sudoku is not $R \times C$-equivalent to $\mathtt{A}_1$. In other words, the symmetry τ cannot belong to $R \times C$. Note that $\tau^2 = \mathrm{id}$, so that $\tau^{-1} = \tau$, and $Z = \{\mathrm{id}, \tau\}$ is a group of mini-Sudoku symmetries isomorphic to $\mathbb{Z}_2$.

Rotating 90° clockwise is the result of first reflecting across the main diagonal and then reflecting across the vertical axis. Applying these two symmetries in the other order results in a rotation of 270°. Thus the rotations by 90° and 270° are compositions of τ and symmetries in $R \times C$. Rotation by 180° is the composition of the reflections across the horizontal and vertical axes (in either order). So this rotation is in $R \times C$. Specifically, it is the result of reversing the orders of the rows and the columns.

For example, rotating $\mathtt{B}_1$ by 90°, 180° and 270° clockwise we get the mini-Sudokus

$$\mathtt{S} = \begin{array}{|c|c|c|c|}\hline 4&2&3&1\\\hline 3&1&4&2\\\hline 1&4&2&3\\\hline 2&3&1&4\\\hline\end{array} \qquad \mathtt{T} = \begin{array}{|c|c|c|c|}\hline 2&1&3&4\\\hline 3&4&1&2\\\hline 1&2&4&3\\\hline 4&3&2&1\\\hline\end{array} \qquad \mathtt{U} = \begin{array}{|c|c|c|c|}\hline 4&1&3&2\\\hline 3&2&4&1\\\hline 2&4&1&3\\\hline 1&3&2&4\\\hline\end{array}$$

with partition types $[\mathtt{S}] = [\mathtt{U}] = (\{\beta, \gamma\}, \{\alpha, \alpha\})$ and $[\mathtt{T}] = (\{\alpha, \alpha\}, \{\beta, \gamma\})$. Since $[\mathtt{B}_1] = (\{\alpha, \alpha\}, \{\beta, \gamma\})$, neither $\mathtt{S}$ nor $\mathtt{U}$ is $R \times C$-equivalent to $\mathtt{B}_1$. This means that the corresponding symmetries, rotation by 90° and 270°, are not in $R \times C$.

What about $\mathtt{T}$? As suggested above, $\mathtt{T}$ can be obtained from $\mathtt{B}_1$ by reversing the orders of both the rows and the columns:

$$\mathtt{B}_1 = \begin{array}{|c|c|c|c|}\hline 1&2&3&4\\\hline 3&4&2&1\\\hline 2&1&4&3\\\hline 4&3&1&2\\\hline\end{array} \longrightarrow \begin{array}{|c|c|c|c|}\hline 4&3&1&2\\\hline 2&1&4&3\\\hline 3&4&2&1\\\hline 1&2&3&4\\\hline\end{array} \longrightarrow \begin{array}{|c|c|c|c|}\hline 2&1&3&4\\\hline 3&4&1&2\\\hline 1&2&4&3\\\hline 4&3&2&1\\\hline\end{array} = \mathtt{T}$$

There is just one symmetry in D_4 that we have not yet discussed, namely reflection across the other diagonal (from top right to bottom left). We leave the reader the task of showing that this symmetry is not in $R \times C$, but is, nonetheless, τ composed with a symmetry in $R \times C$.

Note that τ interchanges the rows and the columns of any mini-Sudoku. It sends Row 1 to Column 1, Row 2 to Column 2, etc. This implies also that τ interchanges row symmetries and column symmetries. For example, if $\rho \in R$ is the row symmetry that interchanges Row 1 and Row 2, then $\sigma = \tau\rho\tau$ is the column symmetry that interchanges Column 1 and Column 2. This equation can be written as $\sigma\tau = \tau\rho$, which shows that, even though τ does not commute with elements of $R \times C$, it does so at the cost of interchanging rows and columns. As a consequence, any symmetry that can be obtained by composing τ and elements of $R \times C$ in any order can be written in the form $\tau\mu$ with $\mu \in R \times C$. (The same symmetry can also be written in the form $\nu\tau$ with $\nu \in R \times C$ where ν and μ are the same except for the interchange of rows and columns.)

We now have 64 symmetries in $R \times C$, and 64 more symmetries of the form $\tau\mu$ with $\mu \in R \times C$. In the second category are the rotations by 90° and 270°, as well as the reflections across the diagonal axes. Together, these symmetries form a group H of order 128. Since τ does not commute with all elements of $R \times C$, we know that H is not the direct product of $R \times C$ and $Z = \{\mathrm{id}, \tau\}$. Instead, H is a semi-direct product [**9**] of these groups:

$$H = (R \times C) \rtimes Z.$$

In other words, $R \times C$ is normal in H and has trivial intersection with Z.

Naturally, we will say that two mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ are *H-equivalent* if one can be obtained from the other by applying one of the symmetries in H.

Is $\mathtt{A}_1$ H-equivalent to $\mathtt{B}_1$ or $\mathtt{C}_1$? That is, is there some symmetry in H that yields $\mathtt{B}_1$ or $\mathtt{C}_1$ when applied to $\mathtt{A}_1$? Once again, using partition types, we can show that the answer is no.

Since the group H acts on mini-Sudokus, it also acts on partition types of mini-Sudokus. By Rule 1, symmetries in H that are also in $R \times C$ leave partition types unchanged. Because τ interchanges the rows and columns of mini-Sudokus, this symmetry interchanges the associated row and column partitions of any partition type. For example, since $[\mathtt{C}_1] = (\{\alpha, \gamma\}, \{\beta, \beta\})$, we have $[\tau(\mathtt{C}_1)] = (\{\beta, \beta\}, \{\alpha, \gamma\})$.

In view of the nature of the symmetry τ, it is natural to call $\tau(\mathtt{X})$ the *transpose* of $\mathtt{X}$ and $[\tau(\mathtt{X})]$ the *transpose* of $[\mathtt{X}]$. So we use the notation $[\mathtt{X}]^T = [\tau(\mathtt{X})]$. Thus $[\mathtt{X}]^T$ is obtained from $[\mathtt{X}]$ by switching its two entries. Now, if mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ are H-equivalent, then $\mathtt{X}$ is $R \times C$-equivalent to $\mathtt{Y}$ or to $\tau(\mathtt{Y})$. We therefore say that $[\mathtt{X}]$ and $[\mathtt{Y}]$ are *H-equivalent* if $[\mathtt{X}] = [\mathtt{Y}]$ or $[\mathtt{X}] = [\mathtt{Y}]^T$. Define $[\mathtt{X}]_H$ to be the H-equivalence class of $[\mathtt{X}]$. Note that we now have two H-equivalences: H-equivalence of mini-Sudokus and H-equivalence of partition types.

With Rule 1, we have the following:

RULE 2. *If mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ are H-equivalent, then $[\mathtt{X}]_H = [\mathtt{Y}]_H$.*

Since we have $[\mathtt{A}_1]^T = (\{\beta, \beta\}, \{\alpha, \alpha\})$, $[\mathtt{B}_1]^T = (\{\beta, \gamma\}, \{\alpha, \alpha\})$ and $[\mathtt{C}_1]^T = (\{\beta, \beta\}, \{\alpha, \gamma\})$, no pair of the mini-Sudokus $\mathtt{A}_1$, $\mathtt{B}_1$ and $\mathtt{C}_1$ is H-equivalent.

Is $[\cdot]_H$ a complete invariant with respect to H-equivalence? In other words, are there mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ that have H-equivalent partition types, but which are not H-equivalent? We will see the answer to this question shortly.

Relabeling symmetries

There is yet one other way to create a new mini-Sudoku from a given mini-Sudoku—simply relabel it, that is, apply a permutation of the set $\{1, 2, 3, 4\}$ to its entries. For example, starting with $\mathtt{A}_1$, we could interchange 1 and 2 to get

$$\mathtt{V} = \begin{array}{|c|c|c|c|}\hline 2&1&3&4\\\hline 3&4&2&1\\\hline 1&2&4&3\\\hline 4&3&1&2\\\hline\end{array}.$$

Since $[\mathtt{V}] = (\{\alpha, \alpha\}, \{\gamma, \gamma\})$, $\mathtt{V}$ is not H-equivalent to $\mathtt{A}_1$, and so this relabeling symmetry, switching 1 and 2, is not in H. We have found a new symmetry! Since there are $4! = 24$ different permutations of $\{1, 2, 3, 4\}$ forming the group S_4, there is a corresponding group $L \cong S_4$ of relabeling symmetries of mini-Sudokus.

What do the relabeling symmetries do to the partitions α, β, and γ? The answer is that these symmetries permute them. For example, interchanging 1 and 2 takes α to α, β to γ, and γ to β. This is why this particular labelling symmetry takes $[\mathtt{A}_1] = (\{\alpha, \alpha\}, \{\beta, \beta\})$ to $[\mathtt{V}] = (\{\alpha, \alpha\}, \{\gamma, \gamma\})$.

For another example, consider the relabeling symmetry $\lambda \in L$ that maps 2 to 3, 3 to 4, and 4 to 2, leaving 1 fixed. In cycle notation, we would write $\lambda = (2, 3, 4)$. This symmetry takes α to β, β to γ, and γ back to α. It is not hard to see that each of the six permutations of α, β, and γ comes from exactly 4 relabeling symmetries in L. Hence we say that two partition types are *L-equivalent* if one can be obtained from the other by a permutation of α, β and γ.

We claim that there is no element of H that has the same effect on all mini-Sudokus as the relabeling symmetry $\lambda = (2, 3, 4)$. To show that this is so, we pick a mini-Sudoku

with partition type $(\{\alpha, \alpha\}, \{\beta, \gamma\})$—for example, $\mathtt{B}_1$ will do. Then λ acting on this mini-Sudoku produces a mini-Sudoku with partition type $(\{\beta, \beta\}, \{\gamma, \alpha\})$. By Rule 2, the new mini-Sudoku is not H-equivalent to the original one, and so λ cannot be in H. (Another way to see that $\lambda \notin H$ is to use Lagrange's theorem [**6, 9**]. No element of order 3 can be in H, since $|H| = 128$.) Are *any* of the symmetries in L also in H? It turns out that only the identity symmetry is in both groups. We sketch a proof of this fact, leaving the details to the reader. First observe that if $\lambda \in L$ and $\sigma \in H$, then applying first λ and then σ has the same effect as applying first σ and then λ. In other words, $\lambda\sigma = \sigma\lambda$. So if $\sigma \in L \cap H$, then σ is in the center of L. But L is isomorphic to S_4, whose center contains only the identity.

Counting symmetries How many symmetries have we identified? We know that relabeling symmetries commute with symmetries in H, so combining all of these symmetries gives a group isomorphic to the direct product of H and L. Therefore we define

$$\textit{The mini-Sudoku Symmetry Group} = G \cong H \times L.$$

This group has order $|G| = |H| \cdot |L| = 128 \cdot 24 = 3072$. Since G contains all mini-Sudoku symmetries that we wish to consider, instead of saying that mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ are G-equivalent, we will just say that they are *equivalent*. Equivalence of this type is what is meant by "essentially the same" in the introduction and in [**8, 10**].

Why should G be *the* group of symmetries that we, and others, have chosen to consider? Here's why. Let M be the set of all sixteen cells in a 4×4 grid. Then a mini-Sudoku is nothing more and nothing less than a function $f : M \to \{1, 2, 3, 4\}$ that obeys the Rule of One. An element $\lambda \in L$ acts on a mini-Sudoku f by pre-composition, sending f to $\lambda \circ f$. Likewise, H is a subgroup of the group S_M of all bijective functions from M to itself, and an element $\sigma \in H$ acts on a mini-Sudoku f by post-composition, sending f to $f \circ \sigma$. We have chosen H with care, so that $f \circ \sigma$ necessarily still obeys the Rule of One; hence we shall say that every element of H is *mini-Sudoku-preserving*. It is tedious but straightforward to verify the converse, that every mini-Sudoku-preserving element of S_M is in H. So we have not chosen G arbitrarily at all—it is the set of all mini-Sudoku symmetries that can be obtained by permuting cells and permuting labels.

(For 9×9 Sudokus, the group generated by the row and column symmetries together with the rotations and reflections has order 3,359,232, and there are 9! relabeling symmetries. Hence the Sudoku symmetry group has order $3{,}359{,}232 \cdot 9! = 1{,}218{,}998{,}108{,}160$. We remark that the row symmetry group for an $n^2 \times n^2$ Sudoku is an n-fold wreath product [**10**].)

Note that, like the groups H and L, the group G acts on mini-Sudokus and also on partition types. Define $[\mathtt{X}]_G$ to be the G-equivalence class of $[\mathtt{X}]$. In other words, $[\mathtt{X}]_G = [\mathtt{Y}]_G$ if and only if $[\mathtt{X}]$ is L-equivalent to $[\mathtt{Y}]$ or to $[\mathtt{Y}]^T$.

RULE 3. *If mini-Sudokus* $\mathtt{X}$ *and* $\mathtt{Y}$ *are equivalent, then* $[\mathtt{X}]_G = [\mathtt{Y}]_G$.

Now we can see whether $\mathtt{A}_1$, $\mathtt{B}_1$, and $\mathtt{C}_1$ are equivalent. If a mini-Sudoku $\mathtt{X}$ is equivalent to $\mathtt{A}_1$, then, by Rule 3, $[\mathtt{X}]$ is $(\{\alpha, \alpha\}, \{\beta, \beta\})$, $(\{\beta, \beta\}, \{\alpha, \alpha\})$, $(\{\alpha, \alpha\}, \{\gamma, \gamma\})$, $(\{\gamma, \gamma\}, \{\alpha, \alpha\})$, $(\{\beta, \beta\}, \{\gamma, \gamma\})$ or $(\{\gamma, \gamma\}, \{\beta, \beta\})$. So, in particular, $\mathtt{X}$ cannot be $\mathtt{B}_1$ or $\mathtt{C}_1$, and so $\mathtt{A}_1$ is not equivalent to either of these mini-Sudokus.

What about the equivalence of $\mathtt{B}_1$ and $\mathtt{C}_1$? If $\mathtt{X}$ is equivalent to $\mathtt{B}_1$, then, by Rule 3, $[\mathtt{X}]$ is $(\{\alpha, \alpha\}, \{\beta, \gamma\})$, $(\{\beta, \gamma\}, \{\alpha, \alpha\})$, $(\{\beta, \beta\}, \{\alpha, \gamma\})$, $(\{\alpha, \gamma\}, \{\beta, \beta\})$, $(\{\gamma, \gamma\}, \{\alpha, \beta\})$, or $(\{\alpha, \beta\}, \{\gamma, \gamma\})$. Because $[\mathtt{C}_1] = (\{\alpha, \gamma\}, \{\beta, \beta\})$, it is possible that $\mathtt{B}_1$ and $\mathtt{C}_1$ are equivalent. Since we have not (yet) shown that the converse of Rule 3 holds, we do not yet know whether $\mathtt{B}_1$ is equivalent to $\mathtt{C}_1$ or not.

If these mini-Sudokus are equivalent, then the partition types of $\mathtt{B}_1$ and $\mathtt{C}_1$ suggest how to construct a symmetry that takes one to the other. There will have to be a relabeling symmetry that interchanges α and β, composed with the transposition τ, composed (perhaps) with some row and column symmetry:

$$[\mathtt{B}_1] = (\{\alpha,\alpha\},\{\beta,\gamma\}) \longrightarrow (\{\beta,\beta\},\{\alpha,\gamma\}) \longrightarrow (\{\alpha,\gamma\},\{\beta,\beta\}) = [\mathtt{C}_1].$$

In fact, by choosing the relabeling symmetry, $\lambda \in L$, which interchanges 2 and 3, no row or column symmetry is needed:

$$\mathtt{B}_1 = \begin{array}{|c|c|c|c|}\hline 1&2&3&4\\\hline 3&4&2&1\\\hline 2&1&4&3\\\hline 4&3&1&2\\\hline\end{array} \xrightarrow{\lambda} \begin{array}{|c|c|c|c|}\hline 1&3&2&4\\\hline 2&4&3&1\\\hline 3&1&4&2\\\hline 4&2&1&3\\\hline\end{array} \xrightarrow{\tau} \begin{array}{|c|c|c|c|}\hline 1&2&3&4\\\hline 3&4&1&2\\\hline 2&3&4&1\\\hline 4&1&2&3\\\hline\end{array} = \mathtt{C}_1.$$

This shows that $\mathtt{B}_1$ and $\mathtt{C}_1$ are equivalent, and hence that $\mathtt{B}_1$, $\mathtt{B}_2$, $\mathtt{B}_3$, $\mathtt{B}_4$, $\mathtt{C}_1$, $\mathtt{C}_2$, $\mathtt{C}_3$, and $\mathtt{C}_4$ are all in the same equivalence class. Since $\mathtt{A}_1$ and $\mathtt{B}_1$ are not equivalent, there must be a second equivalence class containing $\mathtt{A}_1$, $\mathtt{A}_2$, $\mathtt{A}_3$, and $\mathtt{A}_4$.

Now we are ready for the main (and only) theorem of this article.

THEOREM. *There are exactly two equivalence classes of mini-Sudokus:*

$\mathcal{C}_1$: *All mini-Sudokus with the following partition types:*

$$\begin{array}{lll}(\{\alpha,\alpha\},\{\beta,\beta\}) & (\{\alpha,\alpha\},\{\gamma,\gamma\}) & (\{\beta,\beta\},\{\gamma,\gamma\})\\ (\{\beta,\beta\},\{\alpha,\alpha\}) & (\{\gamma,\gamma\},\{\alpha,\alpha\}) & (\{\gamma,\gamma\},\{\beta,\beta\})\end{array}$$

$\mathcal{C}_2$: *All mini-Sudokus with the following partition types:*

$$\begin{array}{lll}(\{\alpha,\alpha\},\{\beta,\gamma\}) & (\{\beta,\beta\},\{\alpha,\gamma\}) & (\{\gamma,\gamma\},\{\alpha,\beta\})\\ (\{\beta,\gamma\},\{\alpha,\alpha\}) & (\{\alpha,\gamma\},\{\beta,\beta\}) & (\{\alpha,\beta\},\{\gamma,\gamma\})\end{array}$$

Proof. We know already that there are at least two distinct equivalence classes.

Let $\mathtt{X}$ be a mini-Sudoku. By applying a suitable relabeling symmetry, the top left box of $\mathtt{X}$ can be put in the form $\begin{array}{|c|c|}\hline 1&2\\\hline 3&4\\\hline\end{array}$, and so $\mathtt{X}$ is equivalent to one of the 12 mini-Sudokus $\mathtt{A}_1, \mathtt{A}_2, \ldots, \mathtt{C}_3, \mathtt{C}_4$. From the above discussion, $\mathtt{X}$ is equivalent to either $\mathtt{A}_1$ or $\mathtt{B}_1$. But we have seen already that, if $\mathtt{X}$ is equivalent to $\mathtt{A}_1$, then its partition type is as described in $\mathcal{C}_1$, and if $\mathtt{X}$ is equivalent to $\mathtt{B}_1$, then its partition type is as described in $\mathcal{C}_2$. ∎

(A similar argument in [**8**] purports to demonstrate the same result; in fact, the line of reasoning in that article shows only that there are *at most* two equivalence classes. What is missing is an invariant to distinguish the two classes.)

There are exactly 24 mini-Sudokus L-equivalent to each of $\mathtt{A}_1, \mathtt{A}_2, \ldots, \mathtt{C}_3, \mathtt{C}_4$, and hence there are $4 \cdot 24 = 96$ mini-Sudokus in $\mathcal{C}_1$ and $8 \cdot 24 = 192$ mini-Sudokus in $\mathcal{C}_2$.

Since we now know the partition types that are in each of the equivalence classes, it is easy to see that the converse of Rule 3 holds, that is, mini-Sudokus $\mathtt{X}$ and $\mathtt{Y}$ are equivalent if and only if $[\mathtt{X}]$ and $[\mathtt{Y}]$ are equivalent.

Notice that the mini-Sudokus $\mathtt{A}_1$, $\mathtt{A}_2$, $\mathtt{A}_3$, and $\mathtt{A}_4$ have either two or four distinct entries on the main diagonal, whereas the mini-Sudokus $\mathtt{B}_1, \mathtt{B}_2, \ldots, \mathtt{C}_3, \mathtt{C}_4$ have exactly three distinct entries on the main diagonal. Since any mini-Sudoku is L-equivalent to one of these 12, and the number of distinct entries on the main diagonal is unchanged by relabeling symmetries, it now quite easy to tell which equivalence class a mini-Sudoku $\mathtt{X}$ belongs to. If $\mathtt{X}$ has two or four distinct entries on the main diagonal it must be L-equivalent to $\mathtt{A}_1$, $\mathtt{A}_2$, $\mathtt{A}_3$, or $\mathtt{A}_4$, and so $\mathtt{X}$ is in $\mathcal{C}_1$. If $\mathtt{X}$ has three distinct entries on the

main diagonal it must be L-equivalent to one of $\mathtt{B}_1, \mathtt{B}_2, \ldots, \mathtt{C}_3, \mathtt{C}_4$, and so $\mathtt{X}$ is in $\mathcal{C}_2$. Hence the diagonal entries of $\mathtt{X}$ suffice to determine its equivalence class.

It is, of course, easier to count entries along the main diagonal of a mini-Sudoku than to write down its partition type. So why bother with partition types at all? The reason is that they are better suited for weaker forms of equivalence, such as $R \times C$-equivalence and H-equivalence. In fact, the following converses to Rules 1 and 2 assert that $[\cdot]$ is a complete invariant of mini-Sudokus, modulo $R \times C$-equivalence and that $[\cdot]_H$ is a complete invariant with respect to H-equivalence:

PROPOSITION. *Let* $\mathtt{X}$ *and* $\mathtt{Y}$ *be mini-Sudokus.*

(1) *If* $[\mathtt{X}] = [\mathtt{Y}]$, *then* $\mathtt{X}$ *and* $\mathtt{Y}$ *are* $R \times C$*-equivalent.*

(2) *If* $[\mathtt{X}]_H = [\mathtt{Y}]_H$, *then* $\mathtt{X}$ *and* $\mathtt{Y}$ *are* H*-equivalent.*

Proof. (1) Suppose $[\mathtt{X}] = [\mathtt{Y}]$. Recall that Row 1 and Row 2 of $\mathtt{X}$ are associated with the same partition—either α, β, or γ—and that the same is true of Rows 3 and 4. For convenience, we will call these four partitions, in order, the *row partitions* of $\mathtt{X}$.

It may be that the row partitions of $\mathtt{X}$ match (row for row) the row partitions of $\mathtt{Y}$, in which case let $\mathtt{Y}_1 = \mathtt{Y}$. If this is not the case, then, since $[\mathtt{X}] = [\mathtt{Y}]$, applying the *blockwise* row symmetry that switches Rows 1 and 2 with Rows 3 and 4 to $\mathtt{Y}$, yields a mini-Sudoku $\mathtt{Y}_1$ whose row partitions match those of $\mathtt{X}$. Similarly, by applying a blockwise column symmetry to $\mathtt{Y}_1$ if necessary, we obtain a mini-Sudoku $\mathtt{Y}_2$ such that both the row and column partitions of $\mathtt{X}$ and $\mathtt{Y}_2$ match up, and such that $\mathtt{Y}_2$ is $R \times C$-equivalent to $\mathtt{Y}$.

The two row symmetries that switch Rows 1 and 2, and Rows 3 and 4, and the two column symmetries that switch Columns 1 and 2, and Columns 3 and 4, can be used to put any mini-Sudoku into the form

1	*	*	*
*	*	*	*
*	*	1	*
*	*	*	*

.

Moreover, this can be done without changing row and column partitions. So by applying this procedure to $\mathtt{Y}_2$, we obtain a mini-Sudoku $\mathtt{Y}_3$, which is $R \times C$-equivalent to $\mathtt{Y}_2$, such that $\mathtt{X}$ and $\mathtt{Y}_3$ have this special form, in addition to having matching row and column partitions.

Since $\mathtt{X}$ and $\mathtt{Y}_3$ have the same top row partition, it follows that they have the same entries in Row 1, Column 2. Similarly, using the leftmost column partition, we see they have the same entries in Row 2, Column 1. Thus, they have the same upper-left block. Likewise, we find that they have the same lower-right block. We can then use the Rule of One to fill in the remaining entries in the grid and conclude that $\mathtt{X} = \mathtt{Y}_3$. The result follows.

(2) We know that $[\mathtt{X}] = [\mathtt{Y}]$ or $[\mathtt{X}] = [\mathtt{Y}]^{\mathrm{T}}$. So by (1), $\mathtt{X}$ is $R \times C$-equivalent to $\mathtt{Y}$ or to $\mathtt{Y}^{\mathrm{T}}$. In either case, $\mathtt{X}$ is H-equivalent to $\mathtt{Y}$. ∎

More mini-Sudoku puzzles

Though we now know how many mini-Sudokus there are and how many of them are essentially different, many mini-Sudokus puzzles remain for the reader to solve. Here are some suggestions:

1. According to the theorem, if $\mathtt{X}$ is a mini-Sudoku, then $[\mathtt{X}] = (\{\alpha, \alpha\}, \{\alpha, \alpha\})$, $[\mathtt{X}] = (\{\alpha, \beta\}, \{\alpha, \beta\})$, $[\mathtt{X}] = (\{\alpha, \alpha\}, \{\alpha, \beta\})$, and $[\mathtt{X}] = (\{\alpha, \beta\}, \{\alpha, \gamma\})$ are not possible. Show directly that it is not possible for a row partition to equal a column partition.
2. For a mini-Sudoku $\mathtt{X}$, let $\det \mathtt{X}$ be the determinant of $\mathtt{X}$ thought of as a 4×4 matrix. Then $\det \mathtt{X}$ is unchanged or changes sign under the symmetries in H. Hence, if $\mathtt{X}$ and $\mathtt{Y}$ are H-equivalent, then $|\det \mathtt{X}| = |\det \mathtt{Y}|$. Is the converse true?

   It might be useful to replace the entries, 1, 2, 3 and 4, by variables, w, x, y and z, so that $\det \mathtt{X}$ is a polynomial in the four variables. For example,

$$\det \mathtt{A}_1 = \det \begin{bmatrix} w & x & y & z \\ y & z & w & x \\ x & w & z & y \\ z & y & x & w \end{bmatrix}$$
$$= -(w + x - y - z)(w + y - x - z)(w + z - x - y)(w + x + y + z)$$

   How do these determinants change under relabeling symmetries? Can such determinants be used to determine whether mini-Sudokus are H-equivalent or equivalent?
3. Prove that $R \cap C = (R \times C) \cap Z = \{\mathrm{id}\}$.
4. Apply the reasoning from our article to Cayley-Sudoku tables [**2**]. How many 4×4 Cayley-Sudoku tables are there? Is there a natural group action on the set of all Cayley-Sudoku tables of a fixed size? With respect to that group action, how many equivalence classes are there, and can you find a complete invariant to distinguish them?

REFERENCES

1. S. F. Bammel and J. Rothstein, The Number of 9×9 Latin squares, *Discrete Math.* **11** (1975) 93–95. doi:10.1016/0012-365X(75)90108-9
2. J. Carmichael, K. Schloeman, and M. B. Ward, Cosets and Cayley-Sudoku Tables, *Math. Mag.* **83** (2010) 130–139, 147–148. doi:10.4169/002557010X482899
3. T. Davis, *The Mathematics of Sudoku*, http://www.geometer.org/mathcircles/sudoku.pdf, October 2008.
4. J.-P. Delahaye, The Science behind Sudoku, *Sci. Amer.* (June 2006) 80–87.
5. B. Felgenhauer and F. Jarvis, Mathematics of Sudoku I, *Math. Spectrum* **39** (2006) 15–22.
6. J. Gallian, *Contemporary Abstract Algebra*, 7th ed., Brooks/Cole, Belmont, CA, 2010.
7. B. Hayes, Unwed Numbers, *Amer. Sci.* **94** (2006) 12–15. doi:10.1511/2006.1.12
8. A. Herzberg and R. R. Murty, Sudoku Squares and Chromatic Polynomials, *Notices Amer. Math. Soc.* **54** (2007) 708–717.
9. J. Rotman, *An Introduction to the Theory of Groups*, 4th ed., Springer Verlag, New York, 1999.
10. E. Russell and F. Jarvis, Mathematics of Sudoku II, *Math. Spectrum* **39** (2006) 54–58.

Summary Using a little group theory and ideas about equivalence relations, this article shows that there are only two essentially different 4 by 4 Sudoku grids. This can be compared with the 5,472,730,538 essentially different 9 by 9 Sudoku grids found by Jarvis and Russell with the aid of a computer algebra system.

CARLOS ARCOS has just finished his master's thesis on the topological properties of affine spaces at California State University, Los Angeles under the supervision of Mike Krebs.

GARY BROOKFIELD got his Ph.D. at the University of California, Santa Barbara in 1997 and was a visiting professor at the University of Wisconsin, Madison, the University of Iowa and the University of California, Riverside. He is currently an associate professor at California State University, Los Angeles. His interests are in ring and module theory, as well as in the revival of the theory of equations.

MIKE KREBS is currently an assistant professor of mathematics at California State University, Los Angeles. His (current) primary research area is algebraic graph theory, in particular the theory of expander graphs.

NOTES

Gershgorin Disk Fragments

AARON MELMAN
Department of Applied Mathematics
School of Engineering
Santa Clara University
Santa Clara, CA 95053
amelman@scu.edu

Computing the eigenvalues of a general complex matrix can be hard, but finding regions in the complex plane that contain them is surprisingly easy. One way to obtain eigenvalue inclusion regions is to form the Gershgorin disks, centered at the diagonal elements of the matrix. In this note we focus on a lesser known complementary result about where not to look for eigenvalues, which was obtained by Parodi ([**4**]) and Schneider ([**5**]).

We give a standard proof of Gershgorin's theorem and show how it can be continued in a natural way to derive the Parodi-Schneider inclusion regions, which consist of fragments of the Gershgorin disks. After deriving them, we will give some examples, including an application to the location of polynomial zeros.

Parodi ([**4**]) obtained his result as a consequence of a theorem in [**3**]. The Parodi and Schneider results are summarized and discussed in [**7**, p. 73–79, 95]. Gershgorin's theorem originally appeared in [**1**], but a proof can also be found in textbooks ([**2**, p. 344–345]).

Historically, Lucien Lévy was already aware in 1881 of an equivalent formulation of Gershgorin's theorem, albeit for real matrices only, and independent rediscoveries of this theorem were made in subsequent years by many mathematicians ([**6**]). Eigenvalue inclusion sets are a rich subject area and we refer to [**2**], [**7**], and the many references therein for more advanced results.

Disk fragments Gershgorin's theorem states that all the eigenvalues of an $n \times n$ complex matrix $A = [a_{ij}]$ are contained in the union of n disks, each centered at a diagonal element a_{pp} of the matrix. The radius of each disk is equal to the deleted row sum corresponding to the diagonal element, which is the sum of the absolute values of all nondiagonal elements in that row:

$$R'_p(A) = \sum_{\substack{j=1 \\ j \neq p}}^{n} |a_{pj}|.$$

(Since all of our sums cover $j = 1, \ldots, n$, from now on we will write $\sum_{j \neq p}$, for example, instead of $\sum_{\substack{j=1 \\ j \neq p}}^{n}$.) The accompanying Gershgorin disk is

$$\Gamma_p^R(A) = \left\{ z \in \mathbb{C} : |z - a_{pp}| \leq R'_p(A) \right\}.$$

Math. Mag. **83** (2010) 123–129. doi:10.4169/002557010X482880.

Gershgorin's theorem is:

THEOREM 1. *All the eigenvalues of the $n \times n$ complex matrix A are located in*

$$\bigcup_{p=1}^{n} \Gamma_p^R(A) \equiv \Gamma^R(A).$$

An analogous statement holds for the columns of the matrix because the eigenvalues of A and A^T are identical. To avoid repetition we concentrate on the rows only.

As illustration, in FIGURE 1 one finds the Gershgorin disks $\Gamma_i^R(A_1)$ (left, shaded area) that make up the Gershgorin set $\Gamma^R(A_1)$ (right, shaded area) for the matrix

$$A_1 = \begin{pmatrix} 14 & i & 0 & 18-2i \\ 0 & 16 & 4+i & 0 \\ 1+i & 4+i & 11 & 0 \\ 14+i & 0 & 1+i & 10 \end{pmatrix},$$

where $i^2 = -1$. The white dots in the Gershgorin set indicate the eigenvalues of the matrix. We note that although the union of all disks must contain all eigenvalues, an individual disk need not contain an eigenvalue.

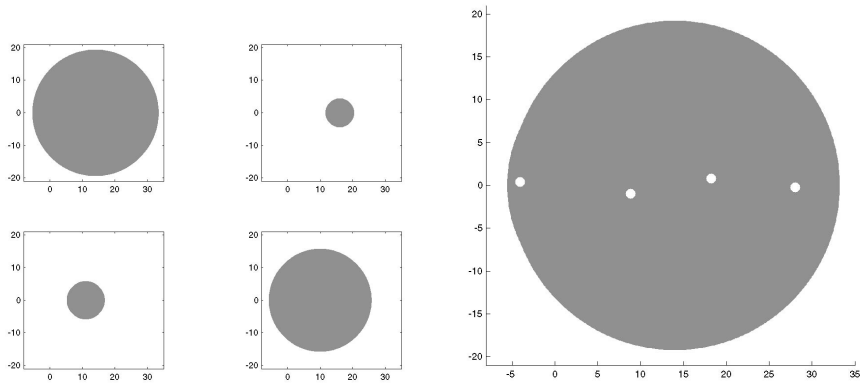


Figure 1 The four Gershgorin disks (left) for the matrix A_1 and their union (right).

The proof of Gershgorin's theorem relies on basic matrix concepts only.

Assume that λ is an eigenvalue of the $n \times n$ complex matrix $A = [a_{ij}]$ with corresponding eigenvector x, i.e., $Ax = \lambda x$. Since x is an eigenvector, it has at least one nonzero component. Let x_p be a component of x with the largest absolute value, so that $|x_p| \geq |x_j|$ for all $j = 1, 2, \ldots, n$ and $x_p \neq 0$. Because $(Ax)_p = (\lambda x)_p$, we have

$$\lambda x_p = a_{pp} x_p + \sum_{j \neq p} a_{pj} x_j.$$

From this it follows that

$$(\lambda - a_{pp}) x_p = \sum_{j \neq p} a_{pj} x_j.$$

Taking absolute values on both sides, using the triangle inequality, and dividing through by $|x_p|$ yields

$$|\lambda - a_{pp}| \leq \sum_{j \neq p} |a_{pj}| \frac{|x_j|}{|x_p|} \leq \sum_{j \neq p} |a_{pj}| = R_p'(A),$$

because $|x_j|/|x_p| \leq 1$ for all $j \neq p$, i.e., λ lies in a disk with center a_{pp} and radius $R'_p(A)$. We do not know which p each eigenvalue corresponds to, so we must take the union of all such disks to obtain a region that is guaranteed to contain all eigenvalues. Here ends the standard proof.

But why stop here? Why not also consider the other components? For any $q \neq p$, we have

$$\lambda x_q = a_{qp}x_p + a_{qq}x_q + \sum_{j\neq p,q} a_{qj}x_j,$$

so that

$$a_{qp}x_p = (\lambda - a_{qq})x_q - \sum_{j\neq p,q} a_{qj}x_j.$$

Taking absolute values as before and dividing through by $|x_p|$ yields

$$|a_{qp}| \leq |\lambda - a_{qq}|\frac{|x_q|}{|x_p|} + \sum_{j\neq p,q} |a_{qj}|\frac{|x_j|}{|x_p|} \leq |\lambda - a_{qq}| + \sum_{j\neq p,q} |a_{qj}|,$$

because we still have $|x_j|/|x_p| \leq 1$ for all $j \neq p$. We finally obtain the additional inequality

$$|\lambda - a_{qq}| \geq |a_{qp}| - \sum_{j\neq p,q} |a_{qj}| = 2|a_{qp}| - R'_q(A). \tag{1}$$

This inequality has the effect of excluding λ from an open disk centered at a_{qq} and having radius $2|a_{qp}| - R'_q(A)$. We call this an *exclusion disk*. It is nontrivial only when the radius is positive. Note that the Gershgorin disk is determined by row p, while the exclusion disk is determined by row q—with a special role for a_{qp}.

There is an exclusion disk for every $q \neq p$. Forming their union and removing that union from the Gershgorin disk gives a new, smaller inclusion set corresponding to p. Such a set is a *disk fragment*.

Define

$$\Delta^R_{pq}(A) = \left\{z \in \mathbb{C} : |z - a_{qq}| \geq 2|a_{qp}| - R'_q(A)\right\},$$

$$\Delta^R_p(A) = \bigcup_{\substack{q=1\\ q\neq p}}^{n} \Delta^R_{pq}(A),$$

and

$$\Omega^R_p(A) = \Gamma^R_p(A) \setminus \Delta^R_p(A).$$

We have proved the following theorem:

THEOREM 2. *All the eigenvalues of an $n \times n$ complex matrix A are located in the union of n disk fragments*

$$\bigcup_{p=1}^{n} \Omega^R_p(A) \equiv \Omega^R(A).$$

This union is contained in the Gershgorin set $\Gamma^R(A)$.

Theorem 2 is precisely the Parodi-Schneider result.

The right-hand side of

$$|\lambda - a_{qq}| \geq 2|a_{qp}| - R'_q(A)$$

with given q is strictly positive for at most one index p. This is trivially true for $n = 2$. To show that it is true for $n \geq 3$, assume that there are two such indices, namely, p and ℓ, i.e.,

$$2|a_{qp}| > R'_q(A) \quad \text{and} \quad 2|a_{q\ell}| > R'_q(A). \tag{2}$$

Defining the nonnegative number $\alpha = R'_q(A) - |a_{qp}| - |a_{q\ell}|$, we can rewrite (2) as

$$|a_{qp}| - |a_{q\ell}| > \alpha \quad \text{and} \quad |a_{q\ell}| - |a_{qp}| > \alpha.$$

Since a real number and its opposite cannot be both strictly greater than zero, we have arrived at a contradiction.

The inequality in (1) is trivially satisfied if $2|a_{qp}| \leq R'_q(A)$. It is easy to find matrices where, for given q, $2|a_{qp}| \leq R'_q(A)$ for all $p \neq q$. The matrices

$$\begin{pmatrix} 2 & 4 & 4 \\ 2 & 1 & 2 \\ 0 & 4 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 2 \\ 1 & 1 & 0 \end{pmatrix}$$

are examples of that for $q = 1, 2$ and $q = 1, 2, 3$, respectively.

On the other hand, it is also possible that $\Omega_k^R(A) = \emptyset$ for some index k, which would mean that for no eigenvector does the kth component of any eigenvector associated with it have the largest absolute value.

Although individual disk fragments $\Omega_p^R(A)$ can look very different from Gershgorin disks, it is quite common for their union to coincide with the Gershgorin set. This is illustrated in FIGURE 2 for the matrix

$$A_2 = \begin{pmatrix} 1 & i & 0 & 10 \\ 0 & 16 & 4+i & 0 \\ 2i & 4 & 10 & 0 \\ 4 & 0 & 1+i & 10 \end{pmatrix},$$

where one finds, on the left (shaded area), the sets $\Omega_p^R(A_2)$ and, on the right (shaded area), their union, which is the same as the Gershgorin set. The eigenvalues of A_2 are indicated by white dots.

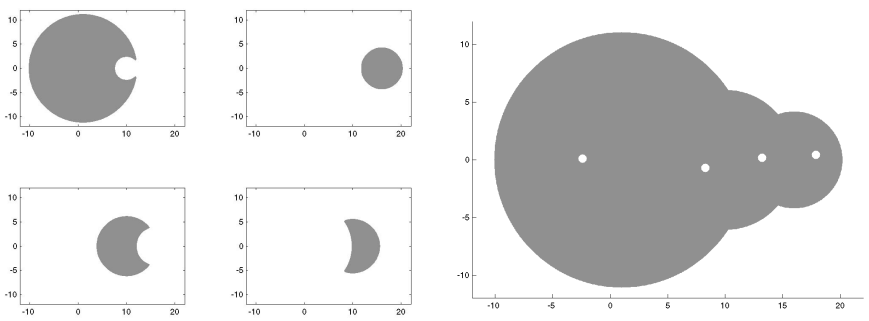


Figure 2 The four Gershgorin disks (left) for the matrix A_2 and their union (right).

The more interesting cases are the ones for which the union of disk fragments is different from the Gershgorin set, as for the example in FIGURE 3. There one finds, on the left (shaded area), the sets $\Omega_p^R(A_1)$ for the matrix A_1 used in FIGURE 1 and, on the right (shaded area), their union. The eigenvalues of A_1 are once again indicated by white dots.

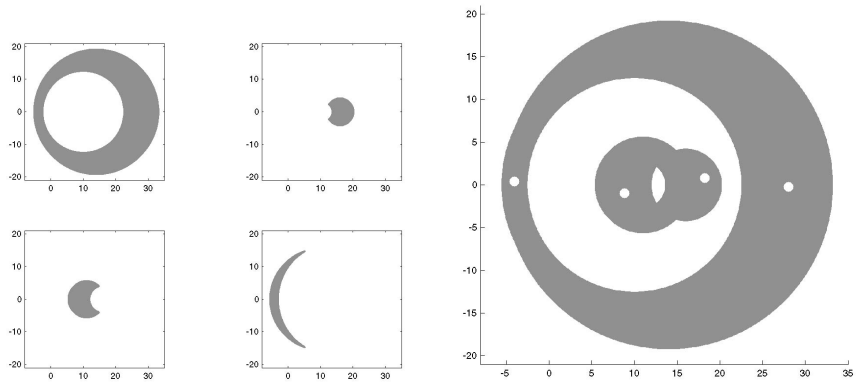


Figure 3 The four Gershgorin disk fragments (left) and their union (right) for the matrix A_1.

An invertibility criterion Just as for Gershgorin's theorem, Theorem 2 gives an invertibility criterion for matrices: zero does not lie in the eigenvalue inclusion set. For Gershgorin's theorem this leads to the condition of strict diagonal dominance (the Lévy-Desplanques theorem ([**2**, p. 302])). Likewise, requiring that zero not lie in $\Omega^R(A)$ leads directly to the following theorem.

THEOREM 3. *An $n \times n$ complex matrix A is nonsingular if for each $p = 1, 2, \ldots, n$ either $|a_{pp}| > R'_p(A)$ or $|a_{qq}| < 2|a_{qp}| - R'_q(A)$ for some $q \neq p$.*

This is a sufficient but not a necessary conditon for invertibility.

FIGURES 1 and 2 make clear that the Lévy-Desplanques theorem cannot determine the invertibility of A_1, whereas the Gershgorin disk fragment set clearly shows A_1 to be invertible.

More examples Consider the matrices

$$A_3 = \begin{pmatrix} 11 & 1 & 1 & 11+i \\ 0 & 2 & 14+i & 0 \\ 1+i & 10+i & 2+i & 1 \\ 12+i & i & 1 & 12+i \end{pmatrix}$$

and

$$A_4 = \begin{pmatrix} 15+i & 1+i & 0 & 14+i \\ 0 & 11 & 4+i & 0 \\ 0 & 9 & 14+i & i \\ 14 & 1 & 0 & 14 \end{pmatrix}.$$

Their Gershgorin disk fragment sets (Ω^R) are the shaded areas in FIGURE 4. Their respective Gershgorin sets would cover all of the interior white areas.

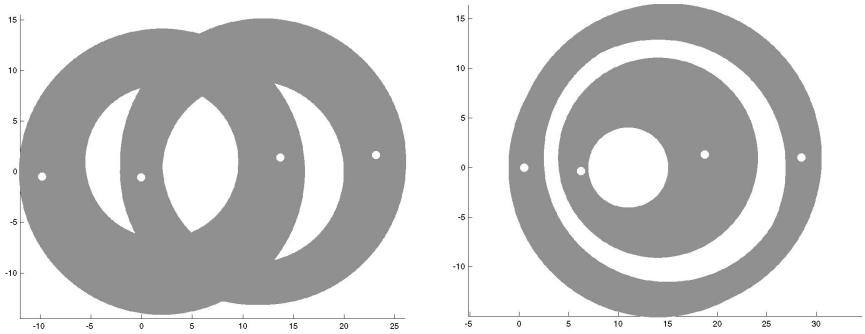


Figure 4 The Gershgorin disk fragment sets for the eigenvalues of A_3 (left) and A_4 (right).

Zeros of polynomials Eigenvalue inclusion sets can be used to locate zeros of polynomials by using the polynomial's companion matrix, whose characteristic polynomial is the given polynomial ([**2**, p. 146]). Thus, its eigenvalues are the zeros of the polynomial. The companion matrix of a monic polynomial $p(z) = z^n + \alpha_{n-1}z^{n-1} + \cdots + \alpha_1 z + \alpha_0$ is

$$C(p) = \begin{pmatrix} 0 & 0 & \dots & 0 & -\alpha_0 \\ 1 & 0 & \dots & 0 & -\alpha_1 \\ 0 & 1 & \dots & 0 & -\alpha_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -\alpha_{n-1} \end{pmatrix}.$$

FIGURE 5 shows Γ^R (left, shaded area) and Ω^R (right, shaded area) for $C(p_1)$, with $p_1(z) = z^4 - z^3 + 0.2z^2 - 0.1z + 2$. The zeros are indicated by the white dots.

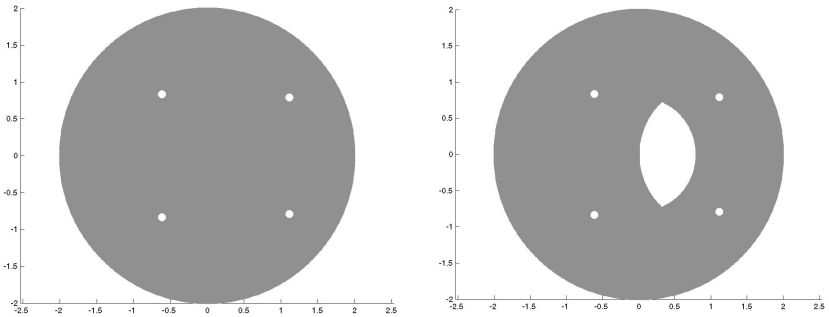


Figure 5 Γ^R (left) and Ω^R (right) for $C(p_1)$.

The simple structure of the companion matrix makes it easy to consider special cases. For monic polynomials with $|\alpha_j| < 1$ for all $j = 0, \dots, n-1$, one can show that

$$\bigcup_{i=2}^{n-2} \Omega_i^R(A) = \left\{ z \in \mathbb{C} : 1 - \max_{2 \le j \le n-2} |\alpha_j| \le |z| \le 1 + \max_{1 \le j \le n-3} |\alpha_j| \right\},$$

so that the union of all but three disk fragments is given by an annulus centered at the origin. That makes this case simple enough to draw the Gershgorin disk fragment set

Ω^R by hand with just paper and compass, regardless of the degree of the polynomial. In FIGURE 6, the shaded areas show the Gershgorin disk fragment sets for the companion matrices of $p_2(z) = z^8 + 0.5z^7 + 0.2z^6 + 0.15z^5 + 0.3z^4 + 0.1z^3 + 0.2z^2 + 0.1z + 0.8$ (left) and $p_3(z) = z^8 + 0.01z^7 + 0.02z^6 + 0.04z^5 + 0.02z^4 + 0.01z^3 + 0.02z^2 + 0.04z + 1$ (right), with the zeros of the polynomials indicated by the white dots. The respective Gershgorin sets cover all of the interior unshaded area.

Finally, the zeros of $p(z) = z^n + 1$ are just the nth roots of -1. The Gershgorin set Γ^R for $C(p)$ is the unit disk, whereas the Gershgorin disk fragment set Ω^R for $C(p)$ is the unit circle.

We conclude by noting that exclusion sets can also be obtained for a polynomial $p(z)$ of degree n with a nonzero constant term by applying Gershgorin's theorem to the companion matrix of the polynomial $z^n p(1/z)$ whose zeros are the reciprocals of the zeros of $p(z)$ ([**2**, p. 318]).

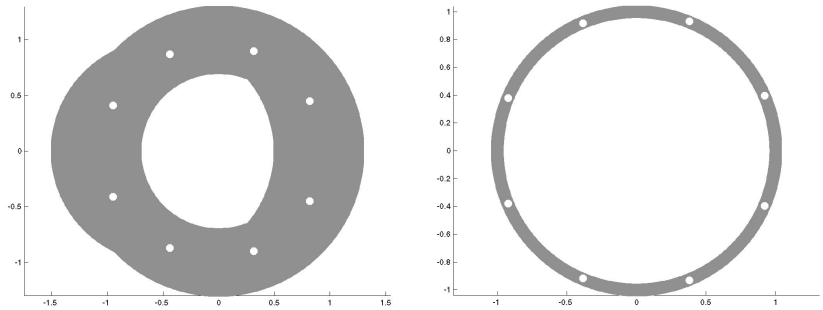


Figure 6 Ω^R for $C(p_2)$ (left) and $C(p_3)$ (right).

REFERENCES

1. S. Gerschgorin, Über die Abgrenzung der Eigenwerte einer Matrix, *Izv. Akad. Nauk SSSR, Ser. Mat.* (1931) 749–754.
2. R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1988.
3. M. Müller, Ein Kriterium für das Nichtverschwinden von Determinanten. *Math. Z.* **51** (1948) 291–293. doi: 10.1007/BF01181595
4. M. Parodi, Sur quelques propriétés des valeurs caractéristiques des matrices carrées, *Mémor. Sci. Math* **118** (1952), Gauthier-Villars, Paris.
5. H. Schneider, Regions of exclusion for the latent roots of a matrix, *Proc. Amer. Math. Soc.* **5** (1954) 320–322. doi:10.2307/2032247
6. Olga Taussky, A recurring theorem on determinants, *Amer. Math. Monthly* **56** (1949) 672–676. doi:10.2307/2305561
7. R. S. Varga, *Geršgorin and His Circles*, Springer Verlag, Berlin, 2004.

Summary Eigenvalue inclusion regions for a general complex matrix can be found by forming the Gershgorin disks, centered at the diagonal elements of the matrix. We give a standard proof of Gershgorin's theorem and show how it can be can be continued in a natural way to derive a lesser known result obtained by Parodi and Schneider, which uses fragments of the Gershgorin disks. We provide some examples, including an application to the location of polynomial zeros.

AARON MELMAN received his M.S. from the Technion–Israel Institute of Technology in 1986 and Ph.D. from Caltech in 1992. He taught at Ben-Gurion University in Israel and is currently teaching at Santa Clara University in California. His research interests are in numerical analysis and linear algebra, with a particular fondness for structured matrices. In his spare time, or whatever spare time his kids allow him to have, he is also very interested in linguistics and history.

Cosets and Cayley-Sudoku Tables

JENNIFER CARMICHAEL
Chemeketa Community College
Salem, OR 97309
jcarmic5@my.chemeketa.edu

KEITH SCHLOEMAN
Oregon State University
Corvallis, OR 97331
schloemk@lifetime.oregonstate.edu

MICHAEL B. WARD
Western Oregon University
Monmouth, OR 97361
wardm@wou.edu

The wildly popular Sudoku puzzles [**2**] are 9×9 arrays divided into nine 3×3 subarrays or blocks. Digits 1 through 9 appear in some of the entries. Other entries are blank. The goal is to fill the blank entries with the digits 1 through 9 in such a way that each digit appears exactly once in each row and in each column, and in each block. TABLE 1 gives an example of a completed Sudoku puzzle.

TABLE 1: A completed Sudoku puzzle

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

One proves in introductory group theory that every element of any group appears exactly once in each row and once in each column of the group's operation or Cayley table. (In other words, any Cayley table is a Latin square.) Thus, every Cayley table has two-thirds of the properties of a Sudoku table; only the subdivision of the table into blocks that contain each element exactly once is in doubt. A question naturally leaps to mind: When and how can a Cayley table be arranged in such a way as to satisfy the additional requirements of being a Sudoku table? To be more specific, group elements labeling the rows and the columns of a Cayley table may be arranged in any order. Moreover, in defiance of convention, row labels and column labels need not be in the same order. Again we ask, when and how can the row and column labels be arranged so that the Cayley table has blocks containing each group element exactly once?

For example, TABLE 2 shows that the completed Sudoku puzzle in TABLE 1 is actually a Cayley table of $\mathbb{Z}_9 := \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ under addition modulo 9. (We use 9 instead of the usual 0 in order to maintain the Sudoku-like appearance.)

Math. Mag. **83** (2010) 130–139. doi:10.4169/002557010X482899.

TABLE 2: A Cayley table of $\mathbb{Z}_9$ with Sudoku properties

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

As a second example, consider A_4, the alternating group on 4 symbols. We seek to arrange its elements as row and column labels so that the resulting Cayley table forms a Sudoku-like table, one in which the table is subdivided into blocks such that each group element appears exactly once in each block (as well as exactly once in each column and in each row, which, as noted, is always so in a Cayley table). TABLE 3 shows such an arrangement with 6×2 blocks. (In constructing the table, we operate with the row label on the left and column label on the right. Permutations are composed right to left. For example, the entry in row $(14)(23)$, column (134) is $(14)(23)(134) = (123)$.)

We say TABLES 2 and 3 are Cayley-Sudoku tables of $\mathbb{Z}_9$ and A_4, respectively. In general, a *Cayley-Sudoku table* of a finite group G is a Cayley table for G subdivided into uniformly sized rectangular blocks in such a way that each group element appears exactly once in each block.

Uninteresting Cayley-Sudoku tables can be made from any Cayley table of any group by simply defining the blocks to be the individual rows (or columns) of the table. Our goal in this note is to give three methods for producing interesting tables using cosets, thereby uncovering new applications of this fundamental idea. (Any introductory group theory text reviews the concept of cosets, for example, Gallian [**1**, Ch. 7].)

Cosets Revisit TABLE 2. The cyclic subgroup generated by 3 in $\mathbb{Z}_9$ is $\langle 3\rangle = \{9, 3, 6\}$. The right and left cosets of $\langle 3\rangle$ in $\mathbb{Z}_9$ are $\langle 3\rangle + 9 = \{9, 3, 6\} = 9 + \langle 3\rangle$, $\langle 3\rangle + 1 = \{1, 4, 7\} = 1 + \langle 3\rangle$, and $\langle 3\rangle + 2 = \{2, 5, 8\} = 2 + \langle 3\rangle$. With only a little prompting, we quickly see that the columns in each block are labeled by elements of right cosets of $\langle 3\rangle$ in $\mathbb{Z}_9$. Each set of elements labeling the rows of a block contains exactly one element from each left coset. Equivalently, the row labels partition $\mathbb{Z}_9$ into *complete sets of left coset representatives* of $\langle 3\rangle$ in $\mathbb{Z}_9$. (Momentarily we shall see why we bothered to distinguish between right and left.)

Reexamining TABLE 3 reveals a similar structure. Consider the subgroup $H := \langle (12)(34)\rangle = \{(1), (12)(34)\}$. We brush-up on composing permutations (right to left) by calculating the right coset $H(123) = \{(1)(123), (12)(34)(123)\} = \{(123), (243)\}$ and the corresponding left coset $\{(123)(1), (123)(12)(34)\} = \{(123), (134)\}$. In that fashion, we find the right cosets to be

$$H(1) = \{(1), (12)(34)\}, \qquad H(142) = \{(142), (134)\},$$
$$H(13)(24) = \{(13)(24), (14)(23)\}, \qquad H(132) = \{(132), (143)\},$$
$$H(123) = \{(123), (243)\}, \qquad H(234) = \{(234), (124)\},$$

TABLE 3: A Cayley table of A_4 with Sudoku properties

	(1)	(12)(34)	(13)(24)	(14)(23)	(123)	(243)	(142)	(134)	(132)	(143)	(234)	(124)
(1)	(1)	(12)(34)	(13)(24)	(14)(23)	(123)	(243)	(142)	(134)	(132)	(143)	(234)	(124)
(13)(24)	(13)(24)	(14)(23)	(1)	(12)(34)	(142)	(134)	(123)	(243)	(234)	(124)	(132)	(143)
(123)	(123)	(134)	(243)	(142)	(132)	(124)	(143)	(234)	(1)	(14)(23)	(12)(34)	(13)(24)
(243)	(243)	(142)	(123)	(134)	(143)	(234)	(132)	(124)	(12)(34)	(13)(24)	(1)	(14)(23)
(132)	(132)	(234)	(124)	(143)	(1)	(13)(24)	(14)(23)	(12)(34)	(123)	(142)	(134)	(243)
(143)	(143)	(124)	(234)	(132)	(12)(34)	(14)(23)	(13)(24)	(1)	(243)	(134)	(142)	(123)
(12)(34)	(12)(34)	(1)	(14)(23)	(13)(24)	(243)	(123)	(134)	(142)	(143)	(132)	(124)	(234)
(14)(23)	(14)(23)	(13)(24)	(12)(34)	(1)	(134)	(142)	(243)	(123)	(124)	(234)	(143)	(132)
(134)	(134)	(123)	(142)	(243)	(124)	(132)	(234)	(143)	(14)(23)	(1)	(13)(24)	(12)(34)
(142)	(142)	(243)	(134)	(123)	(234)	(143)	(124)	(132)	(13)(24)	(12)(34)	(14)(23)	(1)
(234)	(234)	(132)	(143)	(124)	(13)(24)	(1)	(12)(34)	(14)(23)	(142)	(123)	(243)	(134)
(124)	(124)	(143)	(132)	(234)	(14)(23)	(12)(34)	(1)	(13)(24)	(134)	(243)	(123)	(142)

while the left cosets are

$$(1)H = \{(1), (12)(34)\}, \qquad (243)H = \{(243), (142)\},$$
$$(13)(24)H = \{(13)(24), (14)(23)\}, \qquad (132)H = \{(132), (234)\},$$
$$(123)H = \{(123), (134)\}, \qquad (143)H = \{(143), (124)\}.$$

This time we know what to expect. Sure enough, the columns in TABLE 3 are labeled by the elements of the distinct *right* cosets of H in A_4 while the row labels partition A_4 into complete sets of *left* coset representatives of H in A_4.

Those examples illustrate our first general construction.

Before proceeding, let us agree upon a convention for labeling a Cayley table. When a set is listed in a row or column of the table, it is to be interpreted as the individual elements of that set being listed in separate rows or columns, respectively. For example, under that convention, the rows and columns of TABLE 2 could be labeled

	{9, 3, 6}	{1, 4, 7}	⋯
{9, 1, 2}			
{3, 4, 5}			
⋮			

where the label {9, 1, 2} is interpreted as the elements 9, 1, and 2 listed vertically, one per row, and {9, 3, 6} is interpreted as the elements 9, 3, and 6 listed horizontally, one per column.

CAYLEY-SUDOKU CONSTRUCTION 1. *Let G be a finite group. Assume H is a subgroup of G having order k and index n (so that $|G| = nk$). If $Hg_1, Hg_2, \ldots, Hg_n$ are the n distinct right cosets of H in G, then arranging the Cayley table of G with columns labeled by the cosets $Hg_1, Hg_2, \ldots, Hg_n$ and the rows labeled by sets $T_1, T_2, \ldots, T_k$ (as in* TABLE 4*) yields a Cayley-Sudoku table of G with blocks of dimension $n \times k$ if and only if $T_1, T_2, \ldots, T_k$ partition G into complete sets of left coset representatives of H in G.*

TABLE 4: Construction 1 using right cosets and left coset representatives

	Hg_1	Hg_2	⋯	Hg_n
T_1				
T_2				
⋮				
T_k				

Furthermore, if $y_1H, y_2H, \ldots, y_nH$ are the n distinct left cosets of H in G, then arranging the Cayley table of G with rows labeled by the cosets $y_1H, y_2H, \ldots, y_nH$ and the columns labeled by sets $R_1, R_2, \ldots, R_k$ yields a Cayley-Sudoku table of G with blocks of dimension $k \times n$ if and only if $R_1, R_2, \ldots, R_k$ partition G into complete sets of right coset representatives of H in G.

Note that the second version of the Construction 1 is dual to the first, obtained by reversing the left with right and rows with columns.

Now let us prove the correctness of the construction for the case of right cosets. An arbitrary block of the table, indexed by $T_h = \{t_1, t_2, \ldots, t_n\}$ and Hg_i, is the given the following table.

	Hg_i
t_1	t_1Hg_i
t_2	t_2Hg_i
$\vdots$	$\vdots$
t_n	t_nHg_i

The elements in the block are the elements of the set $B := t_1Hg_i \cup t_2Hg_i \cup \cdots \cup t_nHg_i = (t_1H \cup t_2H \cup \cdots \cup t_nH)g_i$, the equality being a routine exercise. We want to show that the elements of G appear exactly once in that block if and only if T_h is a complete set of left coset representatives of H in G.

If T_h is a complete set of left coset representatives, then $t_1H \cup t_2H \cup \cdots \cup t_nH = G$. So we have $B = Gg_i = G$. Thus every element of G appears in every block. But the number of entries in the block is nk and the order of G is nk, so every element of G appears exactly once in each block.

On the other hand, if every element appears exactly once in each block, then $B = G$ and that gives us $G = Gg_i^{-1} = Bg_i^{-1} = t_1H \cup t_2H \cup \cdots \cup t_nH$. There are n cosets in this union, each having order k. Therefore, since the union of those cosets is the entire group G and $|G| = nk$, we must have n distinct cosets in the union. Thus, $\{t_1, t_2, \ldots, t_n\}$ is a complete set of left coset representatives of H in G, as claimed.

The first punch line of this section is that any proper nontrivial subgroup of a finite group gives rise to interesting Cayley-Sudoku tables by using Construction 1. Therefore, any finite group having a proper nontrivial subgroup (which is to say, any finite group whose order is not a prime) admits an interesting Cayley-Sudoku table.

We now introduce another construction using cosets. This construction is *like-handed,* that is, it uses left cosets and left coset representatives (or right and right). Construction 1 is *cross-handed,* meaning that it uses right cosets and left coset representatives or left cosets and right coset representatives.

For this construction, we recall a standard group theoretic definition. For any subgroup H of a group G and for any $g \in G$, $g^{-1}Hg := \{g^{-1}hg : h \in H\}$ is called a *conjugate* of H and is denoted H^g. It is routine to show H^g is a subgroup of G. Instead of using cross-handed coset representatives, it works to use like-handed coset representatives of conjugates.

CAYLEY-SUDOKU CONSTRUCTION 2. *Assume H is a subgroup of G having order k and index n. Also suppose $t_1H, t_2H, \ldots, t_nH$ are the distinct left cosets of H in G. Arranging the Cayley table of G with columns labeled by the cosets $t_1H, t_2H, \ldots, t_nH$ and the rows labeled by sets $L_1, L_2, \ldots, L_k$ yields a Cayley-Sudoku table of G with blocks of dimension $n \times k$ if and only if $L_1, L_2, \ldots, L_k$ are complete sets of left coset representatives of H^g for all $g \in G$.*

Since one or two interesting subtleties arise, we verify the correctness of Construction 2. Consider an arbitrary block in the table, indexed by $L_i := \{g_{i1}, g_{i2}, \ldots, g_{in}\}$ and t_jH. Note, however, that $t_jH = t_jHt_j^{-1}t_j = H^{t_j^{-1}}t_j$. Thus, our block is

$$g_{i1}H^{t_j^{-1}}t_j \cup g_{i2}H^{t_j^{-1}}t_j \cup \cdots \cup g_{1n}H^{t_j^{-1}}t_j = \left(g_{i1}H^{t_j^{-1}} \cup \cdots \cup g_{in}H^{t_j^{-1}}\right)t_j.$$

TABLE 5: Construction 2 using left cosets and left coset representatives

	t_1H	t_2H	$\cdots$	t_nH
L_1				
L_2				
$\vdots$				
L_k				

Suppose $L_1, L_2, \ldots, L_k$ are complete sets of left coset representatives of H^g for all $g \in G$, then $L_i = \{g_{i1}, g_{i2}, \ldots, g_{in}\}$ is a complete set of left coset representatives of $H^{t_j^{-1}}$. Therefore, just as in the verification of Construction 1, we can show every element of G appears exactly once in the block.

Conversely, suppose every element of G appears exactly once in each block. Once again arguing as in Construction 1, we conclude each L_i is a complete set of left coset representatives of H^{t_j} for every j.

In order to finish, we need a (known) result of independent group theoretic interest. Namely, with notation as in the construction, for every $g \in G$, there exists t_j such that $H^g = H^{t_j^{-1}}$. To see this, let $g \in G$, then g^{-1} is in some left coset of H, say t_jH. Thus, $g^{-1} = t_jh$ for some $h \in H$. Armed with the observation $hHh^{-1} = H$ (easily shown since H is a subgroup), we hit our target: $H^g = g^{-1}Hg = (t_jh)H(h^{-1}t_j^{-1}) = t_j(hHh^{-1})t^{-1} = t_jHt_j^{-1} = H^{t_j^{-1}}$. Combining this with the preceding paragraph, we can conclude $L_1, L_2, \ldots, L_k$ are complete sets of left coset representatives of H^g for all $g \in G$, as claimed.

We invite the reader to formulate and verify a right-handed version of Construction 2. We also raise an interesting and, evidently, nontrivial question for further investigation. Under what circumstances can one decompose a finite group G in the way required by Construction 2?

There is one easy circumstance. If H is a normal subgroup of a G, then it is not difficult to show $H^g = H$ for every $g \in G$ [**1**, Ch. 9]. Thus, decomposing G into complete sets of left coset representatives of H will do the trick. Sadly, in that case, Construction 2 gives the same Cayley-Sudoku table as Construction 1 because the left cosets indexing the columns equal the corresponding right cosets by normality.

Happily, we know of one general circumstance in which we can decompose G in the desired way to obtain new Cayley-Sudoku tables. It is described in the following proposition, stated without proof.

PROPOSITION. *Suppose the finite group G contains subgroups $T := \{t_1, t_2, \ldots, t_n\}$ and $H := \{h_1, h_2, \ldots, h_k\}$ such that $G = \{th : t \in T, h \in H\} := TH$ and $T \cap H = \{e\}$, then the elements of T form a complete set of left coset representatives of H and the cosets $Th_1, Th_2, \ldots, Th_k$ decompose G into complete sets of left coset representatives of H^g for every $g \in G$.*

In other words, from the Proposition, Construction 2 applies when we set $L_i := Th_i$ and use the left cosets $t_1H, t_2H, \ldots, t_nH$. Let us try it out on the group S_4, which can be decomposed in terms of a subgroup of order 8 and a subgroup generated by a single permutation of order 3.

Let $H = \langle(123)\rangle = \{(1), (123), (132)\}$ and $T = \{(1), (12)(34), (13)(24), (14)(23), (24), (1234), (1432), (13)\}$. One can check (by brute force, if necessary) that H and T

are subgroups of S_4 satisfying the hypotheses of the Proposition. Therefore, according to Construction 2, the following table yields a Cayley-Sudoku table of S_4.

	H	$(12)(34)H$	$(13)(24)H$	$(14)(23)H$	$(24)H$	$(1234)H$	$(1432)H$	$(13)H$
T								
$T(123)$								
$T(132)$								

Seeking to be convinced that is a new Cayley-Sudoku table, not of the kind produced by Construction 1, we examine the sets indexing the columns and rows. In Construction 1, the sets indexing columns are right cosets of some subgroup or else the sets indexing the rows are left cosets of some subgroup. In our table, the only subgroup indexing the columns is H and most of the remaining index sets are not right cosets of H. For example, $(12)(34)H \neq H(12)(34)$ and so it is not a right coset of H. Similar consideration of the sets indexing the rows shows Construction 1 was not on the job here.

Extending Cayley-Sudoku tables Our final construction shows a way to extend a Cayley-Sudoku table of a subgroup to a Cayley-Sudoku table of the full group.

CAYLEY-SUDOKU CONSTRUCTION 3. *Let G be a finite group with a subgroup A. Let $C_1, C_2, \ldots, C_k$ partition A and $R_1, R_2, \ldots, R_n$ partition A in any way such that the following table is a Cayley-Sudoku table of A.*

	C_1	C_2	$\cdots$	C_k
R_1				
R_2				
$\vdots$				
R_n				

If $\{l_1, l_2, \ldots, l_t\}$ and $\{r_1, r_2, \ldots, r_t\}$ are complete sets of left and right coset representatives, respectively, of A in G, then arranging the Cayley table of G with columns labeled with the sets $C_i r_j$, $i = 1, \ldots, k$, $j = 1, \ldots, t$ and the b^{th} block of rows labeled with $l_j R_b$, $j = 1, \ldots, t$, for $b = 1, \ldots, n$ (as in TABLE 6*) yields a Cayley-Sudoku table of G with blocks of dimension $tk \times n$.*

Proving the correctness of Construction 3 is quite like the proof for Construction 1. We leave it as an exercise for the reader and proceed to an example. Working in the group $\mathbb{Z}_8 := \{0, 1, 2, 3, 4, 5, 6, 7\}$ under addition modulo 8, let us apply Construction 1 to form a Cayley-Sudoku table for the subgroup $\langle 2 \rangle$ and then extend that table to a Cayley-Sudoku table of $\mathbb{Z}_8$ via Construction 3.

Observe that $\langle 4 \rangle = \{0, 4\}$ is a subgroup of $\langle 2 \rangle = \{0, 2, 4, 6\}$. The left and right cosets of $\langle 4 \rangle$ in $\langle 2 \rangle$ are $0 + \langle 4 \rangle = \{0, 4\} = \langle 4 \rangle + 0$ and $2 + \langle 4 \rangle = \{2, 6\} = \langle 4 \rangle + 0$. Thus, $\{0, 2\}$ and $\{4, 6\}$ partition $\langle 2 \rangle$ into complete sets of right coset representatives. Applying Construction 1, wherein elements of left cosets label the rows and right coset representatives label the columns, yields TABLE 7.

Now the left and right cosets of $\langle 2 \rangle$ in $\mathbb{Z}_8$ are $0 + \langle 2 \rangle = \{0, 2, 4, 6\} = \langle 2 \rangle + 0$ and $1 + \langle 2 \rangle = \{1, 3, 5, 7\} = \langle 2 \rangle + 1$. Accordingly, $\{0, 1\}$ is a complete set of left and right coset representatives of $\langle 2 \rangle$ in $\mathbb{Z}_8$. According to Construction 3, TABLE 8 should be (and is, much to our relief) a Cayley-Sudoku table of $\mathbb{Z}_8$. For easy comparison with TABLE 6, rows and columns are labeled both with sets and with individual elements.

TABLE 6: Construction 3

	C_1r_1	C_2r_1	$\cdots$	C_kr_1	C_1r_2	$\cdots$	C_kr_2	$\cdots$	C_1r_t	$\cdots$	C_kr_t
l_1R_1 l_2R_1 $\vdots$ l_tR_1											
l_1R_2 $\vdots$ l_tR_2											
$\vdots$											
l_1R_n $\vdots$ l_tR_n											

TABLE 7: Construction 1 applied

	0	2	4	6
0	0	2	4	6
4	4	6	0	2
2	2	4	6	0
6	6	0	2	4

We chose $\mathbb{Z}_8$ for our example because it has enough subgroups to make Construction 3 interesting, yet the calculations are easy to do and the resulting table fits easily on a page. In one sense, however, the calculations are too easy. Since $\mathbb{Z}_8$ is abelian, all the corresponding right and left cosets of any subgroup are equal. (In other words, all the subgroups are normal.) Thus, the role of right versus left in Construction 3 is obscured. The interested reader may wish to work out an example where right and left cosets are different. For instance, in S_4, one could consider the subgroup $A := \{(1), (12)(34), (13)(24), (14)(23), (24), (1234), (1432), (13)\}$. Use Construction 1 with the subgroup $\langle(24)\rangle$ of A to obtain a Cayley-Sudoku table of A, then

TABLE 8: A Cayley-Sudoku table of $\mathbb{Z}_8$ from Construction 3

		$\{0,2\}+0$		$\{0,2\}+1$		$\{4,6\}+0$		$\{4,6\}+0$	
		0	2	1	3	4	6	5	7
$0+\{0,4\}$	0	0	2	1	3	4	6	5	7
	4	4	6	5	7	0	2	1	3
$1+\{0,4\}$	1	1	3	2	4	5	7	6	0
	5	5	7	6	0	1	3	2	4
$0+\{2,6\}$	2	2	4	3	5	6	0	7	1
	6	6	0	7	1	2	4	3	5
$1+\{2,6\}$	3	3	5	4	6	7	1	0	2
	7	7	1	0	2	3	5	4	6

apply Construction 3 to extend that table to a Cayley-Sudoku table of S_4. The associated computations are manageable (barely, one might think by the end!) and the roles of right and left are more readily apparent.

TABLE 8 is a new sort of Cayley-Sudoku table, one not produced by either of Constructions 1 or 2. To see why, recall that in Construction 1 and 2 (including the right-handed cousin of 2), either the columns or the rows in the blocks are labeled by cosets of a subgroup. One of those cosets is, of course, the subgroup itself. However, we easily check that none of the sets labeling columns or rows of the blocks in TABLE 8 are subgroups of $\mathbb{Z}_8$.

A puzzle The ubiquitous Sudoku leads many students to treat the familiar exercise of filling in the missing entries of a partial Cayley table as a special sort of Sudoku puzzle. In a recent group theory course taught by the third author, several students explained how they deduced missing entries in such an exercise [**1**, p. 55, exercise 25], by writing "I Sudokued them." Meaning they applied Sudoku-type logic based on the fact that rows and columns of a Cayley table contain no repeated entries.

We extend that notion by including a Cayley-Sudoku puzzle for the reader. It requires both group theoretic and Sudoku reasoning. The group theory required is very elementary. (In particular, one need not use the classification of groups of order 8.)

The puzzle has three parts, one for entertainment and two to show this is truly a new sort of puzzle. First, complete TABLE 9 with 2×4 blocks as indicated so that it becomes a Cayley-Sudoku table. Do not assume *a priori* that TABLE 9 was produced by any of Constructions 1–3. Second, show group theoretic reasoning is actually needed in the puzzle by completing TABLE 9 so that it satisfies the three Sudoku properties for the indicated 2×4 blocks but is not the Cayley table of any group. Third, show Sudoku reasoning is required by finding another way to complete TABLE 9 so that it is a Cayley table of some group, but not a Cayley-Sudoku table.

TABLE 9: A Cayley-Sudoku puzzle (answers on p. 147)

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

Ideas for further study By exhaustive (in more ways than one) analysis of cases, the authors can show that the only 9×9 Cayley-Sudoku tables are those resulting from Construction 1. Is the same true for $p^2 \times p^2$ Cayley-Sudoku tables where p is a prime?

All the constructions of Cayley-Sudoku tables known to the authors, including some not presented in this paper, ultimately rely on cosets and coset representatives. Are there Cayley-Sudoku constructions that do not use cosets and coset representatives?

Related to the previous question, how does one create a single block of a Cayley-Sudoku table? That is, if G is a group with subsets (not necessarily subgroups) K and H such that $|G| = |K||H|$, what tractable conditions guarantee that $KH = G$?

Can a Cayley-Sudoku table of a group be used to construct a Cayley-Sudoku table of a subgroup or a factor group? Can a Cayley-Sudoku table of a factor group be used to construct a Cayley-Sudoku table of the original group?

Are there efficient algorithms for generating interesting Cayley-Sudoku puzzles?

Making the definition of Cayley-Sudoku tables less restrictive can lead to some interesting examples. For instance if the definition of Cayley-Sudoku tables is altered so that the individual blocks of the table do not have to be of fixed dimension we obtain in TABLE 10 an example of a *generalized Cayley-Sudoku* table of the group $\mathbb{Z}_8$.

TABLE 10: A Generalized Cayley-Sudoku table

	0	4	1	3	5	7	2	6
0	0	4	1	3	5	7	2	6
1	1	5	2	4	6	0	3	7
2	2	6	3	5	7	1	4	0
3	3	7	4	6	0	2	5	1
4	4	0	5	7	1	3	6	2
5	5	1	6	0	2	4	7	3
6	6	2	7	1	3	5	0	4
7	7	3	0	2	4	6	1	5

What is a construction method for such generalized Cayley-Sudoku tables? How about for *jigsaw Cayley-Sudoku tables* wherein the blocks are not rectangles?

Perhaps most interesting of all, find other circumstances under which Construction 2 applies.

Acknowledgment This note is an outgrowth of the senior theses of the first two authors written under the supervision of the third author. We thank one another for many hours of satisfying and somewhat whimsical mathematics. We also thank the students at DigiPen Institute of Technology, where the first author was a guest speaker, for suggesting that we make a Cayley-Sudoku puzzle.

REFERENCES

1. J. A. Gallian, *Contemporary Abstract Algebra*, 7th ed., Brooks/Cole, Belmont, CA, 2010.
2. R. Wilson, The Sudoku Epidemic, *FOCUS* **26** (2006) 5–7.

Summary The popular Sudoku puzzles are 9 by 9 tables divided into nine 3 by 3 sub-tables or blocks. Digits 1 through 9 appear in some of the entries. Other entries are blank. The goal is to fill the blank entries with digits 1 through 9 in such a way that each digit appears exactly once in each row and in each column and in each block. Cayley tables are group operation tables. As such, each group element always appears exactly once in each row and in each column, two thirds of being a Sudoku-like table. Cayley-Sudoku tables are Cayley tables arranged in such a way as to satisfy the additional requirement. Namely, the Cayley table may be divided into blocks with each group element appearing exactly once in each block. We show three ways to construct nontrivial Cayley-Sudoku tables for finite groups, present a Cayley-Sudoku puzzle to solve, and give several suggestions for additional investigations.

Proof Without Words: Mengoli's Series

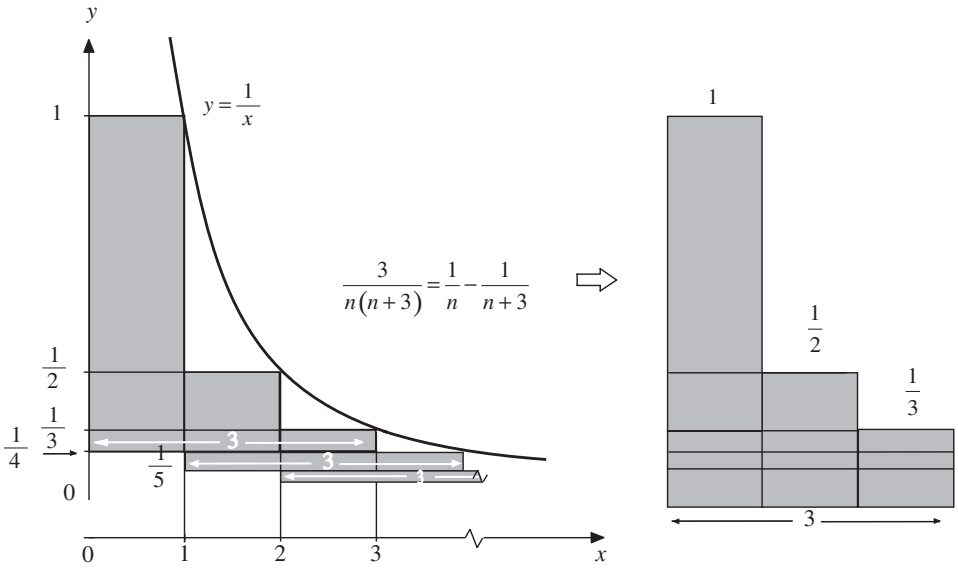
Pietro Mengoli posed the problem of summing the series below and found the sum for the cases $1 \le k \le 10$. [**1**]

$$\sum_{n=1}^{\infty} \frac{k}{n(n+k)} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k}$$

We illustrate the case $k = 3$.

$$\frac{3}{n(n+3)} = \frac{1}{n} - \frac{1}{n+3} \Rightarrow$$

$$\sum_{n=1}^{\infty} \frac{3}{n(n+3)} = 1 - \cancel{\frac{1}{4}} + \frac{1}{2} - \cancel{\frac{1}{5}} + \frac{1}{3} - \cancel{\frac{1}{6}} + \cancel{\frac{1}{4}} - \cdots = 1 + \frac{1}{2} + \frac{1}{3}$$



—ÁNGEL PLAZA
ULPGC, 35017-LAS PALMAS G.C., SPAIN

REFERENCE

1. Pietro Mengoli, *Novae quadraturae arithmeticae, seu de additione fractionum*. Bologna, 1650.

Math. Mag. **83** (2010) 140. doi:10.4169/002557010X482907.

Triangle Equalizers

DIMITRIOS KODOKOSTAS
Technological Education Institute of Larissa
Larissa, 42110 Greece
dkodokostas@gmail.com

An interesting question from Euclidean plane geometry concerns the existence of lines called *equalizers* that bisect *both* the perimeter *and* the area of a given region. The answer is that equalizers always exist. Some issues regarding the meaning of length and area should be handled, but relaxing them, a quick proof is not too difficult [**2**].

Expanding on the question of existence and focusing on triangles, G. Berzsenyi in *Quantum* [**1**] conjectured that no triangle has more than three equalizers. A companion conjecture by the Emeritus Professor H. Bailey states that no triangle has exactly two equalizers. Our aim in this article is to obtain the number of equalizers for an arbitrary triangle, proving the first conjecture (Proposition 1) and disproving the second one (Proposition 2). This will provide at the same time a fairly detailed account of the location of the equalizers. As a bonus, we witness the unexpected role of the angle $2\arcsin(-1+\sqrt{2}) \approx 48°56'23''$ in pinning down the number of triangle equalizers.

Searching for triangle equalizers As we look for the equalizers of a triangle, we use its *perimeter* and *area splitters*, that is lines that bisect the perimeter or the area of the triangle. With this terminology, a line is an equalizer whenever it is both an area and a perimeter splitter.

The reader may have seen *splitter* defined in a more restricted way, as a line through a *vertex* bisecting the perimeter, whereas a line with this property through the midpoint of a side has been called a *cleaver* [**3**]. We intend no such restrictions.

Interestingly, the equalizers are just the area splitters through the triangle's incenter. This fact alone greatly reduces the possible locations for an equalizer. Its proof is quite easy, and readers may wish to find it before proceeding.

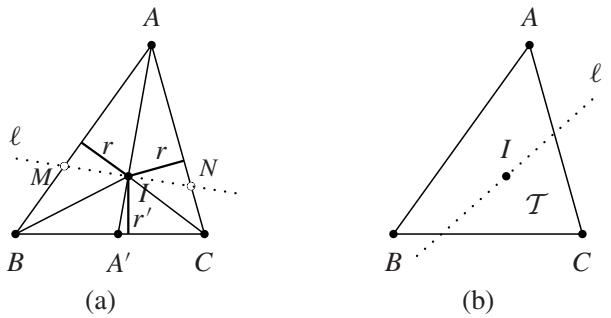


Figure 1 (a) All equalizers must go through the incenter (b) Any line through the incenter of a triangle always cuts off from it a triangular region

In $\triangle ABC$, any candidate for an equalizer, or a perimeter or area splitter, is a line ℓ missing at least one of the vertices, say A, intersecting the sides by this vertex at two points, say AB at M and AC at N as in FIGURE 1(a). The common point I of ℓ with the interior angle bisector AA' from A, has the same distance r from the sides AB and AC. Let r' be its distance from BC. Fixing the notation for the sequel, we denote by

Math. Mag. **83** (2010) 141–146. doi:10.4169/00255701OX482916.

a, b, c the side lengths of $\triangle ABC$, and by $\mathcal{A}(F)$ the area of a region F. Trivially, ℓ is a perimeter splitter of $\triangle ABC$ if and only if

$$AM + AN = \frac{a+b+c}{2}. \tag{1}$$

Also, ℓ being an area splitter of $\triangle ABC$ is equivalent to $\mathcal{A}(MAI) + \mathcal{A}(IAN) = \mathcal{A}(MIB) + (BIC) + \mathcal{A}(CIN)$, and writing each area as half the product of the side opposite to I with its corresponding altitude, we find that ℓ is an area splitter of $\triangle ABC$ if and only if

$$(2(AM + AN) - (c + b))\, r = ar'. \tag{2}$$

So, if ℓ is an equalizer of $\triangle ABC$, that is, both a perimeter and an area splitter, then (1) and (2) imply $r' = r$, thus I is the incenter of $\triangle ABC$.

Conversely, if I is the incenter of $\triangle ABC$, that is $r = r'$, then ℓ being an area splitter of $\triangle ABC$ is equivalent by (2) to $[2(AM + AN) - (c + b)])r = ar$, and so to $AM + AN = (a + b + c)/2$ which by (1) is equivalent to ℓ being a perimeter splitter of the triangle. We have proved:

LEMMA 1.

(a) *Any equalizer of a triangle goes through the incenter of the triangle.*

(b) *A line through the incenter of a triangle is an area splitter if and only if it is a perimeter splitter.*

(c) *The equalizers of a triangle are the area splitters through its incenter.*

Lines through the incenter Any line ℓ through the incenter I of $\triangle ABC$ splits it into two regions (FIGURE 1(b)), one of them triangular, say $\mathcal{T}$. Then by Lemma 1, ℓ is an equalizer exactly when $\mathcal{T}$ occupies half of the area Δ of $\triangle ABC$. Our plan is to rotate ℓ around I and spot positions for which ℓ becomes an equalizer by comparing $\mathcal{A}(\mathcal{T})$ with $\mathcal{A}(ABC)/2 = \Delta/2$.

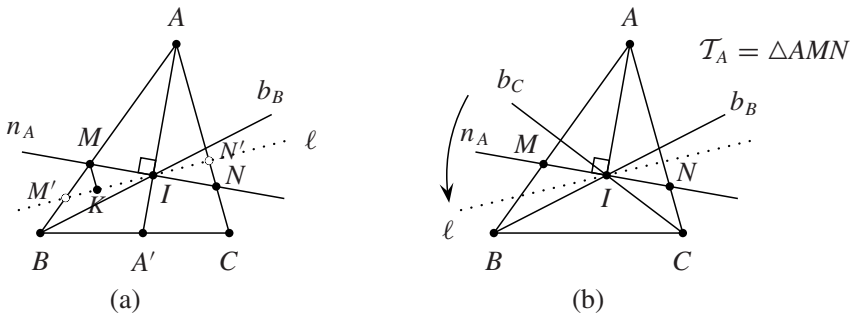


Figure 2 (a) The area $\mathcal{A}(AM'N')$ increases as ℓ turns towards B. (b) How many equalizers lie between angle bisectors b_B and b_C?

As ℓ rotates about the incenter of $\triangle ABC$ and cuts off various triangles, it passes six important landmarks, as far as area is concerned. These are the three angle bisectors and the three lines through the incenter that are normal to these bisectors.

Let b_A, b_B, b_C be the bisectors and n_A, n_B, n_C their corresponding normals through I. For convenience, we will refer to n_A as the *normal for* A, instead of the longer but more accurate "line through the incenter that is normal to the bisector through A." Every line through I, other than these six, lies in a sector delimited by two of them.

At the risk of spoiling the drama, we reveal that n_A can be an equalizer only when $A \le 2\arcsin(-1+\sqrt{2})$, and then only if B is chosen carefully.

It turns out that the area of the triangle cut off, $\mathcal{A}(\mathcal{T})$, has local maxima at the bisectors and local minima at the normals.

LEMMA 2. *For lines ℓ through the incenter I of $\triangle ABC$, the area $\mathcal{A}(\mathcal{T})$ of the triangle $\mathcal{T}$ cut off from $\triangle ABC$ by ℓ increases as we rotate ℓ from the normal for a vertex towards the bisector through another vertex.*

Proof. For any line ℓ between n_A and b_B as in FIGURE 2(a), we first show that ℓ cuts off more area than n_A. When we move from n_A to $\ell = M'N'$, the cut-off triangle gains $\triangle IMM'$ and loses $\triangle INN'$. Draw MK parallel to AC to cut $M'I$ at K and observe that triangles $\triangle IMK$ and $\triangle INN'$ are congruent. The amount of area gained is the difference, $\mathcal{A}(\triangle M'MK) = MM' \cdot MK \cdot \sin(\angle M'MK)/2$. As M' moves toward B, this extra area grows larger, so $\mathcal{A}(\mathcal{T})$ increases as claimed. ∎

We now compare $\Delta/2$ with $\mathcal{A}(\mathcal{T})$ for each of these landmark lines. This allows us to count the number of equalizers in all regions $\langle b_i, n_j\rangle$ between some bisector b_i from a vertex and the normal n_j for another vertex: If for both positions of ℓ we have $\mathcal{A}(\mathcal{T}) > \Delta/2$, then there can be no equalizer in that region. The same is true if $\mathcal{A}(\mathcal{T}) < \Delta/2$ for both positions of ℓ. On the other hand, if $\mathcal{A}(\mathcal{T}) > \Delta/2$ for one position of ℓ, and $\mathcal{A}(\mathcal{T}) < \Delta/2$ for the other, then there exists exactly one position of ℓ in $\langle b_i, n_j\rangle$ for which $\mathcal{A}(\mathcal{T}) = \Delta/2$, that is, there exists a unique equalizer ℓ in $\langle b_i, n_j\rangle$.

This area comparison is easy whenever ℓ is a bisector. For example, for the bisector b_A and the two triangles $AA'B$, $AA'C$ formed (FIGURE 2(b)), we have

$$\frac{\mathcal{A}(AA'B)}{\mathcal{A}(AA'C)} = \frac{A'B}{A'C} = \frac{c}{b}.$$

So c is greater, equal to, or less than b as $\mathcal{A}(AA'B)$ is greater, equal to, or less than $\Delta/2$. Thus, a bisector is an equalizer exactly when the two sides by the vertex of the bisector are equal in length.

As an immediate consequence of all the above we have:

PROPOSITION 1. *All equalizers of a triangle ABC go through its incenter, and their number is analyzed in* TABLE 1 *where $a \le b \le c$, $\Delta = \mathcal{A}(ABC)$ and $\mathcal{T}_A$ is the triangular region cut off from $\triangle ABC$ by the normal line for A.*

In particular, Berzenyi's conjecture is true, that is no triangle can have more than 3 *equalizers.*

Whenever ℓ is a normal line, the comparison between $\mathcal{T}$ and $\Delta/2$ is much harder. Assuming without loss of generality that $a \le b \le c$ as in Proposition 1, we need only check the case of n_A. From now on it will be more convenient to work with angles rather than with sides and to transform $a < b \le c$ to its equivalent $A < B \le C$ which in turn is equivalent to

$$A \in (0, \pi/3),\ B \in (A, (\pi - A)/2] \tag{3}$$

The crucial comparison is given in the following Lemma:

LEMMA 3. *Let $A_0 = 2\arcsin(-1+\sqrt{2})(\approx 48°56'23'')$. Then for all $A \in (0, A_0]$ there exists a unique root B_A of $\cot(x/2)\tan(A/2 + x/2)\cos^2(A/2) - 2 = 0$ in $(A, (\pi - A)/2]$, and $B_A = (\pi - A)/2$ exactly when $A = A_0$.*

If $\mathcal{A}(ABC) = \Delta$ is the area of an $A < B \le C$-angled triangle ABC and $\mathcal{A}(\mathcal{T}_A)$ is the triangular region cut off from the triangle by the normal line of A, then:

TABLE 1: The number of equalizers of $\triangle ABC$, $(a \le b \le c)$ with bisectors b_A, b_B, b_C and normals n_A, n_B, n_C. By $\langle b_A, b_B\rangle$, we mean the set of lines in the region between b_A and b_B that does not contain C or AB, and excluding n_C. Similarly, $\langle n_A, b_B\rangle$ is the set of lines between n_A and b_B, and so on. The triangle cut off from $\triangle ABC$ by n_A is $\mathcal{T}_A$ and $\Delta = \mathcal{A}(ABC)$.

<table>
<tr><th></th><th>bisectors</th><th>normals</th><th>$\langle b_B, b_C\rangle$</th><th>$\langle b_C, b_A\rangle$</th><th>$\langle b_A, b_B\rangle$</th><th>$\triangle ABC$</th></tr>
<tr><td rowspan="3">$a < b \le c$</td><td rowspan="3">0 (when $b \neq c$)
1 : b_A (when $b = c$)</td><td>0 (if $\frac{\Delta}{2} > \mathcal{A}(\mathcal{T}_A)$)</td><td>$1 \in \langle n_A, b_B\rangle$
$1 \in \langle n_A, b_C\rangle$</td><td rowspan="3">$1 \in \langle n_B, b_A\rangle$ (when $b \neq c$)
0 (when $b = c$)</td><td rowspan="3">0</td><td rowspan="3">3 or 2 or 1</td></tr>
<tr><td>1 : n_A (if $\frac{\Delta}{2} = \mathcal{A}(\mathcal{T}_A)$)</td><td>0</td></tr>
<tr><td>0 (if $\frac{\Delta}{2} < \mathcal{A}(\mathcal{T}_A)$)</td><td>0</td></tr>
<tr><td>$a = b < c$</td><td>1 : b_C</td><td>0</td><td>$1 \in \langle n_A, b_B\rangle$</td><td>0</td><td>$1 \in \langle n_C, b_B\rangle$</td><td>3</td></tr>
<tr><td>$a = b = c$</td><td>3: b_A, b_B, b_C</td><td>0</td><td>0</td><td>0</td><td>0</td><td>3</td></tr>
</table>

(1) *If $A \in (0, A_0]$, then $\frac{\Delta}{2}$ is greater than, equal to, or less than $\mathcal{A}(\mathcal{T}_a)$ as the angle B is less than, equal to, or greater than B_A.*

(2) *If $A \in (A_0, \pi/3)$, then* $\dfrac{\Delta}{2} > \mathcal{A}(\mathcal{T}_a)$.

Granted this Lemma, we restate Proposition 1 in terms of angles rather than sides:

PROPOSITION 2. *All equalizers of a triangle ABC go through its incenter and their number is analyzed in* TABLE 2, *where $A \le B \le C$, $A_0 = 2\arcsin(-1 + \sqrt{2})$ ($\approx 48°56'23''$), and for all $A \in (0, A_0]$ the number B_A is the unique root of $\cot(x/2)\tan(A/2 + x/2)\cos^2(A/2) - 2 = 0$ in $(A, (\pi - A)/2]$.*

Moreover, all entries except for $A < B = C$, $A = A_0$, $B_A < B$ can occur, and these are all that can. In particular, Bailey's conjecture is not true. In other words there exist triangles with exactly 2 equalizers, one being the normal line corresponding to the smallest angle.

It only remains to prove Lemma 3. It will pay off to employ a bit of trigonometry and calculus. To this end:

Note that $\mathcal{T}_A$ is isosceles with apex A and corresponding attitude AI (FIGURE 2b), so $\mathcal{A}(\mathcal{T}_A) = AI^2 \tan(A/2)$. But it is well known (and easy to verify) that $AI = 4R\sin(B/2)\sin(C/2)$ where R is the circumradius of $\triangle ABC$. So

$$\mathcal{A}(\mathcal{T}_a) = \frac{16R^2 \sin(\frac{A}{2})\sin^2(\frac{B}{2})\sin^2(\frac{C}{2})}{\cos(\frac{A}{2})}$$

On the other hand, it is standard knowledge that

$$\Delta = 2R^2 \sin(A)\sin(B)\sin(C),$$

so

$$\frac{\Delta}{2} - \mathcal{A}(\mathcal{T}_a) = \frac{8R^2 \sin\left(\frac{A}{2}\right)\sin^2\left(\frac{B}{2}\right)\sin^2\left(\frac{C}{2}\right)}{\cos(\frac{A}{2})}\left(\cot\left(\frac{B}{2}\right)\cot\left(\frac{C}{2}\right)\cos^2\left(\frac{A}{2}\right) - 2\right)$$

Since the fraction on the right is positive and $C = \pi - A - B$, the sign of $\Delta/2 - \mathcal{A}(\mathcal{T}_A)$ is that of $\cot(B/2)\tan((A + B)/2)\cos^2(A/2) - 2$. Recall that by

TABLE 2: The number of equalizers of $\triangle ABC$, $(a \le b \le c)$ with bisectors b_A, b_B, b_C and normals n_A, n_B, n_C. By $\langle b_A, b_B\rangle$, we mean the set of lines in the region between b_A and b_B that does not contain C or AB, and excluding n_C. Similarly, $\langle n_A, b_B\rangle$ is the set of lines between n_A and b_B, and so on. The angle A_0 is $2\arcsin(-1+\sqrt{2}) \approx 48°56'23''$ and B_A is the unique root of $\cot(x/2)\tan(A/2+x/2)\cos^2(A/2) - 2 = 0$ in $(A, (\pi - A)/2]$.

<table>
<tr><th colspan="3"></th><th>bs</th><th>ns</th><th>$\langle b_B, b_C\rangle$</th><th>$\langle b_C, b_A\rangle$</th><th>$\langle b_A, b_B\rangle$</th><th>$\triangle ABC$</th></tr>
<tr><td rowspan="5">$A<B\le C$</td><td rowspan="4">$0<A\le A_0$</td><td>$B<B_A$</td><td rowspan="4">0
$(B\neq C)$

$1 : b_A$
$(B=C)$</td><td rowspan="2">0</td><td>$1 \in \langle n_A, b_B\rangle$
$1 \in \langle n_A, b_C\rangle$</td><td rowspan="4">$1 \in \langle n_B, b_A\rangle$
$(B\neq C)$

0
$(B=C)$</td><td rowspan="4">0</td><td>3</td></tr>
<tr><td>$B=B_A$</td><td>$1 : b_A$</td><td>0</td><td>2</td></tr>
<tr><td>$B_A<B$</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>$A_0<A<\frac{\pi}{3}$</td><td rowspan="2">0</td><td>$1 \in \langle n_A, b_B\rangle$
$1 \in \langle n_A, b_B\rangle$</td><td>3</td></tr>
<tr><td>$A=B<C$</td><td colspan="2"></td><td>$1 :\ b_C$</td><td>0</td><td>$1 \in \langle n_A, b_B\rangle$</td><td>0</td><td>$1\in\langle n_B, b_A\rangle$</td><td>3</td></tr>
<tr><td>$A=B=C$</td><td colspan="2"></td><td>3 : all</td><td>0</td><td>0</td><td>0</td><td>0</td><td>3</td></tr>
</table>

(3) $A \in (0, \pi/3)$ and $B \in (A, (\pi - A)/2]$. Thus for any such A, $\Delta/2 - \mathcal{A}(\mathcal{T}_A)$ has the same sign as $f(B)$ where f is the following function of the variable B :

$$f(B) = \cot\left(\frac{B}{2}\right)\tan\left(\frac{A+B}{2}\right)\cos^2\left(\frac{A}{2}\right) - 2, \quad B \in \left[A, \frac{\pi - A}{2}\right].$$

It is more or less straightforward to determine the sign of $f(B)$. Note first that f is strictly decreasing since

$$f'(B) = \frac{-\cos^2\left(\frac{A}{2}\right)\sin\left(\frac{A}{2}\right)\cos\left(\frac{A}{2}+B\right)}{2\sin^2\left(\frac{B}{2}\right)\cos^2\left(\frac{A+B}{2}\right)} < 0, \quad \text{for } B \in \left[A, \frac{\pi - A}{2}\right].$$

Next, notice that the sign of f on the left endpoint A of its domain of definition is always positive since after some trivial calculations

$$f(A) = \frac{2\left(\cos^2\left(\frac{A}{2}\right) - 1\right)^2}{\cos(A)} > 0. \tag{4}$$

Similar calculations for the right endpoint $(\pi - A)/2$ of the domain give

$$f\left(\frac{\pi - A}{2}\right) = \sin^2\left(\frac{A}{2}\right) + 2\sin\left(\frac{A}{2}\right) - 1.$$

The sign of this number can be determined for all $A \in (0, \pi/3)$. Indeed, since $A/2 \in (0, \pi/6)$, the range of $\sin(A/2)$ is the interval $(0, 1/2)$. Now the binomial $g(x) = x^2 + 2x - 1$ vanishes on its roots, $r_1 = -1 - \sqrt{2} < 0 < r_2 = -1 + \sqrt{2} < 1/2$, is negative between these roots, and positive elsewhere. So $f((\pi - A)/2)$ is negative, zero, or positive depending on whether $0 < \sin(A/2) < r_2$, $\sin(A/2) = r_2$, or $r_2 < \sin(A/2) < 1/2$ respectively. If we call A_0 the solution of $\sin(A/2) = r_2 = -1 + \sqrt{2}$ in $(0, \pi/3)$, so that $A_0 = 2\arcsin(-1 + \sqrt{2})$, the last result can be summarized as follows:

$$\left(f\left(\frac{\pi - A}{2}\right) <,\ =,\ >\ 0\right) \Leftrightarrow \left(A \in (0, A_0),\ A = A_0,\ A \in \left(A_0, \frac{\pi}{3}\right)\right). \tag{5}$$

Relations (4) and (5) for the sign of f on the endpoints of its domain of definition, along with the fact that f is strictly decreasing in this domain $[A, (\pi - A)/2]$ imply that:

- If $A \in (0, A_0]$, there exists a unique root B_A of $f(B)$ in $(A, (\pi - A)/2]$ with $f(B) > 0$ for $B \in (A, B_A)$, and $f(B) < 0$ for $B \in (B_A, (\pi - A)/2]$. By the way, clearly B_A equals $(\pi - A)/2$ exactly when $A = A_0$.
- If $A \in (A_0, \pi/3)$, then $f(B) > 0$ for all $B \in [A, (\pi - A)/2]$.

Finally, from the fact that $\Delta/2 - \mathcal{A}(\mathcal{T}_A)$ and $f(B)$ have the same sign, these statements prove Lemma 3.

Concluding remarks As noted, Bailey's conjecture is false because the normal line corresponding to the smallest angle A of a triangle ABC can sometimes be its equalizer. This happens whenever A is less or equal to A_0, and the angle next in size, say B, assumes a unique value B_A. The rarity of such triangles is quite obvious and the lack of visual evidence, even with the help of a computer sketching program, seems to be the origin of the conjecture, as the story goes in Quantum.

We now know what kind of triangles exhibit the rare property of possessing exactly two equalizers. Their existence, as well as all other properties mentioned, can be verified with the help of a program like Geometer's Sketchpad (within its given computational limitations). We construct triangles ABC on a fixed base BC, a movable vertex A, a line ℓ through the incenter of $\triangle ABC$, the normal line n_A and the intersections of this line with the sides of $\triangle ABC$. Then we measure $\angle A$, $\mathcal{A}(ABC)$, $\mathcal{A}(\mathcal{T}_A)$, and $\mathcal{A}(\mathcal{T})$, where $\mathcal{T}_A$, and $\mathcal{T}$ are the triangles cut off from $\triangle ABC$ by n_A and ℓ. With this set-up, we can move the vertex A and rotate ℓ around the incenter, monitoring the values of $\mathcal{A}(ABC)/2 - \mathcal{A}(\mathcal{T}_A)$ and $\mathcal{A}(ABC)/2 - \mathcal{A}(\mathcal{T})$. A value of 0 means that n_A or ℓ is respectively an equalizer. To satisfy $a \le b \le c$ we must keep the vertex A outside the circles centered at B, C with radii equal to BC, and also keep it on the same side as C of the perpendicular bisector of BC.

One of the surprises is the angle $A_0 = 2\arcsin(-1 + \sqrt{2}) \approx 48°56'23''$ arising in connection with the quite symmetric question about the number of equalizers of a triangle. It is not known to the author if A_0 appears elsewhere in the literature with an equally important role.

REFERENCES

1. G. Berzsenyi, The equalizer of a triangle, a clever line that does double duty, *Quantum* **7** (March–April 1997) 51.
2. M. Gardner, *Penrose Tiles to Trapdoor Ciphers*, MAA, 1993.
3. R. Honsberger, *Episodes in Nineteenth and Twentieth Century Euclidean Geometry*, MAA, 1995.

Summary A triangle equalizer is a line bisecting both its area and perimeter. We provide a detailed account of equalizer locations, showing that there exist triangles with exactly one, two, or three equalizers, but no more. Triangles with exactly two equalizers are quite rare: Their smallest angle is less or equal to a particular angle of approximately 49 degrees, and the angle next in size is the unique root of a trigonometric equation in a specific interval depending on the smallest angle.

Puzzle Solutions for "Cosets and Cayley-Sudoku Tables"

In each solution to the problems on p. 138, the original puzzle entries appear in bold face for easy identification.

Part 1 TABLE 1 shows the solution. Clearly, it satisfies the three Sudoku properties. We will show it is the Cayley table of D_4, the dihedral group of order 8. We will regard D_4 as the group of symmetries of a square. Let R_{90} be a counterclockwise rotation about the center of the square and let H be a reflection across a line through the center of the square that is parallel to a side of the square. The eight elements of D_4 are R_{90}^0, R_{90}^1, R_{90}^2, R_{90}^3, HR_{90}^0, HR_{90}^1, HR_{90}^2, and HR_{90}^3. Numbering those elements 1 through 8 in the order given and then calculating the Cayley table gives TABLE 1. Thus, we have a Cayley-Sudoku table as claimed. By the way, it was obtained by applying Construction 1 to the subgroup $\langle R_{90} \rangle$.

TABLE 1: Cayley-Sudoku puzzle solution

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

Part 2 TABLE 2 visibly satisfies the Sudoku conditions. It is not a Cayley table. For otherwise, $1 \cdot 7 = 7$ would imply 1 is the identity, but $1 \cdot 2 \neq 2$.

TABLE 2: Sudoku-not-Cayley puzzle solution

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

Math. Mag. **83** (2010) 147–148. doi:10.4169/002557010X482899-s.

Part 3 TABLE 3 does not satisfy the Sudoku conditions. Blocks contain repeated entries. It is, however, a Cayley table. One can check that it is again the Cayley table of D_4. Just change the labeling of R_{90}^2 from 3 to 5 and of H from 5 to 3. TABLE 3 is the recalculated Cayley table.

TABLE 3: Cayley-not-Sudoku puzzle solution

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

The First Quickie

Mathematics Magazine, Volume 23, Number 4 (March–April 1950), pp. 210, 211.

QUICKIES

From time to time as space permits this department will publish problems which may be solved by laborious methods, but which with the proper insight may be disposed of with dispatch. Readers are urged to submit their favorite problems of this type, together with the elegant solution and the source, if known.

Q 1. After a typist had written ten letters and had addressed the ten corresponding envelopes, a careless mailing-clerk inserted the letters in the envelopes at random, one letter per envelope. What is the probability that exactly nine letters were inserted in the proper envelopes?

A 1. If nine letters are in the correct envelopes, the tenth must be also, so the probability is zero.

PROBLEMS

BERNARDO M. ÁBREGO, *Editor*
California State University, Northridge

Assistant Editors: SILVIA FERNÁNDEZ-MERCHANT, California State University, Northridge; JOSÉ A. GÓMEZ, Facultad de Ciencias, UNAM, México; ROGELIO VALDEZ, Facultad de Ciencias, UAEM, México; WILLIAM WATKINS, California State University, Northridge

PROPOSALS

To be considered for publication, solutions should be received by September 1, 2010.

1841. *Proposed by H. A. ShahAli, Tehran, Iran.*

Let $n \geq 3$ be a natural number. Prove that there exist n pairwise distinct natural numbers such that each of them divides the sum of the remaining $n - 1$ numbers.

1842. *Proposed by Bianca-Teodora Iordache, student, National College "Carol I," Craiova, Romania.*

In the interior of a square of side-length 3 there are several regular hexagons whose sum of perimeters is equal to 42 (the hexagons may overlap). Prove that there are two perpendicular lines such that each one of them intersects at least five of the hexagons.

1843. *Proposed by José Heber Nieto, Universidad del Zulia, Maracaibo, Venezuela.*

For every positive integer n, let S_n denote the set of permutations of the set $N_n = \{1, 2, \ldots, n\}$. For every $1 \leq j \leq n$, the permutation $\sigma \in S_n$ has a *left to right maximum* (LRM) at position j, if $\sigma(i) < \sigma(j)$ whenever $i < j$. Note that all $\sigma \in S_n$ have a LRM at position 1. Let M be a subset of N_n. Prove that the number of permutations in S_n with LRMs at exactly the positions in M is equal to

$$\prod_{k \in N_n \setminus M} (k - 1),$$

where an empty product is equal to 1.

Math. Mag. **83** (2010) 149–153. doi:10.4169/002557010X482925.

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. Submitted problems should not be under consideration for publication elsewhere.

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

Solutions and new proposals should be mailed to Bernardo M. Ábrego, Problems Editor, Department of Mathematics, California State University, Northridge, 18111 Nordhoff St, Northridge, CA 91330-8313, or mailed electronically (ideally as a LaTeX or pdf file) to mathmagproblems@csun.edu. All communications, written or electronic, should include **on each page** the reader's name, full address, and an e-mail address and/or FAX number.

1844. *Proposed by Marian Tetiva, National College "Gheorghe Roşca Codreanu," Bîrlad, Romania.*

Let ABC be a triangle with $a = BC$, $b = AC$, and $c = AB$. Prove that

$$\frac{a^2 + b^2 + c^2}{2 \cdot \text{Area}(ABC)} \geq \sec\frac{A}{2} + \sec\frac{B}{2} + \sec\frac{C}{2}.$$

1845. *Proposed by Albert F. S. Wong, Temasek Polytechnic, Singapore.*

Evaluate

$$\int_0^1 \left\{\frac{1}{x}\right\}^2 dx,$$

where $\{\alpha\} = \alpha - \lfloor\alpha\rfloor$ denotes the fractional part of α.

Quickies

Answers to the Quickies are on page 153.

Q999. *Proposed by Hongbiao Zeng, Fort Hays State University, Hays, KS.*

Let ABC be a triangle and O its circumcenter. Suppose that O and A are on the same side of $\overline{BC}$. Prove that if $\text{Area}(ABC) = 2 \cdot \text{Area}(OBC)$, then $\triangle ABC$ is a right triangle.

Q1000. *Proposed by Michael W. Botsko, Saint Vincent College, Latrobe, PA.*

Let D be an open and connected subset of $\mathbb{C}$ and let f and g be continuous complex-valued functions defined on D such that $f(z)g(z) = 0$ and $|f(z)| + |g(z)| \neq 0$ for all z in D. Show that either f or g is identically zero on D.

Solutions

A trigonometric identity for the square root triangle — April 2009

1816. *Proposed by Mehmet Sahin, Ankara University of Science, Ankara, Turkey.*

Let ABC be a triangle with $a = BC$, $b = CA$, and $c = AB$. Let $A'B'C'$ be another triangle with $B'C' = \sqrt{a}$, $C'A' = \sqrt{b}$, and $A'B' = \sqrt{c}$. Prove that

$$\sin\left(\frac{1}{2}A\right)\sin\left(\frac{1}{2}B\right)\sin\left(\frac{1}{2}C\right) = \cos A' \cos B' \cos C'.$$

Solution by Philip Benjamin, Middlesex County College, Edison, NJ.

We require two formulas, the half-angle formula for the sine function: $\sin(t/2) = \sqrt{(1 - \cos t)/2}$ for $0 \leq t \leq 2\pi$, and the law of cosines: $\cos A = (b^2 + c^2 - a^2)/(2bc)$. In triangle ABC,

$$\sin\left(\frac{1}{2}A\right) = \sqrt{\frac{1 - \cos A}{2}} = \sqrt{\frac{2bc - b^2 - c^2 + a^2}{4bc}} = \sqrt{\frac{(a - b + c)\,(a + b - c)}{4bc}}$$

with similar formulas for the other angles of triangle ABC. Thus

$$\sin\left(\frac{1}{2}A\right)\sin\left(\frac{1}{2}B\right)\sin\left(\frac{1}{2}C\right) = \frac{(a + b - c)(a - b + c)(-a + b + c)}{8abc}.$$

In triangle $A'B'C'$, the law of cosines gives $\cos A' = (b + c - a)/(2\sqrt{bc})$ with similar formulas for the other angles of $A'B'C'$. Thus

$$\cos A' \cos B' \cos C' = \frac{(a + b - c)(a - b + c)(-a + b + c)}{8abc}$$

and the identity is proved.

Also solved by Tamin Alkhonaini; Armstrong Problem Solvers; Herb Bailey; Michel Bataille (France); J. C. Binz (Switzerland); Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Australia); Stan Byrd and Ossama A. Saleh; Robert Calcaterra; Elsie M. Campbell, Dionne T. Bailey, and Charles Diminnie; Minh Can; Michael J. Caulfield; The Constant Math Party; Tim Cross (United Kingdom); Robert D. Crise; Chip Curtis; Ranjan Dahal; Daniele Degiorgi (Switzerland); M. J. Englefield (Australia); John Ferdinands; Dimitri Fleischman; Jason Fornkohl; Rohollah Garmanjani (Portugal); Leon Gerber; David Getling (Germany); Michael Goldenberg and Mark Kaplan; James R. Henderson; Eugene A. Herman; John G. Heuver (Canada); Brian Hogan; Peter Hohler (Switzerland); Joel Iiams; Tom Jager; Bradley Jones; Young Ho Kim (Korea); Victor Y. Kutsenok; Harris Kwong; David P. Lang; Kee-Wai Lau; Kim McInturff; William McNeary; James Meyer; Ruthven Murgatroyd; Shoeleh Mutameni; Daniel Narrias-Villar (Chile); Northwestern University Math Problem solving Group; Peter Nüesch (Switzerland); J. Oelschlager; Jennifer Pajda; Éric Pité (France); Cosmin Pohoata (Romania); Xavier Retnam; Joel Schlosberg; Mark H. Schultz; C. R. Selvaraj and Suguna Selvaraj; Raul A. Simon (Chile); Nicholas C. Singer; Earl A. Smith; Albert Stadler (Switzerland); John Sumner and Aida Kadic-Galeb; James Swenson; Marian Tetiva (Romania); R. S. Tiberio; Dave Trautman; Anibal Velozo (Chile); Michael Vowe (Switzerland); G. Gerard Wojnar; John. B. Zacharias; and the proposer.

Fibonaccis and game scores **April 2009**

1817. *Proposed by Marcos Donnantouni, La Plata, Argentina and José Nieto, Maracaibo, Estado Zulia, Venezuela.*

A TV game show has a format in which contestants are asked questions and give answers. Each contestant starts with a score of 0 points. A contestant's score is then calculated as follows: after giving a correct answer, the score is increased by 1; after a wrong answer the score is divided by 2. If a contestant responds to n questions, how many different scores are possible? (As an example, for $n = 3$ there are seven possible scores : 0, $\frac{1}{4}$, $\frac{1}{2}$, 1, $\frac{3}{2}$, 2, and 3.)

Solution by Benjamin V. C. Collins and James Swenson, University of Wisconsin-Platteville, Platteville, WI.

There are $F_{n+3} - 1$ possible scores after n questions, where $(F_0, F_1, \dots) = (0, 1, \dots)$ is the Fibonacci sequence.

Encode the contestant's responses by a word $a = a_1 a_2 \dots a_n$, where $a_k = R$ if the kth question is answered correctly, and $a_k = W$ otherwise. For any words a and b, we claim that the words $aRRWb$ and $WaWRb$ have the same length and yield the same score for the contestants. To see this, suppose that the word a yields score x. Then $aRRW$ yields $\frac{x+2}{2}$, while $WaWR$ yields $\frac{x}{2} + 1$; this proves the claim.

By induction, if the score x can be achieved, then it can be achieved by a word that does not contain RRW as a subword. Such a word is called *nice*; it is of the form $(\prod_{i=1}^{k} R^{\varepsilon_i} W)R^j$, where each $\varepsilon_i \in \{0, 1\}$ and $j, k \geq 0$. Because the word $(\prod_{i=1}^{k} R^{\varepsilon_i} W)R^j$ yields the score $(\sum_{i=1}^{k} \varepsilon_i \cdot 2^{i-1-k}) + j$, it follows that no two nice words yield the same score. Hence possible scores are in one-to-one correspondence with nice words.

The word $(\prod_{i=1}^{k} R^{\varepsilon_i} W)R^j$ contains k letters W and $j + \sum_{i=1}^{k} \varepsilon_k$ letters R, so there are exactly $n - k - j$ of the ε_i equal to 1; this means that there are $\binom{k}{n-k-j}$ nice words of length n containing k letters W and ending with R^j. Altogether there are

$$\sum_{k=0}^{n} \sum_{j=0}^{n-k} \binom{k}{n-k-j} = \sum_{k=0}^{n} \sum_{j=0}^{k} \binom{k-j}{j} = \sum_{k=0}^{n} F_{k+1} = F_{n+3} - 1$$

nice words of length n.

Also solved by Michael Abram, Michel Bataille (France), Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Australia), Robert Calcaterra, John Christopher, G.R.A.20 Problem Solving Group (Italy), Joel Iiams, Tom Jager, Omran Kouba (Syria), Lafayette College Problem Group, Jacob Richey, Albert Stadler (Switzerland), John Sumner and Aida Kadic-Galeb, Marian Tetiva (Romania), Dave Trautman, Todd G. Will, and the proposers. There were three incomplete solutions.

Counting two different ways **April 2009**

1818. *Proposed by Cosmin Pohoata, Tudor Vianu National College of Informatics, Bucharest, Romania.*

Let $n, k, i, i_1, i_2, \ldots, i_k$ be positive integers with $n \geq i = i_1 + i_2 + \cdots + i_k$. Prove that 2^{n-i} is a factor of

$$\sum_{j=0}^{n} \left(\binom{n}{j} \prod_{r=1}^{k} \binom{j}{i_r} \right).$$

Solution by Timothy Woodcock, Stonehill College, Easton, MA.

The expression counts the set of all $(k+1)$-tuples $\mathbf{t} = (S, A_1, \ldots, A_k)$ where S is a j-element subset of $N = \{1, \ldots, n\}$, and for $1 \leq r \leq k$, A_r is an i_r-element subset of S.

We may also construct a general $\mathbf{t}$ by first forming $\mathbf{u} = (A_1, \ldots, A_k)$, where each A_r is an i_r-element subset of N, and then require $S \supseteq \cup_{r=1}^{k} A_r$. Letting $|\mathbf{u}|$ denote the cardinality of the union, there are $2^{n-|\mathbf{u}|}$ subsets of $N \setminus \cup_{r=1}^{k} A_r$ that can be joined to $\cup_{r=1}^{k} A_r$ to form a suitable set S. Thus the total number of possibilities for $\mathbf{t}$ is equal to $\sum_{\mathbf{u}} 2^{n-|\mathbf{u}|}$. But for any $\mathbf{u}$, $|\mathbf{u}| \leq i_1 + \cdots + i_k = i$, so 2^{n-i} divides $2^{n-|\mathbf{u}|}$.

Also solved by Michel Bataille (France), Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Australia), Robert Calcaterra, G.R.A.20 Problem Solving Group (Italy), Santhosh Karnik, Omran Kouba (Syria), Northwestern University Math Problem Solving Group, Albert Stadler (Switzerland), John Sumner and Aida Kadic-Galeb, Marian Tetiva (Romania), and the proposer. There was a solution with no name.

Irreducibility of x in $\mathbb{Z}_m[x]$ **April 2009**

1819. *Proposed by Jody M. Lockhart and William P. Wardlaw, U.S. Naval Academy, Annapolis, MD.*

An element a of a ring R is reducible in R if there are elements b and c in R, neither of which are units in R, such that $a = bc$. If a is not reducible then we say a is irreducible. For each integer $m > 1$, let $\mathbb{Z}_m[x]$ denote the ring of polynomials over the ring $\mathbb{Z}_m$ of integers modulo m. For which integers $m > 1$ is the polynomial x irreducible in $\mathbb{Z}_m[x]$?

Solution by Bruce S. Burdick, Roger Williams University, Bristol, RI.

The answer is the numbers m that are prime powers. Suppose $m = p^k$ with p prime and $k \geq 1$, and suppose that there are polynomials $f(x), g(x) \in \mathbb{Z}_m[x]$ such that $f(x)g(x) = x$. Let the coefficients of the polynomials be given by

$$f(x) = f_i x^i + \cdots + f_1 x + f_0 \quad \text{and} \quad g(x) = g_j x^j + \cdots + g_1 x + g_0.$$

Then $f_1 g_0 + f_0 g_1 = 1$ and $f_0 g_0 = 0$. The second equation implies that either f_0 and g_0 are both powers of p or one of them is 0. But if they are both powers of p, then the first equation becomes impossible modulo m. Thus one of f_0 or g_0 is 0. Suppose that $f_0 = 0$. Then $f(x)/x$ is a polynomial in $\mathbb{Z}_m[x]$ and it follows from the equation $(f(x)/x)g(x) = 1$ that $g(x)$ is a unit. This proves that x is irreducible.

Conversely, suppose $m = ab$ with a and b relatively prime and both greater than 1. Then in $\mathbb{Z}_m[x]$, $(ax + b)(bx + a) = (a^2 + b^2)x$. The numbers $a^2 + b^2$ and ab are relatively prime whenever a and b are, so $a^2 + b^2$ has an inverse c in $\mathbb{Z}_m$. It follows that $(cax + cb)(bx + a) = x$. Neither $cax + cb$ nor $bx + a$ are units in $\mathbb{Z}_m[x]$ because otherwise a or b would be a unit in $\mathbb{Z}_m$, contradicting the fact that they are both zero-divisors in $\mathbb{Z}_m$. Thus x is reducible.

Also solved by Michel Bataille (France), Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Australia), Robert Calcaterra, Thomas Craven, Jim Delany, Robert L. Doucette, John Ferdinands, Florida Southern College Modern Algebra Class, Joel Haack, Joel Iiams, Tom Jager, David P. Lang, Rick Mabry, Vadim Ponomarenko, Joel Schlosberg, Nicholas C. Singer, Alin A. Stancu and Andrew S. Wilson, John Sumner and Aida Kadic-Galeb, James Swenson, Bob Tomper, Naveed Zaman, and the proposers. There were five incorrect submissions.

Zero trace implies zero product **April 2009**

1820. *Proposed by Christopher J. Hillar, Texas A&M University, College Station, TX.*

A real positive semidefinite matrix is a symmetric matrix with all eigenvalues non-negative. Prove that if P and Q are real positive semidefinite $n \times n$ matrices with $\mathrm{tr}(PQ) = 0$, then $PQ = 0$.

Solution by Cosmin Pohoata, Bucharest, Romania.

By the Cholesky Decomposition Theorem, the positive semidefinite matrices P and Q can be written as $P = XX^t$ and $Q = YY^t$, where X and Y are real-valued matrices. Thus,

$$0 = tr(PQ) = tr(XX^tYY^t) = tr(Y^tXX^tY) = tr[(Y^tX)(Y^tX)^t] = tr(AA^t),$$

where $A = Y^tX$. Because $A = (a_{ij})$ has real entries and $tr(AA^t) = \sum_{i,j} a_{ij}^2 = 0$, it follows that $A = 0$. Hence $A^t = \left(Y^tX\right)^t = X^tY = 0$ and $PQ = XX^tYY^t = 0$.

Also solved by Oskar M. Baksalary (Poland) and Götz Trenkler (Germany), Michel Bataille (France), Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Australia), Paul Budney, Robert Calcaterra, Chip Curtis, Michael J. Englefield (Australia), Michael Goldenberg and Mark Kaplan, Eugene A. Herman, Tom Jager, Omran Kouba (Syria), Éric Pité (France), and the proposer. There was one incorrect submission.

Answers

Solutions to the Quickies from page 150.

A999. Let M be the midpoint of $\overline{AB}$, ℓ the line parallel to $\overline{BC}$ through M, and ℓ' the perpendicular bisector of $\overline{AB}$. Because O and A are on the same side of $\overline{BC}$ and $\mathrm{Area}(ABC) = 2 \cdot \mathrm{Area}(OBC)$, it follows that O is in ℓ. If $O = M$, then ABC is a triangle with right angle at C. Otherwise, since both O and M are in ℓ and ℓ', it follows that $\ell = \ell'$ and so $\overline{AB}$ is perpendicular to $\overline{BC}$, that is, ABC is a triangle with right angle at B.

A1000. Let $S_f = \{z \in D : f(z) \neq 0\}$ and $S_g = \{z \in D : g(z) \neq 0\}$. Because f is continuous, it follows that $D \setminus S_f = f^{-1}(\{0\})$ is a closed set, so S_f is an open set. Similarly S_g is an open set. The conditions on D imply that for all z in D, $f(z) = 0$ if and only if $g(z) \neq 0$. Therefore $D = S_f \cup S_g$ and $S_f \cap S_g = \emptyset$. Because D is connected, it follows that $S_f = \emptyset$ or $S_g = \emptyset$, that is, either f or g is identically zero on D.

REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

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

Carter, Nathan, *Visual Group Theory*, MAA, 2009; xiii + 297 pp, $71.95 (member price: $57.50). ISBN 978-0-88385-757-1. Accompanying free graphing software: GroupExplorer 2.2, for Windows XP, Mac OS X, and Linux. `http://groupexplorer.sourceforge.net/`.

This is a beautiful book, with magnificent color illustrations. It "teaches you to know groups," through admirably many examples of Cayley diagrams, multiplication tables, and other diagrams, focussing mainly on finite groups. Its aim is not to present the theorems and proofs of group theory but to equip intuition with a stock of examples and teach "how to make conjectures about groups and prove or refute them." The exercises bear out the purpose; most ask the reader to investigate a conjecture, few ask for proofs. A group is first defined as a collection of actions and only later as a set with a binary operation. Topics include subgroups, products, quotients, homomorphisms, Sylow theory, and Galois theory; rings are not mentioned. The book opens with a discussion of the moves in Rubik's Cube but never returns to it; commutators, the key to solving the Cube, are defined only in an exercise late in the book, without mention of their use with the Cube. However, the author is right to claim that this is an ideal book for a student beginning to learn about groups. It could also serve for an "abstract algebra light" course, perhaps even for students who do not intend to become mathematics majors but who sincerely want to learn how mathematicians think, are curious how mathematics is not necessarily about numbers, and can put up with a little symbolism and reasoning as they explore.

Erlich, Yaniv, Kenneth Chang, Assaf Gordon, Roy Ronen, Oron Navon, Michelle Rooks, and Gregory J. Hannon, DNA Sudoku: Harnessing high-throughput sequencing for multiplexed specimen analysis, *Genome Research* 19 (2009) 1243–1253, `http://genome.cshlp.org/content/early/2009/05/15/gr.092957.109.full.pdf+html`; mathematical supplement at `http://genome.cshlp.org/content/early/2009/05/15/gr.092957.109/suppl/DC1`. CSHL scientists harness logic of "Sudoku" math puzzle to vastly enhance genome-sequencing capability (press release), `http://www.cshl.edu/public/releases/09_sudoku.html`.

DNA-sequencing machines can analyze DNA from many different specimens simultaneously; the trick is to match each sequence back to its specimen. Rather than assign an identifying code to each specimen, the authors group specimens into overlapping pools and tag the pools. The authors choose the number of pools, and the numbers of pools in which each specimen occurs, so as to minimize costs while maximizing probability of correct matching. They have applied the method to 40,000 bacterial clones using 384 pools. The major mathematical tool that allows recovery of the matching information is the Chinese remainder theorem from number theory. "Many elements of this approach were reminiscent of seeking the solution to a Sudoku puzzle, which led us to dub this strategy 'DNA Sudoku.'" There may be no direct connection to methods for solving Sudoku puzzles, but the authors have given a catchy name to a useful method employing mathematics from number theory.

Math. Mag. **83** (2010) 154–155. doi:10.4169/002557010X492735.

Blackburn, Simon R., The geometry of perfect parking, http://personal.rhul.ac.uk/uhah/058/perfect_parking.pdf . Devlin, Keith, The formula for perfect parallel parking, http://www.npr.org/templates/story/story.php?storyId=122880263 . Devlin, Keith, Is math a socialist plot?, http://www.maa.org/devlin/devlin_02_10.html.

How long does a parking space have to be for you to park? Author Blackburn uses the Pythagorean theorem to answer; and in a recent appearance on National Public Radio (NPR), Devlin talked about Blackburn's paper. That might have been the end of it, except for responses at the NPR Website. As Devlin notes in his teaser-titled column, a science story must cite an application—despite no such demand for "sports, music, movies, entertainment and the arts, none of which are 'good for anything' in the sense that science stories are supposed to live up to." Indeed, his original radio piece cited automated parking (already a reality on some car models); but that mention was cut from the broadcast. Absent mention of any application, respondents wrote "I don't need a math formula to tell me how to park," and "This kind of research is a waste of time." Devlin's conclusion: They thought they were supposed to put numbers into the formula "and work out the answer," a conviction bred from the phony examples that pass as real-life applications in textbooks. Devlin makes the crucial point that "no one uses mathematical formulas in their day-to-day life. . . . [t]hese days, it's not people who 'do the math,' it's the various devices we buy, use, and carry around with us." He pleads, "don't use unrealistic, fake scenarios and tell the students they are seeing 'How math is really used.'" He claims that the discussants (and by extension, most adults) "are rooted [in] their belief that math is something you do at school to solve irrelevant problems but is of no use in the real world. . . [they] have absolutely no idea how mathematics is used in today's world. . . . *They don't even know what it is or what it is used for.*" Moreover, "*they have formed this belief after having had at least ten years of almost daily mathematics instruction*. . . [Devlin's emphasis]." That is a damning indictment of mathematics education! From your experience, is it true?

Bressoud, David M., The rocky transition from high-school calculus, *Chronicle of Higher Education* (22 January 2010) A80, http://chronicle.texterity.com/chronicle/20100122a/?pg=80.

Bressoud, President of the MAA, has written several times in *Focus* about trends in calculus and what we should teach students who had some calculus in high school. Here, based on statistics, he describes Advanced Placement (AP) calculus as "not a steppingstone but a stumbling block." Of high school students entering college in 2008, 11% had taken an AP calculus exam and almost 20% had had a calculus course in high school. That sounds fine, except that the increasing numbers of students taking calculus in high school over the past 15 years has been accompanied by a *de*crease in students taking Calculus II in college (except at research universities, with their engineering schools). Why so? Pedagogy? Pushing underprepared students prematurely into calculus? Misalignment of the AP syllabus with college calculus? Focus in high school courses on imitating algorithms rather than on understanding? We don't know, and Bressoud urges us to find out. He also urges a redesign of college Calculus I to be a general overview, including emphasizing why calculus is important and what it is good for. [My perspective: 1. Students take AP calculus for the same reasons as other AP courses (status, competitive edge to get ahead into college, maybe save a semester of college tuition); so expansion of high school offering of calculus has roped in the marginally mathematically motivated. 2. College science and business departments minimize the mathematics required of majors, because some potential majors shy away from mathematics or can't get high grades in mathematics courses (in this era of grade inflation, we grade low). "Early transcendentals" Calculus I—derivatives and integrals of trigonometric, exponential, and logarithmic functions in Calculus I instead of Calculus II—lets departments absolve their majors of taking Calculus II. 3. Departments that require calculus (or other mathematics) for their majors don't make it a prerequisite for any of their courses, thus illustrating that they consider the mathematics slightly desirable but basically inessential.]

38th United States of America Mathematical Olympiad

CECIL ROUSSEAU
University of Memphis
Memphis, TN 38152-3240
rousseac@msci.memphis.edu

STEVEN R. DUNBAR
MAA American Mathematics Competitions
University of Nebraska-Lincoln
Lincoln, NE 68588-0658
sdunbar@maa.org

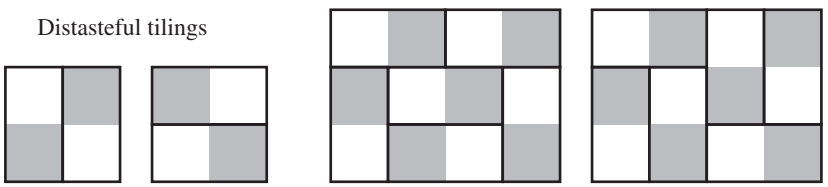
Problems

1. Given circles ω_1 and ω_2 intersecting at points X and Y, let ℓ_1 be a line through the center of ω_1 intersecting ω_2 at points P and Q and let ℓ_2 be a line through the center of ω_2 intersecting ω_1 at points R and S. Prove that if P, Q, R and S lie on a circle then the center of this circle lies on line XY.
2. Let n be a positive integer. Determine the size of the largest subset of

$$\{-n, -n+1, \ldots, n-1, n\}$$

which does not contain three elements a, b, c (not necessarily distinct) satisfying $a + b + c = 0$.

3. We define a *chessboard polygon* to be a polygon whose edges are situated along lines of the form $x = a$ or $y = b$, where a and b are integers. These lines divide the interior into unit squares, which are shaded alternately grey and white so that adjacent squares have different colors. To tile a chessboard polygon by dominoes is to exactly cover the polygon by non-overlapping 1×2 rectangles. Finally, a *tasteful tiling* is one which avoids the two configurations of dominoes shown on the left below. Two tilings of a 3×4 rectangle are shown; the first one is tasteful, while the second is not, due to the vertical dominoes in the upper right corner.



(a) Prove that if a chessboard polygon can be tiled by dominoes, then it can be done so tastefully.

(b) Prove that such a tasteful tiling is unique.

Math. Mag. **83** (2010) 156–158. doi:10.4169/002557010X492744.

4. For $n \geq 2$ let $a_1, a_2, \ldots, a_n$ be positive real numbers such that

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

Prove that $\max(a_1, a_2, \ldots, a_n) \leq 4\min(a_1, a_2, \ldots, a_n)$.

5. Trapezoid $ABCD$, with $\overline{AB} \parallel \overline{CD}$, is inscribed in circle ω and point G lies inside triangle BCD. Rays AG and BG meet ω again at points P and Q, respectively. Let the line through G parallel to $\overline{AB}$ intersect $\overline{BD}$ and $\overline{BC}$ at points R and S, respectively. Prove that quadrilateral $PQRS$ is cyclic if and only if $\overline{BG}$ bisects $\angle CBD$.

6. Let $s_1, s_2, s_3, \ldots$ be an infinite, nonconstant sequence of rational numbers, meaning it is not the case that $s_1 = s_2 = s_3 = \ldots$. Suppose that $t_1, t_2, t_3, \ldots$ is also an infinite, nonconstant sequence of rational numbers with the property that $(s_i - s_j)(t_i - t_j)$ is an integer for all i and j. Prove that there exists a rational number r such that $(s_i - s_j)r$ and $(t_i - t_j)/r$ are integers for all i and j.

Solutions Following are the essential ideas for each problem. For interested readers, detailed solutions with figures and multiple approaches developed by the USAMO Committee are at the website of the MAA American Mathematics Competitions: `http://www.unl.edu/amc/e-exams/e8-usamo/archiveusamo.shtml`.

1. Let ω denote the circumcircle of P, Q, R, S and let O denote the center of ω. The XY is the radical axis of circles ω_1 and ω_2. Then show that O has equal power to the two circles using that $\ell_1 \perp OO_2$.

   This problem was suggested by Ian Le. The solution was contributed by Zuming Feng.

2. The maximum size is n if n is even, and $n + 1$ if n is odd, achieved by the subset

$$\left\{-n, \ldots, -\left\lfloor \frac{n}{2} \right\rfloor - 1, \left\lfloor \frac{n}{2} \right\rfloor + 1, \ldots, n\right\}.$$

   This problem was suggested by Kiran Kedlaya with Tewodros Amdeberhan.

3. Prove the first part by induction on the number n of dominoes in the tiling.

   Suppose now that there are two tasteful tilings of a given chessboard polygon. By overlaying these two tilings obtain chains of overlapping dominoes, since every square is part of one domino from each tiling. For example, a chain of length one indicates a domino common to both tilings. A chain of length two cannot occur, since these arise when a 2×2 block is covered by horizontal dominoes in one tiling and vertical dominoes in the other, and one of these configurations will be distasteful. Since the tilings are distinct a chain of length three or more must occur; let R be the region consisting of such a chain along with its interior, if any. Argue that the chain must include a horizontal domino along its lowermost row having a white square on the left. Now focus on the tiling that includes this WB domino.

   The two squares above the WB domino must be part of region R. Deduce the existence of a horizontal WB domino on the next row up. Repeat this argument until reaching a horizontal WB domino in region R for which the two squares immediately above it are not both in region R. Show that this is impossible, so two tasteful tilings are not possible.

   This problem was suggested by Sam Vandervelde.

4. Let $m = \min(a_1, a_2, \ldots, a_n)$ and $M = \max(a_1, a_2, \ldots, a_n)$. Without loss of generality, $a_1 = m$ and $a_n = M$. Then $n = 2$ case follows immediately. For $n \geq 3$ use the Cauchy-Schwarz Inequality and reduce to the $n = 2$ case. This problem was suggested by Titu Andreescu.

5. First prove the "if" part. Let rays CG and DG meet ω again at E and F, respectively. Let R_1 denote the intersection of segments BD and QE, and let S_1 denote the intersection of segments BC and QF. Applying Pascal's theorem to cyclic hexagon $BDFQEC$ to show R_1, G, S_1 are collinear. Deduce that $EBGR_1$ is cyclic. Use that $EBGR_1$ and $EBCD$ are cyclic to show $PQRS$ is cyclic.

   Prove the "only if" part by letting γ denote the circumcircle of $PQRS$. Then approach indirectly by assuming that ray BG does not bisect $\angle CBD$ and show a contradiction.

   This problem was suggested by Zuming Feng.

6. For p a prime, define the *p-adic norm* $\|\cdot\|_p$ on rational numbers as follows: for $r \neq 0$, $\|r\|_p$ is the unique integer n for which we can write $r = p^n a/b$ with a, b integers not divisible by p. (By convention, $\|0\|_p = +\infty$.) Repeatedly use the easy to prove fact that for any rational numbers r_1, r_2, we have $\|r_1 \pm r_2\|_p \geq \min(\|r_1\|_p, \|r_2\|_p)$, with equality whenever $\|r_1\|_p \neq \|r_2\|_p$. The condition of the problem implies that

$$\|s_i - s_j\|_p \geq -\|t_i - t_j\|_p$$

   for all i, j and all prime p. Then show that in fact

$$\|s_i - s_j\|_p \geq -\|t_k - t_l\|_p$$

   for all i, j, k, l and all prime p. Now for each prime p, define the integer $f(p) = \min_{i,j} \|s_i - s_j\|_p$. The function $f(p)$ is well-defined and $f(p) = 0$ for all but finitely many primes. Finally, show that $r = \prod_p p^{-f(p)}$, where the product is over all primes.

   This problem was suggested by Gabriel Carroll. The solution was suggested by Lenhard Ng.

2009 Olympiad Results The top twelve students on the 2009 USAMO were (in alphabetical order):

John Berman	12	John T. Hoggard School	Wilmington	NC
Sergei Bernstein	12	Belmont High School	Belmont	MA
Wenyu Cao	10	Phillips Academy	Andover	MA
Robin Chen	12	Pinetree Secondary School	Coquitlam	BC
Vlad Firiou	10	Westford Academy	Westford	MA
Eric Larson	12	South Eugene High School	Eugene	OR
Delong Meng	12	Baton Rouge Magnet High School	Baton Rouge	LA
Qinxuan Pan	12	Thomas S. Wooten High School	Rockville	MD
Panupong Pasupat	12	Deerfield Academy	Deerfield	MA
Toan Phan	11	Taft School	Watertown	CT
David Rush	12	Phillips Exeter Academy	Exeter	NH
David Yang	8	Homeschool	Walnut	CA

Wenyu Cao was the winner of the Samuel Greitzer-Murray Klamkin Award, given to the top scorer on the USAMO. Wenyu was also awarded a college scholarship of $20,000 by the Akamai Foundation. Toan Phan was awarded a scholarship of $15,000 by the Akamai Foundation for second place. Qinxuan Pan won third place and was awarded a scholarship of $10,000 by the Akamai Foundation. All 12 students received a $1,000 Robert P. Balles Award in the form of a savings bond.

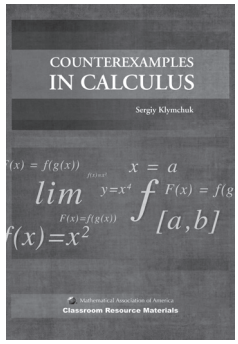
The citation for a solution of outstanding elegance was presented to Evan O'Dorney for his solution on Problem 6. O'Dorney's solution converts the hypotheses of the problem into statements about products of polynomials, and then the conclusion follows from showing that the intersection of a chain of sets is non-empty by using routine facts about coefficients of irreducible polynomials.

New title by the MAA

Counterexamples in Calculus

Sergiy Klymchuk

As a robust repertoire of examples is essential for students to learn the practice of mathematics, so a mental library of counterexamples is critical for students to grasp the logic of mathematics. Counterexamples are tools that reveal incorrect beliefs. Without such tools, learners' natural misconceptions gradually harden into convictions that seriously impede further learning. This slim volume brings the power of counterexamples to bear on one of the largest and most important courses in the mathematics curriculum.
—Professor Lynn Arthur Steen, St. Olaf College, Minnesota, USA, Co-author of *Counterexamples in Topology*



Counterexamples in Calculus serves as a supplementary resource to enhance the learning experience in single variable calculus courses. This book features carefully constructed incorrect mathematical statements that require students to create counterexamples to disprove them. Methods of producing these incorrect statements vary. At times the converse of a well-known theorem is presented. In other instances crucial conditions are omitted or altered or incorrect definitions are employed. Incorrect statements are grouped topically with sections devoted to: Functions, Limits, Continuity, Differential Calculus and Integral Calculus.

This book aims to fill a gap in the literature and provide a resource for using counterexamples as a pedagogical tool in the study of introductory calculus. In that light it may well be useful for

- high school teachers and university faculty as a teaching resource
- high school and college students as a learning resource
- a professional development resource for calculus instructors

Catalog Code: CXC
101pp., Paperbound, 2010
ISBN: 978-0-88385-756-6
List: $45.95
MAA Member: $35.95

Order your copy today!
1.800.331.1622 ● www.maa.org

MAA