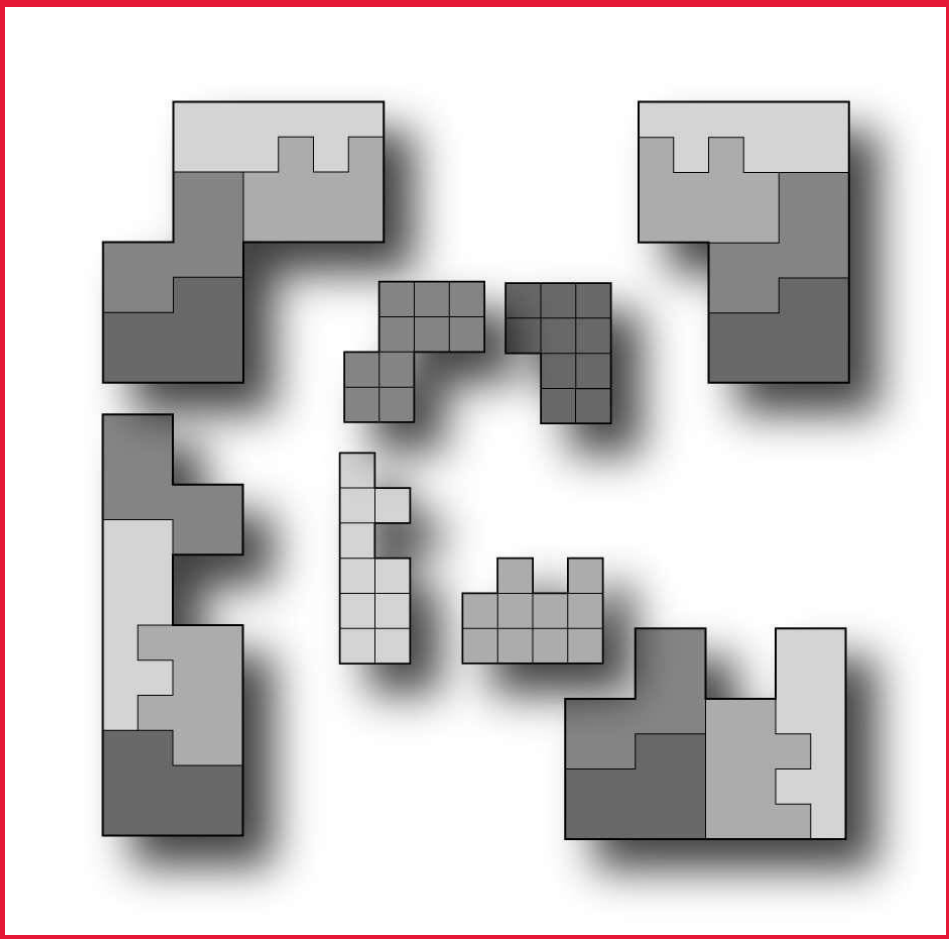
Vol. 87, No. 2, April 2014

MAA

MATHEMATICS MAGAZINE



Four figures tile copies of themselves (p. 100)

- Avoiding arithmetic progressions
- Geometry of cubic polynomials in the complex plane
- Sphere packing, magic tricks, Pascal's triangle

An Official Publication of the **Mathematical Association of America**

EDITORIAL POLICY

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

A more detailed statement of author guidelines appears in this *Magazine*, Vol. 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 Lee Sallows

MATHEMATICS MAGAZINE (ISSN 0025-570X) is published by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, D.C. 20036 and Lancaster, 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: (877) 622-2373
E-mail: tmarmor@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. 87, No. 2, April 2014

MAA

MATHEMATICS MAGAZINE

LETTER FROM THE EDITOR

Everybody likes Ramsey Theory. The problems are easy to understand and the solutions hard to find, but often concrete and especially satisfying.

Certainly these aspects are on display in our first article, in which authors Steve Butler, Ron Graham, and Linyuan Lu report progress on a famous problem. If the integers $\{1, \ldots, n\}$ are colored with r colors, then what fraction of the k-term arithmetic progressions must be monochromatic? Let's take $r = 2$, $k = 3$, and n large. Then a random 2-coloring makes $1/4$ of the 3-term progressions monochromatic. But by using a smarter 2-coloring, the authors can reduce that fraction to $117/548$. Now, can you do better?

A lot of packing and partitioning is happening in this issue. From Owen Byer and Deirdre Smeltzer we have a demonstration of sphere packing. Start with one sphere inside another—not necessarily with the same center—and then pack spheres between them, each tangent to both initial spheres, and each tangent to the same number of other spheres. When can this project be completed? They answer this question elegantly using inversions.

Our cover, by Lee Sallows, shows four shapes, each of which can be constructed from smaller copies of all four shapes. Sallows calls this a *self-tiling tile set,* or more simply, a *setiset,* and such structures were the subject of his article in our December, 2012 issue. He now returns to the subject with new methods and a vastly expanded collection of examples.

Mitsuo Kobayashi's paper presents visual partition-proofs of some series identities, and shows how they can be found in a systematic way.

The article by Frayer, Kwon, Schafhauser, and Swenson takes us in a different direction: They look at cubic polynomials in the complex plane, and consider where the roots and critical points can lie. They are building on two themes that we have seen recently in the Magazine: Polynomial root dragging (February 2008, June 2010) and Marden's Theorem (April and December, 2013). There is more structure than you might guess.

Todd Mateer uses a magic trick to introduce a very practical subject: Reed-Solomon Codes. All you need to perform the trick is provided. These are linear codes, and if you look into their details, you will see a kind of sphere packing here as well. Also in the Notes Section, Patrick Nystedt has a new route to the cosine addition formula, and Robert Schneider has some pretty identities using the golden ratio.

The paper by Tyler Ball, Tom Edgar, and Daniel Juda will be especially satisfying to a few people in particular. Was your high-school science fair project about Pascal's Triangle mod n? Mine was. Like me, you may have produced images like the ones on page 137—all based on prime moduli. But you couldn't make the pattern extend to, for example, $n = 6$. Now you can! Your reward is Figure 4 in this paper. The authors show how this very satisfying figure arises in two different ways—once as a representation of a partial order on the integers, and again from a generalization of binomial coefficients.

Walter Stromquist, Editor

ARTICLES

Unrolling Residues to Avoid Progressions

STEVE BUTLER
Iowa State University
Ames, IA 50011
butler@iastate.edu

RON GRAHAM
University of California, San Diego
La Jolla, CA 92093
graham@ucsd.edu

LINYUAN LU
University of South Carolina
Columbia, SC 29208
lu@math.sc.edu

Mathematics is often described as the study of patterns, some of the most beautiful of which are periodic. In the integers, these periodic patterns take the form of arithmetic progressions—that is, sets of integers that are equally spaced, or have the same common difference. For example, 5, 11, 17, 23, 29 is an arithmetic progression with 5 terms (a "5-AP") where consecutive terms have a difference of 6.

Suppose we divide the numbers $1, 2, \ldots, 28$ into two sets, distinguishing them by "coloring" the numbers red (r) and blue (B). We can then ask how many times we have three equally spaced terms all from the same set. In other words, we can ask how many 3-APs are "monochromatic" in either red or blue.

If we want to find a coloring to *maximize* the number of such occurrences, then the obvious thing to do is to color $1, 2, \ldots, 28$ either all red or all blue. Then every 3-AP is monochromatic. In this case, there are 182 of these triples. (That number comes from Proposition 1, below.)

We can also easily find the *average* number of such occurrences—averaged over all possible colorings. This is the same as asking how many monochromatic 3-APs we would expect if we colored each number by flipping a coin with one side marked r and the other B. If it is a fair coin, then given any three equally spaced numbers, the probability that the coin would always land on the same side is $\frac{2}{8}$. So on average, $\frac{2}{8} \cdot 182 = 45.5$ of the 3-APs are monochromatic.

What if we want to find a coloring to *minimize* the number of monochromatic 3-APs? By the random argument, we can at least find one coloring with at most 45 of these, but we would hope to do better. For the case of $1, 2, \ldots, 28$, one way to color to minimize the number is the following:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
r	r	r	B	B	r	r	r	B	B	B	B	B	B	r	r	r	r	r	r	B	B	B	r	r	B	B	B

Math. Mag. **87** (2014) 83–94. doi:10.4169/math.mag.87.2.83.

A careful count shows 28 monochromatic triples, 14 red and 14 blue. This pattern was found by an exhaustive computer search, and it is unique except for reversal and color-switching.

We want to expand this small problem in three ways. First, we want to look at longer sets of equally spaced terms. So we will look at k-term arithmetic progressions, or k-APs, which consist of the numbers $a, a+d, \ldots, a+(k-1)d$, with $d \geq 1$. (In the case that $d = 0$, this becomes $a, a, \ldots, a$, which can be called a trivial k-AP, but we never count trivial APs.) Second, instead of using only the two colors of red and blue, we want to consider $r \geq 2$ colors. Finally, we are interested in what happens as n is large. (28 is a beautiful number, but most interesting things happen for much larger numbers.)

In particular, we are interested in exploring the following question, where k and r are fixed and n is very large.

QUESTION. *What is the minimum number of monochromatic k-APs that can occur in a coloring of the numbers $1, 2, \ldots, n$ with r colors? Further, what colorings achieve the minimum?*

We are interested in the limit as n approaches infinity, and we will often express the number of monochromatic k-APs as a fraction of all k-APs in $\{1, 2, \ldots, n\}$.

van der Waerden's Theorem

The best case scenario would be if there were a way to color $1, 2, \ldots, n$ with no monochromatic k-APs. However, a classical result of van der Waerden shows that this is impossible as n gets large.

THEOREM 1. (VAN DER WAERDEN [**6, 9**]) *For any positive integers r and k, there is a threshold N so that if $n \geq N$, then any coloring of $1, 2, \ldots, n$ using r colors has a monochromatic arithmetic progression of length k.*

For the case $k = 3$ and $r = 2$, we can color 12345678 as BBrrBBrr and not have any three numbers that form a 3-AP all red or all blue. But for any coloring of $1, 2, \ldots, 9$, there is a monochromatic 3-AP. So in this case, $N = 9$.

As a direct consequence of Theorem 1, for large n we must have at least one monochromatic k-AP when using r colors. Actually, there must be many.

THEOREM 2. (FRANKL, GRAHAM, AND RÖDL [**3**]) *For fixed r and k, there is $c > 0$ so that the number of monochromatic k-APs in any r-coloring of $1, 2, \ldots, n$ is at least $cn^2 + o(n^2)$.*

In this theorem and throughout the paper we will use little-"o" and big-"O" notation to bound the size of the lower-order terms. In general, we say that $f(n) = O(g(n))$ to indicate that for some constants C and n_0, for all $n \geq n_0$, $|f(n)| \leq Cg(n)$; that is, the size of f is bounded by a multiple of g, at least for sufficiently large n. We say that $f(n) = o(g(n))$ if for each $\varepsilon > 0$, there is a sufficiently large n_0 so that if $n \geq n_0$, then $|f(n)| \leq \varepsilon g(n)$; that is, the size of f is eventually smaller than any multiple of g. So the preceding corollary shows that, asymptotically, the number of monochromatic k-APs is at least cn^2. (A more substantial introduction to this notation can be found in many places, including Graham, Knuth, and Patashnik [**5**, Ch. 9].)

Before we give a short sketch of the proof of Theorem 2, we make an observation, which we will use repeatedly throughout the paper.

PROPOSITION 1. *Let $k \geq 3$, and let n be congruent to a mod $k-1$, with $0 \leq a < k-1$. Then the number of k-APs in $1, 2, \ldots, n$ is*

$$\frac{n^2}{2(k-1)} - \frac{n}{2} + \frac{a(k-1-a)}{2(k-1)}.$$

The last term is independent of n and is in the interval $[0, (k-1)/8]$. Using the "O" notation, we can say that the number of k-APs is

$$\frac{n^2}{2(k-1)} + O(n).$$

Proof. Let $n = (k-1)\ell + a$, with $0 \leq a < (k-1)$. We count the number of k-APs. The starting points for a k-AP with sizes of step d are $1, 2, \ldots, n-d(k-1)$. Since $1 \leq d \leq \ell$, we have that the number of k-APs is

$$\begin{aligned}
&\underbrace{n-(k-1)}_{d=1} + \underbrace{n-2(k-1)}_{d=2} + \cdots + \underbrace{n-\ell(k-1)}_{d=\ell} \\
&\quad = \frac{\ell\big[((n-(k-1)) + (n-\ell(k-1))\big]}{2} \\
&\quad = \frac{(n-a)(n-k+1+a)}{2(k-1)} \\
&\quad = \frac{n^2}{2(k-1)} \underbrace{-\frac{n}{2} + \frac{a(k-1-a)}{2(k-1)}}_{=O(n)}.
\end{aligned}$$

For the first equality, we used the old trick of reversing the sum and adding it to itself. There are $\ell = (n-a)/(k-1)$ terms in the sum. The rest involves simple manipulations. ∎

Sketch of proof of Theorem 2. We let N be as in Theorem 1 for the given r and k. Now for large n, we have inside of $1, 2, \ldots, n$ many arithmetic progressions of length N; in particular, we have $\approx n^2/(2N)$ of them. By Theorem 1, inside of each one of these progressions of length N is a monochromatic progression of length k. We have now found many monochromatic k-APs, but need to correct for possible double counting. We note that any monochromatic progression could have been counted at most $\binom{N}{2}$ times. Therefore, we have at least $\approx n^2/N^3$ monochromatic progressions. ∎

We made use of the fact that shifted arithmetic progressions are still arithmetic progressions, one of the reasons why they are so nice to study.

By combining Proposition 1 and Theorem 2, we now conclude that a strictly positive portion of the k-APs in our coloring must be monochromatic as n gets large.

On the other hand, the bound given by this proof of Theorem 2 is a poor bound. We gave a lot away in our counting. Further, the N from Theorem 1 is hard to compute. The best bound for N when $r = 2$ colors with arithmetic progression of length k was given by Tim Gowers [**4**],

$$N \leq 2^{2^{2^{2^{2^{k+9}}}}}.$$

Ron Graham currently offers \$1000 to show that, for this case, $N \leq 2^{k^2}$.

Color randomly

We now set about the task of finding colorings with small proportions of monochromatic k-APs. Equivalently, we are trying to find upper bounds for the constant c in Theorem 2.

A natural first guess is to color "randomly," that is, as in our initial problem to flip a coin and assign a color. By looking over all the colorings, we can find the expected number of arithmetic progressions in a typical coloring; then, some coloring must have at most that many. The advantage of this technique is in its simplicity to produce bounds. Moreover, there are examples of combinatorial problems when a random construction is best. Indeed, for the question we are considering, the random construction was the best-known construction for decades.

PROPOSITION 2. *For n large and r, k fixed, there is a coloring of $1, 2, \ldots, n$ with r colors, which has at most*

$$\frac{1}{2(k-1)r^{k-1}}n^2 + O(n)$$

monochromatic k-APs. In particular, $1/r^{k-1}$ portion of the k-APs are monochromatic.

Proof. We count the total number of monochromatic k-APs in all colorings. To get the average, we divide by r^n, the number of colorings. Some coloring must have at most the average number, and that will give us the bound.

For a moment, focus on a single k-AP. There are r ways to color that k-AP monochromatically, and we can extend each such monochromatic coloring of that k-AP to a coloring of $1, 2, \ldots, n$ by r^{n-k} ways. Therefore, each k-AP will contribute $r \cdot r^{n-k}$ to the total.

On the other hand, by Proposition 1, there are $\frac{1}{2(k-1)}n^2 + O(n)$ k-APs in $1, 2, \ldots, n$ so that the total number of monochromatic k-APs over all possible colorings is

$$\left(\frac{1}{2(k-1)}n^2 + O(n)\right) r^{n-k+1}.$$

Dividing this by r^n gives $\frac{1}{2(k-1)r^{k-1}}n^2 + O(n)$, as required. ∎

Beating random colorings for 3-APs

We can use the coloring found at the beginning for $n = 28$ and "blow it up" to find a coloring for larger values of n. One way to do this is to simply repeat each color m times to get a new coloring of $28m$. For example, when $m = 2$, our initial coloring becomes

rrrrrrBBBBrrrrrrBBBBBBBBBBBBrrrrrrrrrrrrBBBBBBrrrrBBBBBB.

This gives a coloring of $28m$, and we can then simply drop the last few terms to get a coloring for n. (Dropping ≤ 27 terms affects at most $27n$ arithmetic progressions, which is a lower-order term, so it will not significantly affect the portion that are monochromatic when n is large.) A computation shows that this gives $(3/56)n^2 + O(n)$ monochromatic 3-APs, or 21.42% of the 3-APs, compared to the random bound of 25%. (These computations are described by Butler, Costello, and Graham [**2**].)

The best-known coloring for minimizing monochromatic 3-APs uses the same approach, but a larger starting pattern to get 21.35% of the 3-APs monochromatic. This was discovered independently by two different groups at about the same time.

THEOREM 3. (PARRILO-ROBERTSON-SARACINO [**8**]; BUTLER-COSTELLO-GRAHAM [**2**]) *Start with the following coloring of* $1, 2, \ldots, 548$ *and then blow it up by repeating each term with its color* $\ell = \lceil n/548 \rceil$ *times, then deleting any excess terms to get a coloring of* $1, 2, \ldots, n$.

$$\underbrace{r\cdots r}_{28}\underbrace{B\cdots B}_{6}\underbrace{r\cdots r}_{28}\underbrace{B\cdots B}_{37}\underbrace{r\cdots r}_{59}\underbrace{B\cdots B}_{116}\underbrace{r\cdots r}_{116}\underbrace{B\cdots B}_{59}\underbrace{r\cdots r}_{37}\underbrace{B\cdots B}_{28}\underbrace{r\cdots r}_{6}\underbrace{B\cdots B}_{28}$$

Then the number of monochromatic three-term arithmetic progressions in such a coloring is $\frac{117}{2192}n^2 + O(n)$.

For large n, the resulting coloring will look like FIGURE 1.

Figure 1 The best known 2-coloring of $1, 2, \ldots, n$ to minimize monochromatic 3-APs

Butler, Costello, and Graham [**2**] used a computer search to look for similar ways to color $1, 2, \ldots, n$ by subdividing into large blocks to minimize monochromatic 4-APs and 5-APs. For 4-APs, they found a coloring where 10.33% of the 4-APs were monochromatic (random gives 12.5%). Another coloring gave 4.57% of the 5-APs as monochromatic (random gives 6.25%).

Unrolling better solutions

Another way to blow up the coloring from the first example is to place m complete copies of it in succession to get a new coloring of $28m$, and then as before to delete any excess. We will call this process *unrolling* (FIGURE 3 explains the term). So, for example, when $m = 2$, our initial coloring becomes

rrrBBrrrBBBBBBrrrrrrrBBBrrBBBrrrBBrrrBBBBBBrrrrrrrBBBrrBBB.

The surprising thing is that even though this is a great coloring of $1, 2, \ldots, 28$, the unrolling is *not* a great coloring of $1, 2, \ldots, n$. This coloring yields $n^2/4 + O(n)$ monochromatic 3-APs, the same as random! Moreover, no coloring found by unrolling a fixed coloring of $1, 2, \ldots, \ell$ can do better than random for avoiding 3-APs.

For larger values of k, the situation is dramatically different. This was first noted in the paper of Lu and Peng [**7**]. Their original motivation was to find a way to minimize the number of monochromatic 4-APs in colorings of $\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\}$, where we can do arithmetic operations using modular arithmetic. Lu and Peng also showed that the same approach works for coloring $1, 2, \ldots, n$. Their idea was to unroll a special coloring of $\mathbb{Z}_{11}$, shown in FIGURE 2 with a bit that can be arbitrarily colored. This coloring has only trivial monochromatic 4-APs. As a result, when it is unrolled as in FIGURE 3, this would cause any monochromatic 4-AP to have a large gap between consecutive terms—regardless of how the white squares are eventually colored, or whether they are colored at all. Hence, there would be relatively few of these monochromatic 4-APs.

We can be more precise. When the coloring in FIGURE 2 is unrolled, an arithmetic progression is monochromatic if and only if its common difference is a multiple of 11. (Some of them are "monochromatic" in white.) For large n, this means that $1/11$ of the 4-APs are monochromatic, compared to $1/8$ for a random coloring.

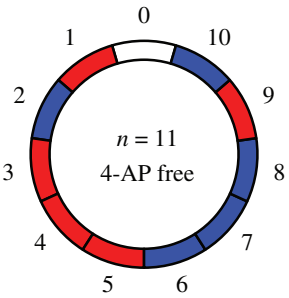


Figure 2 A coloring of $\mathbb{Z}_{11}$ with only trivial monochromatic 4-APs; the white bit can be either color.

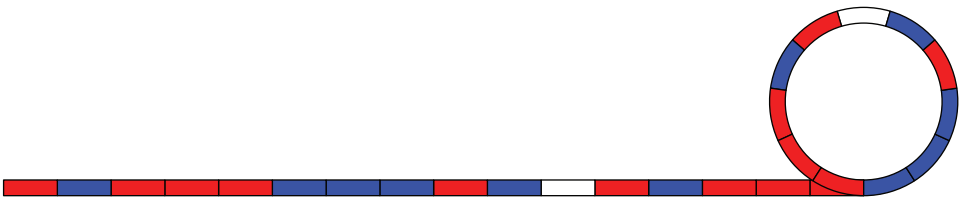


Figure 3 Unrolling the coloring of $\mathbb{Z}_{11}$ to find a coloring of $1, 2, \ldots, n$

There is further improvement in that we still have freedom (and obligation) to color the white squares either red or blue. Our method is recursive. We unroll along the white squares the same pattern we used initially. This still leaves a white square in each 121st position, so we repeat the process for those squares, and so on.

This is much easier to do than it might seem at first glance. The coloring is equivalent to the following: Given any number ℓ, write it in base 11:

$$\ell = \sum_i b_i \cdot 11^i \qquad \text{where } 0 \le b_i \le 10.$$

Let j be the smallest index for which $b_j \neq 0$. Then

$$\text{color } \ell \quad \begin{cases} \text{red} & \text{if } b_j = 1, 3, 4, 5, \text{ or } 9; \\ \text{blue} & \text{if } b_j = 2, 6, 7, 8, \text{ or } 10. \end{cases}$$

It will follow from Theorem 4 below that this coloring has $(1/72)n^2 + O(n)$ monochromatic 4-APs, so that $1/12 \approx 8.33\%$ of the 4-APs are monochromatic. For comparison, when we color randomly, we have $(1/48)n^2 + O(n)$ monochromatic 4-APs, so that $1/8 = 12.5\%$ of the 4-APs are monochromatic.

This coloring is far superior to the block coloring referred to earlier, and also far easier to describe and to implement. Indeed, we can color all of $\mathbb{N}$ and then just take the first n terms with this coloring; in a block coloring such as in FIGURE 1, this would be impossible.

This idea of unrolling works for any pattern. What we need is a pattern where the unrolling produces few monochromatic k-APs. We will focus on colorings of $\mathbb{Z}_m$ that have only trivial monochromatic k-APs. In this case, we get the following count for the number of monochromatic k-APs in an unrolling.

THEOREM 4. *If there is a coloring of $\mathbb{Z}_m$ with r colors, which have only trivial monochromatic k-APs, then for n large there is an r-coloring of $1, 2, \ldots, n$ that has*

$$\frac{1}{2m(k-1)}n^2 + O(n)$$

monochromatic k-APs. This corresponds to a fraction $1/m$ of all k-APs.

If there are r different colorings of $\mathbb{Z}_m$ with r colors, which have only trivial monochromatic k-APs, and these differ only in the coloring of 0, then for n large there is an r-coloring of $1, 2, \ldots, n$ that has

$$\frac{1}{2(m+1)(k-1)}n^2 + O(n)$$

monochromatic k-APs. This corresponds to a fraction $1/(m+1)$ of all k-APs.

Proof. Any monochromatic k-AP can be rolled back into a monochromatic k-AP of $\mathbb{Z}_m$. Since we only have trivial k-APs in our coloring of $\mathbb{Z}_m$, it follows that the difference between any consecutive terms in our arithmetic progression is a multiple of m.

In the first case, when unrolling we color according to the residue class in the coloring of $\mathbb{Z}_m$. In particular, the residue completely determines the color, and so we can simply count our coloring by counting what happens on each of the m residue classes. Each residue consists of a solid coloring of length $(n/m) + O(1)$. Using Proposition 1, each of these m residue classes will contribute

$$\frac{n^2}{2m^2(k-1)} + O(n)$$

monochromatic k-APs. This establishes the first case.

In the second case, we will do recursive unrolling (similar to the 4-AP case). Namely, for $i = 1, 2, \ldots, r$ let C_i be the nonzero elements of $\mathbb{Z}_m$ colored with color i. For each ℓ, we find the first nonzero digit in the base m expansion of ℓ and color ℓ with the ith color, if the digit is in C_i. Initially, when we subdivide by residue classes, then as before $m-1$ of the residue classes will be colored monochromatically. The residue class for 0 will look like the unrolling for the coloring of $(n/m) + O(1)$.

If we let $F(n)$ be the number of monochromatic k term arithmetic progressions, then the above analysis shows that

$$F(n) = F\left(\frac{n}{m}\right) + (m-1)\frac{1}{2(k-1)}\left(\frac{n}{m}\right)^2 + O(n).$$

Iteratively applying this, we have

$$\begin{aligned} F(n) &= \frac{(m-1)n^2}{2(k-1)} \sum_{i \geq 1} \left(\frac{1}{m^2}\right)^i + O(n) \\ &= \frac{(m-1)n^2}{2(k-1)} \cdot \frac{1}{m^2-1} + O(n) \\ &= \frac{1}{2(m+1)(k-1)}n^2 + O(n). \end{aligned}$$

■

Using residues to color

To be able to apply Theorem 4, we must find a coloring of $\mathbb{Z}_m$ that contains only trivial arithmetic progressions of length k. Furthermore, given a k, we want m to be as large as possible.

A close examination of the coloring of $\mathbb{Z}_{11}$ reveals that the nonzero elements colored red, $\{1, 3, 4, 5, 9\}$, are precisely the nonzero *quadratic residues* of $\mathbb{Z}_{11}$, that is, values r for which there is some $x \in \mathbb{Z}_{11}$ satisfying $x^2 = r$. Similarly, the nonzero elements colored blue, $\{2, 6, 7, 8, 10\}$, are the non-residues.

This suggests that we can look at colorings of $\mathbb{Z}_p$ where p is a prime and use the residues to help us determine the coloring. This has two important advantages.

- Suppose that $a, a + d, \ldots, a + (k-1)d$ is a nontrivial monochromatic k-AP, i.e., $d \neq 0 \pmod{p}$. Then, we can multiply each term by d^{-1} to get the arithmetic progression $ad^{-1}, ad^{-1} + 1, \ldots, ad^{-1} + (k-1)$, consisting of k consecutive elements of $\mathbb{Z}_p$. Further, since we started with a sequence that consisted either of only residues or non-residues, the resulting sequence must also consist of only residues or non-residues. (This is because if d^{-1} is a residue, then multiplication by d^{-1} sends residues to residues and non-residues to non-residues; a similar situation occurs if d^{-1} is not a quadratic residue.)

  This shows that to determine the length of the longest arithmetic progression in a coloring by quadratic residues, it suffices to find the longest run of residues or non-residues. This speeds things up tremendously and makes large computer searches possible.
- It is known that for $\mathbb{Z}_p$, there are no long runs of consecutive residues or non-residues. In particular, Burgess [**1**] showed that the maximum number of consecutive quadratic residues or non-residues for a prime p is $O(p^{1/4}(\log p)^{3/2})$.

  (The result of Burgess is not strong enough to guarantee the type of colorings we need for large k. If we compare the random bound with Theorem 4, then we need for the length of the longest monochromatic progression in $\mathbb{Z}_p$ to be $\leq \log_2(p)$ to yield a coloring that is better than random for some k. On the other hand, the result of Burgess is a general bound and does not preclude the possibility that for some primes p—or even for most primes p—the length of the longest run might be smaller than $\log_2(p)$.)

When we want to use more than two colors, we can use higher-order residues in place of quadratic residues to look for colorings that have only trivial monochromatic k-APs. In general, suppose p is a prime, so that $\mathbb{Z}_p^*$ (the invertible elements of $\mathbb{Z}_p$) is a group of order $p - 1$. Further, if $r \mid (p-1)$, then there is a unique subgroup of index r, namely

$$S = \{x^r \mid x \in \mathbb{Z}_p, x \neq 0\}.$$

Now $S \cup \{0\}$ may be written simply as

$$S \cup \{0\} = \{x^r \mid x \in \mathbb{Z}_p\}.$$

If there are only trivial monochromatic k-APs in $S \cup \{0\}$, then there can be only trivial k-APs in $yS \cup \{0\}$ for any $y \in \mathbb{Z}_p^*$ (since otherwise, if we multiply the progression by y^{-1}, we find a nontrivial k-AP in $S \cup \{0\}$, a contradiction). In particular, by choosing $y_1 = 1, y_2, \ldots, y_r$ so that the collection of $y_i S$ form the cosets of $\mathbb{Z}_p^*/S$, we form a coloring of $\mathbb{Z}_p$ with r colors that has no k-AP regardless of the color assignment of 0. We have now established the following.

THEOREM 5. *Suppose that p is an odd prime and $r \mid (p-1)$. If $\{x^r : x \in \mathbb{Z}_p\}$ contains only trivial k-APs, then there is a coloring with r colors where 0 can be colored arbitrarily, and that contains only trivial monochromatic k-APs.*

In some cases, we can combine two good colorings to form a larger coloring.

THEOREM 6. *For $i = 1, 2$, let $\mathcal{C}_i$ be a coloring of $\mathbb{Z}_{m_i}$ using r_i colors, where 0 can be colored arbitrarily and containing no nontrivial k-APs. Then there exists a coloring $\mathcal{C}$ of $\mathbb{Z}_{m_1 m_2}$ using $r_1 r_2$ colors, where 0 can be colored arbitrarily and containing no nontrivial k-APs.*

Proof. We show how to color $\mathbb{Z}_{m_1 m_2}$ using the colors (c_1, c_2), where c_1 is a color from $\mathcal{C}_1$ and c_2 is a color from $\mathcal{C}_2$. Given $0 \le z < m_1 m_2$, it can uniquely be written as $xm_2 + y$, where $0 \le x < m_1$ and $0 \le y < m_2$. This gives a map from $\mathbb{Z}_{m_1 m_2} \to \mathbb{Z}_{m_1} \times \mathbb{Z}_{m_2}$ with the following property: If $z_1 \mapsto (x_1, y_1)$ and $z_2 \mapsto (x_2, y_2)$, then $z_1 + z_2 \mapsto (x', y_1 + y_2 \pmod{m_2})$.

The coloring is now defined by the following: Given $z \in \mathbb{Z}_{m_1 m_2}$, we color it as $(c_1(x), c_2(y))$, where $c_1(x)$ is the color of x in $\mathcal{C}_1$ and $c_2(y)$ is the color of y in $\mathcal{C}_2$.

It remains to show that this has no nontrivial monochromatic k-APs. Suppose that $z_1, z_2, \ldots, z_k$ were a monochromatic k-AP where $z_i \mapsto (x_i, y_i)$. Then $y_1, y_2, \ldots, y_k$ is a monochromatic k-AP in $\mathcal{C}_2$. But this is only possible if $y_1 = y_2 = \cdots = y_k$.

We can think of a k-AP as found by sewing together several 3-APs. So we have for $2 \le i \le k - 1$ (with what we have about the equality of the y_i),

$$0 = z_{i-1} - 2z_i + z_{i+1} = m_2(x_{i-1} - 2x_i + x_{i+1}) \pmod{m_1 m_2},$$

which implies

$$x_{i-1} + x_{i+1} = 2x_i \pmod{m_1}.$$

This shows that we now have a sequence of 3-APs glued together in $\mathbb{Z}_{m_1}$; that is, $x_1, x_2, \ldots, x_k$ is a monochromatic k-AP in $\mathbb{Z}_{m_1}$. But this is only possible if $x_1 = x_2 = \cdots = x_k$.

We now conclude that $z_1 = z_2 = \cdots = z_k$; that is, there are no nontrivial monochromatic k-APs. Finally, we note that our argument does not depend on how we choose to color 0. ∎

By computer search, we can now find several colorings by either using residues (as in Theorem 5) or combinations of colorings from residues (as in Theorem 6). The results of the search are shown in TABLE 1. We indicate the best known value of m. In each case, the resulting number of monochromatic k-APs is $\frac{1}{2(m+1)(k-1)} n^2 + O(n)$; the resulting portion of monochromatic k-APs is $1/(m+1)$. We note that for each entry, this is the best-known coloring for minimizing the number of monochromatic k-APs in a coloring of $1, 2, \ldots, n$ using r colors.

The random bound is $1/r^{k-1}$. Each entry m in the table is significantly larger than r^{k-1}, which means that we we have beaten the random bounds significantly.

Open problems and variations

There is no reason why we must use residues for our colorings. As an example, if we work with $r = 3$, then -1, 0, and 1 are always cubic residues and so we cannot avoid 3-APs. On the other hand, there are colorings of $\mathbb{Z}_{12}$ using three colors (shown in FIGURE 4) which, by Theorem 4, when unrolled have $(1/48)n^2 + O(n)$ monochromatic 3-APs, so 8.33% of the colorings will be monochromatic. In a random coloring we would expect 11.11% of the 3-APs to be monochromatic. We still have a lot to learn about ways to color $\mathbb{Z}_m$ beyond residues, which give efficient unrollings.

TABLE 1: Best known m for which there exists a coloring using residues of $\mathbb{Z}_m$ with r colors, where 0 can be colored arbitrarily, and containing no k-APs

m	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$
$k=3$			37		103	
$k=4$	11	97	349	751	3259	1933
$k=5$	37	241	2609	6011	14173	30493
$k=6$	139	1777	139^2	49391	$139 \cdot 1777$	317969
$k=7$	617	7309	617^2	230281	$617 \cdot 7309$	
$k=8$	1069	34057	1069^2			
$k=9$	3389	116593				
$k=10$	11497	463747				
$k=11$	17863					
$k=12$	58013					
$k=13$	136859					
$k=14$	239873					
$k=15$	608789					
$k=16$	1091339					

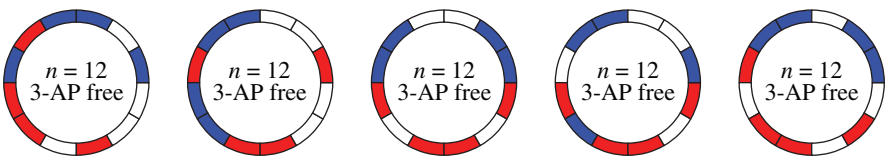


Figure 4 Colorings of $\mathbb{Z}_{12}$ with no nontrivial monochromatic 3-AP, using three colors

Further, all of our work in answering the original question about finding the minimum number of monochromatic arithmetic progressions has been in constructing colorings that give upper bounds. Much less is known about the lower bounds, and in particular *no* coloring is currently known to be optimal. For the simplest nontrivial case when $r = 2$ and $k = 3$, the best-known lower bound is due to Parrilo, Robertson, and Saracino [**8**]. They showed that if $f(n)$ is the minimal number of monochromatic 3-APs in a 2-coloring of $1, 2, \ldots, n$, then $f(n)$ lies in the interval

$$\frac{1675}{32768}n^2 + o(n^2) \approx 0.05111n^2 \leq f(n) \leq 0.05338n^2 \approx \frac{117}{2192}n^2 + O(n).$$

We conjecture that the upper bound, which is based on Theorem 3 above, is actually the correct limiting result.

CONJECTURE 1. *Each coloring of* $1, 2, \ldots, n$ *with two colors has at least* $\frac{117}{2192}n^2 + O(n)$ *monochromatic three-term arithmetic progressions.*

In general, we have provided some evidence to support that we can always beat the random bound for minimizing the number of monochromatic k-APs. While this evidence is compelling, we must be careful not to be misled by these small cases. We will need a more general machinery to establish the following.

CONJECTURE 2. *For each $k \geq 3$ and $r \geq 2$, there is a coloring of $1, 2, \ldots, n$ with r colors so that the minimum portion of monochromatic k-APs that occur is $cn^2 + o(n^2)$, where $c < \frac{1}{2(k-1)r^{k-1}}$.*

Besides considering ways to minimize the number of monochromatic k-APs, we can also try to work with other patterns. For example, we can try to avoid $a, a + 2d, a + 3d, a + 5d$ (which can be thought of as a 6-AP with some terms skipped over). In this case, the best-known coloring with two colors for unrolling is found by a coloring using the residues of $\mathbb{Z}_{13}$ and is shown in FIGURE 5 (as before, the white bit can be arbitrary). Similar calculations to Theorem 4 will show that this unrolling has $(1/140)n^2 + O(n)$ monochromatic patterns, so that 7.14% of the patterns are monochromatic. A random coloring would give 12.5% of the patterns monochromatic.

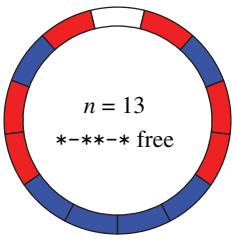


Figure 5 A coloring of $\mathbb{Z}_{13}$ with no nontrivial monochromatic $*-**-*$ pattern; the white bit can be either color.

We have explored a small corner of the colorful world of avoiding monochromatic arithmetic progressions. There are many beautiful open problems and variations left to explore. We look forward to seeing more progress in this area.

REFERENCES

1. D. A. Burgess, On character sums and primitive roots, *Proceedings of the London Mathematical Society* **12** (1962) 179–192.
2. Steve Butler, Kevin Costello, and Ron Graham, Finding patterns avoiding many monochromatic constellations, *Experimental Mathematics* **19** (2010) 399–411.
3. Peter Frankl, Ron Graham, and Vojtěch Rödl, Quantitative theorems for regular systems of equations, *Journal of Combinatorial Theory, Series A* **47** (1988) 246–261.
4. W. T. Gowers, A new proof of Szemerédi's theorem, *Geometric and Functional Analysis* **11** (2001) 465–588.
5. Ron Graham, Don Knuth, and Oren Patashnik, *Concrete Mathematics*, 2nd ed., Addison Wesley, Boston, MA, 1994.
6. Bruce Landman and Aaron Robertson, *Ramsey Theory on the Integers*, American Mathematical Society, Providence, RI, 2004.
7. Linyuan Lu and Xing Peng, Monochromatic 4-term arithmetic progressions in 2-colorings of Z_n, *Journal of Combinatorial Theory, Series A* **119** (2012) 1048–1065, http://www.sciencedirect.com/science/article/pii/S0097316511001932
8. Pablo Parrilo, Aaron Robertson, and Dan Saracino, On the asymptotic minimum number of monochromatic 3-term arithmetic progressions, *Journal of Combinatorial Theory, Series A* **115** (2008) 185–192, http://arxiv.org/abs/math/0609532
9. Bartel Leendert van der Waerden, Beweis einer Baudetschen Vermutung, *Nieuw Arch. Wisk.* **15** (1927) 212–216.

Summary We look at the problem of coloring $1, 2, \ldots, n$ with r colors to minimize the portion of monochromatic k-term arithmetic progressions. By using residues to color $\mathbb{Z}_m$ and then unrolling, we produce the best-known colorings for several small values of r and k.

STEVE BUTLER received his Ph.D. from the University of California, San Diego in 2008, under the direction of Fan Chung. After holding an NSF Mathematical Sciences Postdoctoral Fellowship at UCLA, he came to Iowa State University where he is currently an Assistant Professor in the Department of Mathematics. He enjoys teaching a variety of discrete math courses including, most recently, the mathematics of juggling.

RON GRAHAM is the Irwin and Joan Jacobs Professor in the Department of Computer Science and Engineering of the University of California, San Diego, and the Chief Scientist of California Institute for Telecommunications and Information Technology. He is a former president of the Mathematical Association of America and the American Mathematical Society and has been awarded numerous honors for his extensive groundbreaking work in discrete mathematics. Most recently, he co-wrote *Magical Mathematics* with Persi Diaconis, which was awarded the 2013 Euler Book Prize.

LINYUAN LU received his Ph.D. from the University of California, San Diego in 2002, under the direction of Fan Chung. He is currently a Professor in the Department of Mathematics at the University of South Carolina. His research interests are in spectral graph theory, random graph theory, and extremal combinatorics.

Math Bite: Indeterminate Forms

Before embarking on L'Hospital's rule, many teachers show that indeterminate forms ∞/∞ and $0/0$ are not automatically 1 by using simple examples like $\lim_{n\to\infty}(2n/n)$ and $\lim_{n\to\infty}((2/n)/(1/n))$. The exponential indeterminate forms 1^∞, ∞^0, and 0^0 are another matter. Comparably simple examples for these forms don't appear in calculus texts. For n a natural number, these examples suffice:

$$\begin{aligned} 2 = \lim_{n\to\infty} 2^{n/n} &= \lim_{n\to\infty} \left(2^{1/n}\right)^n && \text{(this is } 1^\infty\text{)} \\ &= \lim_{n\to\infty} \left(2^n\right)^{1/n} && \text{(this is } \infty^0\text{)} \\ &= \lim_{n\to\infty} \left(2^{-n}\right)^{-1/n} && \text{(this is } 0^0\text{).} \end{aligned}$$

Summary The indeterminate forms 1^∞, ∞^0, and 0^0 are not automatically 1.

Mark Lynch
Millsaps College, Jackson, MS 39210
lynchmj@millsaps.edu

Math. Mag. **87** (2014) 94. doi:10.4169/math.mag.87.2.94.

A 3-D Analog of Steiner's Porism

OWEN D. BYER
Eastern Mennonite University
Harrisonburg, VA 22802
byer@emu.edu

DEIRDRE L. SMELTZER
Eastern Mennonite University
Harrisonburg, VA 22802
smeltzed@emu.edu

Steiner's Porism, the classical result discovered by Jacob Steiner in the nineteenth century, is treasured for its beauty, for its simplicity, and as one of the great applications of inversion in the plane. In this note, we present a Steiner-like result in 3-space.

STEINER'S PORISM IN THE PLANE. *Consider fixed circles $\mathcal{C}$ and $\mathcal{D}$ in the plane, $\mathcal{C}$ interior to $\mathcal{D}$. Suppose a chain of circles $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \ldots, \mathcal{C}_n$ can be formed such that*

(i) *for $i < n$, $\mathcal{C}_i$ is tangent to $\mathcal{C}_{i+1}$, and $\mathcal{C}_n$ is tangent to $\mathcal{C}_1$; and*

(ii) *for $1 \leq i \leq n$, $\mathcal{C}_i$ is tangent to $\mathcal{C}$ and $\mathcal{D}$.*

Then any circle inscribed between $\mathcal{C}$ and $\mathcal{D}$ is one circle in a chain of n circles satisfying conditions (i) *and* (ii).

FIGURE 1 shows two closed chains (called *Steiner chains*) of length ten for a specific pair, $\mathcal{C}$ and $\mathcal{D}$.

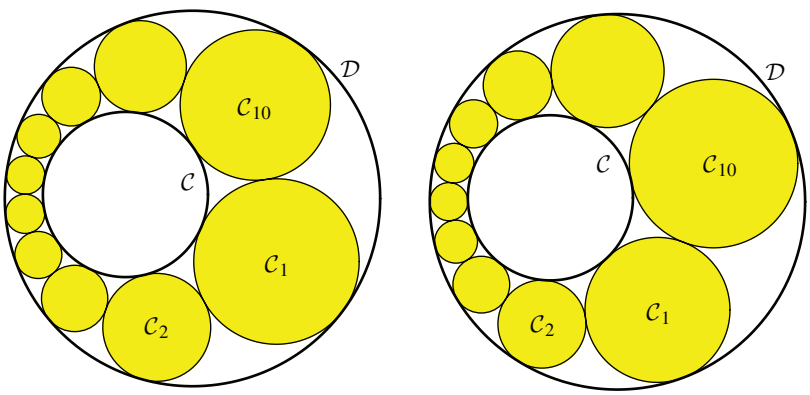


Figure 1 An example of Steiner's Porism with $n = 10$

In the closed chains of FIGURE 1, the circles $\mathcal{C}_i$ do not overlap. But it is possible for the chain of circles to wrap around the inner circle more than once, as in FIGURE 2. In such a situation, overlaps do occur.

A common proof of this result (see, for example, [**1**]) uses an appropriate inversion of the plane, which essentially allows us to transform $\mathcal{C}$ and $\mathcal{D}$ into concentric circles, $\mathcal{C}'$ and $\mathcal{D}'$, as in FIGURE 3. Since $\mathcal{C}'$ and $\mathcal{D}'$ are concentric, the presence of a chain of circles, and number of circles in the chain, depend only on the radii of $\mathcal{C}'$ and $\mathcal{D}'$, and not on the initial placement of ${\mathcal{C}_1}'$.

Math. Mag. **87** (2014) 95–99. doi:10.4169/math.mag.87.2.95.

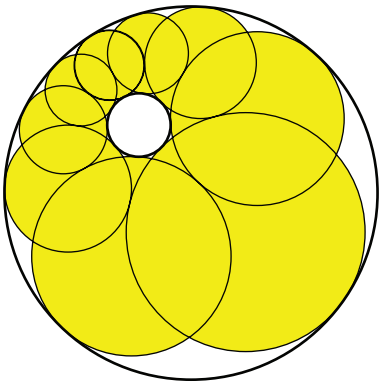


Figure 2 A Steiner chain with overlapping circles

Over the years, the Steiner chain result has been extended to three dimensions in various ways. Soddy's Hexlet is a well-known result first published in 1937 by Frederick Soddy, but stated as a Sangaku problem more than a century earlier. For this problem, consider two externally tangent spheres, $\mathcal{S}_1$ and $\mathcal{S}_2$, together with a third sphere, $\mathcal{S}$, to which $\mathcal{S}_1$ and $\mathcal{S}_2$ are internally tangent. Any closed chain of spheres in which each sphere is tangent internally to $\mathcal{S}$ and externally to $\mathcal{S}_1$ and $\mathcal{S}_2$ will contain exactly six spheres.

A *ring of spheres,* as defined by H.S.M. Coxeter in his 1952 paper [**2**], is a chain of spheres in which each sphere is tangent to its immediate predecessor and successor in the chain. Coxeter's work examined pairs of finite interlocked rings of spheres, in which each sphere in one ring is tangent to each sphere in the other ring. Soddy's Hexlet is a specific case of Coxeter's interlocked ring of spheres.

More recently, E. Roanes-Macias and E. Roanes-Lozano [**5**] extended the Steiner chains problem to three dimensions by constructing chains of spheres in such a way that each sphere in the chain is externally tangent to two fixed spheres, α and β, and internally tangent to a third fixed sphere, γ; no restrictions are placed on the placement of α, β, and γ relative to one another. Roanes-Macias and Roanes-Lozano establish conditions for the existence of such three-dimensional Steiner chains and prove that if, for a given configuration of fixed spheres α, β, and γ, a chain of n spheres meeting the tangency requirements is closed, then all such chains of n spheres for α, β, and γ are closed.

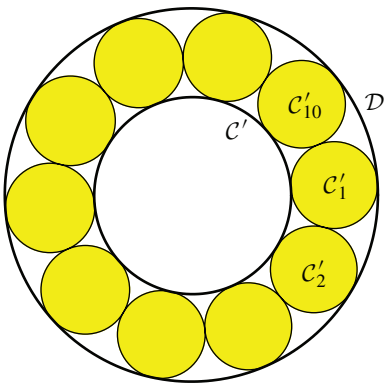


Figure 3 A Steiner chain between concentric circles $\mathcal{C}'$ and $\mathcal{D}'$

In this paper, we interpret an extension of Steiner's result to three dimensions by considering packings rather than chains. We begin with two fixed spheres, $\mathcal{S}$ and $\mathcal{T}$, one interior to the other, and then inscribe a collection of n spheres between them. We identify requirements on these new spheres in order to match the spirit of the 2-space result, in which a collection of circles fits tightly between the inner and outer circles. Specifically, we want each of the n inscribed spheres to be tangent to both $\mathcal{S}$ and $\mathcal{T}$ and to have a Steiner-type chain around it consisting of other spheres in the collection. The following result shows that if this is possible for a given n, then it will always be possible, regardless of where the first inscribed sphere is placed.

3-D VERSION OF STEINER'S PORISM. *Consider a sphere $\mathcal{S}$, interior to sphere $\mathcal{T}$. Suppose a set of n spheres $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n\}$ can be inscribed between $\mathcal{S}$ and $\mathcal{T}$ such that for $1 \leq i \leq n$,*

(i) *$\mathcal{S}_i$ is tangent to both $\mathcal{S}$ and $\mathcal{T}$, and*

(ii) *$\mathcal{S}_i$ is tangent to each sphere in a nonempty subset $\mathcal{A}_i \subset \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n\}$, where each sphere in $\mathcal{A}_i$ is externally tangent to exactly two other spheres in $\mathcal{A}_i$.*

Then, any sphere inscribed between $\mathcal{S}$ and $\mathcal{T}$ is one sphere in a set of n spheres satisfying conditions (i) and (ii). Furthermore, $|\mathcal{A}_i|$ is constant.

FIGURE 4 is an illustration with $n = 12$ and $\mathcal{A}_1 = \{\mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4, \mathcal{S}_5, \mathcal{S}_6\}$. (This image was created by Bob Allanson. More 3-dimensional images are at his website, http://members.ozemail.com.au/~llan.)

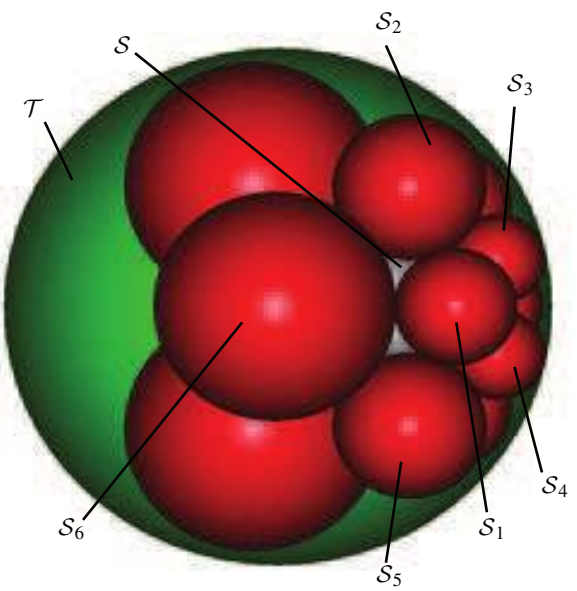


Figure 4 3-D Analog of a Steiner chain

Notice that the spheres in each $\mathcal{A}_i$ form a closed chain around $\mathcal{S}_i$, and the set $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n\}$ creates what could be viewed as a "packing" between $\mathcal{S}$ and $\mathcal{T}$. Obviously, such a packing will not always exist for a given $\mathcal{S}$ and $\mathcal{T}$, but when it does, each sphere in the set will be tangent to a fixed number, $|\mathcal{A}_i|$, of other spheres.

Our proof of the theorem uses spherical inversion in a similar fashion to the way inversion is used to prove the 2-dimensional result, so we first review the concepts and

properties of inversion. The following definition is actually valid in n-space. Since the ensuing properties of inversion in 3-space correspond to similar results with respect to inversion in 2-space (see, for example, [**1**], [**3**], or [**4**]), we omit their proofs.

DEFINITION. Let O be a fixed point and $r > 0$ a fixed number (radius). The inversion $I = I(O, r)$ maps every point $A \neq O$ onto a point A', such that A' is on the ray OA and $OA \cdot OA' = r^2$.

We list some properties of inversions in 3-space:

1. If $\mathcal{S}$ is a sphere not containing O, then the image of $\mathcal{S}$ under inversion is a sphere not containing O.
2. Points of tangency are preserved under inversion.
3. Given two non-intersecting spheres, an inversion exists under which their images are concentric spheres.

We now turn to the proof of the theorem.

Proof. Consider an inversion that maps $\mathcal{S}$ and $\mathcal{T}$ to concentric spheres, $\mathcal{S}'$ and $\mathcal{T}'$. Then the images of the $\mathcal{S}_i$ will form a set of congruent spheres, $\{\mathcal{S}_1', \mathcal{S}_2', \ldots, \mathcal{S}_n'\}$. Because points of tangency between the $\mathcal{S}_i$ are preserved under the inversion, each $\mathcal{S}_i'$ is tangent to $\mathcal{S}'$, $\mathcal{T}'$, and to the image of each sphere in $\mathcal{A}_i$. The symmetry and tangency of congruent spheres dictate that $|\mathcal{A}_i|$ is constant.

To show that the placement of an initial sphere, $\mathcal{S}_1$, is not important, let $\mathcal{T}_1'$ be a sphere between $\mathcal{S}'$ and $\mathcal{T}'$ and tangent to each of them. There is a rotational mapping centered at the mutual center of $\mathcal{S}'$ and $\mathcal{T}'$ that sends $\mathcal{S}_1'$ to $\mathcal{T}_1'$. Under this mapping, the images of the $\mathcal{S}_i'$ form a set of n congruent spheres satisfying properties (i) and (ii), regardless of the location of $\mathcal{T}_1'$. Thus, the location of $\mathcal{S}_1'$ is irrelevant, which implies that the same is true of the initial placement of its preimage $\mathcal{S}_1$ between $\mathcal{S}$ and $\mathcal{T}$. ■

Recall that in the 2-dimensional Steiner's Porism result, it is possible for the chain of inserted circles to wrap around the inner circle more than once before closing with $\mathcal{C}_n$ tangent to $\mathcal{C}_1$, resulting in overlapping circles. We do not know whether it is even possible for overlapping spheres to meet the conditions of our theorem, but we note that nothing in the statement or proof of our 3-dimensional result requires that the interiors of the spheres $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n$ be disjoint. However, when they are disjoint, we get a somewhat surprising restriction on n.

In the above proof, consider the (convex) polyhedron whose vertices correspond to the n centers of disjoint $\mathcal{S}_i'$, with edges joining the centers of adjacent $\mathcal{S}_i'$. As was noted, the $\mathcal{S}_i'$ are congruent, each having radius r equaling half the difference between the radius of $\mathcal{T}'$ and the radius of $\mathcal{S}'$. All edges will have length $2r$ and, due to the symmetry of the arrangement, all vertices will have the same degree, namely $|\mathcal{A}_i|$. Therefore, the faces of the polyhedron will be congruent regular polygons, identically joined at each vertex, which implies that the polyhedron will be one of the five Platonic solids. Furthermore, if $\mathcal{S}_i'$ is any sphere in the set $\{\mathcal{S}_1', \mathcal{S}_2', \ldots, \mathcal{S}_n'\}$ and $\mathcal{S}_j'$ and $\mathcal{S}_k'$ are images of spheres that are adjacent in $\mathcal{A}_i$, then $\mathcal{S}_i'$, $\mathcal{S}_j'$, and $\mathcal{S}_k'$ are all mutually tangent. Thus, the equilateral triangle joining their centers must be a face of the polyhedron. The only regular polyhedra with triangular faces are the tetrahedron, the octahedron, and the icosahedron, which proves the following corollary.

COROLLARY. *In the previous theorem, if the interiors of the spheres* $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n$ *are disjoint, then* $(n, |\mathcal{A}_i|)$ *must be one of the pairs* $(4, 3)$, $(6, 4)$, *or* $(12, 5)$.

We invite the reader to verify that spheres $\mathcal{S}$, $\mathcal{T}$, and $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n$ can actually be constructed, meeting the three conditions of the theorem for $n = 4$, $n = 6$, and $n = 12$.

Acknowledgments The authors thank Sam Vandervelde and Jeremy Dover for insightful comments that led to a cleaner and more focused paper.

REFERENCES

1. O. Byer, F. Lazebnik, and D. L. Smeltzer, *Methods for Euclidean Geometry*, Mathematical Association of America, Washington, DC, 2010.
2. H. S. M. Coxeter, Interlocked rings of spheres, *Scripta Mathematica* **18** (1952) 113–121.
3. H. S. M. Coxeter, *Introduction to Geometry*, 2nd ed., Wiley, New York, 1969.
4. N. A. Court, *Modern Pure Solid Geometry*, 2nd ed., Chelsea, New York, 1964.
5. E. Roanes-Macías, E. Roanes-Lozano, 3D extension of Steiner chains problem, *Mathematical and Computer Modelling* **45** (2007) 137–148, http://www.sciencedirect.com/science/article/pii/S0895717706001877.

Summary Steiner's Porism, the classical result on chains of circles discovered by Jacob Steiner in the nineteenth century, is treasured for its beauty, for its simplicity, and as one of the great applications of inversion in the plane. In this note, we extend his result to a packing of spheres in 3-space, along with a surprising connection to regular polyhedra.

More On Self-Tiling Tile Sets

LEE SALLOWS
Johannaweg 12, 6523 MA, Nijmegen
The Netherlands
lee.sal@inter.nl.net

Form is not emptiness;
emptiness is not form.

In an earlier article in this MAGAZINE [**7**], I introduced the idea of a *self-tiling tile set*, or hereinafter, *setiset*. A setiset of order n is a set of n planar shapes, each of which can be tiled with smaller replicas of the complete set of n shapes. It is required that the scaling factor be the same for each piece. FIGURE 1 shows an example of order 4 using distinctly shaped pieces. In fact, the definition given in the previous article demanded that the n pieces be distinct, whereas here no such requirement is made. This less stringent definition means that a setiset may be composed of identical pieces, in which case it becomes nothing less than a *rep-tile*, an animal that will figure prominently below.

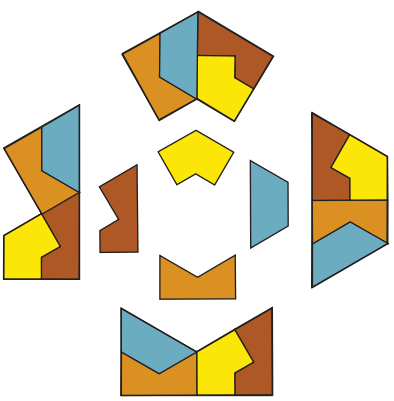


Figure 1 A setiset of order 4.

Since the pieces in any order-n setiset are all paved by the same n shapes, their areas are equal. And since the increase in scale from single piece to compound copy entails an n-fold increase in area, the linear scale factor involved is $\sqrt{n}$. Hence, in the case of pieces with integral side-lengths, $\sqrt{n}$ must be an integer. A setiset of integer-sided pieces, such as polyominoes, using a non-square number of tiles is thus impossible, so that after $n = 4$, the next possibility becomes $n = 9$.

Note also that the shapes employed do not have to be *connected* regions. Disjoint pieces composed of two or more separated islands are equally permissible, if aesthetically less appealing, although our human predeliction for connected pieces is difficult to defend on mathematical grounds. Setisets that include such pieces may thus be described as *disconnected* or *weakly-connected* (when piece islands join only at a point),

Math. Mag. **87** (2014) 100–112. doi:10.4169/math.mag.87.2.100.

while specimens composed of connected pieces that are also distinct in shape we shall call *perfect*.

How are setisets produced? In the absence of any known alternative, in the previous article I described a computer program that searched for solutions using polyominoes of a given size. Practical considerations meant that its trawls were restricted to sets of four pieces only, using polyominoes of sizes up to 8. These were major limitations.

Since then, however, a fresh insight has revealed a way to produce such sets using almost any (non-prime) number of pieces, and in a far greater variety of shapes. Moreover, the method is simple to the point of child's play and requires no computer. My main purpose here is thus to describe this technique, as well as to present a selection of the many new setisets it has brought to light.

Extracting setisets from rep-tiles

A *rep-tile* of order n (or *rep-n*) is a planar shape that can be tiled with n smaller duplicates of itself. It is easily seen that all triangles and parallelograms are rep-tiles, whereas a specimen of any order, n, may be created simply by stacking n rectangles of size $1 \times \sqrt{n}^{-1}$ on top of each other. The size of the resulting pile is then $\sqrt{n} \times 1$, which has the same shape as its components, the property required, albeit that the resulting figure is less than exciting. Skipping rep-tiles of orders 2 and 3, which are similarly dull, FIGURE 2 presents some more attractive examples of order 4, labeled R1 to R12. This list of examples is not exhaustive, but includes most of the interesting specimens I could discover in the literature and on the web [**4, 3, 8, 2, 1, 5, 6**].

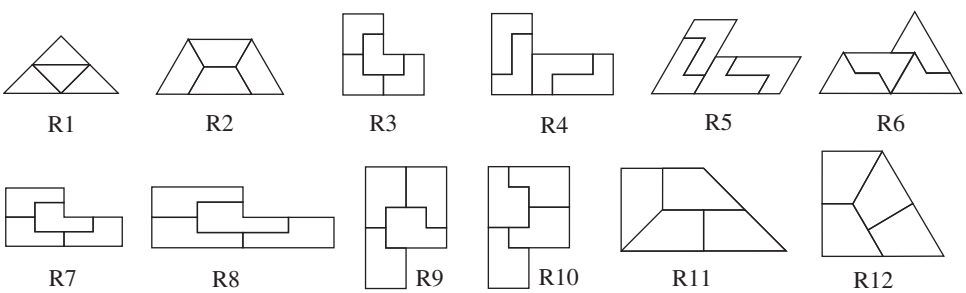


Figure 2 Some rep-tiles of order 4.

Consider rep-tile R6 in FIGURE 2, the hoary "sphinx." It is of order 4. As a rep-tile, we think of it as a single shape divided into units. However, considered as a collection of separate pieces, these units also form a setiset—one in which the four pieces are of the same shape.

Seen thus, we might wonder: Is there some way in which the identical pieces in this setiset can be reshaped so as to render them distinct, but without losing their self-replicating property?

There is. In fact, there is more than one way. The trick is to introduce a second copy of the same setiset. By combining these two sets in different ways, we can form distinct shapes that preserve their self-replicating property. In practice, what this boils down to is dissecting the rep-tiles appropriately.

FIGURE 3 shows three copies of the sphinx, each with its four sphinxlets. Suppose now that the rep-tile is dissected into two smaller pieces of equal area, each composed of a pair of sphinxlets. The result is a sphinx partitioned into two parts, such as *A*

and B, having different shapes. Taking next a second copy of the same rep-tile and repeating the process, but now using a different cut along the edges of the sphinxlets, we will then have four pieces, all different in shape. A further repeat will exhaust all possibilities, to leave us with a total of six distinct pieces, all shown in FIGURE 3.

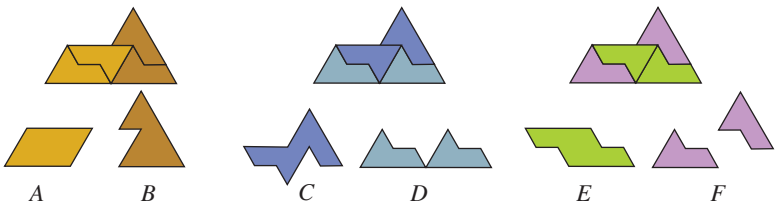


Figure 3 Three ways to dissect the sphinx.

Four of the resulting pieces are connected, A, B, C, E, one weakly-connected, D, and one disjoint, F. Now the pair A and B will together tile the bigger sphinx. The same goes for C and D. So the four pieces A, B, C, and D together tile two separate sphinxes. But two sphinxes are exactly the units from which A, B, C, and D are constructed. So, if these four pieces can pave two sphinxes, and two sphinxes are all we need to create A or B or C or D, then $\{A, B, C, D\}$ must be a self-tiling set. The conclusion is confirmed in FIGURE 4.

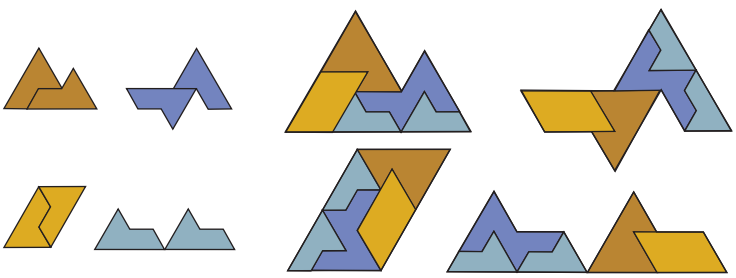


Figure 4 A setiset derived from the sphinx.

The four pieces are each composed of 2×6 unit triangles, and are thus dodeciamonds. To assist the eye, the outlines of the component sphinxlets in the smaller pieces have been left visible. See how the orientation of the two sphinxes making up each small piece is exactly mirrored by that of the larger (two-tone) sphinxes that compose their compound copies.

Above, I said there is more than one way to achieve the goal. This is because $\{A, B, C, D\}$ is not the only set of four we could have chosen. Beside the pairs A, B and C, D, the pair E, F also tiles the sphinx. Hence $\{A, B, E, F\}$ and $\{C, D, E, F\}$ are also setisets, although piece F may be thought less attractive because it is disconnected. Similarly, it is perhaps regrettable that one of the pieces, D, in our first setiset is itself only weakly connected, but the rep-tile structure affords us no better alternative.

Happily, however, this is but the start of our exploration. The rep-tile dissection technique just described is applicable to any rep-tile of order 4, and there remain several more specimens in FIGURE 2 to examine.

Not every rep-4 rep-tile can be dissected into six distinct pairs as above. Readers can check for themselves that rep-tiles R1, R2, R3, R4, and R5 in FIGURE 2 yield at

most five different pair-shapes. Four is sufficient for them to form a setiset, but they all include weakly-connected or disjoint pieces.

Turning next to rep-tile R7, a hexomino, inspection reveals it as the first to admit of dissections resulting in four fully connected pieces, to be seen in the setiset of FIGURE 5. Again, the orientation of the two hexominoes that make up each smaller piece are repeated in the compound constructions. This is a characteristic feature of every setiset of order 4 derived via the rep-tile dissection method.

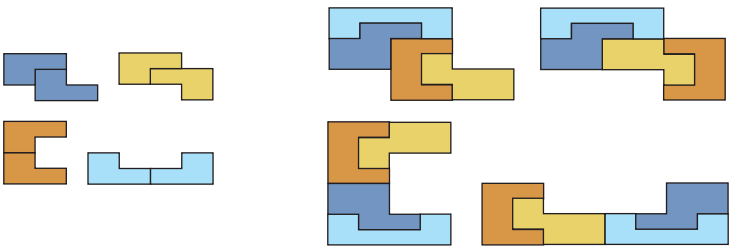


Figure 5 A setiset based on rep-tile *R7* in Figure 2.

As might be expected, the close resemblance between rep-tiles R7 and R8 in FIGURE 2 is echoed in the setisets they give rise to, the one resulting from R8 being an exact analog of the one shown in FIGURE 5. Surprising perhaps is that R3, which again resembles the structure of both R7 and R8, fails even to yield four distinct connected pieces.

A more interesting case is rep-tile R9, which has the shape of a P-pentomino. FIGURE 6 shows the various smaller rep-tile pairs into which it can be dissected, with again one non-connected piece (F).

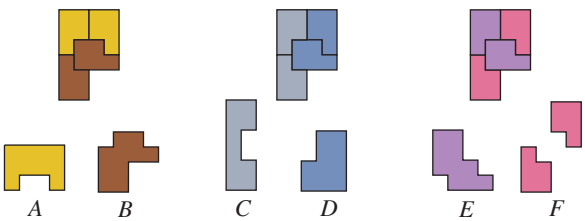


Figure 6 One of the two *P*-shaped rep-tiles dissected thrice.

Leaving aside F (along with its complement, E), there remain four fully-connected pieces that together pave the pentomino twice, which is all that is required of them to create a setiset. It can be seen in FIGURE 8, at left.

There exists, however, a second version of this same rep-tile in which the paving is very slightly different; see R10 in FIGURE 2. The difference between the two versions is so insignificant as to be routinely overlooked. Nevertheless, even such a tiny difference makes for the distinct set of pairwise dissections seen in FIGURE 7, as comparison with FIGURE 6 will show.

Here again, aside from piece Z, there remain four connected shapes that can pave two P-pentominoes. The associated setiset can be seen at right in FIGURE 8, alongside that using pieces from FIGURE 6. However, bearing in mind that a setiset results from any four pieces able to tile the rep-tile twice, there remain yet two more connected-piece setisets to be extracted here. They are those that arise by taking one pair from

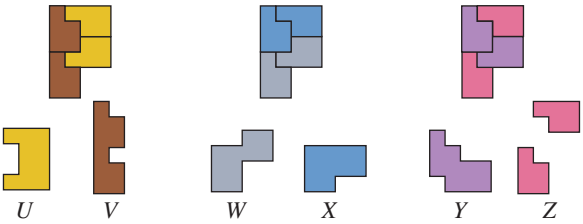


Figure 7 The alternative *P*-shaped rep-tile dissected thrice.

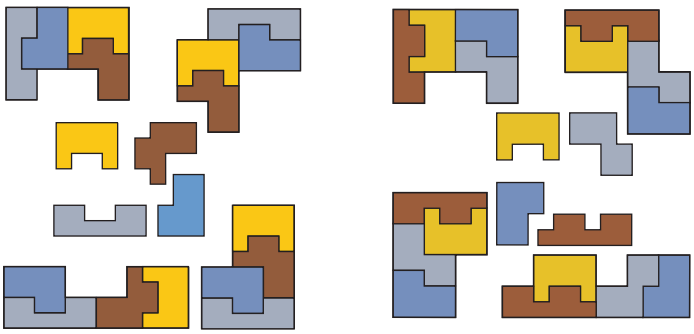


Figure 8 The setisets derived from Figure 6 (left) and Figure 7 (right).

FIGURE 6 together with one pair from FIGURE 7. The resulting two new solutions based upon $\{A, B\} \cup \{W, X\}$ and $\{C, D\} \cup \{U, V\}$ are pictured in FIGURE 9.

Of course, $\{A, B\} \cup \{U, V\}$ and $\{C, D\} \cup \{W, X\}$ also form setisets, but contain two pieces that are alike: A and U in the first, D and X in the second. Even so, the P-pentomino rep-tile proves itself a remarkably rich source of setisets.

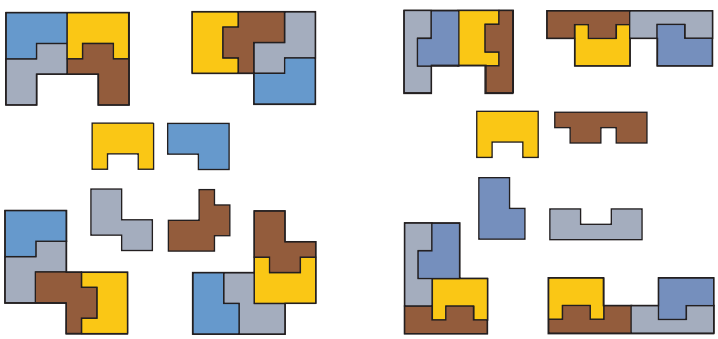


Figure 9 The pieces in Figures 6 and 7 yield two further setisets.

The rep-tile R11 in FIGURE 2 is a polyform built up from units having non-integral side-lengths: 1, 1, 2, and $\sqrt{2}$. It can be dissected into two connected halves in two different ways to result in the setiset of FIGURE 10.

This brings us to the final example in FIGURE 2, rep-tile R12, the shape of which is one quarter (quadrant?) of a regular hexagon. Once again, decomposing the rep-tile into two connected halves in two different ways results in a further setiset of order 4. It is the one with which we started, pictured in FIGURE 1.

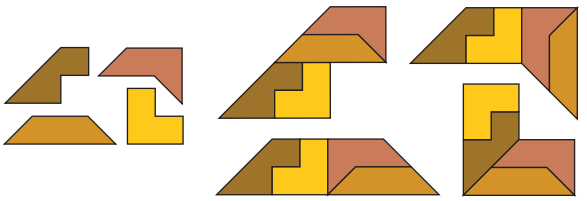


Figure 10 A setiset with non-integral side length pieces.

Higher-order setisets

The dissection method can be applied to rep-tiles beyond order 4, but not in every case. This follows from the fact that the pieces in any setiset are equal in area. Thus, if a set is derived from a rep-tile (not all are), the number of rep-tile units contained in each piece must be the same (and also be greater than 1). Hence, n, the order of the rep-tile, must be divisible by an integer, showing that prime orders will not work. Also, as the number of pieces goes up, so do the number of dissections required. Hence, if d is the number of times the rep-tile is to be dissected, and p is the number of parts into which it is split, then $d \times p = n$. A first step in constructing a setiset of order n is therefore to look at the factors of n so as to see how many dissections of what kind will be needed.

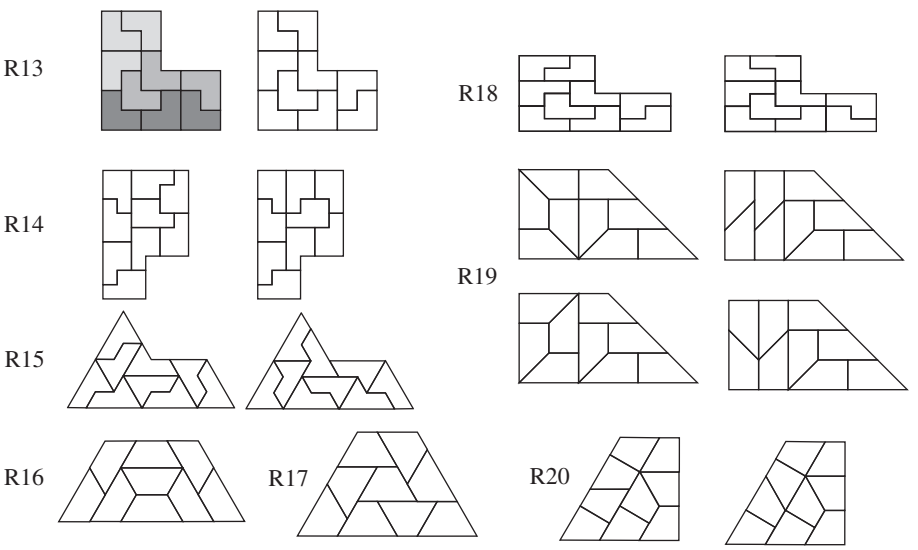


Figure 11 Some rep-tiles of order 9.

FIGURE 11 shows some rep-tiles of order 9, labeled R13–R20. Note the return of shapes earlier seen in FIGURE 2, in fact all except R17. These are among other rep-tiles that can be inflated to form new specimens of higher order. Several of these examples admit more than one paving. FIGURE 12 shows a setiset of 9 pieces (above) derived from rep-tile R13. Being of order $9 = 3 \times 3$, the only possible values for d and p are $d = 3$ and $p = 3$, indicating a need for three dissections into three pieces. The piece shapes seen were arrived at by trial and error. The dissections of the rep-tile that give rise to them can be seen in the L-shapes at the bottom of FIGURE 12. Deciding on how to make these dissections may look like a difficult task, but is really no more than a trivial puzzle. The requirement here was to thrice cut the rep-tile into thirds without

producing any repeated shapes. For example, the shading of R13 at left in FIGURE 11 shows an alternative dissection into three pieces. Had we started with this, the next question would be: Can two further dissections into thirds be found so as to yield a total of nine distinct pieces? For rep-tiles of small order such as this, trial and error suffices to find out. For much larger rep-tiles, a computer program could be written to do the job. In fact, here the goal of nine distinct pieces cannot be achieved, forcing a retry with a new dissection. Even now, both pavings of R13 in FIGURE 11 are needed to attain nine distinct pieces.

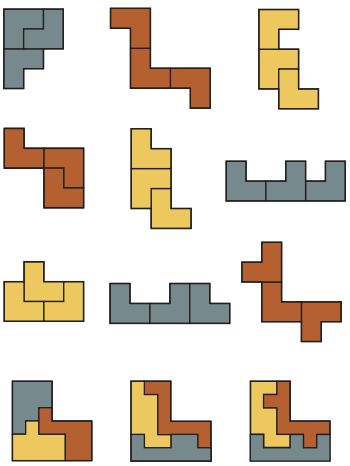


Figure 12 A setiset of order 9.

Returning to FIGURE 12, rather than show the nine pavings of each piece, which would make for an over-intricate picture, the self-tiling property of the set is made manifest more simply. As shown by the thin black lines, every one of the nine pieces can be tiled by three identical L-shapes. Below is seen that these same L-shapes are paved by the complete set of nine pieces. Hence, the latter form a setiset, QED. Note that the three compound L-shapes can tile any piece shape in 3! different ways.

FIGURE 13 shows two comparable perfect setisets derived from rep-tiles R18 and R19 in FIGURE 11. The exotic piece shapes might suggest wizardry to the uninitiated, whereas the method of patiently dissecting the rep-tiles into nine non-congruent regions of three units remains as simple as described above.

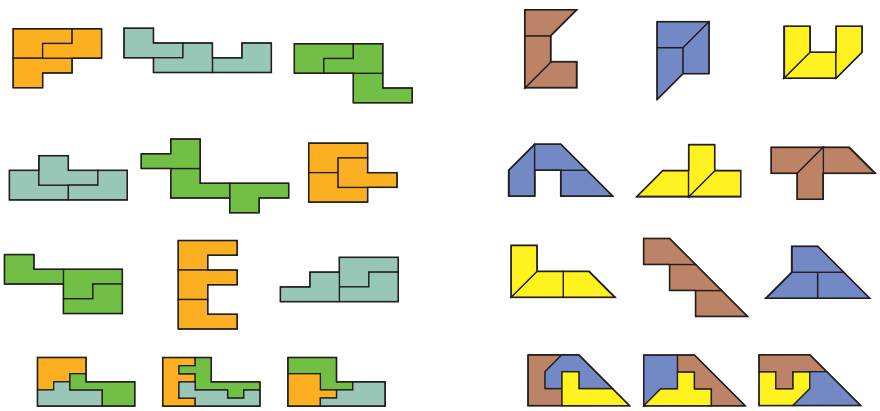


Figure 13 Order 9 setisets based on rep-tiles *R*18 (left) and *R*19 (right).

Not every rep-tile can be dissected so as to yield nine connected pieces. R15, R16, and R17 fall into this category, even though R15, the order-9 sphinx, comes in two versions that make for distinct dissections. It is important to realize that any distinct pavings of a single rep-tile may all be used in deriving the same setiset. As was just seen with FIGURE 12, the extra flexibility thus conferred can be crucial to success. Note that the second paving of R13 (right) results simply by flipping its component 2×3 rectangle at bottom right. This trick of flipping bilaterally symmetric regions can spawn many variant tilings. Four variants among others are shown for R19 in FIGURE 11.

Above, we saw that the rep-4 P-pentomino rep-tile gave rise to four different setisets showing connected pieces. FIGURE 14 shows how the two variants of R14, which is the order-9 version of the same rep-tile, can be dissected into three non-congruent thirds in a total of 11 different ways. Any three of these dissections will form a perfect setiset, provided they include nine distinct pieces. Analysis reveals 130 such cases, a single example of which gives rise to FIGURE 15 left. FIGURE 15 right shows a final example of order 9 based upon R20 in FIGURE 11. By now, the reader should be able to take in the self-tiling property at a glance.

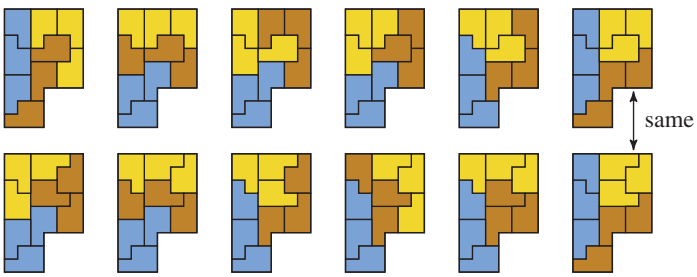


Figure 14 The two variants of rep-tile *R*14 can be dissected in 11 different ways.

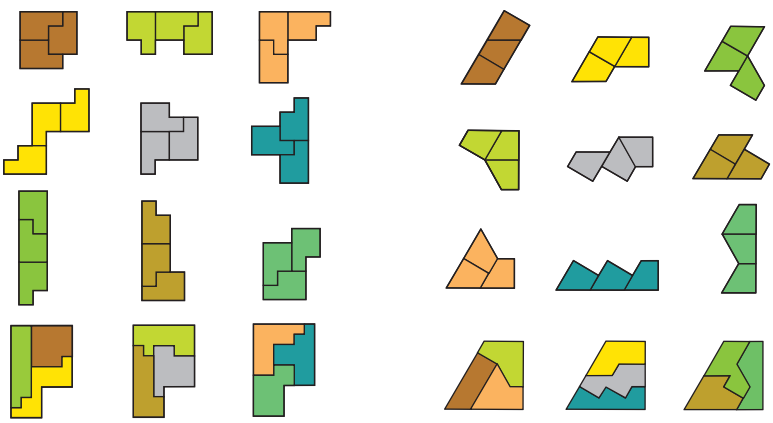


Figure 15 Setisets derived from rep-tiles *R*14 (left) and *R*20 (right).

So much for the unproblematic case of order 9. Similarly, order 16 presents no difficulties. However, a serious obstacle met with when it comes to the non-square, non-prime orders 6, 8, 12 and 14, is the absence of suitable rep-tiles. For example, a search of the literature reveals but one rep-tile of order 6, which is the one formed by a stack of six $1 \times \sqrt{6}$ rectangles. However, since $6 = 2 \times 3$, either $d = 2$ and $p = 3$, or $d = 3$ and $p = 2$, meaning two dissections into three thirds or three dissections into two halves, both of which are impossible to realize. Happily, in this case further

thought turned up what I believe to be a previously unknown L-shaped rep-tile of order 6. It can be seen at left in FIGURE 16. In the center is an order-6 setiset it gives rise to, at right of which are seen the two distinct pavings of the rep-tile required. Alas, although alternatives exist, at least one disjoint piece is unavoidable.

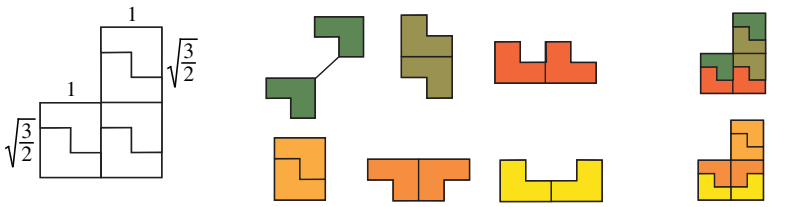


Figure 16 A (new?) order 6 rep-tile and associated setiset that includes a disjoint piece.

Likewise, the only setiset of order 8 yet found is one including four disconnected pieces that is based on a $\sqrt{8} \times 4$ rectangular rep-tile. Excessive as so many disjoint pieces may seem, this does at least furnish a solution, whereas when it comes to orders 12 and 14, solutions are lacking completely.

FIGURE 17 shows an unusual-looking setiset of order 10. It is derived from the rep-tile shown beneath, a right-angled triangle with sides of length 1, 3, and $\sqrt{10}$. Here the factors of 10 dictate five dissections into two halves. The alternative of two dissections into five fifths proves impossible. At left, an example shows how the ten pieces tile one of their own number. If it seems that some pieces are merely flipped versions of others, don't be deceived; all are distinct.

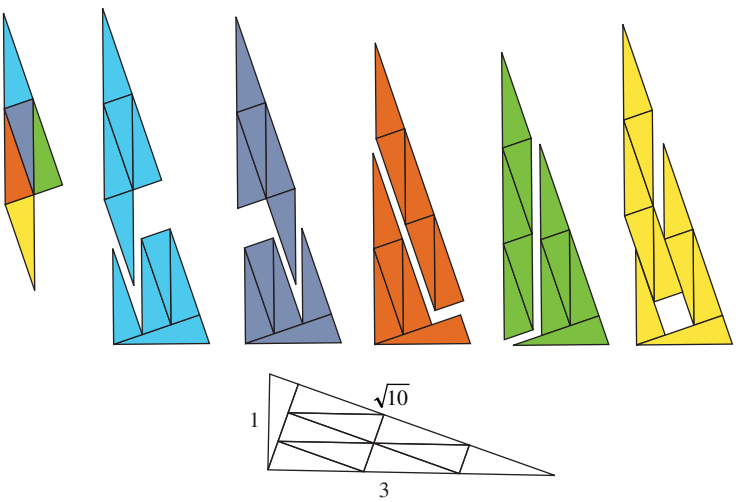


Figure 17 A setiset of order 10 based on the rep-tile shown below.

I conclude this brief tour of the setiset flora with two final examples of order 16. From $16 = 4 \times 4 = 2 \times 8 = 8 \times 2$, we find there are three possibilities. FIGURE 18 makes use of the L-shaped rep-tile already encountered in order-4 and order-9 form, but now of order 16. Four-fold dissections of the L into four non-congruent quarters result in 16 distinct pieces that tile themselves. Finding a rep-tile that can be dissected twice into eight distinct eighths, or eight times into two distinct halves are challenges I leave to the reader.

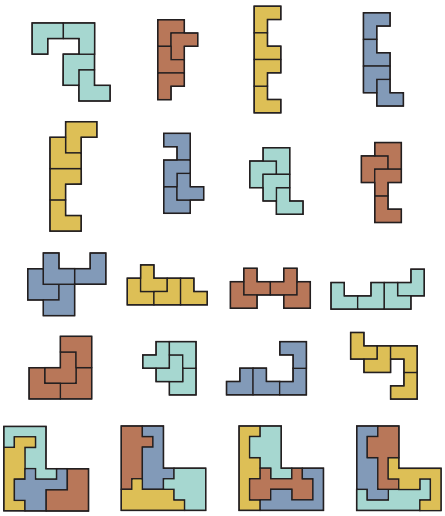


Figure 18 A setiset of order 16.

A search for the setiset in FIGURE 19 was undertaken after noticing that FIGURE 18 resembles a 4×4 geomagic square. For that is exactly what FIGURE 19 is: The four pieces in each row and column (but not the diagonals) all pave a 4×8 rectangular target. Moreover, each of the 16 pieces is itself composed of four 2×4 rectangles. An enlarged copy of any piece can thus be assembled using either the four row or the four column targets, which are of similar shape.

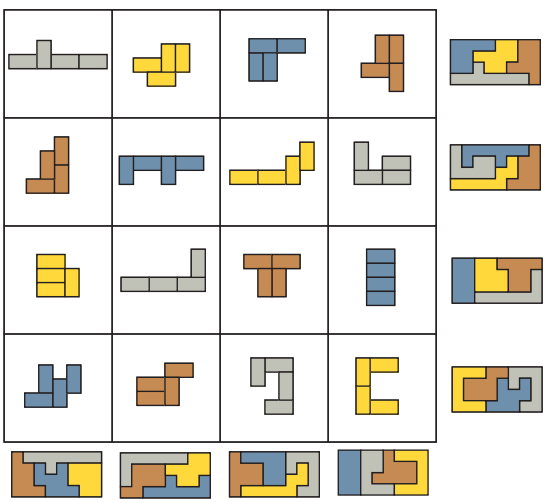


Figure 19 An order 16 setiset arranged so as to form a 4×4 geomagic square.

Double-rep-tiles and self-constructors

Rep-tiles are not the only tile sets to exhibit self-similarity. Besides rep-tiles, there are sets that include more than one kind of shape. For example, a set of two distinct shapes, each of which can be tiled by two smaller duplicates of both, is what I call a *double*-rep-tile.

FIGURE 20 shows an example in which the areas of distinct color are pentiamonds of two kinds: A-shaped (orange) or B-shaped (yellow). Here, the double-rep-tile is composed of the two twice-sized or "big-pentiamonds," A and B, seen at top. Both A and B are each tiled by two A-shaped and two B-shaped pentiamonds. The two tints of orange help to distinguish distinct components. As with rep-tiles, the self-similar properties of double-rep-tiles yield setisets.

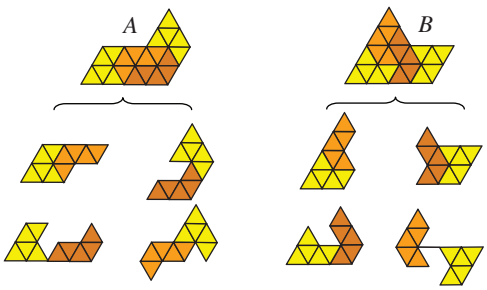


Figure 20 A double-rep-tile: two shapes tiled by two smaller duplicates of both.

Below A and B are seen their two possible dissections into two halves each composed of an A- and a B-pentiamond. They include a weakly-connected and disjoint piece. Since an A-pair will unite to form A, and a B-pair to form B, any two such pairs yield a setiset of order 4. Among the four possibilities, the two pairs immediately under A and B furnish one using connected pieces, shown in FIGURE 21 at left.

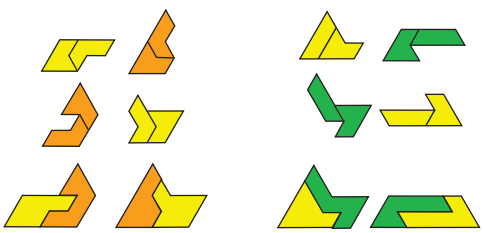


Figure 21 Two setisets derived from two double-rep-tiles.

Unlike setisets derived from rep-tiles, in which pieces are composed of identical units, here they are built up from distinctly-shaped pairs. At the right of FIGURE 21 is a setiset derived from a second double-rep-tile, again composed of two big-pentiamonds.

These double-rep-tiles were discovered with the aid of a computer program that identifies every possible tiling of a big-pentiamond by four pentiamonds. A similar program was used to find a pair of double-rep-tiles based upon polyominoes. Both use P-shaped and V-shaped big-pentominoes. The P-shaped one can be tiled in two different ways, as FIGURE 22 shows.

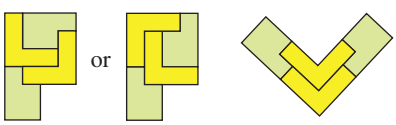


Figure 22 A double rep-tile using P and V pentominoes.

The dissections of these pieces result in three setisets having connected pieces. FIGURE 23 shows that the three share two pieces in common, while two of them differ in one piece only.

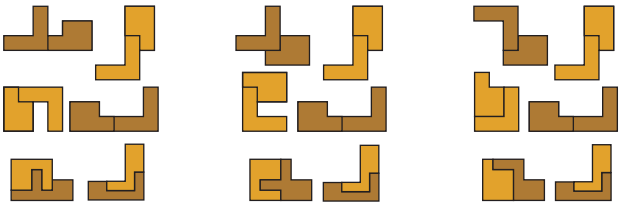


Figure 23 Three order 4 setisets derived from the double-rep-tile of Figure 22.

As before, every piece is composed of two distinct units. After looking at these specimens, it is not difficult to imagine setisets whose pieces are composed of three or more distinct units, a contingency that invites us to contemplate *multi-rep-tiles*. On the other hand are piece sets in which the number of copies of each component shape is not the same in each case. For example, FIGURE 24 shows a remarkable set of three distinct big-pentiamonds labeled A, B, and C. If we let a, b, and c stand for their corresponding pentiamonds, and write $P, Q \leftarrow r, s$ to indicate that both P and Q are tiled by r as well as by s, then the properties of A, B, and C as seen in FIGURE 24 are as follows: $A, B, C \leftarrow 2a + b + c, a + 2b + c, a + b + 2c$, where the 2's indicate two copies.

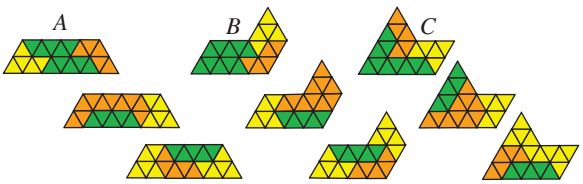


Figure 24 A remarkable set of self-tiling pentiamonds.

No setisets result from dissecting these pieces, but instead can be found a pair of what I tentatively dub *self-constructors* of order 6. FIGURE 25 shows one example. Each of the six pieces in the set is paved by some four of them.

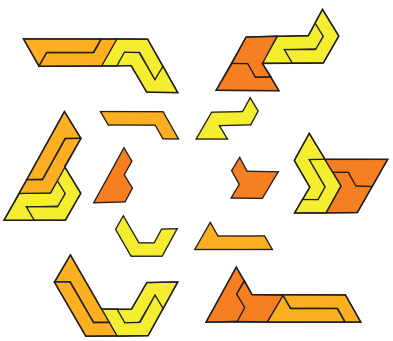


Figure 25 A "self-constructor." Each piece is paved by some four of them.

Note that the three shapes in FIGURE 24 do not qualify as a "triple-rep-tile," which would be a set of three distinct shapes, each of which is tiled by three smaller copies of themselves. I have no idea whether such creatures even exist, although there seems no reason to exclude the possibility. But what FIGURE 24 does demonstrate is that besides multi-rep-tiles of higher degree, there may exist still stranger animals sporting pieces composed of duplicates that occur *in differing numbers*. As to what peculiar forms of self-replicator these may yet lead in the future, who will say?

REFERENCES

1. Andrew Clarke, The Poly Pages, http://recmath.org/PolyPages/.
2. Steven Dutch, Rep-Tiles, http://www.uwgb.edu/dutchs/symmetry/reptile1.htm.
3. Martin Gardner, *The Unexpected Hanging and Other Mathematical Diversions,* University of Chicago Press, Chicago, 1991.
4. Solomon Golomb, Replicating figures in the plane, *Math. Gazette* **48** No. 366 (1964) 403–412.
5. Stewart Hinsley, Rep-Tiles, http://www.meden.demon.co.uk/Fractals/reptiles.html.
6. Michael Reid, http://math.cos.ucf.edu//~reid/index.html.
7. Lee Sallows, On self-tiling tile sets, *Math. Mag.* **85** (2012) 323–333, http://dx.doi.org/10.4169/math.mag.85.5.323.
8. Eric Weisstein, Rep-Tile, http://mathworld.wolfram.com/Rep-Tile.html.

Summary This is an extension of the author's previous work on self-tiling tile sets. Hitherto, the only known method for discovering the latter was by means of a brute-force computer search. A new pencil and paper method for extracting such sets from rep-tiles is introduced, and a plethora of new examples presented.

LEE SALLOWS is a retired electronics engineer. Now that he no longer goes out to work, he stays home and does exciting things with vacuum cleaners, furniture polish, and suchlike. In between times he looks after two splendid cats, Yorick and Twinkle, who are crazy about recreational math, and from whom he gets his best ideas.

The Geometry of Cubic Polynomials

CHRISTOPHER FRAYER
University of Wisconsin-Platteville
Platteville, WI 53818
frayerc@uwplatt.edu

MIYEON KWON
University of Wisconsin-Platteville
Platteville, WI 53818
kwonmi@uwplatt.edu

CHRISTOPHER SCHAFHAUSER
University of Nebraska-Lincoln
Lincoln, NE 68588
cschafhauser2@math.unl.edu

JAMES A. SWENSON
University of Wisconsin-Platteville
Platteville, WI 53818
swensonj@uwplatt.edu

In 2009, Dan Kalman won the MAA's Lester R. Ford Award for an article [**7**] on what he has called "the most marvelous theorem in mathematics." This result was first proved in 1864 by Jörg Siebeck [**11**], and geometric proofs were given by Bôcher and Grace, independently, in 1892 and 1901 respectively, but Kalman names the result after Morris Marden, who proved a generalized version in 1945 [**9**].

MARDEN'S THEOREM. *Given a triangle $\triangle r_1 r_2 r_3$ in the complex plane, there is a unique inscribed ellipse, called the Steiner ellipse, which is tangent to the sides of the triangle at the midpoints of the sides. If*

$$p(z) = (z - r_1)(z - r_2)(z - r_3),$$

then the roots of $p'(z)$ are the foci of the Steiner ellipse of $\triangle r_1 r_2 r_3$, and the root of $p''(z)$ is the centroid of $\triangle r_1 r_2 r_3$.

This is indeed a marvelous theorem. What we find so attractive about it is the connection between analysis and geometry: A critical point is a focus of an ellipse! Before learning this, all we had known about the critical points of a complex polynomial was the Gauss–Lucas theorem, which guarantees that the critical points of any polynomial lie in the convex hull of its roots.

Kalman's article was particularly exciting for those of us who were already intrigued by the "polynomial root dragging" introduced by Bruce Anderson [**1**]. Anderson and others [**2, 3, 4, 5**] had investigated how the critical points of a polynomial depend on its roots, when all roots are real.

We wondered: What happens to the critical points of a complex cubic polynomial when its roots move? We investigated using GeoGebra, a freely-available software package for plane geometry. We set the roots of a polynomial in motion on the screen, and watched the critical points move. You can find an assortment of our GeoGebra notebooks at our website (http://www.uwplatt.edu/~swensonj/gocp/). We invite you to play with these figures as you read.

On the real line, there are relatively few options: Just decide how far to drag each root to the right. The complex plane offers more degrees of freedom. To focus our

Math. Mag. **87** (2014) 113–124. doi:10.4169/math.mag.87.2.113.

efforts, we made a choice. Through any three points r_1, r_2, r_3 in the complex plane, there is a unique circle. By changing coordinates, we can take this circle to be the unit circle and fix $r_3 = 1$. We set r_1 and r_2 in motion around the unit circle and traced the loci of the critical points.

We were surprised to see that as the roots varied, the trajectories of the critical points always avoided a certain disk. Further investigation explained this fact (Theorem 4 below) and revealed a level of structure by which we were again surprised: A critical point of such a polynomial almost always determines the polynomial uniquely (Theorem 7). It is our pleasure to share with you the geometry (especially Theorem 3) behind these analytic facts. It is a connection that we hope Marden would have enjoyed.

Centers

Let $p(z)$ be a cubic polynomial with its roots on the unit circle in $\mathbb{C}$, having a root at 1 (that is, $p(1) = 0$). We assume that p is monic: This can be achieved without changing the roots of $p(z)$ or of its derivatives. Our results can be applied to polynomials whose roots lie elsewhere by changing the coordinate system. Thus, from now on, we will assume that p belongs to the family of cubic polynomials

$$\Gamma = \{q : \mathbb{C} \to \mathbb{C} \mid q(z) = (z-1)(z-r_1)(z-r_2),\ |r_1| = |r_2| = 1\}.$$

Before studying the zeros of $p'(z)$, we take a moment to investigate the zeros of $p''(z)$.

DEFINITION. Given $p \in \Gamma$ and $g \in \mathbb{C}$, we say g is the *center* of $p(z)$, provided that $p''(g) = 0$.

Since p has degree 3, it is clear that every $p \in \Gamma$ has a unique center. More interestingly, we can prove something like a converse to this fact.

THEOREM 1. *Let $g \in \mathbb{C}$.*

1. *$p \in \Gamma$ has center $\frac{1}{3}$ if and only if $p(z) = (z-1)(z-r_1)(z-r_2)$ with $r_2 = -r_1$.*
2. *If $0 < \left|g - \frac{1}{3}\right| \le \frac{2}{3}$, then there is a unique polynomial $p \in \Gamma$ with center g.*
3. *If $\left|g - \frac{1}{3}\right| > \frac{2}{3}$, there is no $p \in \Gamma$ with center g.*

In words, the center of any $p \in \Gamma$ must lie in the closed disk bounded by the dashed circle in FIGURE 1. Conversely, almost every point in that closed disk is the center of a unique $p \in \Gamma$. The single exception, the point $1/3$ at the center of the dashed circle, is the center of infinitely many polynomials—one for each pair $\{r, -r\}$ on the unit circle.

Proof. By Marden's theorem, this claim is related to the construction of a triangle given a vertex, the centroid, and the circumcircle.

Suppose that g is the center of $p(z) \in \Gamma$. Then by Marden's theorem, or by algebra, g is the centroid of triangle $\triangle r_1 r_2 r_3$, where r_1 and r_2 are points to be constructed on the unit circle, and $r_3 = 1$.

Though we do not know where r_1 and r_2 are, let us denote the midpoint of $\overline{r_1 r_2}$ by w. Then w lies in the closed unit disk. The segment $\overline{r_3 w}$ is a median of $\triangle r_1 r_2 r_3$, so $g = \frac{2}{3}(w) + \frac{1}{3}(1)$. By the triangle inequality, it is immediate that $\left|g - \frac{1}{3}\right| \le \frac{2}{3}$.

Consequently, to begin the construction, we solve the previous equation for w and let $w = (3g-1)/2$. (In geometric terms, given g, we may construct the midpoint w' of $\overline{r_3 g}$; then w is the reflection of w' over g.)

Assume first that $g \neq 1/3$, so $w \neq 0$. We know that $\overline{r_1 r_2}$ is a chord of the unit circle, so its perpendicular bisector must pass through the origin O and through the midpoint w.

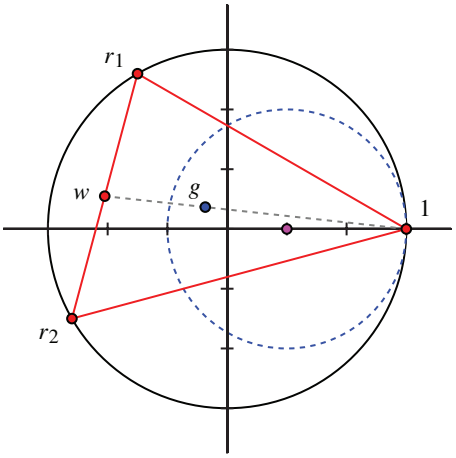


Figure 1 Construction of $\triangle r_1 r_2 1$ given center g (Theorem 1)

Hence, we find r_1 and r_2 by constructing the line L through w, perpendicular to $\overline{Ow}$. Since w lies in the closed unit disk, L intersects the unit circle in two points (counting by multiplicity if $|w| = 1$); these are r_1 and r_2.

On the other hand, suppose that $w = 0$ and $g = 1/3$. Since $w = 0$ is the midpoint of $\overline{r_1 r_2}$, we have $r_2 = -r_1$. Conversely, we can compute directly that $1/3$ is the center of $p(z) = (z-1)(z^2 - r_1^2) = z^3 - z^2 - r_1^2 z + r_1^2$. ■

Theorem 1 draws our attention to a circle that is internally tangent to the unit circle at 1. Such circles will be very important to us, so let us fix some notation. Given $\alpha > 0$, we denote by T_α the circle of diameter α that passes through 1 and $1 - \alpha$ in the complex plane; see FIGURE 2. Symbolically,

$$T_\alpha = \left\{ z \in \mathbb{C} : \left| z - \left(1 - \frac{\alpha}{2}\right) \right| = \frac{\alpha}{2} \right\}.$$

In this notation, Theorem 1 says that a point $g \in \mathbb{C}$ will be the center of zero, one, or infinitely many polynomials $p \in \Gamma$, depending on whether g lies outside, inside, or on the circle $T_{4/3}$.

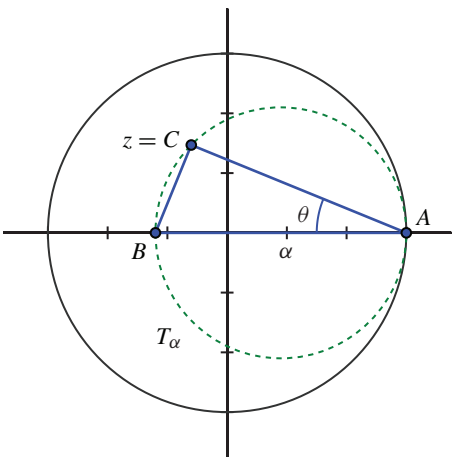


Figure 2 Each value of z lies on a unique T_α.

Critical points

We return to our original questions: Where can the critical points of $p \in \Gamma$ lie, and to what extent do they determine p?

Regarding the first question, Saff and Twomey show in [**10**] that p has at least one critical point in the closed disk bounded by T_1. Moreover, they show that if c_1 and c_2 are the critical points of p, and if $c_1 \neq 1$ is on T_1, then c_2 is the complex conjugate of c_1 and hence also lies on T_1.

We shall take advantage of a delightful geometric symmetry to extend the results of Saff and Twomey. Let's begin by walking through a couple of examples, to which we shall refer later.

EXAMPLE A. Suppose $p \in \Gamma$ has a critical point at 1. Then we know that p has a double root at 1. (Though this may be well known, we provide a proof: Since $p(z) = (z-1)q(z)$ for some quadratic polynomial q, the product rule gives $p'(z) = q(z) + (z-1)q'(z)$. Substituting $z = 1$, we have $0 = q(1)$, as we claimed.) Explicitly, there is some r on the unit circle such that $p(z) = (z-1)^2(z-r)$. Hence

$$p'(z) = 3(z-1)\left(z - \frac{2r+1}{3}\right).$$

We conclude that $p \in \Gamma$ has a critical point at 1 if and only if $p(z) = (z-1)^2(z-r)$ for some r on the unit circle, and in this case the other critical point is $\frac{1}{3} + \frac{2}{3}r \in T_{4/3}$.

We have considered this example in detail because, in many of our later proofs, when we consider a critical point c, technical issues will force us to assume that $c \neq 1$.

EXAMPLE B. Suppose now that $p \in \Gamma$ has a critical point at some $c \neq 1$ on the unit circle. By the Gauss–Lucas theorem, c is in the convex hull of the roots of p, and so c must be a root of p. It follows that p has a double root at c, so $p(z) = (z-c)^2(z-1)$ and

$$p'(z) = 3(z-c)\left(z - \frac{c+2}{3}\right).$$

To summarize, $p \in \Gamma$ has a critical point at $c \neq 1$ on the unit circle if and only if $p(z) = (z-c)^2(z-1)$, and in this case the other critical point is $\frac{2}{3} + \frac{1}{3}c \in T_{2/3}$.

Observe that each $z \neq 1$ in the unit disk lies on a unique T_α with $0 < \alpha \leq 2$. For what follows, it will be very useful to know how to find the value of α corresponding to a given z.

THEOREM 2. *Let $z \in \mathbb{C}$ with $\mathrm{Re}(z) < 1$. We have $z \in T_\alpha$ if and only if*

$$\frac{1}{\alpha} = \mathrm{Re}\left(\frac{1}{1-z}\right).$$

Proof. Name points as in FIGURE 2, and let θ denote the measure of $\angle BAC$, taking $\theta > 0 \iff \mathrm{Im}(z) > 0$. Set $r = |z-1|$, so that

$$\mathrm{Re}\left(\frac{1}{1-z}\right) = \mathrm{Re}\left(\frac{1}{r}e^{i\theta}\right) = \frac{\cos\theta}{r}.$$

Then $\angle ACB$ intercepts a diameter of T_α, and must therefore be a right angle. By right-triangle trigonometry, $\cos\theta = r/\alpha$, completing the proof. ∎

We are now ready to prove our main theorem, a symmetry result for the critical points of p. We prove a more general statement than we need, because it is no more difficult to do so.

THEOREM 3. *Let $f(z) = (z-1)(z-z_1)\cdots(z-z_n)$, where $|z_k| = 1$ for each k. Let $c_1, \ldots, c_n$ denote the critical points of $f(z)$, and suppose that $1 \neq c_k \in T_{\alpha_k}$ for each k. Then*

$$\sum_{k=1}^{n} \frac{1}{1-c_k} = 2\sum_{k=1}^{n} \frac{1}{1-z_k},$$

and

$$\sum_{k=1}^{n} \frac{1}{\alpha_k} = n. \tag{1}$$

Arithmetically, equation (1) says that the "harmonic mean" of the α_k is 1.

Proof. We evaluate $\operatorname{Re}(f''(1)/f'(1))$ in two different ways. To begin with, $f'(z) = (n+1)\prod_{k=1}^{n}(z-c_k)$. By logarithmic differentiation,

$$\frac{f''(1)}{f'(1)} = \frac{f''(z)}{f'(z)}\Big|_{z=1} = \sum_{k=1}^{n} \frac{1}{1-c_k},$$

and Theorem 2 implies

$$\operatorname{Re}\left(\frac{f''(1)}{f'(1)}\right) = \operatorname{Re}\left(\sum_{k=1}^{n} \frac{1}{1-c_k}\right) = \sum_{k=1}^{n} \frac{1}{\alpha_k}.$$

On the other hand, if we write $f(z) = (z-1)g(z)$, then the product rule and logarithmic differentiation yield

$$\frac{f''(1)}{f'(1)} = \frac{2g'(1)}{g(1)} = 2\sum_{k=1}^{n} \frac{1}{1-z_k}.$$

But since $|z_k| = 1$, we have $z_k \in T_2$, and we can apply Theorem 2 to obtain

$$\operatorname{Re}\left(\frac{f''(1)}{f'(1)}\right) = 2\sum_{k=1}^{n} \operatorname{Re}\left(\frac{1}{1-z_k}\right) = 2\sum_{k=1}^{n} \frac{1}{2} = n. \qquad \blacksquare$$

COROLLARY. *Let $p \in \Gamma$, and let $c_1 \neq 1$ and $c_2 \neq 1$ be the critical points of p. If c_1 lies on T_α and c_2 lies on T_β, then*

$$\frac{1}{\alpha} + \frac{1}{\beta} = 2. \tag{2}$$

We claim that the corollary has a clear geometric meaning: T_β is the inversion of the circle T_α across T_1 (FIGURE 3). Recall (see [**13**], for example) that if C is a circle centered at O and X is a point distinct from O, then the *inversion* of X over C is the point Y on the ray $\overrightarrow{OX}$ such that $|OX| \cdot |OY| = r^2$: The product of the distances from O to X and Y equals the square of the radius of C.

To prove the claim, suppose that $1 \neq c_1 \in T_\alpha$ and $1 \neq c_2 \in T_\beta$. Then, since $\beta > 0$, we know that $\alpha > \frac{1}{2}$. Since T_1 and T_α are symmetric across the real axis and pass

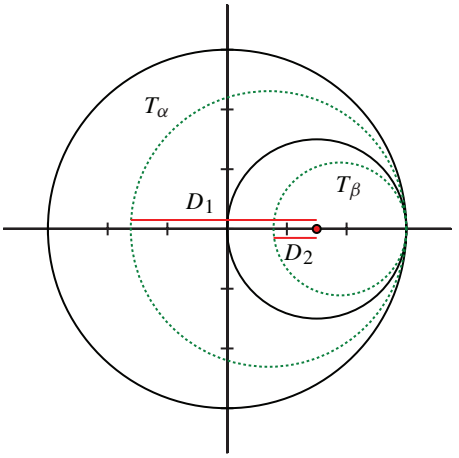


Figure 3 T_β is the inversion of T_α across T_1.

through 1, the same is true of the inversion of T_α across T_1. This being the case, it is enough to show that the product of the distances D_1 and D_2 to $1/2$ from $1-\alpha$ and $1-\beta$ equals $1/4$, the square of the radius of T_1. In fact, this is true, because (2) gives $\beta = \alpha/(2\alpha-1)$, and then

$$\left|(1-\alpha)-\frac{1}{2}\right| \cdot \left|(1-\beta)-\frac{1}{2}\right| = \left|\frac{2\alpha-1}{2} \cdot \frac{1}{2(2\alpha-1)}\right| = \frac{1}{4}.$$

We can generalize our results to give information about a non-normalized cubic polynomial. Given $\triangle ABC$ in the complex plane, let T_α and T_β be the circles tangent at A to the circumcircle of $\triangle ABC$ that pass through the foci of the Steiner ellipse. The radius of the circumcircle then equals the harmonic mean of the diameters of T_α and T_β.

We think the corollary is intrinsically attractive, but even better, it is also useful! First, it allows us to prove our original observation: There is a "desert" in the unit disk, the open disk $T_{2/3} = \left\{z \in \mathbb{C} : \left|z - \frac{2}{3}\right| < \frac{1}{3}\right\}$, in which critical points cannot occur.

THEOREM 4. *No polynomial $p \in \Gamma$ has a critical point strictly inside $T_{2/3}$* (FIGURE 4).

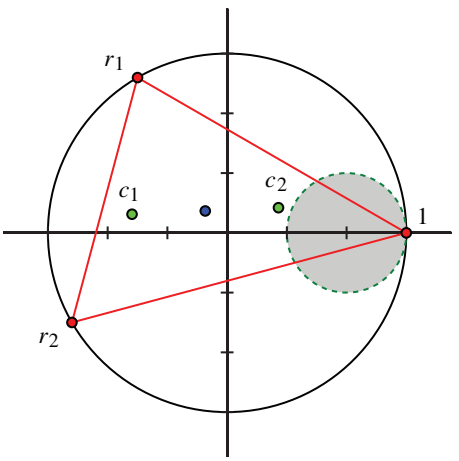


Figure 4 No $p \in \Gamma$ has a critical point in the desert (Theorem 4).

Proof. If c is strictly inside $T_{2/3}$, then c lies on T_α for some $\alpha \in \left(0, \frac{2}{3}\right)$. Suppose for contradiction that c is a critical point of some polynomial $p \in \Gamma$. Then the other critical point of p lies on T_β, where (by our corollary) $\frac{1}{\beta} = 2 - \frac{1}{\alpha} < 2 - \frac{3}{2} = \frac{1}{2}$. But then $\beta > 2$, which is impossible, by the Gauss–Lucas theorem. ∎

Recall that Saff and Twomey [**10**] had shown that every $p \in \Gamma$ has at least one critical point on or inside T_1. Our corollary lets us say a bit more.

THEOREM 5. *Let $c_1 \neq 1$ and $c_2 \neq 1$ be the critical points of $p \in \Gamma$. If c_1 lies on T_1, then c_2 also lies on T_1. Otherwise, c_1 and c_2 are on opposite sides of T_1.*

Proof. Let $c_1 \in T_\alpha$ and $c_2 \in T_\beta$. Then $\frac{1}{\alpha} + \frac{1}{\beta} = 2$, so $\alpha = 1$ if and only if $\beta = 1$ and $\alpha < 1$ if and only if $\beta > 1$. ∎

A critical point determines $p \in \Gamma$ (almost always)

Once again, suppose $p \in \Gamma$, with roots r_1, r_2, and 1, and let c be a critical point of p. Then

$$\begin{aligned} p(z) &= (z - r_1)(z - r_2)(z - 1) \\ &= z^3 - (r_1 + r_2 + 1)z^2 + (r_1 + r_2 + r_1 r_2)z - r_1 r_2 \end{aligned}$$

so that

$$0 = p'(c) = 3c^2 - 2c(r_1 + r_2 + 1) + (r_1 + r_2 + r_1 r_2). \tag{3}$$

Assuming that $r_1 \neq 2c - 1$, we get

$$r_2 = \frac{(2c - 1)r_1 + (2c - 3c^2)}{r_1 + (1 - 2c)}, \tag{4}$$

which gives r_2 as a function of r_1.

DEFINITION. Given $c \in \mathbb{C}$, we define

$$f_c(z) = \frac{(2c - 1)z + (2c - 3c^2)}{z + (1 - 2c)}.$$

We let S_c denote the image of the unit circle under f_c.

THEOREM 6. *Suppose $p(z) = (z - r_1)(z - r_2)(z - 1) \in \Gamma$ and $1 \neq c \in \mathbb{C}$. Then p has a critical point at c if and only if $f_c(r_1) = r_2$.*

When $c \neq 1$, we have $(f_c)^{-1} = f_c$. If $c = 1$, we have $f_1(z) = \frac{z-1}{z-1} = 1$ when $z \neq 1$, and $(f_1)^{-1}$ does not exist. Hence, we suppose $c \neq 1$ in what follows.

Observe that f_c is a linear fractional transformation, that is, a function of the form $g(z) = \frac{Az+B}{Cz+D}$. It is well known [**14**] that any invertible linear fractional transformation maps circles (and lines) to circles (and lines), so S_c is a circle—or a line when there is some $z \in T_2$ for which the denominator $z + (1 - 2c) = 0$. Recall that T_2 is the unit circle, so S_c is a line if and only if

$$|1 - 2c| = 1 \iff \left|\frac{1}{2} - c\right| = \frac{1}{2} \iff c \in T_1. \tag{5}$$

Let us pause to study an important special case.

EXAMPLE C. Suppose that $c \neq \pm 1$ but $c \in T_2$. We already know that S_c is a circle. Direct calculations yield

$$f_c(c) = c, \quad f_c(1) = \frac{3c-1}{2} \text{ and } f_c(-1) = \frac{3c^2-1}{2c}.$$

Therefore, for $z \in \{1, -1, c\}$,

$$\left| f_c(z) - \frac{3}{2}c \right| = \frac{1}{2},$$

so S_c is the circle of radius $1/2$ centered at $3c/2$, which is externally tangent to T_2 at c.

We saw in Example B that if $p \in \Gamma$ has a critical point at c, then the other critical point of p is $(2+c)/3$, and an argument just like the last one shows that $S_{(2+c)/3}$ is the circle of radius $1/2$ centered at $c/2$, which is internally tangent to T_2 at c.

Now, f_c maps T_2 onto S_c, and since $(f_c)^{-1} = f_c$, f_c also maps S_c onto T_2. Hence, f_c restricts to a one-to-one correspondence from $S_c \cap T_2$ to itself, and if c is a critical point of p, then we have $\{r_1, r_2\} \subseteq S_c \cap T_2$. We can use this fact to classify the polynomials $p \in \Gamma$ having a critical point at any given $c \neq 1$ in the closed unit disk.

Because S_c and T_2 are circles, there are four cases to consider:

1. S_c and T_2 are disjoint;
2. S_c and T_2 are tangent;
3. S_c and T_2 intersect in two distinct points;
4. $S_c = T_2$.

In the first case, there can be no $p \in \Gamma$ with a critical point at c, because no point in $\mathbb{C}$ is eligible to be r_1 (or r_2). In the second case, if $S_c \cap T_2 = \{r\}$, then it is necessary that $r_1 = r = r_2$ and $p(z) = (z-1)(z-r)^2$. Conversely, when p is of this type, we have seen in Examples B and C that $S_c \cap T_2 = \{r\}$.

In the third case, we assume that $S_c \cap T_2 = \{a, b\}$ for some pair $a \neq b$. Suppose for the sake of contradiction that $f_c(a) = a$. Since f_c is a permutation of $S_c \cap T_2$, we have $f_c(b) = b$. By Theorem 6, c is a critical point of $p_a(z) = (z-a)^2(z-1)$ and of $p_b(z) = (z-b)^2(z-1)$. We showed in Example C that

$$c \in \{a, (2+a)/3\} \cap \{b, (2+b)/3\}.$$

However, $(2+a)/3$ is on the segment from 1 to a, and $(2+b)/3$ is on the segment from 1 to b; since $a \neq b$, $c \in \{a, (2+a)/3\} \cap \{b, (2+b)/3\} = \phi$, a contradiction. It follows that $f_c(a) = b$ and $f_c(b) = a$, and so $p(z) = (z-1)(z-a)(z-b)$ is the only polynomial with a critical point at c.

To handle the last case, suppose that $S_c = T_2$. In this case, whenever $|r| = 1$, we have $|f_c(r)| = 1$, in particular, $|f_c(1)| = 1 = |f_c(-1)|$. On writing $c = x + iy$, these equations become:

$$y^2 = \frac{4}{9} - \left(x - \frac{1}{3}\right)^2$$

$$4(x^2+y^2) = 9(x^2+y^2)^2 - 6(x^2-y^2) + 1.$$

By elimination, we eventually obtain

$$0 = 3x^2 - 2x - 1 = (3x+1)(x-1).$$

Since $c \neq 1$ by assumption, $x = -1/3$, and by back-substitution $y = 0$. That is, if $S_c = T_2$, then $c = -1/3$.

Conversely, we claim that $S_{-1/3} = T_2$. Clearing fractions, $f_{-1/3}(z) = \frac{-5z-3}{3z+5}$, and a routine calculation shows that $|f_{-1/3}(i)| = 1$. Now we know that $S_{-1/3}$ is a circle containing $f_{-1/3}(1)$, $f_{-1/3}(-1)$, and $f_{-1/3}(i)$, each of which has modulus 1. This implies the claim.

It follows that if $c = -1/3$ is a critical point of p, then there is some $r \neq 1$ on the unit circle for which $p(z) = (z-1)(z-r)\left(z - f_{-1/3}(r)\right)$. On the other hand, it is easy to check that every such polynomial has a critical point at $c = -1/3$.

It remains to show that if $c \notin \{1, -1/3\}$ lies on T_α for some $\alpha \in (2/3, 2)$, then $|S_c \cap T_2| = 2$, so that there exists some (necessarily unique) $p \in \Gamma$ with a critical point at c. Suppose to the contrary that no such $p \in \Gamma$ exists. Then S_c and T_2 are disjoint. So S_c lies either entirely inside or entirely outside the circle T_2. Assume that S_c lies inside T_2; the other case is similar.

We will study how S_c changes as we "drag" $c \neq 1$ through $X := \{z : z \in T_\alpha,\ \alpha \in (2/3, 2]\}$. Define a (continuous) path $\gamma : [0, 1] \to X$ with $\gamma(0) = c$ and $\gamma(1) \in T_2$, such that $\gamma(t)$ lies in the interior of X when $t < 1$. Since $S_{\gamma(1)}$ is externally tangent to T_2 (by Example C) and $S_{\gamma(0)}$ lies inside T_2, there must be a $t_0 \in (0, 1)$ where $S_{\gamma(t_0)}$ is internally tangent to T_2. But then $\gamma(t_0) \in T_{2/3}$, a contradiction.

We have proved the following theorem.

THEOREM 7. *Let $c \in \mathbb{C}$.*

1. *If $c \notin \{1, -1/3\}$ lies on T_α for some $\alpha \in [2/3, 2]$, then there is a unique $p \in \Gamma$ with a critical point at c.*
2. *If c lies strictly inside $T_{2/3}$, or strictly outside T_2, then there is no $p \in \Gamma$ with a critical point at c.*
3. *$p \in \Gamma$ has a critical point at 1 if and only if $p(z) = (z-1)^2(z-r)$ for some r on the unit circle.*
4. *$p \in \Gamma$ has a critical point at $-1/3$ if and only if $p(z) = (z-1)(z-r)\left(z + \frac{5r+3}{3r+5}\right)$ for some r on the unit circle.*

As an application of Theorem 7, we give an independent proof of Saff and Twomey's result on conjugate critical points; see FIGURE 5.

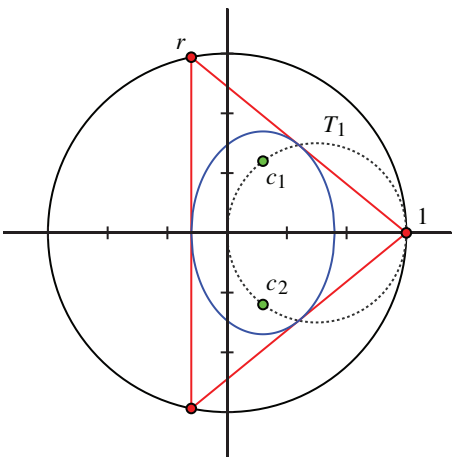


Figure 5 If $1 \neq c_1 \in T_1$, then $c_2 = \overline{c_1}$.

COROLLARY. *Let c_1 and c_2 be the critical points of $p \in \Gamma$. If $1 \neq c_1 \in T_1$, then $c_2 = \overline{c_1}$.*

Proof. Given $c_1 \in T_1$, we have $\frac{1}{1-c_1} = 1 + yi$ by Theorem 2, so $\frac{1}{c_1} = 1 - \left(\frac{1}{y}\right) i$. Then, from (5) and Example C, we know that S_{c_1} is the line through $f_{c_1}(1) = \frac{3}{2}c_1 - \frac{1}{2}$ and

$$f_{c_1}(-1) = \frac{3}{2}c_1 - \frac{1}{2c_1} = \frac{3}{2}c_1 - \frac{1}{2} - \left(\frac{1}{2y}\right) i.$$

These differ by a purely imaginary number, so S_{c_1} is vertical. Thus $p(z) = (z-1)(z-r)(z-\overline{r})$ for some $r \in T_2$ whose real part is $\frac{3}{2}\mathrm{Re}(c_1) - \frac{1}{2}$. We can see from (3) that $c_1 + c_2 = \frac{2}{3}(1 + r + \overline{r}) = 2\mathrm{Re}(c_1)$, and subtraction yields $c_2 = \overline{c_1}$. ∎

Constructions and conclusions

We began this article by thinking of a cubic polynomial as a triangle, inscribing it in a circle, and choosing coordinates in such a way that the polynomial belonged to Γ. We conclude by returning to this general context: We emphasize the geometric content of our earlier results by stating a pair of geometric theorems, which assert the existence of a variety of compass-and-straightedge constructions.

THEOREM 8. *The triangle $\triangle ABC$ can be constructed given a vertex and any two of the following five points: the other vertices, the centroid, and the foci of the Steiner ellipse.*

THEOREM 9. *The triangle $\triangle ABC$ can be constructed given the circumcircle, a vertex, and either the centroid or a focus of the Steiner ellipse.*

We prove these theorems one case at a time. The first four constructions prove Theorem 8. Recall (see [**12**]) that it is possible to construct sums, differences, products, quotients, and square roots of known points in the complex plane.

CONSTRUCTION. *$\triangle ABC$, given vertex A and the foci F_1 and F_2 of the Steiner ellipse.*

Proof. We work backward using Marden's theorem and the quadratic formula. Treating A, F_1, and F_2 as points in the complex plane, construct $\sigma_1 = A - \frac{3}{2}(F_1 + F_2)$ and $\sigma_2 = r_1\sigma_1 + 3F_1F_2$. Then construct B and C at $(-\sigma_1 \pm \sqrt{\sigma_1^2 - 4\sigma_2})/2$. ∎

CONSTRUCTION. *$\triangle ABC$, given vertex A, centroid G, and a focus F_1 of the Steiner ellipse.*

Proof. Let F_2 be the reflection of F_1 over G. Apply the previous construction, using F_1 and F_2 to find B and C. ∎

CONSTRUCTION. *$\triangle ABC$, given vertex A, vertex B, and centroid G.*

Proof. Construct $C = 3G - A - B$. ∎

CONSTRUCTION. *$\triangle ABC$, given vertex A, vertex B, and a focus F of the Steiner ellipse.*

Proof. As in (4), construct $C = \frac{3F^2 - 2F(A+B) + AB}{2F - A - B}$. ∎

This completes the proof of Theorem 8. The remaining constructions prove Theorem 9.

CONSTRUCTION. *$\triangle ABC$, given vertex A, circumcircle S, and centroid G.*

Proof. We proceed as in Theorem 1. Construct the point O at the center of S. Draw line L through A and G, and let M be the midpoint of $\overline{AG}$. With center G, draw a circle S' of radius AM. Let S' intersect L at the point $X \neq M$. Draw line $\overline{XO}$ and construct the line L' perpendicular to $\overline{XO}$ through X. Then L' intersects S at the points B and C. ∎

To complete the proof of Theorem 9, we need to show that we can construct $\triangle ABC$ given a vertex, circumcircle, and a focus of the Steiner ellipse. To understand the main idea of the construction, let us recall how we showed that a critical point almost always determines $p \in \Gamma$. With this in mind, assume that $A = 1$, $S = T_2$ is the circumcircle and that $F \notin \{1, 1/3\}$ is a focus of the Steiner ellipse. If X is any point on S other than $A = 1$, use a known construction to construct a point X' such that $\triangle AXX'$ has a Steiner ellipse with focus at F. We cannot guarantee that X' lies on S, so this may not be the desired triangle. However, as we let X trace out the circle S, X' traces out the circle S_F. Therefore, since $F \notin \{1, 1/3\}$ is the focus of a Steiner ellipse, $S \cap S_F = \{B, C\}$.

CONSTRUCTION. *$\triangle ABC$, given vertex A, circumcircle S, and a focus F of the Steiner ellipse.*

Proof. Let distinct points X, Y, and Z, different from A, be given on the circle S. As above, construct points X', Y', and Z' such that $\triangle AXX'$, $\triangle AYY'$, and $\triangle AZZ'$ each have a Steiner ellipse with a focus at F. Construct S' to be the circumcircle of $\triangle X'Y'Z'$. Then S' and S intersect at the points B and C. ∎

This completes our last proof, but as always, some questions remain unanswered. It would be especially nice to learn more about polynomials of higher degree. Preliminary results suggest that some subset of the polynomials of the form

$$p(z) = (z-1)^j (z-r_1)^k (z-r_2)^\ell,$$

with $\{r_1, r_2\} \subseteq T_2$ and $\{j, k, \ell\} \subseteq \mathbb{N}$, should be amenable to the same type of analysis. For example, if $p(z) = (z-1)(z-r_1)^k(z-r_2)^\ell$, we have the following critical points:

- $c_1 = r_1$ with multiplicity $k-1$;
- $c_2 = r_2$ with multiplicity $\ell - 1$;
- two non-trivial critical points, $c_3 \in T_\alpha$ and $c_4 \in T_\beta$.

The analogue of (2) in this case is

$$\frac{1}{\alpha} + \frac{1}{\beta} = 1 + \frac{k+\ell}{2}.$$

We are sure that much more is waiting to be discovered.

Acknowledgment We are grateful to Ryan Knuesel for his work on the software that first helped us visualize the trajectories of the critical points of polynomials $p \in \Gamma$.

REFERENCES

1. Bruce Anderson, Polynomial root dragging, *Amer. Math. Monthly* **100** (1993) 864–866, http://dx.doi.org/10.2307/2324665.

2. Matthew Boelkins, Justin From, and Samuel Kolins, Polynomial root squeezing, *Math. Mag.* **81** (2008) 39–44, http://dx.doi.org/10.2307/27643078.
3. Christopher Frayer, More polynomial root squeezing, *Math. Mag.* **83** (2010) 218–222.
4. ———, Squeezing polynomial roots a nonuniform distance, *Missouri J. Math. Sci.* **22** (2010) 124–129.
5. Christopher Frayer and James A. Swenson, Polynomial root motion, *Amer. Math. Monthly* **117** (2010) 641–646, http://dx.doi.org/10.4169/000298910X496778.
6. Pamela Gorkin and Elizabeth Skubak, Polynomials, ellipses, and matrices: two questions, one answer, *Amer. Math. Monthly* **118** (2011) 522–533, http://dx.doi.org/10.4169/amer.math.monthly.118.06.522.
7. Dan Kalman, An elementary proof of Marden's theorem, *Amer. Math. Monthly* **115** (2008) 330–338.
8. Morris Marden, *Geometry of Polynomials*, 2nd ed., American Mathematical Society, Providence, RI, 1966.
9. ———, A note on the zeros of the sections of a partial fraction, *Bull. Amer. Math. Soc.* **51** (1945) 935–940.
10. E. B. Saff and J. B. Twomey, A note on the location of critical points of polynomials, *Proc. Amer. Math. Soc.* **27** (1971) 303–308.
11. Jörg Siebeck, Uber eine neue analytische Behandlungweise der Brennpunkte, *Journal f ur die riene und angewandte Mathematik* **64** (1864) 175–182.
12. Eric W. Weisstein, "Constructible number," from MathWorld—A Wolfram Web Resource. http://mathworld.wolfram.com/ConstructibleNumber.html
13. ———, "Inversion," from MathWorld—A Wolfram Web Resource. http://mathworld.wolfram.com/Inversion.html
14. ———, "Linear fractional transformation," from MathWorld—A Wolfram Web Resource. http://mathworld.wolfram.com/LinearFractionalTransformation.html

Summary We study the critical points of a complex cubic polynomial, normalized to have the form $p(z) = (z-1)(z-r_1)(z-r_2)$ with $|r_1| = 1 = |r_2|$. If T_γ denotes the circle of diameter γ passing through 1 and $1-\gamma$, then there are $\alpha, \beta \in [0, 2]$ such that one critical point of p lies on T_α and the other on T_β. We show that T_β is the inversion of T_α over T_1, from which many geometric consequences can be drawn. For example, (1) a critical point of such a polynomial almost always determines the polynomial uniquely, and (2) there is a "desert" in the unit disk, the open disk $\{z \in \mathbb{C} : |z - \frac{2}{3}| < \frac{1}{3}\}$, in which critical points cannot occur.

CHRIS FRAYER received his B.S. from Grand Valley State University and an M.S. and Ph.D. from the University of Kentucky. He is an associate professor of mathematics at University of Wisconsin-Platteville. His mathematical interests include geometry of polynomials, knot theory, and differential equations. In his free time, he enjoys spending time with his family, rock climbing, and running long distances.

MIYEON KWON received her Ph.D. from the University of Alabama in 2004. She is currently an associate professor of mathematics at the University of Wisconsin-Platteville.

JAMES A. SWENSON is an associate professor of mathematics at the University of Wisconsin-Platteville. He attended Augustana University in South Dakota, and earned his Ph.D. from the University of Minnesota. In his down time, he most enjoys being with his family; he also loves fiction and choral music. Despite living in Wisconsin, he is an avid fan of the Minnesota Twins, Vikings, Timberwolves, Lynx, Wild, and Gophers.

CHRISTOPHER SCHAFHAUSER received his B.S. from the University of Wisconsin-Platteville and is currently in graduate school at the University of Nebraska-Lincoln.

NOTES

A Reed–Solomon Code Magic Trick

TODD D. MATEER
Howard Community College
Columbia, Maryland 21044
tmateer@howardcc.edu

Richard Ehrenborg [**1**] has provided a nice magic trick that can be used to illustrate many properties of Hamming codes. His paper includes a set of manipulatives that can be used to implement the trick. An improved version of this Hamming code magic trick is described in a recent paper in *Math Horizons* [**3**].

In this paper, we introduce a similar magic trick that is based on a Reed–Solomon code. It can be used to introduce Reed–Solomon codes to engineering students, and yet is simple enough to use with high school students. The ideas behind the magic trick are also the basis for the error correction technique used on CDs and DVDs.

The magic trick

The magic trick uses the four cards shown in FIGURE 1.

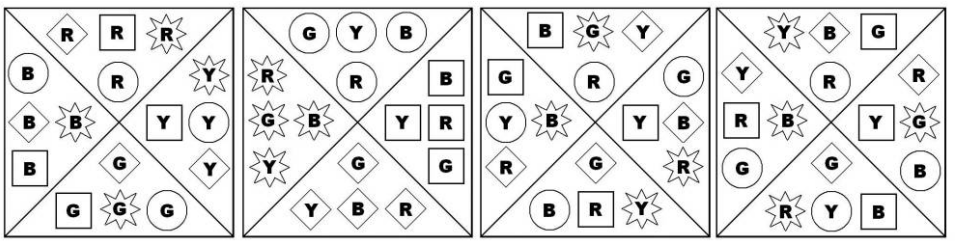


Figure 1 Cards used for magic trick

On each card are sixteen different symbols: four shapes, each in four different colors. In this paper, we use the colors Red (R), Yellow (Y), Green (G), and Blue (B) along with the shapes Circle, Square, Diamond, and Star. Any colors and any shapes with the proper symmetry would serve as well. We have put the first letter of the color in the center of each symbol as an extra cue.

A volunteer is selected from the audience and asked to silently pick a color from {Red, Yellow, Blue, Green} and a shape from {Circle, Star, Diamond, Square}. The magician shows the volunteer the four cards, and instructs the volunteer to report the quadrant (Up, Left, Right, Down) where the chosen symbol is found on each card. However, the volunteer is also instructed to lie—that is, give an incorrect response—on one of the cards. Following each response, the magician inconspicuously rotates the

Math. Mag. **87** (2014) 125–131. doi:10.4169/math.mag.87.2.125.

card so that the chosen section is in the Up position. As soon as the volunteer responds to the fourth card, the magician tells the volunteer which symbol was selected as well as which card was reported incorrectly.

The secret

The cards are designed so that when they are rotated according to the above instructions, the chosen symbol appears in the same position within the upper section on three of the four cards. None of the remaining symbols will have this property. The magician can then easily tell which symbol was chosen. For the card on which the volunteer lied, the chosen symbol does not appear in its proper position, allowing the magician to recognize the card that was reported incorrectly.

For example, suppose that the volunteer responded "Up Down Up Right" to the cards from FIGURE 1. After the four cards have been rotated, they will appear as shown in FIGURE 2. The magician sees that the Red Diamond is in the upper left position in the top quadrant of the first, second, and fourth cards. So the magician declares that the chosen symbol was the Red Diamond. Since this symbol does not appear in this position on the third card, the magician declares that the third card was reported incorrectly.

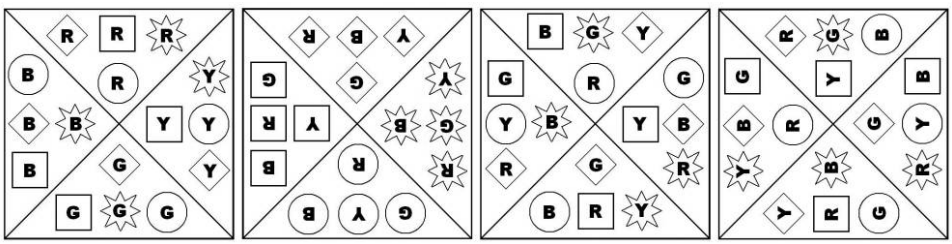


Figure 2 Example of the magic trick

There is nothing special about the Up position in this trick; the magician can rotate the cards in some other direction if desired for variety.

Sometimes, the magician can identify the chosen symbol after the third card and also knows that the volunteer will lie on the fourth card. In this case, the magician can wait for the fourth response or can say "You are about to lie to me, aren't you?" before the fourth card is shown.

The trick works no matter which symbol is chosen and which card is reported incorrectly. The magic is based on properties of a particular Reed–Solomon code. In the remainder of this paper we introduce the Reed–Solomon code and explain why the trick works.

Error-correcting codes

The magic trick is based on one of the simplest Reed–Solomon codes.

On the first two cards, the volunteer can choose any of four possible sections (Up, Left, Right, Down). If the person does not lie on the first two cards, the responses to these two cards uniquely determine the chosen symbol. Let us assign the symbol 0 to represent the upper quadrant, 1 to represent the right quadrant, A to represent the lower

quadrant, and B to represent the left quadrant. (These choices will be important later.) We can now represent each symbol using a *codeword* determined by the location of the symbol on the first two cards. For example, if the cards are presented according to FIGURE 1, then the red diamond would be represented using the codeword [0, A].

Suppose that the third card is added and the person responds honestly on each of the three cards. Our code could be modified to use all three replies as given in FIGURE 3. Each three-component vector is now a codeword of the new three-letter code. While each codeword has three letters, only the first two letters are really needed to represent each symbol.

Symbol	Codeword	Symbol	Codeword
R◯	UUU = [0 , 0 , 0]	G◯	DUR = [A , 0 , 1]
R■	URD = [0 , 1 , A]	G■	DRL = [A , 1 , B]
R◇	UDL = [0 , A , B]	G◇	DDD = [A , A , A]
R★	ULR = [0 , B , 1]	G★	DLU = [A , B , 0]
Y◯	RUL = [1 , 0 , B]	B◯	LUD = [B , 0 , A]
Y■	RRR = [1 , 1 , 1]	B■	LRU = [B , 1 , 0]
Y◇	RDU = [1 , A , 0]	B◇	LDR = [B , A , 1]
Y★	RLD = [1 , B , A]	B★	LLL = [B , B , B]

Figure 3 A (3, 2, 2) Reed–Solomon code based on the magic trick

In general, a code such as this one can be represented using the two parameters (n, k). Here, n is the total number of letters in each codeword and k is the number of *information components*, that is, the number of letters actually needed to identify the symbol. In the case of the magic trick, the codewords in FIGURE 3 represent a (3, 2) code, since each codeword consists of 3 letters and the first 2 of these components identify the chosen symbol (the color and the shape). A popular set of "real-world" Reed–Solomon code parameters is (255, 223), where 223 components of information are encoded into each 255-letter codeword.

During transmission, a codeword can become distorted. In the case of the magic trick, the codeword is distorted by the one lie of the volunteer. In the case of the CD or DVD, the distortions are in the form of scratches or fingerprints. The extra $n - k$ components added to the k components of information can be used to correct a limited number of these mistakes. For this reason, these codes are called *error-correcting codes*. It is the job of a decoder to determine which codeword best represents the received set of letters after transmission and possible distortion. The best representation is typically the codeword that differs from what was received in the minimum number of positions.

If we study the codewords in FIGURE 3, we see that each codeword differs from any other codeword in at least 2 positions. We call this the *minimum distance* of the code and use the parameter d to represent it. The minimum distance is frequently given as a third parameter of an error-correcting code. So the code of FIGURE 3 can be called a (3, 2, 2) code. It can be shown [**4**] that a code with minimum distance d can correct up to $\lfloor (d-1)/2 \rfloor$ errors. In the case of the magic trick, this tells us that the (3, 2, 2) code constructed using the first three cards cannot correct any errors.

To see this, suppose that the volunteer lied on one of the first three cards. For example, suppose that the responses of the volunteer were "Left, Up, Right". Using the quadrant symbol assignments mentioned earlier, these responses can be represented as [B, 0, 1]. Observe in FIGURE 3 that there are three codewords that differ from [B, 0, 1] in one position ([A, 0, 1], [B, 0, A], and [B, A, 1]). Since each of

$\{[A, 0, 1], [B, 0, A], [B, A, 1]\}$ differs from any of the other two codewords in this set in exactly two positions, the vector $[B, 0, 1]$ can be considered to be exactly halfway between any two of these codewords. Without any further information, we cannot determine which of these three possible codewords is closest to $[B, 0, 1]$.

Although the $(3, 2, 2)$ code cannot correct any errors, it is capable of detecting that at least one component of $[B, 0, 1]$ was incorrect. Since the code has this capability, it is considered to be an example of an error-detecting code.

Because the $(3, 2, 2)$ code was not sufficient to make the magic trick work, a fourth card was added. Each symbol is now encoded using the code as given in FIGURE 4. It is still true that any two components suffice to identify the symbol. The reader can verify from FIGURE 4 that each codeword differs from any other codeword in at least 3 positions. Thus, this code can be expressed using the parameters $(4, 2, 3)$. Since an error-correcting code can correct up to $\lfloor (d-1)/2 \rfloor$ errors, this code is capable of correcting one error.

Symbol	Codeword	Symbol	Codeword
R◯	UUUU = [0 , 0 , 0 , 0]	G◯	DURL = [A , 0 , 1 , B]
R■	URDL = [0 , 1 , A , B]	G■	DRLU = [A , 1 , B , 0]
R◇	UDLR = [0 , A , B , 1]	G◇	DDDD = [A , A , A , A]
R★	ULRD = [0 , B , 1 , A]	G★	DLUR = [A , B , 0 , 1]
Y◯	RULD = [1 , 0 , B , A]	B◯	LUDR = [B , 0 , A , 1]
Y■	RRRR = [1 , 1 , 1 , 1]	B■	LRUD = [B , 1 , 0 , A]
Y◇	RDUL = [1 , A , 0 , B]	B◇	LDRU = [B , A , 1 , 0]
Y★	RLDU = [1 , B , A , 0]	B★	LLLL = [B , B , B , B]

Figure 4 A $(4, 2, 3)$ Reed–Solomon code based on the magic trick

For example, considering the earlier example of "Up Down Up Right", $[0, A, 0, 1]$, for the four responses, now only one of the 16 possible codewords, $[0, A, B, 1]$, differs from $[0, A, 0, 1]$ in at most one position. As long as the volunteer provided at most one incorrect answer to the four cards, the responses will always differ from a valid codeword in at most one position.

In general, the problem of decoding a codeword is not easy. We need to determine both the location of the error(s) added to the codeword, as well as the correct value at each error location. In the simplified case of this magic trick, we can step through the four cards, ignoring each one in its turn, to see if the same symbol appears in the upper section of the remaining three cards. For those familiar with error-correcting codes, we are essentially "erasing" each of the four components in the codeword, one at a time. The remaining three cards form a $(3, 2, 2)$ code. Recall that this code cannot correct any errors, but it can detect whether or not an error occurred. If we erase the correct card, then the remaining three cards should decode to a valid codeword (i.e., one of the symbols should appear in the upper section of the three remaining cards). If we erase the wrong card, then the $(3, 2)$ code will detect the error and we should try a different erasure. Since there are only four choices for the incorrect card, this erasure decoding procedure finds the incorrect card efficiently. It should be noted, however, that this technique is not efficient for larger codes.

Reed–Solomon codes

The cards used in the trick are based on a particular finite field—that is, an algebraic system with finitely many elements, together with operations that satisfy the usual

axioms for addition and multiplication. A finite field is sometimes called a *Galois field* because the research of Èvariste Galois was influential in this branch of mathematics.

Let us consider the set of four elements $\{0, 1, A, B\}$ and define the operations $\oplus$ and $\odot$ according to the arithmetic tables given in FIGURE 5. This is a finite field. For the rest of this paper, we will use the symbol $\mathbb{F}_4$ to represent the finite field with elements $\{0, 1, A, B\}$ and operations $\oplus, \odot$.

$\oplus$	0	1	A	B
0	0	1	A	B
1	1	0	B	A
A	A	B	0	1
B	B	A	1	0

$\odot$	0	1	A	B
0	0	0	0	0
1	0	1	A	B
A	0	A	B	1
B	0	B	1	A

Figure 5 Arithmetic tables of $\mathbb{F}_4$

We will now say what we mean by a *Reed–Solomon Code*. In practice, a code is defined by how we select its codewords, how we encode content into the codewords, and how we decode the codewords to recover the content.

Every step depends on representing vectors as polynomials. Given any vector with n entries, we associate it with a polynomial (of degree at most $n-1$) whose coefficients are the entries of the vector. For example, if $[0, A, B]$ is a vector whose entries are elements of $\mathbb{F}_4$, we associate it with the polynomial $0x^2 + Ax + B$.

To construct a Reed–Solomon Code with parameters n and k, we first select a finite field, and then we select a particular *generating polynomial* of degree $n-k$ with coefficients from this finite field. In the case of the three-letter $(3, 2, 2)$ code used above, we use $n = 3$, $k = 2$, the finite field $\mathbb{F}_4$, and the generating polynomial

$$g(x) = x - A.$$

For a more general Reed–Solomon Code, the generating polynomial is typically

$$g(x) = \big(x - a\big)\big(x - a^2\big)\big(x - a^3\big) \cdots \big(x - a^{n-k}\big),$$

where a is an element whose powers generate the multiplicative group of the field.

The codewords are the vectors corresponding to polynomials of degree $n-1$ that are multiples of $g(x)$. For example, in our $(3, 2, 2)$ code, the vector $[0, A, B]$ is a codeword because its associated polynomial is a multiple of $g(x)$:

$$0x^2 + Ax + B = (x - A)(0x + A)$$

(all calculations being done in $\mathbb{F}_4$). You can see this codeword in FIGURE 3.

To encode a message, we must first represent the message as a vector of length k (the number of information components) whose entries are in the finite field. We then interpret the vector as a polynomial, the *message polynomial*, $m(x)$. Now to find the codeword for the message, we divide $m(x) \cdot x^{n-k}$ by $g(x)$ to obtain quotient $q(x)$ and remainder $r(x)$. Then the codeword $w(x)$ is given by

$$w(x) = q(x) \cdot g(x) = m(x) \cdot x^{n-k} - r(x).$$

In the magic trick, a message is one of our 16 symbols, or equivalently, an ordered pair $[c, s]$ where c is a color and s is a shape. Here, each of the four colors is represented as an element from the set $\mathbb{F}_4 = \{0, 1, A, B\}$. In this paper, 0 represents Red,

1 represents Yellow, A represents Green, and B represents Blue. Similarly, each of the four shapes is also represented as an element from $\mathbb{F}_4$. Here, 0 represents Circle, 1 represents Square, A represents Diamond, and B represents Star. For example, the message $[0, A]$ represents the Red Diamond, so that the message polynomial is

$$m(x) = 0x + A.$$

In general, the symbol $[c, s]$ becomes the polynomial $m(x) = cx + s$.

Now divide $m(x) \cdot x$ by $x - A$ using long division. In the general case of $m(x) = cx + s$, we find that

$$\begin{aligned} m(x) \cdot x &= c \cdot x^2 + s \cdot x \\ &= \big[(c \cdot x + (s \oplus (A \odot c))\big] \cdot (x - A) + \big[(c \odot B) \oplus (A \odot s)\big], \end{aligned}$$

so the quotient is $q(x) = c \cdot x + (s \oplus (A \odot c))$ and the remainder is $r(x) = (c \odot B) \oplus (A \odot s)$. Hence, the corresponding codeword associated with the message $[c, s]$ is given by

$$\begin{aligned} w(x) &= q(x)(x - A) \\ &= [\, c,\ s,\ (c \odot B) \oplus (A \odot s)\,]. \end{aligned} \tag{1}$$

Repeating this calculation for all of the possibilities for $[c, s]$, we get a (3, 2, 2) Reed–Solomon code.

For example, the message polynomial associated with the Red Diamond would be $0 \cdot x + A$, since 0 represents Red and A represents Diamond. Using (1) to compute the third component, the codeword associated with the Red Diamond is $[0, A, (0 \odot B) \oplus (A \odot A)] = [0, A, 0 \oplus B] = [0, A, B]$.

To construct the first three cards, we associate each element of $\mathbb{F}_4 = \{0, 1, A, B\}$ with one of the four quadrants on each card. In this paper, the upper quadrant is associated with 0, the right quadrant is associated with 1, the bottom quadrant is associated with A, and the left quadrant is associated with B. Now place each of the 16 symbols somewhere in the correct quadrant on each of the three cards. For example, the Red Diamond, with codeword $[0, A, B]$, would be placed in the upper section of the first card, the bottom section of the second card, and the left section of the third card.

The fourth card As mentioned earlier, the (3, 2, 2) Reed–Solomon code has minimum distance 2 and thus cannot correct any errors. An (n, k, d) Reed–Solomon code can be extended to an $(n + 1, k, d + 1)$ code (minimum distance $d + 1$) by appending a fourth component called a *parity check* to the codeword. In a parity check, the sum of all of the codeword components should be zero. For the magic trick, a formula that can be used to compute the parity check component for each symbol is given by $p = c \oplus s \oplus t$, where t is the third component of the codeword, i.e., $(A \odot s) \oplus (c \odot B)$. Summing the four components, we see that $c \oplus s \oplus t \oplus p = c \oplus s \oplus t \oplus (c \oplus s \oplus t) = 0$.

In the case of the Red Diamond, $t = B$, so the parity check would be computed as $p = 0 \oplus A \oplus B = 1$. The (4, 2, 3) codeword associated with the Red Diamond is given by $[c, s, t, p] = [0, A, B, 1]$. The rest of the (4, 2, 3) code is given in FIGURE 4. The fourth card is constructed using the parity check component for each of the codewords. Since the parity check component of the Red Diamond is 1, then the Red Diamond would be placed on the right quadrant of the fourth card.

The order of the symbols within each section of the cards was chosen so that any chosen symbol would appear in the same position within the upper section after rotation. Once the assignment of the symbols to sections was made using the (4, 2, 3)

Reed–Solomon code, the assignment of the symbols within the sections was implemented by logical deduction and a little trial and error.

Application to CD and DVD players

This magic trick is a significantly simplified representation of how a CD or DVD player can process data on a disc containing fingerprints or scratches. Instead of representing the data as a finite field with 4 elements, a finite field of size 256 is typically used. This finite field consists of the elements $\{0, 1, \alpha, \alpha^2, \ldots, \alpha^{254}\}$ and much larger arithmetic tables.

More powerful Reed–Solomon codes are able to correct many errors per codeword. Typically, a (255, 223) code with generating polynomial

$$g(x) = \big(x - \alpha\big)\big(x - \alpha^2\big)\big(x - \alpha^3\big) \cdots \big(x - \alpha^{255-223}\big)$$

is used. It can be mathematically proven that this code has minimum distance $d = 33$. Using the formula $(d - 1)/2$, we see that this code is capable of correcting up to 16 errors that may be present in the 255 components of each codeword.

Instead of the erasure decoding technique used in the magic trick, sophisticated methods such as the Berlekamp–Massey algorithm and the Euclidean algorithm are typically used to solve this problem. It is not generally known, however, that these two techniques are equivalent (see [**2**]).

REFERENCES

1. Richard Ehrenborg, Decoding the Hamming Code, *Math Horizons*, special issue on *Codes, Cryptography, and National Security*, **13** (2006) 16–17.
2. Todd D. Mateer, On the equivalence of the Berlekamp–Massey and Euclidean algorithms for algebraic decoding, *Proceedings of the 2011 Canadian Workshop on Information Theory (IEEE),* Kelowna, British Columbia, Canada, May 2011.
3. ———, A magic trick based on the Hamming Code, *Math Horizons* **21**(2) (November, 2013) 9–11, http://dx.doi.org/10.4169/mathhorizons.21.2.9.
4. Todd K. Moon, *Error Correction Coding: Mathematical Methods and Algorithms*, John Wiley & Sons, New York, 2005.

Summary A magic trick based on properties of a (4, 2, 3) extended Reed–Solomon code is introduced. We then discuss relevant concepts of the Reed–Solomon code to explain why the trick works and how it was designed. This magic trick can be used as a teaching device in introductory error-correcting codes courses.

Downloadable images of the cards used in this trick appear in an article supplement at the Magazine's website. Through the end of 2014, the images can also be downloaded from http://www.mathematicsmagazine.org.

If you are going to perform this trick, you might want to experiment with different methods of identifying the selected symbol. With practice—and depending on your memory skills—you might be able to avoid handling the cards, or to present them in random order.

Your method of identifying the card amounts to a decoding algorithm for this Reed-Solomon code. Fast, effective decoding algorithms can be tricky, even for a very simple code!

A Golden Product Identity for e

ROBERT P. SCHNEIDER
Emory University
Atlanta, GA 30322
robert.schneider@emory.edu

Consider the number e, the base of the natural logarithm, a constant of fundamental significance in calculus and the sciences. A spectacular connection between e and π is arrived at through the theory of complex numbers, highlighted by Euler's identity [**1**]

$$e^{i\pi} = -1.$$

In this note, we relate e to another famous constant via functions from number theory. Let τ denote the golden ratio (often denoted by ϕ in the popular literature), the number

$$\tau = \frac{1+\sqrt{5}}{2}.$$

Discovered in antiquity, the golden ratio has fascinated mathematicians and laypersons alike over the centuries [**3**]. Among many nice properties that τ possesses, it is related to its own reciprocal by the identity $\tau = 1 + 1/\tau$.

In the theory of numbers, the Euler phi function $\varphi(n)$ counts the number of positive integers less than an integer n that are co-prime to n. The classical Möbius function $\mu(n)$ is defined at a positive integer n as

$$\mu(n) = \begin{cases} 1 & \text{if } n = 1 \\ 0 & \text{if } n \text{ is not squarefree} \\ 1 & \text{if } n \text{ is squarefree, having an even number of prime factors} \\ -1 & \text{if } n \text{ is squarefree, having an odd number of prime factors.} \end{cases}$$

Both φ and μ are at the heart of deep results [**4**]. These two arithmetic functions, intimately connected to the factorization of integers, are related through the identity [**2**]

$$\varphi(n) = n \sum_{d|n} \frac{\mu(d)}{d}.$$

(The sum on the right has one term for each positive divisor d of n, including $d = 1$ and $d = n$.)

Here we prove an interesting infinite product expansion relating the phi function and Möbius function to the golden ratio; moreover, we show that the constant e can be written in terms of these three number-theoretic objects. In the course of the proof, we find a curious inverse relationship between φ and μ.

THEOREM. *We have the infinite product identity*

$$e = \prod_{n=1}^{\infty} \left(1 - \frac{1}{\tau^n}\right)^{\frac{\mu(n)-\varphi(n)}{n}}.$$

Math. Mag. **87** (2014) 132–134. doi:10.4169/math.mag.87.2.132.

While the golden ratio is related to many marvelous self-similar phenomena, it is not particularly central to research in number theory. However, e, μ, and φ pervade the study of prime numbers and other arithmetic topics; it is intriguing to find these important entities intersecting τ in this product formula.

We begin the proof of the theorem by stating a few well-known identities. For any positive integer n, the following two identities hold [**2**]:

$$\sum_{d|n} \varphi(d) = n \qquad \sum_{d|n} \mu(d) = \begin{cases} 1 & \text{if } n = 1 \\ 0 & \text{if } n > 1. \end{cases}$$

In addition to these, we have the Maclaurin series

$$-\log(1-y) = \sum_{j=1}^{\infty} \frac{y^j}{j}$$

when $0 < y < 1$, where $\log x$ denotes the natural logarithm of a real variable $x > 0$. We use these expressions to prove two lemmas, leading to a quick proof of the main result.

LEMMA 1. *On the interval* $0 < x < 1$*, we have the identities*

$$-\sum_{k=1}^{\infty} \frac{\varphi(k)}{k} \log\left(1 - x^k\right) = \frac{x}{1-x} \tag{1}$$

$$-\sum_{k=1}^{\infty} \frac{\mu(k)}{k} \log\left(1 - x^k\right) = x. \tag{2}$$

Proof of Lemma 1. To prove (1), observe that the left-hand side of the identity is equal to

$$\sum_{k=1}^{\infty} \frac{\varphi(k)}{k} \sum_{j=1}^{\infty} \frac{x^{kj}}{j}.$$

For any $n \geq 1$, the coefficient of x^n in the preceding expression is

$$\frac{\sum_{k|n} \varphi(k)}{n} = 1.$$

Therefore, the left side of (1) is equal to the geometric series

$$x + x^2 + x^3 + \cdots = \frac{x}{1-x}.$$

To prove (2), observe that the left-hand side of the identity is equal to

$$\sum_{k=1}^{\infty} \frac{\mu(k)}{k} \sum_{j=1}^{\infty} \frac{x^{kj}}{j}.$$

For any $n \geq 1$, the coefficient of x^n in the preceding expression is

$$\frac{\sum_{k|n} \mu(k)}{n} = \begin{cases} 1 & \text{if } n = 1 \\ 0 & \text{if } n > 1. \end{cases}$$

Every coefficient but the first equals zero; therefore the left side of (2) is equal to x.

(These infinite series are absolutely convergent when $0 < x < 1$, justifying the changes in order of summation.) ∎

The proof of Lemma 1 makes use of basic properties of φ, μ, geometric series, and power series. Using these identities, the following lemma displays a kind of inverse relationship between the Euler phi function and Möbius function with respect to the golden ratio: The two sums in Lemma 1 are reciprocals precisely when x is equal to $1/\tau$.

LEMMA 2. *We have the pair of reciprocal summation identities*

$$\tau = -\sum_{k=1}^{\infty} \frac{\varphi(k)}{k} \log\left(1 - \frac{1}{\tau^k}\right)$$

$$\frac{1}{\tau} = -\sum_{k=1}^{\infty} \frac{\mu(k)}{k} \log\left(1 - \frac{1}{\tau^k}\right).$$

Proof of Lemma 2. Observing that $0 < 1/\tau < 1$, we make the substitution $x = 1/\tau$ in the two identities of Lemma 1. We use the identity $1/\tau = \tau - 1$ to simplify the right-hand side of (1), as

$$\frac{1/\tau}{1 - 1/\tau} = \frac{1}{\tau - 1} = \frac{1}{1/\tau} = \tau.$$ ∎

Proof of the Theorem. We subtract the second identity in Lemma 2 from the first identity and observe that $\tau - 1/\tau = 1$, to find

$$\sum_{k=1}^{\infty} \frac{\mu(k) - \varphi(k)}{k} \log\left(1 - \frac{1}{\tau^k}\right) = 1.$$

By applying the exponential function to both sides of this identity, we arrive at the theorem. ∎

Similar product identities for $\exp(x)$, $\exp\left(\frac{x}{1-x}\right)$, and $\exp\left(\frac{x^2}{1-x}\right)$ are obtained by applying the exponential function to the identities in Lemma 1. More generally, it is evident from the above proof that for any arithmetic function f, we have the identity

$$\prod_{n=1}^{\infty} (1 - x^n)^{-\frac{f(n)}{n}} = \exp\left(\sum_{n=1}^{\infty} \frac{(1 * f)(n)}{n} x^n\right),$$

where $(1 * f)(n) = \sum_{d|n} f(d)$, as usual.

Acknowledgments The author thanks David Leep at the University of Kentucky for his guidance during the preparation of this report, and for noting a simplified proof of Lemma 1, as well as Cyrus Hettle, Richard Ehrenborg, Ken Ono, Adele Lopez, and Andrew Granville for their useful suggestions.

REFERENCES

1. W. Dunham, *Euler: The Master of Us All*, Mathematical Association of America, Washington, DC, 1999.
2. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers,* Oxford University Press, Oxford, 1980.
3. M. Livio, *The Golden Ratio*, Broadway Books, New York, 2002.
4. J. Stopple, *A Primer of Analytic Number Theory: From Pythagoras to Riemann*, Cambridge University Press, Cambridge, 2003.

Summary We prove an infinite product representation for the constant e involving the golden ratio, the Möbius function, and the Euler phi function—prominent players in number theory.

Dominance Orders, Generalized Binomial Coefficients, and Kummer's Theorem

TYLER BALL
Pacific Lutheran University
Tacoma, WA 98847
ballta@plu.edu

TOM EDGAR
Pacific Lutheran University
Tacoma, WA 98847
edgartj@plu.edu

DANIEL JUDA
University of Arkansas
Fayetteville, AR 72701
dpjuda@uark.edu

Reducing Pascal's Triangle modulo a prime leads to some fascinating images and interesting patterns. Unfortunately, reducing the triangle modulo a composite number does not yield analogous patterns. In the current work, we investigate Pascal's Triangle mod p (for p prime) via an ordering on the natural numbers dependent on the base p expansions of numbers. These two objects are connected by a famous, yet relatively unknown, theorem due to Lucas. It turns out that these orderings can be defined for any base $b \geq 2$ (not just prime); as such, replacing the usual binomial coefficients by certain generalized binomial coefficients, as defined by Knuth and Wilf in [**7**], leads to a similar connection between the orders and the associated triangles. These new coefficients also provide a uniform description of carries in base b arithmetic allowing us to provide a generalization of another famous theorem due to Kummer. In the process we describe the interaction between the orders of interest, arithmetic in base b, and the relevant family of generalized binomial coefficients, which come from a simple integer sequence.

Base b dominance orders

We begin by introducing a class of partial orders on the set of natural numbers, $\mathbb{N} = \{0, 1, 2, \ldots\}$. We first encountered these orders while studying the combinatorics of ovals and their application to coding theory [**5**], and our interest was rekindled while reading about an application of the same orders to what is called "Dismal Arithmetic" [**2**].

Recall that for any integer $b \geq 2$, we can write each natural number in its base b expansion as

$$n = \sum_{i=0}^{k} n_i b^i$$

where $0 \leq n_i < b$ for each i and $n_k \neq 0$. We call k the *length* of n in base b, denoted $\text{len}_b(n)$. For ease, we take n_i to be zero when $i > k$. Since base 10 is standard, we will write numbers using the usual base 10 representation and use finite sequences to describe the base b expansion of a number, that is, $n = (n_k, \ldots, n_1, n_0)_b$.

Math. Mag. **87** (2014) 135–143. doi:10.4169/math.mag.87.2.135.

DEFINITION. For $b \in \mathbb{N}$ with $b \geq 2$, let $\ll_b$ be the relation on $\mathbb{N}$ defined by

$$n \ll_b m \quad \iff \quad n_i \leq m_i \text{ for all } i.$$

In other words, we compare the digits in the base b expansions of n and m. In this case we will say *m b-dominates n* (or *n is dominated by m* in base b).

EXAMPLE. Let $n = 5273$ and $m = 15041$. Then $5273 \ll_6 15041$, since $5273 = (0, 4, 0, 2, 2, 5)_6$ and $15041 = (1, 5, 3, 3, 4, 5)_6$. On the other hand, since $5273_8 = (1, 2, 2, 3, 1)_8$, $15041_8 = (3, 5, 3, 0, 1)_8$, and $3 > 0$, it follows that $5273 \not\ll_8 15041$. We can also check that $15041 \not\ll_8 5273$.

The relation $\ll_b$ is reflexive, antisymmetic, and transitive on $\mathbb{N}$. This means that $\ll_b$ is a *partial order* and $(\mathbb{N}, \ll_b)$ is a *partially ordered set* ("poset"). We say that *y covers x* if $x \ll_b y$ and there is no element $z \in \mathbb{N} \setminus \{x, y\}$ with $x \ll_b z \ll_b y$. The b-dominance order can be visualized using a graph known as a *Hasse diagram*; the vertices correspond to the elements of $\mathbb{N}$, the edges represent covering relations, and the other relations are determined by the transitive property. See FIGURE 1 for a picture of the Hasse diagram of 3-dominance up to 26. For general information about order theory, consult [**10**]; for the many properties of the dominance order, see [**3**].

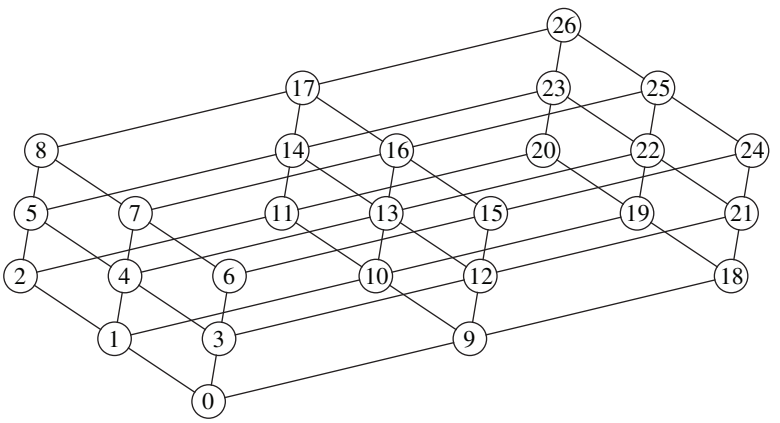


Figure 1 Hasse diagram of 3-dominance order restricted to $\{0, \ldots, 26\}$

Remarkably, p-dominance for p prime can be determined using binomial coefficients. This claim arises from a famous theorem of Lucas.

THEOREM 1. (LUCAS) *Let $n, m, p \in \mathbb{N}$ with p prime, $m = (m_\ell, \ldots, m_0)_p$, and $n = (n_k, \ldots, n_0)_p$. Then*

$$\binom{m}{n} \equiv \prod_{i=0}^{\max\{k,\ell\}} \binom{m_i}{n_i} \pmod{p},$$

where we interpret $\binom{m_i}{n_i}$ as zero when $m_i < n_i$.

We omit the proof, since there are many sources containing different and interesting proofs [**1, 4, 6, 8**].

Now, fix a prime p. It is well known that the binomial coefficient $\binom{a}{b}$ is never a multiple of p when $0 \leq b \leq a < p$. Suppose $m = (m_\ell, \ldots, m_0)_p$ and $n = (n_k, \ldots, n_0)_p$. If $n \not\ll_p m$, then $n_j > m_j$ for some j, which means $\binom{m_j}{n_j} = 0$ by definition, making

$\prod \binom{m_i}{n_i} \equiv 0 \pmod{p}$. On the other hand, if $n \ll_p m$, it follows that $\binom{m_i}{n_i} \not\equiv 0 \pmod{p}$ for each i; consequently, $\prod \binom{m_i}{n_i} \not\equiv 0 \pmod{p}$, since p is prime. We have thus seen that

$$n \ll_p m \quad \Longleftrightarrow \quad \prod_{i=0}^{\max(k,\ell)} \binom{m_i}{n_i} \not\equiv 0 \pmod{p}.$$

Applying Lucas' theorem gives

$$n \ll_p m \quad \Longleftrightarrow \quad \binom{m}{n} \not\equiv 0 \pmod{p}. \tag{1}$$

This result allows us to visualize p-dominance using Pascal's Triangle mod p. For example, when $p = 2$, each entry of Pascal's Triangle is shaded light if the entry is 0 (mod 2) and shaded dark if the entry is nonzero (mod 2). Equation (1) then states that $n \ll_2 m$ if and only if the nth entry in the mth row of this triangle is dark. FIGURE 2 shows this construction mod 2 and also mod p for $p = 3, 5, 7$. All of these images were created using Sage [**11**].

As guaranteed by Lucas' Theorem, the shaded entries represent instances of dominance. For instance, in Pascal's Triangle mod 2, the seventh row is completely dark, and we conclude that 7 2-dominates every number less than or equal to 7, which is

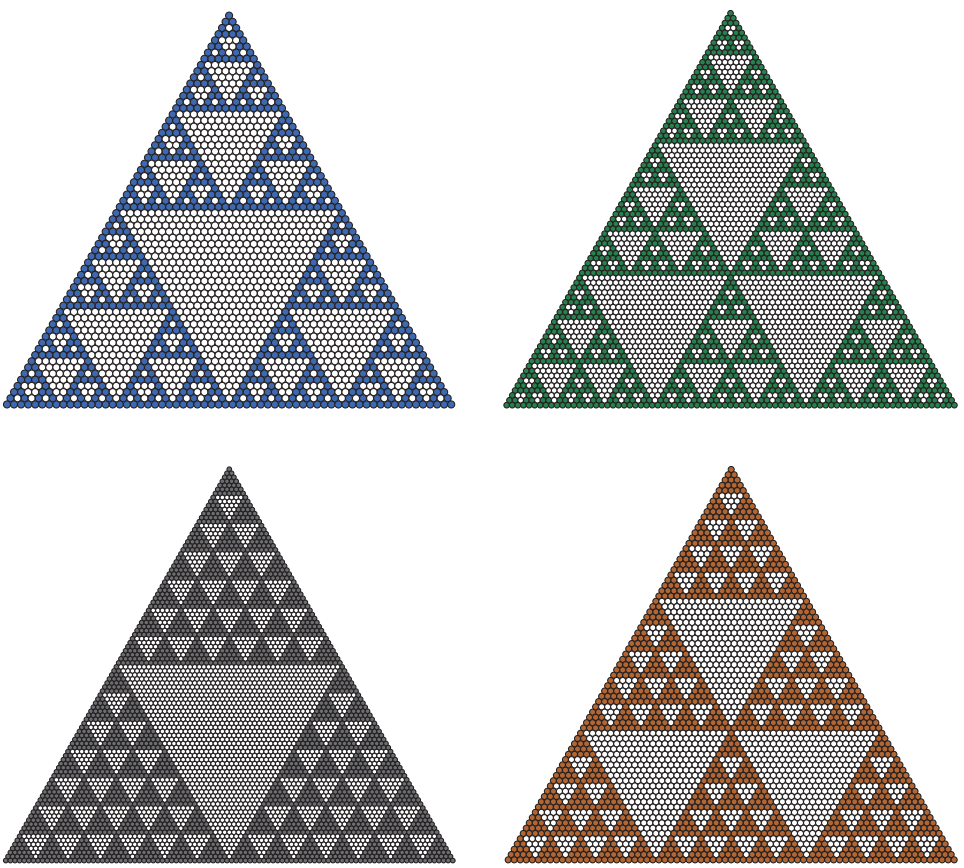


Figure 2 Pascal's Triangle mod 2, mod 3, mod 5, and mod 7 (in clockwise order from top left) with zero entries in white

true since $7 = (1, 1, 1)_2$. Furthermore, we can compare Pascal's Triangle mod 3 to the Hasse diagram for 3-dominance given in FIGURE 1.

Unfortunately, the characterization given in equation (1) does not hold for b-dominance when b is composite. For example, Pascal's Triangle mod 6, illustrated in FIGURE 3, has no apparent relationship with the 6-dominance order. For example, $2 \ll_b 4$. However, entry 2 in row 4 of the triangle is white, since $\binom{4}{2} \equiv 0 \pmod 6$. In what follows, we introduce a class of generalized binomial coefficients as defined in [**7**] to create "generalized" Pascal's Triangles. These coefficients provide the desired characterization of b-dominance, which can be visualized using the associated triangles. For instance, FIGURE 4 shows 6-dominance given by one of these triangles; this image contains the regularity we had hoped to see in Pascal's Triangle mod 6.

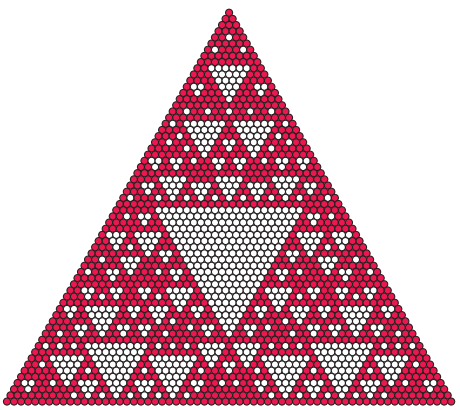


Figure 3 Pascal's Triangle mod 6, which seems to be unrelated to 6-dominance

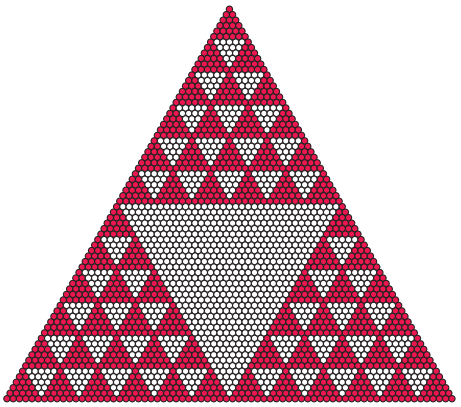


Figure 4 The "6-binomial triangle" mod 6, which does correspond to 6-dominance

Rank, "Carries," and Kummer's Theorem for prime bases

The relevant feature of the b-dominance poset to our discussion is its rank function. Note that b-dominance has a unique minimal element, 0. The *rank function* of the b-dominance order is the function r such that $r(0) = 0$ and, for $y \neq 0$, $r(y)$ is the size of the largest set $\{x_1, \ldots, x_n\}$ of nonzero integers with the property that

$$0 \ll_b x_1 \ll_b x_2 \ll_b \cdots \ll_b x_n = y.$$

(This longest chain may not exist for general partial orders, but does exist for b-dominance as it is a "lower-finite" poset [**3**].)

It turns out that the rank function is given by the sum-of-digits function, which has been well studied, so that for $n = (n_k, \ldots, n_0)_b$ we have

$$\text{rank}_b(n) = \sum_{i=0}^{k} n_i.$$

This function provides a link between base b arithmetic and the b-dominance order.

To understand the connection, we give a formal definition of a base b carry: For $m, n \in \mathbb{N}$ with $n \leq m$, the *base b carries* when adding n and $m - n$, denoted $\epsilon_i^{m,n,b}$,

are defined by $\epsilon_{-1}^{m,n,b} = 0$ and for all $i \geq 0$,

$$\epsilon_i^{m,n,b} = \begin{cases} 1 & \text{if } n_i > m_i, \\ 1 & \text{if } n_i = m_i \text{ and } \epsilon_{i-1}^{m,n,b} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

For ease of notation, we will refer to $\epsilon_i^{m,n,b}$ as ϵ_i. When $n \leq m$, the base b expansion of m can be obtained from the expansions of n and $m - n$, since

$$m_i = n_i + (m-n)_i + \epsilon_{i-1} - b\epsilon_i. \tag{2}$$

Next, we define

$$\kappa_b(m,n) = \sum_{i=0}^{\text{len}_b(m)} \epsilon_i$$

to be the total number of carries when adding n and $m - n$ in base b. We see that equation (2) can be rewritten as $n_i + (m-n)_i - m_i = b\epsilon_i - \epsilon_{i-1}$, which implies that

$$\begin{aligned} \text{rank}_b(n) + \text{rank}_b(m-n) - \text{rank}_b(m) &= \sum_{i=0}^{\text{len}_b(m)} (n_i + (m-n)_i - m_i) \\ &= \sum_{i=0}^{\text{len}_b(m)} (b\epsilon_i - \epsilon_{i-1}). \end{aligned}$$

Since $\epsilon_{-1} = \epsilon_{\text{len}_b(m)} = 0$, we obtain the following important theorem.

THEOREM 2. (RANK-DIFFERENCE) *Let $n, m, b \in \mathbb{N}$ with $b \geq 2$ and $n \leq m$. Then*

$$\text{rank}_b(n) + \text{rank}_b(m-n) - \text{rank}_b(m) = (b-1)\kappa_b(m,n).$$

Computing $\text{rank}_b(n)$ amounts to measuring the length of the chain from 0 to n in the Hasse diagram of the order. Therefore, if we have the Hasse diagram for b-dominance, we can easily determine how many carries occur when adding two numbers n and $m - n$ in base b. For instance, we use FIGURE 1 to see that when adding 7 and 11 in base 3, we will encounter

$$\frac{\text{rank}_3(7) + \text{rank}_3(11) - \text{rank}_3(18)}{3-1} = \frac{3+3-2}{2} = 2$$

carries. Consequently, if $n \ll_b m$, then adding n and $m - n$ in base b will not involve any carries.

COROLLARY. *Let $n, m, b \in \mathbb{N}$ with $b \geq 2$ and $n \leq m$. The following are equivalent.*

(a) $n \ll_b m$

(b) $\text{rank}_b(m-n) = \text{rank}_b(m) - \text{rank}_b(n)$

(c) $\kappa_b(m,n) = 0$.

The Rank-difference theorem plays an essential role in the proof of Kummer's Theorem, which is only applicable for prime bases [**6**].

THEOREM 3. (KUMMER) *Let $n, m, p \in \mathbb{N}$ with p prime and $n \leq m$. Then the largest power of p dividing $\binom{m}{n}$ is $\kappa_p(m,n)$; that is, it is the total number of carries when adding n and $m - n$ in base b.*

Since the Rank-difference theorem holds for all b, we will use it to obtain an analog of Kummer's Theorem for composite bases in the next section.

Generalized binomial coefficients and new results

We will use a system of generalized factorials and generalized binomial coefficients introduced by Knuth and Wilf [**7**]. Given any sequence

$$C = (C_1, C_2, C_3, \ldots)$$

of natural numbers, we define the *C-factorial* by

$$(m!)_C = \begin{cases} C_m C_{m-1} \cdots C_1 & \text{when } m \neq 0 \\ 1 & \text{when } m = 0. \end{cases}$$

In effect, each C_i does the job that i would do in a standard factorial. In turn, we use the C-factorial to define the generalized binomial coefficients associated to C, called the *C-nomial coefficients*, which are given by

$$\binom{m}{n}_C = \begin{cases} \dfrac{(m!)_C}{(n!)_C((m-n)!)_C} & \text{when } 0 \leq n \leq m \\ 0 & \text{otherwise.} \end{cases}$$

These constructions work for any sequence, but the generalized binomial coefficients cannot be expected to be integers. In [**7**], the authors define the notion of a "regularly divisible sequence" and show that sequences with this property have integral generalized binomial coefficients. They discuss certain sequences satisfying the regularly divisible property and demonstrate that the Gaussian coefficients as well as the so-called Fibonomial coefficients (constructed using the Fibonacci sequence) are all integers. Moreover, they prove the following lemma, which provides a condition implying integrality of generalized binomial coefficients.

LEMMA [**7**]. *The C-nomial coefficients corresponding to a sequence C are integers if C satisfies the condition that* $\gcd(C_i, C_j) = C_{\gcd(i,j)}$ *for all* $i, j > 0$.

With these ideas in hand, we turn to the construction of the generalized binomial coefficients relevant to our discussion. For any $b, m \in \mathbb{N}$ with $b \geq 2$, we let $v_b(m) = q$ where $m = b^q z$ and b does not divide z. In other words, $v_b(m)$ represents the exponent of the largest power of b dividing m. When p is prime, v_p is well studied due to its importance for p-adic integers.

DEFINITION. Let $b \in \mathbb{N}$ with $b \geq 2$. Define the sequence $S^b = (S_1^b, S_2^b, \ldots)$ by

$$S_n^b = b^{v_b(n)},$$

that is, S_n^b is the largest power of b dividing n.

EXAMPLE. When $b = 6$, the sequence is given by

$$S^6 = (1, 1, 1, 1, 1, 6, 1, 1, 1, 1, 1, 6, 1, 1, 1, 1, 1, 6, \ldots).$$

The first entry that is not 1 or 6 is $S_{36}^6 = 36$.

We have added a few examples of these sequences to the OEIS (when $b = 4$, see A234957, and when $b = 6$, see A234959 in [**9**]).

For the rest of this section, let b be fixed so that $S = S^b$ is defined. For ease of notation, in place of $(m!)_S$ we use $(m!)_b$ and in place of $\binom{m}{n}_S$ we use $\binom{m}{n}_b$. We call the latter numbers the *b-binomial coefficients*.

THEOREM 4. *Let $b \in \mathbb{N}$ with $b \geq 2$. Each b-binomial coefficient is an integer.*

Proof. Using the fact that $\min\{v_b(i), v_b(j)\} = v_b(\gcd(i, j))$ for all b, i and j, we apply the previous lemma. In particular, let $m, n \in \mathbb{N}$. Then

$$\begin{aligned}\gcd(S_m, S_n) &= \gcd(b^{v_b(m)}, b^{v_b(n)}) \\ &= b^{\min\{v_b(m), v_b(n)\}} \\ &= b^{v_b(\gcd(m,n))} = S_{\gcd(m,n)}.\end{aligned}$$

∎

Since the b-binomial coefficients are integers, we investigate each coefficient's residue class mod b. We are particularly interested in the case when $\binom{m}{n}_b \not\equiv 0 \pmod{b}$. To proceed, we need the following auxiliary result.

LEMMA. *Let $n, b \in \mathbb{N}$ with $b \geq 2$. Then*

$$v_b\left((n!)_b\right) = \frac{1}{b-1}(n - \mathrm{rank}_b(n)).$$

Proof. Let $l = \mathrm{len}_b(n)$. According to the definition of S^b, we see that

$$v_b\left((n!)_b\right) = \sum_{i=1}^{l} \left\lfloor \frac{n}{b^i} \right\rfloor.$$

Now,

$$\sum_{i=1}^{l} \left\lfloor \frac{n}{b^i} \right\rfloor = \sum_{i=1}^{l} \sum_{j=i}^{l} n_j b^{j-i}.$$

Note that $n_j b^i$ appears exactly once in this sum if and only if $i < j$. Therefore,

$$\sum_{i=1}^{l} \sum_{j=i}^{l} n_j b^{j-i} = \sum_{j=1}^{l} \sum_{i=0}^{j-1} n_j b^i = \sum_{j=1}^{l} n_j \sum_{i=0}^{j-1} b^i = \sum_{j=1}^{l} n_j \left(\frac{b^j - 1}{b-1}\right).$$

Finally,

$$\begin{aligned}\sum_{i=1}^{l} n_i \left(\frac{b^i - 1}{b-1}\right) &= \frac{1}{b-1}\left[\sum_{i=1}^{l}(n_i b^i) - \sum_{i=1}^{l} n_i\right] \\ &= \frac{1}{b-1}\left[n - n_0 - \sum_{i=1}^{l} n_i\right] = \frac{1}{b-1}\left[n - \sum_{i=0}^{l} n_i\right].\end{aligned}$$

∎

Recall that $\kappa_b(m, n)$ represents the number of carries when adding n and $m - n$ in base b and that Kummer's Theorem, stated above, can be formulated as

$$v_p\left(\binom{m}{n}\right) = \kappa_p(m, n).$$

Using the b-binomial coefficients, we obtain the following analogous result, which can be viewed as a simplified version of Kummer's Theorem that applies to both prime and composite bases.

THEOREM 5. *Let $m, n, b \in \mathbb{N}$ with $b \geq 2$ and $n \leq m$. Then the highest power of b that divides $\binom{m}{n}_b$ is the number of carries when adding n and $m - n$ in base b.*

Proof. As $\binom{m}{n}_b = \frac{(m!)_b}{(n!)_b((m-n)!_b}$, we apply the lemma and simplify to get

$$\begin{aligned} v_b\left(\binom{m}{n}_b\right) &= \left(\frac{1}{b-1}\right)(\text{rank}_b(m-n) + \text{rank}_b(n) - \text{rank}_b(m)) \\ &= \kappa_b(m, n), \end{aligned}$$

where the last equality follows by the Rank-difference theorem. ∎

This result unlocks the promised characterization of b-dominance. We know that $n \ll_b m$ if and only if $\kappa_b(m, n) = 0$, and the latter condition is now equivalent to $\binom{m}{n}_b \not\equiv 0 \pmod{b}$. Therefore, we have

$$n \ll_b m \quad \Longleftrightarrow \quad \binom{m}{n}_b \not\equiv 0 \pmod{b}. \tag{3}$$

When p is prime, $\binom{m}{n}_p \not\equiv 0 \pmod{p}$ if and only if $\binom{m}{n} \not\equiv 0 \pmod{p}$; thus, we obtain the consequence of Lucas' Theorem given in equation (1).

Using the b-binomial coefficients, we create a "generalized" Pascal's Triangle by placing the coefficient $\binom{m}{n}_b$ in the nth entry of the mth row; we call this triangle the b-binomial triangle. By equation (3) this triangle, reduced mod b, describes b-dominance: If we shade the zero entries white and other entries dark, the dark entries in row m correspond to the natural numbers b-dominated by m. (We can revisit FIGURE 4 to see this triangle when $b = 6$).

We finish by investigating the set $(\ll_b m) := \{n \in \mathbb{N} \mid n \ll_b m\}$, often called a *principal lower set*. If $m = (m_{\text{len}_b(m)}, \ldots, m_0)_b$ then every $n \ll_b m$ has base b expansion of the form $n = (n_{\text{len}_b(m)}, \ldots, n_0)_b$, where $0 \leq n_i \leq m_i$ for all i. Thus, to construct $n \ll_b m$, there are $m_i + 1$ choices for the ith coefficient, n_i. Since each choice of coefficient is independent from the others, we have

$$|(\ll_b m)| = \prod_{i=0}^{\text{len}_b(m)} (m_i + 1).$$

Accordingly, this implies that the number of nonzero entries in the mth row of b-binomial triangle mod b is $\prod(m_i + 1)$. As previously noted, $\binom{m}{n}_p \not\equiv 0 \pmod{p}$ if and only if $\binom{m}{n} \not\equiv 0 \pmod{p}$ for p prime; consequently, the previous statement recovers a well-known result about the number of nonzero entries in Pascal's Triangle mod p. This result can be found in [**4**].

Acknowledgment The authors thank the M. J. Murdock Charitable Trust and the Pacific Lutheran University Division of Natural Sciences for their generous support.

REFERENCES

1. Peter G. Anderson, Arthur T. Benjamin, and Jeremy A. Rouse, Combinatorial proofs of Fermat's, Lucas's, and Wilson's theorems, *Amer. Math. Monthly* **112** (2005) 266–268.
2. David Applegate, Marc Lebrun, and N. J. A. Sloane, *Dismal Arithmetic*, arXiv:1107.1130v2, 2011.

3. Tyler Ball and Daniel Juda, Dominance over $\mathbb{N}$, *Rose Hulman Undergraduate Math Journal* **14**(2) (2013), Article 2.
4. N. J. Fine, Binomial coefficients modulo a prime, *Amer. Math. Monthly* **54** (1947) 589–592.
5. David G. Glynn, A condition for the existence of ovals in $PG(2; q)$; q even, *Geom. Dedicata* **32**(2) (1989) 247–252.
6. Andrew Granville, Arithmetic properties of binomial coefficients I: Binomial coefficients modulo prime powers, in *Organic mathematics*, CMS Conf. Proc. vol. 20, Amer. Math. Soc., 1997, 253–276.
7. Donald E. Knuth and Herbert S. Wilf, The power of a prime that divides a generalized binomial coefficient, *J. Reine Angew. Math.* **396** (1989) 212–219.
8. Edouard Lucas, Theorie des Fonctions Numeriques Simplement Periodiques, *Amer. J. Math.* **1**(4) (1878) 289–321.
9. The On-Line Encyclopedia of Integer Sequences, http://oeis.org.
10. Bernd S. W. Schröder, *Ordered Sets, An introduction*, Birkhauser, Boston, 2003.
11. W. A. Stein *et al.*, Sage Mathematics Software (Version 4.8), The Sage Development Team, 2012, http://www.sagemath.org.

Summary We discuss the connections between a family of partial orders known as b-dominance orders, arithmetic in base b, and generalized binomial coefficients. In particular, we investigate theorems of Lucas and Kummer in relation to these orders and attempt to extend and explain these theorems using a family of generalized binomial coefficients derived from a simple integer sequence.

A Golden Connection

In this issue (page 132) we describe some connections between the golden ratio, the constant e, and important functions in number theory. Here is another connection. Ramanujan found this beautiful continued fraction identity involving the golden ratio, here called ϕ:

$$\cfrac{1}{1+\cfrac{e^{-2\pi}}{1+\cfrac{e^{-4\pi}}{1+\cfrac{e^{-6\pi}}{1+\cdots}}}} \quad = \quad \left(\sqrt{\phi\sqrt{5}}-\phi\right) e^{2\pi/5}.$$

This exotic formula connects the golden ratio to both π and e, weaving together in a single expression the three most famous constants in mathematics.

It is an instance of the Rogers-Ramanujan continued fraction; see Andrews, The meaning of Ramanujan now and for the future, *The Ramanujan Journal* **20** (2009) 257–273.

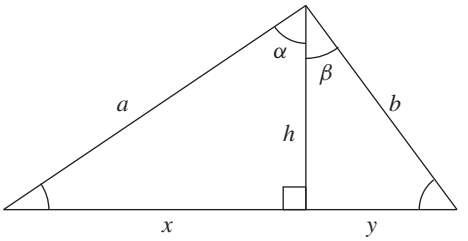
Robert P. Schneider

Math. Mag. **87** (2014) 143. doi:10.4169/math.mag.87.2.143.

A Proof of the Cosine Addition Formula Using the Law of Cosines

PATRIK NYSTEDT
Department of Engineering Science, University West
Trollhättan, Sweden
Patrik.Nystedt@hv.se

It is fairly well known (see, e.g., [**1**] or [**2**, p. 39]) that the following triangle can be used to prove the sine addition formula in the case when α and β are acute.



In fact, using the area formula on the largest triangle and then splitting up this area into the two right triangles shows that

$$\frac{ab\sin(\alpha+\beta)}{2} = \frac{xh}{2} + \frac{hy}{2}$$

and hence that

$$\sin(\alpha+\beta) = \frac{x}{a}\cdot\frac{h}{b} + \frac{h}{a}\cdot\frac{y}{b} = \sin\alpha\cos\beta + \cos\alpha\sin\beta.$$

What seems to be less known is that the same triangle can be used to prove the cosine addition formula, using the law of cosines:

$$(x+y)^2 = a^2 + b^2 - 2ab\cos(\alpha+\beta).$$

Indeed, if we solve for $\cos(\alpha+\beta)$ and use Pythagoras' theorem, then we get that

$$\cos(\alpha+\beta) = \frac{a^2+b^2-(x+y)^2}{2ab} = \frac{a^2-x^2+b^2-y^2-2xy}{2ab} = \frac{2h^2-2xy}{2ab}$$
$$= \frac{h}{a}\cdot\frac{h}{b} - \frac{x}{a}\cdot\frac{y}{b} = \cos\alpha\cos\beta - \sin\alpha\sin\beta.$$

REFERENCES

1. Christopher Brueningsen, Proof Without Words, *Math. Mag.* **66** (1993) 135.
2. Roger B. Nelsen, *Proof Without Words II,* Mathematical Association of America, Washington, DC, 2000.

Summary We give a proof of the cosine addition formula using the law of cosines.

Math. Mag. **87** (2014) 144. doi:10.4169/math.mag.87.2.144.

A Dissection Proof of Leibniz's Series for $\pi/4$

MITSUO KOBAYASHI
California State Polytechnic University, Pomona
Pomona, CA 91768
mkobayashi@csupomona.edu

Archimedes [**4**, p. 233] discovered a method of subdividing a parabolic segment into infinitely many triangles, which allowed him to express the total area as a geometric series. He was thus able to determine the area of the parabolic segment. Since then, there has been much interest in expressing the area of interesting figures as a sum of infinitely many subdivided areas, not only to determine the areas of the original figures, but also to discover series expressions for their areas. For instance, Lord Brouncker [**1**] studied certain regions bounded by a hyperbola by decomposing the regions into rectangles. (One of Brouncker's decompositions was recently rediscovered and appears as a Proof Without Words in [**5**].) This partition allowed Lord Brouncker to discover the series

$$\ln 2 = \left(1 - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{5} - \frac{1}{6}\right) + \left(\frac{1}{7} - \frac{1}{8}\right) + \cdots,$$

where consecutive terms are paired so that each pair is a positive value representing the area of a rectangle.

We may also try to proceed in reverse. That is, suppose we have an infinite series expression for a quantity known to be the area of a geometric figure. Can we partition the region so that the areas of the subdivisions correspond to terms of the series? Inspired by the work of Lord Brounker, Viggo Brun [**2, 3**] devoted considerable effort to finding a method of partitioning the unit circle into areas corresponding to the terms of Leibniz's series

$$\frac{\pi}{4} = \left(1 - \frac{1}{3}\right) + \left(\frac{1}{5} - \frac{1}{7}\right) + \left(\frac{1}{9} - \frac{1}{11}\right) + \cdots,$$

where consecutive terms are paired as in Brounker's series so that each pair of terms has positive value. Using polar coordinates, Brun found that the sequence of functions

$$r = B_n(\theta) := \tan^{2n}\left(\frac{\theta}{2}\right), \qquad n = 0, 1, 2, \ldots,$$

partitions a quarter circle into parts with areas following the terms of Leibniz's series for $\pi/4$ (FIGURE 1).

In this article, we describe a simple method of partitioning a region below a function using a power series of the function to arrive at an infinite series with terms that are the areas of the parts. We then apply this method to recover Brun's partition as well as to discover a new partition of the quarter circle related to his.

A partition for e

We first demonstrate the method by showing how we may visualize the series for Euler's number e,

$$e = \sum_{n=0}^{\infty} \frac{1}{n!}.$$

Math. Mag. **87** (2014) 145–150. doi:10.4169/math.mag.87.2.145.

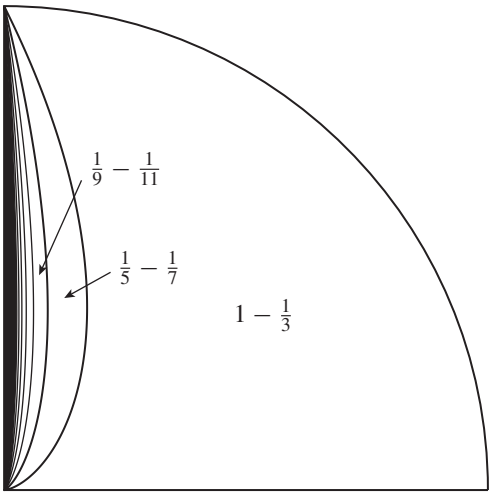


Figure 1 Brun's partition of the quarter circle according to Leibniz's series

We first need a region with area e. We will simply use the region bounded by the function $f(x) = e^x + 1$ and the x-axis in the interval $[0, 1]$, since

$$\int_0^1 (e^x + 1)\, dx = e.$$

In addition, we consider the Taylor series of $e^x + 1$ about $x = 0$. Denote the degree n Taylor polynomial by s_n. Thus,

$$s_0 = 2, \qquad s_1 = 2 + x, \qquad s_2 = 2 + x + \frac{x^2}{2}, \qquad s_3 = 2 + x + \frac{x^2}{2} + \frac{x^3}{6},$$

and so on. The graphs of these polynomials slice through our region, effecting an infinite partition. This is a consequence of the fact that the Taylor polynomials are all positive and increasing with n, for each x in the interval $(0, 1]$. Denoting the new areas $a_0, a_1, a_2, \ldots$ as in FIGURE 2, we have $a_0 = 2$ and, for $n \geq 1$,

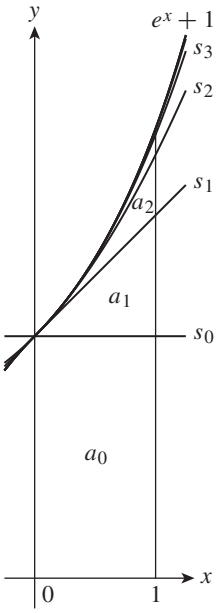
$$a_n = \int_0^1 (s_n - s_{n-1})\, dx = \int_0^1 \frac{x^n}{n!}\, dx = \frac{1}{(n+1)!}.$$

Observe that this method will work whenever we have a series representation for a function that is nonnegative on an interval, such that the terms of the series are also nonnegative in that interval. Starting with the region below the curve, we define the partition using the graphs of the partial sums of the series. Integrating the function as well as the series termwise, results in the area of the whole region along with the areas of the parts. In this case, we will say we have found a *series partition* for a region. When we perform such a series partition and arrive at an interesting series, then we have successfully associated this series with an illustration of it.

A series partition for $\pi/4$

We now seek a series partition for a region having area $\pi/4$. Since Leibniz's series was originally discovered by plugging in $x = 1$ into the Maclaurin series for $\arctan x$,

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots,$$



$$\int_0^1 (e^x + 1)\,dx = a_0 + a_1 + a_2 + \cdots$$

$$e = 2 + \frac{1}{2!} + \frac{1}{3!} + \cdots$$

Figure 2 A series partition for e

it is natural to consider the integral over [0, 1] of the derivative series

$$f(x) = \frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + x^8 - x^{10} + \cdots .$$

If we follow Brouncker and Brun's pairing of terms, the corresponding partial sums are

$$s_1 = 1 - x^2,$$
$$s_2 = 1 - x^2 + x^4 - x^6,$$
$$s_3 = 1 - x^2 + x^4 - x^6 + x^8 - x^{10}, \ldots,$$

or, in general,

$$s_n = \frac{1 - x^{4n}}{1 + x^2}.$$

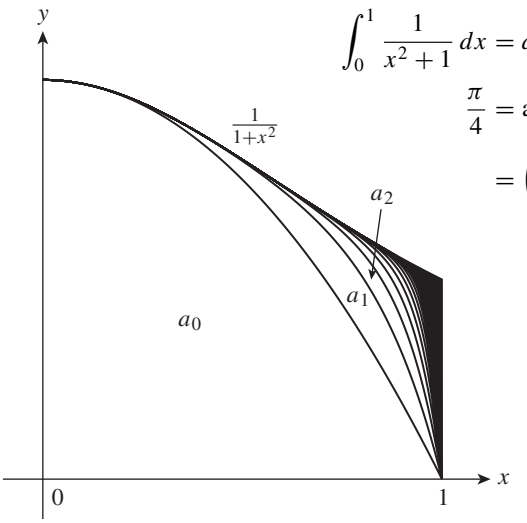
Since these polynomials are positive and increasing with n for each fixed x in the interval (0, 1), the graphs form a series partition of our region, as shown in FIGURE 3.

A mapping into the circle

Although we have found a series partition for the desired series, the region involved is certainly not a quarter circle. We now show how we can map the polynomials into the desired region in such a way that areas are preserved.

For each x in [0, 1], we first choose $\theta = \theta(x)$ so that the area of the sector of the unit circle between 0 and θ is equal to the area below $f(x) = (1 + x^2)^{-1}$ in $[0, x]$, namely,

$$\frac{\theta}{2} = \arctan x.$$



$$\int_0^1 \frac{1}{x^2+1}\,dx = a_0 + a_1 + a_2 + \cdots$$

$$\frac{\pi}{4} = \arctan 1$$

$$= \left(1 - \frac{1}{3}\right) + \left(\frac{1}{5} - \frac{1}{7}\right) + \left(\frac{1}{9} - \frac{1}{11}\right) + \cdots$$

Figure 3 A series partition for $\pi/4$

Then, for each s_n, we seek to determine r_n so that for each x and each $\theta = 2\arctan x$, the area bounded by s_n and the x-axis in $[0, x]$ is equal to the polar-coordinate area bounded by r_n in $[0, \theta]$, namely,

$$\int_0^x s_n(x)\,dx = \frac{1}{2}\int_0^\theta r_n^2(\theta)\,d\theta.$$

Differentiating each side of this expression with respect to x, we get

$$s_n = \frac{1}{2}r_n^2\frac{d\theta}{dx} = \frac{r_n^2}{1+x^2},$$

and we find that

$$r_n = \sqrt{s_n(1+x^2)}.$$

Thus, we have shown that

$$(x, y) \mapsto (r, \theta) \qquad \text{where} \qquad r = \sqrt{y(1+x^2)}, \qquad \theta = 2\arctan x$$

is an area-preserving transformation from the region below $(x^2+1)^{-1}$ in $[0, 1]$ to the quarter circle.

This transformation produces a new Brun-like partition of the quarter circle by means of

$$s_n = \frac{1-x^{4n}}{1+x^2}, \qquad x = \tan\left(\frac{\theta}{2}\right)$$

to yield

$$r_n = B_n'(\theta) := \sqrt{1 - \tan^{4n}\left(\frac{\theta}{2}\right)}, \qquad 0 \le \theta \le \frac{\pi}{2},$$

for $n = 1, 2, 3, \ldots$. We also define $B_0'(\theta) = 0$. This partition is shown in FIGURE 4.

For instance, the area below $B_1'(\theta)$ is

$$\frac{1}{2}\int_0^{\pi/2} \left(B_1'(\theta)\right)^2\,d\theta = \frac{1}{2}\int_0^{\pi/2} \left(1 - \tan^4\left(\frac{\theta}{2}\right)\right)\,d\theta = 1 - \frac{1}{3},$$

which is the first term of Leibniz's series, as expected.

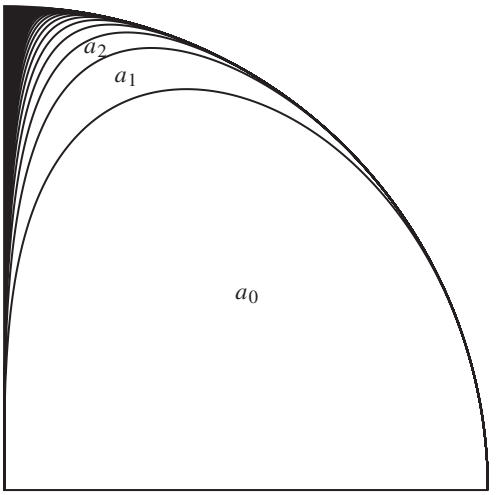


Figure 4 A mapping of the series partition into the quarter circle

However, this is not the only way to produce an area-preserving transformation. Rather than mapping low values of y to low values of r, we can map y to r in reverse: Namely, map $y = 0$ onto the circle and $y = f(x)$ to the interior of the circle. This amounts to replacing our original integral expression by

$$\int_0^x s_n(x)\,dx = \frac{1}{2}\int_0^\theta \left(1 - r_n^2(\theta)\right)\,d\theta.$$

Repeating our calculations, we now get the transformation

$$(x, y) \mapsto (r, \theta), \qquad r = \sqrt{1 - y(1 + x^2)}, \qquad \theta = 2\arctan x,$$

resulting in

$$r_n(\theta) = \sqrt{x^{4n}} = \tan^{2n}\left(\frac{\theta}{2}\right),$$

which is Brun's original partition for the quarter circle.

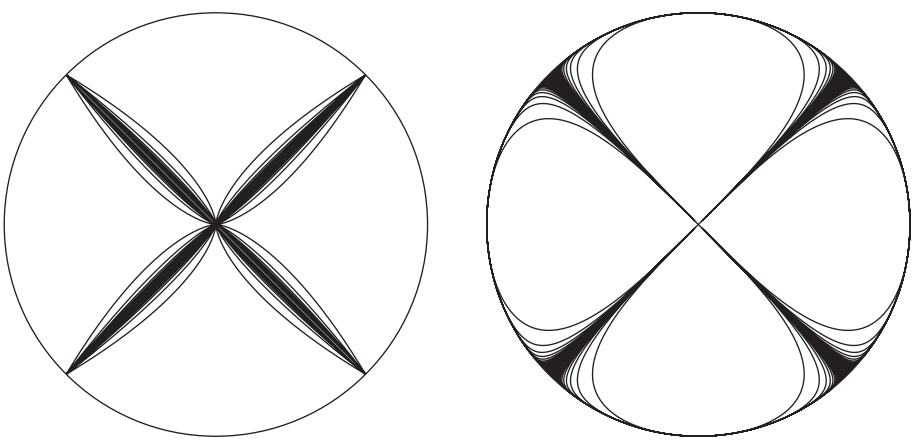
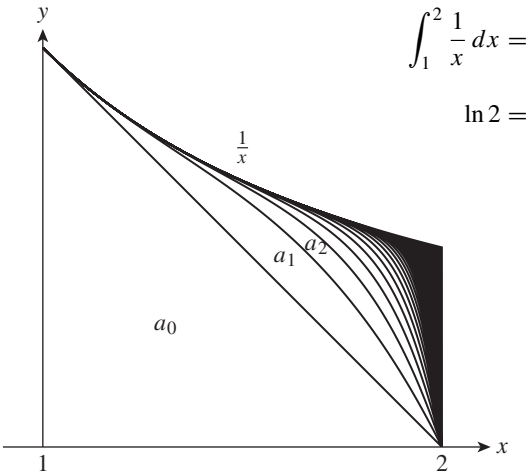


Figure 5 Two partitions of the unit circle based on Leibniz's series



$$\int_1^2 \frac{1}{x}\,dx = a_0 + a_1 + a_2 + \cdots$$

$$\ln 2 = \left(1 - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{5} - \frac{1}{6}\right) + \cdots$$

Figure 6 A series partition for $\ln 2$

Brun concluded his paper by presenting a variation of his partition, in which he used $\tan^{2n}(\theta)$ to partition an eighth of a circle and reflecting into the next eighth, then continuing around the complete circle so that the final result produced a highly symmetric picture. We display both his and our new partitions in FIGURE 5 for comparison. In these pictures, the largest regions each have areas $1 - 1/3$; then pairs of the next largest regions have areas $1/5 - 1/7$, and so on.

A parting partition

We conclude by observing that Lord Brouncker's region is also amenable to series partition. We simply consider the Taylor series for $1/x$ about $x = 1$, which yields the partition in FIGURE 6 for $\ln 2$. This partition can be thought of as the smooth version of Lord Brouncker's observation, and we have come full circle, as it were.

Acknowledgment The author thanks Paul Pollack for introducing him to the work of Brun.

REFERENCES

1. W. Brouncker, The Squaring of the Hyperbola, by an Infinite Series of Rational Numbers, Together with Its Demonstration, *Philosophical Transactions* **3** (1668) 645–649.
2. V. Brun, On the problem of partitioning the circle so as to visualize Leibniz' formula for π, *Nordisk Mat. Tidskr.* **3** (1955) 159–166.
3. ———, Leibniz' formula for π deduced by a "mapping" of the circular disc, *Nordisk Mat. Tidskr.* **18** (1970) 73–81.
4. T. L. Heath, editor, *The Works of Archimedes*, Dover, New York, 1912.
5. M. Hudelson, Proof without words: the alternating harmonic series sums to $\ln 2$, *Math. Mag.* **83** (2010) 294, http://dx.doi.org/10.4169/002557010X521831.

Summary Inspired by Lord Brouncker's discovery of his series for $\ln 2$ by dissecting the region below the curve $1/x$, Viggo Brun found a way to partition regions of the unit circle so that their areas correspond to terms of Leibniz's series for $\pi/4$. Brun's argument involved *ad hoc* methods which were difficult to find. We develop a method based on usual techniques in calculus that leads to Brun's result and that applies generally to other related series.

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; EUGEN J. IONASCU, Columbus State University; 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, 2014.

1941. *Proposed by Spiros P. Andriopoulos, Third High School of Amaliada, Eleia, Greece.*

Find all pairs of nonzero real valued functions $(u(x), v(x))$ that satisfy both of the following equations:

$$\left(\frac{u}{v}\right)' = \frac{u'}{v'} \quad \text{and} \quad u'v' = uv.$$

1942. *Proposed by Marcel Chiriţă, Bucharest, Romania.*

Let $n \geq 2$ be an integer. Prove that

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

is not an integer.

1943. *Proposed by Mihály Bencze, Brason, Romania*

Let n be a positive integer, and $a_1, a_2, \ldots, a_n$ be positive real numbers. Set $A = \frac{1}{n}\sum_{k=0}^{n} a_k$ and $a_{n+k} = a_k$ for $1 \leq k \leq n-1$. Prove that

$$\sum_{k=0}^{n-1} \sqrt{a_{k+1} + \sqrt{a_{k+2} + \sqrt{a_{k+3} + \cdots + \sqrt{a_{k+n}}}}} \leq n\sqrt{A + \sqrt{A + \sqrt{A + \cdots + \sqrt{A}}}}.$$

Math. Mag. **87** (2014) 151–158. doi:10.4169/math.mag.87.2.151.

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.

1944. *Proposed by Mowaffaq Hajja and Mostafa Hayajneh, Yarmouk University, Irbid, Jordan.*

For any triangle XYZ with side lengths x, y, and z, let $\nu(\triangle XYZ) = x^2 + y^2 + z^2$.

(a) Let $ABCD$ be a convex quadrilateral with diagonals intersecting at P. Prove that if

$$\nu(\triangle PAB) = \nu(\triangle PBC) = \nu(\triangle PCD) = \nu(\triangle PDA), \tag{1}$$

then $ABCD$ is a rhombus.

(b) Classify the quadrilaterals $ABCD$ for which there exists a point P (not necessarily the intersection of the diagonals) such that (1) holds.

1945. *Proposed by Michael W. Botsko, Saint Vincent College, Latrobe, PA.*

Let f be continuous on $[0, 1]$ and let $\int_0^1 f(x)dx = 0$. In addition, let $\phi(x)$ be differentiable on $[0, 1]$ with $\phi(0) = 0$ and $\phi'(x) > 0$ on $(0, 1)$. Prove that there exists $x_0 \in (0, 1)$ such that

$$\int_0^{x_0} \phi(x) f(x)dx = 0.$$

Quickies

Answers to the Quickies are on page 157.

Q1039. *Proposed by Cezar Lupu, University of Pittsburgh, Pittsburgh, PA.*

Let $x_1, x_2, \ldots, x_n$ be real numbers in the interval $(0, 1)$ and let $s = x_1 + x_2 + \cdots + x_n$. For $1 \le i \le n$, let $s_i = s - x_i$. Prove that

$$s_1^{x_1} + s_2^{x_2} + \cdots + s_n^{x_n} > n - 1.$$

Q1040. *Proposed by Ovidiu Furdui, Technical University of Cluj-Napoca, Cluj, Romania.*

Let $a < b$ be positive real numbers. Calculate

$$\lim_{n\to\infty} \frac{1}{n} \sqrt[n]{\int_a^b \ln^n(1 + e^{nx})dx}.$$

Solutions

Variety of remainders — April 2013

1916. *Proposed by H. A. ShahAli, Tehran, Iran.*

Let M and n be positive integers. For every integer d, let r_d be the remainder obtained when n is divided by d. Let $R_n = \{r_d : 1 \le d \le \lfloor n/2 \rfloor\}$. Prove that $|R_n| > M$ holds for all sufficiently large positive integers n.

Solution by Robert Calcaterra, University of Wisconsin-Platteville, Platteville, WI.
Suppose that m is an integer such that $\lceil n/3 \rceil < m \le \lfloor n/2 \rfloor$. Then $n/3 < m \le n/2$ and so $2 \le n/m < 3$. Therefore, for each m in the interval, the quotient when dividing by

m is 2, and so the remainders $r_m = n - 2m$ must be distinct for distinct m. It follows that

$$|R_n| \geq \left\lfloor \frac{n}{2} \right\rfloor - \left\lceil \frac{n}{3} \right\rceil \geq \frac{n-1}{2} - \frac{n+2}{3} = \frac{n-7}{6},$$

and so $\lim_{n\to\infty} |R_n| = \infty$.

Editor's Note. By considering the range of values m such that $3 \leq n/m < 4$ and the principle of inclusion and exclusion, it can be proved that $|R_n| \geq (5/24)n - O(1)$. The remainders obtained from the range of values m such that $4 \leq n/m < 5$ are all included in the range obtained from $2 \leq n/m < 3$. The next range that provides new values comes from the values m such that $5 \leq n/m < 6$. In this case, the principle of induction yields $|R_n| \geq (79/360)n - O(1)$. Following this approach, it is possible to prove that $\lim_{n\to\infty} |R_n|/n$ exists. As pointed by Omran Kouba, numerical evidence suggests that this the value of the limit is close to 0.229.

Also solved by Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Australia), Paul Budney, CMC328, Robert L. Doucette, Dmitry Fleischman, Daniel Fritze (Germany), Eugene A. Herman, Cliff S. Holroyd, Tom Jager, Omran Kouba (Syria), Elias Lampakis (Greece), Reiner Martin (Germany), Peter McPolin (Northern Ireland), José Heber Nieto (Venezuela), Joel Schlosberg, Thomas Steinberger, Traian Viteam (Chile), Edward T. White, Timothy Woodcock, and the proposer.

Generalized mean inequalites **April 2013**

1917. *Proposed by Ovidiu Furdui, Technical University of Cluj-Napoca, Cluj, Romania.*

Let $\alpha > 0$ and let a and b be real numbers such that $b > a$. Find the value of

$$\lim_{n\to\infty} \int_a^b \sqrt[n]{(x-a)^n + (b-x)^n + \alpha((x-a)(b-x))^{n/2}}\, dx.$$

Solution by Charles Martin, Vanderbilt University, Nashville, TN.
We show that $L = (3/4)(b-a)^2$. Let us observe that the integrand

$$f_n(x) := \sqrt[n]{(x-a)^n + (b-x)^n + \alpha[(x-a)(b-x)]^{n/2}},$$

for $x \in [a, b]$ and fixed n, is a symmetric function with respect to $x = c := (a+b)/2$. Hence, it is enough to show that $\lim_{n\to\infty} \int_a^c f_n(x)dx = L/2$.

For $x \in [a, c]$, we have $0 \leq x - a \leq b - x$, which implies

$$b - x \leq f_n(x) \leq (2+\alpha)^{1/n}(b-x).$$

Since $\int_a^c (b-x)dx = (3/8)(b-a)^2$, integrating the above inequality on $[a, c]$ gives

$$\frac{3(b-a)^2}{8} \leq \int_a^c f_n(x)dx \leq (2+\alpha)^{1/n}\frac{3(b-a)^2}{8}.$$

Letting $n \to \infty$, because $(2+\alpha)^{1/n} \to 1$, the Squeeze Principle gives us the claimed value for L.

Editor's Note. Michael Goldenberg and Mark Kaplan sent us an elegant solution based on known results about the power mean inequalities of positive numbers:

$$M_n := \left(\frac{\alpha_1 a_1^n + \alpha_2 a_2^n + \cdots + \alpha_k a_k^n}{\alpha_1 + \cdots + \alpha_k}\right)^{1/n},$$

where $a_j > 0$ and $\alpha_j > 0$ for $1 \le j \le k$. It is known that $\{M_n\}$ is non-decreasing and convergent to $\max_{1\le j\le k}(a_j)$. By Dini's theorem, since the sequence $f_n/(2+\alpha)^{1/n}$ is non-decreasing and pointwise convergent to a continuous function, it must converge uniformly. Hence, $\int_a^b f_n(x)dx$ converges to

$$\int_a^b \max(x-a, b-x, \sqrt{(x-a)(b-x)})dx = \frac{3(b-a)^2}{4}.$$

Tomas Persson and Mikael P. Sundqvist showed that $L = (3/4)(b-a)^2$, even if $\alpha \in [-2, 0]$. They used a different but clever analysis in the special case $\alpha = -2$.

Also solved by Hatef I. Arshagi, Michel Bataille (France), Robert Calcaterra, Robert L. Doucette, John N. Fitch, J. A. Grzesik, Eugene A. Herman, Michael Goldenberg and Mark Kaplan, Omran Kouba (Syria), Elias Lampakis (Greece), Paolo Perfetti (Italy), Tomas Persson (Sweden) and Mikael P. Sundqvist (Sweden), Amit K. Verma (India), Haohao Wang and Jerzy Wojdylo, and the proposer. There were four incomplete solutions.

Solving a cubic geometrically **April 2013**

1918. *Proposed by Kent Holing, StatoilHydro, Trondheim, Norway.*

A triangle is given with side lengths a, b, c, and with inradius and circumradius r and R, respectively. Let $s = \frac{1}{2}(a+b+c)$ denote the semiperimeter of the triangle.

(a) Solve the cubic $x^3 - (r+4R)x^2 + s^2x - rs^2 = 0$ exactly by starting with the monic cubic having roots $s-a$, $s-b$, and $s-c$.

(b) Using (a), give a geometrical interpretation of the roots of the given cubic.

Solution by Peter McPolin, St. Mary's University College, Belfast, Northern Ireland.
The roots of the given cubic are the exradii r_a, r_b, and r_c of the given triangle, where $r_a = rs/(s-a)$, $r_b = rs/(s-b)$, and $r_c = rs/(s-c)$.

Let $p(x)$ denote the monic cubic having roots $s-a$, $s-b$, and $s-c$, that is, $p(x) = (x-(s-a))(x-(s-b))(x-(s-c))$. Upon expansion, we obtain $p(x) = x^3 - \sigma_1 x^2 + \sigma_2 x - \sigma_3$, where σ_1, σ_2, and σ_3 are the elementary symmetric functions of the roots:

$$\begin{aligned}
\sigma_1 &= (s-a) + (s-b) + (s-c) = 3s - (a+b+c) = s,\\
\sigma_2 &= (s-a)(s-b) + (s-b)(s-c) + (s-c)(s-a)\\
&= 3s^2 - 2s(a+b+c) + ab + bc + ca = -s^2 + ab + bc + ca, \text{ and}\\
\sigma_3 &= (s-a)(s-b)(s-c).
\end{aligned}$$

To express σ_2 and σ_3 in terms of r, s, and R, we make use of Heron's formula for the area A of a triangle, i.e., $A^2 = s(s-a)(s-b)(s-c)$, and the following two well-known triangle formulas $A = rs$ and $abc = 4Rrs$. We obtain

$$\begin{aligned}
r^2s^2 = A^2 &= s(s-a)(s-b)(s-c)\\
&= s(s^3 - s^2(a+b+c) + s(ab+bc+ca) - abc)\\
&= s(-s^3 + s(ab+bc+ca) - 4Rrs),
\end{aligned}$$

giving $ab+bc+ca = s^2 + r^2 + 4Rr$, and hence $\sigma_2 = r^2 + 4Rr$. It also follows from the above that $\sigma_3 = r^2 s$. Thus the cubic $p(x)$ may be expressed as

$$p(x) = x^3 - sx^2 + (r^2 + 4Rr)x - r^2 s.$$

Comparing this form of $p(x)$ with the given cubic $x^3 - (r + 4R)x^2 + s^2x - rs^2$, we see that they are related by means of a simple reciprocal transformation:

$$\left(\frac{-x^3}{r^2 s}\right) p\left(\frac{rs}{x}\right) = x^3 - (r + 4R)x^2 + s^2x - rs^2.$$

It follows that the roots of the given cubic are $rs/(s - a)$, $rs/(s - b)$, and $rs/(s - c)$. These are well-known expressions for the exradii r_a, r_b, and r_c, respectively, of the three excircles of a triangle having side lengths a, b, and c.

Also solved by Arkady Alt, George Apostolopoulos (Greece), Brian D. Beasley, Michel Bataille (France), Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Australia), Robert Calcaterra, Robert L. Doucette, J. A. Grzesik, John G. Heuver, Omran Kouba (Syria), Elias Lampakis (Greece), Raúl A. Simón (Chile), Nicholas C. Singer, Traian Viteam (Chile), Michael Vowe (Switzerland), John Zacharias, and the proposer.

Symmetrical packings of circles **April 2013**

1919. *Proposed by Abdurrahim Yilmaz, Middle East Technical University, Ankara, Turkey.*

Let $n \geq 3$ be a positive integer. A circle of radius 1 centered at the origin is inscribed by a circular arrangement of n congruent circles, where every two consecutive circles in the arrangement are tangent. This arrangement is in turn inscribed by a second circular arrangement of n congruent circles in two different ways, as shown in the figures below. In FIGURE 1, the line through the centers of tangent circles in different arrangements goes through the origin. In FIGURE 2, each circle in the second arrangement is tangent to two consecutive circles in the first arrangement. Suppose that this process continues indefinitely, as shown in the figures. For each of the two cases, find the area enclosed by the union of all circles when $n \to \infty$.

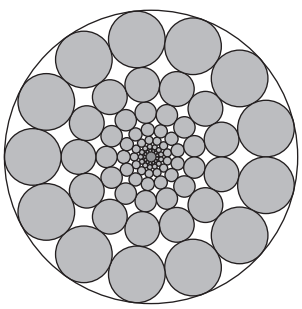


Figure 1

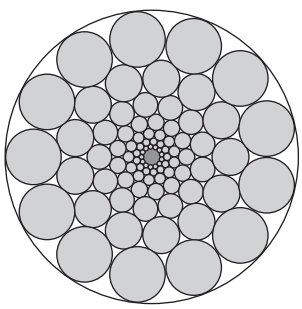


Figure 2

Solution by Tim McDevitt, Sciences, Elizabethtown College, Elizabethtown, PA.
For the two cases shown in FIGURES 1 and 2, the arrangements of n tangent circles are inscribed inside concentric circles of decreasing radius R_k, $k = 0, 1, 2, \ldots$, where $R_0 = 1$ to coincide with the outermost circle. Let $\theta_n = \pi/n$; for the n circles in an arrangement to be tangent to each other, they must have radius

$$r_k = \frac{\sin\theta_n}{1 + \sin\theta_n} R_k, \tag{1}$$

which we deduce from triangle OAB in FIGURE 3. For the case illustrated in FIGURE 1, $R_{k+1} = R_k - 2r_k$, which leads to the recurrence relation

$$R_{k+1} = \frac{1 - \sin\theta_n}{1 + \sin\theta_n} R_k,$$

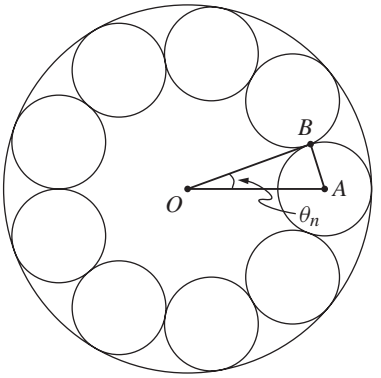


Figure 3

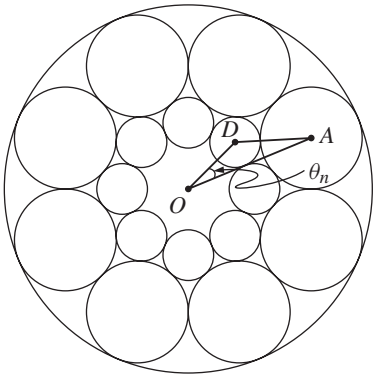


Figure 4

and thus

$$R_k = \left(\frac{1-\sin\theta_n}{1+\sin\theta_n}\right)^k.$$

The sum of the areas of all of the circles in FIGURE 1 is easily computed, since the resulting series is geometric:

$$\sum_{k=0}^{\infty} n\pi r_k^2 = n\pi\left(\frac{\sin\theta_n}{1+\sin\theta_n}\right)\sum_{k=0}^{\infty}\left[\left(\frac{1-\sin\theta_n}{1+\sin\theta_n}\right)^2\right]^k = \frac{\pi}{4}n\sin\theta_n = \frac{\pi^2}{4}\frac{\sin\theta_n}{\theta_n},$$

which approaches $\pi^2/4$ when $n \to \infty$. For the case shown in FIGURE 2, a similar analysis follows, except that it is more difficult to relate R_{k+1} to R_k. Focusing on angle AOD, the law of cosines applied to triangle OAD in FIGURE 4 gives

$$(r_k + r_{k+1})^2 = (R_{k+1} - r_{k+1})^2 + (R_k - r_k)^2 - 2(R_{k+1} - r_{k+1})(R_k - r_k)\cos\theta_n,$$

which, using (1), reduces to

$$(1-2f_n)R_{k+1}^2 - 2[f_n^2 + (1-f_n)^2\cos\theta_n]R_kR_{k+1} + (1-2f_n)R_k^2 = 0,$$

where $f_n = \sin\theta_n/(1+\sin\theta_n)$. Applying the quadratic formula and simplifying yields

$$R_{k+1} = \frac{\cos\theta_n + \sin^2\theta_n - \sin\theta_n\sqrt{1+2\cos\theta_n}}{\cos^2\theta_n}R_k,$$

where the minus solution is chosen so that $R_{k+1} < R_k$. This implies that

$$R_k = \left(\frac{\cos\theta_n + \sin^2\theta_n - \sin\theta_n\sqrt{1+2\cos\theta_n}}{\cos^2\theta_n}\right)^k,$$

and thus the sum of the areas of the circles in FIGURE 2 is

$$\sum_{k=0}^{\infty} n\pi r_k^2 = n\pi\left(\frac{\sin\theta_n}{1+\sin\theta_n}\right)^2\sum_{k=0}^{\infty}\left[\left(\frac{\cos\theta_n + \sin^2\theta_n - \sin\theta_n\sqrt{1+2\cos\theta_n}}{\cos^2\theta_n}\right)^2\right]^k$$

$$= \frac{\pi^2\sin^2\theta_n\cos^4\theta_n}{\theta_n(1+\sin\theta_n)^2(\cos^4\theta_n - (\cos\theta_n + \sin^2\theta_n - \sin\theta_n\sqrt{1+2\cos\theta_n})^2)},$$

which approaches $\pi^2/(2\sqrt{3})$ as $n \to \infty$.

Also solved by Elton Bojaxhiu (Albania) and Enkel Hysnelaj (Australia), Robert Calcaterra, William Dickinson, Robert L. Doucette, John N. Fitch, Dmetry Fleischmann, J. A. Grzesik, Eugene A. Herman, Hidefumi Katsuura and Edward Schmeichel, Omran Kouba (Syria), Victor Kutsenok, Hosam M. Mahmoud, Howard Cary Morris, Ángel Plaza (Spain) and José M. Pacheco (Spain), Kevin A. Roper, John Zacharias, and the proposer. There were two incomplete or incorrect solutions.

An identity of sums of binomials **April 2013**

1920. *Proposed by Ilya Bluskov, University of Northern British Columbia, Prince George, Canada.*

For every positive integer n, prove that

$$\sum_{m=0}^{n}\sum_{r=0}^{n-m}\frac{(-2)^r}{m!r!}\binom{2n-m-r}{n}=\sum_{r=0}^{n}\frac{(-1)^r}{r!}\binom{2n-r}{n}.$$

Solution by Thomas Steinberger, Ohio Northern University, Ada, OH.
Let $k=m+r$. It follows that the double sum above becomes

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

By interchanging the sums, we have

$$\begin{aligned}\sum_{k=0}^{n}\frac{1}{k!}\binom{2n-k}{n}\sum_{m=0}^{k}\frac{(-2)^{k-m}k!}{m!(k-m)!}&=\sum_{k=0}^{n}\frac{1}{k!}\binom{2n-k}{n}\sum_{m=0}^{k}(-2)^{k-m}\binom{k}{m}\\&=\sum_{k=0}^{n}\frac{1}{k!}\binom{2n-k}{n}(1-2)^k\\&=\sum_{k=0}^{n}\frac{(-1)^k}{k!}\binom{2n-k}{n},\end{aligned}$$

as desired.

Also solved by Aleams Barra (Indonesia), Michel Bataille (France), Robert Calcaterra, Eddie Cheng, Robert L. Doucette, Daniel Fritze (Germany), Omran Kouba (Syria), José Heber Nieto (Venezuela), Ángel Plaza (Spain), Joel Schlosberg, Timothy Woodcock, William Xiao, and the proposer.

Answers

Solutions to the Quickies from page 152.

A1039. We use Bernoulli's inequality for real exponents: $(1+x)^r \le 1+rx$ for $x > -1$ and $0 < r < 1$. For every $1 \le i \le n$,

$$\left(\frac{1}{s_i}\right)^{x_i}=\left(1+\left(\frac{1}{s_i}-1\right)\right)^{x_i}\le 1+\left(\frac{1}{s_i}-1\right)x_i=\frac{s-s_ix_i}{s_i}.$$

It follows that

$$s_i^{x_i}\ge\frac{s_i}{s-s_ix_i}>\frac{s_i}{s}.$$

Adding these inequalities for $1 \le i \le n$, we obtain

$$s_1^{x_1} + s_2^{x_2} + \cdots + s_n^{x_n} > \frac{1}{s}\sum_{i=1}^{n} s_i = \frac{1}{s}\sum_{i=1}^{n}(s - x_i) = \frac{1}{s}(ns - s) = n - 1.$$

A1040. The limit equals b. First, note that the integrand satisfies the following inequalities: $n^n x^n = \ln^n(e^{nx}) \le \ln^n(1 + e^{nx}) \le \ln^n(1 + e^{nb})$. Thus,

$$\sqrt[n]{\frac{b^{n+1} - a^{n+1}}{n+1}} = \frac{1}{n}\sqrt[n]{\int_a^b n^n x^n dx} \le \frac{1}{n}\sqrt[n]{\int_a^b \ln^n(1 + e^{nx})dx}$$
$$\le \frac{\ln(1 + e^{nb})}{n} \cdot \sqrt[n]{b - a}.$$

A direct calculation shows that

$$\lim_{n\to\infty} \frac{\ln(1 + e^{nb})}{n} \cdot \sqrt[n]{b - a} = b$$

and

$$\lim_{n\to\infty} \sqrt[n]{\frac{b^{n+1} - a^{n+1}}{n+1}} = \lim_{n\to\infty} b^{\frac{n+1}{n}} \cdot \lim_{n\to\infty} \sqrt[n]{1 - \left(\frac{a}{b}\right)^{n+1}} \cdot \lim_{n\to\infty} \frac{1}{\sqrt[n]{n+1}} = b.$$

It follows, based on the Squeeze Theorem, that

$$\lim_{n\to\infty} \frac{1}{n}\sqrt[n]{\int_a^b \ln^n(1 + e^{nx})dx} = b.$$

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.

Mackenzie, Dana, Prime clusters may go all the way: Major advance made toward proving legendary conjectures, *Science News* (11 January 2014) 12.

Bounded gaps between primes, http://michaelnielsen.org/polymath1/index.php?title=Bounded_gaps_between_primes.

Zhang, Yitang, Bounded gaps between primes, *Annals of Mathematics* 179 (2014) (3) (May 2014) 1121–1174. Abstract at http://dx.doi.org/10.4007/annals.2014.179.3.7.

Polymath, D.H.J., New equidistribution estimates of Zhang type, and bounded gaps between primes, http://arxiv.org/abs/1402.0811.

Green, Ben, Bounded gaps between primes, http://arxiv.org/abs/1402.4849.

Granville, Andrew, Primes in intervals of bounded length, *AMS 2014 Current Events Bulletin* http://www.ams.org/meetings/lectures/currentevents2014.pdf.

Are there infinitely many pairs of primes that differ by k or less? The twin prime conjecture is the case $k = 2$; the current best result is $k = 252$. But wait, there's more! If the Elliott Halberstam (EH) conjecture is true, then $k = 12$; and if the generalized conjecture (GEH) is true, then $k = 6$. The GEH asserts that the primes are "nicely distributed" in a particular sense (but it cannot achieve $k = 4$). This work is part of a more general research program, a crowdsourced wiki project Polymath8b that has promulgated its first paper (by "D.H.J. Polymath"). That project is to determine the least quantity H_m for which there are infinitely many intervals $n, n + 1, \ldots, n + H_m$ that contain $m + 1$ primes, with the twin prime conjecture being related to $m = 1$. Green gives an exposition of Zhang's paper and subsequent developments, as does Granville, who presented his paper at the Baltimore meeting in January.

Zorn, Paul, Paul Zorn on communicating mathematics, http://mathcomm.org/archives/paul-zorn-on-communicating-mathematics-i-an-artistic-view/.

I hope that the seven parts of this retiring MAA presidential address will appear at some point in a single document, preferably in the *Chronicle of Higher Education*. Zorn asserts that communicating mathematics is "a valid form of mathematics in its own right," and he proposes that "mathematical exposition be considered—. . . by departments, colleges, universities. . . —as a fully respectable mathematical activity"—that is, one worthy of tenure and promotion. This call to value what THIS VERY MAGAZINE publishes is welcome, but Zorn may be "preaching to the choir" (as Zorn himself notes). My perspective is that deans and tenure/promotion committees tend to favor other values, either research (at a university) or service to the institution (at a liberal arts college, in lieu of publication). Zorn notes in particular the difficulties in communicating the fine distinctions that are absolutely necessary in mathematics, citing as an example the problems students have in parsing the standard calculus textbook formulation of the definition of continuity. (However, he fails to criticize, as a logician would, the inherent confusion in putting some of the quantification at the start of the sentence and some at the end, rather than gathering it all into prenex normal form.)

Math. Mag. **87** (2014) 159–160. doi:10.4169/math.mag.87.2.159.

Biegert, Mark, Math Encounters Blog, http://mathscinotes.wordpress.com/.

I have not been a great fan of blogs, perhaps it is because it is hard to find the few that regularly provide enlightenment amid all the dreck on the Internet. In any case, here is a blog that I enjoy and subscribe to, by an engineer whose touchstone is "I stumbled upon some math today.... These are everyday problems where a little bit of math really goes a long way." Its author is willing to model any phenomenon that strikes his curiosity, dig thoroughly into relevant details and sources, and do whatever mathematics is required to get a back-of-the-envelope (or better) estimate. Examples include sewer math (the cost of leaking manholes), tire pressure (variation with temperature), T-shirt fabric, skydiving, hardware implementation of an approximation of a logarithm function, chili math (decibel scale for capsaicin), and Olympic rowing (dimensional analysis). If you too have a fondness for "Fermi problems," tune in (or whatever it is that blog-readers do).

Mazur, Joseph, *Enlightening Symbols: A Short History of Mathematical Notation and Its Hidden Powers*, Princeton University Press, 2014; 296 pp, $27.95. ISBN 978-0-691-15463-3.

This very readable book explores through history the mathematical symbolization of numbers and of algebraic operations, with remarks and observations on the power of symbols and on the nature of historical revisionism. Author Mazur's investigation and sources regarding the introduction of arabic numerals into Europe are worth reading in light of celebrations of Fibonacci as the introducer (e.g., Keith Devlin's *The Man of Numbers: Fibonacci's Arithmetic Revolution*, 2011). (Note to book designer: I, and likely others, would have preferred a larger type size in place of line-and-a-half spacing.)

Maor, Eli, and Eugen Jost, *Beautiful Geometry*, Princeton University Press, 2014; xiv + 187 pp, including almost 60 full-page color plates, $27.95. ISBN 978-0-691-15099-4.

"We attempt here to depict a wide selection of geometric theorems in an artistic way." Of course, the Pythagorean theorem appears multiple times; but also appearing among the more than 50 selections are Steiner's porism, Celtic motifs, Morley's theorem, Pick's theorem, Sierpiński's triangle, and Ceva's theorem, with several theorems illustrated by multiple instantiations of them. Most of the illustrations are set against a uniform dark background, and I cannot tell if the needs of the colorblind have been taken into account.

Hart, George W., and Reza Sarhangi (eds.), *Bridges Enschede: Mathematics, Music, Art, Architecture, Culture Conference Proceedings*. Phoenix, AZ: Tessellations Publishing, 2013. http://archive.bridgesmathart.org/2013/.

Like the better-known Gathering 4 Gardner (G4G), held every other year in Atlanta by invitation only, the Bridges conference collects an eclectic group of mathematicians and mathematical hobbyists. G4G focuses on games and puzzles; Bridges features mathematical connections in art, music, architecture, and science. Anyone may attend Bridges, which meets annually; submitted papers (10-page maximum) are refereed. Unlike G4G, Bridges papers are published online shortly after the conference and soon thereafter in print. This volume from the 2013 conference contains 115 papers, including ones on braids, choreography, ant paintings, and the Moore–Penrose inverse in art. You can search the entire archive of papers 1998–2013 at http://archive.bridgesmathart.org/.

Pukelsheim, Friedrich, *Proportional Representation: Apportionment Methods and Their Applications*, Springer, 2014; xxi + 234 pp, $69.99 (P). ISBN 978-3-319-03855-1.

How can the allocation of seats in a parliament reflect most faithfully the votes cast? I would not have thought that rules for rounding fractions could generate so much mathematics nor have so great an effect on election results. This book explains the methods employed in various European elections, uses data from real elections to illustrate the resulting biases, provides software, and suggests electoral reforms. At Pukelsheim's recommendation, a method of double proportionality, intended to incorporate both local representation and overall proportionality, was instituted in several Swiss cantons and in the 2013 revision of the German election law. [Full disclosure: I have often been a guest of Friedrich Pukelsheim at the Augsburg University Institute for Mathematics, and I made minor suggestions regarding the text of the book.]

MAA MATHFEST
August 6-9, 2014

Join us in Portland, Oregon for MAA MathFest 2014

The largest annual summertime gathering of mathematicians

Invited Addresses

Earle Raymond Hedrick Lecture Series
Speaker: Bjorn Poonen, Massachusetts Institute of Technology

AMS-MAA Joint Invited Address
Speaker: Sara Billey, University of Washington

MAA Invited Address
Speaker: Ricardo Cortez, Tulane University

MAA Invited Address
Speaker: Erika Camacho, Massachusetts Institute of Technology and Arizona State University

MAA Invited Address
Speaker: Keith Devlin, Stanford University

James R. C. Leitzel Lecture
Speaker: Joseph Gallian, University of Minnesota Duluth

AWM-MAA Etta Z. Falconer Lecture
Speaker: Marie A. Vitulli, University of Oregon

Pi Mu Epsilon J. Sutherland Frame Lecture
Speaker: Keith Devlin, Stanford University

The Jean Bee Chan and Peter Stanek Lecture for Students
Speaker: Jack Graver, Syracuse University

NAM David Harold Blackwell Lecture
Speaker: Mark Lewis, Cornell University

Martin Gardner Centennial Lecture
Speaker: Persi Diaconis, Stanford University

Mathematical Association of America • maa.org/meetings/mathfest

MAA
MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth St., NW
Washington, DC 20036

CONTENTS

ARTICLES

NOTES

PROBLEMS

REVIEWS