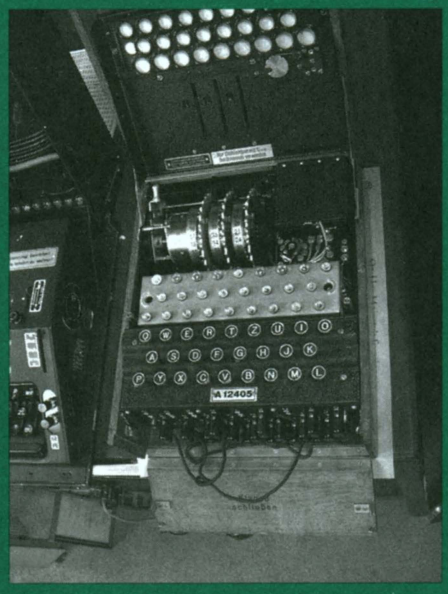
Vol. 80, No. 4, October 2007

MATHEMATICS MAGAZINE



Marian Rejewski and the Enigma Machine

- Polish Mathematicians Finding Patterns in Enigma Messages
- Seeing Dots: Visibility of Lattice Points
- The Lost Cousin of the Fundamental Theorem of Algebra

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

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

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

Submit new manuscripts to Allen Schwenk, Editor, Mathematics Magazine, Department of Mathematics, Western Michigan University, Kalamazoo, MI, 49008. Manuscripts should be laser printed, with wide line spacing, and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should mail three copies and keep one copy. In addition, authors should supply the full five-symbol 2000 Mathematics Subject Classification number, as described in *Mathematical Reviews*.

The **cover image** pays tribute to the Polish mathematician Marian Rejewski, who spotted patterns in German Enigma messages in the 1930s. Using elementary group theory, he was able to construct a model of Enigma and to determine the settings of the rotors. See the article by Chris Christensen for more details.

AUTHORS

Chris Christensen teaches mathematics at Northern Kentucky University. His mathematical genealogy is a sequence of algebraic geometers; his thesis advisor was Professor Shreeram S. Abhyankar of Purdue University. Chris' interest in cryptology began more than a decade ago when he read Robert Harris' novel *Enigma*. He was hooked. Since that time, he has enjoyed teaching cryptology and watching mathematics majors and nonmathematics majors discover mathematics by finding patterns in encrypted messages.

Joshua D. Laison received his BA from Oberlin College in 1996 and his PhD from Dartmouth College in 2001. As a Visiting Assistant Professor at Colorado College, he has involved himself in a number of research projects with students, and this paper is the happy outcome of one such project. His interests include graph theory and a wide variety of games and puzzles.

Michelle Schick received her BA from Colorado College in 2004 and is currently working on her MS at Kansas State University. This work was part of her undergraduate distinction project at Colorado College. She shows an early interest in Number Theory.

Vol. 80, No. 4, October 2007

MATHEMATICS MAGAZINE

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

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

Advertising correspondence should be addressed to

advertising@maa.org

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

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

ARTICLES

Polish Mathematicians Finding Patterns in Enigma Messages

CHRIS CHRISTENSEN
Northern Kentucky University
Highland Heights, KY 41099
christensen@nku.edu

Whenever there is arbitrariness, there is also a certain regularity. There is no avoiding it. Marian Rejewski [**14b**, p. 235].

This is a story about heroes. Its heroes are three Polish mathematicians who in the decade before World War II broke German Enigma messages. It seems rare that mathematicians are heroes of stories, and it seems even rarer that they are heroes because they are mathematicians. A recent exception is Robert Harris' novel *Enigma* [**7**] (and the 2002 Michael Apted film *Enigma* that was based upon it). In *Enigma*, which is based upon the work of the British World War II codebreakers at Bletchley Park, the hero is Tom Jericho, a mathematician whose successes are loosely based upon the work of Alan Turing. (The novel *Enigma* was reviewed by Peter Hilton who served at Bletchley Park from 1942 until the end of the war in Europe in the June 1996 *Notices of the American Mathematical Society* [**8**].)

World War II seems to mark a change in cryptology. Although mathematician Werner Kunze was recruited as a cryptologist by Germany in World War I[1] and there are examples of mathematicians studying codes and ciphers throughout the history of cryptology, World War II seems to mark the point at which cipher bureaus began to recruit mathematicians for their problem solving abilities—for their abilities to find patterns. The Government Code and Cipher School at Bletchley Park recruited many mathematicians. Probably their two most famous recruits are Alan Turing [**10**] and Gordon Welchman [**27**]. The United States Signals Intelligence Service, which was organized by William Friedman, had among its first recruits two mathematicians, Frank Rowlett [**23**] and Abraham Sinkov [**26**], and statistician Solomon Kullback [**17**]. The heroes of our story Jerzy Różycki (Roozh-IT-ski), Henryk Zygalski (Zig-AHL-ski), and Marian Rejewski (Rey-EF-ski)[2] were mathematics students at Poznań University when they were recruited into a cryptology course in 1929.

The story of the Polish mathematicians' success against Enigma is well known to cryptologists. Rejewski was able to use elementary theorems about permutations to determine the wiring of the Enigma rotors and to determine the Enigma settings. "If ever there was a real-world story problem handed to mathematics teachers on a silver platter, this would be it." [**21**, p. 371]

We will return to the work of the three Polish mathematicians, but first we will take a moment to examine substitution ciphers and the operation of Enigma.

[1] "Kunze was presumably the first professional mathematician to serve in a modern cryptanalytic bureau." [**2**, p. 85]

[2] Guides to pronunciation are taken from [**14**].

Substitution Ciphers

SAKSP VPAPV YWAVH QLUS

A substitution cipher is a method of concealment that replaces, for example, each letter of a plaintext message with another letter. Here is the key to a simple substitution cipher:

Plaintextletters :	abcdefghijklmnopqrstuvwxyz
Ciphertextletters :	EKMFLGDQVZNTOWYHXUSPAIBRCJ.

The key gives the correspondence between a plaintext letter and its replacement ciphertext letter. (It is traditional to use small letters for plaintext and capital letters for ciphertext.) Using this key, every plaintext letter a would be replaced by ciphertext E, every e by L, etc. The key describes a permutation of the alphabet. Just as in abstract algebra courses, the internal structure of the permutation is revealed when it is written as a product of disjoint cycles. In this case, our permutation consists of a 10-cycle, two 4-cycles, one 3-cycle, two 2-cycles, and a 1-cycle.

(aeltphqxru)(bknw)(cmoy)(dfg)(iv)(jz)(s)

There are 26! = 403,291,461,126,605,635,584,000,000 possible keys for such simple substitution ciphers. The security of ciphers often depends on the cipher "having a large key space"—having too many keys for the cryptanalyst to do a brute force attack of trying all the keys. This is certainly the case for our substitution cipher. If the cryptanalyst tried one key per second, it would take 4,667,725,244,520,898,560,000 days to try all possible keys. Yet, such keys are used to encipher the cryptograms that appear regularly in newspapers and puzzle books, and these cryptograms are routinely broken in a few minutes. What makes it possible to break these ciphers?

Patterns. Every language has rules so that the language "makes sense." These rules create patterns in messages that can be exploited by cryptanalysts. Usually cryptograms that appear in newspapers preserve word length and punctuation, but even without that information these simple substitution ciphers can be solved. The letter e is the most frequent letter in plaintext English. If we used the key that was described above, we would expect that the most frequent ciphertext letter would be L. Now, it might not be, but it is likely that the most frequent ciphertext letter corresponds to one of e, t, a, o, i, n, or s. Using letter frequencies and other patterns, simple substitution ciphers are usually quickly solved. Such an attack on ciphertext is called frequency analysis.

Here is a more secure method of enciphering. Instead of using the same permutation to replace each letter of the plaintext, we will have a collection of permutations and will use one permutation to determine the replacement for the first letter, another permutation to determine the replacement for the second letter, etc. Certainly there are enough permutations available to use a different permutation for each plaintext letter. This is the idea for the cipher called a one-time pad; it is the only provably secure cipher. But, there are practical problems that makes it difficult to implement this idea by hand—keeping track of the order in which the permutations will be used and communicating the order to an authorized receiver. The one-time pad is provably secure because it uses a random ordering of the permutations; there are no patterns for the cryptanalyst to discover. The classical Vigenère cipher, which was developed in the Sixteenth Century, is based on a similar idea, but it uses a small number of permutations—typically many fewer than the number of characters in the message. The key for a Vigenère cipher prescribes a rotation among the permutations—permutation number one, permutation number two, ... , permutation number n—repeated as necessary depending

on the length of the plaintext message. The Vigenère cipher was broken in the Nineteenth Century by using frequency analysis to discover patterns that become evident in the sets of ciphertext letters that are enciphered using the same permutation.

Enigma is a mechanical way to generate a large number of permutations. Although it was one of the first, Enigma was not the only machine cipher. For example, during World War II, the United States used William Friedman's SIGABA, and the British used TypeX. In fact, rotor machines dominated cryptography from the 1920s into the 1970s. Enigma began as a device to protect commercial communications.

Enigma

> *If you have no good coding system, you are always running a considerable risk. Transmitted by cable or without wire, your correspondence will always be exposed to every spy, your letters, to being opened and copied, your intended or settled contracts, your offers and important news to every inquisitive eye. Considering this state of things, it is almost inconceivable that persons interested in those circumstances should delay securing themselves better against such things. Yet, ciphering and deciphering has been a troublesome art hitherto.... Now, we can offer you our machine "Enigma", being a universal remedy for all those inconveniences.*
>
> Mid-1920s Enigma sales brochure reprinted in the July 2001 *Cryptologia*. See [**28**, p. 246].

Although Enigma was only one of a family of machine ciphers, it has attracted the most interest because of the exciting stories of the "duels" between the machine and, first, the Polish and, then, the British codebreakers. The story of the solution of Enigma began to become visible in 1974 with the publication of *The Ultra Secret* by F. W. Winterbotham [**29**]. Since that time much has been written about Enigma and the duel. Because of the secret nature of military cryptography and cryptanalysis, that story is often muddled and contradictory, but there is a clear trail from Arthur Scherbius' 1918 patent of a machine designed to protect commercial communications to the German military Enigma of World War II.

Here is how Enigma works. The Enigma machine consists of four visible components: a keyboard, a plugboard, a rotor system, and a lampboard. (See the front cover and Figure 1.) Enigma has both electric and mechanical parts. The executive summary of its operation is that the operator pushes a plaintext letter on the keyboard and the corresponding ciphertext letter is lighted on the lampboard.

Forget for a moment about the mechanical part of Enigma and follow the electrical action from the keyboard to the lampboard in Figure 2.

When the operator pushes a key on the keyboard (`A` is the key in the diagram), an electrical current passes from the key to the plugboard. The plugboard looks like an old telephone switchboard. There are 26 sockets—one for each letter of the keyboard.

Throughout the war, the Enigma machine evolved and the methods for using it changed. Different branches of the German military used different models of the machine, and the same model was used in a different manner by different branches. So, a description of how Enigma operated is dependent on who was using it and when they were using it. This description applies to the Enigma that the Polish mathematicians were attacking in 1932.

When the Polish mathematicians began their attack on Enigma, six plugs were in use. Each plug would connect (in a way prescribed by the key) one letter on the plugboard to another. The effect of the plugboard was to swap six pairs of letters and let

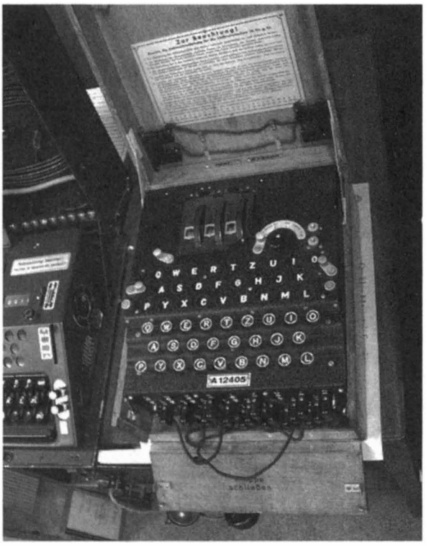


Figure 1 Closed Enigma

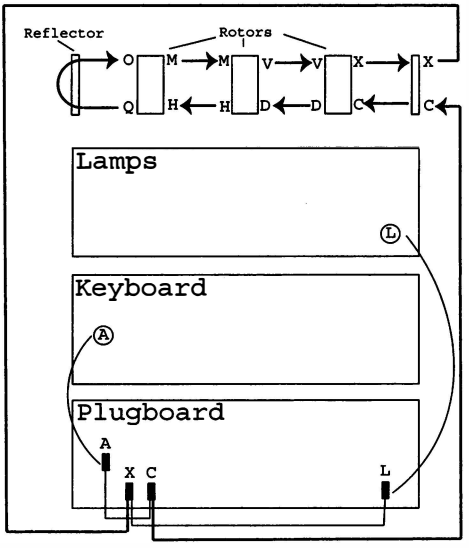


Figure 2 Enigma Diagram

the remaining 14 letters pass through unchanged. The plugboard consisted of six transpositions; 14 letters were fixed by the plugboard permutation. Later in the war, more plugs were used. In the diagram, the plugboard permutation includes the transposition `(AC)`; so, `A` is replaced by `C` by the plugboard. (Not all versions of Enigma had a plugboard.)

After passing through the plugboard (*Steckerbrett*, steckerboard), the electrical charge passed into the rotor system. In 1932, the rotor system consisted of three rotors and a reflector. Each rotor permuted the letters of the alphabet. The right-hand side of each rotor had 26 spring-loaded input terminals arranged around the disk; the left-hand side had 26 flat circular output terminals. Each input was wired to an output. The wiring determined the permutation. At the time that the Polish mathematicians began attacking Enigma, the machine had only three rotors; later the machine had as many as eight rotors from which either three or four were installed depending on the type of Enigma in use. Rejewski and the other Polish mathematicians did not know the wirings of the rotors. As we will see later, one of Rejewski's remarkable feats was his determination of the rotor wirings from intercepted messages alone. The three rotors were labelled I, II, and III; the labels identified the rotors but did not correspond to the positions of the rotors in the machine. When placing the rotors in Enigma, all six orderings of the three rotors were possible. In Figure 2, the rotors have been installed in the order I, III, II. The permutations accomplished by the three rotors are:

Rotor I `(aeltphqxru)(bknw)(cmoy)(dfg)(iv)(jz)(s)`
Cycles 10 4 4 3 2 2 1

Rotor II `(a)(bj)(cdklhup)(esz)(fixvyomw)(gr)(nt)(q)`
Cycles 8 7 3 2 2 2 1 1

Rotor III `(abdhpejt)(cflvmzoyqirwukzsg)(n)`
Cycles 17 8 1

Following the diagram, the electrical charge enters the rotor system as `C`. `C` enters the right-hand rotor and exits as `D`, `D` enters the middle rotor and exits as `H`, `H` enters

the left-hand rotor and exits as Q, and then Q enters the reflector at the left of the rotor system.

The reflector was "half a rotor." There were only 26 contacts on the right-hand side of the reflector. Internally, the 26 contacts were joined in pairs by wires to create a permutation consisting of 13 disjoint transpositions. At the time that we are considering, Enigma had just one reflector (reflector A). Its wiring was also not known to the Poles. It creates the following permutation:

Reflector A (ae)(bj)(cm)(dz)(fl)(gy)(hx)(iv)(kw)(nr)(oq)(pu)(st)

In the diagram, Q enters the reflector and exits as O.

Then the electrical charge passes backwards through the rotor system. The O enters the left-hand rotor, passes backwards through it, and exits as M. Then M enters the middle rotor, passes backwards through it, and exits as V. Next V enters the right-hand rotor, passes backwards through it, and exits as X.

The X then passes through the plugboard where it is changed to L, and lamp L lights. The operator would substitute ciphertext L for plaintext A.

This is an unduly complicated way to do a single permutation, but the point of the process is that the mechanical portion of Enigma allows for the generation of a long sequence of different permutations. Each time that a letter on the keyboard is pressed, before enciphering begins, the right-hand rotors turns one letter forward. The output side of the right-hand rotor has a notch that causes the middle rotor to turn forward. Like the odometer of a car, the middle rotor will turn forward one letter once during every 26 turns of the right-hand rotor. Similarly, there is a notch on the output side of the middle rotor that causes the left-hand rotor to turn forward one letter once during every 26 turns of the middle rotor. The theoretical maximum of $26^3 = 17576$ permutations is not actually achieved by Enigma because the mechanical movement of the rotors is such that the middle rotor can "double step"—it can rotate forward on two subsequent presses on the keyboard [**6**]. So, $26 \times 25 \times 26 = 16900$ keys can be pressed on the keyboard before the rotor system returns to the initial permutation. For a given setup of Enigma, 16900 substitution ciphers are generated in order; the period of Enigma is 16900.

The Polish Mathematicians

The King hath note of all that they intend,
By interception, which they dream not of.
—Henry V, Act II, Scene 1[3]

In the 1930s, in direct contradiction of the Versailles Treaty of 1919, Germany was rearming and was looking to reclaim its "lost" territories in the east—territories that were at that time part of Poland. The nervous Poles followed the German buildup by monitoring German radio transmissions. But the Germans had learned from their cryptological mistakes of World War I and were using better encryption—they were using Enigma. Unfortunately for Poland, "there were few persons adept at cryptology in Poland at this time." [**14**, p. 2]

To solve the problem of the lack of cryptologists, in 1929, the Polish government selected some mathematics students from Poznań University to participate in a cryptology course. Poznań was selected because of its location in an area where students

[3]This quote appears at the beginning of "The History of Hut 8, 1939–1945" by Patrick Mahon. Mahon served in Hut 8 (German Naval cryptanalysis) at Bletchley Park from 1941 until the end of the war; he was director of Hut 8 from 1944 until the end of the war. Alan Turing was the first director of Hut 8.

would be German speakers. Why mathematics students were selected is not clear, but, in a manner that was similar to the later recruiting done by the British for Bletchley Park, the Polish codebreakers were recruited by teachers and colleagues.

> Well, one day or one evening, I don't remember which, one of the younger mathematics students came up to me and said that on such-and-such a day, at such-and-such an hour, Professor [Zdzisław] Krygowski [director of Poznań University's Mathematics Insitute] wanted me to come to the Institute. This student had some sort of list, and he would go and tell each of the persons on the list about this. Not everyone was invited, only a certain number [of] selected students. What the criteria were, I can only guess .... I expect it wasn't Professor Krygowski who selected the students but rather Section II [the Intelligence Section of the Polish General Staff] that had made the selection. Probably there had been correspondence between Section II and Professor Krygowski, and on the basis of this correspondence Professor Krygowski had given them a list of all the third- and fourth-year students ... who were close to graduating, and then Section II had by its own methods conducted some kind of selection. In any case, not all the students were selected .... Marian Rejeweski [**14b**, p. 229].

Among the students who were selected were Jerzy Różycki, Henryk Zygalski, and Marian Rejewski.

On March 1, 1929, Rejewski (who is pictured on the front cover) received his master of philosophy in mathematics. Without having completed the cryptology course, because of an interest in actuarial mathematics, he went to Göttingen for a period of training. He returned to Poznań in October, 1930, and took a position as a teaching assistant. He also began work at the Poznań office of the Polish Cipher Bureau [Biuro Szyfrów, BURE-oh SHIF-roof].

During the Summer of 1932, the Poznań office was disbanded and Rejewski, Różycki, and Zygalski (the latter two had just graduated) became employees of the Cipher Bureau in Warsaw.

So begins the story of the Polish mathematicians and their duel with the Enigma machine. The most important of these was Rejewski, and in what follows we will focus on two applications of the theory of permutations to the attack on Enigma—determining the order of the Enigma rotors and determining the wiring of the Enigma rotors.

Setting Up Enigma

> *When two Enigma machines are set to the same key and their three wheels are in the same positions, the electrical connections through their steckerboards and scramblers will produce the same thirteen pairings of the twenty-six letters of the alphabet. ... Thus, if pressing letter-key* K *on one of the machines causes lamp* P *to be lit, then pressing letter-key* P *on the other machine will cause lamp* K *to be lit.* [**27**, p. 45]

Two Enigma operators could communicate only if their Enigma machines were set up using the same key. Daily keys were provided to the operators in a book, for example, for a month at a time. There were several settings which made up the Enigma key. In 1932, the following made up the key.

Plugboard: The key specified which 6 pairs of letters were to be connected on the plugboard. For example, `CO DI FR HU JW LS`.

Rotor order: The key specified the order in which the rotors were placed in the rotor system (from left to right). For example, I III II.

Ring setting: There was a ring around the circumference of each rotor on which the letters of alphabet A, B, . . . , Z or the numbers 01, 02, . . . , 26 were engraved. This ring could be rotated around the circumference and then held in place with a pin. The ring setting of the key indicated the letter of the alphabet on the ring that corresponded to the position of the pin. For example, P K M. The purpose of the ring setting was to set the letters on the ring with respect to the internal wiring of the rotor. The permutations that were given in Section 3 for each rotor assume that the ring setting for each is A. Another effect was to position the turnover notch. The notch was in a fixed position on the left side of each rotor. Changing the ring setting changed the position of the turnover with respect to the internal wiring of the rotors.

Groundsetting: This portion of the key specified the position of each rotor at the beginning of sending or receiving a transmission. The groundsetting indicated which letter on each ring should be visible in the windows above the three rotors, for example, N K U. These settings made up the key.

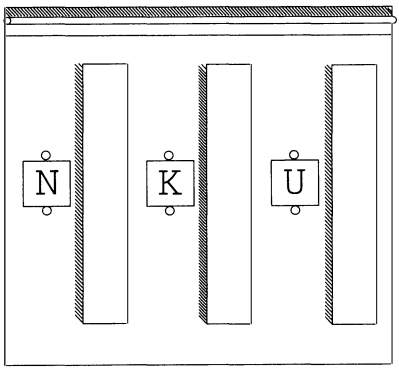


Figure 3 Enigma Rotor Cover closed with setting at NKU

The Number of Enigma Keys

[If] a man were able to adjust, day and night, a new key at every minute, it would take him 4000 years to try all those possibilities through on[e] after another.

Mid-1920s Enigma sales brochure reprinted in the July 2001 *Cryptologia.* See [**28**, p. 252].

The security of Enigma depends on its having a large key space. The size of the keyspace equals the number of possible plugboard settings × the number of possible rotor orders × the number of possible ring settings × the number of possible ground settings.

The number of possible plugboard settings: Assume that n plugs are being used. There are

$$\frac{[26 \times 25] \times [24 \times 23] \times [22 \times 21] \times \cdots \times [(26 - 2n + 2) \times (26 - 2n + 1)]}{2^n \times n!}$$

ways to connect n plugs into the plugboard. Here is a table which shows the number of connections for each of the possible number of plugs.

n	Number of connections	n	Number of connections
0	1	7	1,305,093,289,500
1	325	8	10,767,019,638,375
2	44,850	9	53,835,098,191,875
3	3,453,450	10	150,738,274,937,250
4	164,038,875	11	205,552,193,096,250
5	5,019,589,575	12	102,776,096,548,125
6	100,391,791,500	13	7,905,853,580,625

When the Poles began to attack Enigma, six plugs were in use. So, there were 100,391,791,500 ways to connect the six plugs into the plugboard. Later the Germans used ten plugs.

The number of possible rotor orders: There are six ways to arrange the three rotors in order in the rotor system.

The number of possible ring settings: Only the positions of the notches on the right-hand and middle rotors contributed to the cryptographic security of Enigma. So, we will say that there are $26^2 = 676$ possible ring settings.

The number of possible groundsettings: There are $26^3 = 17576$ choices of the letters to appear in the windows.

So, effectively, the number of possible keys was

$$100,391,791,500 \times 6 \times 676 \times 17576 = 7,156,755,732,750,624,000$$

which would seem to be secure enough.

Enigma Ciphers

> ... *we shall see that cryptography is more than a subject permitting mathematical formulation, for indeed it would not be an exaggeration to state that* abstract *cryptography is* identical *with abstract mathematics.*
>
> A. A. Albert [**1**, p. 903]

There are $26! = 403,291,461,126,605,635,584,000,000$ simple substitution cipher permutations, but there are many fewer possible Enigma substitution cipher permutations. Consider the diagram (Figure 4) "Enigma's functional circuit" that is based upon a figure in [**14e**, p. 274] and uses Rejewski's notation. S represents the plugboard (*Steckerbrett*); N represents the right-hand, or fast, rotor; M represents the middle rotor; L represents the left-hand, or slow, rotor; and R represents the reflector.

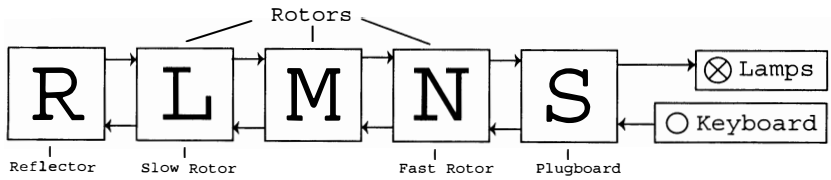


Figure 4 Enigma's functional circuit

Think of S, N, M, L, and R as permutations. There is one permutation missing from Enigma's functional circuit. There should be a permutation P: N $\leftarrow$ S corresponding to the motion of the fast rotor which moves forward one letter every time a key is pressed. P = (abcdefghijklmnopqrstuvwxyz). Rejewski's attack on Enigma uses only the first six ciphertext letters; so, he assumes that the middle and left-hand rotor do not move, but if they do move, his method will not work. Because the middle rotor turns only once in every 26 turns of the fast rotor, it is reasonable to assume that the middle rotor and the left-hand rotor do not move during the first six encryptions. For the first enciphered letter, the permutation is

$$\mathrm{SPNMLRL^{-1}M^{-1}N^{-1}P^{-1}S^{-1}} = \mathrm{(SPNML)R(SPNML)^{-1}}.$$

Rejewski composes his permutations from left to right, and we will follow his notation.

For the second enciphered letter, the permutation is

$$\mathrm{SP^2NMLRL^{-1}M^{-1}N^{-1}P^{-2}S^{-1}} = \mathrm{(SP^2NML)R(SP^2NML)^{-1}}.$$

Whether the middle and left-hand rotors move or not, an Enigma permutation is always a conjugate of the reflector. So, an Enigma permutation is always a product of 13 disjoint transpositions. There are no more than

$$\frac{\binom{26}{2}\binom{24}{2}\cdots\binom{2}{2}}{13!} = 7{,}905{,}853{,}580{,}625$$

such permutations, many fewer than the 26! possible simple substitution permutations.

The fact that every Enigma permutation is a product of 13 disjoint transpositions is what permits Enigma to encipher and decipher in the same mode. Every Enigma permutation is self-reciprocal.

But, being self-reciprocal can also be a weakness. The reflector permutation guarantees that every Enigma permutation is self-reciprocal, but it also guarantees that no letter can be enciphered as itself. The latter was useful information for British cryptanalysts. The cryptanalysts who attacked Enigma would know, for example, that ciphertext T did not correspond to plaintext t. The same rule usually applies to cryptograms that appear in newspapers (so-called "aristocrats")—no letter ever substitutes for itself. With such a rule, we would know, for example, that the trigraph JFE could not represent plaintext the.

The Entry Permutation

Q		W		E		R		T		Z		U		I		O
	A		S		D		F		G		H		J		K	
P		Y		X		C		V		B		N		M		L

The Enigma Keyboard[4]

There is another permutation that was not considered in the "Enigma functional circuit"—the entry permutation. For the original, commercial Enigma, the entry permutation corresponded to the order of the keys on the Enigma keyboard q $\rightarrow$ A, w $\rightarrow$ B, e $\rightarrow$ C, . . . :

Output from Plugboard:	abcdefghijklmnopqrstuvwxyz
Entry into Rotors:	JWULCMNOHPQZYXIRADKEGVBTSF.

[4]The arrangement of the keys on an Enigma keyboard differs slightly from the arrangement on a keyboard today.

In the process of solving for the wirings of the rotors, Rejewski assumed that the entry permutation—the permutation from the plugboard into the right-hand rotor—was the same as for the commercial Enigma, but permutations that should have been similar were not.

> ... it finally occurred to me [Marian Rejewski] that the cause of my failure may have been merely a mistaken assumption as to the connections of the entry drum. [**14d**, p. 257]

Dillwyn ("Dilly") Knox, a British codebreaker who was also attacking Enigma, was also stumped by the entry permutation.

> I [Marian Rejewski] have the fullest grounds to believe that the British cryptologists were unable to overcome the difficulties caused by the connections in the entry drum. When the meeting of Polish, French, and British cipher bureau representatives took place in Poland in July 1939, the first question that the British cryptologist Dillwyn Knox asked was: What are the connections in the entry drum? Knox's niece, Penelope Fitzgerald states in her book *The Knox Brothers*, published in 1978, that Knox was furious when he learned how simple it was.
>
> What ... were the connections in the entry drum? It turned out later that they can be found by deduction, but in December 1932, or perhaps in the first days of 1933, I obtained those connections by guessing. I assumed that, since the keyboard keys were not connected with the successive contacts in the entry drum in the order of the letters on the keyboard, then maybe they were connected in alphabetical order; that is, the permutation caused by the entry drum was an identity and need not be taken into account at all. The hypothesis turned out to be correct. [**14d**, pp. 257 & 258]

The permutation from the plugboard to the rotor system was:

Output from Plugboard:	`abcdefghijklmnopqrstuvwxyz`
Entry into Rotors:	`ABCDEFGHIJKLMNOPQRSTUVWXYZ.`

Peter Twinn was one of the first mathematicians recruited to Bletchley Park. Twinn was working with Knox when the Poles revealed the secret of the wiring from the keyboard to the entry drum. In *The Telegraph* [November 17, 2004] obituary of Twinn, he is quoted as saying:

> I know in retrospect it sounds daft. It was such an obvious thing to do. Rather a silly thing, that nobody, not Dilly Knox, not Alan Turing, ever thought it worthwhile trying.

Sometimes it is good to guess.

Rotor Order and Groundsetting

> *The double encipherment of each text setting ... was a gross error. It enabled us to attack the million-odd combinations of wheel order and ring settings without bickering about the vast number of steckerboard cross-connections in which the German experts had placed their trust ....*
>
> Gordon Welchman [**27**, p. 164]

The quote refers to techniques used by the codebreakers at Bletchley Park, but it also applies to the power of Rejewski's methods. Rejewski was able to discover patterns in the Enigma messages and apply the theory of permutations to defeat the plugboard and determine the rotor order and groundsetting.

Recall that the effective number of Enigma keys is

$$100,391,791,500 \times 6 \times 676 \times 17576 = 7,156,755,732,750,624,000$$

where 100,391,791,500 corresponds to the number of possible connections of the plugboard. The product $6 \times 17576 = 105,456$ corresponds to the number of possible rotor orders and groundsettings. (We will, as Rejewski did at this point, ignore the $26^2 = 676$ ring settings. Recall that the ring settings set the position of the turnovers [and we are assuming that turnover did not occur], and the ringsettings set the relation between the letters on the circumferences of the wheels with respect to the internal wiring [and Rejewski had other methods to determine that relationship].) Rejewski was able to reduce the large number of keys to the smaller 105,456, which is not small but is more manageable than 7,156,755,732,750,634,000.

Recall that Enigma was designed to generate a long sequence of simple substitution ciphers. The goal was to defeat frequency analysis by effectively using a different permutation to encipher each plaintext letter of a message. That is a good idea. But, there is still a problem, and the problem is called "depth." Say that every Enigma operator sets up his machine according to the instructions and begins every message with the same groundsetting—the same letters appearing in the windows on top of the Enigma. Permutation P_1 will encipher the first letter of every message, P_2 will encipher the second letter of every message, ... , P_{50} will encipher the fiftieth letter of every message, ... This would happen for every message enciphered by every operator. If it were possible to intercept a large number of messages, say 100, then the first letter of each message would have been enciphered with P_1. If we stripped off the first letter of each message we would have 100 ciphertext letters each enciphered with the same simple substitution cipher. We could apply frequency analysis (perhaps, modified for this letter to use frequencies of initial letters) to this collection and have a chance of determining P_1. And, we could proceed similarly for the second letter of each message, the third letter of each message, etc. This is called depth. Although there might not be repetition of ciphers within a message, there is repetition within the collection of 100 messages.

Prior to World War I, most cryptanalysis was done by lone cryptanalysts working in "Black Chambers" attacking individual ciphertext messages. The use of radio in World War I changed the nature of cryptanalysis. Suddenly there were hundreds or thousands of messages that could be attacked by teams of cryptanalysts.

German Enigma procedures were designed to defeat the problem of depth. At the time that the Poles first encountered Enigma, Enigma procedures required that each Enigma message be enciphered using a different setting of the rotors—different letters appearing in the windows on top of the machine. It was left to each operator to determine the three-letter message setting. If each message were enciphered with a different message setting, depth would not occur. But, how would the message setting be sent from the sender to the receiver? How would the key be distributed? The solution that the Germans decided upon was to use Enigma to encipher the message setting—the three-letter message setting was enciphered using the groundsetting. Because radio transmission was subject to garbling, the operators sent the message setting twice. So, preceding each ciphertext message were six letters that were two copies of the message setting enciphered with the first six permutations beginning with the groundsetting. Rejewski calls these first six permutations A, B, C, D, E, and F.

For example, say we have decided our message setting will be NKU. After setting up Enigma according to the instructions given in the key, we first encipher nkunku. Let us assume that these letters encipher to JHNQBG. These six letters would be sent in the preamble to the ciphertext. When the receiving operator received the transmission, he would set his machine according to the instructions given in the key. Beginning with the groundsetting, he would press the keys JHNQBG. The lamps nkunku should light. The operator would then set his rotors to NKU, and enter the ciphertext; the plaintext message should appear.

It was in these enciphered double message settings that Rejewski discovered a pattern.

Rejewski would not have known the message setting NKU, but he would have known that the first letter, say ?, of the message setting was changed to J by permutation A and changed to Q by permutation D. A: ? $\rightarrow$ J and D : ? $\rightarrow$ Q . Because Enigma ciphers are self-reciprocal, we know that AD: J $\rightarrow$? $\rightarrow$ Q.

The composition AD changes J to Q. Similarly, BE changes H to B, and CF changes N to G.

Now what remains is to collect enough ciphertext messages.

> If we have a sufficient number of messages (about eighty) for a given day, then, in general, all the letters of the alphabet will occur in all six places at the openings of the messages. Marian Rejewski [**14e**, p. 274]. Cf. [**14d**, p. 234].

Here is a list of 65 enciphered double message settings AUQ AMN, ... , ZSJ YWG taken from [**2**, p. 390].

AUQ	AMN	IND	JHU	PVJ	FEG	SJM	SPO	WTM	RAO
BNH	CHL	JWF	MIC	QGA	LYB	SJM	SPO	WTM	RAO
BCT	CGJ	JWF	MIC	QGA	LYB	SJM	SPO	WTM	RAO
CIK	BZT	KHB	XJV	RJL	WPX	SUG	SMF	WKI	RKK
DDB	VDV	KHB	XJV	RJL	WPX	SUG	SMF	XRS	GNM
EJP	IPS	LDR	HDE	RJL	WPX	TMN	EBY	XRS	GNM
FBR	KLE	LDR	HDE	RJL	WPX	TMN	EBY	XOI	GUK
GPB	ZSV	MAW	UXP	RFC	WQQ	TAA	EXB	XYW	GCP
HNO	THD	MAW	UXP	SYX	SCW	USE	NWH	YPC	OSQ
HNO	THD	NXD	QTU	SYX	SCW	VII	PZK	YPC	OSQ
HXV	TTI	NXD	QTU	SYX	SCW	VII	PZK	ZZY	YRA
IKG	JKF	NLU	QFZ	SYX	SCW	VQZ	PVR	ZEF	YOC
IKG	JKF	OBU	DLZ	SYZ	SCW	VQZ	PVR	ZSJ	YWG

Consider the first and fourth letters of each indicator. We can notice that the composition cipher AD replaces A by A, B by C, C by B, D by V, E by I, F by K, G by Z, etc. The composition cipher AD is

abcdefghijklmnopqrstuvwxyz
ACBVIKZTJMXHUQDFLWSENPRGOY.

In terms of disjoint cycles,

$$\mathrm{AD} = (\mathrm{a})(\mathrm{bc})(\mathrm{dvpfkxgzyo})(\mathrm{eijmunqlht})(\mathrm{rw})(\mathrm{s}),$$

and the lengths of the cycles are 10 10 2 2 1 1.

Similarly,

$$\mathrm{BE} = (\mathrm{axt})(\mathrm{blfqveoum})(\mathrm{cgy})(\mathrm{d})(\mathrm{hjpswizrn})(\mathrm{k}),$$

and the lengths of the cycles are 9 9 3 3 1 1.

$$\mathrm{CF} = \texttt{(abviktjgfcqny)(duzrehlxwpsmo)},$$

and the cycles are 13 13.

Rejewski saw that the disjoint cycles assume a very characteristic form, "generally different for each day [i.e., for each groundsetting] . . . ." [**14e**, p. 274] Furthermore, Rejewski realized that the cycle structure is not affected by the plugboard. For example, consider

$$\mathrm{A} = \mathrm{SPNMLRL}^{-1}\mathrm{M}^{-1}\mathrm{N}^{-1}\mathrm{P}^{-1}\mathrm{S}^{-1} \quad \text{and} \quad \mathrm{D} = \mathrm{SP}^4\mathrm{NMLRL}^{-1}\mathrm{M}^{-1}\mathrm{N}^{-1}\mathrm{P}^{-4}\mathrm{S}^{-1}.$$

$$\begin{aligned}\mathrm{AD} &= \left(\mathrm{SPNMLRL}^{-1}\mathrm{M}^{-1}\mathrm{N}^{-1}\mathrm{P}^{-1}\mathrm{S}^{-1}\right)\left(\mathrm{SP}^4\mathrm{NMLRL}^{-1}\mathrm{M}^{-1}\mathrm{N}^{-1}\mathrm{P}^{-4}\mathrm{S}^{-1}\right)\\ &= \mathrm{S}\left(\mathrm{PNMLRL}^{-1}\mathrm{M}^{-1}\mathrm{N}^{-1}\mathrm{P}^{-1}\right)\mathrm{S}^{-1}\mathrm{S}\left(\mathrm{P}^4\mathrm{NMLRL}^{-1}\mathrm{M}^{-1}\mathrm{N}^{-1}\mathrm{P}^{-4}\right)\mathrm{S}^{-1}\\ &= \mathrm{SP}_1\mathrm{P}_4\mathrm{S}^{-1}\end{aligned}$$

where $\mathrm{P}_1 = \mathrm{PNMLRL}^{-1}\mathrm{M}^{-1}\mathrm{N}^{-1}\mathrm{P}^{-1}$ and $\mathrm{P}_4 = \mathrm{P}^4\mathrm{NMLRL}^{-1}\mathrm{M}^{-1}\mathrm{N}^{-1}\mathrm{P}^{-4}$ are each determined only by the rotor order and groundsetting. Because of the theorem from elementary permutation theory that the disjoint cycle structure of a permutation and a conjugate of the permutation are the same, the disjoint cycle structure of AD is the same as it would be if there were no plugboard; the effect of the plugboard has been nullified!

Similarly, the disjoint cycle structure of BE and CF is not affected by the plugboard. Rejewski can determine the rotor order and ground setting without considering the 100,391,791,500 possible plugboard connections. Momentarily, he also ignored the 676 ring settings, and he is, therefore, left with "only" the possible $6 \times 17576 = 105{,}456$ rotor orders and groundsettings.

Rejewski assumed that the middle (and left-hand) rotor did not turn during these six permutations. Because the middle rotor turned only once during 26 turns of the right-hand rotor, this was a reasonable assumption. If a turnover did occur, his method would not work.

For each of the 105,456 settings, the Poles determined the characteristic disjoint cycles. To do this they devised a machine called a cyclometer. (See Figure 6, p. 261.)

The cyclometer consisted of two sets of Enigma rotors. One of the six rotor orders (e.g., I III II) was selected and both sets of Enigma rotors were arranged in that order. Then the first set of rotors was set to a groundsetting (e.g., NKU), and the second set of rotors was stepped three positions beyond the groundsetting (NKX). So, the rotors were set up as if they were permutations A and D. Again, it was assumed that the middle rotor did not turn during the six indicator permutations.

A charge was applied to one of the letters, say A. The charge passed through the first rotor system and the output of the first rotor system passed through the corresponding lamp, say N. Then N entered the second rotor system and the output of the second rotor system, say J, passed through the corresponding lamp and entered the first rotor system. This process continues until the charge returns to A. The diagram[5] shows the situation when (ajqe) is a cycle of the permutation created by the cyclometer. Because that permutation is conjugate to AD, AD also contains a 4-cycle. Notice that applying current to any of A, J, Q, or E would result in the same cycle. Also notice that this 4-cycle results in the lighting of eight lamps; G, N, H, and S also light and correspond to another 4-cycle of the permutation AD. If a charge were applied to G, N, H, or S, the same lamps would light.

[5]This diagram is based upon an example and diagram in [**4**]

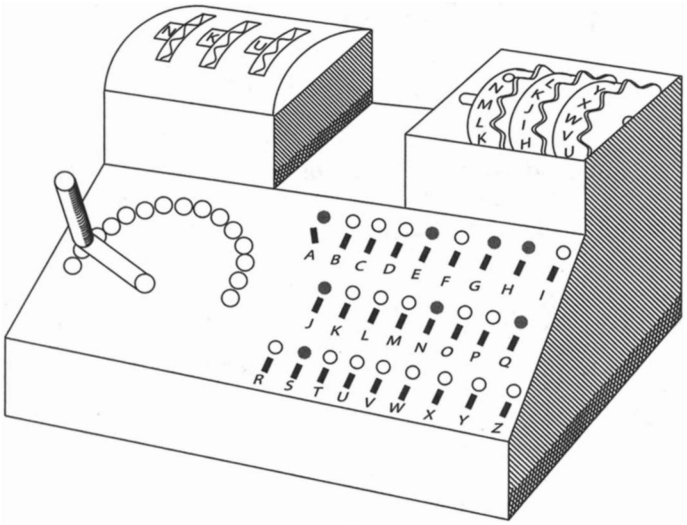


Figure 5 Cyclometer

The cyclometer is equipped with a rheostat so that the amount of current in the circuit can be varied according to the number of lamps that are lit. (In case many lamps are lit, the current can be increased to strengthen the light coming from the lamps; if few lamps are lit, the lower current would be less likely to burn out the filaments.)

The eight lamps that are lit told the Poles that AD contained two 4-cycles.

Then a charge was applied to a letter other than A, J, Q, E, G, N, H and S; and another pair of disjoint cycles was determined. This process was continued until the lengths of all of the disjoint cycles of AD were known.

Then both rotors were moved forward one position—to NKV on the first rotor and NKY on the second rotor. The permutation that the cyclometer now creates is conjugate to BE, and the lengths of the disjoint cycles of BE are determined. Then each rotor is advanced forward one more position to create a permutation conjugate to CF, and the lengths of the disjoint cycles of CF are determined.

The Poles catalogued the lengths of the disjoint cycles to all $6 \times 17576 = 105{,}456$ possible rotor orders and groundsettings. These lists of the lengths of disjoint cycles were called the characteristics of the permutation. Apparently no copies of their catalogue still exist; so, it is not known how the Polish mathematicians ordered the characteristics.

The mapping from rotor orders and groundsettings to characteristics is not one-to-one. Several rotor orders and groundsettings can result in the same characteristic.

Rejewski describes the use of the cyclometer:

> One had to note on a card the position of the drums and the number of bulbs that were lit, and to order the cards themselves in a specified way, for example by the lengths of the cycles.
>
> This job took a long time, over a year, since we carried it out along with our normal work at reconstructing daily keys using the grille [another method of cryptanalysis used by the Poles]. Once all six card catalogues [one for each of the six possible orders of the rotors] were ready, though, obtaining a daily key was usually a matter of ten to twenty minutes. The card told the drum positions [the letters appearing in the window on the top of the Enigma], the box from which the card had been taken told the drum sequence [the ordering of the rotors], and

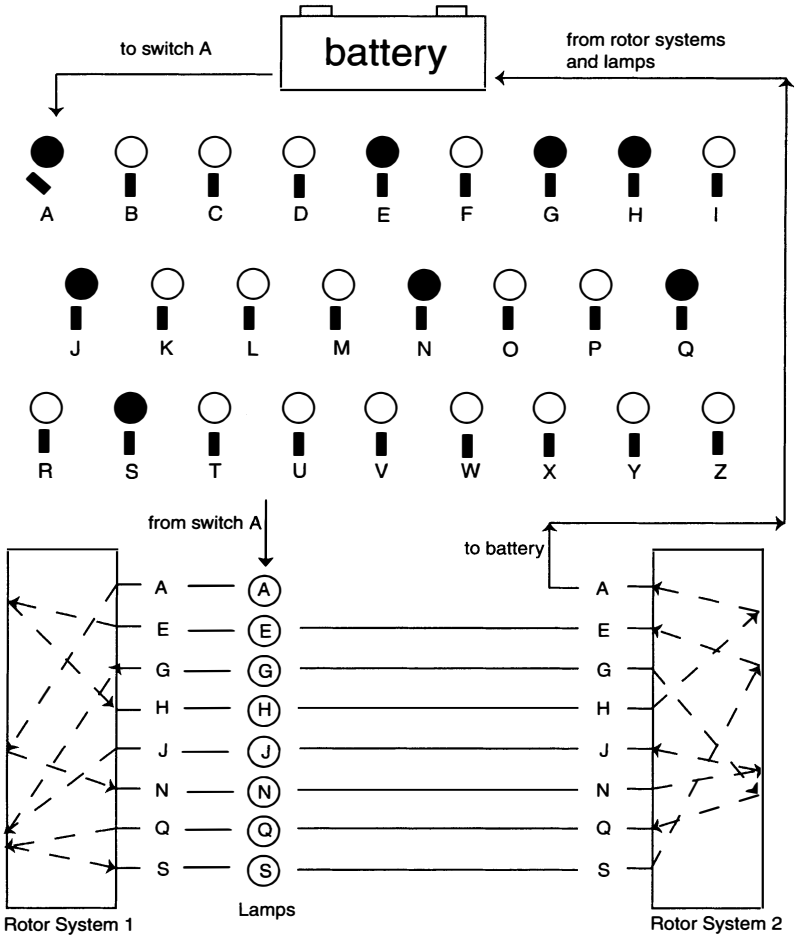


Figure 6 Cyclometer diagram

the permutation S [the permutation resulting from the plugboard] was obtained by comparing the letters of the cycles of permutations AD, BE, CF, which were obtained by tapping on the machine's keyboard. [**14d**, pp. 263 & 264]

How far from being one-to-one is the mapping from rotor orders and groundsettings to characteristics? Carter conducted a modern reconstruction of a portion of a catalogue (see [**4**, p. 10f]); he used one rotor order and reflector B, which was not the reflector in use when the Poles were assembling their catalogue.[6] Carter comments:

It now seems apparent that the use of the catalogue to determine the daily starting positions, from the composite cycle pattern could not have been an entirely straightforward procedure. In bad cases, the number of possibilities given by the catalogue would have been daunting and, if attempted, would have required the subsequent checking of large numbers of possible alternative starting positions. For the majority of patterns however, the starting positions would have been

[6]The wiring of reflector A had not yet been rediscovered when Carter constructed his catalogue. The wiring was reconstructed and published in 2000.

found immediately from the catalogue, or at most after checking on only a few alternatives [**4**, p. 11].

Kuhl [**16**] considered all rotor orders and groundsettings and used reflector A, which was the reflector in use when the Poles were assembling their catalogue.

Soon after the Poles completed the catalogue there was a change in Enigma.

> Unfortunately, on 2 November 1937, when the card catalogue was ready, the Germans exchanged the reversing drum [reflector] that they had been using, which they designated by the letter A, for another drum, a B drum, and consequently, we had to do the whole job over again, after first reconstructing the connections in drum B, of course. Marian Rejewski [**14d**, p. 264]

Rejewski's Theorems

> *Nonetheless, the Polish mathematicians at B.S.-4* [Biuro Szyfrów-4, the German cipher office]—*thanks to the cycle principle discovered by Marian Rejewski, ... were able to quickly distinguish total chaos from the merely ostensible chaos that resulted when initially ordered impulses flowed through the machine's innards.* [**14**, pp. 42 and 43]

In addition to the theorem that conjugation preserves disjoint cycle structure, Rejewski in his two papers [**14d**] and [**14e**] explicitly states four theorems and uses another.

THEOREM 1. (THEOREM ON THE PRODUCTS OF TRANSPOSITIONS) *If two permutations of the same degree consist solely of disjoint transpositions, then their product will consist of disjoint cycles of the same length in even numbers.*

He argues its proof as follows:

$X = (a_1a_2)(a_3a_4)(a_5a_6)\ldots(a_{2k-3}a_{2k-2})(a_{2k-1}a_{2k})$
and $Y = (a_2a_3)(a_4a_5)(a_6a_7)\ldots(a_{2k-2}a_{2k-1})(a_{2k}a_1)$,
then $XY = (a_1a_3a_5\ldots a_{2k-3}a_{2k-1})(a_{2k}a_{2k-2}\ldots a_6a_4a_2)$.

"If, in this manner, we have not exhausted all the letters in the permutation, we continue our procedure until we have done so." [**14e**, p. 278]

Composing permutations is a routine activity in abstract algebra courses, but what Rejewski needed to do was factor the permutations AD, BE, and CF.

THEOREM 2. (CONVERSE TO THE THEOREM ON THE PRODUCT OF TRANSPOSITIONS) *If a permutation of even-numbered degree includes cycles of the same length in even numbers, then this permutation may be regarded as a product of two permutations, each consisting solely of disjoint transpositions.*

Recall that each of AD, BE, and CF satisfy the conditions of this theorem. Its proof is immediate from what was noted above.

Given $XY = (a_1a_3a_5\ldots a_{2k-3}a_{2k-1})(a_{2k}a_{2k-2}\ldots a_6a_4a_2)$,
then we can write $X = (a_1a_2)(a_3a_4)(a_5a_6)\ldots(a_{2k-3}a_{2k-2})(a_{2k-1}a_{2k})$
and $Y = (a_2a_3)(a_4a_5)(a_6a_7)\ldots(a_{2k-2}a_{2k-1})(a_{2k}a_1)$.

Rejewski notes two other results that follow from the proof of his *Theorem on the Product of Transpositions*:

THEOREM 3. *Letters entering into one and the same transposition of permutation X or Y, enter always into two different cycles of the permutation XY.*

and

THEOREM 4. *If two letters found in two different cycles of the same length of the permutation XY belong to the same transposition, then the letters adjacent to them (one to the right, the other to the left) also belong to the same transposition.*

Lawrence [**20**] proves a generalization of Rejewski's factoring method.

Rejewski also notes one more fact about conjugation. Rejewski does not call this a theorem, but we will here. Say, we consider a conjugation of the permutation H.

THEOREM 5. *If* $H(i) = j$*; i.e.,* $H = (\ldots i\, j \ldots)$*; then* $T^{-1}HT = (\ldots T(i)T(j) \ldots)$*. Notice that this implies that* $H = (\ldots i\, j \ldots)$ *and* $T^{-1}HT = (\ldots T(i)\ T(j) \ldots)$ *have the same disjoint cycle decomposition.*

For a proof, Rejewski notes that $T(i)(T^{-1}HT) = i(HT) = H(i)T = T(j)$.

In particular, we note that this means that the entries of the permutations can be ordered so that

$$H = (\ \ldots\ldots\ \ i\ \ j\ \ \ldots\ldots\ )$$
$$T^{-1}HT = (\ \ldots\ \ T(i)\ \ \ T(j)\ \ \ \ldots\ )$$

which describes the permutation T in two-row notation.

These theorems are used by Rejewski to determine the wiring of the right-hand (fast) rotor using the disjoint cycle description of AD, BE, and CF.

Finding the Wiring of the Right-Hand Rotor—The Fast Rotor

> *Still working in isolation, Rejewski's next step was to develop a mathematical representation of the working Enigma machine. He was hoping that the knowledge of permutations A to F would enable him to work out the wiring of the wheels. He had reduced the problem to a set of six equations involving three unknown permutations, and he was wondering whether they could be solved, when, at just the right moment, he was given four documents from the German traitor Asche.*
>
> Gordon Welchman [**27**, p. 210]

Rejewski was also able to use the enciphered double indicators to determine the wiring of the right-hand (fast) rotor. This was accomplished by solving systems of equations that resulted from the patterns determined by the composed permutations AD, BE, and CF.

To see how Rejewski did this, we will closely follow his example [**14e**, p. 281f].

First, recall that Rejewski was able to determine the composed permutations provided that he had enough messages—provided that in the collection of 6-letter indicators each letter occurred at least once in each of the first three positions. Recall that we have determined the composed permutations to be:

AD = `(a)(bc)(dvpfkxgzyo)(eijmunqlht)(rw)(s)`
BE = `(axt)(blfqveoum)(cgy)(d)(hjpswizrn)(k)`
CF = `(abviktjgfcqny)(duzrehlxwpsmo)`.

Rejewski wants to factor these permutations into A, B, C, D, E, and F. His description of what he did is terse:

> We assume that thanks to the theorem on the product of permutations, combined with a knowledge of encipherer's habits, we know separately the permutations A through F. [**14e**, p. 282]

A = (as)(br)(cw)(di)(ev)(fh)(gn)(jo)(kl)(my)(pt)(qx)(uz)
B = (ay)(bj)(ct)(dk)(ei)(fn)(gx)(hl)(mp)(ow)(qr)(su)(vz)
C = (ax)(bl)(cm)(dg)(ei)(fo)(hv)(ju)(kr)(np)(qs)(tz)(wy)
D = (as)(bw)(cr)(dj)(ep)(ft)(gq)(hk)(iv)(lx)(mo)(nz)(uy)
E = (ac)(bp)(dk)(ez)(fh)(gt)(io)(jl)(ms)(nq)(rv)(uw)(xy)
F = (aw)(bx)(co)(df)(ek)(gu)(hi)(jz)(lv)(mq)(ns)(py)(rt).

It is easy to see that the factors do not violate any of the theorems, but how did Rejewski factor them?

Let us consider factoring AD = (a)(bc)(dvpfkxgzyo)(eijmunqlht)(rw)(s).

First, consider the two 1-cycles. From Theorem 2, if $(a_1)(a_2)$ in XY, (a_1a_2) appears in X and (a_2a_1) appears in Y.

(a)(s) appears in AD; so, (as) appears in both A and D.

Next, consider the two 2-cycles. From the Theorem 2, if $(a_1a_3)(a_4a_2)$ appears in XY, then $(a_1a_2)(a_3a_4)$ appears in X and $(a_2a_3)(a_4a_1)$ appears in Y.

AD contains two transpositions, and there are two possible orders of the elements within them: (bc)(rw) or (bc)(wr). [Although the order of the elements is not important to writing the permutation as a product of disjoint cycles, it is important to the factoring.]

Therefore, either (br)(cw) appears in A and (rc)(wb) appears in D, or (bw)(cr) appears in A and (wc)(rb) appears in D. There are two possibilities.

Finally, consider the two 10-cycles.

From Theorem 2, if $(a_1a_3a_5\ldots a_{2k-3}a_{2k-1})\,(a_{2k}a_{2k-2}\ldots a_6a_4a_2)$ appears in XY, then $(a_1a_2)\,(a_3a_4)\,(a_5a_6)\ldots(a_{2k-3}a_{2k-2})\,(a_{2k-1}a_{2k})$ appears in X and $(a_2a_3)\,(a_4a_5)\,(a_6a_7)\ldots(a_{2k-2}a_{2k-1})\,(a_{2k}a_1)$ appears in Y.

For (dvpfkxgzyo)(eijmunqlht), there are ten possible orders:

Order number 1:
AD =(dvpfkxgzyo)(eijmunqlht)
A = (dt)(hv)(pl)(fq)(kn)(xu)(gm)(zj)(yi)(oe)
D = (tv)(hp)(lf)(qk)(nx)(ug)(mz)(jy)(io)(ed)
⋮
Order number 3:
AD =(dvpfkxgzyo)(jmunqlhtei)
A = (di)(ve)(pt)(fh)(kl)(xq)(gn)(zu)(ym)(oj)
D = (iv)(ep)(tf)(hk)(lx)(qg)(nz)(uy)(mo)(jd)
⋮
Order number 10:
AD =(dvpfkxgzyo)(teijmunqlh)
A = (dh)(vl)(pq)(fn)(ku)(xm)(gj)(zr)(ye)(ot)
D = (hv)(lp)(qf)(nk)(ux)(mg)(jz)(iy)(eo)(td).

So there are $1 \times 2 \times 10 = 20$ possible factorizations of AD. Here Rejewski gets some help from the Enigma operators.

> ... it is a well-known phenomenon that man, as a being endowed with consciousness and memory, cannot imitate chance perfectly, and it is the cryptologist's task, among other things, to discover and make proper use of these deviations from chance.
>
> Marian Rejewski [**14d**, p. 254]

Just as our selection of NKU for our message setting earlier in this paper was not random, Enigma operators did not usually choose random 3-letter strings for their message settings. Often they used initials, patterns in rows or diagonals of the keyboard, etc. Rejewski was able to exploit his knowledge of the operators' habits to reduce the number of possible factorizations. Eventually he was able to arrive at the factorizations

A = (as)(br)(cw)(di)(ev)(fh)(gn)(jo)(kl)(my)(pt)(qx)(uz)
B = (ay)(bj)(ct)(dk)(ei)(fn)(gx)(hl)(mp)(ow)(qr)(su)(vz)
C = (ax)(bl)(cm)(dg)(ei)(fo)(hv)(ju)(kr)(np)(qs)(tz)(wy)
D = (as)(bw)(cr)(dj)(ep)(ft)(gq)(hk)(iv)(lx)(mo)(nz)(uy)
E = (ac)(bp)(dk)(ez)(fh)(gt)(io)(jl)(ms)(nq)(rv)(uw)(xy)
F = (aw)(bx)(co)(df)(ek)(gu)(hi)(jz)(lv)(mq)(ns)(py)(rt).

In terms of the individual permutations of the Enigma circuit, we have

$A = SPNP^{-1}MLRL^{-1}M^{-1}PN^{-1}P^{-1}S^{-1}$
$B = SP^2NP^{-2}MLRL^{-1}M^{-1}P^2N^{-1}P^{-2}S^{-1}$
$C = SP^3NP^{-3}MLRL^{-1}M^{-1}P^3N^{-1}P^{-3}S^{-1}$
$D = SP^4NP^{-4}MLRL^{-1}M^{-1}P^4N^{-1}P^{-4}S^{-1}$
$E = SP^5NP^{-5}MLRL^{-1}M^{-1}P^5N^{-1}P^{-5}S^{-1}$
$F = SP^6NP^{-6}MLRL^{-1}M^{-1}P^6N^{-1}P^{-6}S^{-1}$

where P is the entry permutation and

P = (abcdefghijklmnopqrstuvwxyz)
P^2 = (acegikmoqsuwy)(bdfhjlnprtvxz)
P^3 = (adgjmpsvybehknqtwzcfilorux)
P^4 = (aeimquycgkosw)(bfjnrzvdhlptx)
Etc.

Rejewski substitutes $Q = MLRL^{-1}M^{-1}$. This permutation is a factor of each of A, B, C, D, E, and F because Rejewski assumed that no turnover occurred during the double encipherment of the message setting; so, the middle and left-hand rotor are assumed to be fixed.

$A = SPNP^{-1}QPN^{-1}P^{-1}S^{-1}$
$B = SP^2NP^{-2}QP^2N^{-1}P^{-2}S^{-1}$
$C = SP^3NP^{-3}QP^3N^{-1}P^{-3}S^{-1}$
$D = SP^4NP^{-4}QP^4N^{-1}P^{-4}S^{-1}$
$E = SP^5NP^{-5}QP^5N^{-1}P^{-5}S^{-1}$
$F = SP^6NP^{-6}QP^6N^{-1}P^{-6}S^{-1}$

The unknowns are Q, S, N, and their inverses. Rejewski wants to determine N.

As it turned out, the Polish Cipher Bureau had information that made S, the plugboard permutation, known.

> ... I had a set of six equations with three unknowns—S, N, and Q. And just as I was wondering how to solve this set, quite unexpectedly on 9 December 1932, at just the right moment, I was given a photocopy of two tables of daily keys for September and October 1932.
>
> Now, the situation had changed radically. Since the key tables also contained the daily changes in the commutator connections, I could now regard S as known and transfer it ... to the left side of the set ...
>
> Marian Rejewski [**14d**, p. 256]

The French had purchased the information (along with other information about Enigma) from the German traitor Hans Thilo Schmidt (code name Asché). In 1932, German Enigma procedures called for changing the order of the three rotors once per quarter. Because September is in one quarter of the year and October is in the next, the information from Schmidt provided the plugboard connections when two different rotors were in the right-hand rotor location.

Kahn [**13**, p. 66] claims that " ... the Poles had a stroke of luck. The Germans changed the rotors every three months, or quarter of a year. Fortunately, the keys that Schmidt had supplied straddled two different quarters." And, Budiansky [**3**, p. 102] echoes and strengthens Kahn's statement: " ... if it were not for the changes in the rotor order, Rejewski would have hit another impasse ... ." Lawrence [**18**] suggests that even if Rejewski only had received data for one rotor order he still would have been able to determine the Enigma wiring. In [**19**], Lawrence considers whether Rejewski needed the information obtained from Asché to solve his six equations and obtain the wiring of the rotors. But, through Asché, information about S was available to Rejewski for two different quarters, and he did use it to determine the wiring of two Enigma rotors.

> Also known, thanks to materials obtained by intelligence, are the plug connections S for the given day:
>
> $$S = \texttt{(ap)(bl)(cz)(fh)(jk)(qu)}.$$
>
> Marian Rejewski [**14e**, p. 282]

So, the remaining unknowns are Q and N.
Rejewski transfers S to the left side of each of the six equations.

$$\begin{aligned}
S^{-1}AS &= PNP^{-1}QPN^{-1}P^{-1}\\
S^{-1}BS &= P^2NP^{-2}QP^2N^{-1}P^{-2}\\
S^{-1}CS &= P^3NP^{-3}QP^3N^{-1}P^{-3}\\
S^{-1}DS &= P^4NP^{-4}QP^4N^{-1}P^{-4}\\
S^{-1}ES &= P^5NP^{-5}QP^5N^{-1}P^{-5}\\
S^{-1}FS &= P^6NP^{-6}QP^6N^{-1}P^{-6}
\end{aligned}$$

Then, because he also knows P, he transfers it to the other side of each equation.

$$\begin{aligned}
U &= P^{-1}S^{-1}ASP &&= NP^{-1}QPN^{-1}\\
V &= P^{-2}S^{-1}BSP^2 &&= NP^{-2}QP^2N^{-1}\\
W &= P^{-3}S^{-1}CSP^3 &&= NP^{-3}QP^3N^{-1}\\
X &= P^{-4}S^{-1}DSP^4 &&= NP^{-4}QP^4N^{-1}\\
Y &= P^{-5}S^{-1}ESP^5 &&= NP^{-5}QP^5N^{-1}\\
Z &= P^{-6}S^{-1}FSP^6 &&= NP^{-6}QP^6N^{-1}
\end{aligned}$$

Actually, Rejewski needs only the first four of these. He substitutes for S, the various powers of P, and A, B, C, and D and determines U, V, W, and X.

$U = (ax)(bu)(ck)(dr)(ej)(fw)(gi)(lp)(ms)(nz)(oh)(qt)(vy)$
$V = (ar)(bv)(co)(dh)(fl)(gk)(iz)(jp)(mn)(qy)(su)(tw)(xe)$
$W = (as)(bz)(cp)(dq)(eo)(fw)(gj)(hl)(iy)(kr)(mu)(nt)(vx)$
$X = (ap)(bf)(cu)(dv)(ei)(gr)(ho)(jn)(ky)(lx)(mz)(qs)(tw)$

Next, Rejewski forms products.

$UV = (NP^{-1}QPN^{-1})(NP^{-2}QP^2N^{-1}) = NP^{-1}(QP^{-1}QP)PN^{-1}$
$VW = NP^{-2}(QP^{-1}QP)P^2N^{-1}$
$WX = NP^{-3}(QP^{-1}QP)P^3N^{-1}$

Rejewski notes that (because each is a conjugate of $QP^{-1}QP$) "the products have the same configuration of cycles, which is as it should be." [**14e**, p. 282]

$UV = (aepftybsnikod)(rhcgzmuvqwljx)$
$VW = (akjcevzydlwnu)(smtfhqibxopgr)$
$WX = (aqvloikgnwbmc)(puzftjryehxds)$

He then eliminates the common expression $QP^{-1}QP$ between UV and VW

$$\begin{aligned} VW &= NP^{-2}(QP^{-1}QP)P^2N^{-1} \\ &= NP^{-1}N^{-1}(NP^{-1}(QP^{-1}QP)PN^{-1})NPN^{-1} \\ &= NP^{-1}N^{-1}(UV)NPN^{-1} \\ &= (NPN^{-1})^{-1}(UV)(NPN^{-1}) \end{aligned}$$

and similarly between VW and WX.

$$WX = (NPN^{-1})^{-1}(VW)(NPN^{-1})$$

Because $VW = (NPN^{-1})^{-1}(UV)(NPN^{-1})$, Theorem 5 can be used to find several possibilities for NPN^{-1}. Similarly, because $WX = (NPN^{-1})^{-1}(VW)(NPN^{-1})$, Theorem 5 can be used to find possibilities for NPN^{-1}.

> We should ... write VW beneath product UV in every possible way, and likewise, product WX beneath product VW. Of all these possible ways, one will give the same result in both cases. This will be the expression that we need. Writing VW beneath UV, and WX beneath VW, in every possible way is rather tedious. However, there are various tricks and technical means that make this subscription unnecessary, but whose description and, especially, justification would take us too far afield. It will suffice to say that products UV, VW, and WX should be subscribed in the following way:
>
> $UV = (aepftybsnikod)(rhcgzmuvqwljx)$
> $VW = (ydlwnuakjcevz)(ibxopgrsmtfhq)$
> $VW = (ydlwnuakjcevz)(ibxopgrsmtfhq)$
> $WX = (uzftjryehxdsp)(caqvloikgnwbm)$
>
> For, in both cases, we obtain for NPN^{-1} the same expression:
>
> $NPN^{-1} = (ayuricxqmgovskedzplfwtnjhb)$
>
> Marian Rejewski [**14e**, pp. 282 & 283]

To find N, Rejewski uses Theorem 5 again.

> Subscribing beneath permutation NPN^{-1} permutation P in all possible ways, of which there are twenty-six, we will obtain [using Theorem 5] twenty-six variants of the permutation N. For example, one variant is [**14e**, p. 283]:
>
> $$NPN^{-1} = (\texttt{ayuricxqmgovskedzplfwtnjhb})$$
> $$P = N^{-1}(NPN^{-1})N = (\texttt{abcdefghijklmnopqrstuvwxyz})$$
>
> For this variant, after the upper row has been placed in alphabetical order, we obtain:
>
> $$N = \begin{pmatrix} \texttt{abcdefghijklmnopqrstuvwxyz} \\ \texttt{azfpotjyexnsiwkrhdmvclugbq} \end{pmatrix}$$

The Polish Doubles

> *After only a month of continuous and highly concentrated effort, [Rejewski] had worked out the electrical connections of the three wheels that were used at that time in the German Enigma. He was able to have a replica of the machine constructed.*
>
> Gordon Welchman [**27**, p. 15]

By December, 1932, Rejewski knew the wiring of the Enigma rotors and was able to determine the settings based upon the double encipherment of the message indicators. By the middle of January, 1933, the Poles were able to read Enigma messages.

> In 1938, the situation was aggravated. The Germans changed the encryption procedure on January 15, and introduced on December 15 a fourth and a fifth rotor, which now gave 60 instead of the previous 6 possible rotor orders.
>
> The Poles had to find out the wiring of the new rotors quickly, and they were lucky. Among the traffic they regularly decrypted were signals from the S.D. (*Sicherheitsdienst*), the intelligence service of the Nazi Party. The S.D. did not change their encryption procedure, but introduced the new rotors in December 1938. These rotors came from time to time into the position of the fast rotor and their wiring would be reconstructed the same as previously with the first three rotors. [**2**, p. 395]

Soon the Poles had several Enigma "doubles" built.

Fearing that war would begin soon, the Poles met, on July 24 and 25, 1939, with British and French cryptologists in the B.S.-4 facility in Kabackie Woods outside Warsaw.[7] It was at this meeting that the Poles revealed the extent of their abilities to read Enigma and told the French and British that each would receive a Polish-made Enigma double.

One of the Polish Enigma doubles is now on display in the Sikorski Polish Museum in London. (See Figure 7, p. 269.) The Enigma plugboard is not visible behind the lampboard. This is a 3-rotor Enigma, but the machine had five rotors from which the three in use were chosen. The two rotors on the right are in storage; the three rotors on the left are installed.

At the beginning of September, 1939, Poland was attacked by Germany.

[7] A photograph of the site as it now exists may be found in [**11**].

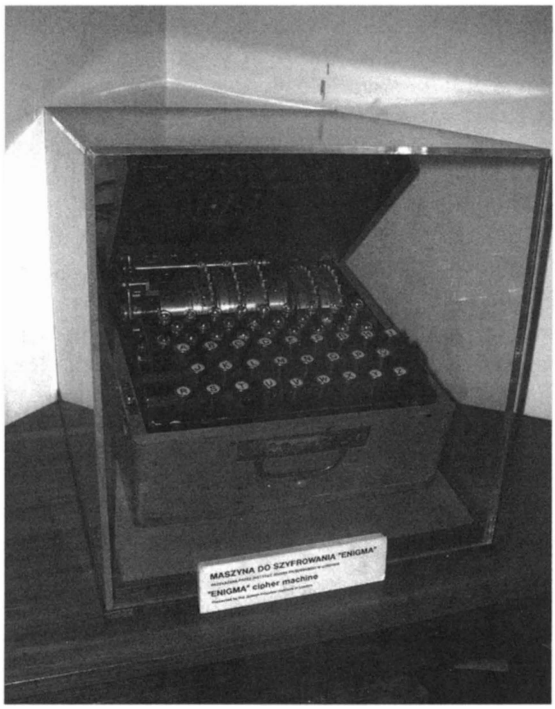


Figure 7 Polish Enigma Double

What Happened to the Polish Mathematicians?

Mathematicians are often thought of as being rather remote individuals, indulging in activities which have little or no relevance to real life.

Frank Carter [**4**, p. 4]

On September 5, 1939, B.S.-4 was told to evacuate Warsaw on a special train. The Polish mathematicians crossed the border into Romania, traveled through Italy, and eventually crossed the border into France. On October 20, 1939, the Polish mathematicians, from a site not far from Paris, resumed their attack on the German ciphers. On June 22, 1940, French Premier Pétain signed an armistice which divided France; on June 24 the Poles were flown to North Africa. In Algiers, they took on new identities and returned to France to resume signal intelligence in Vichy France. They operated from a site near the town of Uzès near the Mediterranean coast. The Poles occasionally spent two- or three-month periods at the North African station, and on January 9, 1942, Różycki died when the French ship *Lamoricière* carrying him and other staff back to France from Algeria was sunk.

Just prior to the German occupation of the free zone of France, Rejewski and Zygalski fled to the Italian zone, then back to France, and "on the night of 29 January 1943, ... set out with [a] smuggler for the [Spanish] border." [**14**, p. 150] On the trip, the smuggler demanded from them at gunpoint more money for the trip. Upon arriving in neutral (but sympathetic to Germany) Spain, the Poles were arrested. Upon their release they made their way to Madrid.

Near the end of July, they made their way to Portugal and were taken by boat to a British naval vessel waiting off the coast.

For the remainder of the war, Rejewski and Zygalski worked at a Polish Signals Battalion in Boxmoor near London. The British codebreakers at Bletchley Park were now routinely breaking Enigma; the Poles worked on German S.S. and S.D. ciphers.

Stuart Milner-Barry, who was playing chess for the British team in Argentina when war broke out, became a codebreaker and later became director of Hut 6 (German Army and Air Force cryptanalysis); he speculates about why the Polish codebreakers were not invited to Bletchley Park.

> It was always a mystery to me that the Polish contingent was not incorporated at Bletchley during the war, where they would no doubt have made an invaluable contribution; but in fact they were side-tracked in France and had to be evacuated when the Germans overran the whole of the country. I can only assume there were security doubts, and I believe the Poles continued to operate their own organization, but I feel there must have been a sad waste of resources somewhere.
>
> Stuart Milner-Barry [**9**, pp. 92 & 93]

After the war, Rejewski returned to Poland in November, 1946.

> ... for reasons of practical and family nature, it proved difficult for Rejewski to find employment as a mathematician at an institution of higher learning, and, in the early postwar period, he felt it imprudent to apply for a job in cryptology .... for 20 years [Rejewski] worked in the administrations of various concerns in Bydgoszcz, and in February 1967 retired [**14**, p. 224].

Rejewski died in 1980.

Henryk Zygalski remained in England after the war and taught in London. He died in 1978.

Bletchley Park is now a museum that honors the work of the British codebreakers. Outside the Bletchley Park cottage in which the British codebreakers made their first break into Enigma is a tablet that honors the work of the Polish codebreakers. A copy of that tablet has been placed on the west wall of the former Ministry of War office in Pilsudski Square in Warsaw where the Polish codebreakers worked.

> *This plaque commemorates the work of Marian Rejewski, Jerzy Różycki, and Henryk Zygalski, mathematicians of the Polish intelligence service, in first breaking the Enigma code. Their work greatly assisted the Bletchley Park code breakers and contributed to the allied victory in World War II.*[8]

For further study. Beginning with the publication of *The Ultra Secret* in 1974 [**29**], some information about Enigma has become public. Although other information is still classified, there are many websites and papers and books about Enigma. Here are some to use for further study.

There are many Enigma websites; some include virtual Enigma machines. Two sites to start with are the official website of Bletchley Park:

`http://www.bletchleypark.org.uk/`

and Tony Sale's World War II Codes and Ciphers:

`http://www.codesandciphers.org.uk/`

[8]The English version of the statement on the tablet honoring the Polish codebreakers at Bletchley Park.

The National Security Agency's website:

`http://www.nsa.gov/history/histo00007.cfm`

contains downloadable publications about cryptological history including Enigma.

Wikipedia is also an excellent reference for cryptological topics.

There are also many books. The standard reference for cryptological history is *The Codebreakers* by David Kahn [**12**].

When Kahn's book appeared in 1967, Enigma was unknown to the public. The revised and updated book published in 1997 contains some material about Enigma, but his *Seizing The Enigma: The race to break the German U-boat codes 1939–1943* [**13**], which was published in 1991, is a more complete history of Enigma.

Simon Singh's *The Code Book: The Evolution of Secrecy from Mary, Queen of Scots to Quantum Cryptograph* [**25**] has prompted some popular interest in cryptology, but should be read or used with some caution (see, for example [**21**]).

Two good readable recent histories of World War II codebreaking are *Enigma: The Battle for the Code* by Hugh Sebag-Montefiore [**24**] and *Battle of Wits: The complete story of codebreaking in World War II* by Stephen Budiansky [**3**].

The history of the Polish codebreakers is written in *Enigma: How the German Machine Cipher Was Broken, and How it was Read by the Allies in World War Two* by Wladyslaw Kozaczuk (translated by Christopher Kasparek) [**14**] and also in *Enigma: How the Poles Broke the Nazi Code* [**15**].

The Hut Six Story by Gordon Welchman [**27**] and *Codebreakers: The Inside Story of Bletchley Park* edited by F.H. Hinsley and Alan Stripp [**9**] are good starting points for understanding the work of the British codebreakers at Bletchley Park.

Two mathematical papers by Rejewski about the solution of Enigma ([14d] and [14e]) appear as appendices to [14]. They are also available on several internet sites.[9] Similar results appear in Rejewski's paper *An Application of the Theory of Permutations in Breaking the Enigma* [**22**]. This paper is also available on several internet sites.

Frank Carter, a mathematician who is now a Bletchley Park volunteer, has written several papers describing the mathematics used by the World War II cryptanalysts. His papers are available as either Bletchley Park Trust reports or on the Bletchley Park website. In particular, two of his papers [**4**] and the technical report "The Polish recovery of the Enigma Rotor wiring" (which is available on the Bletchley Park website and appeared just after this paper was written) discuss the mathematical work of Rejewski.

A complete coverage of cryptology from a mathematician's viewpoint is contained in *Decrypted Secrets: Methods and Maxims of Cryptology* by F.L. Bauer [**2**]. It is hoped that this paper's gentle introduction would encourage readers to examine Bauer's excellent book.

Cryptologia is a quarterly journal devoted to all aspects of cryptology. The journal began publishing in 1977, and its back issues contain many articles about the history and mathematics of Enigma. *Cryptologia* is published by the Taylor & Francis Group.

Mathematicians Did Not Win the War

Polish penetration into the secrets of the Enigma began in earnest when Rejewski realized the application of a simple property of permutations—namely, that if G and P are permutations, then the permutation defined by PGP^{-1} has the same

[9]The paper that is Appendix E also appears in *Cryptologia*, **VI**, number 1, (January 1989), 1–18; and in Cipher Deavours; David Kahn; Louis Kruh; Gregg Mellen; and Brian Winkel; editors, 1989, *Cryptology: Machines, History, & Methods*, Artech House, Boston, 1989, pp. 310–327.

cycle structure as the permutation G. No doubt practitioners of group theory should introduce this property of permutations to students as "the theorem that won World War II." Cipher A. Deavours [**5**, pp. 229 & 232].

To paraphrase many others, no theorem won the war. The war was won by those who served in the various Allied military services, but the information gleaned from Enigma helped the Allies win the war, and the breaking of Enigma began with Polish mathematicians who found patterns in Enigma messages.

Acknowledgments. I thank Northern Kentucky University student Dan Meredith for creating the illustrations and enhancing the photographs for this article. Dman41786@yahoo.com

This paper results from two talks given by the author. One talk was presented at the Pi Mu Epsilon initiation ceremony at Austin Peay State University, December, 2004, and the other was presented at the Mathematical Association of America, Kentucky Section, Spring 2005.

REFERENCES

1. A.A. Albert, Some Mathematical Aspects of Cryptography (address to the AMS on November 22, 1941), *A. A. Albert Collected Mathematical Papers*, AMS, 1993, pp. 903–920.
2. F.L Bauer, *Decrypted Secrets: Methods and Maxims of Cryptology, Second Edition*, Springer-Verlag, Berlin, 2000.
3. Stephen Budiansky, *Battle of Wits: The complete story of codebreaking in World War II*, The Free Press, New York, 2000.
4. Frank Carter, The first breaking of Enigma: Some of the pioneering techniques developed by the Polish Cipher Bureau, *The Bletchley Park Trust Reports*, number 10, 1999.
5. Cipher A. Deavours, Afterward to How Polish Mathematicians Deciphered the Enigma by Marian Rejewski, *Annals of the History of Computing*, **3**, Number 3, (July 1981), 229–232.
6. David Hamer, Enigma: Actions Involved in the "Double-Stepping" of the Middle Rotor, *Cryptologia*, **XXI**, number 1 (January 1997), 47–50.
7. Robert Harris, *Enigma*, Ballatine Books, New York, 1995.
8. Peter Hilton, Enigma, *Notices of the American Mathematical Society*, **43**, number 6, (June 1996), 681–682.
9. F.H. Hinsley and Alan Stripp, editors, *Codebreakers: The Inside Story of Bletchley Park*, Oxford University Press, Oxford, 1994.
10. Andrew Hodges, *Alan Turing: The Enigma*, Vintage, London, 1992.
11. David Kahn, The Polish Enigma Conference and Some Excursions, *Cryptologia*, **XXIX**, number 2 (April 2005), 121–126.
12. David Kahn, *The Codebreakers: The story of secret writing, revised and updated*, Scribner, New York, 1996.
13. David Kahn, *Seizing the Enigma: The race to break the German U-boat codes 1939–1943*, Barnes and Noble, New York, 1991.
14. Wladyslaw Kozaczuk, *Enigma: How the German machine cipher was broken, and how it was read by the Allies in World War Two*, translated by Christopher Kasparek, University Publications of America, 1984. [14b] Appendix B, A conversation with Marian Rejewski by Richard Woytak. [14d] Appendix D, How the Polish mathematicians broke Enigma by Marian Rejewski. [14e] Appendix E, The mathematical solution of the Engima cipher by Marian Rejewski.
15. Wladyslaw Kozaczuk and Jerzy Straszak, *Enigma: How the Poles Broke the Nazi Code*, Hippocrene Books, New York, 2004.
16. Alex Kuhl, Rejewski's Catalog, *Cryptologia*, to appear.
17. Solomon Kullback, *Statistical Methods in Cryptanalysis*, Aegean Park Press, Walnut Creek, CA, 1976.
18. John Lawrence, The versatility of Rejewski's method: Solving for the wiring of the second rotor, *Cryptologia*, **XXVIII**, number 2 (April 2004), 149–152.
19. John Lawrence, A Study of Rejewski's Equations, *Cryptologia*, **XXIX**, number 3 (July 2005), 233–247.
20. John Lawrence, Factoring for the Plugboard—Was Rejewski's Proposed Solution for Breaking the Enigma Feasible?, *Cryptologia*, **XXVIX**, number 4 (October 2005), 343–366.
21. Jim Reeds, Review of The Code Book: The Evolution of Secrecy from Mary, Queen of Scots to Quantum Cryptography, *Notices of the American Mathematical Society*, **47**, number 3, (March 2000), 369–372.
22. M. Rejewski, An Application of the Theory of Permutations in Breaking the Enigma Cipher, *Applicaciones Mathematicae* **16**, number 4, (1980), 543–559.
23. Frank Rowlett, *The Story of Magic: Memoirs of an American cryptologic pioneer*, Aegean Park Press, Walnut Creek, CA, 1998.

24. Hugh Sebag-Montefiore, *Enigma: The Battle for the Code*, Phoenix, London, 2000.
25. Simon Singh, *The Code Book: The Evolution of Secrecy from Mary, Queen of Scots to Quantum Cryptography*, Doubleday, New York, 1999.
26. Abraham Sinkov, *Elementary Cryptanalysis: A mathematical approach*, Mathematical Association of America, 1968.
27. Gordon Welchman, *The Hut 6 Story*, M and M Baldwin, Cleobury Mortimer, Shropshire, England, 1998.
28. Brian J. Winkel, Cipher A. Deavours, David Kahn, Louis Kruh, editors, *The German Enigma Machine: Beginnings, Success, and Ultimate Failure*, Artech House, Boston, 2005.
29. F.W. Winterbotham, *The Ultra Secret*, Dell, New York, 1974.

Why Richard Cory[1] Offed Himself
or
One Reason to Take a Course in Probability

A hypochondriac at heart, he thought
(Though symptom free) he had a dire disease,
And after fruitless weeks of worry, sought
Some test to take to set his mind at ease.

He forthwith found one that would do the trick,
And accurate (at oh point nine) to tell
Those having the disease that they were sick,
And just the same, the well that they were well.

One crucial point he failed to note was this:
That of a hundred like him, only one
Had the disease, and this slip made him miss
The implication when the test was done

And positive! Therefore, consumed with dread,
And now convinced his blackest fears were right
(By faulty logic fatally misled[2]),
He shattered silence that calm summer night.

J. D. Memory
Professor of Physics, Emeritus
North Carolina State University
jmemory@nc.rr.com

[1] "Richard Cory" is a frequently anthologized poem by E. A. Robinson

[2] An example of the False Positive Fallacy: On average, of 1000 Corys, ten would have the disease, yielding nine true positives and one false negative. Of the remaining 990, there would be 99 false positives and 891 true negatives. The false positives outnumber the true positives by a factor of eleven. So if D denotes having the disease and P denotes testing positive, we learn from the poem that $\Pr(P \mid D) = 0.9$, whereas $\Pr(D \mid P) = 9/108$, or about 0.083. Richard Cory shot himself "that calm summer night," because he confused $\Pr(P \mid D)$ with $\Pr(D \mid P)$.

Seeing Dots: Visibility of Lattice Points

JOSHUA D. LAISON
Willamette University
Salem, OR 97301
jlaison@willamette.edu

MICHELLE SCHICK
Kansas State University
Manhattan, KS 66506-2602

Consider a photographer who is commissioned to take publicity pictures of a high school marching band while standing near them on the field. He wishes to take the smallest number of photographs such that every person is visible and unobstructed in at least one of the photographs, so as not to upset any member of the PTA.

A marching band's traditional formation is a grid pattern, with exactly five feet between each student, horizontally and vertically as viewed from above. We abstract the problem by representing each member of the band by an integer lattice point, and say that she is viewed in the photograph if there is no other student on the line segment between her and the photographer, who is required to stand at a lattice point. Notice that this is a simplification, since some lines of sight will pass very close to another lattice point without passing through it. We represent the band as an $r \times s$ rectangle of integer lattice points with corners at $(1, 1)$, $(r, 1)$, $(1, s)$, and (r, s), denoted by $\Delta_{r,s}$, and the photographer's vantage points as a small number of integer lattice points outside the band formation, one for each picture. An example of a particular picture is shown in FIGURE 1. Note that the photographer captures three of the four students in this shot.

Given these criteria, what is the smallest number of pictures the photographer needs to take of an $r \times s$ marching band? It turns out that this simple question has a not-so-simple answer. Although we provide a number of answers for some small values of r and s, using the Chinese Remainder Theorem as a key ingredient, much of the work on this problem remains open. We will also briefly discuss the analogous problem for marching bands in higher-dimensional space.

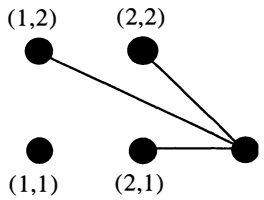


Figure 1 From the point $(3, 1)$, the photographer can see three of the four students in this band, formalized as the set $\Delta_{2,2}$.

Formally, two distinct integer lattice points P and Q are *mutually visible* if there are no other integer lattice points on the line segment joining P and Q. Equivalently, $P = (a_1, a_2)$ and $Q = (b_1, b_2)$ are mutually visible if and only if $\gcd(a_1 - b_1, a_2 - b_2) = 1$ [**10**]. So for example, $(3, 5)$ and $(1, 2)$ are mutually visible, since $\gcd(3 - 1, 5 - 2) = 1$. Recall that for any integer k, $\gcd(k, 0) = k$. So if P and Q have the same x-coordinate, then $a_1 - b_1 = 0$, and they are mutually visible if and only if their y-coordinates differ by 1.

We will use this equivalent criterion more frequently than the definition in our arguments.

Weak lattice point visibility

As a warm-up question, we ask where the photographer might position himself to take a small number of photographs of the band in an empty field, so that every person is clearly visible in one of the photographs. Recall that we require the photographer to stand at a lattice point. We say that the rectangle of lattice points $\Delta_{r,s}$ is *weakly visible* from the point P if no line segment between P and a point in $\Delta_{r,s}$ passes through another point in $\Delta_{r,s}$. It turns out that it takes only a single point to weakly view a rectangle of any size.

THEOREM 1. *The set $\Delta_{r,s}$ is weakly visible from the point P whose coordinates are $(rs - s + r, s + 1)$.*

Proof. Consider the set of all lines $\mathcal{L}$ between pairs of points in $\Delta_{r,s}$. We claim that P is not on any line in $\mathcal{L}$. From this claim, the theorem follows.

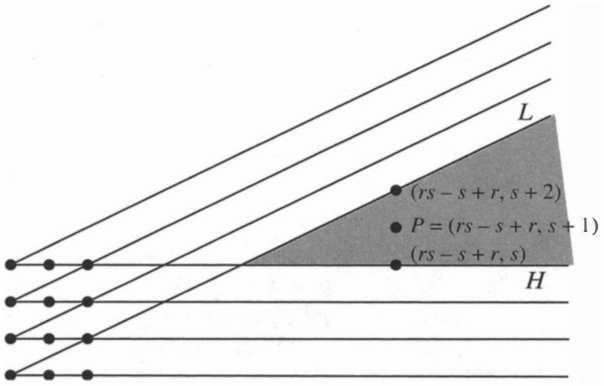


Figure 2 The point $(rs - s + r, s + 1)$ weakly views all of $\Delta_{r,s}$.

The line L connecting the points $(1, 1)$ and $(r, 2)$ is the line with the smallest positive slope in $\mathcal{L}$. The line H with equation $y = s$ is the horizontal line with largest y-intercept in $\mathcal{L}$. Therefore, no lines in $\mathcal{L}$ cross through the region below L and above H. But since L contains the point $(rs - s + r, s + 2)$ and H contains the point $(rs - s + r, s)$, P is below L and above H, as illustrated in FIGURE 2. ■

It is interesting to note that the situation only improves in higher dimensions. If A, B, and C are three points in n-dimensional space, and their projections onto the xy-plane are not collinear, then A, B, and C are also not collinear. Therefore if Δ is an n-dimensional box of integer lattice points, with x and y dimensions r and s, respectively, then Δ is still weakly visible from the point $P = (rs - s + r, s + 1, 0, 0, \ldots, 0)$.

Also note that the point we have found might be quite far from the rectangle or box we wish to view, so the people might appear quite small in the photograph.

OPEN PROBLEM 1. *Find the lattice point(s) that weakly view the rectangle $\Delta_{r,s}$ (or higher-dimensional box Δ) and are closest to it.*

External lattice point visibility

Now that we have answered our question in an empty field, we turn to a more interesting question: What if the band is marching down a crowded street? We can represent the crowd by adding in the rest of the integer lattice points in the plane. Now to view a member of the band, the photographer must have a line of sight that doesn't pass through any member of the band or of the crowd. Again, we ask how few photographs are needed to view every face in the band. Note that this question is not interesting if we are allowed to take our photographs from noninteger points, so we restrict ourselves to integer points.

Formally, we say that $\Delta_{r,s}$ is *externally visible* from the set of integer lattice points $P_1, P_2, \ldots, P_k$ not in $\Delta_{r,s}$, if for every point Q in $\Delta_{r,s}$, Q and P_i are mutually visible for some i. We then define $e(r, s)$ to be the smallest number of points outside $\Delta_{r,s}$ from which $\Delta_{r,s}$ is externally visible. We are interested in the exact value of $e(r, s)$ for all positive integers r and s.

As a useful visualization technique, we consider a single photograph of an arbitrarily large band, with the photographer standing at the point $(0, 0)$ and facing toward the first quadrant. FIGURE 3 shows the points that are visible from this position as black squares, and the points that are not visible as white squares. To obtain the pattern of points visible and invisible to a photographer facing toward the second, fourth, or third quadrants, we reflect FIGURE 3 across the y-axis, across the x-axis, or through the origin, respectively. To obtain the pattern of points visible to a photographer facing directly along an axis, we place one additional black square at $(1, 0)$, $(-1, 0)$, $(0, 1)$, or $(0, -1)$, since only the band member closest to the photographer in that row or column will be visible.

To obtain the pattern of points visible and invisible to a photographer standing at the point (a, b), we simply relabel the x-axis starting with a and the y-axis starting

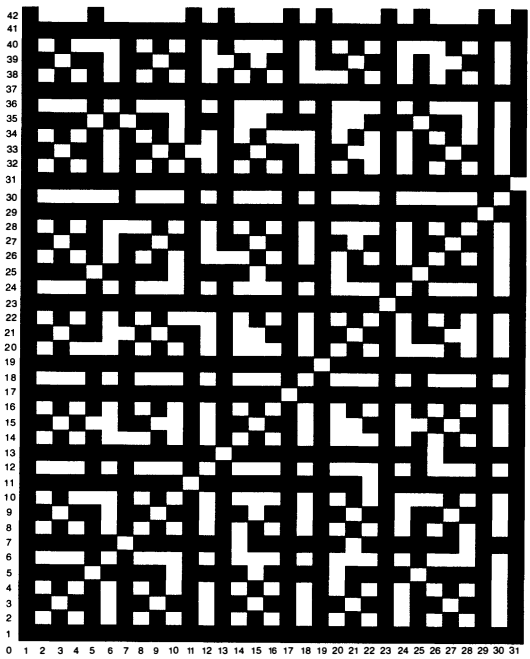


Figure 3 Lattice points visible from the point (0, 0) appear as black squares, and invisible points appear as white squares.

with b. Thus, we can simulate multiple photographs by overlaying several copies of FIGURE 3, copied onto transparency film. Many of our small values of $e(r, s)$ were obtained in this way.

We can now make some general observations about the values of $e(r, s)$. By symmetry, $e(r, s) = e(s, r)$. Seeing $\Delta_{r,s}$ means seeing every subset of $\Delta_{r,s}$, therefore $e(r, s) \leq e(t, u)$ if $r \leq t$ and $s \leq u$. The point $(0, 0)$ can view any point of the form $(1, k)$, which means that $e(1, s) = 1$ for all positive integers s.

One point is not, however, sufficient to view a rectangle even as small as 2×2. To see this, let (a, b) be an integer lattice point not in $\Delta_{2,2}$. Suppose that a is odd and b is even. Then $a - 1$ and $b - 2$ are both even, so $\gcd(a - 1, b - 2) > 1$. Hence by our criterion, (a, b) and $(1, 2)$ are not mutually visible. In an analogous way, every point (a, b) cannot view at least one of the points in $\Delta_{2,2}$, so $e(2, 2) > 1$. The points $(0, 0)$ and $(0, 1)$ externally view $\Delta_{2,2}$, so $e(2, 2) = 2$. We can generalize this idea to larger rectangles by use of the Chinese Remainder Theorem.

THEOREM 2. (CHINESE REMAINDER THEOREM) *Let $m_1, m_2, \ldots, m_k$ be pairwise relatively prime positive integers. Then the system of congruences*

$$x \equiv n_1 \bmod m_1, x \equiv n_2 \bmod m_2, \ldots, x \equiv n_k \bmod m_k$$

has a unique solution modulo $m_1 m_2 \cdots m_k$.

Detailed introductions to the Chinese Remainder Theorem and modular arithmetic appear in most introductory books on number theory [**6, 11, 12**].

Herzog and Stewart [**9**] use this theorem to investigate patterns of visible lattice points in larger rectangles. They define a *pattern* P to be a subset of integer lattice points in the plane. They define a particular pattern P to be *realizable* if P can be translated by some vector in $\mathbb{Z}^2$ such that every point in P is visible from the origin. In other words, P is realizable if it can be externally viewed from a single point. They define a *complete square modulo m* to be a set $S = \{(x_k, y_k)\}$ of m^2 integer lattice points in the plane, such that $\{(x_k \bmod m, y_k \bmod m)\} = \{(a, b) \mid 0 \leq a, b \leq m - 1\}$. For example, the four points $(1, 5)$, $(7, 2)$, $(8, 3)$, and $(4, 4)$ form a complete square modulo 2, since their remainders modulo 2 are $(1, 1)$, $(1, 0)$, $(0, 1)$, and $(0, 0)$, respectively. Note that any integer lattice point (a, b) must be congruent to one of these points modulo 2, and hence does not view that point. This idea is at the heart of the following theorem.

THEOREM 3. (HERZOG AND STEWART, 1971) *A given pattern P is realizable if and only if P fails to contain a complete square modulo p for every prime p.*

We can see that if P contains a complete square modulo any prime p, then P cannot be externally viewed from a single point, since any candidate point (a, b) must be congruent modulo p to one of the points in the pattern, and hence cannot view that point. To prove the converse, Herzog and Stewart use the Chinese Remainder Theorem to construct an external point (a, b) that views all points in the pattern. This point is constructed so that a is not congruent to any of the x-coordinates of points in P modulo any prime smaller than the size of the pattern, and analogously for b.

Using similar techniques, other authors have obtained asymptotic results concerning the number of lattice points needed to view a rectangle of lattice points, as the rectangle gets large. Abbott defines $f(n)$ to be the smallest number of points in $\Delta_{n,n}$ required to view the rest of $\Delta_{n,n}$. This version of the problem may be thought of as *internal visibility*. Abbott [**1**] proves that $\log n / 2 \log \log n < f(n) < 4 \log n$ for large n. His proof uses the Chinese Remainder Theorem to establish the lower bound, again by using a system of congruences to construct the set of viewing points. The upper

bound follows, after some algebraic manipulations, from the following classical result of analytic number theory [5].

THEOREM 4. *The set of integer lattice points visible from the origin has density* $6/\pi^2$.

Chen and Cheng [7] define $g(n)$ to be the smallest number of lattice points in the plane, internally or externally, required to view $\Delta_{n,n}$. Chen and Cheng also base their construction on the Chinese Remainder Theorem to prove that $g(n) \geq k \log n / \log \log n$, where k approaches $\pi^2/6$.

However, no previous authors seem to have investigated external visibility alone, or more significantly, the exact numbers of points required to view various size rectangles of lattice points. While informative for very large rectangles, the results of Abbot, Chen, and Cheng give little indication of how many points might be required to view a rectangle of any given small size. In particular, it is unlikely that any marching band would be large enough to find these results useful, even one from a large state university. This seemed to us to be a golden opportunity to tackle an unexplored and interesting problem. We leave plenty of questions for interested readers to continue our investigations.

Lower bounds. Let's start by using Theorem 3 to demonstrate that two points are not enough to externally view $\Delta_{6,6}$. Let $P = (a_1, a_2)$ and Q be any two integer lattice points outside of $\Delta_{6,6}$. We consider the set $S = \{(a_1 + 3j, a_2 + 3k) \mid j, k \in \mathbb{Z} \text{ not both } 0\}$. No point in S is visible from P. There are exactly four points in $S \cap \Delta_{6,6}$, and they form a complete square modulo 2. By Theorem 3, these four points cannot all be externally viewed from Q. Therefore $e(6, 6) > 2$, and we will discover below that in fact $e(6, 6) = 3$.

Not only that, a 6×6 rectangle is the smallest rectangle that requires three points. Consider the points $(0, 0)$ and $(0, -1)$. If a and b are positive integers, $(0, 0)$ is mutually visible with any point (a, b) such that a and b are relatively prime, and $(0, -1)$ is mutually visible with any point (a, b) such that a and $b + 1$ are relatively prime. If a is an integer from 1 to 5, then for every positive integer b, either a and b are relatively prime, or a and $b + 1$ are relatively prime. Therefore $e(5, s) = 2$ for all positive integers s. We might also choose to define $\Delta_{5,\infty}$ as the set of all lattice points with x-coordinate from 1 to 5, and positive y-coordinate. Then by the same reasoning, $e(5, \infty) = 2$.

We now know exactly which rectangles can be viewed from two external points: the rectangles with fewer than six points along one of their sides. Continuing our search for values of $e(r, s)$, our next question is, what is the largest rectangle viewable from three external points?

We try adding a third point to the two we already have, and we find that the points $(0, 0)$, $(0, -1)$, and $(15, -2)$ externally view $\Delta_{14,s}$ for any positive integer s. We have already seen that $(0, 0)$ and $(0, -1)$ view all points (a, b) with either $\gcd(a, b) = 1$ or $\gcd(a, b + 1) = 1$. The points in $\Delta_{14,s}$ not satisfying either of these criteria have one of the following forms: $(6, 6n + 2)$, $(6, 6n + 3)$, $(10, 10n + 4)$, $(10, 10n + 5)$, $(12, 6n + 2)$, $(12, 6n + 3)$, $(14, 14n + 6)$, or $(14, 14n + 7)$, where n is a nonnegative integer. These points are shown as squares in FIGURE 4. By brute force, we can check that every one of these points is visible from $(15, -2)$.

We display a picture of $\Delta_{14,15}$ in FIGURE 4. The circles in the figure represent those points that cannot be viewed from $(0, 0)$. The line segments on the left illustrate how to view representative points among these from $(0, -1)$. (Drawing segments for *all* such points would make the diagram unnecessarily cluttered.) The squares mark the

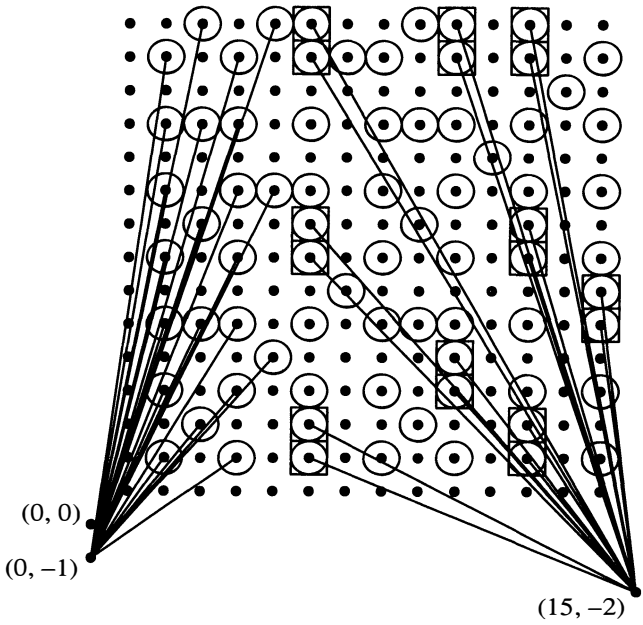


Figure 4 Viewing $\Delta_{14,15}$ from the points $(0, 0)$, $(0, -1)$, and $(15, -2)$.

points that can be seen from neither $(0, 0)$ nor $(0, -1)$. The second set of line segments illustrates that these points are visible from $(15, -2)$.

Using the same ideas we used to show that $e(2, 2) > 1$ and $e(6, 6) > 2$, we can prove a general lower bound for $e(r, s)$.

THEOREM 5. *Let p_i be the ith prime number, so that $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, and so on, and let $r_j \in \mathbb{Z}^+$. If $n = p_1^{r_1} p_2^{r_2} \dots p_i^{r_i}$, then $e(n, n) > i$.*

Proof. Let $P_1 = (x_1, y_1), \dots, P_i = (x_i, y_i)$ be i points not in $\Delta_{n,n}$. By the Chinese Remainder Theorem, there exist numbers a and $b \leq n$ such that $a \equiv x_k \bmod p_k$, and $b \equiv y_k \bmod p_k$, for each k between 1 and i. Therefore $x_k - a$ and $y_k - b$ are both divisible by p_k, and so (a, b) is not externally visible from any of the points P_1 through P_i. Thus, $e(n, n) > i$. ∎

COROLLARY 1. *If $r = p_1^{r_1} p_2^{r_2} \dots p_i^{r_i}$ and $s = p_1^{r_1} p_2^{r_2} \dots p_j^{r_j}$, then $e(r, s) > \min\{i, j\}$.*

Note that if n is the product of the first i primes, then $i = \omega(n)$, the number of primes in the prime factorization of n. In this case, i is approximately equal to $\log n / \log \log n$ [**8**], and the bound we obtain in Theorem 5 looks very similar to the lower bounds on internal and internal-external visibility obtained by Abbot, Chen, and Cheng [**1, 7**].

More specifically, Theorem 5 tells us that $e(30, 30) > 3$. FIGURE 5 allows us to see a connection between Theorem 3 and Theorem 5 in this case. Let P be any point outside $\Delta_{30,30}$. The circled points in the figure represent a set of points in $\Delta_{30,30}$ that are not externally visible from the single external point $P = (0, 0)$. The light and dark gray disks pick out a pattern of those circles that includes 25 complete squares modulo 2. Note that if the location of P were changed, the location of these 25 complete squares modulo 2 would correspondingly change, but such complete squares would still be present among the points not viewable from P. Since we have already seen that no complete square modulo 2 is visible from a single point, any second external point

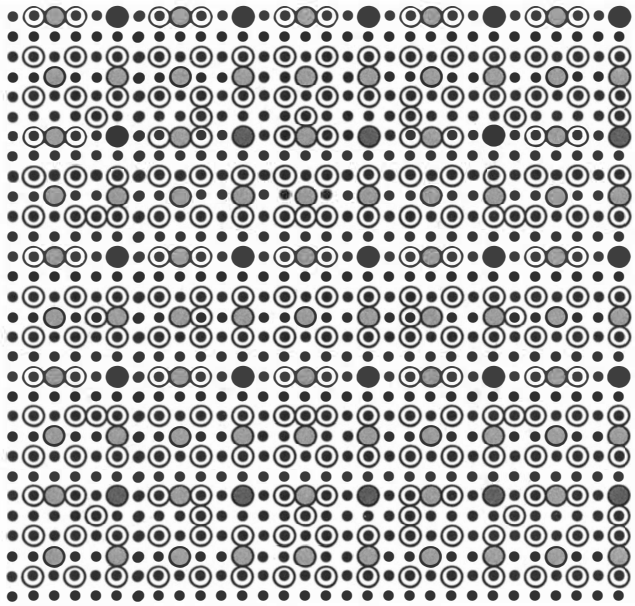


Figure 5 Complete squares in $\Delta_{30,30}$.

must miss one point from each of these complete squares. The invisible point appears in the same location in every complete square modulo 2 because, for instance, if the viewing point is congruent to (0, 0) mod 2, it cannot view the point in each complete square that is congruent to (0, 0) mod 2. These points missed by the second external point are indicated by the dark grey circles. But notice that the dark grey circles form a complete square modulo 5, requiring at least two additional points to view the entire rectangle.

So $e(14, s) = 3$, and $e(30, 30) > 3$. We remark that (6, 0), (6, −1), and (5, −16) externally view $\Delta_{23,19}$, so $e(23, 19) = 3$, but these points do not externally view $\Delta_{23,20}$.

OPEN PROBLEM 2. *Find* $e(23, 20)$.

OPEN PROBLEM 3. *Find the largest value of* n *such that* $e(n, n) = 3$.

Now that we have a general lower bound for $e(r, s)$ we ask for a general upper bound. Theorem 6 represents our best efforts in this direction, although it involves a sequence, $\delta(n)$, whose terms are not all known.

First suppose that i is a positive integer. We let $\rho(i)$ be the size of the longest sequence of consecutive positive integers sharing a common factor with i. For example, $\rho(p) = 1$ for any prime p, since no two consecutive positive integers have the same prime factor. Also, $\rho(6) = 3$ since the consecutive integers 2, 3, and 4 all share a common factor with 6, but four consecutive numbers must include an odd number not divisible by 3.

Then we define $\delta(n) = \max\{\rho(i) \mid i \leq n\}$, in other words, the largest value of $\rho(i)$ for any i between 1 and n. Equivalently, $\delta(n)$ is the size of the longest vertical line of lattice points that are *not* visible from (0, 0) in the rectangle $\Delta_{n,\infty}$.

THEOREM 6. $e(n, n) \leq \delta(n) + 1$.

Proof. By the definition of $\delta(n)$, for any $a < n$, if we order the points in $\Delta_{n,n}$ with x-coordinate a by increasing y-value, then there are at most $\delta(n)$ of these points in sequence not visible from $(0, 0)$.

Therefore we may take the external points $(0, 0), (0, 1), \ldots, (0, \delta(n))$. For any point (a, b) in $\Delta_{n,n}$, one of the integers $b, b-1, \ldots, b-\delta(n)$ is relatively prime to a. Say that $b-k$ is relatively prime to a. Then $(0, k)$ and (a, b) are mutually visible. Therefore the $\delta(n)+1$ points listed can view all of $\Delta_{n,n}$, and $e(n, n) \leq \delta(n)+1$. ■

OPEN PROBLEM 4. *Find a closed formula for* $\delta(n)$.

When n is the product of the first j primes, let $\gamma(j) = \rho(n)$. In other words, $\gamma(j)$ is the size of the longest sequence of consecutive positive integers such that each is divisible by one of the first j primes. The sequence $\gamma(j)$ is sequence A058989 in the *On-Line Encyclopedia of Integer Sequences* [**13**], of which only the first 24 terms are known. This leads us to suspect that Open Problem 4 is hard.

However, we do know a few small values of $\delta(n)$. In particular, we have $\delta(5) = 1$ and $\delta(6) = 3$, so Theorem 6 gives us $e(5, 5) \leq 2$, which is exact, but $e(6, 6) \leq 4$, which is not exact.

TABLE 1: The values of $e(r, s)$ for $1 \leq r, s \leq 30$.

$e(r, s)$	1	2–5	6–14	15–19	20–29	30	n
1	1	1	1	1	1	1	1
2–5		2	2	2	2	2	2
6–14			3	3	3	3	3
15–23				3	3–4	3–4	3–4
24–29					3–4	3–4	3–4
30						4–6	4–6

We summarize the known values of $e(r, s)$ in Table 1. When a specific value is not known, the range of possible values is given. Note that since $e(r, s) = e(s, r)$ as remarked above, the table is symmetric, and so values below the diagonal are omitted.

CONJECTURE 1. *If* $r \leq s$ *and* $r \leq t$ *then* $e(r, s) = e(r, t)$.

OPEN PROBLEM 5. *Find the values of* n *for which* $e(n, n) > e(n-1, n-1)$.

Marching bands in space

We can make an easy generalization to n-dimensional space from our formal definitions, although the rocket packs required to support a three-dimensional marching band formation are as yet prohibitively expensive, and dimensions larger than three are even worse. As noted earlier, weak visibility does not become any more interesting in higher dimensions. There are, however, some interesting things we can say about external visibility in higher dimensions.

Analogously, let $\Delta_{r_1,r_2,\ldots,r_n}$ be the n-dimensional box of lattice points with corner at $(1, 1, \ldots, 1)$, and let $e_n(r_1, r_2, \ldots, r_n)$ be the smallest number of points outside of $\Delta_{r_1,r_2,\ldots,r_n}$ required to externally view $\Delta_{r_1,r_2,\ldots,r_n}$. We add the subscript n so that the

dimension we're working in is clear. Herzog and Stewart proved this generalization of Theorem 3.

THEOREM 7. (HERZOG AND STEWART, 1971) *A given pattern P is realizable if and only if P fails to contain a complete hypercube modulo p for every prime p.*

Note that in particular, this means that $e_n(2, 2, \ldots, 2) = 2$. However, since $e(1, 2) = 1$, $e_n(1, 2, 2, \ldots, 2) = 1$ for all n. More generally, if r_i and r_j are the smallest of the numbers $r_1, r_2, \ldots, r_n$, then $e_n(r_1, r_2, \ldots, r_n) \leq e(r_i, r_j)$, since we can take the points required to externally view Δ_{r_i, r_j} in two dimensions, and add arbitrary additional coordinates to increase their dimension to n, and these new points will externally view $\Delta_{r_1, r_2, \ldots, r_n}$.

So our values of $e(r, s)$ in two dimensions are upper bounds for the values of $e_n(r_1, r_2, \ldots, r_n)$ in n dimensions.

OPEN PROBLEM 6. *Find an example of integers r, s, and t for which $r \leq s < t$ and $e(r, s) > e(r, s, t)$, or prove that one does not exist.*

REFERENCES

1. H. L. Abbott, Some Results in Combinatorial Geometry, *Discrete Math.* **9** (1974) 199–204.
2. Sukumar Das Adhikari, R. Balasubramanian, On a question regarding visibility of lattice points, *Mathematika* **43** (1996) 155–158.
3. Sukumar Das Adhikari, Yong-Gao Chen, On a question regarding visibility of lattice points II, *Acta Arith.* **LXXXIX** 3 (1999) 279–282.
4. Sukumar Das Adhikari, Yong-Gao Chen, On a question regarding visibility of lattice points III, *Discrete Math.* **259** (2002) 251–256.
5. T. M. Apostol, *Introduction to analytic number theory*, Springer, New York, NY, 1976.
6. D. M. Burton, *Elementary number theory*, Fifth edition, McGraw Hill, New York, NY, 2002.
7. Yong-Gao Chen, Lin-Feng Cheng, Visibility of lattice points, *Acta Arith.* **107** 3 (2003) 203–207.
8. G. H. Hardy and E. M. Wright, *An introduction to the theory of numbers*, Fifth edition, Oxford Univ. Press, New York, NY, 1979.
9. Fritz Herzog, B. M. Stewart, Patterns of Visible and Nonvisible Lattice Points, *Amer. Math. Monthly* **78** (1971) 487–496.
10. David Rearick, Mutually visible lattice points, *Norske Vid. Selsk. Forh.* **39** 6 (1966) 41–45.
11. K. H. Rosen, *Elementary number theory and its applications*, Fourth edition, Addison-Wesley, Reading, MA, 2000.
12. Joseph H. Silverman, *A Friendly Introduction to Number Theory*, Third edition, Pearson Prentice Hall, Upper Saddle River, NJ, 2006
13. N. J. A. Sloan, *The On-Line Encyclopedia of Integer Sequences*, `http://www.research.att.com/~njas/sequences/`

In his book, *Legends, Lies, and Established Myths of American History*, Richard Shenkman (refuting the notion that undergraduate behavior has deteriorated lately) notes:

> In 1830 students at Yale revolted over a change in the teaching of mathematics in an incident dubbed the "Conic Sections Rebellion." Before it was over, forty-three students—about half the class—had been expelled.

Daniel Moran
(Professor Emeritus)
Michigan State University

NOTES

Faro Shuffles and the Chinese Remainder Theorem

ARNE LEDET
Texas Tech University
Lubbock, TX 79409-1042

If m and n are natural numbers with greatest common divisor 1, the *Chinese Remainder Theorem* (or at least one version of it) states that for any two integers a and b there exists an integer x such that

$$x \equiv a \pmod{m} \quad \text{and} \quad x \equiv b \pmod{n},$$

and that this x is unique modulo mn.

There are various reformulations of this result, for instance that the map $[x]_{mn} \mapsto ([x]_m, [x]_n)$ from $\mathbb{Z}_{mn}$ to $\mathbb{Z}_m \times \mathbb{Z}_n$ is a ring isomorphism. However, for the purposes of this paper, the most useful formulation is the following:

We set up an $m \times n$ grid of squares, and number the squares according to this rule: We let square number 1 be the one in the upper left-hand corner, and get from square number i to square number $i + 1$ by going one step down and one to right. If this is not possible (at the lower and right-hand edges of the grid), we "wrap around" to the opposite edge and continue. Then we will in fact number all mn squares. Moreover, the squares in the ith row will have numbers that are congruent to i modulo m, and the squares in the jth column will have numbers that are congruent to j modulo n.

For example, a 3×5 grid will be numbered as follows:

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

Note that the numbers in the ith row are in fact all congruent to i modulo 3, and that similarly the numbers in the jth column are all congruent to j modulo 5. Thus, if we need to find, say, a number that is congruent to 2 modulo 3 and congruent to 3 modulo 5, all we have to do is look at position (2, 3), and we see that 8 is a solution. Furthermore, among the numbers $1, \ldots, 15$ (that is, working modulo 15) this is the only solution, since a solution would have to be in the second row and the third column.

It is easy to prove the Chinese Remainder Theorem in this reformulation: In carrying out the numbering of the squares, we must eventually stop, because we come to a square that has already been numbered. This must be the square in the upper left-hand corner, since otherwise we could back-track a step to conclude that the last square we visited had in fact also be numbered previously. But in order to reach the upper left-hand corner, the number of steps must be both a multiple of m (to get us back to the first row) and a multiple of n (to get us back to the first column), that is, a multiple

of mn (since m and n have greatest common divisor 1). So it happens after mn steps, prior to which we must have numbered mn different squares, that is to say all of them.

If we imagine a deck of cards with m suits and n values in each suit, we can lay out the cards in an $m \times n$ array in such a way that all the cards in a row have the same suit, and all the cards in a column have the same value. For an ordinary deck ($m = 4$ and $n = 13$), this might give us

♠A, ♠2, ♠3, ♠4, ♠5, ♠6, ♠7, ♠8, ♠9, ♠10, ♠J, ♠Q, ♠K,
◇A, ◇2, ◇3, ◇4, ◇5, ◇6, ◇7, ◇8, ◇9, ◇10, ◇J, ◇Q, ◇K,
♣A, ♣2, ♣3, ♣4, ♣5, ♣6, ♣7, ♣8, ♣9, ♣10, ♣J, ♣Q, ♣K,
♡A, ♡2, ♡3, ♡4, ♡5, ♡6, ♡7, ♡8, ♡9, ♡10, ♡J, ♡Q, ♡K.

Of course, here it is not necessary that m and n have greatest common divisor 1. However, if they do, the Chinese Remainder Theorem can be illustrated by picking up the cards according to the numbering of the grid. For instance, the layout above would give a sequence

♠A, ◇2, ♣3, ♡4, ♠5, ◇6, ♣7, ♡8, ..., ♠10, ◇J, ♣Q, ♡K.

In this way, the cards are arranged such that the suits and the values go through repeating cycles independently of each other. (In the case of the example, this would be ♠◇♣♡ and A, 2, 3, ..., J, Q, K, respectively.)

Clearly, the nature of this arrangement (the independent cycling of suits and values) is unchanged if the deck is cut. Also, it makes it possible to work out which card occupies a given position i, simply by reducing i modulo m and n and interpreting the results as a suit and a value. For example, in the sequence above we would have ♠ = 1, ◇ = 2, ♣ = 3, and ♡ = 4, as well as giving Aces, Jacks, Queens, and Kings the more-or-less customary values of 1, 11, 12, and 13. If we then want to identify, say, the thirtieth card (starting the count from ♠A), the calculation would be: $30 \equiv 2 \pmod 4$, so the card is a diamond; and $30 \equiv 4 \pmod{13}$, so the card is a four; hence, the thirtieth card is ◇4. (If we want to work out where a given card is located, that is of course possible as well: Take ◇4 as an example. It is a four, so it must be in position 4, $17 = 4 + 13$, $30 = 4 + 2 \cdot 13$, or $43 = 4 + 3 \cdot 13$. And it is a diamond, so the position must be $\equiv 2 \pmod 4$. Of the numbers 4, 17, 30 and 43, this singles out 30.)

If, instead of simply taking the values in the straightforward A–K order, we pick some other arrangement, it is possible to make the deck look randomized and well-shuffled to anyone not paying *too* much attention, while still enabling calculations like the ones above, or simply making it easy to identify a missing card. One such arrangement is to let the values cycle through A, 4, 7, 10, K, 3, 6, 9, Q, 2, 5, 8, J (increasing in steps of 3 modulo 13), that is, sorting the deck as

♠A, ◇4, ♣7, ♡10, ♠K, ◇3, ♣9, ♡Q, ..., ♠2, ◇5, ♣8, ♡J.

This ordering of the deck is sometimes presented in books on card magic, and is known as *Si Stebbins* order.

In the case of a deck with an odd (composite) number of cards, the Chinese Remainder Theorem works well with the *Faro shuffle* (also known as the perfect riffle shuffle), and it is the purpose of this note to point this out. This is by no means a deep observation, but does not seem to have been made in the literature.

The Faro shuffle. The Faro shuffle is a method for shuffling a deck of cards where the deck is first divided into two packets. These are then meshed together, letting the cards interleave perfectly. For this to be possible, it is of course necessary that the two packets either contain the same number of cards, or differ in size by a single card. The

shuffle is named after the card game Faro (or Pharaoh), which was very popular in the 19th century, but seems to have gone out of style since. The rules are described by Morris [**8**, Ch. 1]. Also, game compendia such as Scarne's *Encyclopedia of Card Games* will generally contain a description, and of course there are numerous web sites devoted to the subject.

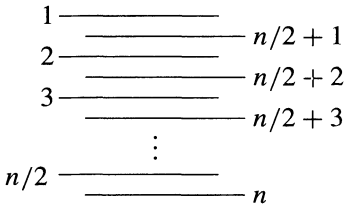
If we simply number the cards in an n-card deck from 1 through n, the Faro shuffle can be expressed as a permutation:

DEFINITION. Let n be a natural number.

(a) If n is even, the *out-Faro* shuffle on an n-card deck is the perfect riffle shuffle that leaves the top and bottom cards in place. As a permutation, it is given by

$$f(i) = \begin{cases} 2i - 1, & \text{if } i \le n/2 \\ 2i - n, & \text{if } i > n/2. \end{cases}$$

The cards are then rearranged as



(b) If n is even, the *in-Faro* shuffle on an n-card deck is the perfect riffle shuffle that shuffles the top and bottom cards into the deck. As a permutation, it is given by

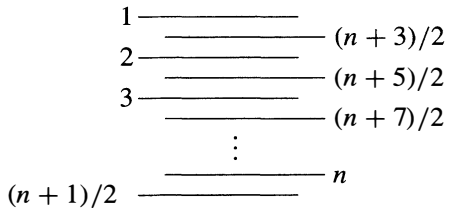
$$g(i) = \begin{cases} 2i, & \text{if } i \le n/2 \\ 2i - n - 1, & \text{if } i > n/2. \end{cases}$$

To see how this rearranges the cards, simply interchange the two columns of numbers in the illustration to (a) above.

(c) If n is odd, the *Faro* or *straddle* shuffle is the perfect riffle shuffle. It will leave either the top or bottom card in place, and if the invariant card is number 1, the permutation is

$$h(i) = \begin{cases} 2i - 1, & \text{if } i \le (n+1)/2 \\ 2i - n - 1, & \text{if } i > (n+1)/2. \end{cases}$$

The rearrangement of the cards is as follows:



We see that the three types of Faro shuffles are closely related: If we simply ignore the top and bottom cards in the out-Faro shuffle, or the invariant card in the straddle shuffle, the remaining cards are subjected to an in-Faro shuffle. And moreover, we see

from the definition that the in-Faro shuffle can be described very easily in terms of modular arithmetic: The card in position i is moved to position $2i \pmod{n+1}$. In particular, the *order* of the in-Faro shuffle on n cards (n even), that is, the number of times the shuffle must be performed before the deck is back in its original order, equals the order of 2 modulo $n+1$, that is, the smallest natural number d such that $2^d \equiv 1 \pmod{n+1}$.

Thus we have

PROPOSITION. (a) *If n is even, the order of the n-card out-Faro shuffle equals the order of 2 modulo $n-1$.*

(b) *If n is even, the order of the n-card in-Faro shuffle equals the order of 2 modulo $n+1$.*

(c) *If n is odd, the order of the n-card Faro shuffle equals the order of 2 modulo n.*

For instance, the out-Faro on an ordinary 52-card deck has order 8, since $2^8 \equiv 1 \pmod{51}$ and $2^4 \not\equiv 1 \pmod{51}$; the in-Faro has order 52. This last fact was observed by Johnson [**6**]. That eight out-Faroes will restore a deck of cards to its original order is well-known in the magical literature; a thorough description—complete with tables listing the ordering of the deck after each shuffle—can be found in Hugard and Braue [**5**, Ch. I.16].

An alternative description of Faro shuffles was given recently by Scully [**10**], using binary expansions. This description is particularly well suited to studying the cycle structure of the Faro shuffle, that is, subsets of the deck that are preserved under the shuffle.

Note. The Faro shuffle is not completely trivial to perform. However, the knack can be acquired with practice, should one wish to do so. A good reference for learning it is Appendix 2 of Morris [**8**], where he explains a Faro shuffle with numerous illustrations. It is also possible to find descriptions on-line, simply by searching on "Faro shuffle." The best way is probably to find someone who can do it, and get them to demonstrate.

For an even-numbered deck, the two possible Faro shuffles (in and out) generally speaking do not mix well, and tend to generate a fairly large group of permutations of the deck. This is described in detail by Diaconis, Graham, and Kantor [**3**].

For an odd-numbered deck, on the other hand, the two straddle shuffles work well together, as shown by Golomb [**4**]: If the cards are numbered as $0, \ldots, n-1$ instead of $1, \ldots, n$, one of the shuffles is the map $i \mapsto 2i \pmod{n}$, as for the in-Faro above (card number 0 being the invariant card), and the other is

$$i \mapsto (n-1) - 2[(n-1) - i] \equiv 2i + 1 \pmod{n}.$$

This means that performing k straddle shuffles gives a map of the form $i \mapsto 2^k i + a \pmod{n}$ for some a. In particular, straddle shuffles can be mixed with cuts, that is, with maps of the form $i \mapsto i + c \pmod{n}$, and still yield basically the same outcome as the shuffles alone. (This is manifestly *not* true for an even number of cards: Golomb [**4**] shows that Faro shuffles and cuts together produce *all* possible permutations of an even-numbered deck.)

Reversing the deck. As mentioned, 2 has order 52 modulo 53. A consequence of this is that $2^{26} \equiv -1 \pmod{53}$, which again means that twenty-six in-Faroes will reverse the order of a 52-card deck. (This is never used by magicians: Even with practice, it takes some seven minutes, and would seriously slow down any performance.)

More generally, in-Faroes can be used to reverse the order of an n-card deck (for n even) whenever -1 is a power of 2 modulo $n+1$. For instance, if $n = 2^k$ the in-Faro has order $2k$, and k in-Faroes will reverse the order. (The out-Faro has order k in this case, and the group generated by the in- and out-Faroes is quite small and easy to handle. See Lemma 4 in Diaconis, Graham, and Kantor [3] for details.) For decks with ≤ 52 cards, an in-Faro reversal can be performed for $n = 2$, 4, 8, 10, 12, 16, 18, 24, 26, 28, 32, 36, 40, 42, and 52.

The Chinese Remainder Theorem, once again. Let N be an odd number, factored as $N = mn$. If $\gcd(m, n) = 1$, multiplication by 2 modulo N corresponds to multiplication by 2 modulo both m and n: Consider the 3×5 grid from before, and let us number the squares from 0 to 14 instead of from 1 to 15:

0	6	12	3	9
10	1	7	13	4
5	11	2	8	14

Numbering the rows from 0 to 2, and the columns from 0 to 4, we then again have that the numbers in the ith row are congruent to i modulo 3, and that the numbers in the jth column are congruent to j modulo 5.

Now, multiplying all the numbers by 2, we get

0	12	9	6	3
5	2	14	11	8
10	7	4	1	13

Thus, the numbers in the ith row are now congruent to $2i$ modulo 3, and the numbers in the jth column are congruent to $2j$ modulo 5.

This will work in general: If the numbers in the ith row were all congruent to i modulo m, they will of course be congruent to $2i$ modulo m after we multiply by 2, and similarly for the columns. In particular, this shows that the order of 2 modulo N must be the least common multiple of the orders modulo m and n: Let A, B, and C denote the order of 2 modulo m, n, and N, respectively. Then we must have $A \mid C$ to ensure that the numbers in the ith row are congruent to i modulo m, and $B \mid C$ to ensure that the numbers in the jth column are congruent to j modulo n. The smallest C satisfying these two conditions is $C = \operatorname{lcm}(A, B)$.

In terms of Faro shuffles, this result can be expressed as follows: Suppose again that the deck consists of m suits, with n values of each, where $\gcd(m, n) = 1$. We then arrange the deck in accordance with the Chinese Remainder Theorem as before, with the suits and the values running through their cycles independently. For instance, if we consider an ordinary deck with the spades removed (hence $N = 39 = 3 \cdot 13$), it could be sorted as

♣A, ◇2, ♡3, ♣4, ◇5, ♡6, ♣7, ◇8, ♡9, ♣10, ◇J, ♡Q, ♣K,
◇A, ♡2, ♣3, ◇4, ♡5, ♣6, ◇7, ♡8, ♣9, ◇10, ♡J, ♣Q, ◇K,
♡A, ♣2, ◇3, ♡4, ♣5, ◇6, ♡7, ♣8, ◇9, ♡10, ♣J, ◇Q, ♡K.

Now, performing a Faro shuffle (of either kind) on this deck is, in a way, equivalent to performing m Faro shuffles, one for each packet of n cards: If we ignore the suits and simply consider the deck to be $1, 2, 3, \ldots, n, 1, 2, 3, \ldots$, the deck after the shuffle

consists of m packets of n cards, each of which is ordered as $1, 2, 3, \ldots, n$ would be after a Faro shuffle. With the 39 cards above, the result would be

$$\begin{array}{l}\clubsuit\text{A},\ \heartsuit 8,\ \diamondsuit 2,\ \clubsuit 9,\ \heartsuit 3,\ \diamondsuit 10,\ \clubsuit 4,\ \heartsuit\text{J},\ \diamondsuit 5,\ \clubsuit\text{Q},\ \heartsuit 6,\ \diamondsuit\text{K},\ \clubsuit 7,\\ \heartsuit\text{A},\ \diamondsuit 8,\ \clubsuit 2,\ \heartsuit 9,\ \diamondsuit 3,\ \clubsuit 10,\ \heartsuit 4,\ \diamondsuit\text{J},\ \clubsuit 5,\ \heartsuit\text{Q},\ \diamondsuit 6,\ \clubsuit\text{K},\ \heartsuit 7,\\ \diamondsuit\text{A},\ \clubsuit 8,\ \heartsuit 2,\ \diamondsuit 9,\ \clubsuit 3,\ \heartsuit 10,\ \diamondsuit 4,\ \clubsuit\text{J},\ \heartsuit 5,\ \diamondsuit\text{Q},\ \clubsuit 6,\ \heartsuit\text{K},\ \diamondsuit 7.\end{array}$$

Here, the sequence A, 8, 2, 9, 3, 10, 4, J, 5, Q, 6, K, 7 is exactly the result of a straddle shuffle of A, 2, 3, ..., K. Similarly, the suits are shuffled: $\clubsuit\heartsuit\diamondsuit$ is the result of shuffling $\clubsuit\diamondsuit\heartsuit$. (Since the cards in the table are organized in three rows of thirteen cards, rather than the other way around, the arrangement of the suits before and after is not as obvious as the arrangement of the values.)

In particular, cards that are n apart (counted cyclically) will have the same value after the shuffle, just as they had before, and cards that are m apart will have the same suit. This property is also (of course) preserved by cuts. Hence, a deck arranged in this way can be cut and straddle shuffled arbitrarily, and still be an arranged deck.

Remark. This above principle, for $51 = 3 \times 17$, is the basis for the "Chart of Seventeen" described by Hugard and Braue [**5**, Ch. I.16]: When performing an ordinary out-Faro, there are seventeen groups of cards that are preserved by the shuffle; these correspond to residue classes modulo 17, except that the group containing the top and bottom cards has four cards in it. (But that hardly matters, since the top and bottom cards are unaffected by the shuffle.)

Example. Take an ordinary deck and remove the spades. Order the remaining 39 cards as above, cycling through suits and values independently. The sequences can be picked at random, for example as

$$\begin{array}{l}\clubsuit 4,\ \diamondsuit\text{A},\ \heartsuit 7,\ \clubsuit 8,\ \diamondsuit\text{K},\ \heartsuit 2,\ \clubsuit 6,\ \diamondsuit 9,\ \heartsuit\text{Q},\ \clubsuit 3,\ \diamondsuit 5,\ \heartsuit 10,\ \clubsuit\text{J},\\ \diamondsuit 4,\ \heartsuit\text{A},\ \clubsuit 7,\ \diamondsuit 8,\ \heartsuit\text{K},\ \clubsuit 2,\ \diamondsuit 6,\ \heartsuit 9,\ \clubsuit\text{Q},\ \diamondsuit 3,\ \heartsuit 5,\ \clubsuit 10,\ \diamondsuit\text{J},\\ \heartsuit 4,\ \clubsuit\text{A},\ \diamondsuit 7,\ \heartsuit 8,\ \clubsuit\text{K},\ \diamondsuit 2,\ \heartsuit 6,\ \clubsuit 9,\ \diamondsuit\text{Q},\ \heartsuit 3,\ \clubsuit 5,\ \diamondsuit 10,\ \heartsuit\text{J}.\end{array}$$

As observed, giving the deck any number of shuffles and cuts will not change the nature of this arrangement, and the cuts can of course be performed by somebody else.

To turn this into a magic trick, one can do as follows: After enough shuffles and cuts to satisfy the spectator, ask him to cut the deck as often as he pleases and then take the top card. We may assume the cards to be ordered as above, from top to bottom, making his card the $\clubsuit 4$. Then ask him to pick up about half of the deck. You now pick up the rest, casually fanning out the cards with one hand. This shows you the $\heartsuit\text{J}$ at the bottom, and further up the $\diamondsuit\text{J}$ followed by the $\heartsuit 4$ and $\clubsuit\text{A}$, telling you that his card is a four, and a club, that is, the $\clubsuit 4$. You then ask him to shuffle his card into his packet, illustrating by folding your cards together, taking the top card, inserting it into your packet, and shuffling. When you have both shuffled your packets, give yours to the spectator and ask him to shuffle them together. Since you already know his card, he can pretty much do as he pleases with the deck, and you can then proceed to reveal the card in whatever way you prefer.

(Arranging the values as in the Si Stebbins order, with the cards increasing by a fixed amount modulo 13, will unfortunately not work here: All the numbers $1, \ldots, 12$ are powers of 2 modulo 13, so no matter what step-value you start with, eventually the cards will just be in order A, 2, ..., Q, K, and if the spectators see the deck at that point, it will be obvious that the cards are sorted.)

Remark. If you have the memory for it, this trick can be performed with 51 cards as well. Just leave, say, the $\heartsuit\text{K}$ in the box, and declare the $\heartsuit\text{A}$–Q to be the 14 through 17

of the other three suits, in some systematic way. Then you have redefined the deck to consist of three suits, with seventeen values each, and everything proceeds as above.

Example. Once people have learned the mathematics of the Faro shuffle, one of its standard uses (getting the cards in a pre-arranged order while appearing to shuffle them) will of course fall flat. However, it is possible to do a Faro trick that *depends* on the audience knowing the math:

Take an ordinary deck and remove the spades. Order the remaining 39 cards as follows, with the ♣A visible:

♣A, ◇2, ♡3, ♣4, ◇5, ♡6, ♣7, ◇8, ♡9, ♣10, ◇J, ♡Q, ♣K,
◇A, ♡2, ♣3, ◇4, ♡5, ♣6, ◇7, ♡8, ♣9, ◇10, ♡J, ♣Q, ◇K,
♡A, ♣2, ◇3, ♡4, ♣5, ◇6, ♡7, ♣8, ◇9, ♡10, ♣J, ◇Q, ♡K.

This is the same arrangement used above. Spread out the cards for all to see, pointing out the ordering (of both suits and values). Then perform a straddle shuffle, preserving the ♣A, and ask what the order of the shuffle is. While this is worked out, you perform an additional five shuffles, making sure to count them out loud.

The Faro shuffle on 39 cards has order 12, which is perhaps most easily seen by noting that 2 has order 12 modulo 13 and order 2 modulo 3.

Now point out that the shuffle you have performed is essentially an in-Faro on 38 cards, with the ♣A playing no real role. Also, note that sometimes performing in-Faroes will reverse the order of a deck. Obviously, if such a reversal happens after d in-Faroes, then the order of the in-Faro must be $2d$.

You have performed six in-Faroes on the 38 cards, and the order of the in-Faro is 12. So, have the 38 cards reversed their order?

No, since $2^6 = 64 \equiv 25 \not\equiv -1 \pmod{39}$.

However, when you spread out the cards again, they are seen to be in the order

♣A, ◇K, ♡Q, ♣J, ◇10, ♡9, ♣8, ◇7, ♡6, ♣5, ◇4, ♡3, ♣2,
◇A, ♡K, ♣Q, ◇J, ♡10, ♣9, ◇8, ♡7, ♣6, ◇5, ♡4, ♣3, ◇2,
♡A, ♣K, ◇Q, ♡J, ♣10, ◇9, ♡8, ♣7, ◇6, ♡5, ♣4, ◇3, ♡2.

This should fool most people, who are unlikely to notice that only the values have reversed their order, while the suits are in exactly the same sequence as before.

(And if you do not want to explain it afterwards, just make it an exercise.)

REFERENCES

1. I. Adler, Make up your own card tricks, *J. Recr. Math.* **6** (1973) 87–91.
2. S. M. Cohen and P. R. Coe, Card shuffling in discrete mathematics, *Coll. Math. J.* **26** (1995) 224–227.
3. P. Diaconis, R. L. Graham, and W. M. Kantor, The mathematics of perfect shuffles, *Adv. Appl. Math.* **4** (1983) 175–196.
4. S. W. Golomb, Permutations by cutting and shuffling, *SIAM Review* **3** (1961) 293–297.
5. J. Hugard and F. Braue, *Expert Card Technique: Close-Up Table Magic*, Dover, New York, 1974 (reprint of the 1944 edition).
6. P. B. Johnson, Congruences and card shuffling, *Amer. Math. Monthly* **63** (1956) 718–19.
7. S. Medvedoff and K. Morrison, Groups of perfect shuffles, this MAGAZINE **60** (1987) 3–14.
8. S. B. Morris, *Magic Tricks, Card Shuffles and Dynamic Computer Memories*, MAA, 1998.
9. S. Ramnath and D. Scully, Moving cards from position i to position j with perfect shuffles, this MAGAZINE **69** (1996) 361–365.
10. D. Scully, Perfect shuffles through dynamical systems, this MAGAZINE **77** (2004) 101–117.

The Lost Cousin of the Fundamental Theorem of Algebra

TIMO TOSSAVAINEN
Department of Teacher Education
Joensuu University
Savonlinna, Finland
timo.tossavainen@joensuu.fi

Have you ever reflected on the mystery of learning mathematics when you are struggling with a new concept or theorem, and suddenly, it may show up from a new angle and reveal something familiar that you can relate to? Thereafter, the pieces of an unfinished puzzle usually fall nicely into place. This is a story of a couple of such moments and of conquering an intriguing theorem on exponential functions.

The other day my elderly colleague asked me to check whether a lemma, which we would need elsewhere, is true in certain cases. The lemma was the following one.

LEMMA. *For $n \geq 2$, let $\kappa_1 > \cdots > \kappa_n > 0$ and $t_1 > \cdots > t_{n-1} \geq 0$ and suppose that $a_1, \ldots, a_n$ are real numbers with $a_1 > 0$. If the function*

$$f(t) = \sum_{j=1}^{n} a_j \kappa_j^t$$

satisfies $f(t_1) = \cdots = f(t_{n-1}) = 0$, then $f(t) > 0$ for all $t > t_1$.

The lemma seemed to be just another technical proposition, one of many, belonging to the folklore of real analysis. But since I have taken some courses in real analysis, naturally I accepted the challenge.

Rather soon I was able to prove it for $n = 2, 3, 4$, the cases that were the most interesting with respect to our linear algebraic research. And since it was only a technical lemma, I am a little embarrassed to say it now, I just thought to leave it at that.

But the lemma did not leave me in peace. I had a constant feeling that I hadn't yet figured out the deepest essence of the lemma. So, I had to return to it and find out what could have escaped my notice. Sooner or later, something made me think of the fundamental theorem of algebra and its well-known consequence, namely, that a polynomial of degree n has at most n roots. Then, almost immediately, the lemma started to take another shape. And now, if you think carefully enough, you certainly notice that it is very closely related to the following conjecture.

CONJECTURE. *For $n \in \mathbb{N}$ and $j = 0, \ldots, n$, let $0 < \kappa_0 < \cdots < \kappa_n$ and $a_j \in \mathbb{R}$ so that $a_n \neq 0$. Then the function $f : \mathbb{R} \to \mathbb{R}$,*

$$f(t) = \sum_{j=0}^{n} a_j \kappa_j^t$$

has at most n zeros.

There it was! A statement that is very similar to the famous result on the number of the roots of a polynomial. Recall that a polynomial of degree n is a sum of n power functions and a constant function, possibly with some coefficients equal to zero. In other words, the similarity lies in the relation between the number of the terms in the sum and the upper bound for the number of the zeros.

At this point, I remembered that I had proved the lemma only for $n = 2, 3, 4$, not for all $n \geq 2$. The conjecture would not be of any interest if it held only for a few values of n. How should I proceed now?

The first thing that came to my mind was to ask whether a sum of $n + 1$ exponential functions could be presented, similarly as a polynomial of degree n, as a product of at most n simple terms whose zeros are easily detected. To my disappointment, I was not able to find such a presentation. My second thought then was to try to construct an example of a sum of $n + 1$ exponential functions that has exactly n zeros; or even more which would mean the failure of the conjecture.

A method for constructing examples. Now, clearly, an example to be constructed would be most useful if it applied for all $n \geq 1$. Hence I decided to try a recursive method.

Soon I noticed that I have to overcome at least two problems. First, suppose that

$$f_k(t) = a_0\kappa_0^t + \cdots + a_k\kappa_k^t$$

has exactly l zeros $t_1 < \cdots < t_l$ which we may know or not. Now, adding $a_{k+1}\kappa_{k+1}^t$ to f_k implies that $f_{k+1}(t_j) = 0$ does not hold anymore for $j = 1, \ldots, l$. But on the other hand, if $|a_{k+1}\kappa_{k+1}^t|$ were small enough on the interval $[t_1, t_l]$, then the number of the zeros might be controlled somehow when shifting from f_k to f_{k+1}. However, this approach requires that we know the interval where the zeros of f_k lie.

The second problem is that usually we do not know the zeros of f_k and thus the interval $[t_1, t_l]$. This is simply due to the fact that, in most cases, we are not able to find a complete solution to the equation $f_k(t) = 0$. Fortunately, it is possible to gain some information on the zeros of f_k simply by exploring how f_k changes sign. Recall that f_k is continuous for every $k \geq 0$.

For the sake of simplicity, I also decided to use only exponential functions whose bases are at least one. Here is what I did.

For $1 \leq \kappa_0 < \kappa_1$, it is easy to find constants $a_0, a_1 \in \mathbb{R}$ so that

$$f_1(t) = a_0\kappa_0^t + a_1\kappa_1^t$$

has exactly one zero. Suppose then that

$$f_k(t) = a_0\kappa_0^t + \cdots + a_k\kappa_k^t,$$

where $1 \leq \kappa_0 < \cdots < \kappa_k$, has at least k zeros $t_1 < \cdots < t_k$ whose existence has been verified by observing that, for some $0 \leq \delta_0 < \cdots < \delta_k$,

$$f_k(\delta_j) = (-1)^j y_j, \tag{1}$$

in which all y_j's have the same sign and

$$b_k = \min\{|y_j| : j = 0, 1, \ldots, k\} > 0.$$

If we now fix $\kappa_{k+1} > \kappa_k$ and choose a_{k+1} so that a_k and a_{k+1} have different signs and

$$|a_{k+1}| \leq \frac{b_k}{2\kappa_{k+1}^{\delta_k}},$$

then, by the triangle inequality and the fact that κ_{k+1}^t is increasing, $f_{k+1}(\delta_j)$ and $f_k(\delta_j)$ have the same sign and

$$|f_{k+1}(\delta_j)| \geq \frac{b_k}{2} \tag{2}$$

for every $j = 1, \ldots, k$. This means that, similar to (1), also $f_{k+1}(\delta_j)$'s oscillate about the t-axis but never equal zero. Hence f_{k+1} has at least k zeros in $[\delta_0, \delta_k]$. Moreover, since κ_{k+1}^I is not bounded, there exists $\delta_{k+1} > \delta_k$ so that

$$\frac{f_{k+1}(\delta_{k+1})}{f_{k+1}(\delta_k)} \leq -1. \tag{3}$$

This implies that f_{k+1} has another zero in $[\delta_k, \delta_{k+1}]$. So we conclude that f_{k+1} must have at least $k+1$ zeros in $[\delta_0, \delta_{k+1}]$ altogether. Further, (2) and (3) guarantee that we can proceed with adding new exponential functions as many times as we want to. Here is an explicit example.

EXAMPLE. For any natural number n and $t \in \mathbb{R}$, let

$$f_n(t) = 1^t - \frac{2^t}{2} + \frac{4^t}{4^2} - \frac{8^t}{8^3} + \cdots + (-1)^n 2^{n(t-n)} = \sum_{j=0}^{n} (-1)^j 2^{j(t-j)}.$$

I claim that f_n alternates sign at the sequence $t = 0, 2, 4, \ldots, 2n$. Let us verify this for all natural numbers n. For $t = 0$, the leading term is 1^t and it dominates the sum of the absolute values of the other terms. Similarly, for $t = 2n$ the last term dominates. For any even integer $t = 2k$ with $1 \leq k \leq n-1$, the three terms $j = k-1, k, k+1$ sum to zero. The terms for the leading tail, that is the terms $j = 0, \ldots, k-2$ (if any exist), have the sum of their absolute values dominated by the term $j = k-2$. Likewise the terms for the trailing tail, the terms $j = k+2, \ldots, n$ (if any exist), have the sum of their absolute values dominated by the term $j = k+2$. Thus the sum of the two tails has the same sign as the term $j = k$, specifically $(-1)^k$, because the terms $j = k-2, k, k+2$ all have the same sign.

For instance, for $n = 5$, we have

$$f_5(t) = 1^t - \frac{2^t}{2} + \frac{4^t}{4^2} - \frac{8^t}{8^3} + \frac{16^t}{16^4} - \frac{32^t}{32^5}$$

and

$$\begin{aligned}
f_5(0) &= 1 - 2^{-1} + 2^{-4} - 2^{-9} + 2^{-16} - 2^{-25} > 0, \\
f_5(2) &= 1 - 2^1 + 2^0 - 2^{-3} + 2^{-8} - 2^{-15} < 0, \\
f_5(4) &= 1 - 2^3 + 2^4 - 2^3 + 2^0 - 2^{-5} > 0, \\
f_5(6) &= 1 - 2^5 + 2^8 - 2^9 + 2^8 - 2^5 < 0, \\
f_5(8) &= 1 - 2^7 + 2^{12} - 2^{15} + 2^{16} - 2^{15} > 0 \\
f_5(10) &= 1 - 2^9 + 2^{16} - 2^{21} + 2^{-24} - 2^{25} < 0.
\end{aligned}$$

Clearly, f_5 has at least 5 zeros.

All in all, I was able to construct a sum of $n+1$ exponential functions that has at least n zeros but not able to confirm that it had no more than n zeros. An important conclusion follows: if the conjecture is true, it is sharp in the sense that the upper bound for the number of zeros is the best possible.

A well-established example often leads the way to the proof. It is nearly a law of nature in mathematics that a well-motivated example fitting a theorem is half the battle of proving the theorem. Therefore, I decided to revise the above construction once again.

First I noticed that the recursive method I had chosen is based on the idea of mathematical induction. Hence, if I could verify the statement of the conjecture in the case $n = k + 1$ using the same statement for $n = k$, the proof would be essentially done. Moreover, whenever I had not been able to count the number of the zeros directly, I had managed by counting something else that is connected to the zeros. And now remember what Rolle's theorem implies. If you know the number of the zeros of the derivative of a differentiable function, then you can bound the number of the zeros of the original function. And again, almost miraculously, the above construction began to change into the argument that converts our conjecture into a theorem!

And now afterward, when I reflect on the proof, I still find it very elegant. The elegance, in my opinion, comes from the fact that it relies only on a few very basic results of classical real analysis and still it reveals quite an interesting property of exponential functions.

To my surprise, I have not been able to find this result in the literature. However, I imagine that several mathematicians may have noticed it in the course of history, and yet we do not know if any actually did. Anyway, I take the liberty to name this theorem and call it *the lost cousin of the fundamental theorem of algebra* due to an apparent resemblance between that famous result and our theorem.

The proof. Let us consider first the case $n = 1$. By writing $\lambda_1 = \kappa_1/\kappa_0$, we have

$$f(t) = \kappa_0^t\left(a_1\lambda_1^t + a_0\right) = \kappa_0^t g(t),$$

where $f(t) = 0$ if and only if $g(t) = 0$. Since $\lambda_1 > 1$ and $a_1 \neq 0$, there is at most one value $t = t_1$ such that $g(t_1) = 0$.

Assume then that the claim holds for some $n = k \geq 1$. Similar to the above, we denote $\lambda_j = \kappa_j/\kappa_0$ in order to have $f(t) = \kappa_{k+1}^t g(t)$, where

$$g(t) = \sum_{j=1}^{k+1} a_j\lambda_j^t + a_0,$$

and $\lambda_{k+1} > \cdots > \lambda_1 > 1$. Again, $f(t) = 0$ if and only if $g(t) = 0$.

Now, for all $t \in \mathbb{R}$, the derivative of g is

$$g'(t) = \sum_{j=1}^{k+1} a_j(\ln\lambda_j)\lambda_j^t = \sum_{j=0}^{k} \beta_j\mu_j^t,$$

where $\mu_j = \lambda_{j+1}$, $\beta_j = a_{j+1}\ln\lambda_{j+1}$ and $\beta_k \neq 0$. By the assumption, there exist at most k distinct numbers $\delta_1 < \cdots < \delta_k$ such that $g'(\delta_1) = \cdots = g'(\delta_k) = 0$. Thus, by Rolle's theorem, there are at most $k + 1$ distinct numbers $t_1 < \cdots < t_{k+1}$ such that $g(t_1) = \cdots = g(t_{k+1}) = 0$. The theorem follows. ∎

There is still one thing to tell. My elderly colleague, who seems to know me quite well, reminds me every now and then that after having proved a theorem I should always check whether there is another one around the corner. So, let us consider the following question. Since exponential and logarithm functions share many important analytical properties (all of them are continuous, differentiable, integrable and, except for the constant function, strictly monotone etc.), do we find another lost cousin by replacing the exponential functions with logarithm functions (to different bases) in our theorem?

Well, the logarithm functions have certain arithmetical properties that eventually forces us to answer the question with "No". Let us consider the sum

$$g_n(t) = \sum_{j=0}^{n} a_j \log_{\kappa_j} t,$$

where $1 < \kappa_0 < \kappa_1 < \cdots < \kappa_n$ and a_j's are real numbers so that $a_n \neq 0$. Changing logarithms to the same base e gives us

$$g_n(t) = \sum_{j=0}^{n} a_j \frac{\ln t}{\ln \kappa_j} = \ln t^{\alpha},$$

where

$$\alpha = \sum_{j=0}^{n} \frac{a_j}{\ln \kappa_j}.$$

Now, depending on whether α is greater than, less than, or equal to zero, g_n is, respectively, strictly increasing with $g_n(1) = 0$, strictly decreasing with $g_n(1) = 0$, or $g_n(t) = 0$ for every $t > 0$.

Acknowledgments. The author wishes to thank one referee for improving the original example, and all the referees for their useful comments.

REFERENCES

1. R. Courant and Fr. John, *Introduction to Calculus and Analysis I*, Springer, Berlin, 1999.
2. S.W. Drury, J.K. Merikoski, V. Laakso and T. Tossavainen, On Nonnegative Matrices with Given Row and Column Sums. *Electron. J. Linear Algebra* **10** (2003) 280–290.

Not Mixing Is Just as Cool

SAM NORTHSHIELD
SUNY Plattsburgh
Plattsburgh, NY 12901

Newton's law of cooling is a staple of the Calculus curriculum; it is usually presented as a first or second example of a separable differential equation. In that context, the law states that the rate of change of the temperature T of, say, a quantity of fluid is proportional to the difference between the fluid's temperature and the ambient temperature T_∞:

$$\frac{dT}{dt} = -k(T - T_\infty). \tag{1}$$

This is easily solved (part of the difficulty in solving it is dealing with initial conditions):

$$T(t) = T_\infty + (T_0 - T_\infty)e^{-kt} \tag{2}$$

where $T_0 := T(0)$ is the temperature at time $t = 0$.

The following problem is, for many students, a challenging application of Newton's law even given the formula (2).

PROBLEM. *Which results in a cooler drink:*

(1) *Pour a cup of coffee, wait five minutes, and then add an ounce of cold milk or*

(2) *Pour a cup of coffee, add an ounce of cold milk, and then wait five minutes?*

One challenging aspect of this problem is that a law for the temperature of mixed fluids must be either known *a priori* or else invented during the solution of the problem. We shall give it. If we mix two fluids (that are thermodynamically "similar") then the temperature of the mixed fluid is the average of the temperatures but weighted according to quantity. For example 3 oz. of water at 100 degrees mixed with 2 oz. of water at 150 degrees results in 5 oz. of water at $(3 \cdot 100 + 2 \cdot 150)/5 = 120$ degrees. In general,

PRINCIPLE A. *If Q_A units of fluid at temperature T_A is mixed with Q_B units of fluid at temperature T_B, then the resulting mix has temperature*

$$T := \frac{T_A Q_A + T_B Q_B}{Q_A + Q_B}. \tag{3}$$

We will now solve the problem in two ways, one via Newton's law, the other through intuition.

Solution 1. Let T_c and T_m denote the initial temperatures of the coffee and milk respectively. Let's assume that a cup of coffee is the 'standard' 6 oz. though, as it turns out, this assumption will not affect the answer to the question.

In case 1, after 5 minutes of cooling, equation (2) predicts that the temperature of the coffee is

$$T(5) = T_\infty + (T_c - T_\infty)e^{-5k} \tag{4}$$

and, after mixing with milk, equation (3) predicts a final temperature of $(6T(5) + T_m)/7$ or

$$T_1 := \frac{1}{7}(6T_\infty + T_m) + \frac{6}{7}(T_c - T_\infty)e^{-5k}. \tag{5}$$

In case 2, mixing the coffee and milk yields 7 oz. of fluid at initial temperature $(6T_c + T_m)/7$ which, after cooling for 5 minutes, has temperature

$$T_2 := T_\infty + \frac{1}{7}(6T_c + T_m - 7T_\infty)e^{-5k}. \tag{6}$$

Taking the difference and simplifying, we find

$$T_2 - T_1 = \frac{1}{7}(1 - e^{-5k})(T_\infty - T_m). \tag{7}$$

Hence, assuming that "cold" means colder than the ambient temperature T_∞, Case 1 yields a cooler drink. ■

Note that equation (7) implies that the difference of temperature in Case 1 and Case 2 is independent of the temperature of the coffee. It is, rather, the relative amount of milk that makes the difference. That is, the 7 in equation (7) comes from the ratio 1:6 of milk to coffee. We would replace the 7 by 9 if we used an 8 oz. cup of coffee.

Consider now an 'intuitive' solution:

Solution 2. Suppose we take 1 oz. of milk and allow it to warm up for 5 minutes while simultaneously allowing 6 oz. of coffee to cool. Then we mix them. Since *it*

makes no difference if the fluids were already mixed or not, the temperature of the 7 oz. mix is the same as T_2. This is clearly warmer than if we do not allow the milk to warm up; that is, we keep it in the fridge until we mix it which is Case 1. Hence $T_1 < T_2$.

Although solution 2 is short and elegant, it lacks the rigor of solution 1 (especially in the phrase "since it makes no difference if the fluids are mixed or not"). It is not our goal to show the correctness of either method (not least because it is difficult to justify Newton's law or its assumptions thermodynamically) but rather to show that the two solutions are *equally* correct!

THEOREM. *Not only does Newton's law imply that "it makes no difference if the fluids are mixed or not" but also, if there is* ***any*** *law of cooling for which it "makes no difference if the fluids are mixed or not", then it must be Newton's Law.*

We shall henceforth assume that there is *some* law of cooling. What we mean by that is that given an ambient temperature T_∞ and an object (e.g., a quantity of fluid), the future temperature of the object depends only on its present temperature and the elapsed time. Furthermore, we shall insist it is reasonable that temperature changes monotonically and continuously, converging at ∞ to T_∞. A mathematical way of saying all this is that there exists *some* monotonic and continuous function f with limit T_∞ at ∞ such that

$$\text{If } T(0) = f(t_0) \text{ then, for all } t,\ T(t) = f(t_0 + t). \tag{8}$$

Select any two temperatures $T_1 > T_2 > T_\infty$. Also let us identify the unique choices t_1 and t_2 for which $T_1 = f(t_1)$ and $T_2 = f(t_2)$. Finally, let $r = (T_2 - T_\infty)/(T_1 - T_\infty)$. Upon clearing the denominator, we observe that $(1 - r)T_\infty + rT_1 = T_2$. Let us envision mixing two quantities of fluid, an amount r of temperature T_1 with a quantity $1 - r$ at room temperature T_∞. If we mix first, we get $rT_1 + (1 - r)T_\infty = T_2 = f(t_2)$, and then allowing the fluid to cool for a time period t, gives $f(t_2 + t)$. But waiting first gives an amount r of $f(t_1 + t)$ to be mixed with an amount $1 - r$ of T_∞ yielding $rf(t_1 + t) + (1 - r)T_\infty$. If mixing does not matter, these must be equal so

$$rf(t_1 + t) + (1 - r)T_\infty = f(t_2 + t).$$

This can be rewritten:

$$f(t_2 + t) - T_\infty = r(f(t_1 + t) - T_\infty). \tag{9}$$

Subtract $T_2 - T_\infty = r(T_1 - T_\infty)$, and divide by t to get

$$\frac{f(t_2 + t) - T_2}{t} = r\frac{f(t_1 + t) - T_1}{t}.$$

Assuming that f is differentiable, we may take the limit as $t \to 0$ to find

$$f'(t_2) = rf'(t_1). \tag{10}$$

Now set $t = 0$ in (9) and divide equation (10) by these equal quantities to get

$$\frac{f'(t_2)}{f(t_2) - T_\infty} = \frac{f'(t_1)}{f(t_1) - T_\infty}.$$

Since this holds for any time $t_2 > t_1$, the fractions must have a constant value, say $-k$,

$$\frac{f'(t_2)}{f(t_2) - T_\infty} = -k.$$

Upon integrating and exponentiating, we have

$$\begin{aligned} \ln(f(t_1 + t) - T_\infty) &= -kt + c \\ f(t_1 + t) &= T_\infty + Ce^{-kt} \\ T(t) = f(t_1 + t) &= T_\infty + (T_1 - T_\infty)e^{-kt}. \end{aligned}$$

We have in fact derived Newton's Law of Cooling. However in this approach we assumed, quite reasonably, that f is differentiable. We can avoid this assumption and succeed using the weaker assumption that f is continuous by using the technique of *functional equations*, a topic seldom seen in the undergraduate curriculum. To use this approach, set $s = t_2 - t_1$ in equation (9) and divide by the same equation with t replaced by 0 to get

$$\frac{f(t_1 + s + t) - T_\infty}{f(t_1 + s) - T_\infty} = \frac{f(t_1 + t) - T_\infty}{f(t_1) - T_\infty}.$$

Next multiply both sides by

$$\frac{f(t_1 + s) - T_\infty}{f(t_1) - T_\infty}$$

to give us

$$\frac{f(t_1 + s + t) - T_\infty}{f(t_1) - T_\infty} = \frac{f(t_1 + s) - T_\infty}{f(t_1) - T_\infty} \cdot \frac{f(t_1 + t) - T_\infty}{f(t_1) - T_\infty}.$$

This suggests that we can define a new function $g(t) := (f(t_1 + t) - T_\infty)/(f(t_1) - T_\infty)$ for $t \geq 0$ to reveal the functional equation

$$g(s + t) = g(s)g(t), \qquad g(0) = 1. \tag{11}$$

This equation is (one of several) known as Cauchy's equation (see [**1**]) and, assuming only the continuity of $g(x)$ it turns out that there exists a real number k such that

$$g(t) = e^{-kt}. \tag{12}$$

This is remarkable, since infinite differentiability then follows from the much weaker condition of continuity (in fact, it follows from the even weaker condition of boundedness on an interval and continuity at a single point!—see [**1**] for details). We sketch a proof assuming continuity for all x. As a first consequence, we can choose $c \neq 0$ with $g(c) > 0$. By induction and equation (11),

$$g(nc) = g(c)^n.$$

Since $g(c) = g(n(c/n)) = g(c/n)^n$, taking the nth root of both sides,

$$g(c/n) = g(c)^{1/n}.$$

Then $g(mc/n) = g(c/n)^m = g(c)^{m/n}$ and so, for every positive rational x,

$$g(xc) = g(c)^x. \tag{13}$$

Since $g(-xc)g(xc) = g(-xc + xc) = g(0) = 1$, equation (13) holds for all rational numbers x. By the continuity of g, the two functions on either side of equation (13) are defined and continuous for each real number x and agree on the dense set of rational

numbers, and so equation (13) holds for all x. Since there is some k such that $g(c) = e^{-kc}$, we may rewrite (13) as equation (12).

To apply this to the previous problem, recall that $g(t) := (f(t_1 + t) - T_\infty)/(f(t_1) - T_\infty)$ satisfies equation (11). Using its solution (12), and the definition $T(t) := f(t_1 + t)$, we may write

$$T(t) = T_\infty + (T(0) - T_\infty)e^{-kt}$$

for some k—Newton's Law!

Going the other way, assume Newton's law (2). Then two similar fluids at temperatures a and b respectively satisfy temperature laws

$$\begin{cases} T_1(t) = T_\infty + (a - T_\infty)e^{-kt} \\ T_2(t) = T_\infty + (b - T_\infty)e^{-kt} \end{cases}$$

respectively. If mixed in the proportion $r : (1 - r)$, the resulting fluid has temperature law:

$$T(t) := T_\infty + (ra + (1 - r)b - T_\infty)e^{-kt} = rT_1(t) + (1 - r)T_2(t)$$

which shows that mixing does not matter. ■

Equation (2) can also reasonably be called "Newton's law of heating" when $T(0) < T_\infty$. The extension to that case follows from Principle A and that mixing does not matter: if a fluid at initial temperature $T_1(0) = T_\infty + c$ and another at temperature $T_2(0) = T_\infty - c$ are mixed in equal proportions, then the mixed temperature is constant T_∞. Since the first fluid obeys Newton's law ($T_1(t) = T_\infty + ce^{-kt}$) and since mixing does not matter, $T_2(t) = T_\infty - ce^{-kt}$. Since $c = T_\infty - T_2(0)$,

$$T_2(t) = T_\infty + (T_2(0) - T_\infty)e^{-kt}.$$

We have purposely not addressed many of the assumptions necessary for Newton's law to give even a reasonable approximation to reality. That which makes Newton's law of cooling interesting (to Calculus teachers at least) is its simplicity and, as we tried to show, its inevitability given a few basic principles and a little knowledge of the Calculus or of functional equations.

What then was Newton's intuition? In his paper of 1701, written in Latin, no equations appear. He wrote, however (this quoted from [**2**]): "the iron was laid not in a calm air, but in a wind that blew uniformly upon it . . . for thus equal parts of air were heated in equal times, and received a degree of heat proportional to the heat of the iron." The results of this experiment in 'forced convection' led to the empirical law equivalent to (1) and (2). We recommend the paper [**2**] and its references for further information.

Acknowledgment. I thank the editor for extensive and detailed suggestions for the improvement of this article.

REFERENCES

1. J. Aczel, *Lectures on Functional Equations and their Applications*, Academic Press, New York, 1966.
2. R.H.S. Winterton, Newton's law of cooling, *Contemporary Physics*, **40**(3) (1999) 205–212.

Polynomial Congruences and Density

GERRY MYERSON
Macquarie University
NSW 2109 Australia
gerry@math.mq.edu.au

What do the solutions of a congruence look like, as the modulus varies? Let $f(t)$ be a polynomial with integer coefficients, let the solutions of $f(t) \equiv 0 \pmod{m}$, if there are any, be $r_1, r_2, \ldots, r_k$, with $0 \le r_j \le m-1$; how are the numbers $r_1, r_2, \ldots, r_k$ distributed within the set $\{0, 1, \ldots, m-1\}$? For a fixed f, how does the answer change as m increases?

In order to compare answers for different values of the modulus, it is convenient to divide through by the modulus. Thus, we'll define

$$S_f(m) = \{r/m : 0 \le r \le m-1, \gcd(r, m) = 1, \text{ and } m \text{ divides } f(r)\}$$

This makes $S_f(m)$ a (finite, possibly empty) subset of $[0, 1)$ for each m, so the sets for various m are directly comparable.

We'll also let S_f be the union of the sets $S_f(m)$ over all positive integers m. It is this set that will concern us.

Polynomials of degree zero are supremely boring in this context, so let's start by looking at polynomials of degree one. Let $f(t) = at + b$. If $at + b \equiv 0 \pmod{m}$, then $at + b = mv$ for some integer v, so

$$\frac{t}{m} = \frac{v}{a} - \frac{b}{am} = \frac{v}{a} + O(m^{-1}).$$

Thus, most of the points in S_f are very close to one or another of the points $0, 1/a, 2/a, \ldots, 1$. Moreover, if v is relatively prime to a then there will be infinitely many m such that $mv \equiv b \pmod{a}$, and for such m there will be a point in $S_f(m)$ close to v/a. Thus, we have a pretty good idea of what S_f looks like; very crowded near the points v/a with v relatively prime to a (and perhaps also near the other points of the form v/a—we encourage the reader to look at this more closely), very sparse everywhere else.

Matters get considerably more complicated when we go to quadratics and polynomials of yet higher degree. Hooley [**1**] proved that if f is an irreducible polynomial of degree at least 2, with integer coefficients, then the sequence formed by ordering S_f by increasing denominator is uniformly distributed in $[0, 1)$. What does that mean?

Consider, for example, the polynomial $f(t) = t^2 + 1$. The relevant sequence is then $0/1, 1/2, 2/5, 3/5, 3/10, 7/10, 5/13, 8/13, 4/17, 13/17, 7/25, 18/25, 5/26, 21/26$, etc. In fact, it makes no difference for our purposes how fractions with the same denominator are ordered, so long as the denominators are in nondecreasing order.

A sequence $u_1, u_2, \ldots$ of numbers in $[0, 1)$ is said to be uniformly distributed in $[0, 1)$ if in the limit each subinterval I of $[0, 1)$ contains a proportion of terms of the sequence equal to the length of I. That is to say, $u_1, u_2, \ldots$ is uniformly distributed in $[0, 1)$ means $\lim_{n\to\infty} \#\{i \le n : a \le u_i < b\}/n = b - a$ for all a and b with $0 \le a < b \le 1$. Chapter 21 of Roberts [**2**] provides a gentle introduction to the theory of uniformly distributed sequences.

Hooley's proof uses tools at a level beyond that suitable for this MAGAZINE. The main purpose of this paper is to prove a weaker, but still interesting, result, using only readily accessible methods.

THEOREM. *Let $f(t) = t^e g(t)$ where e is a nonnegative integer, g is a polynomial of degree at least 2 with integer coefficients, and $g(0) \neq 0$. Define T_f by*

$$T_f = \{r/m : \gcd(r, m) = 1, \text{ and } m \text{ divides } f(r)\}.$$

Then T_f is dense in the reals.

It is easy to see that if a sequence is uniformly distributed in $[0, 1)$ then the set underlying the sequence is dense in $[0, 1)$. Our theorem is thus an immediate consequence of Hooley's, provided only that f have an irreducible factor of degree at least two. We will prove it without this requirement, and without reference to Hooley's result.

We note that the theorem is best possible, in the sense that if f is a polynomial with integer coefficients but doesn't satisfy the hypothesis of the theorem then it is easy to see from the introductory remarks on first degree polynomials that T_f is not dense in the reals.

Our proof proceeds along the following lines. Given f satisfying the hypotheses, and given a real number x, we choose rational numbers h_1/k_1 and h_2/k_2 close to x. Then $(h_1a + h_2b)/(k_1a + k_2b)$ is also close to x for all positive integers a and b. We show how to choose a and b in such a way that if $r = h_1a + h_2b$ and $m = k_1a + k_2b$ then m divides $f(r)$.

First we show that a "weighted mediant" of two fractions lies between the fractions and is in lowest terms:

LEMMA 1. *If h_1, h_2, k_1, k_2, a and b are positive integers and $h_1/k_1 < h_2/k_2$ then*

$$\frac{h_1}{k_1} < \frac{h_1a + h_2b}{k_1a + k_2b} < \frac{h_2}{k_2}.$$

Moreover, if $h_1k_2 - h_2k_1 = -1$ and a and b are relatively prime then $h_1a + h_2b$ and $k_1a + k_2b$ are relatively prime.

Proof. The inequalities are evident on viewing the weighted mediant as a weighted average of the two fractions h_1/k_1 and h_2/k_2 with positive weights k_1a and k_2b:

$$\frac{h_1a + h_2b}{k_1a + k_2b} = \frac{(k_1a)(h_1/k_1) + (k_2b)(h_2/k_2)}{k_1a + k_2b}$$

If $h_1k_2 - h_2k_1 = -1$, let $r = h_1a + h_2b$ and $m = k_1a + k_2b$ and solve for a and b; $a = h_2m - k_2r$, $b = k_1r - h_1m$. Now any common divisor of r and m divides both a and b, and the last assertion of the lemma follows. ∎

Next we show that good approximations to x can be chosen to satisfy certain divisibility and coprimality conditions.

LEMMA 2. *Given positive reals x and ϵ, a positive integer c, a nonzero integer c', and an integer $n \geq 2$, there are positive integers h_1, h_2, k_1, k_2 and d such that d is relatively prime to c', $k_2 = cd^{n-1}$, $h_1k_2 - h_2k_1 = -1$, and $|x - h_i/k_i| < \epsilon$ for $i = 1, 2$.*

Proof. Given any positive integer D, there is a positive integer H such that $|x - H/D^{n-1}| \leq 1/(2D^{n-1})$. Let $d = D^2$, let $h_2 = cD^{n-1}H + 1$, and let $k_2 = cd^{n-1} = cD^{2n-2}$; then

$$|x - h_2/k_2| \leq |x - H/D^{n-1}| + 1/(cD^{2n-2}) \leq 1/(2D^{n-1}) + 1/(cD^{2n-2}).$$

Moreover, $\gcd(h_2, k_2) = 1$, so there are positive integers h_1 and k_1 such that $h_1k_2 - h_2k_1 = -1$, which entails that

$$|x - h_1/k_1| \le |x - h_2/k_2| + h_2/k_2 - h_1/k_1 = |x - h_2/k_2| + 1/(k_1k_2)$$
$$\le 1/(2D^{n-1}) + 2/(cD^{2n-2}).$$

Now it suffices to choose D relatively prime to c' and large enough to ensure $1/(2D^{n-1}) + 2/(cD^{2n-2}) < \epsilon$. ∎

Finally, we need a version of the Remainder Theorem, which we present without proof.

LEMMA 3. *Given a polynomial $f(t)$ of degree n and real numbers a and b, $a \neq 0$, there is a polynomial $q(t)$ such that*

$$a^n f(t) = (at + b)q(t) + a^n f(-b/a).$$

Moreover, if a, b, and the coefficients of f are integers, then so are the remainder and the coefficients of q.

Proof of the Theorem. Let $f(t) = t^e g(t) = c_0t^n + c_1t^{n-1} + \cdots + c_n$ satisfy the hypotheses. Note that if T_g is dense then so is T_f, so we may assume without loss of generality that $e = 0$. Thus, $c_n \neq 0$. We may assume that c_0 is positive, since we may replace f with $-f$, if need be. If m divides $f(r)$ then also m divides $f(r + mQ)$ for any integer Q, so it is enough to prove that T_f is dense in the positive reals (indeed, in $[0, 1)$).

Let x and ϵ be positive. By Lemma 2, there are positive integers h_1, h_2, k_1, k_2, and d such that $\gcd(d, c_n) = 1$, $k_2 = c_0d^{n-1}$, $h_1k_2 - h_2k_1 = -1$, and $|x - h_i/k_i| < \epsilon$ for $i = 1, 2$. Note that $\gcd(k_1, k_2) = 1$, whence $\gcd(d, k_1) = 1$.

Recall that we want to choose positive integers a and b in such a way that if $r = h_1a + h_2b$ and $m = k_1a + k_2b$ then m divides $f(r)$. We claim this can be achieved by letting $b = d + k_1s$, where s is any multiple of c_n large enough to guarantee $k_1^n f(b/k_1) > k_2b$, and then defining a by $k_1^n f(b/k_1) = k_1a + k_2b$.

Clearly, a is positive. Moreover, a is an integer, because

$$a = c_1b^{n-1} + c_2b^{n-2}k_1 + \cdots + c_nk_1^{n-1} + b(c_0b^{n-1} - k_2)/k_1 \tag{1}$$

and

$$(c_0b^{n-1} - k_2)/k_1 = c_0(b^{n-1} - d^{n-1})/k_1 = c_0s(b^{n-1} - d^{n-1})/(b - d).$$

From $b = d + k_1s$ we deduce

$$\gcd(b, k_1) = \gcd(d, k_1) = 1. \tag{2}$$

We claim that a and b are relatively prime. For let p be a prime dividing both a and b. Then by (1) p divides c_n or k_1. By (2), p divides c_n. We chose s to be a multiple of c_n, so p divides s. We defined b by $b = d + k_1s$, so p divides d. But now we have reached a contradiction, as d was chosen relatively prime to c_n.

Let $r(t) = h_1t + h_2b$ and $m(t) = k_1t + k_2b$. By Lemma 1, $r(a)$ and $m(a)$ are relatively prime, and $|x - r(a)/m(a)| < \epsilon$. All that remains is to prove $m(a)$ divides $f(r(a))$.

By Lemma 3, there is a polynomial $q(t)$ with integer coefficients such that $k_1^n f\big(r(t)\big) = m(t)q(t) + k_1^n f\big(r(-k_2b/k_1)\big)$. Thus,

$$k_1^n f\big(r(t)\big) - m(t)q(t) = k_1^n f\big(r(-k_2b/k_1)\big) = k_1^n f\big(h_1(-k_2b/k_1) + h_2b\big)$$
$$= k_1^n f\big((-h_1k_2 + h_2k_1)(b/k_1)\big) = k_1^n f(b/k_1)$$
$$= k_1a + k_2b = m(a).$$

Evaluating at $t = a$, we see that $m(a)$ divides $k_1^n f(r(a))$. Any common prime divisor of $m(a)$ and k_1^n divides $k_2 b$, but k_1 is relatively prime to b by (2) and also to k_2. Thus $m(a)$ is relatively prime to k_1, so it divides $f(r(a))$, and we are done. ■

REFERENCES

1. Christopher Hooley, On the distribution of the roots of polynomial congruences, *Mathematika* **11** (1964) 39–49, MR 29 #1173.
2. Joe Roberts, *Elementary Number Theory*, MIT Press, 1977, MR 58 #16472.

A Curious Way to Test for Primes

DENNIS P. WALSH
Middle Tennessee State University
Murfreesboro, TN, 37132
dwalsh@mtsu.edu

Upon hearing that you are a student of mathematics, a cab driver says to you, "Check this out. The second derivative of e^x is e^x, right? And e^x evaluated at 0 is equal to 1, right? Therefore 2 has got to be a prime number." Your first reaction is a condescending chuckle for the cabby who seems to dabble in mathematics and appears to make a jumble of it. "Well, at least she's logically correct—2 is a prime number," you mumble to yourself with some smugness. But the cab driver hears you and takes a long route to your destination. You end up paying a hefty fare. The cab driver smirks as she drives off.

In fact, the smirking cabby stated a specific case of the theorem we provide below, a theorem which offers an unusual characterization of prime numbers based on differentiation. The cab driver could give another specific case by stating, "The third derivative of $e^x + e^{x^2/2}$ evaluated at $x = 0$ is 1, and thus 3 is a prime number." Offering another example, she could state accurately that, "The fourth derivative of $e^x + e^{x^2/2} + e^{x^3/3}$ evaluated at 0 does not equal 1, and hence 4 is not a prime number." You can probably guess the pattern. We give it below in a theorem with a surprisingly simple proof that uses the series expansion of $e^{x^k/k}$ and the power rule for differentiation.

THEOREM. *For each positive integer $n > 1$, define the function g_n by $g_n(x) = \sum_{k=1}^{n-1} e^{x^k/k}$. A positive integer n is prime if and only if $\frac{d^n}{dx^n} g_n(0) = 1$.*

Proof. Let n be a positive integer greater than 1. Note that $e^{x^k/k}$ has a series expansion given by

$$e^{x^k/k} = \sum_{j=0}^{\infty} \frac{(x^k/k)^j}{j!} = \sum_{j=0}^{\infty} \frac{x^{kj}}{k^j j!}$$

for all real x. Hence, we have

$$g_n(x) = \sum_{k=1}^{n-1} \sum_{j=0}^{\infty} \frac{x^{kj}}{k^j j!},$$

and upon differentiating g_n we obtain

$$\begin{aligned}\frac{d^n}{dx^n}g_n(x) &= \frac{d^n}{dx^n}\sum_{k=1}^{n-1}\sum_{j=0}^{\infty}\frac{x^{kj}}{k^j\,j!}\\ &= \sum_{k=1}^{n-1}\sum_{j=0}^{\infty}\frac{1}{k^j\,j!}\frac{d^n}{dx^n}(x^{kj})\\ &= \sum_{k=1}^{n-1}\sum_{j=0}^{\infty}\frac{(kj)(kj-1)(kj-2)\cdots(kj-n+1)x^{kj-n}}{k^j\,j!}I[kj\geq n],\end{aligned}$$

where the indicator function $I[s]$ takes the value 1 if statement s is true and the value 0 otherwise. Upon evaluating $\frac{d^n}{dx^n}g_n(x)$ at $x=0$, we see that every term in the inner sum above vanishes except for the terms where $kj=n$. Thus, using $k|n$ to denote the statement "k divides n," we get

$$\begin{aligned}\frac{d^n}{dx^n}g_n(0) &= \sum_{k=1}^{n-1}\frac{n(n-1)(n-2)\cdots(n-n+1)}{k^{n/k}(n/k)!}I[k|n]\\ &= \sum_{k=1}^{n-1}\frac{n!}{k^{n/k}(n/k)!}I[k|n].\end{aligned}$$

If n is prime, the only divisor of n that is less than or equal to $n-1$ is 1, in which case the summation above collapses to the single term for $k=1$. Hence, when n is prime, we obtain

$$\frac{d^n}{dx^n}g_n(0) = \frac{n!}{1^{n/1}(n/1)!} = 1.$$

If n is not prime and $n>1$, there exist positive integers k and r, both in the interval $[2, n-1]$, such that $n=kr$. Thus, if n is not prime and $n>1$, we have

$$\frac{d^n}{dx^n}g_n(0) \geq 1 + \frac{(kr)!}{k^r r!} > 1.$$

We conclude with *Maple* code below that performs the primality test for small n.

```
eval(diff(sum(exp(t^k/k), k = 1..n - 1), t$n), t = 0);
```

Replace n in the code with the specific positive integer you wish to test. If the output is 1, then n is a prime. If the output is not 1, then n is not a prime.

Excitement from an Error

LINDA MARIE SALIGA
University of Akron
Akron, OH 44325-4002

A very lively discussion ensued after my advanced calculus class read the note "On Sequences" by Julio Cano that appeared in the *American Mathematical Monthly* in 1968 [**1**]. When I made this assignment I simply wanted my students to read some mathematics other than their text book. I had skimmed the note to make sure the students would have the needed background to understand it, but didn't realize that an error was present. I had no idea that discovering this error in a published note would bring out such excitement and passion for mathematics from my students. I will give a synopsis of the discussion to illustrate the thought process of the class. Will you see the error before it is revealed?

We start by stating the proposition, and corresponding argument of its correctness, from Cano's paper.

PROPOSITION [**1**]. *If $\{a_k\}$ is a bounded sequence of real numbers such that $a_k \neq 0$, then there is a subsequence $\{b_n\}$ of $\{a_k\}$ such that $\{b_{n+1}/b_n\}$ converges.*

Argument as presented by Cano [**1**]. We consider two cases:

Case 1: There is a positive ε such that for infinitely many k, $|a_k| \geq \varepsilon$. Let $\{b_n\}$ be the sequence of $\{a_k\}$ consisting of the a_k with $|a_k| \geq \varepsilon$. Then $|b_{n+1}/b_n| \leq (\sup_k |a_k|)/\varepsilon$ for all n.

Case 2: For every $\varepsilon > 0$, ultimately $|a_k| < \varepsilon$. Then there is a subsequence $\{b_n\}$ of $\{a_k\}$ such that $|b_{n+1}| < |b_n|$. Now in both cases the sequence of ratios $\{b_{n+1}/b_n\}$ is bounded, and thus by the Bolzano-Weierstrass theorem there exists a convergent subsequence. ■

The first comment was, "Something doesn't seem right." The students realized that Cano proved that a *subsequence* of $\{b_{n+1}/b_n\}$ converges, not that $\{b_{n+1}/b_n\}$ converges as stated in the proposition. Now I heard things like, "But this was published, so it has to be right," or "why would they publish an incorrect proof." The students decided that they had to be wrong—if it was published it had to be right! I suggested that we follow the steps of the proof with some sequences of real numbers.

When we used $\{a_k\} = \left\{-2, -\frac{3}{2}, \frac{4}{3}, \frac{5}{4}, -\frac{6}{5}, -\frac{7}{6}, \frac{8}{7}, \frac{9}{8}, -\frac{10}{9}, -\frac{11}{10}, \ldots\right\}$ as the original sequence in the proposition, it fell into Case 1 of Cano's argument. Choosing $\varepsilon = 1$ allowed us to set $b_n = a_k$. Then $\{b_{n+1}/b_n\} = \left\{\frac{3}{4}, -\frac{8}{9}, \frac{15}{16}, -\frac{24}{25}, \frac{35}{36}, \ldots\right\}$, which does not converge but has two convergent subsequences; $\{b_{n+1}/b_n\}$ for n odd and $\{b_{n+1}/b_n\}$ for n even. This is exactly what Cano said in the proof, but it is not what the statement of the proposition claims. "Great! We have a counterexample, Cano's proposition is false!" was the sentiment of the class at this point.

Not so fast! Looking at the odd terms of $\{a_k\}$ we get

$$\left\{\frac{b_{n+1}}{b_n}\right\} = \left\{\frac{4/3}{-2}, \frac{-6/5}{4/3}, \frac{8/7}{-6/5}, \frac{-10/9}{8/7}, \ldots\right\} = \left\{\frac{-2}{3}, \frac{-9}{10}, \frac{-20}{21}, \ldots\right\}$$

which does converge as Cano wanted.

Now what? Let's take a break and think about this. . . . After about a month's time, we returned to this discussion. Many different ideas were presented; some dead end

"proofs" and others examples that turned out not to be counterexamples. In the end, we came up with the following proof.

Proof of Proposition. Since $\{a_k\}$ is bounded, by the Bolzano-Weierstrass theorem it has a convergent subsequence, say $\{c_k\}$. We consider two cases.

Case 1: $\{c_k\} \to 0$ as $n \to \infty$. In this case, for each $\varepsilon > 0$ there is a natural number N such that $|c_k| < \varepsilon$ for all $k \geq N$. Let $b_1 = c_1$ and choose b_2 such that $|b_2| \leq \frac{1}{2}|b_1|$. Choose b_3 such that $|b_3| \leq \frac{1}{2^2}|b_2|$. Continue in this manner choosing b_n such that $|b_n| \leq \frac{1}{2^{n-1}}|b_{n-1}|$. Now $\frac{|b_{n+1}|}{|b_n|} \leq \frac{1}{2^n} \to 0$ as $n \to \infty$ and b_{n+1}/b_n converges to 0.

Case 2: $\{c_k\} \to c \neq 0$ as $n \to \infty$. In this case we are going to show that $\left\{\frac{c_{k+1}}{c_k}\right\}$ converges. Let M be the bound of $\{c_k\}$. Since $c \neq 0$, there is a natural number N_1 such that $|c_n| > \frac{|c|}{2}$ for all $n > N_1$. For $\varepsilon > 0$ choose a second natural number N_2 such that $|c_k - c_n| < \frac{c^2\varepsilon}{8M}$ for all $k, n > N_2$. If $N = \max\{N_1, N_2\}$ and $k, n > N$, we show $\left\{\frac{c_{k+1}}{c_k}\right\}$ is Cauchy as follows:

$$\begin{aligned}
\left|\frac{c_{k+1}}{c_k} - \frac{c_{n+1}}{c_n}\right| &= \left|\frac{c_{k+1}c_n - c_kc_{n+1}}{c_kc_n}\right| \\
&= \left|\frac{c_{k+1}c_n - c_{k+1}c_k + c_{k+1}c_k - c_kc_{n+1}}{c_kc_n}\right| \\
&= \left|\frac{(c_n - c_k)\,c_{k+1} + (c_{k+1} - c_{n+1})\,c_k}{c_kc_n}\right| \\
&\leq \frac{|c_n - c_k|\,|c_{k+1}| + |c_{k+1} - c_{n+1}|\,|c_k|}{|c_kc_n|} \\
&< \frac{\frac{c^2\varepsilon}{8M}M + \frac{c^2\varepsilon}{8M}M}{\frac{c^2}{4}} = \varepsilon.
\end{aligned}$$

Therefore, in both cases, there is a subsequence $\{b_n\}$ of $\{a_k\}$ such that $\{b_{n+1}/b_n\}$ converges.

■

Cano's proposition is correct, but there was an error in his "proof." The moral of this story: Even mathematicians make mistakes, but an error can be serendipity!

REFERENCES

1. Julio Cano, On Sequences, *Amer. Math. Monthly* **75** (1968) 645.

Shanille Practices More

HEATHER ANDERTON
Houghton College
Houghton, NY 14744
heather.anderton@houghton.edu

RICHARD JACOBSON
Houghton College
Houghton, NY 14744
jake.jacobson@houghton.edu

The following problem appeared in the 63rd annual Putnam Exam on December 7, 2002.

> Shanille O'Keal shoots free throws on a basketball court. She hits the first and misses the second, and thereafter the probability that she hits the next shot is equal to the proportion of the shots she has hit so far. What is the probability she hits exactly 50 of her first 100 shots? [**1**]

A natural generalization of this problem would be to increase the number of shots in determining her skill level. Suppose her skill level is determined by the fact that she made a and missed b shots. Then what is the probability, based on the criteria above, that she will make k out of the next n shots?

Richey and Zorn recently developed a solution to this problem involving an approach in a statistical setting; see this MAGAZINE [**2**]. Their argument was an interesting application of Bayesian analysis. It is the purpose of this note to solve this problem using a direct tree diagram approach.

At this point we introduce some combinatorial notation that is convenient, but not used universally. For natural numbers a and k we define the falling factorial to be $a_{[k]} = a(a-1)(a-2)\cdots(a-k+1)$. This is another name for the number of permutations $P(a,k)$. Similarly, we define the rising factorial $a^{[k]} = a(a+1)(a+2)\cdots(a+k-1)$. This notation will be useful later in the paper.

Let $[a:b]$ denote Shanille's skill level at the beginning of her shooting the n additional shots. The probability of making k of these shots is the sum of the probabilities of the distinct ways in which she can make those k additional shots. The number of these different ways can be counted by specifying which of the n shots were made. It follows that there are $C(n,k)$ distinct paths.

The probability associated with any particular path will be the product of the n probabilities of making or missing a basket on any specific attempt. The probability of making (missing) a basket is the number of previous shots made (missed) divided by the number of previous attempts. The denominator of this product will be $(a+b)(a+b+1)\cdots(a+b+n-1) = (a+b)^{[n]}$. The factors in the numerator will be $a, a+1, a+2, \ldots, a+k-1$ for the events when the shot is made and $b, b+1, b+2, \ldots, b+n-k-1$ for the $n-k$ events when the shot is missed. The product of these factors can be rearranged so the numerator is $a^{[k]}b^{[n-k]}$ for each of the $C(n,k)$ paths.

This leads to the theorem:

Let $[a:b]$ denote a skill level of making a shots and missing b shots in $a+b$ attempts. Let $P([a:b];k,n)$ be the probability of making exactly k of the next n shots.

Then

$$P([a:b];k,n)=\frac{C(n,k)a^{[k]}b^{[n-k]}}{(a+b)^{[n]}}.$$

Acknowledgement. We thank the referees for their very helpful comments.

REFERENCES

1. 63rd Annual William Putnam Mathematical Competition, this MAGAZINE **76** (2003) 76–80.
2. Matthew Richey and Paul Zorn, this MAGAZINE **78** (2005) 354–367.

Dearest Blaise

1

1 1

1 he 1

1 saw the 1

1 pure simple idea 1

1 truly distilling arithmetic later 1

1 poetic mathematization uncharacteristically sentimentalized Pascal 1

Caleb J. Emmons
Pacific University
Emmons@pacificu.edu

PROBLEMS

ELGIN H. JOHNSTON, *Editor*
Iowa State University

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

Proposals

To be considered for publication, solutions should be received by March 1, 2008.

1776. *Proposed by Leon Gerber, St. John's University, Jamaica, NY.*

Let n be an odd integer, let $f_n(x) = (1+x)^n - (1+nx)$, and let x_n be the unique negative solution to $f_n(x) = 0$. It is easy to show that f_n has a positive relative maximum at $x = -2$. Prove that the sequence $\{f_n(-2)\}$ is increasing, and that $\lim_{n\to\infty} x_n = -2$.

1777. *Proposed by Richard A. Jacobson, Houghton College, Houghton, NY.*

The graph of the relation $|x+y| + |x-y| = 2$ is a square of side length 2. Find positive integer n and real constants a_k, b_k, c_k, $1 \le k \le n$ such that the graph of the relation $\sum_{k=1}^{n} |a_k x + b_k y + c_k| = 2$ is

a. a regular hexagon of side length 2.

b. a regular dodecagon of side length 2.

1778. *Proposed by Jody M. Lockhart and William P. Wardlaw, U.S. Naval Academy, Annapolis, MD.*

Let q be a positive integer power of a prime and let F_q denote the finite field of q elements. For each positive integer n and each $\gamma \in F_q$, find the number of $n \times n$ matrices over F_q with determinant γ.

1779. *Proposed by Will Gosnell, Amherst, MA, and Herb Bailey, Rose Hulman Institute of Technology, Terre Haute, IN.*

Let ABC be a triangle with $BC = a$, $CA = b$, and $AB = c$, let $\theta = \angle ACB$, and let $k = c/b$. Alice walks from C to A at a constant speed, and Bob walks from B to A, also at constant speed. The sum of the two travel speeds is numerically equal to a, and

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

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

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

for each the travel time is numerically equal to $k^{3/2}$. For which values of k can a value of θ be found so that the given conditions can be satisfied? (This problem generalizes the problem with $\theta = 90°$ posed by Will Gosnell in the February 2005 issue of *Math Horizons*.)

1780. *Proposed by Yiu Tung Poon, Iowa State University, Ames, IA.*

For positive integer k, define odd(k) to be the number of odd digits in the (base-ten) expansion of 2^k. Prove that

$$\sum_{k=1}^{\infty} \frac{\text{odd}(k)}{2^k} = \frac{1}{9}.$$

Quickies

Answers to the Quickies are on page 314.

Q973. *Proposed by Michael Andreoli, Miami-Dade College, North Campus, Miami, FL.*

Prove that for positive real numbers a, b, c,

$$\frac{1}{2a} + \frac{1}{2b} + \frac{1}{2c} \geq \frac{1}{a+b} + \frac{1}{b+c} + \frac{1}{c+a}.$$

Q974. *Proposed by Ricardo García-Pelayo, Universidad Politécnica de Madrid, Madrid, Spain.*

Prove that

$$\sum_{j=2}^{\infty} (-1)^j \frac{\zeta(j)}{j+1} = 1 + \frac{\gamma - \ln(2\pi)}{2},$$

where γ is the Euler-Mascheroni constant and ζ is the Riemann zeta function, defined for $j = 2, 3, \ldots$ by $\zeta(j) = \sum_{k=1}^{\infty} \frac{1}{k^j}$.

Solutions

A combinatorial identity. **October 2006**

1751. *Proposed by Iliya Bluskov, University of Northern British Columbia, Prince George, BC, Canada*

Let $k_1, k_2, \ldots, k_n$ be integers with $k_i \geq 2$, $i = 1, 2, \ldots, n$, and let $N = \sum_{i=1}^{n} \binom{k_i}{2}$. Prove that

$$\sum_{1 \leq i < j \leq n} \binom{k_i}{2}\binom{k_j}{2} + 3\sum_{i=1}^{n} \binom{k_i + 1}{4} = \binom{N}{2}. \tag{1}$$

I. *Solution by Robert Doucette, McNeese State University, Lake Charles, LA.*

Let $x_i = \binom{k_i}{2}$. Then,

$$\binom{\sum_{i=1}^n x_i}{2} = \frac{1}{2}\sum_{i=1}^n x_i\left(\sum_{i=1}^n x_i - 1\right)$$

$$= \sum_{1\le i<j\le n} x_i x_j + \frac{1}{2}\sum_{i=1}^n x_i(x_i-1) = \sum_{1\le i<j\le n} x_i x_j + \sum_{i=1}^n \binom{x_i}{2}.$$

The desired result follows from

$$\binom{\binom{k_i}{2}}{2} = \frac{1}{2}\frac{k_i(k_i-1)}{2}\left(\frac{k_i(k_i-1)}{2} - 1\right)$$

$$= \frac{1}{8}(k_i+1)k_i(k_i-1)(k_i-2) = 3\binom{k_i+1}{4}.$$

II. *Solution by Arthur T. Benjamin and Andrew Carman (student) Harvey Mudd College, Claremont, CA.*

We give a combinatorial proof.

A city has n districts and, for $i = 1, \ldots, n$, district i has k_i representatives on the city council. The number of ways to select a *ticket*, an unordered pair of representatives from the same district is $N = \sum_{i=1}^n \binom{k_i}{2}$.

On the right side of (1), $\binom{N}{2}$ counts the ways to pick two different tickets. We claim the left side also gives this count.

The number of ticket pairs that come from two different districts is $\sum_{1\le i<j\le n}\binom{k_i}{2}\binom{k_j}{2}$. The number of ticket pairs using four different people from district i is $3\binom{k_i}{4}$, because once we have chosen ticket members a, b, c, d, there are three ways to select the running mate of a. The number of ticket pairs involving three different people from district i is $3\binom{k_i}{3}$, since once we have chosen ticket members a, b, c, there are three ways to pick which representative appears on both tickets. Since $\binom{k_i}{4} + \binom{k_i}{3} = \binom{k_i+1}{4}$, the total number of ways to pick a pair of tickets from the same district is $3\sum_{i=1}^n \binom{k_i+1}{4}$. The result now follows.

Also solved by JPV Abad, Steve Abbott, Armstrong Problem Solvers, Ovidiu Bagdasar (Romania), Michel Bataille (France), J. C. Binz (Switzerland), Robert Calcaterra, Johann Chen, Haiwen Chu, Con Amore Problem Group (Denmark), Chip Curtis, Knut Dale (Norway), Joe DeMaio, M. N. Deshpande (India), Fejéntáltuka Szeged Problem Solving Group (Hungary), G.R.A.20 Problem Solving Group (Italy), Ralph P. Grimaldi, Enkel Hysnelaj (Australia) and Elton Bojaxhiu (Albania), Ronald A. Kopas, Harris Kwong, Elias Lampakis (Greece), Kathleen E. Lewis, McDaniel College Problem Group, Kim McInturff, William Moser (Canada), Michael Natanson, Rob Pratt, Nicholas C. Singer, Albert Stadler, Marian Tetiva (Romania), Thomas R. Wilkerson, Stuart V. Witt, Paul Zwier, and the proposer.

Lots of limit points. **October 2006**

1752. *Proposed by John Sternitzky and Robert Calcaterra, University of Wisconsin Platteville, Platteville, WI.*

Let $\mathbb{R}$ be the real line with the standard topology. Prove that every uncountable subset of $\mathbb{R}$ has uncountably many limit points.

Many readers submitted a solution similar to the following.

Let $S \subset \mathbb{R}$ be a set with at most countably many limit points, let L be the set of points in S that are limit points of S, and let $I = S\backslash L$ be the set of isolated points

in S. We prove that I is at most countable. It will then follow that $S = L \cup I$ is at most countable. Let $x \in I$. Because x is an isolated point in S we can find rational numbers r_x, s_x so that $(r_x, s_x) \cap S = \{x\}$. Thus, each $x \in I$ corresponds to a unique open interval with rational endpoints. It follows that I is finite or countably infinite. Hence, if $S \subseteq \mathbb{R}$ is uncountable then it must have an uncountable number of limit points.

Note. Readers noted that this result appears (for $\mathbb{R}^n$) as an exercise in *Principles of Mathematical Analysis, Third Edition,* by Walter Rudin, and as an exercise in the text *Mathematical Analysis* by G. E. Silov.

Solved by Alejandro Aguado, Michael R. Bacon, Michel Bataille (France), Michael W. Botsko, Paul Budney, Bruce S. Burdick, Haiwen Chu, Con Amore Problem Group (Denmark), A. K. Desai (India), Fejéntaláltuka Szeged Problem Solving Group (Hungary), Eugene A. Herman, Enkel Hysnelaj (Australia) and Elton Bojaxhiu (Albania), Danrun Huang, Elias Lampakis (Greece), Thomas C. Lominac, Rick Mabry, Eric Mbakop, McDaniel College Problems Group, Eric Mbakop, Valerian M. Nita, Northwestern University Math Problem Solving Group, Michael Ontiveros, Paolo Perfetti (Italy), Albert Stadler (Switzerland), Richard Stephens, Marian Tetiva (Romania), Dave Trautman, Xiaoshen Wang, and the proposers. There was one incorrect submission.

A subseries of the harmonic series. **October 2007**

1753. *Proposed by John C. George, Eastern New Mexico University, Portales, NM.*

Let n be a positive integer and let S_n be the set of all positive integers whose (base ten) digit sum is n. Determine the convergence or divergence of the series

$$\sum_{k \in S_n} \frac{1}{k}.$$

Solution by Marian Tetiva, Bîrlad, Romania.

It is a very well known result that the equation

$$x_1 + \cdots + x_j = n$$

has $\binom{n+j-1}{n}$ solutions in non-negative integers, two solutions being considered different even if they differ only by order. Therefore the number of numbers with j digits that sum up to n (denote the set of these numbers by $S_{n,j}$) is at most $\binom{n+j-1}{n}$, since the digits $x_1, x_2, \ldots, x_j$ of such a number must satisfy the above equation, *and* the conditions $x_1 \in \{1, \ldots, 9\}$, $x_2, \ldots, x_j \in \{0, 1, \ldots, 9\}$. In addition, any number with j digits, is greater than or equal to 10^{j-1}. It follows that

$$\sum_{k \in S_n} \frac{1}{k} = \sum_{j \geq 1} \sum_{k \in S_{n,j}} \frac{1}{k} \leq \sum_{j \geq 1} \frac{1}{10^{j-1}} \binom{n+j-1}{n} \leq \frac{2^n}{n!} \sum_{j \geq 1} \left(\frac{1}{5}\right)^{j-1},$$

where the last inequality follows from $\binom{n+j-1}{n} < 2^{n+j-1}$. Therefore the given series is convergent and its sum is at most

$$\frac{2^n}{n!} \sum_{j \geq 1} \left(\frac{1}{5}\right)^{j-1} = \frac{5}{4} \cdot \frac{2^n}{n!}.$$

Also solved by Michael Andreoli, Armstrong Problem Solvers, Michael R. Bacon, H. Bailey and R. Schoemacher, Michel Bataille (France), Robert Calcaterra, Haiwen Chu, John Cobb, Con Amore Problem Group (Denmark), Fejéntaláltuka Szeged Problem Solving Group (Hungary), Dmitry Fleischman, Marty Getz and Dixon Jones, Michael Goldenberg and Mark Kaplan, G.R.A.20 Problem Solving Group (Italy), Eugene A. Herman, Enkel Hysnelaj (Australia) and Elton Bojaxhiu (Albania), Douglas E. Iannucci, Harris Kwong, Elias Lampakis (Greece), David P. Lang, Kee-Wai Lau (China), Kathleen E. Lewis, Marvin Littman, David Lovit, Rick

Mabry, Eric Mbakop, McDaniel College Problems Group, Michael Nathanson, Northwestern University Math Problem Solving Group, Nadeeka De Silva, Nicholas C. Singer, Albert Stadler (Switzerland), Paul J. Zwier, and the proposer. There were two incorrect submissions and one solution with no name.

Between the geometric and arithmetic means. **October 2006**

1754. *Proposed by Mihály Bencze, Săcele-Négyfalu, Romania.*

Let $a_1, a_2, \ldots, a_n$ be positive real numbers. Prove that

$$\sqrt[n]{\prod_{k=1}^{n} a_k} \le \ln\left(1 + \sqrt[n]{\prod_{k=1}^{n}(e^{a_k} - 1)}\right) \le \frac{1}{n}\sum_{k=1}^{n} a_k.$$

Solution by Michel Bataille, Rouen, France.

If $x_1, x_2, \ldots, x_n$ are positive real numbers, then

$$1 + \sqrt[n]{x_1 x_2 \cdots x_n} \le \sqrt[n]{(1 + x_1)(1 + x_2)\cdots(1 + x_n)}. \tag{1}$$

This follows readily from Jensen's inequality applied to the convex function $f(t) = \ln(1 + e^t)$ with $t_k = \ln x_k$, $1 \le k \le n$. Setting $x_k = e^{a_k} - 1$, $1 \le k \le n$, (1) gives

$$1 + \sqrt[n]{\prod_{k=1}^{n}(e^{a_k} - 1)} \le e^{(1/n)(a_1 + a_2 + \cdots + a_n)},$$

and the right inequality follows by taking logarithms.

Next consider the function g defined by $g(t) = \ln\left(\exp(e^t) - 1\right)$. Then

$$g''(t) = \frac{e^t e^{e^t}\left(e^{e^t} - e^t - 1\right)}{(e^{e^t} - 1)^2},$$

and it is clear that this expression is positive for all real t. Hence g is convex on $\mathbb{R}$. By Jensen's inequality we then have

$$g\left(\frac{1}{n}\sum_{k=1}^{n} \ln a_k\right) \le \frac{1}{n}\sum_{k=1}^{n} g(\ln a_k)$$

or, with $G = \sqrt[n]{a_1 a_2 \cdots a_n}$,

$$\ln(e^G - 1) \le \ln \sqrt[n]{(e^{a_1} - 1)(e^{a_2} - 1)\cdots(e^{a_n} - 1)}.$$

Thus, $e^G - 1 \le \sqrt[n]{\prod_{k=1}^{n}(e^{a_k} - 1)}$ and the left inequality

$$G \le \ln\left(1 + \sqrt[n]{\prod_{k=1}^{n}(e^{a_k} - 1)}\right)$$

follows.

Also solved by Michael Andreoli, Ovidiu Bagdasar (Romania), Erhard Braune (Germany), Robert Calcaterra, Haiwen Chu, Gordon Crandall, Chip Curtis, Knut Dale (Norway), José Luis Díaz-Barrero (Spain), Charles R. Diminnie, Prithwijit De (Ireland), Robert L. Doucette, Fejéntaláltuka Szeged Problem Solving Group (Hungary), Ovidui Furdui, G.R.A.20 Problem Solving Group (Italy), Enkel Hysnelaj (Australia) and Elton Bojaxhiu (Albania), Elias Lampakis (Greece), David P. Lang, Eric Mbakop, Northwestern University Math Problem Solving Group, Paolo Perfetti (Italy), Teodora-Liliana Rădulescu and Vicentiu Rădulescu (Romania), Albert Stadler (Switzerland), Tony Tam, Thomas R. Wilkerson, and the proposer.

An asymptotic formula **October 2006**

1755. *Proposed by Michel Bataille, Rouen, France.*

Let a, b, $c > 0$ with $b > c$. Prove that, as $n \to \infty$,

$$\frac{(a+b)^{a+b}(2a+b)^{2a+b}\cdots(na+b)^{na+b}}{(a+c)^{a+c}(2a+c)^{2a+c}\cdots(na+c)^{na+c}} \sim \lambda(na)^{n\alpha+\beta},$$

for some positive real numbers λ, α, and β.

Solution by Eugene A. Herman, Grinnell College, Grinnell, IA.

Let E_n denote the given expression on the left. It suffices to show that

$$\log E_n = (n\alpha+\beta)(\log n + \log a) + \text{constant} + o(1) \tag{1}$$

for some positive real numbers α and β. To derive (1), first note that

$$\begin{aligned}
\log E_n &= \sum_{j=1}^{n}[(ja+b)\log(ja+b) - (ja+c)\log(ja+c)] \\
&= \sum_{j=1}^{n}\left[(ja+b)\left(\log a + \log\left(j+\tfrac{b}{a}\right)\right) - (ja+c)\left(\log a + \log\left(j+\tfrac{c}{a}\right)\right)\right] \\
&= n(b-c)\log a + a\sum_{j=1}^{n} j\log\left(\frac{j+\frac{b}{a}}{j+\frac{c}{a}}\right) + \sum_{j=1}^{n}\left[b\log\left(j+\tfrac{b}{a}\right) - c\log\left(j+\tfrac{c}{a}\right)\right].
\end{aligned} \tag{2}$$

Next, we expand the second and third terms in the sum (2) with the help of the following Taylor remainder formulas for $\log(1+x)$ with $x > 0$:

$$\begin{aligned}
&\log(1+x) = x + R_2(x), && \text{where } |R_2(x)| \le \tfrac{1}{2}x^2 \\
&\log(1+x) = x - \tfrac{1}{2}x^2 + R_3(x), && \text{where } |R_3(x)| \le \tfrac{1}{3}x^3,
\end{aligned}$$

and Stirling's formula:

$$\log n! = \left(n+\tfrac{1}{2}\right)\log n - n + \text{constant} + o(1).$$

Thus, with $k = \dfrac{b-c}{a}$, we have

$$\begin{aligned}
\sum_{j=1}^{n} j\log\left(\frac{j+\frac{b}{a}}{j+\frac{c}{a}}\right) &= \sum_{j=1}^{n} j\log\left(1+\frac{k}{j+\frac{c}{a}}\right) \\
&= \sum_{j=1}^{n} j\left(\frac{k}{j+\frac{c}{a}} - \frac{k^2}{2\left(j+\frac{c}{a}\right)^2} + R_3\left(\frac{k}{j+\frac{c}{a}}\right)\right) \\
&= \sum_{j=1}^{n}\left(k - \frac{k\frac{c}{a}}{j+\frac{c}{a}} - \frac{k^2}{2(j+\frac{c}{a})} + \frac{k^2\frac{c}{a}}{2(j+\frac{c}{a})^2}\right) + \text{constant} + o(1)
\end{aligned}$$

$$= nk - \sum_{j=1}^{n} \frac{1}{j}\left(k\frac{c}{a} + \frac{k^2}{2}\right) + \frac{c}{a}\sum_{j=1}^{n} \frac{1}{j(j+\frac{c}{a})}\left(k\frac{c}{a} + \frac{k^2}{2}\right)$$

$$+ \text{constant} + o(1)$$

$$= nk - \left(k\frac{c}{a} + \frac{k^2}{2}\right)\log n + \text{constant} + o(1). \tag{3}$$

In addition, for any $q > 0$, we have

$$\sum_{j=1}^{n} \log(j+q) = \sum_{j=1}^{n} \left(\log j + \log\left(1 + \tfrac{q}{j}\right)\right)$$

$$= \log n! + \sum_{j=1}^{n}\left(\frac{q}{j} + R_2\left(\frac{q}{j}\right)\right)$$

$$= \left(n + \tfrac{1}{2}\right)\log n - n + q\log n + \text{constant} + o(1). \tag{4}$$

Substituting (3) and (4) into (2), we conclude

$$\log E_n = n(b-c)\log a + ank - \left(kc + \frac{ak^2}{2}\right)\log n + (b-c)\left(\left(n+\tfrac{1}{2}\right)\log n - n\right)$$

$$+ \frac{b^2-c^2}{a}\log n + \text{constant} + o(1)$$

$$= n(b-c)\log a + n(b-c) - \left(\frac{b-c}{a}c + \frac{(b-c)^2}{2a}\right)\log n + n(b-c)\log n$$

$$+ \frac{1}{2}(b-c)\log n - n(b-c) + \frac{b^2-c^2}{a}\log n + \text{constant} + o(1)$$

$$= n(b-c)\log a + n(b-c)\log n + \frac{b-c}{2a}(a+b+c)\log n$$

$$+ \text{constant} + o(1).$$

This establishes (1) with $\alpha = b - c$ and $\beta = \dfrac{b-c}{2a}(a+b+c)$.

Also solved by Knut Dale (Norway), Enkel Hysnelaj (Australia) and Elton Bojaxhiu (Albania), McDaniel College Problem Group, Albert Stadler (Switzerland), Nicholas C. Singer, and the proposer. There were two incorrect submissions.

Answers

Solutions to the Quickies from page 309.

A973. We have

$$\frac{1}{a} + \frac{1}{b} + \frac{1}{c} = \frac{\frac{1}{a}+\frac{1}{b}}{2} + \frac{\frac{1}{b}+\frac{1}{c}}{2} + \frac{\frac{1}{c}+\frac{1}{a}}{2} \geq \frac{2}{a+b} + \frac{2}{b+c} + \frac{2}{c+a},$$

where the inequality follows from the arithmetic-harmonic mean inequality.

A974. Any positive integer can be written as follows:

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

Therefore:

$$\ln\frac{n^n}{n!} = \sum_{j=1}^{n-1}\ln\left(1+\frac{1}{j}\right)^j = \sum_{j=1}^{n-1}\left(1-\frac{1}{2j}+\frac{1}{3j^2}-\ldots\right)$$
$$= (n-1)-\frac{1}{2}\sum_{j=1}^{n-1}\frac{1}{j}+\sum_{k=2}^{\infty}\frac{(-1)^k\sum_{j=1}^{n-1}\frac{1}{j^k}}{k+1} = n-\frac{\ln n}{2}-\ln C(n).$$

The last term is obtained by applying Stirling's approximation to $\ln(n^n/n!)$, so that $C(n)$ satisfies $\lim_{n\to\infty} C(n) = \sqrt{2\pi}$. We rewrite the last equality:

$$n-1-\frac{1}{2}\left(\sum_{j=1}^{n-1}\frac{1}{j}-\ln n\right)+\sum_{k=2}^{\infty}\frac{(-1)^k\sum_{j=1}^{n-1}\frac{1}{j^k}}{k+1} = n-\ln C(n).$$

The desired result follows by letting $n\to\infty$.

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.

Pilkey, Orrin H., and Linda Pilkey-Jarvis, Mathematical models just don't add up, *Chronicle of Higher Education* (25 May 2007) B12. *Useless Arithmetic: Why Environmental Scientists Can't Predict the Future*, Columbia University Press, 2007; xv + 230 pp. $29.95. ISBN 978–0–231–13212–1.

"We argue that quantitative modeling of complex systems is impossible." (Well, there goes applied mathematics, except for the simplest of phenomena.) The authors are sympathetic to "qualitative models," which provide direction and order-of-magnitude answers to the questions "how, why, and what if." But they are dead set against "quantitative models," dedicated to predicting "where, when, and how much." They offer detailed analyses of several venues of modeling. In terms of recognition of uncertainties, debate about model validity, and usefulness, from worst to best in their opinion are beach lifespans, shoreline erosion rates, water quality of abandoned pit mines, nuclear repositories, fish populations, sea levels, and invasive species. The authors rightly criticize some models for lack of comprehension of all relevant processes, but they tend to blame the modelers and the models for the purely economic and political purposes that demand, deploy, and defend the models. Details of three models are given in an appendix. (Thanks to Darrah Chavey of Beloit College.)

Alexanderson, Gerald L., with Peter Ross (eds.), *The Harmony of the World: 75 Years of Mathematics Magazine*, MAA, 2007; xii + 287 pp, $55.95 (MAA members: $44.95). ISBN 978–0–88385–5607.

How could I fail to review a book of the "best" of THIS MAGAZINE, celebrating its 75th anniversary? Did you know that the Stone-Weierstrass theorem was first published here in 1947/48? This collection gives scant coverage to material since 1990 "since that is probably fairly well-known to potential readers and is generally more accessible." (How far back does your library's run of THIS MAGAZINE go? My institution now routinely discards paper copies of journals once they are available on JSTOR, which contains the full archive of THIS MAGAZINE through 2003, with a moving "wall" of three years until archiving.) The editors introduce each reproduced article with valuable notes about it and the author. The title of the volume comes from an article by Morris Kline that "makes a case for mathematics as a cultural component that should be in everyone's background," a cause echoed 30 years later by Judith Grabiner in her "The centrality of mathematics in the history of Western thought." Although a very brief homage to the Problem Section is included and one book review from 1936 is reproduced, I am naturally chagrined that none of my reviews in this column over the past 30 years got selected as part of the "best" of THIS MAGAZINE. However, I may have done my cause irreparable harm with one of my early reviews, of a 1978 article about "uncrackable" codes; the review read in its entirety (I was magnificently briefer then): "Popular account of new developments in cryptology based on factorizations of large primes." [I thank Ramesh Kapadia and Ian Stewart for drawing this blunder to the world's attention in the last issue of their utterly delightful quasiperiodical *Manifold* #20 (Spring 1980) 5, noting kindly that "*Mathematics Magazine* seems to know something that Euclid didn't." But my review and their remark didn't make it

into their "best," either—though the "ordered pears" figures on the same page did (*Seven Years of* Manifold *1968–1980*, edited by Ian Stewart and John Jaworski, Shiva Publishing, 1981).]

Chartier, Timothy P., Googling Markov, *The UMAP Journal* 27 (2006) (1) 17–30. Sobek, Markus, A survey of Google's PageRank, `http://pr.efactory.de/` . Wills, Rebecca, The math behind the search engine, *Mathematical Intelligencer* 28 (4) (Fall 2006) 6–11. Bryan, Kurt, and Tanya Leise, The $25,000,000,000 eigenvector: The linear algebra behind Google, *SIAM Review* 48 (3) 569–581. Langville, Amy N., and Carl D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, 2006; x + 224 pp, $35. ISBN 978–0–691–12202–1.

The four articles, listed in approximately increasing order of mathematical demand on the reader, explain the linear algebra behind Google's PageRank algorithm; Bryan and Leise even include exercises. The book by Langville and Meyer is a thorough analysis of both PageRank and some alternative ranking algorithms, incorporating Matlab code, citing downloadable datasets and crawler programs, and including a 50-page final chapter setting out the mathematical basis in linear algebra, Perron-Frobenius theory, Markov chains, and more. Sobek considers briefly how to incorporate additional factors in ranking, such as the ones cited in Page's patent (strangely, Langville and Meyer do not include the patent—#6,285,999—among their references). The ubiquitous use of Google and other search engines should help motivate students in linear algebra—provided their instructor does such applications in the course. (The $25 B in the title of one article was the approximate market value of Google when it went public.)

Devlin, Keith, and Gary Lorden, *The Numbers Behind* Numb3rs*: Solving Crimes with Mathematics*, Plume, 2007; x + 236 pp, $13. ISBN 978–0–452–28857–7.

For several semesters, I have taught a one-semester-hour seminar on the mathematics behind current episodes of the TV series *Numb3rs*, so I was utterly delighted to see this book appear. It helps answer the common questions: Is the math in *Numb3rs* real? (Yes.) Did—or could—math actually help solve a crime in the way depicted? (Sometimes, and yes, but "not in one or two 'television days.'") Nine chapters take off from specific episodes to treat geographic profiling, data mining, changepoint detection, Bayesian inference, cryptology, fingerprints, networks, game theory, and casino Blackjack; four others consider topics common in several episodes (statistics, image enhancement) or that are generally relevant (DNA profiling, mathematics in court cases). I would have preferred more suggestions pointing the reader to further reading. An appendix summarizes for each episode of the first three seasons the "primary mathematical contribution to solving the case." By the time you read this, the fourth season will have started!

Gigerenzer, Gerd, *Calculated Risks: How to Know When Numbers Deceive You*, Simon & Schuster, 2002; ix + 310 pp, $20.95. ISBN 0–7432–5423–6.

"One in every 10 (or 9, or 8) women develops breast cancer." That's the slogan used to encourage (scare) women to have regular mammograms. Gigerenzer deconstructs that statistic in the clearest fashion that I have seen, giving a table of frequencies (out of 1,000), by age, of numbers of incidents and deaths from breast cancer, cardiovascular causes, and other causes. This example illustrates his thesis that understanding uncertainty and overcoming innumeracy about risks requires communicating risk intelligibly, largely through the use of what he calls "natural frequencies" rather than conditional probabilities. Other chapters address communication of informed consent for medical procedures, AIDS counseling, wife battering, the prosecutor's fallacy, DNA fingerprinting, exploiting innumeracy, and how to teach clear thinking.

Stewart, Ian, How to Cut a Cake and Other Mathematical Conundrums, Oxford University Press, 2006; xiii + 231 pp, $14.95. ISBN 0–19–920590–6.

"This is a book for the fans, for the enthusiasts, for the people who actively *like* mathematics"—in other words, it is perfect for readers of THIS MAGAZINE. (Author Stewart wanted to title it *Weapons of Math Distraction*.) Like some of his previous books, it is a collection of new edits of 20 of his 96 columns from *Scientific American* and its foreign-language affiliates between 1987 and 2001 ("all known mistakes have been corrected, an unknown number of *new* mistakes have been introduced"). The topics of the columns were "selected primarily for amusement value, not for significance." Like his predecessor Martin Gardner, Stewart manages over and over to highlight the value and the fun of mathematics, with nary an equation in sight.

NEWS AND LETTERS

Carl B. Allendoerfer Awards—2007

The Carl B. Allendoerfer Awards, established in 1976, are made to authors of expository articles published in MATHEMATICS MAGAZINE. The Awards are named for Carl B. Allendoerfer, a distinguished mathematician at the University of Washington and President of the Mathematical Association of America, 1959–60.

Carl V. Lutzer, Hammer Juggling, Rotational Instability, and Eigenvalues, MATHEMATICS MAGAZINE, 79 (2006) 243–250.

Citation The author analyzes the physical phenomenon familiar to tennis players and hammer jugglers that certain objects flip when rotated about one of their three rotational axes—the 'middle' axis. The article is organized so that readers can understand the main ideas in a quick reading, but it also invites the reader to delve into the details. The mathematics is not trivial, but the exposition is lively and engaging.

Lutzer uses Euler's equation, which gives the torque on a spinning object in $\Re^3$ in terms of the angular velocity. The behavior of the rotating system is governed by a 3×3 real symmetric matrix whose eigenvalues are positive and distinct. The author proves the positivity result using normed linear operators, which gives the reader a taste of the kind of mathematical methods typically used in the field.

The stability of any rotation is easy to see when the alignment with the axis of rotation is perfect, but perfection is impossible to achieve in practice. The author then shows that when the system is perturbed slightly, the rotation about the axis corresponding to the dominant eigenvalue remains stable, so there is no flipping along the longest axis. However for the middle eigenvalue, small perturbations rapidly propagate to produce large components along the other axes, so the object flips when rotated along this axis.

The author interweaves mathematical rigor with physical intuition throughout the paper. It could form the basis of a research project for students interested in physical applications of mathematics. In particular, students with some background in linear algebra and multivariable calculus should be able to follow all of the details.

For any reader, the article rewards careful study with a deeper understanding of a widely observed but not widely understood phenomenon.

Biographical Note **Carl Lutzer** is an Associate Professor of Mathematics at the Rochester Institute of Technology, and was selected for inclusion in the "Who's Who Among America's Teachers" in both 2003 and 2004. He was a finalist for RIT's Richard and Virginia Eisenhart Provost's Award for Excellence in Teaching in both 2002 and 2003, and was a 2000–2001 ExxonMobil Project NExT Fellow. He earned his PhD from the University of Kentucky under the direction of Dr. Peter Hislop. His mathematical research interests tend to lie in the analysis of partial differential equations and their application to physics and biology. In addition to mathematics and teaching, he enjoys writing fiction, fencing (the sport, not the barrier), and being a father.

Response from Carl Lutzer I've always enjoyed the articles in MATHEMATICS MAGAZINE, so I was excited when my article was accepted for publication. Win-

ning the Carl B. Allendoerfer Award has redoubled that excitement, and I now find myself speechless—I've struggled for a week to write this simple paragraph! So I'll just say that I am tremendously honored to win this award. I'd like to thank the editor and referees for their helpful suggestions, as well as my friends and colleagues at RIT for being so supportive. Lastly, I'd like to thank Ray Hodges for introducing me to the phenomenon, saying, "Hey, watch this!"

Saul Stahl, The Evolution of the Normal Distribution, MATHEMATICS MAGAZINE, 79 (2006) 96–113.

Citation All students who take a statistics course, and anyone who analyzes data, encounter the normal curve early in their study. This article traces the historical roots of the normal distribution and its early development by mathematicians and statisticians. This fascinating account includes false starts by some very famous mathematicians, disagreements about whether the mean or the median should be used to summarize data (or whether any single number should be used at all), and shows where the word 'normal' came from.

Throughout the 18th and 19th centuries, scientists needed tools to analyze the data they were collecting. The author begins with an early example of the binomial distribution: the Dutch mathematician Willem 's Gravesande analyzed data to decide whether the difference between the number of male and female births in London from 1629 to 1710 was explainable by chance. Such calculations were difficult, and in 1733, De Moivre found a way to approximate the binomial distribution using the normal curve, now familiar to all statistics students.

Much of the motivation for the development of the subject came from attempts to understand the errors associated with observations in astronomy. The author does an admirable job of tracing this history. Gauss plays a central role here, with his calculation of the orbit of Ceres leading to his proof that the normal distribution describes the distribution of observational errors. The author presents Gauss' proof here; a reader with a background in calculus should be able to understand the entire argument.

Stahl does a wonderful job of blending some interesting historical research with mathematical details to appeal to a very wide audience. Along the way, we encounter many familiar mathematical names: Jacob and Daniel Bernouli, Cotes, Simpson, Laplace, and Gauss, of course. The extensive bibliography will be valuable for readers wishing to learn more of the historical details.

Biographical Note **Saul Stahl** was born in 1942 in Antwerp, Belgium. He received his BA from Brooklyn College in 1963, his MA from the University of California at Berkeley in 1966, and his PhD from Western Michigan University in 1975. He served in the Peace Corps in Nepal, worked as a systems programmer for IBM in Endicott, NY and also as a postdoctorate fellow at Wright State University in Fairborn, Ohio. He joined the faculty of the University of Kansas in 1977 where he is now a Professor of Mathematics. Most of his research was done in the area of graph theory. He has written six textbooks at the junior-senior level whose exposition is very much informed by the evolution of their respective subject matters. Saul's current hobby is the Tango Argentino. His concern with the center of gravity of the dancers fits in nicely with his current research on the center of mass in hyperbolic geometry.

Response from Saul Stahl I am honored to receive the Carl B. Allendoerfer Award from the MAA. The gathering of the information for the article that earned me this award was greatly facilitated by the excellent histories written by Anders Hald and Stephen M. Stigler.

47th International Mathematical Olympiad Ljubljana, Slovenia July 12 and 13, 2006

edited by Zuming Feng, Cecil Rousseau and Yufei Zhao

PROBLEMS

1. Let ABC be a triangle with incenter I. A point P in the interior of the triangle satisfies

$$\angle PBA + \angle PCA = \angle PBC + \angle PCB.$$

Show that $AP \geq AI$, and that equality holds if and only if $P = I$.

2. Let $\mathcal{P}$ be a regular 2006-gon. A diagonal of $\mathcal{P}$ is called *good* if its endpoints divide the boundary of $\mathcal{P}$ into two parts, each composed of an odd number of sides of $\mathcal{P}$. The sides of $\mathcal{P}$ are also called *good*.

Suppose $\mathcal{P}$ has been dissected into triangles by 2003 diagonals, no two of which have a common point in the interior of $\mathcal{P}$. Find the maximum number of isosceles triangles having two *good* sides that could appear in such a configuration.

3. Determine the least real number M such that the inequality

$$|ab(a^2 - b^2) + bc(b^2 - c^2) + ca(c^2 - a^2)| \leq M(a^2 + b^2 + c^2)^2$$

holds for all real numbers a, b, and c.

4. Determine all pairs (x, y) of integers such that

$$1 + 2^x + 2^{2x+1} = y^2.$$

5. Let $P(x)$ be a polynomial of degree $n > 1$ with integer coefficients and let k be a positive integer. Consider the polynomial

$$Q(x) = \underbrace{P(P(\ldots(P(x)\ldots))}_{k\ P\text{'s}}$$

Prove that there are at most n integers t such that $Q(t) = t$.

6. Assign to each side b of a convex polygon $\mathcal{P}$ the maximum area of a triangle that has b as a side and is contained in $\mathcal{P}$. Show that the sum of the areas assigned to the sides of $\mathcal{P}$ is at least twice the area of $\mathcal{P}$.

SOLUTIONS

1. We note that $(\angle PBA + \angle PCA) + (\angle PBC + \angle PCB) = \angle B + \angle C$, so $\angle PBA + \angle PCA = \angle PBC + \angle PCB = (\angle B + \angle C)/2$. In triangle PBC, we have

$$\angle BPC = 180^\circ - (\angle PBC + \angle PCB) = 180^\circ - (\angle B + \angle C)/2.$$

It is clear that $\angle IBC + \angle ICB = (\angle B + \angle C)/2$, and so in triangle BCI, $\angle BIC = 180^\circ - (\angle B + \angle C)/2$. We conclude that $\angle BPC = \angle BIC$; that is, points B, C, I, and P lie on a circle.

Let the line AI meet the circumcircle of ABC again at point M. Then M is the midpoint of the arc BC not containing A. It is well known that M is the circumcenter of triangle BCI, and thus $MP = MI$.

In triangle APM, we have $AI + IM = AM \leq AP + PM = AP + IM$, implying that $AI \leq AP$. Equality holds if and only if $AM = AP + PM$; that is, A, P, and M are collinear, or $P = I$.

2. We claim the desired maxmimum is $M = 1003$. Define a *good triangle* to be an isosceles triangle having two good sides. Let $\mathcal{P} = P_1P_2 \ldots P_{2006}$, and let ω denote the circumcircle of $\mathcal{P}$. Without loss of generality, points $P_1, \ldots, P_{2006}$ are arranged in clockwise direction along ω. Then P_iP_j is good if and only if $i - j$ is odd. Since 2006 is even, a good triangle has exactly two good sides.

   First, we construct a triangulation with 1003 good triangles. We can first use the diagonals P_1P_3, $P_3P_5, \ldots, P_{2003}P_{2005}$, and $P_{2005}P_1$ to obtain 1003 good triangles. We can then complete the triangulation by providing an arbitrary triangulation of $P_1P_3 \cdots P_{2005}$ using 1001 diagonals. Therefore, $M \geq 1003$. Next, we show that $M \leq 1003$.

   Let $\widehat{P_iP_j}$ denote the clockwise directed broken line segment $P_iP_{i+1} \ldots P_j$ (where $P_{2006+k} = P_k$). We say $\widehat{P_iP_j}$ is non-major if it contains at most 1003 sides of $\mathcal{P}$.

   Let $P_iP_jP_k$ $(i < j < k)$ be a good triangle, with P_iP_j and P_jP_k being good segments. This means that there are an odd number of sides of $\mathcal{P}$ between P_i and P_j and also between P_j and P_k. We say $\widehat{P_iP_j}$ and $\widehat{P_jP_k}$ *belong to* triangle ABC.

   At least one side in each of these groups does not belong to any other good triangle. This is so because any odd triangle whose vertices are among the points between P_i and P_j has two sides of equal length and therefore has an even number of sides belonging to it in total. Eliminating all sides belonging to any other good triangle in $\widehat{P_iP_j}$ must therefore leave at least one side that belongs to no other good triangle. The same argument applies to $\widehat{P_jP_k}$. Let us assign these two sides (one in $\widehat{P_iP_j}$ and one in $\widehat{P_jP_k}$) to triangle $P_iP_jP_k$.

   To each good triangle we have thus assigned a pair of sides, with no two good triangles sharing an assigned side. It follows that at most 1003 good triangles can appear in the triangulation; that is, $M \leq 1003$.

3. Note that $ab(a^2 - b^2) + bc(b^2 - c^2) + ca(c^2 - a^2) = (a - b)(b - c)(c - a)(a + b + c)$. By symmetry, we may assume that $a > b > c$. Since each side of the inequality has degree four, it suffices to find the smallest M such that

$$(a - b)(b - c)(a - c) \leq M(a^2 + b^2 + c^2)^2 \tag{1}$$

   for real numbers $a > b > c$ with $a + b + c = 1$. Setting $a - b = x$ and $b - c = y$, we have $a - c = x + y$. Since $a + b + c = 1$, we have $(a - b)^2 + (b - c)^2 + (c - a)^2 = 3(a^2 + b^2 + c^2) - 1$. We can rewrite (1) as

$$9xy(x + y) \leq M[x^2 + y^2 + (x + y)^2 + 1]^2 \tag{2}$$

   for positive real numbers x and y.

   The AM-GM Inequality yields

$$\begin{aligned} x^2 + y^2 + (x + y)^2 + 1 &= \left(x^2 + \frac{1}{2}\right) + \left(y^2 + \frac{1}{2}\right) + \frac{(x + y)^2}{2} + \frac{(x + y)^2}{2} \\ &\geq \sqrt{2}x + \sqrt{2}y + 2xy + \frac{(x + y)^2}{2} \\ &\geq 4\sqrt[4]{(\sqrt{2}x)(\sqrt{2}y)(2xy) \cdot \frac{(x + y)^2}{2}} \\ &= 4\sqrt[4]{2x^2y^2(x + y)^2}, \end{aligned}$$

   or $(x^2 + y^2 + (x + y)^2 + 1)^2 \geq 16\sqrt{2}xy(x + y)$, implying that the minimum value of M is equal to $9\sqrt{2}/32$, obtained when $x^2 = y^2 = \frac{1}{2}$. When $x = y = \sqrt{2}/2$,

$\{a, b, c\}$ is an arithmetic progression with sum 1 and common difference $\sqrt{2}/2$; that is,

$$(a, b, c) = \left(\frac{1}{3} + \frac{1}{\sqrt{2}}, \frac{1}{3}, \frac{1}{3} - \frac{1}{\sqrt{2}}\right).$$

4. The answers are $(x, y) = (0, \pm 2)$ and $(x, y) = (4, \pm 23)$. We consider the nontrivial case when both x and y are positive.

The equation rewritten as $2^x(1 + 2^{x+1}) = y^2 - 1 = (y - 1)(y + 1)$ shows that the two numbers $y - 1$ and $y + 1$ have gcd $= 2$, and exactly one of them is divisible by 4. Hence $x \geq 3$ and one of $y - 1$ and $y + 1$ is divisible by 2^{x-1} but not by 2^x. Consequently, we may write

$$y = 2^{x-1}m + \epsilon, \tag{3}$$

where m is odd and $\epsilon = \pm 1$. Plugging this into the original equation we obtain

$$2^x(1 + 2^{x+1}) = (2^{x-1}m + \epsilon)^2 - 1 = 2^{2x-2}m^2 + 2^x m\epsilon,$$

or $1 + 2^{x+1} = 2^{x-2}m^2 + m\epsilon$. It follows that

$$1 - m\epsilon = 2^{x-2}(m^2 - 8). \tag{4}$$

If $\epsilon = 1$, (4) becomes $1 - m = 2^{x-2}(m^2 - 8)$, which has no solution. Thus $\epsilon = -1$, so (4) becomes $1 + m = 2^{x-2}(m^2 - 8) \geq 2(m^2 - 8)$, implying that $2m^2 - m - 17 \leq 0$. Hence $m \leq 3$. On the other hand, $m \neq 1$ by (4). Because m is odd, $m = 3$, leading to $x = 4$ by (4). Substituting these into (3) yields $y = 23$, completing our proof.

5. Let $\mathbb{N}$ denote the set of integers. We define

$$S_P = \{t \mid t \in \mathbb{N} \text{ and } P(t) = t\} \quad \text{and} \quad S_Q = \{t \mid t \in \mathbb{N} \text{ and } Q(t) = t\}.$$

Clearly, S_P is a subset of S_Q. Also note that there are at most n elements in S_P. This is so because $t \in S_P$ if and only if t is a root of polynomial $P(x) - x = 0$ of degree n, which has at most n roots. If $S_Q = S_P$, there is nothing to prove. We assume that S_P is a proper subset of S_Q, and that $t \in S_Q$ but $t \notin S_P$.

Consider the sequence $\{t_i\}_{i=0}^{\infty}$ with $t_0 = t$, $t_{i+1} = P(t_i)$ for every nonnegative integer i. Since $t \in S_Q$, $t_k = Q(t_0) = Q(t) = t = t_0$. Note that for each nonnegative integer i the sequence $\{t_i\}_{i=0}^{\infty}$ satisfies the divisibility relation $(t_{i+1} - t_i) \mid (P(t_{i+1}) - P(t_i)) = t_{i+2} - t_{i+1}$. Since $t_{k+1} - t_k = t_1 - t_0 = P(t) - t \neq 0$, each term in the chain of differences $t_1 - t_0$, $t_2 - t_1, \ldots, t_k - t_{k-1}$, $t_{k+1} - t_k$ is a nonzero divisor of the next one, and since $t_{k+1} - t_k = t_1 - t_0$, all these differences have equal absolute values. Let $t_i = \max\{t_0, t_1, \ldots, t_k\}$. Then $t_{i-1} - t_i = -(t_i - t_{i+1})$, or $t_{i-1} = t_{i+1}$. It is then not difficult to see that $t_{i+2} = t_i$ for every i; that is,

$$t_1 = P(t_0) \quad \text{and} \quad t_0 = P(t_1) \quad \text{or} \quad P(P(t_0)) = t_0.$$

Therefore,

$$S_Q = \{t \mid t \in \mathbb{N} \text{ and } P((P(t)) = t\}.$$

Without loss of generality, we may assume that $t_0 < t_1$. If s_0 is another element in S_Q, let $s_1 = P(s_0)$. (It is possible that $s_0 \in S_P$; that is, $s_1 = s_0$.) We further assume without loss of generality that $s_0 < s_1$ and $t_0 < s_0$; that is, $t_0 < s_0 \leq s_1$ and $t_0 < t_1$. Note that $s_1 - t_0$ divides $P(s_1) - P(t_0) = s_0 - t_1$. We must have $t_0 < s_0 < s_1 < t_1$.

Note that $s_0 - t_1$ also divides $P(s_0) - P(t_1) = s_1 - t_0$, it follows that $s_0 - t_1 = -(s_1 - t_0)$; that is,

$$t_0 + t_1 = s_0 + s_1 = s_0 + P(s_0).$$

In other words, s_0 is a root of the polynomial $P(x) + x = t_0 + t_1$. Since $P(x) + x$ has degree n, there are at most n (integer) roots (including t_0) of $P(x) + x$. Hence there are at most n elements in S_Q, completing our proof.

6. Define the *weight* of a side XY to be the area assigned to it, and define an *antipoint* of a side of a polygon to be one of the points in the polygon farthest from that side (and consequently forming the triangle with greatest area).

LEMMA 1. *For any side XY, Z is an antipoint if and only if the line l through Z parallel to XY does not go through the interior of the polygon. (Note that this means we can assume Z is a vertex, as we shall do so henceforth).*

Proof. Clearly, if Z is an antipoint l must not go through the interior of the polygon. Now if l does not go through the interior of the polygon, assume there is a point Z' farther away from XY than Z. Since the polygon is convex, the point $XZ' \cap l$ is in the interior of the polygon, which is a contradiction. ■

Suppose for the sake of contradiction that the sum of the weights of the sides is less than twice the area of some polygon. Then let S be the non-empty set of all convex polygons for which the sum of the weights is strictly less than twice the area. It is easy to check that no polygon in S can be a triangle, so we may assume all polygons in S have at least 4 sides.

We first prove by contradiction that there is some polygon in S such that all of its sides are parallel to some other side. Suppose the contrary; then consider one of the polygons in S which has the minimal number of sides not parallel to any other side (this exists by the well-ordering principle). Call this polygon $P = A_1A_2 \cdots A_n$, and WLOG let A_nA_1 be a side which is not parallel to any other side of P. Then let A_i be the unique antipoint of A_nA_1, and let A_u and A_v be respective antipoints of $A_{i-1}A_i$ and A_iA_{i+1}. Define X to be the point such that $A_uX || A_{i-1}A_i$, $A_vX || A_iA_{i+1}$.

Now consider the set $T \subset P$ of points that are strictly on the same side of A_uA_v as A_nA_1. First of all, for any side in T, A_i must be its antipoint, since the line through A_i parallel to A_jA_{j+1} does not go through the interior of P. Similarly, any vertex in T is not the antipoint of any side.

We now look at the polygon $P' = A_vA_{v+1} \cdots A_{u-1}A_uX$. First of all, it is clear that P' has fewer sides which are not parallel to any other side than P. The area $[P'] - [P]$ is simply $[A_1A_2 \cdots A_{v-1}A_vXA_uA_{u+1} \cdots A_n]$. The weights of the side A_jA_{j+1} is the same in both P' and P for $v \leq j < u$, but for P', the sum of the weights of the remaining two sides is $[XA_uA_iA_v]$, as A_i is an antipoint of both A_uX and A_vX. Meanwhile, the sum of the weights of remaining sides (i.e. sides in T) for P is $[A_1A_2 \cdots A_{v-1}A_vA_iA_uA_{u+1} \cdots A_n]$. Hence the difference in the sums of weights of P' and P is

$$\begin{aligned}&[XA_uA_iA_v] - [A_1A_2 \cdots A_{v-1}A_vA_iA_uA_{u+1} \cdots A_n]\\ &\quad = [A_1A_2 \cdots A_{v-1}A_vXA_uA_{u+1} \cdots A_n],\end{aligned}$$

the same as the difference in area (and both differences were positive). Therefore, if the sum of weights of P was less than $2[P]$, then certainly the sum of weights of P' must be less than $2[P']$, so that $P' \in S$. However, this contradicts the minimality of the number of non-parallel sides in P, so there exists a polygon in S with opposite sides parallel.

Now, we will let R be the non-empty set of all polygons in S with all sides parallel to the opposite side. Note that all polygons in R must have an even number of sides. We will show that there is a parallelogram in R.

Suppose not, and that $Q = B_1B_2 \cdots B_{2m}$ is one of the polygons in R with the minimal number of sides, and $m \geq 3$. Let $X = B_1B_2 \cap B_{2m-1}B_{2m}$ and $Y = B_{m-1}B_m \cap B_{m+2}B_{m+1}$. Set $Q' = XB_2B_3 \cdots B_{m-1}YB_{m+2} \cdots B_{2m}$. We propose that the increase in the sum of weights going from Q to Q' is at most twice the increase in area, so that $Q' \in R$.

To aid us, we will let h_X and h_Y be the respective distances of X and Y from $B_{2m}B_1$ and B_mB_{m+1}. The increase in weight is

$$[XB_{m+1}B_1] + [XB_{2m}B_m] + [YB_mB_{2m}] + [YB_{m+1}B_1] - [B_1B_{2m}B_m] - [B_{2m}B_mB_{m+1}],$$

which is equal to

$$[XB_1Y] + [XB_{2m}Y] - [B_1B_{2m}B_m] + [YB_mX] + [YB_{m+1}X] - [B_{2m}B_mB_{m+1}]$$

or

$$[XB_1B_{2m}] + \frac{h_Y \cdot B_1B_{2m}}{2} + [YB_mB_{m+1}] + \frac{h_X \cdot B_mB_{m+1}}{2}.$$

On the other hand, the increase in area is $[XB_1B_{2m}] + [YB_mB_{m+1}]$.

It remains to show that the first expression is at most twice the second, or in other words; that is, to show that

$$\frac{h_Y \cdot B_1B_{2m}}{2} + \frac{h_X \cdot B_mB_{m+1}}{2} \leq [XB_1B_{2m}] + [YB_mB_{m+1}] = \frac{h_X \cdot B_1B_{2m}}{2} + \frac{h_Y \cdot B_mB_{m+1}}{2},$$

which is equivalent to $(h_X - h_Y)(B_1B_{2m} - B_mB_{m+1}) \geq 0$.

Noting that triangles $B_1B_{2m}X$ and $B_{m+1}B_mY$ are similar, we have $h_X/h_Y = B_1B_{2m}/B_mB_{m+1}$, so the above inequality holds.

With the inequality proven, we now know that $Q' \in R$, and yet Q' has fewer sides than Q. This contradicts the minimality of the number of sides of Q, so there exists a parallelogram in R. However, the sum of the weights of a parallelogram clearly equals twice its area, so this contradicts the entire existence of S, as desired.

2006 Olympiad Results

The top twelve students on the 2006 USAMO were (in alphabetical order):

Yakov Berchenko-Kogan	Needham B. Broughton High School	Raleigh, NC
Yi Han	Phillips Exeter Academy	Exeter, NH
Sherry Gong	Phillips Exeter Academy	Exeter, NH
Taehyeon Ko	Phillips Exeter Academy	Exeter, NH
Brian Lawrence	Montgomery Blair High School	Silver Spring, MD
Tedrick Leung	North Hollywood High School	N. Hollywood, CA
Richard Mccutchen	Montgomery Blair High School	Silver Spring, MD
Peng Shi	Sir John A. MacDonald Collegiate Institute	Toronto, ON
Yi Sun	The Harker School	San Jose, CA
Arnav Tripathy	East Chapel Hill High School	Chapel Hill, NC
Alex Zhai	University Laboratory High School	Urbana, IL
Yufei Zhao	Don Mills Collegiate Institute	Don Mills, ON

Brian Lawrence, was the winner of the Samuel Greitzer-Murray Klamkin award, given to the top scorer(s) on the USAMO. Brian Lawrence, Alex Zhai, and Yufei Zhao placed first, second, and third, respectively. They were awarded college scholarships of $20000, $15000, $10000, respectively, by the Akamai Foundation. The Clay Mathematics Institute (CMI) award, for a solution of outstanding elegance, and carrying a $5000 cash prize, was presented to Brian Lawrence for his solution to USAMO Problem 5, presented as the third solution to the problem in this book.

The USA team members were chosen according to their combined performance on the 35th annual USAMO, and the Team Selection Test that took place at the Mathematical Olympiad Summer Program (MOSP), held at the University of Nebraska-Lincoln, June 5 - July 1, 2005. Members of the USA team at the 2006 IMO (Ljubljana, Slovenia) were Zachary Abel, Zarathustra (Zeb) Brady, Taehyeon (Ryan) Ko, Yi Sun, Arnav Tripathy, and Alex Zhai. Zuming Feng (Phillips Exeter Academy) and Alex Saltman (Stanford University) served as team leader and deputy leader, respectively. The team was also accompanied by Steven Dunbar (University of Nebraska-Lincoln), as observer of the deputy leader.

There were 498 contestants from 90 countries in the 2006 IMO. Gold medals were awarded to students scoring between 28 and 42 points, silver medals to students scoring between 19 and 27 points, and bronze medals to students scoring between 15 and 18 points. There were 42 gold medalists, 89 silver medalists, 122 bronze medalists, and honorable mentions (awarding to non-medalists solving at least one problem completely). There were 3 perfect papers (Iurie Boreico from Republic of Moldova, Zhiyu Liu from People's Republic of China, and Alexander Magazinov from Russian Federation) on this difficult exam, even though it has two relatively easy entry level problems (in problems 1 and 4). Tripathy's 30 tied for 16th place overall. The team's individual

performances were as follows:

Able	SILVER Medallist
Brady	GOLD Medallist
Ko	SILVER Medallist
Sun	SILVER Medallist
Tripathy	GOLD Medallis
Zhai	SILVER Medallist

In terms of total score (out of a maximum of 252), the highest ranking of the 90 participating teams were as follows:

China	214
Russia	174
Korea	170
Germany	157
USA	154
Romania	152
Japan	146
Iran	145
Moldova	140
Taiwan	136
Poland	133
Italy	132

For more information about the USAMO or the MOSP, contact Steven Dunbar at `sdunbar@math.unl.edu`.

UNITED STATES POSTAL SERVICE — **Statement of Ownership, Management, and Circulation (All Periodicals Publications Except Requester Publications)**

1. Publication Title	2. Publication Number	3. Filing Date
Mathematics Magazine	0025-5700	9/10/07
4. Issue Frequency	5. Number of Issues Published Annually	6. Annual Subscription Price
Bimonthly Except July + August	5	$99

7. Complete Mailing Address of Known Office of Publication (Not printer) (Street, city, county, state, and ZIP+4®): Mathematical Association of America, 1529 18th St, NW, Washington, DC 20036 — Contact Person: G. Waldon; Telephone (Include area code): 202-387-5200

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

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

Publisher (Name and complete mailing address): MAA, 1529 18th St, NW, Washington D.C. 20036

Editor (Name and complete mailing address): Allen Schwenk, Western Michigan University, Kalamazoo, MI 49008

Managing Editor (Name and complete mailing address): Harry Waldon, MAA, 1529 18th St, NW, Washington, D.C. 20036

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

Full Name	Complete Mailing Address
Mathematical Assn of America	1529 18th St, NW, Washington DC 2003

11. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box → X None

Full Name	Complete Mailing Address

12. Tax Status (For completion by nonprofit organizations authorized to mail at nonprofit rates) (Check one)
The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes:
X Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months (Publisher must submit explanation of change with this statement)

PS Form 3526, September 2006 (Page 1 of 3 (Instructions Page 3)) PSN 7530-01-000-9931 PRIVACY NOTICE: See our privacy policy on www.usps.com

13. Publication Title: Mathematics Magazine	14. Issue Date for Circulation Data Below: June 2007	
15. Extent and Nature of Circulation	Average No. Copies Each Issue During Preceding 12 Months	No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies (Net press run)	11,000	11,000
b. Paid Circulation (By Mail and Outside the Mail) (1) Mailed Outside-County Paid Subscriptions Stated on PS Form 3541 (Include paid distribution above nominal rate, advertiser's proof copies, and exchange copies)	10,500	10,500
(2) Mailed In-County Paid Subscriptions Stated on PS Form 3541 (Include paid distribution above nominal rate, advertiser's proof copies, and exchange copies)	0	0
(3) Paid Distribution Outside the Mails Including Sales Through Dealers and Carriers, Street Vendors, Counter Sales, and Other Paid Distribution Outside USPS®	0	0
(4) Paid Distribution by Other Classes of Mail Through the USPS (e.g. First-Class Mail®)	0	0
c. Total Paid Distribution (Sum of 15b (1), (2), (3), and (4))	10,500	10,500
d. Free or Nominal Rate Distribution (By Mail and Outside the Mail) (1) Free or Nominal Rate Outside-County Copies included on PS Form 3541	0	0
(2) Free or Nominal Rate In-County Copies included on PS Form 3541	0	0
(3) Free or Nominal Rate Copies Mailed at Other Classes Through the USPS (e.g. First-Class Mail)	0	0
(4) Free or Nominal Rate Distribution Outside the Mail (Carriers or other means)	100	100
e. Total Free or Nominal Rate Distribution (Sum of 15d (1), (2), (3) and (4))	100	100
f. Total Distribution (Sum of 15c and 15e)	10,600	10,600
g. Copies not Distributed (See Instructions to Publishers #4 (page #3))	400	400
h. Total (Sum of 15f and g)	11,000	11,000
i. Percent Paid (15c divided by 15f times 100)	99 %	99 %

16. Publication of Statement of Ownership
X If the publication is a general publication, publication of this statement is required. Will be printed in the October 2007 issue of this publication. ☐ Publication not required.

17. Signature and Title of Editor, Publisher, Business Manager, or Owner — Date: 9/10/07

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

PS Form 3526, September 2006 (Page 2 of 3)

CONTENTS

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