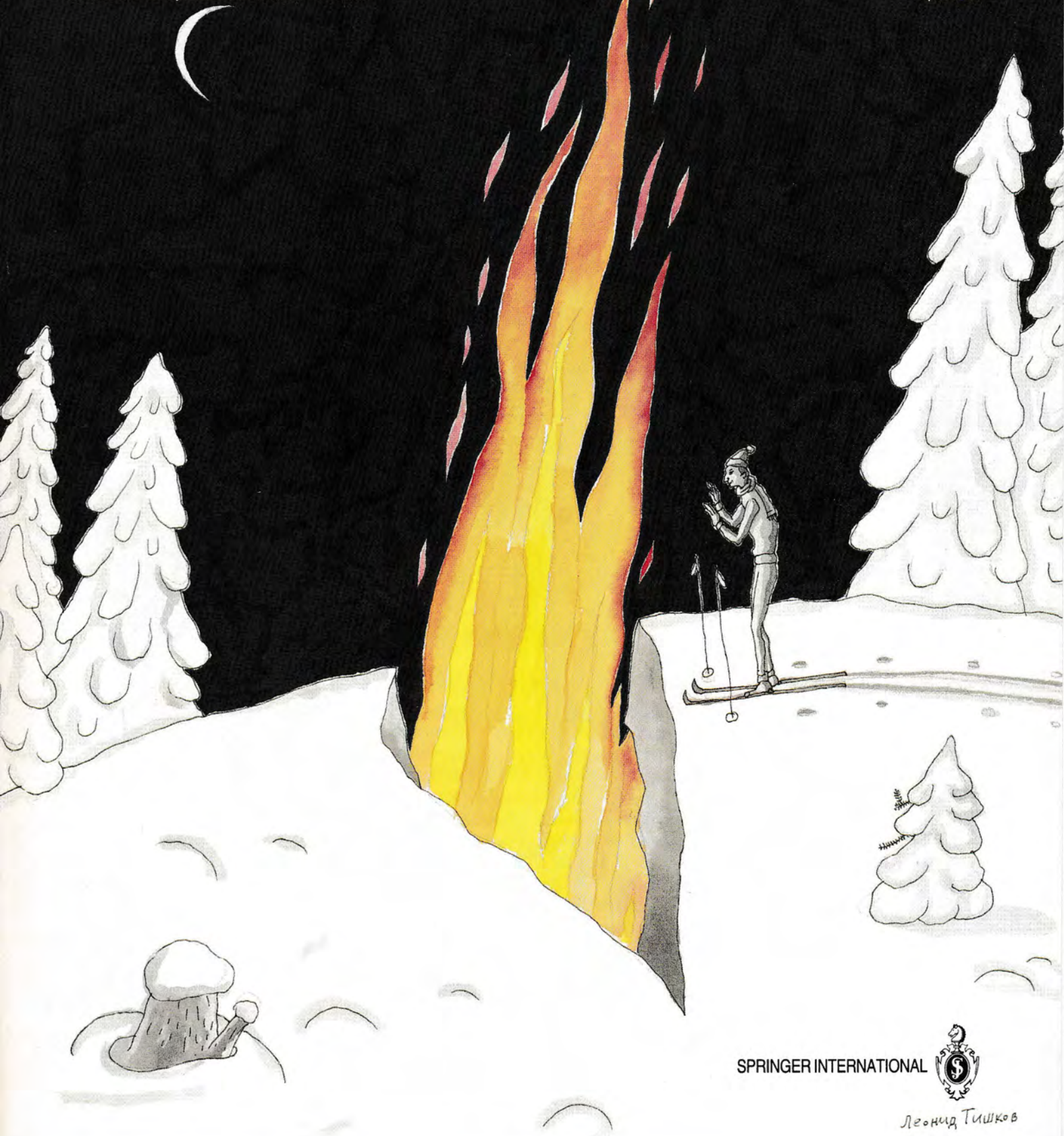


QUANTUM

JANUARY/FEBRUARY 1995

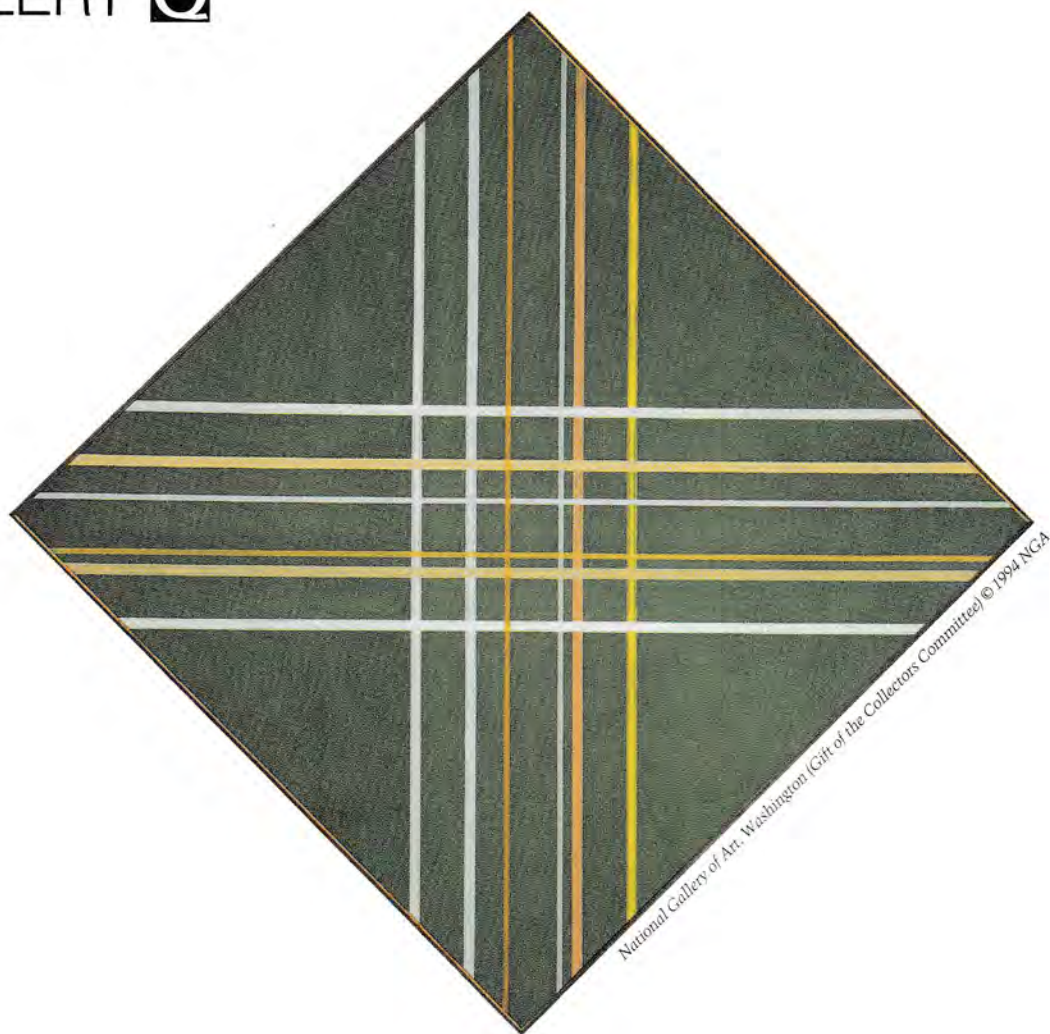
\$5.95



SPRINGER INTERNATIONAL



Леонид Тишков



Another Time (1973) by Kenneth Noland

THE SMART ALECK WILL HAVE A FIELD DAY WITH this painting. "Are you sure the right end is up?" "Hey, that would make a nice flannel shirt!" "I like it, but—could you do it in red?" And who's to say these aren't legitimate responses?

In past visits to Gallery Q we've had occasion to remark on the special challenges presented by abstract art. When it seems that an artist is exploring the medium for its own sake, or solving certain technical problems, we are certainly within our rights to ask: "What's the point?" This is the same question sometimes posed to researchers in pure mathematics. Are the answers the same?

A fundamental question arises: where does the urge to abstraction come from? Is appreciation of the abstract innate or learned? Perhaps you have produced doodles that you found beautiful. They depict nothing, yet they're pleasing. Is "Another Time" a kind of doodle? If so, what are the limits of our appreciation of doodles—

our own and others'? Whose doodles get to hang in the museums of the world?

What we seem to appreciate in abstract art is form, texture, balance—qualities that are difficult to talk about. Some would find "Another Time" both evocative and elegant. It might remind you of the grid of city streets. In fact, it looks a bit like *Quantum's* birthplace, Washington, D.C., which is a very green city laid out in a diamond, and whose x- and y-axes (passing through the Capitol) are off-center. Perhaps a reader who is charmed by the painting will look for instances of the "golden section" (see the January/February 1991 Gallery Q) or find some other explanation for its design.

Turn to page 34 for an exploration of "elegance" in mathematics. As in art, mathematicians "know it when they see it." And for an exercise in "nonpure" mathematics involving networks of streets, see "The School Bus and the Mud Puddles" on page 24.

QUANTUM

JANUARY/FEBRUARY 1995

VOLUME 5, NUMBER 3



Cover art by Leonid Tishkov

While cross-country skiers may dream of a roaring fire in the middle of their trek, our planet generally keeps its inner heat to itself. Hot springs, geysers, and lava flows are relatively rare on the Earth's surface. It's not so hard to figure out why the hot stuff stays bottled up in the planet's interior, or how it occasionally leaks out through cracks in the Earth's crust.

You may have a harder time explaining where the heat comes from. In "Taking the Earth's Temperature," Alexey Byalko explores this question, and along the way he uncovers some interesting facts about the Earth's thermal history and its present structure. In a companion piece, A. G. W. Cameron presents a theory for the creation of the Earth's little sister—the Moon.

We hope our readers in the northern climes have ample opportunity to engage in winter sports and outdoor activities in the coming months, whether or not they enjoy the amenity depicted on our cover. And we wish all our readers everywhere a healthy and happy New Year!

FEATURES

- 4 Deep-down Truths**
Taking the Earth's temperature
by Alexey Byalko
- 10 Mathematical Strands**
Braids and knots
by Alexey Sosinsky
- 16 Simulated Creation**
When a body meets a body
by A. G. W. Cameron
- 24 Tools of the Trade**
The school bus and the mud puddles
by Thomas P. Dence

DEPARTMENTS

- | | |
|---|---|
| 2 Publisher's Page | 41 Innovators
<i>The legacy of Norbert Wiener</i> |
| 21 Brainteasers | 45 Sticking Points
<i>Important components of learning components</i> |
| 23 How Do You Figure? | 48 Math Investigations
<i>Geometry in the pagoda</i> |
| 30 In the Lab
<i>A magical musical formula</i> | 49 Happenings
<i>Programming challenges ... Bulletin Board</i> |
| 32 Kaleidoscope
<i>Surfing the electromagnetic spectrum</i> | 53 Crisscross Science |
| 34 Ruminations
<i>What is elegance?</i> | 54 Answers, Hints & Solutions |
| 36 Physics Contest
<i>Cloud formulations</i> | 62 Toy Store
<i>Triad and true</i> |

It's all Greek to me!

The significance of symbols in math and science

WE USE ALL KINDS OF SYMBOLS. I remember the confusion this first caused me as a novice student, and I wonder if our readers have had similar experiences. Physicists and mathematicians use symbols so routinely, they seldom give them a second thought.

Consider the kinds of symbols we typically use. We use figures like 1, 2, 3 as number symbols to represent specific natural numbers. Then we introduce a zero and a minus sign in front of some them to expand this set to form the integers: -3 , -2 , -1 , 0 , 1 , 2 , 3 ... Now, keep in mind that the minus sign used here doesn't mean the same thing as the minus sign used for the binary operation called subtraction. Can you see how a kid might get confused?

Here's another example. We use letters like a , b , c , d , or x , y , z , to represent numbers. Unless we make it clear when we define them, they could be any kind of numbers, even complex numbers. Sometimes we use symbols like A , B , P , E to represent a point along a line, a location on a surface, or the presence of some object. At other times, the symbol may represent a physical quantity, like E for energy.

On top of the letters themselves, we sometimes add a typographic feature. Take vectors and tensors—they require something extra if they're to be understood. Vectors are written as bold letters, or they carry arrows on top. A tensor might have a double arrow or a subscript of some sort. In fact, subscripts can be

considered the "adjectives" of scientific discourse.

So scientists and mathematicians have, by mutual agreement, created a host of symbols for numbers, objects, points, surfaces, volumes, and physical quantities. It's a lot to keep track of. But wait—there's more.

We need symbols to express the *relationship* between other symbols: less than, equal to, more than, proportional to, and so on. Then we have *operators*—things that change one thing or quantity into another. All of these have their own rules of operation—for instance, the associative and commutative laws may or may not apply to a particular operation for certain kinds of quantities.

To further complicate things, many symbols in the sciences represent not only a number but a unit associated with that physical quantity. This makes science more difficult than the corresponding mathematics.

I remember when I first had to learn symbols. I had trouble transferring from mathematics to physics. The same equation in physics didn't look the same in math. The symbols were different. I was especially troubled by Greek letters, and for

some reason, if I couldn't pronounce the name of the letter, I had trouble with what it meant. When an equation used Greek letters, especially some of the more obscure ones, I had even more trouble with the physics.

To top it all off, a student soon finds that there is no real consistency in the choice or use of symbols as one moves from text to text. You read something where t is a time interval and T is temperature; then you read something else where T is absolute temperature, t is time, and θ is Celsius temperature.

My point is simply that science and mathematics overflow with symbols of all kinds, and those of us who teach or write about science and math should be careful when we describe what those symbols mean. We need to recognize that the novice is often slowed by the effort to assimilate the connection between the symbol and the physical or mathematical quantity it represents.



Symbols are part of a chain of reasoning. If we want young people to learn more than the mere manipulation of symbols, we need to give them direct experience with the *phenomena* underlying the *concepts* represented by *symbols*. Great scientists have that deep understanding of what the symbols represent, and when they see a mathematical expression like ∇V , they see it as a gradient—a vector that gives the direction and magnitude of the greatest rate of change of the voltage V per centimeter or meter. They don't just see this in terms of its defining unit vectors and associated derivatives, although they recognize that forming scalar products along any direction will give the rate of change in that direction. Similarly, understanding symbols in this deep sense allows one to see the equation $\nabla \cdot \mathbf{B} = 0$ as indicating that magnetic fields have no sources. Such field lines must close on one another. There are no magnetic poles if this equation is true.

It's a wondrous thing that a correspondence exists between mathematics, with its symbols and operations, and empirical laws of nature. But students will grasp that only if their teachers can impart a deep sense of what each symbol stands for. Otherwise, it's—like the title says.

And vice versa

"It's all English to me!" That's what many potential readers of *Quantum* around the world might well say. So I'm pleased to announce the publication of a Greek version of *Quantum*. We hope this is but the first in a series of foreign-language editions.

The Greek-language *Quantum* is produced by Katoptro Publications, a publishing house in Athens devoted almost exclusively to scientific and educational titles. In the words of Alex Mamalis, the director of Katoptro: "We believe that *Quantum* is exactly what not only Greek students and teachers but also all the students and teachers of the world need—the ideal magazine." I couldn't agree more.

—Bill G. Aldridge

QUANTUM

THE MAGAZINE OF MATH AND SCIENCE

A publication of the National Science Teachers Association (NSTA)

and Quantum Bureau of the Russian Academy of Sciences

in conjunction with

the American Association of Physics Teachers (AAPT)

and the National Council of Teachers of Mathematics (NCTM)

The National Science Teachers Association is an organization of science education professionals and has as its purpose the stimulation, improvement, and coordination of science teaching and learning.

Publisher

Bill G. Aldridge, Executive Director, NSTA

Associate Publisher

Sergey Krotov, Director, Quantum Bureau,
Professor of Physics, Moscow State University

Founding Editors

Yuri Ossipyan, President, Quantum Bureau
Sheldon Lee Glashow, Nobel Laureate (physics), Harvard University
William P. Thurston, Fields Medalist (mathematics), University of California, Berkeley

Field Editors for Physics

Larry D. Kirkpatrick, Professor of Physics, Montana State University, MT
Albert L. Stasenko, Professor of Physics, Moscow Institute of Physics and Technology

Field Editors for Mathematics

Mark E. Saul, Computer Consultant/Coordinator, Bronxville School, NY
Vladimir Dubrovsky, Associate Professor of Mathematics, Moscow State University

Managing Editor

Timothy Weber

Staff Artist

Sergey Ivanov

Supervising Production Editor

Madeline Kraner

Editorial Consultants

Alexander Buzdin, Professor of Physics, Moscow State University
Yuly Danilov, Senior Researcher, Kurchatov Institute
Larissa Panyushkina, Managing Editor, Quantum Bureau

International Consultant

Edward Lozansky

Advertising Managers

Paul Kuntzler (Washington office)
Bob Vrooman (New York office)

Advisory Board

Bernard V. Khoury, Executive Officer, AAPT
James D. Gates, Executive Director, NCTM
George Berzsenyi, Professor of Mathematics, Rose-Hulman Institute of Technology, IN
Arthur Eisenkraft, Science Department Chair, Fox Lane High School, NY
Karen Johnston, Professor of Physics, North Carolina State University, NC
Margaret J. Kenney, Professor of Mathematics, Boston College, MA
Thomas D. Rossing, Professor of Physics, Northern Illinois University, IL
Alexander Soifer, Professor of Mathematics, University of Colorado–Colorado Springs, CO
Barbara I. Stott, Mathematics Teacher, Riverdale High School, LA
Carol-ann Tripp, Physics Teacher, Providence Country Day School, RI

Quantum (ISSN 1048-8820) is published bimonthly by the National Science Teachers Association in cooperation with Springer-Verlag New York Inc. Volume 5 (6 issues) will be published in 1994–1995. *Quantum* contains authorized English-language translations from *Kvant*, a physics and mathematics magazine published by Quantum Bureau of the Russian Academy of Sciences, as well as original material in English. Editorial offices: NSTA, 1840 Wilson Boulevard, Arlington VA 22201-3000, telephone (703) 243-7100. Production offices: Springer-Verlag New York, Inc., 175 Fifth Avenue, New York NY 10010-7858.

Advertising:

Advertising Representatives: (Washington) Paul Kuntzler (202) 328-5800; (New York) Brian Skepton (212) 460-1575; and G. Probst, Springer-Verlag GmbH & Co. KG, D-14191 Berlin, Germany, telephone (0) 30-82 07-1, telex 185 411.

Second class postage paid at New York, NY, and additional mailing offices. Postmaster: send address changes to: *Quantum*, Springer-Verlag New York, Inc., Journal Fulfillment Services Department, P. O. Box 2485, Secaucus NJ 07096-2485. Copyright © 1995 NSTA. Printed in U.S.A.

Subscription Information:

North America: Student rate: \$15; Personal rate (nonstudent): \$20; Institutional rate: \$34; Single Issue Price: \$5.95. Rates include postage and handling. (Canadian customers please add 7% GST to subscription price. Springer-Verlag GST registration number is 123394918.) Subscriptions begin with next published issue (backstarts may be requested). Bulk rates for students are available. Send orders to *Quantum*, Springer-Verlag New York, Inc., P.O. Box 2485, Secaucus NJ 07096-2485; or call 1-800-SPRINGER (777-4643) (in New York, call (201) 348-4033).

All Countries Outside North America: Subscription rates in U.S. currency as above (all rates calculated in DM at the exchange rate current at the time of purchase) plus postage and handling. SAL (Surface Air-mail Listed) is mandatory for Japan, India, Australia, and New Zealand. Customers should ask for the appropriate price list. Air mail delivery to all other countries is available upon request. Orders may be placed through your bookseller or directly through Springer-Verlag, Postfach 31 13 40, D-10643 Berlin, Germany.



Taking the Earth's temperature

How hot is the heart of our planet?

by Alexey Byalko

THE EARTH'S CRUST IS MOBILE. Everybody knows that it shifts when an earthquake occurs. These shifts are significant, but random and localized. Some earthquakes, though, are caused by the directed, nonrandom motion of the continents and the ocean floor.

The physical mechanism that leads to the directed motion of the Earth's crust has to do with the release of heat stored in the Earth's interior when our planet was formed. That's what this article is about: this heat, the thermal history of the Earth, and the planet's internal structure.

Drilling for knowledge

We gain some understanding of the structure of our planet's interior by means of deep wells. Wells are primarily drilled to extract oil and gas from the depths to the surface, but even if a well comes up dry, we can use it to study the interior of the planet.

Here's how a deep well is created. A drilling solution (water with salts added to increase its density) is pumped under high pressure to the point near the well bottom where a turbine drill rotates against the rock. The drill bits reduce the stone to fragments that are pushed to the surface by the same drilling liquid.

Why not drill a hole all the way to the center of the Earth? We would

learn something about the internal structure of our planet, and maybe we'd find something useful along the way. Unfortunately, it's impossible in principle. Let's convince ourselves that we can't drill a well deeper than 15 km.

First, we need some theory. As you know, any solid material is destroyed if the stresses placed on it are sufficiently high. However, there are two different kinds of stress σ (that is, force per unit area). One is uniform stress—in other words, the ordinary kind of pressure P . According to Pascal's law it compresses material with forces that are equal from every direction. Generally, nothing happens with either liquids or solids under pressure except an increase in density. The other kind of stress is nonuniform *shear* stress. Applied to liquids, shear stress leads to flows with velocities that are roughly proportional to the stress. Applied to solids, it causes nothing at first. But when the shear stress increases, it causes small cracks; then the cracks grow, which sometimes leads to the complete destruction of the solid sample, but generally it just reduces the external forces. The maximum shear stress that a material can sustain without cracking is called its strength $\sigma_{\max}$. The strength of most hard, crystalline rocks from the Earth's crust does not exceed $\sigma_{\max} \cong 2-3 \cdot 10^8$ Pa.

Now let's return to our well. At

any depth H we know the pressure of the surrounding rock: $P_0 = \rho_0 g H$, where ρ_0 is the density of the rock—generally in the range 2.6–2.9 g/cm³. We also know the pressure of the drilling liquid inside the well shaft: $P_w = \rho_w g H$. If one could find a liquid with a density $\rho_w = \rho_0$ that is chemically stable and does not interact with rock (and is also cheap enough), that would be great. Unfortunately, our choice of liquids is rather narrow. In fact, only water, aqueous solutions of salts, and thin suspensions of minerals in water are available to us, but we can't increase the density ρ_w to more than about 1.5 g/cm³. Yes, there are liquids with higher densities (for instance, sulfuric acid, whose density is 1.84 g/cm³), and not all of them are even as aggressive as sulfuric acid. But the search for the magic liquid described above has been in vain.

The pressure difference between the inside of the well and the surrounding wall causes shear stresses in the rock. Naturally, they decrease with distance from the well and reach a maximum near the well wall, equal to $\sigma = \frac{2}{3}(P_0 - P_w) = \frac{2}{3}(\rho_0 - \rho_w)gH$. (I can't prove to you here that the coefficient $\frac{2}{3}$ in this formula is correct—it follows from a more complicated investigation.)

Thus, at the bottom of a deep well the shear stresses start to approach the limit of the rock strength $\sigma_{\max}$ and a deep well collapses under the



pressure of the surrounding rock. The maximum depth of a well drilled in granite ($\rho_0 = 2.7 \text{ g/cm}^3$) is

$$\delta P \equiv (\rho_0 - \rho_w)gH < \sigma_{\max} \\ H < 10\text{--}15 \text{ km.}$$

The deepest well in the world (on Russia's Kola Peninsula) reaches a depth of 12 km—approximately 0.2% of the Earth's radius.

When oil wells are drilled or tunnels are cut through mountains, one discovers that the temperature of the rock increases with depth. This temperature increase is not constant with depth and depends on the location of the well or tunnel. It's not easy to measure it exactly. Seasonal changes in temperature come into play at shallow depths. Also, heat flow is transferred not only by thermal conductivity but by water flowing slowly along the cracks in the rock. At great depths the rock is heated some more when the well partially collapses.

Since there is a limit to the drilling depth, direct measurements of the Earth's temperature are possible only in a thin surface layer. On average, the increase in temperature with depth is $dT/dz = 3 \cdot 10^{-2} \text{ K/m} = 30 \text{ K/km}$. Local temperature gradients actually deviate from this average (the so-called geotherm). They are minimal in old granites—for instance, in the Ural Mountains (about 15 K/km) and maximal in regions of volcanic activity (up to 100 K/km).

The heat flux q is the power emerging from the depths through each square meter of the surface. It's equal to the product of the coefficient of thermal conductivity κ and the temperature gradient: $q = \kappa dT/dz$. If the primary source of the inner heat is deep down, it seems evident that the heat flow is almost constant. Measurements, however, do not provide much support for such a constancy of the heat flow.

The increase in temperature with depth doesn't correspond exactly to the equation for heat conductivity—the heat flux sometimes varies unexpectedly. The reason for such deviations is understandable in principle. Imagine that in the past at

some depth there was a fracture or movement of the layers, reducing the stresses that had accumulated there. Then the temperature around it increases and the calculated heat flux is not constant with depth. Deviations from the average geotherm that occur because of substantial movement of the crust are preserved for a long time. In fact, they provide a historical record of past earthquakes.

There have also been direct measurements of sudden jumps in temperature. On December 7, 1988, the day of the terrible earthquake in Armenia, the temperature in the deep Kola well increased at some levels by 10–15 degrees. After a few days it dropped to the previous level. This shows that the jump was caused by a partial removal of stresses near the walls of the well. But even more importantly, this example elucidates the mechanism itself: the rise in temperature is determined not only by heat flux but by stresses in the rock as well.

Thus, we can only estimate the heat flux in the Earth, assuming it to be constant relative to depth at every location. The coefficients of heat conductivity of the most widely distributed deep rock—basalt and granite—are sufficiently close to each other. We'll take $\kappa = 3 \text{ W/(m} \cdot \text{K)}$ for our estimate. To find the total thermal losses of the Earth—that is, its thermal power—we multiply the thermal conductivity coefficient κ and the near-surface temperature gradient $dT/dz = 0.03 \text{ K/m}$ by the Earth's surface area. Such an estimate of the total thermal power Q of the Earth gives

$$Q = \frac{4\pi R_{\oplus}^2 \kappa dT}{dz} \approx 5 \cdot 10^{13} \text{ W,}$$

where $R_{\oplus}$ is the Earth's radius.

A more accurate summation based on known wells shows that the Earth's thermal power calculated in such a way is equal to $4.2 \cdot 10^{13} \text{ W}$, consisting of $1.1 \cdot 10^{13} \text{ W}$ from the flow through the land surface and $3.1 \cdot 10^{13} \text{ W}$ from the flow through the ocean floor. The contributions

of the land and ocean are approximately proportional to their areas, which means that the densities of the heat flux through the continental and oceanic crusts are virtually the same.

The temperature must certainly continue to increase with depth, since the heat flux from the interior cannot simply vanish. Here's a way to measure temperatures at depths much greater than those of wells.

Diamonds and kimberlites

Sometimes rocks that were formed at a very great depth can be found on the Earth's surface—diamonds, for example. Diamond consists of the chemical element carbon, but its crystal structure differs from other carbon forms—for instance, platelike graphite or amorphous carbon (coal, soot). Soot consists of chains of carbon atoms of a different length with double bonds $\dots = \text{C} = \text{C} = \dots$; graphite consists of bounded flat structures made up of C_6 -type benzene rings. Sometimes carbon crystallizes in the even more exotic C_{60} icosahedral structure called the fullerene.¹ But in diamond, the four bonds of each carbon atom are directed at the vertices of a tetrahedron. This produces the most compact structure its atoms can attain. You can turn graphite into diamond in a lab, but you need to produce pressures and temperatures close to those found at a depth of at least 70 km.

How do natural diamonds end up on the Earth's surface? They are transported upward from the depths by the rather rare expulsion of a low-density rock called kimberlite (named after Kimberley, South Africa, where a large deposit of diamonds was found in a deep, narrow vein of this mineral). Diamonds aren't formed simultaneously with kimberlites—they're swept along by the rapidly rising mass of kimberlite.

In addition to diamonds, other kinds of minerals are found in

¹See "Follow the Bouncing Buckyball" in the May/June 1994 issue of *Quantum*.—Ed.

kimberlite veins. These stones can't survive for long in a medium of kimberlite at high temperatures. They are called xenoliths, from the Greek for "alien stones." The mineral composition and structure of xenoliths tell us much about the state of the Earth's crust at great depths. The very existence of diamonds already gives us some information about the Earth's interior, since they couldn't be formed at low pressures and temperatures.

If a mineral is compressed, its crystal structure changes radically at a certain pressure. The critical pressures and corresponding temperatures for various minerals are well known from laboratory measurements. Each mineral has its own set of such transitions. Moreover, if two grains of minerals come into close contact, the chemical composition of each grain changes depending on the pressure and temperature. All this makes it possible to determine both the pressure and the temperature of the medium where the xenolith was formed. If the xenolith con-

sists of several mineral grains joined together, this gives several pressure-temperature data points. The pressure data make it possible to find the depth at which these processes took place. Sometimes mineralogists are able to determine how rapidly a xenolith changed depth—in effect, they can write its biography.

Figure 1 shows measurements of temperature and pressure for several xenoliths together with the average near-surface temperature gradient of the Earth. Pressures are recalculated for depth on the other vertical scale. The xenoliths whose T - P graphs are given were found in various locations. Notice that some graphs are almost closed. This means that the rock from which the xenolith was formed first descended, then began to rise to the surface.

What causes this heat flux from the depths of the Earth? What is the source of its energy?

The Earth's heat sources

One of the sources of the Earth's inner heat is energy released by ra-

dioactivity. Rock contains a small but measurable amount of uranium. It's especially significant in granite, where it reaches several millionths of the rock's mass. The most widespread uranium isotope, ^{238}U , is the major contributor to the nuclear heat released in rock. With each decay, an alpha particle with an energy of 4.2 MeV is emitted by the nucleus of a uranium atom. After traveling about 10^{-5} m, it is stopped by the rock and passes its energy to it, heating it up.

The half-life of uranium 238 is $T_{238} = 4.47 \cdot 10^9$ years. The products of its decay—even long-lived thorium and radium—decay more rapidly than uranium. So immediately after the first alpha decay the uranium nucleus is transformed into the lead isotope ^{206}Pb by quickly emitting seven more alpha particles and six electrons. The energy of all these particles generated by the nucleus of the decayed uranium is equal to $E_U = 7.1 \cdot 10^{-12}$ J.

Knowing this, we can easily determine that granite with a uranium content of approximately 10^{-6} by mass releases heat at the rate of $1.8 \cdot 10^{-11}$ W/kg, or $5 \cdot 10^{-8}$ W/m³. In addition to the uranium isotopes ^{235}U and ^{238}U , thorium (^{232}Th) and potassium (^{40}K) isotopes also release heat, but their contribution is significantly less.

If the concentration of uranium, thorium, and radioactive potassium in the granite of the continental land masses didn't change with depth, it would provide a significant portion of the total heat flux from the Earth's interior. This tempts us to ascribe the entire heat flux to the decay of uranium in low concentration (even lower than in granite) but present everywhere, all the way down to a depth of several thousand kilometers.

But this isn't the case. First, the heat flux at the bottom of the oceans is almost the same as on the continents, even though there is no granite under the oceanic floor. In the basalt under the sediment on the ocean floor, the uranium content is approximately 1/20 that in

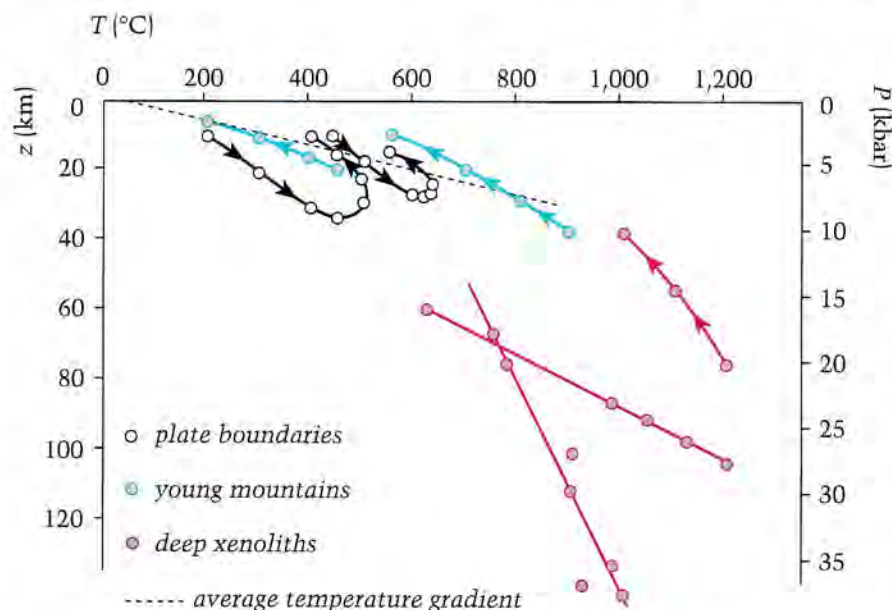


Figure 1

Increase in temperature with depth measured by the mineral changes of xenoliths. The second vertical scale is pressure. The arrows on the curves show how pressure and temperature changed with time. Most of xenoliths moved to the surface (otherwise they would not have been discovered), but xenoliths found near the boundaries of the continental plates recorded their submergence as well. The average temperature gradient of the Earth near its surface is shown by the dotted line. Submergence occurs when the temperature gradients are low and ascendance to the Earth's surface occurs when the temperature gradients are high.

the continental crust.

Besides, we can get an independent estimate of the total amount of uranium in the Earth's crust by looking at the helium content of the atmosphere. When each atom of ^{238}U decays to ^{206}Pb , eight alpha particles are emitted. These particles capture electrons to form eight helium atoms ^4He . Helium—a light gas that doesn't interact with anything—slowly works its way up from the depths along cracks in the crust and is ejected by volcanic eruptions. It doesn't accumulate in the Earth's atmosphere, however—it ascends higher and higher and finally escapes into outer space. Every second the Earth loses 20–30 g of helium. That's the same rate at which helium is formed in the Earth's interior. So the helium flux turns out to be related to the total mass of all the uranium in the Earth. This mass is close to $m_{\text{U}} = 4.5 \cdot 10^{16}$ kg.

The decay of this uranium provides thermal power in the Earth's interior equal to $7 \cdot 10^7$ W. This value is approximately 1/60 the heat flux from the depths. Over the entire history of the planet, all the radioactive elements in its interior have released energy of not more than 10^{30} J. This may seem like a lot of energy, but it's not enough to explain the heat in the Earth's interior.

The main reason for the failure of the radioactive explanation of the Earth's heat is not the divergence of estimates, but the existence of another, more powerful source of heat in the Earth's depths: gravity.

The Earth's history

All the planets were formed by the collisions of smaller celestial bodies. When two bodies of greatly differing mass collide (for instance, a small asteroid and a planet), the released energy is transformed into heat that is mainly distributed near the impact point on the surface of the larger body. This energy is quickly removed by thermal radiation into the atmosphere or into outer space. When a bigger asteroid hits, the crust is compacted to a greater depth, but again, the greater

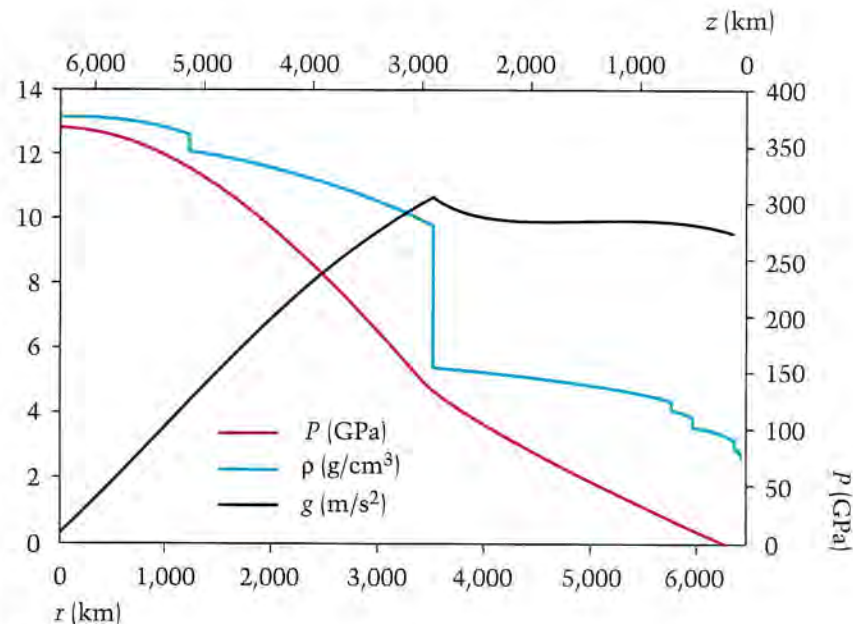


Figure 2

Dependences of density $\rho(r)$, gravitational acceleration $g(r)$, and pressure $P(r)$ on radius, calculated from velocities of seismic waves.

part of the released thermal energy is radiated into outer space. Is it possible to make a cold planet this way? No. From seismic observations (recording sound waves from distant earthquakes) we know the internal structure of our planet (fig. 2). As you can see, density in the Earth's core reaches 13 g/cm^3 . This doesn't mean that the Earth's core consists of lead—it's iron, but it's compressed to such a density by pressure from the upper layers that exceeds $3 \cdot 10^{11}$ Pa.

A hot interior leads to compression for any planet in the process of accumulating mass from smaller bodies. However, the consensus of scientific opinion today is that our planet has a special history. The only way to explain the Moon's present orbit is to suppose that our planet and its satellite were created as a result of a giant collision of two protoplanets about the sizes of Venus and Mars. Read the article by A. G. W. Cameron in this issue and look through the set of pictures of such a collision—they were obtained by means of huge computer calculations.

Another argument supports this idea as well. Look at figure 3. The densities of all the planets in the solar system (except the Earth) fit a smooth curve, whereas the Earth's

density is significantly higher than it "should" be relative to its distance.

The above hypothesis is also attractive for explaining why the Earth and Venus are so different despite their similar masses. Because of a catastrophic collision, the Earth melted, whereas Venus and the other planets seem to have formed from multiple collisions of a much smaller scale. This means that the separation of material in the gravitational field of the Earth occurred more deeply. Also, the high temperatures released water from the interior, where it can exist in crystalline form. Thus, we have oceans on the Earth, but mere traces of water in Venus's atmosphere.

We need now to evaluate how rapidly a heated body cools when heat is transported by thermal conductivity. The cooling time depends on the body's size R [m], heat capacity c_p [J/(kg · K)], thermal conductivity κ [J/(m · s · K)], and density ρ [kg/m³] (the dimensions are given in brackets).² Only one combination that isolates the time dimension can be made out of these four values.

²For a refresher course in dimensional analysis, see "The Power of Dimensional Thinking" in the May/June 1992 issue of *Quantum*.—Ed.

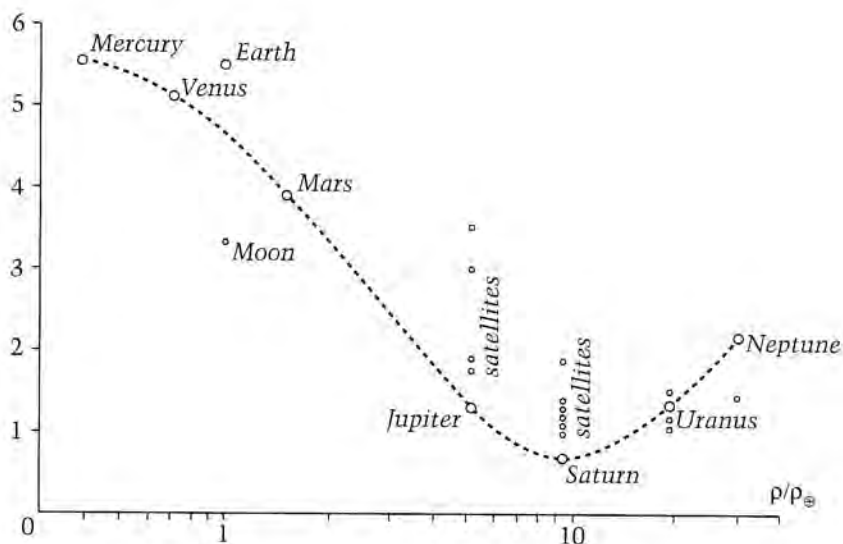


Figure 3

Average density of planets in the solar system plotted against their distance from the Sun. The interpolation shown by the dotted line is made for the planets other than the Earth and results in a nontrivial but smooth dependence. The Earth's density proves to be much higher than its position on the interpolation (4.7 g/m^3), and the Moon's density is much lower.

Thus, the typical time it takes a body to cool an order of magnitude is

$$\tau \equiv \frac{c_p \rho}{\kappa} R^2 [\text{s}].$$

One should certainly be careful making such an estimate for the Earth, whose composition changes with depth, but let's try:

$$\tau_{\oplus} \equiv \frac{R_{\oplus}^2 c_p \rho}{\kappa} \sim 3 \cdot 10^{19} \text{ s} \approx 10^{12} \text{ years}.$$

As it turns out, we get a result that makes no sense: the time far exceeds the age of the solar system (4.6 billion years) and is two orders of magnitude greater than the age of the universe (about 10 billion years). This paradox can't be corrected by making the thermal characteristics more precise. This means that either the Earth's interior is still very hot or our initial assumptions are wrong.

Let's assume the first. We'll estimate the depth h up to which the Earth has cooled during its existence—say for $t_{\oplus} = 10^{17} \text{ s} = 3 \text{ billion years}$. We make the estimate by reworking the above formula for the cooling time:

$$h \equiv \sqrt{\frac{c_p \rho t_{\oplus}}{\kappa}} = 3 \cdot 10^5 \text{ m} = 300 \text{ km}.$$

Three hundred kilometers constitutes a mere 4% of the Earth's radius. But it's a more realistic estimate. It makes sense for the cooling of ancient continental regions like Karelia (the Kola Peninsula). The oldest rocks on the surface there date back 3 billion years. For the Earth as a whole, however, it's physically wrong to suggest that heat was removed from the upper 300 km only and that the Earth's interior remained as hot as it was when the planet was formed. The cooling of the Earth's inner regions occurred differently—not by thermal conductivity but by convection. (And it continues, to a lesser degree, right to the present day.)

Due to thermal expansion, the density of rock depends on temperature: it increases as the rock cools. So the density of the upper, cooled layers of rock is greater in some locations than that of lower layers with a similar chemical composition. In the Earth's gravitational field this leads to instability—it appears to be energetically advantageous for cooled rock to descend and for hotter, lower layers to rise. In other words, convection develops. With convection the transfer of thermal energy by conductivity is supplemented by

heat transferred along with the moving medium.

This convection—the ascent of hot masses and descent of cold masses—causes horizontal motions on the surface of the Earth. Look again on figure 1—traces of some xenoliths show that they descend when their temperature gradient is low (lower than the average, shown by the dotted line) and come back to the Earth's surface when they happen to be in a region with a high temperature gradient. When studying this figure, remember that we can analyze only those rocks that succeed in returning to the surface, so near-circular traces are probably much more common than in actuality.

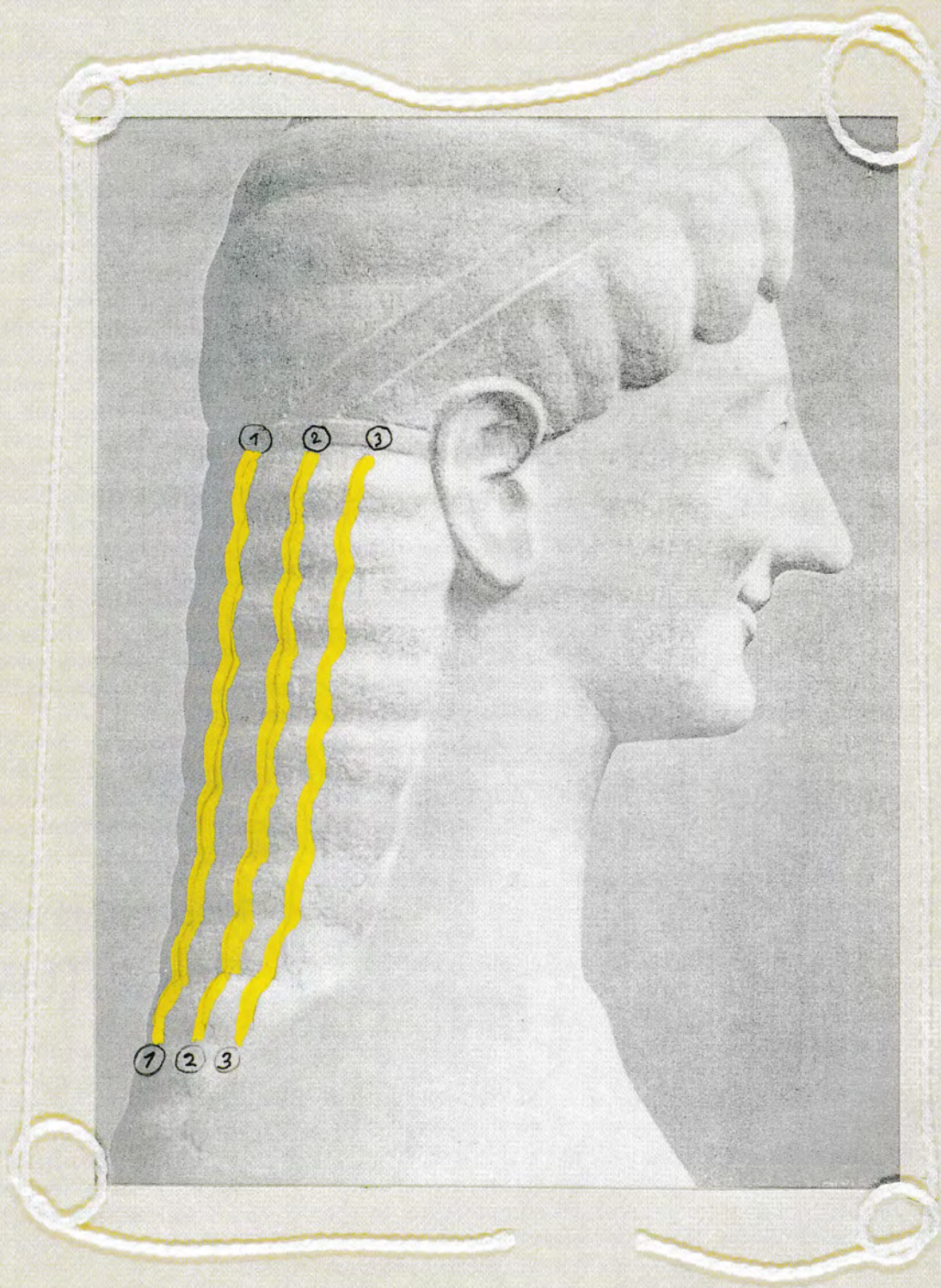
The Earth's crust really does move—slowly but surely. It's not very noticeable, and it was hard to believe it—the rates of displacement were so slow, and people had gotten used to the idea that they lived on firm, stationary ground. But we live on drifting continental plates, and below, in the depths of the Earth, lies a great store of thermal energy that makes the Earth's surface move and causes earthquakes and volcanic eruptions.

As interesting as these phenomena are, they fall outside the framework of this article. Let's come back to them some other time. ●

Yes, you can get back issues of QUANTUM!

Back issues of *Quantum*—from the January 1990 pilot issue on—are available for purchase (except September/October 1990, which is out of print). For more information and prices, call **1 800 SPRINGER (1 800 777-4643)**. Or send your order to:

Quantum Back Issues
Springer-Verlag New York, Inc.
PO Box 2485
Secaucus NJ 07094



Braids and knots

Homespun phenomena with profound implications

by Alexey Sosinsky

BRAIDING AND KNOTTING—is this really mathematics?

A diligent student usually leaves high school with the conviction that mathematics deals with abstract notions, whereas such mundane things as braids and knots have nothing to do with mathematics.

But that's a false impression. Nowadays mathematicians deal not only with lofty problems—number theory, the calculation of space flights, or the study of poetic meters—but also with earthly everyday matters like economics or queuing theory.

And braid theory, too. This real and living theory, which dates back to the 1920s, is not complete yet, and its applications haven't yet been exhausted. And as to its beauty, braid theory doesn't play second fiddle to classical mathematics, which actually stopped learning new tunes in the 16th and 17th centuries and is the only area of mathematics studied in most schools.

I'll begin my story with examples of braids (fig. 1). This is how you can imagine a braid: two rows of n nails are driven in the top and bottom edges of a vertical board (where n may be equal to 1, 2, 3, ...), and each of the top nails is connected to one of the bottom nails with a string. The strings are disjoint and always go down—that is, a string is not allowed to turn upward and head toward the top of the board. The things you see in figure 2 are not braids.

Two braids are considered equivalent (that is, the same) if one

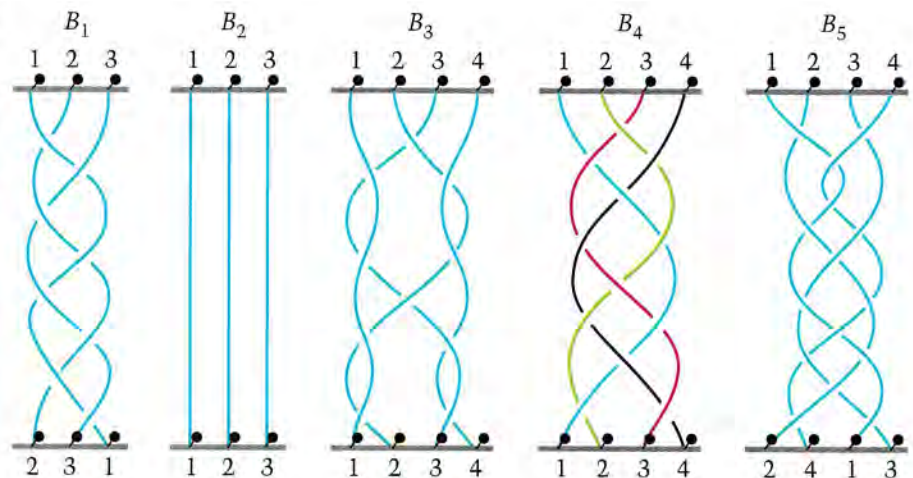


Figure 1

Examples of braids on three and four strands: B_1 —a “girl’s braid”; B_2 —a trivial braid; B_3 —a pure braid; B_4 —a cyclic braid; B_5 —a cyclic braid.

of them can be turned into an exact copy of the other by moving its strings—usually called strands—so that each of their points stays in the same horizontal plane. In so doing, a string can be stretched and shrunk,

but you can't break it or glue it. An example of such a transformation is shown in figure 3.

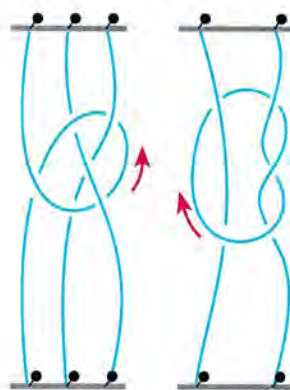


Figure 2

Nonbraids (their strands have ascending sections).

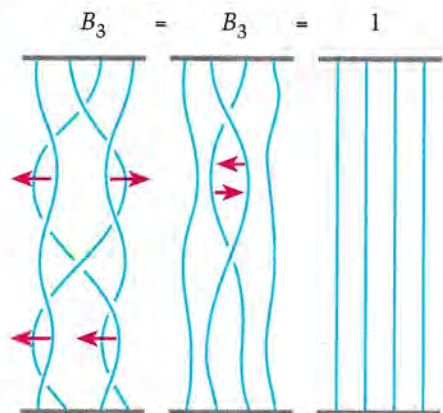


Figure 3

A geometric proof that B_3 is a trivial braid ($B_3 = 1$). By moving its strands “horizontally,” braid B_3 is transformed into a braid with four vertical strands.

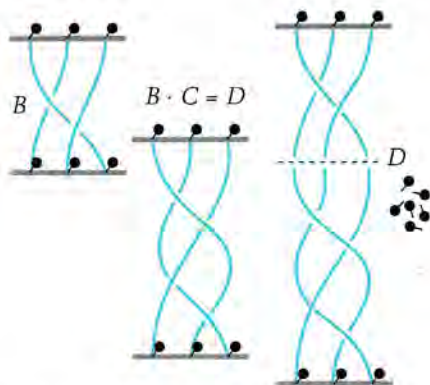


Figure 4

Composition of braids. The top of the second braid is brought into contact with the bottom of the first, and the corresponding strands are pasted together.

In figure 1 the top ends of the strands are supplied with numbers in the usual order—from left to right. At the bottom you see the numbers of the strands again—but here their order isn't necessarily the same. So every braid defines a certain permutation of its strand numbers. For instance, the permutations associated with the braids B_1 , B_4 , B_5 in figure 1 are, respectively,

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix}.$$

Our artist colored one of the braids in figure 1, B_4 . The property that sets it apart is that it defines the identity permutation

$$\begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ 1 & 2 & 3 & \cdots & n \end{pmatrix}$$

(for B_4 , $n = 4$). In other words, this braid preserves the order of its strand numbers. Any braid with this property is called pure. A trivial braid—all of whose strands are vertical lines—is a particular case of a pure braid.

By the way, there are two, not one, trivial braids in figure 1. Two? Yes, indeed, two: the braid B_3 is trivial because it can easily be transformed (see figure 3) into a braid with four vertical strands.

Another kind of braid that should be singled out besides pure braids is, in a certain sense, just the opposite. These are cyclic braids. By definition, they rearrange the strand num-

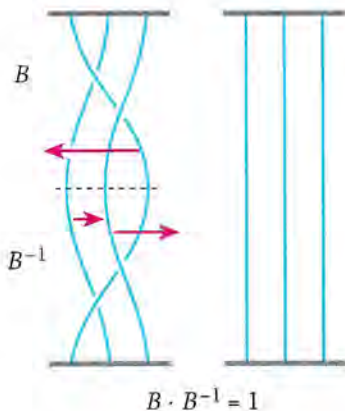


Figure 5

Inverse B^{-1} of a given braid B . It is obtained as the mirror image of B in the horizontal plane through the bottom base of B . Each double point of braid BB^{-1} has a mirror counterpart, and all these pairs of double points can be successively annihilated by straightening strands, moving from the center of the braid to its sides.

bers in a single cycle, as braid B_5 does: $1 \rightarrow 2 \rightarrow 4 \rightarrow 3 \rightarrow 1$.

A braid is one of the simplest geometric objects. Braids easily yield to algebraization by introducing an operation of composition (or multiplication) on those that have the same number of strands. This is quite simple (see figure 4): you bring two braids end to end, glue together the corresponding strands, and remove the now unnecessary nails (the bottom nails of the top braid and the top nails of the bottom braid). This operation is similar to the ordinary multiplication of numbers in several respects. It satisfies the associative property:

$$B_1(B_2B_3) = (B_1B_2)B_3.$$

There is an analog of unity—the *trivial braid* (B_2 in figure 1 for $n = 3$), denoted by 1, such that

$$1 \cdot B = B \cdot 1 = B$$

for any braid B . We can also find an analog for the operation of division: any braid B has an *inverse braid* B^{-1} such that

$$B^{-1} \cdot B = B \cdot B^{-1} = 1.$$

This is not obvious, and I challenge the reader to think of a construction for the braid inverse to a given one. If you fail to do it yourself, look at

the answer in figure 5.

However, the composition of braids is *not commutative*: BC may be not equal to CB (an example is given below).

The algebraic object thus obtained—called the *braid group on n strands*¹—is not very simple, but it has been thoroughly explored. We'll undertake our own investigation of its properties. To this end we'll make use of *elementary braids* S_1 , S_2 , ..., S_{n-1} on n strands (fig. 6).

It turns out that *any braid can be represented as the composition of elementary braids and their inverses*. For instance, it's clear that

$$B_1 = S_1S_2^{-1}S_1S_2^{-1}S_1S_2^{-1}S_1S_2^{-1}.$$

Further,

$$B_3 = S_2S_1S_3^{-1}S_1^{-1}S_3S_2^{-1}S_1S_3S_1^{-1}S_3^{-1}.$$

This becomes obvious after we properly nudge the strands of B_3 so as to move the four double points on the right down slightly (fig. 7).

Exercise 1. Represent the braids B_4 and B_5 in figure 1 as compositions of the elementary braids S_1 , S_2 , S_3 and their inverses.

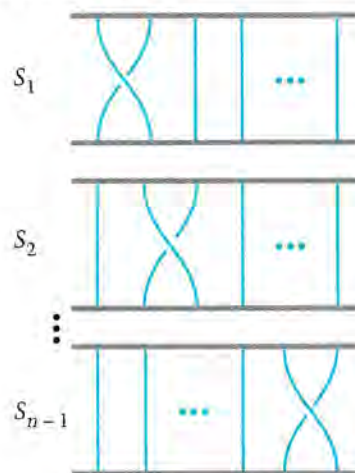


Figure 6

Elementary braids. The i th elementary braid S_i ($i = 1, 2, \dots, n-1$) consists of $n-2$ vertical strands—all except the i th and $(i+1)$ st, which form one crossing with the i th strand above the $(i+1)$ st.

¹You can get acquainted with the general notion of a group, diverse examples of it, and applications in the November/December 1991 issue of *Quantum*.

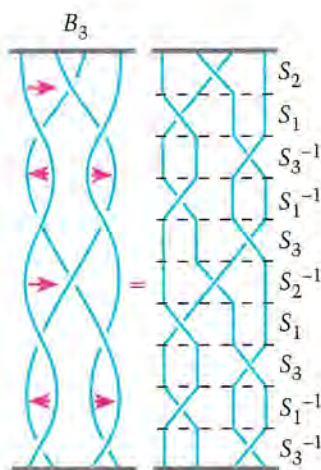


Figure 7

Representation of braid B_3 in terms of the elementary braids S_1, S_2, S_3 and their inverses. By moving the strands slightly, all the double points are moved to different levels; then an algebraic formula for B_3 can simply be read right from the picture: $B_3 = S_2 S_1 S_3^{-1} S_1^{-1} S_3 S_2^{-1} S_1 S_3 S_1^{-1} S_3^{-1}$.

In braid theory, just as in analytical geometry, algebraic notation allows us to replace geometric considerations with absolutely mechanical calculations based on the following identities.

1. Trivial relations:

$$\begin{aligned} S_i S_i^{-1} &= S_i^{-1} S_i = 1, \\ S_i \cdot 1 &= 1 \cdot S_i = S_i \\ (i=1, 2, \dots, n-1). \end{aligned}$$

2. Remote commutativity:

$$S_i S_j = S_j S_i \text{ whenever } |i-j| \geq 2 \\ (i, j=1, 2, \dots, n-1).$$

3. Braid relations:

$$S_i S_{i+1} S_i = S_{i+1} S_i S_{i+1} \\ (i=1, 2, \dots, n-2).$$

These relations are almost self-evident (fig. 8). And they can be used to derive other relations—for instance, the following more general form of remote commutativity.

Exercise 2. Prove that $S_i^u S_j^v = S_j^v S_i^u$ for any $u, v \in \{1, -1\}$, $|i-j| \geq 2$.

Let's examine one more example: a proof of the relation $B_3 = 1$ (established geometrically in figure 3) by direct calculation. We have

$$\begin{aligned} B_3 &= S_2(S_1 S_3^{-1}) S_1^{-1} S_3 S_2^{-1} S_1(S_3 S_1^{-1}) S_3^{-1} \\ &= S_2 S_3^{-1} (S_1 S_1^{-1}) S_3 S_2^{-1} (S_1 S_1^{-1}) (S_3 S_3^{-1}) \end{aligned}$$

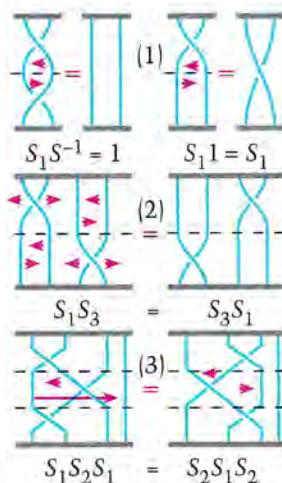


Figure 8

Proofs of the basic relations of braid theory. (1) trivial relations (for braids on two strands); (2) remote commutativity (for braids on four strands); (3) braid relations (for four strands). Proofs for greater numbers of strands are similar.

$$\begin{aligned} &= S_2 S_3^{-1} \cdot 1 \cdot S_3 S_2^{-1} \cdot 1 \cdot 1 \\ &= S_2 (S_3^{-1} S_3) S_2^{-1} \\ &= S_2 S_2^{-1} \\ &= 1. \end{aligned}$$

Here we first used remote commutativity, after which everything cancelled out "by itself" (thanks to trivial relations).

Exercises

3. Prove the identities

$$\begin{aligned} S_1^{-1} S_2^{-1} S_1^{-1} &= S_2^{-1} S_1^{-1} S_2^{-1}, \\ S_1 S_2 S_1 S_2^{-1} S_1^{-1} &= S_3 S_1 S_3^{-1} S_1^{-1} S_2. \end{aligned}$$

4. Prove that $S_1 S_2 \neq S_2 S_1$ for $n \geq 3$.

Why do our relations work so efficiently in solving exercise 3 (as well as in the preceding calculations)? Is it because I happened to pick the right problems, or because there's a certain regularity behind this efficiency? In other words, are relations 1-3 sufficient to prove *all* equalities in braid theory?

It turns out the answer is yes. The creator of braid theory, the German mathematician Emil Artin, proved in 1936 that *any equality in braid theory follows from relations 1-3*. This remarkable theorem allows the fundamental problem of braid theory—the classification

problem—to be solved. That is, *it is possible to give an (infinite) list of braids (without repetitions) and an algorithm that assigns to any braid its number in this list*.

Proofs of these facts are not elementary, so I won't go into them here. I only want to point out that they serve to transform the geometric theory of braids into a calculational science in which any concrete question can be answered, in principle, by a computer.

I can almost hear the sceptical reader say, "So what? And why do we need to solve these 'concrete problems'?"

Well, the point is, braid theory has a lot of applications in mathematics and in other fields. Here I'll expand on only one application, one that I especially value: the application to knot theory.

Let's begin with a number of examples of knots (fig. 9 on the next page). A *knot* is a closed curve in space, smooth or polygonal, that can be arbitrarily twisted and interweaved. It's helpful to imagine that the knot is made in a thin, flexible, stretchable string. Two knots are considered to be the same (equivalent) if one of them can be transformed into an exact copy of the other by moving, bending, stretching, and shrinking the string without tearing it apart. One important kind of knot (actually, not a *genuine* knot at all) is a *trivial knot*—an ordinary, unknotted circle (K_0 in figure 9). In fact, there are two portraits of the trivial knot in this figure: the knot K_7 is trivial as well—it can easily be untangled and turned into a big circle (do it in your mind or using a pencil and eraser!). Not only that, there's yet another pair of equivalent knots in figure 9.

Exercise 5. Find two nontrivial equivalent knots in figure 9.

For mathematicians it's more convenient to have an exact definition for the equivalence of knots rather than a graphic description like the one above. I'll give such a definition for knots that are polygonal paths rather than smoothly bending curves. (This will allow me to avoid some techni-

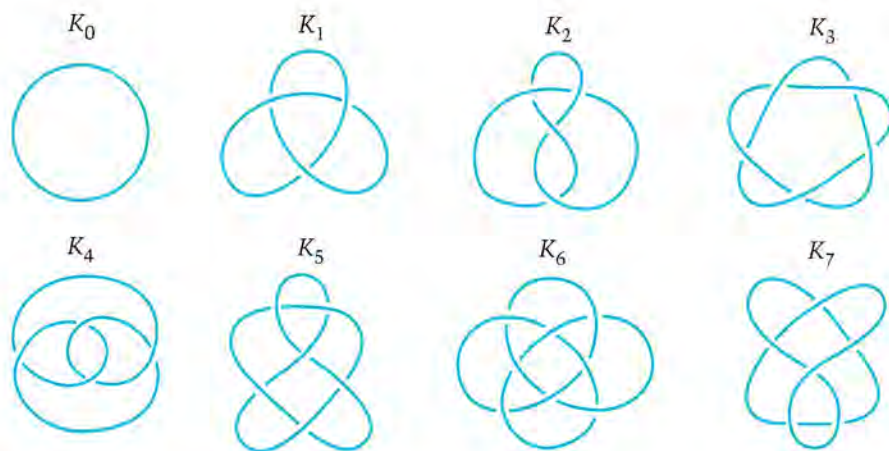


Figure 9

Examples of knots. K_0 —trivial knot; K_1 —trefoil; K_2 —figure-eight knot; K_3 —knot S_1 ; K_4 — S_2 ; K_5 —untitled; K_6 —“turban”; K_7 — S_3 .

cal details without a real loss of generality.) We'll define an *elementary operation* as the replacement of a segment AB of a polygonal knot—for instance, the segment AB in figure 10—with two segments AC and CB , or the reverse transition from ACB to AB , performed under the condition that the knot doesn't contain interior points of triangle ABC . Two knots are *equivalent* if there exists a finite sequence of elementary operations that turns one of them into another (fig. 10). Inspecting this figure, you'll understand without any difficulty that this definition is really adequate to the graphic description above.

Just as with braids, the classification problem can be posed for knots:

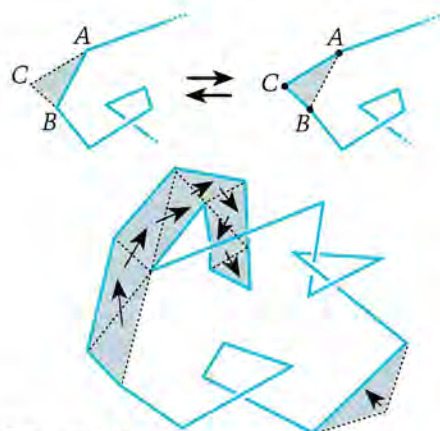


Figure 10

Elementary operations on a knot. A segment AB is replaced by a two-segment path ACB provided that the knot does not intersect the triangle ABC . A sequence of such operations allows the knot to be deformed.

one has to make an (infinite) list of knots (without repetitions) and an algorithm assigning to any knot its number in the list. Although this problem has been solved in principle by now, the solution is so cumbersome that it can't be used in practice. Is it possible to reduce this problem to the already solved classification problem for braids? The following idea suggests itself.

Take a braid, bend it, and glue its ends together (fig. 11). We get a knot. But does such closure of a braid always produce a knot?

Exercises

6. Draw the closures of braids B_1 and B_4 (fig. 1). How many curves do

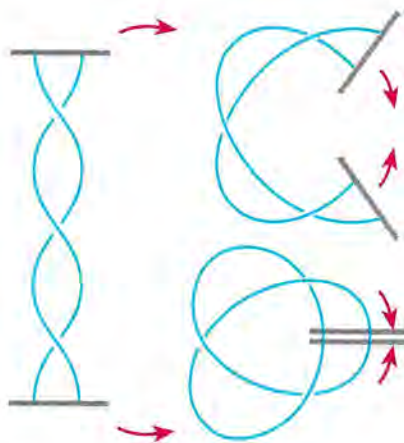


Figure 11

Closure of a braid. Joining the ends of a very simple braid, we get a knot (trefoil).

we obtain in either case? Describe the braids whose closure is one curve—that is, a knot.

7. Find a braid whose closure is the knot (a) K_1 , (b) K_3 , (c) K_4 , (d) K_2 (see figure 9).

So a certain class of braids (namely, cyclic braids, as readers who have done exercise 6 must have understood) yields knots after closure. But is it possible to obtain all knots this way? Yes, as it turns out!

The remarkable American mathematician James W. Alexander, one of the first investigators of knots, proved in 1925 that *any knot is the closure of a certain braid*. Rather than prove it in detail, I'll show the two main techniques used in the proof.

1. *Unraveling*. Draw a knot, choose a direction on it, and take an arbitrary point O not on it (fig. 12). Call a segment of the knot *positive* (with respect to O) if the direction on it is viewed from O as from left to right; for instance, in figure 13a all the segments of the knot except AB and FG are positive. A knot will be called *positive* (with respect to O) if all its segments are positive. For a positive knot, a braid that produces it under closure is easy to find—simply cut the knot at any place and unbend it as shown in figure 12.

2. *Alexander's trick*. Negative segments of a knot (fig. 13b, 13c) are

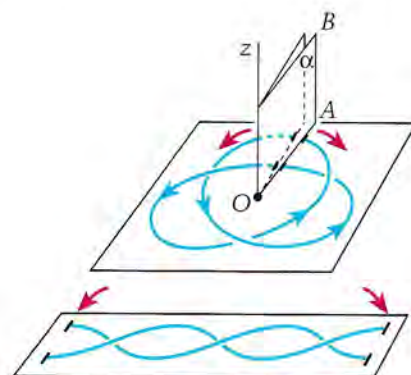


Figure 12

Unbending a positive knot into a braid. The vertical half-plane α duplicates itself, then opens like a book, and the knot unbends into a braid. This operation reverses the closure of a braid shown in figure 11.

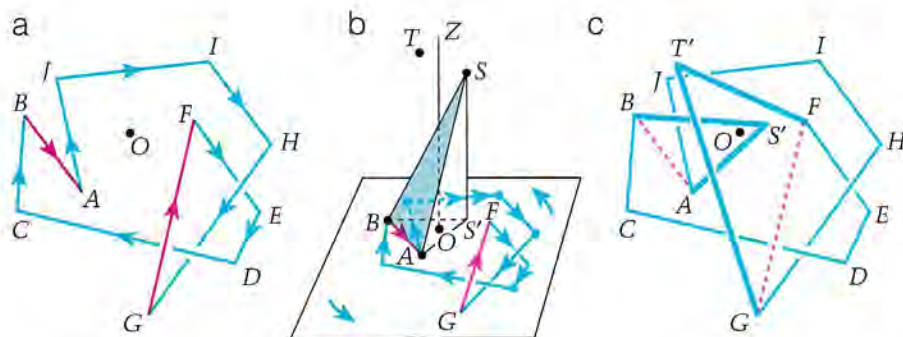


Figure 13

Making a knot positive. (a) The diagram of a knot with two negative edges with respect to point O — AB and FG ; (b) Alexander's trick—replacing a negative edge AB with the polygonal path ASB —reduces the number of negative edges by one; (c) Alexander's trick applied twice turns the knot $ABCDEFGHIJA$ into a knot that is positive with respect to point O .

replaced with pairs of positive segments surrounding point O . After eliminating all negative segments, the first technique is applied.²

And that's how any knot can be proved to be a closure of a certain braid. But braids can be classified. Can we use this to classify knots? Alas, no. The trouble is, the closure of different braids doesn't always give different knots. For example, the three-strand braid in figure 14 differs from the two-strand braid in figure 11, but its closure is a trefoil as well (check it!).

Thus, an attempt to reduce the classification of knots directly to the classification of braids fails. But Alexander's theorem is only the first step in the elaborate train of thought that links together the most beautiful inhabitants of three-dimensional space: braids and knots.

In conclusion I'll try to answer the question "Why do we need all that?" for the sake of readers who don't think that beauty *per se* can be a sufficient reason for studying a subject. To some extent the answer is found in the very history of the creation of braid and knot theories.

Braid theory was invented in the 1920s by the young German algebraist E. Artin at the request of a

²A captious reader may argue that a knot can contain segments that are neither positive nor negative. This sort of nuisance can be avoided by moving the "bad" segments slightly.

fabric mill. He was, as we would now say, a consultant.

The sources of knot theory go back further in time, and the interesting circumstances of its creation are almost forgotten. The systematic study of knot theory was initiated by the great British mathematician and physicist William Thomson, 1st Baron Kelvin. He came to the conclusion that electromagnetic interaction is carried by waves, and later he was struck by an even more daring thought: the interacting particles themselves are waves, too, but since the particles (atoms) are very small and the waves long, the atom-waves must close in on themselves within a small space. So they form little knots that capture all the physical and chemical information about the atom coded in the very same way a knot is tied. Thomson and his students set about to explore knots, beginning their systematic classification by listing them in special tables.

The relay baton in the investigation of knots was picked up by mathematicians in the 20th century, and not for any hope of monetary gain—they were attracted by the sheer beauty of the subject. The subtle invariants they created (we plan to devote a special article to them in an upcoming issue) allowed for a significant advance in knot theory. However, for a long time this field remained a peaceful backwater in mathematics, known mainly to specialists in topology.



Figure 14

Closure of this braid also produces a "trefoil" knot (compare with figure 11). Check this!

In the meantime, braid theory found quite serious applications—for instance, in complex analysis, mechanics, and the physics of elementary particles.

Not that long ago, after work by the English mathematicians John Conway and V. Jones, the Russians V. Turayev and A. Reshetikhin, and the American L. Kauffman, unexpected and deep connections between braid and knot theory, abstract algebra, and physics have been revealed. The peaceful backwater seethed. Physics again! Not only its classical branches (statistical physics, for instance, a model of . . . ice!) happened to be involved, but modern quantum theory as well. And the idea of coding chemical information in small knots (and braids!) reappeared in molecular biology in the course of deciphering amino acids and studying DNA. So—who knows—perhaps there's something to Lord Kelvin's old idea . . .

Does your library have QUANTUM?

If not, talk to your librarian! *Quantum* is a resource that belongs in every high school and college library.

"A first-class 'new' magazine. . . one can appreciate the meaning of quality and imaginative challenge . . . it is for anyone with an interest in science, particularly math and physics. Highly recommended."—*Library Journal*

"It should be in every high school library [and] in most public libraries . . . we owe it to our students to make *Quantum* widely available."—**Richard Askey, Professor of Mathematics at the University of Wisconsin, Madison**

ORDERING INFORMATION
ON PAGE 3

When a body meets a body

The Giant Impact theory for the formation of the Moon

by A. G. W. Cameron

THE MOON HAS BEEN AN object of wonder, and sometimes of worship, throughout the history of the human race, recorded or not. But it is only during the last century that scientists have tried to formulate theories to explain how the Moon was formed. The early theorists considered the origin of the Moon to be a problem quite isolated from the general questions about the origin of the solar system itself, and it is only in recent years that the formation of the Moon has been considered as an element of the broader process of the accumulation of the planets.

The early theories of the formation of the Moon can be divided into three general types. The *fission* theory postulated that the Earth spun about its rotation axis so rapidly that it deformed into the shape of a pear, and that the material in the smaller end then separated from the rest of the planet and went into orbit about it. Such a system has far more angular momentum than is in the entire Earth-Moon system today, and nobody succeeded in finding a mechanism to get rid of the excess angular momentum.

The *companion* theory postulated that the Earth and the Moon were formed together in mutual orbit. But

the Moon has a composition quite dissimilar to the Earth, having very little metallic iron in its composition, and nobody could figure out an accumulation mechanism that would put most of the metallic iron into just one of the two orbiting bodies.

The *capture* theory assumed that the Moon had been formed elsewhere in the solar system, that it had wandered close to the Earth, and that the Earth then captured it. But, again, nobody could understand how a planetary body could be formed anywhere else that would be so unlike the other terrestrial planets as to lack a significant iron core, and those who tried to understand the celestial mechanics of the capture experienced grave difficulties.

This was the situation at the beginning of the 1960s when the United States government created the Apollo program, whose purpose was to land men on the surface of the Moon and return them safely to Earth. There were several different motivations for this decision, but the scientific motivation was to determine which of the above three theories was the correct one.

During the Apollo program there was a psychiatrist, Ian Mitroff, at the University of Pittsburgh who repeatedly interviewed the "lunar scien-

tists" to determine their then current views on the origin of the Moon, hoping that he would be a witness to a perfect demonstration of the scientific method: data would be obtained from analysis of lunar samples and from instruments mounted on the lunar surface, and then the scientists, being logical people, would change their views and converge on the correct theory.

But it was not to be. As the data were obtained, Mitroff noticed that the views were not changing. There was no convergence. Mitroff wrote a book in which he concluded that the lunar scientists were very obstinate people and incapable of responding to scientific evidence. But he did not tell us which of the theories was correct.

Mitroff's mistake was in assuming that the correct theory of the formation of the Moon had to be one of the three classical theories. A proper formulation of the situation would have added a fourth category: "none of the above." All the objections to the three theories that had been raised before Apollo remained objections, and because of the large amount of new data, some additional objections were added. The lunar scientists were very frustrated that the

Art by Sergey Ivanov



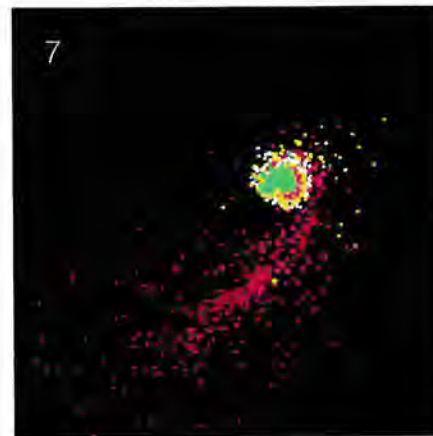
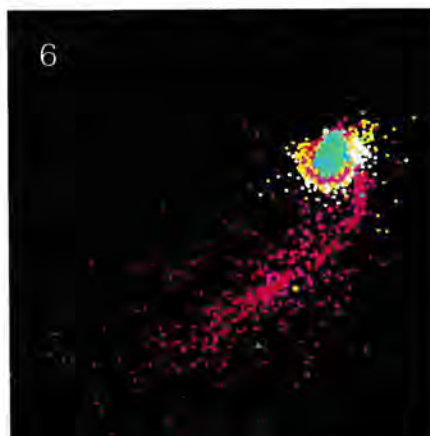
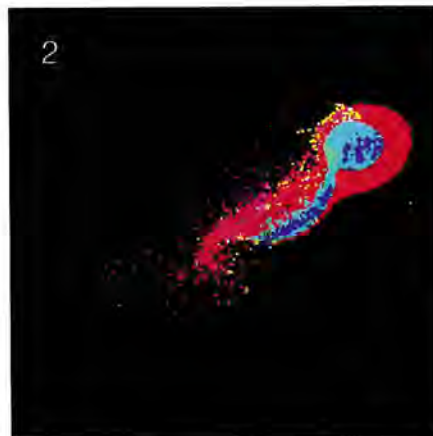
origin of the Moon remained a mystery.

During the next few years the stream of data from the instruments on the Moon ceased, and the rate of analysis of lunar samples continually declined. In 1984 a conference on the origin of the Moon was held at Kona, on the island of Hawaii. The organizers of the meeting found that a surprising number of the abstracts submitted dealt with some aspect of the idea that the Moon was formed as a result of a collision between the Earth and another planetary body at least as large as the planet Mars. This was the "Giant Impact" theory.

The Giant Impact

As is so often true when a new theory is developed, the basic ideas involved in the Giant Impact theory were either incorrectly or inadequately expressed at first. In the middle of the 1970s two ideas were independently developed that were precursors of the Giant Impact theory. At the Planetary Science Institute in Tucson, Arizona, William Hartmann and Donald Davis were investigating the theory of planetary accumulation, and they found that as the planetary bodies grew in size, a broad distribution of body sizes developed, with just one body having the largest size and with increasingly larger numbers of bodies of continually decreasing sizes. When the largest body became comparable in mass to the Earth, they noticed that collisions with it would vaporize the colliding body. So they suggested that a major collision with the Earth could produce an enormous amount of rock vapor that would be thrown high above the site of the collision, and that possibly much of this vapor could go into orbit and condense into the Moon. But they did not provide an explicit method for putting the angular momentum of the Earth-Moon system into this scenario, because when the Earth and the Moon are close together in mutual orbit, most of the Earth-Moon angular momentum must be in the rotation of the Earth.

Meanwhile, in Cambridge, Massa-

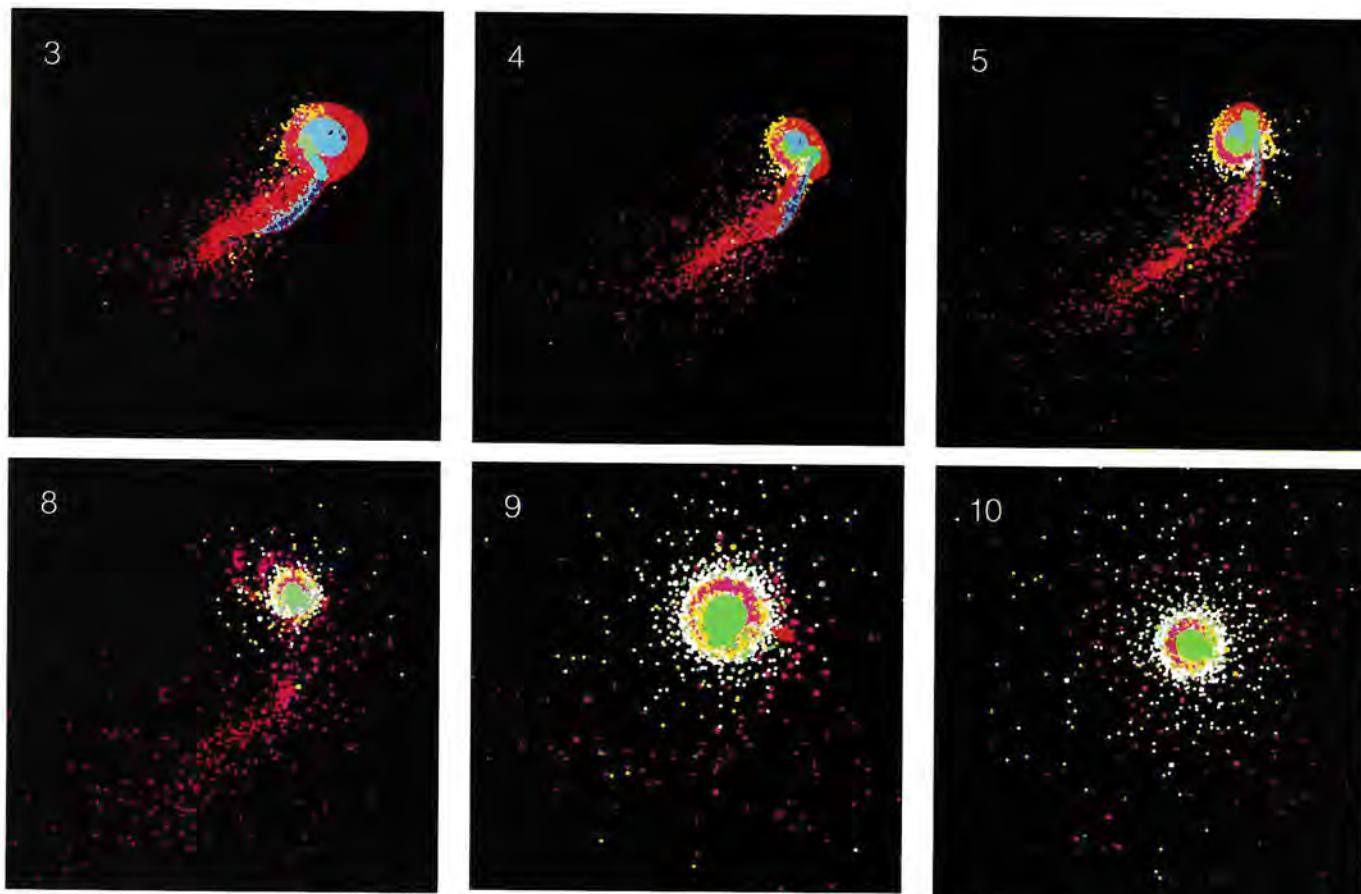


chusetts, William Ward and I were thinking about how to put the angular momentum of the Earth-Moon system into the spin of the Earth. We assumed that a massive body (the impactor) struck the Earth tangent to the equator with a velocity equal to the escape velocity. The mass of the impactor needed to be about the mass of Mars (this is a minimum mass, because if the impact point is closer to the center of the Earth, more mass is required).

It was also apparent to us that a huge cloud of vapor would be created in the collision. The cloud would be formed of a mixture of impactor and Earth materials, and so its center of mass would move at less than escape velocity and probably at less than orbital velocity. We presumed that pressure gradients would cause the cloud to expand, and we assumed that a part of the cloud would be accelerated to circular velocity in this way, thus allowing the Moon to condense and accumulate in orbit (the mass of the Moon is about one eighth of the mass of Mars).

This scheme automatically accounted for the angular momentum of the Earth-Moon system in the spin of the Earth and the motion of the cloud. But we learned later that gas pressure gradients are much less effective than we assumed and that gravitational torques do play a major role, and we had neglected them.

These ideas seemed to make little impression on the workers in planetary science at the time, but this was deceptive, because, as we have seen, when the Kona conference was being organized, many abstracts were submitted dealing in one way or another with the idea of a Giant Impact. But Kona represented a triumph of ideas over substance, because very little of a quantitative nature concerning the Giant Impact was actually presented. People were so delighted to have an alternative to the three classical theories that they were not too critical of the details. I left the meeting with the conviction that the next step must be numerical simulation of the Giant Impact using supercomputers.



Smooth particle hydrodynamics

A simulation of the Giant Impact (or any other major explosion) requires a computer code that treats a problem in hydrodynamics—the motion of fluids subject to external forces such as gravity and pressure gradients. The traditional means for organizing such a calculation is to divide space into a discrete set of compartments (sometimes called a mesh) and to follow the motion of the fluid through the mesh as it is acted upon by the external forces. The pressures are so large and the energy released is so high in a Giant Impact that all material flows like a fluid. Any material that flows outside the mesh is lost.

However, in recent years an alternative scheme for doing numerical hydrodynamics has been developed. This is called smooth (or smoothed) particle hydrodynamics. In this scheme the material is divided into spherical particles that overlap one another. The density distribution in a particle resembles the shape of a

Stages of a computer run modeling a collision of two protoplanets leading to the formation of the present Earth and Moon. The computer run has a target-to-impactor mass ratio of 8 to 2 and a total mass slightly exceeding the present total mass of the Earth and Moon. The initial angular momentum is equal to 1.433 of the present angular momentum in the Earth-Moon system ($3.5 \cdot 10^{34} \text{ kg m}^2/\text{s}$). The collision started with zero velocity at infinity, so that the impact occurred at escape speed. Excesses of mass and momentum are removed by small fragments escaping the system.

Each body was represented by 5,000 interacting smooth particles. Due to the mass difference, the particles of the target are about two times larger than those of the impactor, so you can follow their history. Protoplanets have iron cores and rock envelopes with a mass ratio close to that given by natural abundances of inherent elements. The initial surface temperatures of both colliding bodies were taken to be 2,000 K, which is high enough realistically to represent a history of collisional accumulation and low enough to suppress thermal evaporation of matter from the surface. After the collision, wherever very hot rock appears in a surface region, rock vapor evaporates and forms a hydrostatic atmosphere around the body.

A restricted set of colors is used in the plots: for rock, the lowest internal energy (or temperature) is plotted in dark red, and with increasing internal energy the colors change to light red or pink, yellow, and white. The rock points are plotted first, and then the iron points are plotted so that they will be superposed on the rock points. For iron, the lowest internal energy is dark blue, and with increasing internal energy the colors change to light blue, dark green, and light green. One general phenomenon to notice is that where the collision leaves high internal energy rock at the surface, white particles appear and move away from the area. These are the rock vapor particles that are evaporated and form the extended atmosphere around the system.

The whole time interval from the first to the last picture is about 2.5 days. (These are current Earth days—the Earth's rotation was actually much faster at this moment of creation.) Immediately after this collision the average distance from the Earth is just a few Earth radii; then the Moon gradually withdrew to its present distance, and simultaneously the Earth's rotation slowed.

bell—nearly flat near the center, then falling rapidly at a finite distance from the center. As the density of the material changes, the degree of overlap of the particles correspondingly changes, but the collection of particles maintains a nearly flat density to a good approximation. In this scheme the particles themselves move in response to the external forces.

Results

There is a characteristic outcome to all of the new simulations. In the collision, the Impactor (that is, the body with the lower mass) becomes distorted and elongated; the bulk of it falls onto the Protoearth, including essentially all the central iron, which plunges right through the rock mantle of the Protoearth. The rock part of the Impactor farthest away from the point of impact has a general tendency to go into orbit about the Protoearth as individual particles (but sometimes as clumps of particles).

However, the fallout of the iron core of the Impactor initially accumulates in a relatively small volume on one side of the Protoearth core, forming a rotating iron bar inside the Protoearth. This can exert a powerful torque on clumps of particles outside the Protoearth, in particular that part of the Impactor that would go into orbit.

The internal energy of the particles is indicated through the use of four different colors each for rock and iron. I have found this to be much more useful than using many colors in a pseudo-continuous distribution. The iron is plotted on top of the rock in order to see its behavior. White is used for the highest internal energies for rock particles. It was very striking to see, wherever a collision had heated the surface material of a planetary body, that a cloud of white particles arose above the surface and spread out to surround the planetary body. Even greatly elongated configurations of the Impactor were surrounded by a cloud of white particles after the collision.

Because the Impactor particles

were plotted with smaller radii, it was possible to see that most of the particles (and a majority of the mass) in the white cloud were originally from the Impactor. Evidently when the planetary surface was heated by Impactor rock falling on it, the Impactor particles would tend to be on top of the Protoearth particles and thus they would be the first to evaporate. The white cloud was always densest next to the planetary surface and thinned out away from the surface, as would be expected for an atmosphere. This phenomenon was a major departure from that observed in the previous runs, because particle evaporation and the formation of an atmosphere could not be properly simulated.

The final states of the Protoearth following all of the collisions were remarkably similar. For this reason it is sufficient to show the results of just one case, and I have chosen to show those for the a ratio of the mass of the Protoearth to the Impactor of 8:2.

Discussion

Thus we arrive at a new and quite simple picture of the consequences of a Giant Impact. Wherever the surface of the Protoearth is hit hard by the impact, a very hot magma is produced. From this hot surface, rock evaporates and forms a hot extended atmosphere around the Protoearth. The mean temperature is in excess of 4,000 K out to about eight Earth radii and is in excess of 2,000 K out to about twenty Earth radii. The above description applies to just about any Giant Impact involving an Impactor with at least ten percent of an Earth mass.

A candidate Moon-forming Giant Impact must also possess at least the present value of the Earth-Moon angular momentum, which places constraints on the Impactor. The Impactor needs to have at least 14 percent of an Earth mass in order for the Earth to swallow up the Impactor iron core and avoid getting too much iron in the Moon. But apart from this constraint, it appears from the present simulations that any division of mass

between the Protoearth and the Impactor can produce a promising set of conditions.

But how would the present set of scenarios evolve as time elapses? The computer code in these models does not allow for radiation to be produced and transported. The rationale for omitting it is that the time involved in the simulations is only some hours or a day or two. In reality the scenarios would lead to cooling of the rock vapor atmosphere and precipitation of refractory materials at fairly large distances in the atmosphere.

The only scenario that might work under these conditions would be one in which most of the material in the Moon comes from that part of the Impactor that has been torqued into high Earth orbit. A number of geochemists have expressed dismay at the thought of having to accept such a scenario.

The formation of the Moon as a postcollision consequence of a Giant Impact remains a hypothesis. Our work appears to have characterized the internal effects of a Giant Impact on the Protoearth. It appears to me that understanding the evolution of the external environment is the area where the more interesting challenges lie ahead. 

Grab that chain of thought!

Did an article in this issue of *Quantum* make you think of a related topic? Write down your thoughts. Then write to us for our editorial guidelines. Scientists and teachers in any country are invited to submit material, but it must be written in colloquial English and at a level appropriate for *Quantum's* target readership of high school and college students.

Send your inquiries to:

Managing Editor
Quantum
1840 Wilson Boulevard
Arlington VA 22201-3000

Just for the fun of it!

B131

Letters in digits. After replacing all the letters in a certain word with their numbers in the alphabet, the number 2122122112122 emerged. What was the original word? (A. Savin)



B132

Digits in letters. Solve this number rebus showing the long multiplication of a four-digit number by itself. (L. Mochalov)



B133

It's known from the reports of American astronauts that shadows on the Moon are much darker than on the Earth. Why might that be? (S. Krotov)



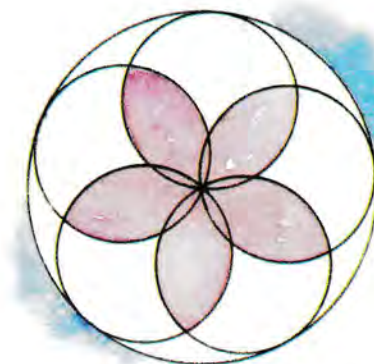
B134

Coloring a grid. A grid consisting of 6×2 squares has 21 nodes. Is it possible to paint the nodes in two colors so that no four nodes of the same color form a rectangle?



B135

Rosette and circumference. Several circles of the same radius are drawn through the center of a circle twice as big, forming a "rosette" (colored pink in the figure). Which of the two is greater—the perimeter of the rosette formed by the smaller circles or the circumference of the big circle? (V. Proizvolov)



ANSWERS, HINTS & SOLUTIONS ON PAGE 57

Art by Pavel Chernusky

Do something different this summer...



Live and learn in

RUSSIA

A unique opportunity awaits high school students and teachers: the Russian-American Mathematics and Science Summer Institute will be held from July 2 to July 22, 1995, at Moscow State University. This exciting program will feature

Two weeks of advanced classes in mathematics, physics, and biology taught in English by prominent Russian professors

Trips to the scientific laboratories of the Russian Academy of Sciences

Cultural and recreational activities

Following the two-week academic program in Moscow will be a one-week cultural program in St. Petersburg, the capital of Russia under the tsars.

Scholarships are available

For more information, please fill out the coupon below and send it to Dr. Edward Lozansky, President, American University in Moscow, 1800 Connecticut Ave. NW, Washington DC 20009, Phone: 202 986-6010, Fax: 202 667-4244, E-mail: lozansky@aol.com

Please send me _____ brochures to distribute among interested high school teachers and students.

Last name _____ First name _____

Address _____ City _____

State _____ Zip code _____ Phone number _____

I am a teacher _____ a student _____.

Challenges in physics and math

Math

M131

Intersection of parabolas. Two distinct quadratic polynomials $f(x)$ and $g(x)$ with a unit leading coefficient satisfy the equality $f(1) + f(10) + f(100) = g(1) + g(10) + g(100)$. Find all the solutions of the equation $f(x) = g(x)$. (A. Perlin)

M132

Winning score. Eight hockey teams compete for inclusion in the final four. (Every pair of teams meets once; a win scores two points, a draw one, a loss zero.) What minimal score ensures passage into the final four? (S. Khodjiyev)

M133

Submafia. Each of an infinite number of gangsters is chasing another one. Prove that we can choose an infinite subset of gangsters in which none of them is chasing another gangster in this subset. (V. Ufnarovsky)

M134

Numbers around decagons. A positive integer is written at each of the twenty vertices of two given regular decagons such that the sum of the numbers around either decagon is 99. Prove that it's possible to mark a number of successive vertices on each decagon (maybe one vertex, but not all) so that the two sums of the marked numbers are equal. (S. Berlov)

M135

*What Napoleon failed to notice.*¹ (a) Three equilateral triangles ABC_1 , BCA_1 , and CAB_1 are constructed externally on the sides of an arbitrary triangle ABC ; the midpoints of the segments A_1B_1 , B_1C_1 , C_1A_1 are labeled C_2 , A_2 , B_2 , respectively. Prove that the lines AA_2 , BB_2 , CC_2 meet at the same point or are parallel. (b) Prove this statement with "equilateral triangles" replaced by any similar isosceles triangles ABC_1 , BCA_1 , CAB_1 with bases AB , BC , CA . (N. Sedrakian, S. Tkachov)

Physics

P131

Car on ice. Due to the small coefficient of friction, a car can't move along a road covered with ice with an acceleration exceeding $a = 0.5 \text{ m/s}^2$. According to the rules of a competition, the car must go from point A

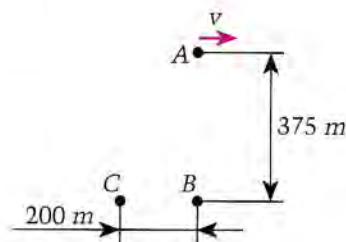


Figure 1

¹Why Napoleon? See page 39 in the July/August 1994 issue of *Quantum*.

to point B, which is located at right angles to the initial velocity of the car (fig. 1), as quickly as possible. What is the minimum time needed for this if the distance $AB = 375 \text{ m}$ and the initial velocity of the car $v = 10 \text{ m/s}$? What is the car's trajectory? Answer the same questions for the case when the finish line is located at point C, where BC is 200 m . (A. Korotkov, E. Yunosov)

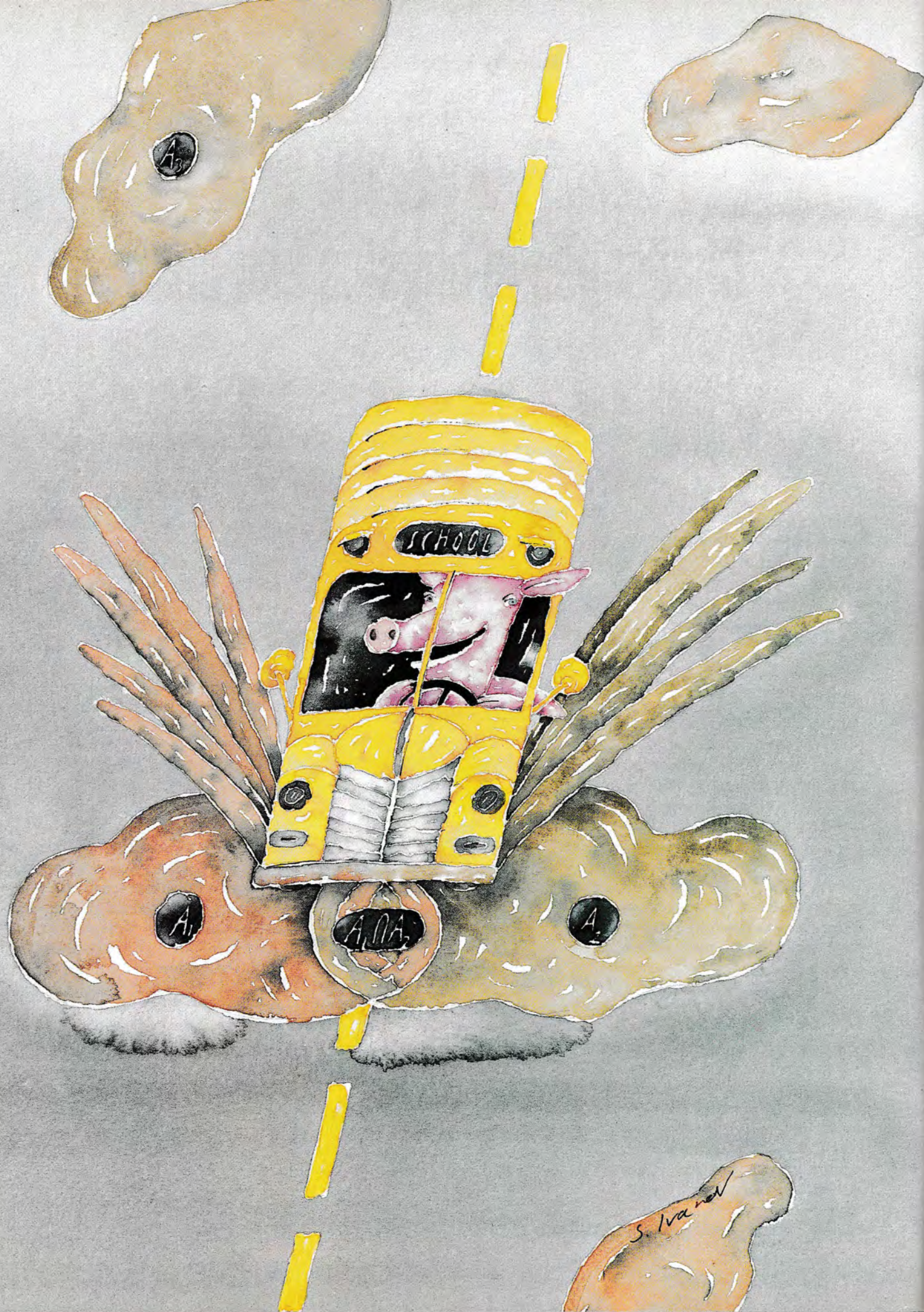
P132

Satellite of the Sun. Calculate the minimum period of revolution of a spaceship around the Sun given that the angular size of the Sun as seen from the Earth is $\alpha = 9.3 \cdot 10^{-3} \text{ rad}$. (M. Gavrilov)

P133

On a distant planet. The surface of a lifeless planet is covered with a thick layer of frozen carbonic acid. It is suggested that an atmosphere can be created on this planet consisting of pure oxygen obtained from the decomposition of the carbonic acid into carbon and oxygen. How long will it take if the decomposition rate is 10^6 moles per second? It is necessary to obtain a pressure $P = 0.2 \text{ atm}$. Consider the temperature near the planet's surface to be $T = 200 \text{ K}$, at which the evaporation of the carbonic acid can be neglected. The mass of the planet

CONTINUED ON PAGE 39



The school bus and the mud puddles

An application of the inclusion–exclusion theorem

by Thomas P. Dence

LET'S START WITH A PROBLEM that has made the rounds. Maybe you've seen it already. Even if you have, it bears repeating.

Suppose you live in Tinytown. The town has two east–west streets and three north–south streets. A map of Tinytown is shown in figure 1. Of course, there may be other roads leading in and out of Tinytown. But we are interested only in those shown in figure 1, because they help you get to school.

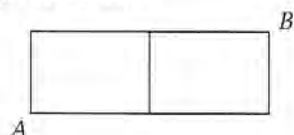


Figure 1

You live at point A and, by walking along the streets, you must get to school at point B. The question is, along how many different routes can you travel? In actuality you could travel infinitely many different routes by merely traveling around a block as often as possible, but we don't want to consider these redundant routes. We want to make the trip as short as possible.

There are really only three routes that are considered acceptable. We can keep track of these routes by indicating which way you must turn

at each intersection (including the intersection at A). At each corner you can choose to go north or east—any other choice would make the route longer than necessary. The three possible routes are

- Route 1: north, east, east;
- Route 2: east, north, east;
- Route 3: east, east, north.

Before reading further, try the following exercise.

Exercise 1. Find the number of shortest possible routes for figures 2 and 3.

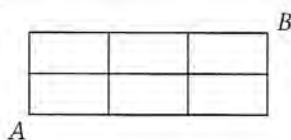


Figure 2

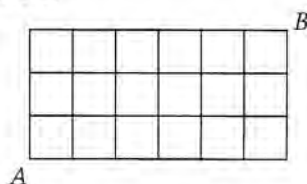


Figure 3

Now suppose we look at a larger town, Midvale. Its map looks like figure 4. Again we ask: how many (shortest possible) routes are there from A to B? This is a little more difficult. First of all, since Midvale is

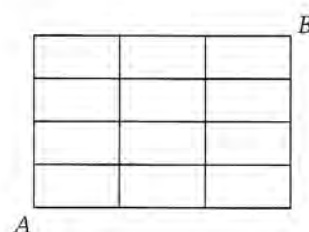


Figure 4

larger than Tinytown, let's start taking the bus from A to B instead of walking. Second, we need to generalize the above argument rather than counting haphazardly.

It's not hard to see that in Midvale, the shortest possible trip will last precisely seven blocks, and that the bus will have to make three turns to the east and four turns to the north. So we can count the possible paths to school by arranging these turns. This means that we are choosing three out of seven turns to be east turns and letting the others be north turns. But there is a formula for such problems. The answer is simply the binomial coefficient $C(7, 3) = \binom{7}{3} = 35$. (Try to list all 35 possibilities—it's a good exercise in keeping yourself organized!)

Exercise 2. Derive a formula for the number of ways to get to school by routes of minimal length in a

town that is m blocks wide and n blocks long.

Detours

Let's look at a town called Mudville (fig. 5). Mudville is a 6×3 town, which means it's 6 blocks wide and 3 blocks long. There happens to be a big mud puddle in the middle of one road, located between C and D , so that our school bus can't get through. As before, we wish to count the number of minimal-length paths from the lower left corner (point A) to the upper right corner (point B). We can do this by

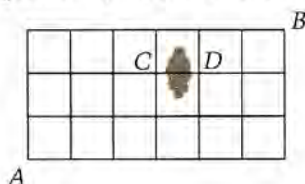


Figure 5

counting all the routes (including those that go through the mud) and subtracting from this number those that definitely do not go through the mud. For the first number, exercise 2 has already given us the answer.

It's $C(9, 3) = \binom{9}{3} = 84$. To count the paths that go through the mud, we first count those that go from A to C . Again, by exercise 2, this number is $\binom{5}{2} = \binom{5}{3} = 10$. Similarly, the number of paths from C to B that pass through the mud, which is the same as the number of paths from D to B ,

must be $\binom{3}{1} = 3$. The product $\binom{5}{3}\binom{3}{1} =$

30 thus counts the number of paths from A to B that use segment CD , and so the number of paths that avoid segment CD is the difference $\binom{9}{3} - \binom{5}{3}\binom{3}{1} = 54$.

We note that this counting strategy will work no matter which street segment is puddle-closed, be it a north-south artery or an east-west thoroughfare, although the actual count will most likely vary.

Exercise 3. Find the number of bus routes from A to B that avoid the mud in figures 6 and 7.

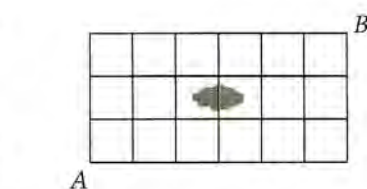


Figure 6

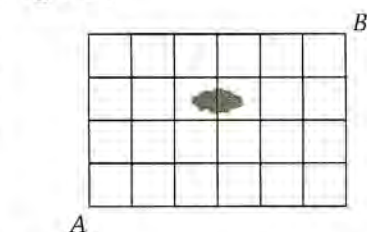


Figure 7

Heavy rains

The nearby town of Sogdale is located on the river, and lots of puddles form there after it rains. Figure 8a shows the town after a heavy rainstorm, when two big puddles have made two blocks impassable for our bus. The drains were repaired, but then another rain caused two new puddles, as shown in figure 8b.

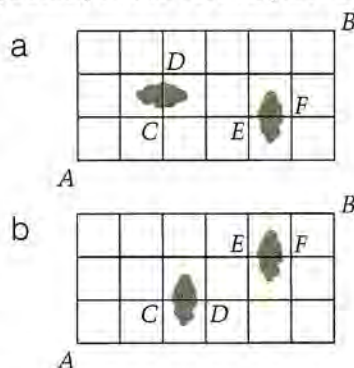


Figure 8

What can we say now about the number of good paths for our trusty school bus? We observe that routes from A to B now fall into one of four distinct categories. They pass through only the first mud puddle, or they pass through only the second mud puddle, or they pass through both (which could happen in figure 8a but not in figure 8b). As our school bus starts out in this more complicated setting, we seek a more general form of counting the number of routes. It's at this point that we start to introduce the terminology from the inclusion-exclusion

theorem. We recall that this theorem gives a count of the number of elements in the finite union of finite sets. To begin, if we have two sets—say, A_1 and A_2 —then the number of elements in their union $|A_1 \cup A_2|$ is given by $|A_1| + |A_2| - |A_1 \cap A_2|$. The Venn diagram in figure 9 is typically used to illustrate this situation.

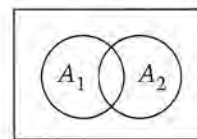


Figure 9

Exercise 4. For a Venn diagram with three sets ($n = 3$) the inclusion-exclusion theorem states that

$$|A_1 \cup A_2 \cup A_3| = |A_1| + |A_2| + |A_3| - |A_1 \cap A_2| - |A_1 \cap A_3| - |A_2 \cap A_3| + |A_1 \cap A_2 \cap A_3|.$$

Verify, for example, that if an element is in exactly two of the sets A_1, A_2, A_3 , it is counted precisely once in the above expression. What about elements that are in exactly one set? In all three sets?

This counting result generalizes to $n \geq 2$ sets (and therefore a more complicated Venn diagram), where the number of elements in the union of n sets is a function of the cardinality of the union of all combinations of k of these sets, as k ranges over all values from 1 to n . Specifically, if $A_1, A_2, \dots, A_n$ denote n arbitrary sets, then we have

$$|A_1 \cup A_2 \cup \dots \cup A_n| = |A_1| - |A_1 \cap A_2| - |A_1 \cap A_3| - \dots - |A_{n-1} \cap A_n| + |A_1 \cap A_2 \cap A_3| + \dots + (-1)^{n+1} |A_1 \cap A_2 \cap \dots \cap A_n|.$$

Returning to our puddle problem illustrated in figure 8, if we define the three sets U, A_1, A_2 by

- U = all good routes from A to B ,
- A_1 = all good routes from A to B that pass through CD ,
- A_2 = all good routes from A to B that pass through EF ,

then A_1 and A_2 are proper subsets of U and, in figure 8a, $A_1 \cap A_2 = \emptyset$, while in figure 8b, $A_1 \cap A_2 \neq \emptyset$. By the inclusion-exclusion theorem (with $n = 2$), the number of good

routes N from A to B is given by

$$|U| - |A_1| - |A_2| + |A_1 \cap A_2|.$$

In the case of figure 8a this reduces to

$$N = \binom{9}{3} - \binom{3}{1}\binom{5}{1} - \binom{5}{1}\binom{3}{1} = 138,$$

while in figure 8b the number of good routes is

$$\begin{aligned} N &= \binom{9}{3} - \binom{3}{1}\binom{5}{2} - \binom{6}{2}\binom{2}{1} \\ &\quad + \binom{3}{1}\binom{2}{1}\binom{2}{1} \\ &= 120. \end{aligned}$$

From the last term we see that there are 12 routes from A to B that pass through both CD and EF . Each of these routes had already been counted twice by belonging to both A_1 and A_2 .

Lots more water

We demonstrate the pattern more fully with a final example (fig. 10). In this setting there are five street sections (numbered 1 to 5) that are closed, and accordingly we define five sets A_i :

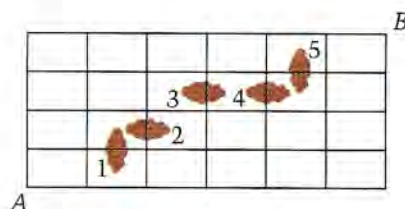


Figure 10

$A_i = \{\text{all routes from } A \text{ to } B \text{ that pass through street section } i\}$

for $i = 1, 2, 3, 4, 5$. It's possible that a route from A to B passes through as many as four of these mudpuddled blocks. In particular, sets $A_1 \cap A_3 \cap A_5$ and $A_1 \cap A_2 \cap A_4 \cap A_5$ are nonempty, with

$$|A_1 \cap A_3 \cap A_5| = \binom{2}{1}\binom{2}{1}\binom{1}{1}\binom{2}{1} = 8$$

and

$$\begin{aligned} |A_1 \cap A_2 \cap A_4 \cap A_5| &= \binom{2}{1}\binom{1}{1}\binom{1}{1}\binom{2}{1} \\ &= 4. \end{aligned}$$

By the inclusion-exclusion theorem the number of good routes N from A to B is given by

$$\begin{aligned} N &= |U| - \sum_i |A_i| + \sum_{i,j} |A_i \cap A_j| \\ &\quad - \sum_{i,j,k} |A_i \cap A_j \cap A_k| \\ &\quad + \sum_{i,j,k,l} |A_i \cap A_j \cap A_k \cap A_l|. \end{aligned}$$

Simplification gives

$$\begin{aligned} \sum |A_i| &= |A_1| + |A_2| + |A_3| + |A_4| + |A_5| \\ &= \binom{2}{1}\binom{7}{3} + \binom{3}{1}\binom{6}{2} + \binom{5}{2}\binom{4}{1} \\ &\quad + \binom{6}{2}\binom{3}{1} + \binom{7}{3}\binom{2}{1} \\ &= 270, \end{aligned}$$

$$\begin{aligned} \sum |A_i \cap A_j| &= |A_1 \cap A_2| + |A_1 \cap A_3| \\ &\quad + |A_1 \cap A_4| + |A_1 \cap A_5| + |A_2 \cap A_3| \\ &\quad + |A_2 \cap A_4| + |A_2 \cap A_5| + |A_3 \cap A_5| \\ &\quad + |A_4 \cap A_5| \\ &= \binom{2}{1}\binom{6}{2} + \binom{2}{1}\binom{2}{1}\binom{4}{1} + \binom{2}{1}\binom{3}{1}\binom{3}{1} \\ &\quad + \binom{2}{1}\binom{4}{2}\binom{2}{1} + \binom{3}{1}\binom{4}{1} + \binom{3}{1}\binom{3}{1} \\ &\quad + \binom{3}{1}\binom{3}{1}\binom{2}{1} + \binom{5}{2}\binom{2}{1} + \binom{6}{2}\binom{2}{1} \\ &= 177 \end{aligned}$$

(since $A_3 \cap A_4 = \emptyset$),

$$\begin{aligned} \sum |A_i \cap A_j \cap A_k| &= |A_1 \cap A_2 \cap A_3| \\ &\quad + |A_1 \cap A_2 \cap A_4| + |A_1 \cap A_2 \cap A_5| \\ &\quad + |A_1 \cap A_3 \cap A_5| + |A_1 \cap A_4 \cap A_5| \\ &\quad + |A_2 \cap A_3 \cap A_5| + |A_2 \cap A_4 \cap A_5| \\ &\quad + (\text{all other terms equal to zero}) \\ &= \binom{2}{1}\binom{4}{1} + \binom{2}{1}\binom{3}{1} + \binom{2}{1}\binom{3}{1}\binom{2}{1} \\ &\quad + \binom{2}{1}\binom{2}{1}\binom{2}{1} + \binom{2}{1}\binom{3}{1}\binom{2}{1} + \binom{3}{1}\binom{2}{1} \\ &\quad + \binom{3}{1}\binom{2}{1} \\ &= 58, \end{aligned}$$

Space contributed by the publisher as a public service.

and

$$\begin{aligned} & \sum |A_i \cap A_j \cap A_k \cap A_l| \\ &= |A_1 \cap A_2 \cap A_3 \cap A_5| \\ &+ |A_1 \cap A_2 \cap A_4 \cap A_5| \\ &+ (\text{other terms equal to zero}) \\ &= \binom{2}{1} \binom{2}{1} + \binom{2}{1} \binom{2}{1} \\ &= 8. \end{aligned}$$

So the number of good routes N from A to B must be

$$N = \binom{10}{4} - 270 + 177 - 58 + 8 = 67.$$

In closing, we note that even though this method can be quite cumbersome for many segments, sometimes early simplification can occur. This happens whenever a large-puddled street CD can be eliminated from consideration because the presence of the other puddled streets would prohibit travel on CD . In figure 11a we have

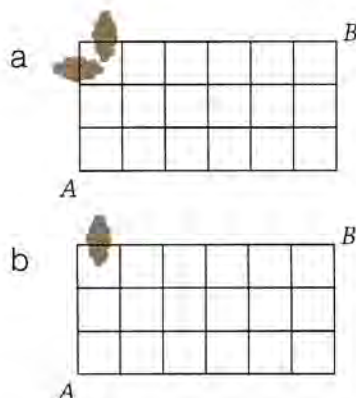


Figure 11

two streets with large puddles and in figure 11b we have only one, yet the two figures are equivalent in the sense that they yield precisely the same set of good routes from A to B .

You may enjoy finding other tricks of counting in the following exercises.

Exercise 5. Determine how many good routes there are from A to B in figure 12.

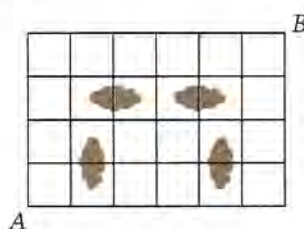


Figure 12

These grid-type layouts provide the framework for a variety of interesting extensions. It's possible, for example, to develop a sensible distance, or metric, function that yields a rich geometry. For instance, the distance between any two points (x_1, y_1) , (x_2, y_2) in the plane can be given by the *school-bus distance* $|x_1 - x_2| + |y_1 - y_2|$. This name is appropriate since distances are measured just east-west and north-south. The set of points that are 2 units from the origin is now not a circle, but a square (fig. 13).

Exercise 6. Using this school-bus distance, draw the locus of points whose distances from the two fixed points $(1, 0)$ and $(-1, 0)$ sum to 4. This is a "school bus ellipse."

In closing, here is another application of the inclusion-exclusion

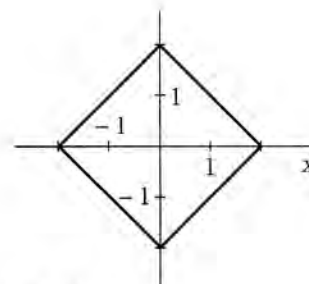


Figure 13

principle—one that comes from the field of number theory. Here we encounter one of the classic functions of mathematics: the Euler ϕ -function, defined as the number of positive integers less than or equal to n that are relatively prime to n . We begin by expressing n in its canonical prime decomposition form:

$$n = p_1^{a_1} p_2^{a_2} \dots p_m^{a_m},$$

where $p_1 < p_2 < \dots < p_m$ are primes and $a_1, a_2, \dots, a_m$ are positive integers. Set $U = \{1, 2, \dots, n\}$, and for each integer i , $1 \leq i \leq m$, let A_i be the subset of integer multiples of p_i that are in U . Each A_i contains n/p_i elements, and if each of these elements is removed from U —

$$U - \bigcup_{i=1}^m A_i$$

—then it's possible that the count of this set does not agree with

$$n - \sum_{i=1}^m \frac{n}{p_i},$$

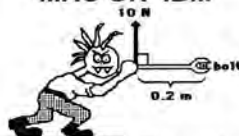
because this latter may count elements common to A_i and A_j twice. So we have to add back in this new tally—

$$n - \sum_{i=1}^m \frac{n}{p_i} + \sum_{i \neq j} \frac{n}{p_i p_j}$$

—to account for a possible nonempty intersection of A_i and A_j . But then we have perhaps counted some elements too many times—those that are multiples of three distinct primes p_i, p_j, p_k . This means that the number $|S|$ of elements in S is approximated by

$$\begin{aligned} |U| - \sum |A_i| + \sum |A_i \cap A_j| \\ - \sum |A_i \cap A_j \cap A_k| \end{aligned}$$

PHYSICS SOFTWARE MAC OR IBM



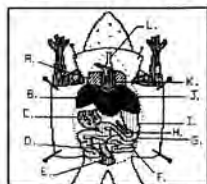
Math Methods	Thermodynamics
Statics	Electricity & Mag.
Motion	Optics
Conservation Laws	Atomic Physics
Circular Motion	Phys. Activity Set

\$35 each/ whole set \$320
Lab pack or Network version \$70 each
Site license \$105 each

Free catalog

Cross Educational Software - P.O. BOX 1536
508 East Kentucky, Ruston, LA 71270
800-768-1969

DISSECTIONS APPLE IBM MACINTOSH



Frog...Earthworm...Grasshopper...
Perch...Crayfish...Starfish...Clam
\$29.95 each -- set of 7 \$190

Circle No. 3 on Reader Service Card

or, equivalently,

$$n - \sum_{i=1}^m \frac{n}{p_i} + \sum_{i \neq j} \frac{n}{p_i p_j} - \sum_{i \neq j \neq k} \frac{n}{p_i p_j p_k}.$$

The inclusion-exclusion principle allows us to continue this line of reasoning, concluding with

$$n - \sum \frac{n}{p_i} + \sum \frac{n}{p_i p_j} + \dots + \sum \frac{n(-1)^m}{p_i p_j \dots p_m}.$$

To illustrate, with $n = 60 = 2^2 \cdot 3 \cdot 5$ we have

$$\begin{aligned} A_1 &= \{2, 4, 6, 8, \dots, 60\}, \\ A_2 &= \{3, 6, 9, 12, \dots, 60\}, \\ A_3 &= \{5, 10, 15, \dots, 60\}, \end{aligned}$$

so

$$\begin{aligned} |S| &= |U| - \sum |A_i| + \sum |A_i \cap A_j| \\ &\quad - \sum |A_i \cap A_j \cap A_k| \\ &= n - (|A_1| + |A_2| + |A_3|) \\ &\quad + (|A_1 \cap A_2| + |A_1 \cap A_3| + |A_2 \cap A_3|) \\ &\quad - |A_1 \cap A_2 \cap A_3| \\ &= 60 - (30 + 20 + 12) \\ &\quad + (10 + 6 + 4) - 2 \\ &= 16. \end{aligned}$$

Indeed, for the number 60,

$$S = \{1, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 49, 53, 59\},$$

and so $|S| = 16 = \phi(60)$.

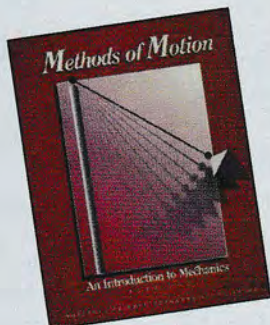
I leave you with one last exercise. It is similar in scope to the preceding exercises, but it differs in its method of solution. It's categorized in the literature as a "derangement."

Exercise 7. Grandma Jones received Christmas cards from each of her four sons, but in her haste in placing the cards back in their envelopes, she got the cards mixed up. How many different ways could she have replaced the cards so not a single card corresponded with the proper envelope? 

ANSWERS, HINTS & SOLUTIONS
ON PAGE 59

Physics Phluency

Let NSTA help you speak the language of physics



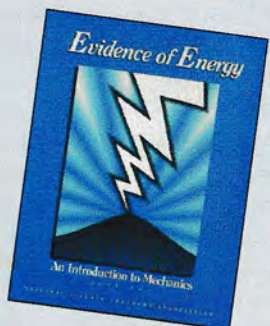
Methods of Motion

An Introduction to Mechanics, Book One

Twenty-seven teacher-created activities aim to simplify the daunting world of Newtonian mechanics for students and teachers.

(grades 6–10, 1992 revised ed., 168 pp.)

#PB039X \$18.50



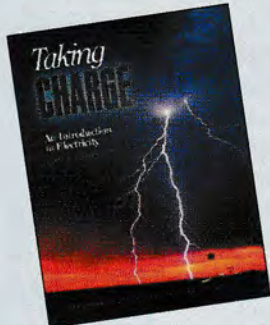
Evidence of Energy

An Introduction to Mechanics, Book Two

The informal hands-on activities in this book use a variety of techniques to combat common misconceptions about mechanics.

(grades 6–10, 1990, 200 pp.)

#PB080X \$17.95



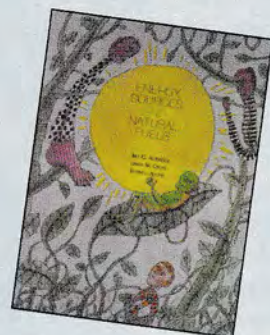
Taking Charge

An Introduction to Electricity

Spark student interest in electricity with 25 hands-on, teacher-tested activities using readily available materials.

(grades 5–10, 1992, 160 pp.)

#PB096X \$18.95



Energy Sources and Natural Fuels

Explore energy, photosynthesis, fossil fuels, and more in this collaboration between NSTA and the Russian Academy of Science. (A teacher's guide and classroom sets also are available.)

(grades 9–10, 1993, 80 pp.)

#PB104X \$12.95

To Order, Call:

(800) 722-NSTA

MasterCard, VISA, Discover, and Purchase Orders

Please make all orders payable in U.S. currency

QJF95

A magical musical formula

Guitar-tuning for the tone-deaf

by P. Mikheyev

TO TELL YOU THE TRUTH, my ear for music isn't perfect. But sometimes I like to strum on the guitar. And the hardest thing for me happens before I even start playing—I have to tune the instrument. To solve the problem once and for all, I grabbed a pen and sat down at my desk. And, lo and behold—the physics of the thing helped me! It turned out that I only need to insert numbers into a certain formula (to be derived later on), turn the pegs of the guitar the calculated number of times, and the instrument is tuned. Now I can forget equations and enjoy the music.

After this little prelude, let's pass from words to deeds and generate the magic formula.

First of all we determine the angular frequency of a string. Let's assume that fundamental oscillations, which contain only half of the wavelength (fig. 1), play the major role. Evidently the angular frequency depends on the tension of the string T , its length L , and its mass m . Using dimensional analysis, we get the formula

$$\omega = A \sqrt{\frac{T}{Lm}},$$

where A is some dimensionless coefficient (prove to yourself that only this combination of T , L , and m has



Figure 1

the correct dimensionality.) The exact solution of the problem results in

$$\omega = \pi \sqrt{\frac{T}{Lm}}.$$

Now let's write the value of the force. When tightened, the string is stretched and deformed. Therefore, an elastic force arises:

$$T = YS \frac{\Delta x}{L},$$

where Δx is the elongation of the string, Y is Young's modulus for this material, and S is the cross-sectional area of the string. Since $m = \rho V = \rho SL$, the angular frequency is

$$\omega = \pi \sqrt{\frac{Y \Delta x}{\rho L^3}}.$$

An important feature should be noted at once. By assuming $m = \rho SL$, we've restricted ourselves to solid, homogeneous strings—that is, the first and second strings of the guitar. The other strings have additional windings to increase their masses.

Now we'll look at the process of stretching a string. The string is wound up on a cylinder of diameter d . The shaft of this cylinder has a gear that we can rotate by turning a tuning peg linked mechanically with it (fig. 2). Let N be the number of turns the cylinder makes when the string is stretched. Then $\Delta x = N\pi d$. I checked experimentally that one turn of the cylinder corresponds to $\alpha = 30$ turns

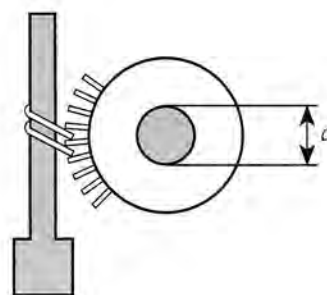


Figure 2

of the tuning peg. So, to rotate the cylinder N times, the peg must make $n = N\alpha$ full turns. Therefore, $\Delta x = n\pi d/\alpha$. Inserting this value into the formula for the angular frequency yields

$$\omega = \pi \sqrt{\frac{\pi Y d n}{\rho L^3 \alpha}}.$$

Since the oscillation frequency f is directly related to the angular frequency ω by the formula $f = \omega/2\pi$, then

$$f = \frac{1}{2} \sqrt{\frac{\pi Y d n}{\rho L^3 \alpha}}.$$

Finally, we obtain the number of turns of the tuning peg:

$$n = \frac{4\rho L^3 \alpha f^2}{\pi Y d} = B f^2,$$

where $B = 4\rho L^3 \alpha / \pi Y d$.

Let's calculate the value of the constant B . Substituting $\alpha = 30$, $\rho = 7,800 \text{ kg/m}^3$, $L = 0.7 \text{ m}$ (for the first string), $d = 5 \cdot 10^{-3} \text{ m}$, $Y = 2 \cdot 10^{11} \text{ Pa}$, we get $B = 10^{-4} \text{ Hz}^{-2}$. Thus, knowing

the frequency of the sound, we can determine how many turns are needed to tune the string. Now we can start to tune the guitar.

The first string should be tuned to E above middle C, or from the physical standpoint, to the frequency $f = 300$ Hz. This corresponds to $n_1 = 9$ turns. The second string produces B below middle C ($f = 247$ Hz), and the tuning peg should be turned $n_2 = 6.1$ turns.

This looks fine, but only the first two strings can be tuned this way. What about the others? Physics helps here, too. From the theory of oscillations and waves the word "beats" may be familiar to you. For those who are encountering this word for the first time, I'll explain what it means. If two oscillations with similar frequencies are superimposed, a very intriguing picture results. Let $x_1 = A \cos \omega_1 t$ and $x_2 = A \cos \omega_2 t$ be two oscillations of equal amplitudes and zero initial phases.

Adding these oscillations together yields a new one:

$$x = 2A \cos \frac{\omega_1 - \omega_2}{2} t \cos \frac{\omega_1 + \omega_2}{2} t,$$

or

$$x = A_0 \cos \omega_m t,$$

where

$$A_0 = 2A \cos \frac{\omega_1 - \omega_2}{2} t.$$

Since $\omega_1 \approx \omega_2$, then $\Delta\omega = \omega_1 - \omega_2$ is a very small value, which means that the amplitude A_0 varies slowly with time. The graph of such an oscillation is shown in figure 3.

If these oscillations are acoustic, we can hear the sound increase and decrease in volume (the beats). They will help to tune the other strings. The unfretted second string should have the same frequency as the third string with one's finger on the fourth fret. Plucking both strings simultaneously, we hear beats as long as the frequencies are close. If the tuning is close, we can measure the period between successive dampings of the sound. Let this time be T . Then the error in the string's tuning is

$$\frac{\Delta f}{f_0} = \frac{1}{T f_0},$$

where f_0 is the standard frequency. Evidently, as $T \rightarrow \infty$ the error becomes negligibly small. The longer the period of the beats, the better the tuning.

Using this method, we can

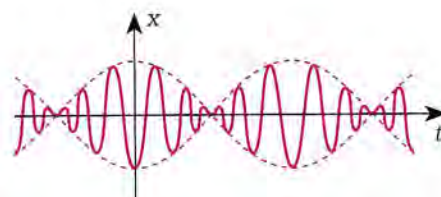


Figure 3

tune all the strings. But there is one pesky detail remaining: a string can't be stretched too much or it will break. What is the upper limit for stretching a string? Suppose the string is made of ordinary steel. The breaking strength for this material is $\sigma_{\max} = 5 \cdot 10^8$ Pa. Since

$$\sigma_{\max} = Y \frac{\Delta x_{\max}}{L},$$

then

$$\Delta x_{\max} = \frac{\sigma_{\max} L}{Y} = 2 \text{ mm}$$

—that is, we can make four turns to obtain a maximum frequency $f = 200$ Hz. Because real strings are made of a special alloy, their limit is higher. This can be determined experimentally by stretching a string until it breaks. The calculations can easily be made according to the formula

$$\sigma_{\max} = \frac{\pi d Y n}{L \alpha}.$$

Breaking a couple of strings, I found the value $\sigma_{\max} = 5 \cdot 10^9$ Pa, which is an order of magnitude greater than the value for strings made of ordinary steel.

As you can see, we can draw much more out of a guitar than just music—we can use it to produce a suite of interesting physical facts. And this theme is far from being exhausted. The most interesting remains before us—an investigation of the acoustic properties of the guitar's resonating sound box. Maybe physics would help here as well? Possibly it would even suggest another shape—one that would produce sounds even more enchanting than those coaxed out of the guitars we inherited from the ages. 

Surfing the electrom

"The experiments described—it seems to me, at least—leave no doubt that

PARADOXICALLY, ELECTROMAGNETIC waves have been known to humankind from time immemorial. People were warmed by thermal rays long before the term "infrared radiation" was coined, and they got sunburnt without knowing anything about the effects of ultraviolet rays on the skin. And they had vision, after all! They not only saw, they conducted experiments with light as a wavelike substance.

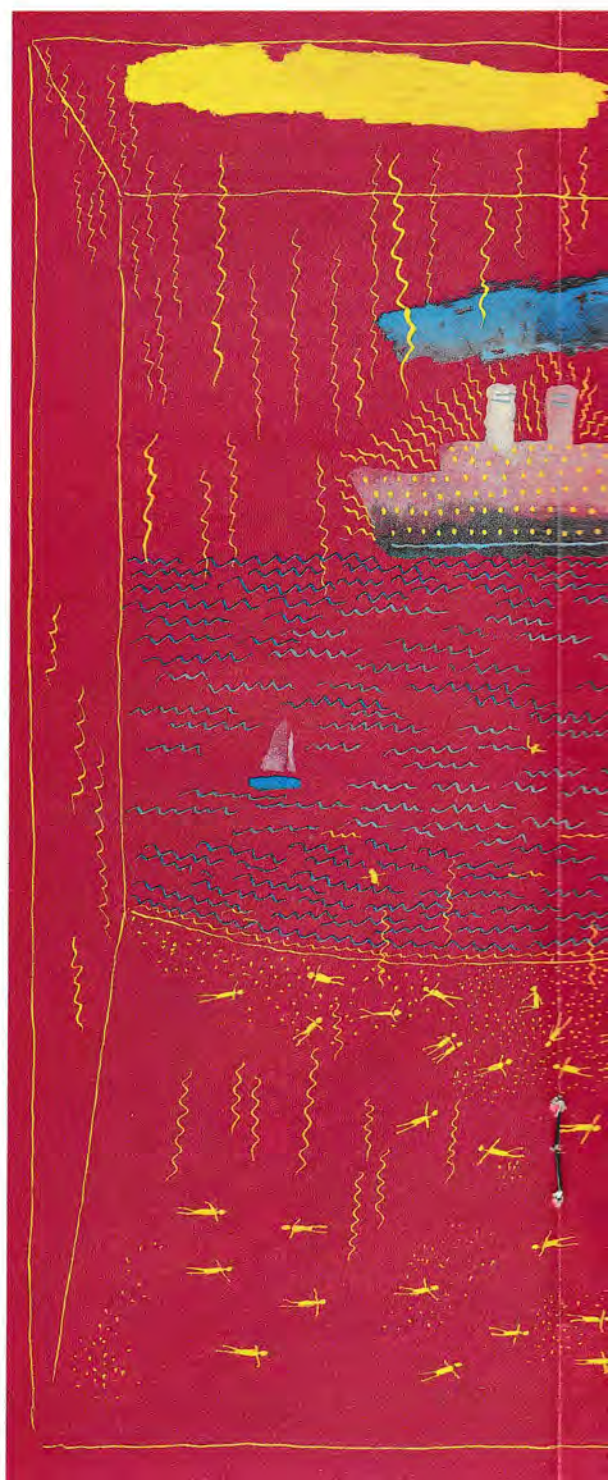
The suspicion that all these kinds of radiation have a common electromagnetic basis led to the development of a scientific theory. In 1865 James Clerk Maxwell generalized the findings of his great forerunners, above all Michael Faraday. Maxwell predicted, among other phenomena, the transmission of information without wires.

Nowadays, when a little more than a century has passed since the creation of Maxwell's theory, electromagnetic waves do more than bring us radio and television signals. People have learned to generate and receive radiation in all parts of a strikingly wide electromagnetic spectrum—from low-frequency radio waves to gamma radiation. These invisible waves—microwaves, ultraviolet rays, infrared rays, X rays—made it possible to "hear" the previously elusive "conversations" among atoms, molecules, stars, and galaxies. Of course, in this article we can touch on only

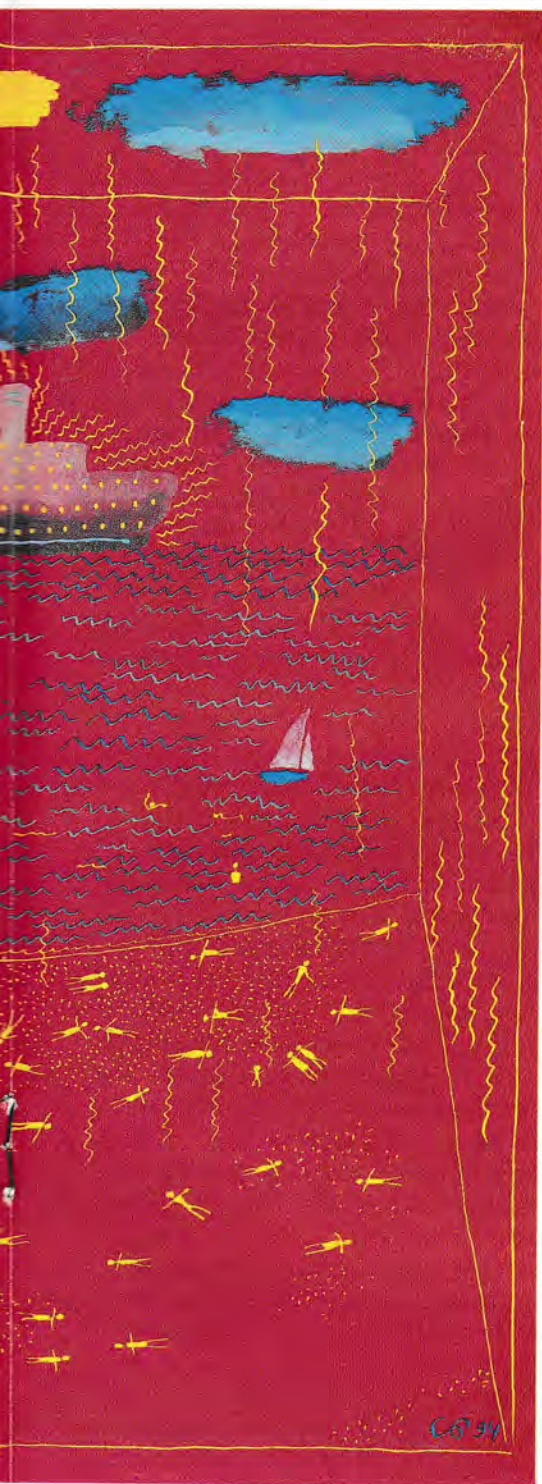
a few ranges of this huge electromagnetic spectrum.

Problems

1. An electrically charged sphere and a permanent magnet are placed near one another. Is there an electromagnetic field in the surrounding space?
2. Why are automobile antennas usually vertical?
3. Why are there so-called black-out zones for short-wave radio communications?
4. Why can radio stations transmitting long and medium waves be heard at far greater distances at night than during the day?
5. Why isn't radio communication with a submarine possible when it's underwater?
6. How was radiolocation first used in astronomy?
7. Why is stable reception of TV signals possible only within line of sight?
8. Why do the temperatures of all bodies in a closed, unheated room reach the same value?
9. Does a piece of iron emit red light when it is white-hot?
10. Why isn't a glass prism suitable for obtaining the spectra of infrared and ultraviolet radiation?
11. Natural vegetation and the artificial kind used for camouflage can be clearly distinguished in photos made by reconnaissance planes, yet they are indistinguishable by



magnetic spectrum



light, heat radiation, and electromagnetic wave motion are equivalent."—Heinrich Herz

direct observation. Why?

12. Why does the continuous X-ray spectrum emitted by an X-ray tube have a distinct boundary at the short-wavelength end?

13. Can the X rays used for detecting manufacturing defects be replaced with gamma radiation from radioactive substances?

Microexperiment

Observe the operation of an electric space heater composed of a spiral filament and a well-polished concave metal surface. What role do you think this surface plays?

It's interesting that . . .

. . . according to pre-Maxwellian opinion, electric and magnetic fields should disappear at zero current (because it was accepted that the changing electric field produced no effect).

. . . the widespread opinion that Herz performed his experiments to confirm Maxwell's theory is mistaken. In the beginning Herz was rather an opponent of this theory and accepted it only in light of evidence he himself obtained.

. . . the transition from an electric lamp with a heated carbon filament to the modern tungsten-filament lamps made it possible to increase the filament temperature by only 400 K. Nevertheless, it more than tripled the portion of the energy emitted in the visible

part of the spectrum—from 0.5% to 1.6%.

. . . infrared and ultraviolet rays, which are invisible to the human eye, are widely used by animals. At a distance of half a meter, some serpents feel changes in temperature as small as a tenth of a degree. And bees "see" ultraviolet rays that show the location of a flower's nectaries.

. . . the beginning of radioastronomy is connected with the work of a Bell Laboratories engineer, C. Yansky, who in 1931 conducted experiments with a rotating antenna to study the interference hampering short-wave radio communication. The noise he investigated seems to originate at the center of our galaxy.

. . . the parabolic antennas of modern radio telescopes are extremely sensitive—they can detect energy fluxes with densities less than $10^{-29} \text{ W} \cdot \text{s/m}^2$.

. . . to investigate the highest-energy gamma rays, astronomers use optical telescopes! Why? As they pass through the atmosphere, these gamma rays produce high-energy electrons, which excite Cherenkov radiation, which finally is recorded by an optical telescope.

—Compiled by A. Leonovich

ANSWERS, HINTS & SOLUTIONS
ON PAGE 57

What is elegance?

Mathematicians say: "I know it when I see it"

by Julia Angwin

IN ARCHITECTURE, THE epitome of elegance might be a Greek temple. In fashion, a Chanel suit. In mathematics, it's a term applied to the best, shortest, most inspired (and inspirational) proofs.

"An elegant proof just hits you between your eyes and fills your heart with joy," explains mathematician Irving Kaplansky.

One of the most elegant mathematicians of all time was Carl Friedrich Gauss. He lived in the period following the rapid expansion and development of mathematics in the 18th century. However, that century was not a period of elegance, according to mathematician Harold Edwards, who studies the history of math. It was Gauss in the 19th century who collected and refined the work done previously.

"He didn't publish anything until it was completely polished," says Edwards, a professor at New York University.

Gauss frustrated his peers by not publishing his proofs until they were perfect, but he thought that a cathedral is not a cathedral until the last scaffolding is down and out of sight. His motto was *Pauca sed matura*—"Few, but ripe."

Elegant proofs come from God, according to the Hungarian mathematician Paul Erdős. His theory is that God has a book containing all the

best proofs, and he lets a mortal one of them. even need to God, you just lieve in the book," he said. "You feel: 'How foolish that

sometimes glimpse "You don't believe in need to be-

I didn't think of it myself."

But, as in other things, elegance is sometimes simply a matter of taste.

"I like these combinatorial things that my colleagues think are a waste of time," laments

John Conway, professor at Princeton University.

However, most mathematicians agree on the basics:

1. Any proof that involves computer number-crunching is not elegant. (The proof of the Four Color Theorem comes to mind.)
2. Elegant proofs are easily understandable.

The classic example of elegance is Euclid's short proof that there are an infinite number of prime numbers. A prime number is one that can be evenly divided only by itself and the number 1. Nonprimes are divisible by primes and so are called composite numbers, since they are composed of primes.

"It's 2,300 years old and there's no better proof," says Kaplansky, director emeritus of the UC-Berkeley Mathematical Science Research Institute.

Euclid's proof is short and bypasses the fact that there is no formula for



Art by Sergey Ivanov

determining prime numbers. Without his proof, one could flounder around trying larger and larger primes without ever determining an upper limit. Euclid simply assumed that there is a final prime number and proves that this assumption leads to a contradiction. Here's the proof.

1. Call the highest prime number Q .
2. Now multiply Q by all the primes leading up to it: $2 \times 3 \times 5 \times 7 \times \dots \times Q$.

3. Let P equal that sum plus 1: $(2 \times 3 \times 5 \times 7 \times \dots \times Q) + 1 = P$.

4. Then P is not divisible by any of the numbers 2, 3, 5, 7, ..., Q , because each divisor would leave a remainder of 1.

5. But P must be divisible by some prime because it is a composite number.

6. But that prime must be larger than Q , because we have used up all the smaller primes. This contradicts step 1.

7. So the assumption must have been false—there must be an infinite number of primes.

Problem 1. Note that Euclid does not claim that his number P is a prime. Indeed, show that $2 \times 3 \times 5 \times 7 \times 11 \times 13 + 1$ is divisible by 59. Can you find a prime divisor of $2 \times 3 \times 5 \times 7 \times 11 \times 13 \times 17$?

Part of the charm of Euclid's proof lies in the fact that the result is incredibly useful. Aside from their role in pure number theory, large prime numbers are used to make and break government codes.

"If the thing you're proving is useful or powerful and yet your proof is simple, that is a great thing," says Conway. A short, concise proof of a less important theorem might not be called elegant, he says. It would simply be cute or interesting.

But the most important quality of an elegant proof is that it makes you think, "Aha! How silly that I didn't think of that."

For example, consider a problem posed by mathematician Ron Graham. Consider a sequence of 101 distinct numbers arranged in any order you like. You can find a subsequence of 11 increasing or decreasing numbers in that set, he says.

First, to get a sense of it, think of the first 100 natural numbers arranged as follows:

91, 92, 93, ..., 100, 81, 82, 83, ..., 90, 71, 72, 73, ..., 80, ..., 1, 2, 3, ..., 10.

This is a sequence of 100 numbers. We can pick one number from each "decade" to create a subsequence of 10 decreasing numbers, such as 95, 85, 75, 65, 55, 45, 35, 25, 15, 5. Or you can pick 10 increasing numbers. But it's impossible to find an increasing or decreasing subsequence of 11 numbers.

So intuitively you can believe that with 101 numbers there will be such a sequence of 11 numbers. But how can we prove it? The uninspired approach would be to check all cases. But we are going to use a cute—ahem, elegant—trick.

We assign to each of the 101 numbers $A_1, A_2, \dots, A_k, \dots, A_{101}$ a pair of integers (i_k, j_k) as follows. Let i_k be the length of the longest increasing subsequence ending in A_k . For example, if the sequence is $\langle 11, 3, 5, 1, 7, 2, \dots \rangle$ and $k = 6$, then the largest increasing sequence ending in $A_k = 2$ is $\langle 1, 2 \rangle$ and $i_6 = 2$.

Similarly, let j_k be the length of the longest decreasing sequence ending in A_k . For our example, if $k = 6$, the longest decreasing sequence ending in $A_6 = 2$ could be either $\langle 11, 5, 2 \rangle$ or $\langle 11, 3, 2 \rangle$ or $\langle 11, 7, 2 \rangle$. In any of these cases, $j_6 = 3$. So for $k = 6$, we have $A_k = (2, 3)$.

Now we can prove that no two pairs of these integers can be the same. For assume the contrary: suppose that $(i_{m'}, i_m) = (i_{n'}, j_n)$ for some subscripts m and n , with $n > m$. Now if $A_n > A_{m'}$, then surely $i_n > i_{m'}$ because otherwise you could just append A_n to the end of an increasing sequence measured by $i_{m'}$. Similarly, if $A_n < A_{m'}$, then $j_n > j_{m'}$.

Now suppose all of the values for i_k and j_k are between 1 and 10. Then you would have 100 pairs. But you have 101 pairs, so the pigeonhole principle¹ guarantees that one of the pairs must contain an 11. This pair

¹See "Pigeons in Every Pigeonhole" in the January 1990 issue of *Quantum*.—Ed.

"points" to a subsequence such as is required in the problem.

Problem 2. Generalize this result to a sequence of $n^2 + 1$ different numbers.

Problem 3. Provide a counterexample to show that if the 101 numbers in the sequence are not distinct, then the result is false.

Don't worry if your proofs aren't elegant—most proofs aren't. But be sure to keep your eye peeled for an unexpected glimpse of God's book.

As John Conway put it: "Sometimes mathematics is like wandering around a strange town, wandering around some streets and suddenly you turn the corner and the view changes—you see the beauty of the whole thing." ◉

STUTTERING... Can be prevented!

1-800-992-9392
STUTTERING
FOUNDATION
OF AMERICA
Box 11048 • Memphis, TN 38111-0748



An Unforgettable Experience

Become An AFS Exchange Student
Call 1-800-AFS INFO
AFS Intercultural Programs
313 East 43rd Street, New York, New York 10017

Cloud formulations

*"Eternal Clouds, let us appear; let us arise
from the roaring depths of Ocean, our father;
let us fly towards the lofty mountains
[and] spread our damp wing
over their forest-laden summits . . ."—Aristophanes*

by Arthur Eisenkraft and Larry D. Kirkpatrick

HOW CAN YOU TAKE TEN thousand gallons of water and suspend them in mid-air? Build a cloud! It seems almost counterintuitive that wet air should be less dense than dry air and float in the sky. But the beauty of the cirrus and cumulus attest to this as we gaze at the myriad shapes and forms above us. Cloud formation reveals to us properties of the environment as well as properties of gases. As we notice that the western slopes of the Rockies are moist while the eastern slopes are deserts, we deduce that the air currents must be coming from the Pacific Ocean.

Last year, I went into a variety store to buy a mylar balloon. When asked what I wanted on the balloon, I wrote down on a piece of paper: $PV = nRT$. The employee was somewhat surprised and asked what the expression meant. After she acknowledged that she had studied high school chemistry, I hoped that she could now learn the ideal gas law within the context of her job. I asked her if anybody had ever bought a balloon during the winter and returned a few minutes later to complain that

the balloon had a leak. She said that this had indeed happened, but that she would explain to the consumer that the balloon would re-inflate as soon as they got it home. In fact, as she was explaining this to the doubting customer, the balloon would inflate before their eyes. The mysterious equation on my balloon could explain this phenomenon.

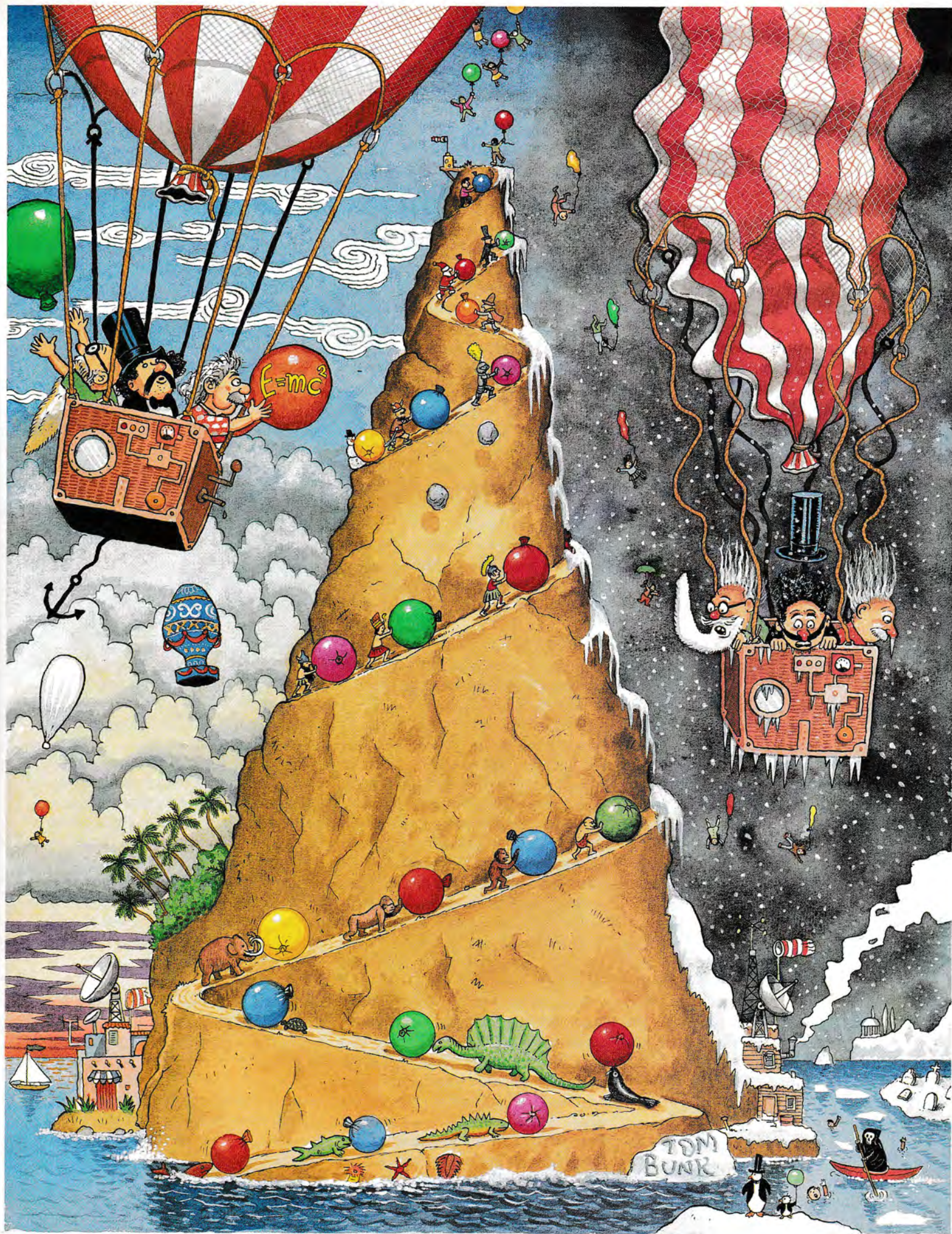
The ideal gas law, $PV = nRT$, describes the relationship between the macroscopic properties of an enclosed gas. In the equation, P is the pressure, V is the volume, T is the temperature of the gas in kelvins, n is the number of moles of gas, and R is the gas constant. In the mylar balloon example, the pressure of the balloon is a constant—the pressure of the atmosphere pushing on the balloon. The balloon is fully inflated inside the store while the temperature of the gas is equal to the store's room temperature. As the new balloon owner steps outside into the cold winter air, the temperature of the helium gas inside the balloon decreases. Since the pressure remains the same, the decrease in temperature is matched by a corresponding

decrease in volume and the mylar balloon appears to be only partially inflated. Stepping back into the store, the balloon will magically become fully inflated again as the gas warms up.

A second illustration of the gas law occurs when a bicycle tire is inflated. In this example, the volume of the tire remains constant. As more and more gas is pumped into the tire, the pressure increases and there is a corresponding increase in the temperature of the tire. Feel the tire and it will be warm. Automobile tires in the winter will be slightly underinflated when you begin your journey but will be just right when the tires warm up from the friction with the road and the flexing of the side walls.

The ideal gas law can also help our readers understand how a pressure cooker works, how our lungs inhale through the movement of the diaphragm, and how a hot air balloon rises and falls in the atmosphere. Physicists and engineers often summarize the behavior of a gas on a P - V diagram, where the pressure is plotted on the y -axis and the volume is on

Art by Tomas Bunk



the x-axis. For example, processes with constant temperatures are hyperbolas, since $PV = nRT$ and nRT is a constant.

Four processes are of special interest. The first three are changes that occur at constant temperature, constant volume, and constant pressure. In the fourth process no heat is transferred into or out of the system. This adiabatic process occurs when the change occurs very quickly—for instance, when sound waves move through the room. The changes in pressure occur so quickly, any heat transfer can be neglected. An adiabatic process also occurs when the system is thermally isolated from its surroundings. In this case the process can be very slow. As an example, a gas confined to an insulated container can expand adiabatically if weight on the piston is slowly removed.

When gases expand adiabatically, we expect that the pressure, the volume, and the temperature will all change. Fortunately, there is a relationship between the pressure and the volume during an adiabatic process: $PV^\gamma = \text{constant}$, where γ is the ratio of the specific heats for the gas and is equal to 1.4 for diatomic gases.

This brief introduction provides the background for this month's contest problem concerning cloud formation on the side of a mountain. The problem is adapted from the XVIII International Physics Olympiad which was held in Jena, East Germany, in 1987 (a few years before the German unification).

Moist air is streaming adiabatically across a mountain range as indicated in figure 1. Equal atmospheric pressures of 100 kPa are measured at meteorological stations

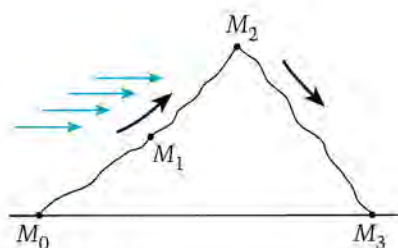


Figure 1

M_0 and M_3 , and a pressure of 70 kPa at station M_2 . The temperature of the air at M_0 is 20°C .

As the air ascends, cloud formation begins at location M_1 , where the pressure is measured to be 84.5 kPa.

A quantity of moist air, with a mass of 2,000 kg over each square meter, ascends the mountain. This moist air reaches the mountain ridge (M_2) after 1,500 s. During this time, 2.45 g of water per kilogram of air precipitates as rain.

A. Determine the temperature T_1 at M_1 , where the cloud forms.

B. Assuming that the atmospheric density decreases linearly with height, what is the height h_1 (M_1)?

C. What temperature T_2 is measured at the ridge of the mountain?

D. Determine the height of the water column precipitated by the air stream in 3 hours, assuming a homogeneous rainfall between points M_1 and M_2 .

E. What temperature T_3 is measured on the back side of the mountain range at station M_3 ? Compare the atmospheric conditions at M_0 and M_3 .

Hints: Assume that the atmosphere is an ideal gas. Influences of the water vapor on the atmospheric density are to be neglected.

The atmospheric density for P_0 and T_0 at station M_0 is $\rho_0 = 1.189 \text{ kg/m}^3$. The specific heat of vaporization of the water within the volume of the cloud is $L_v = 2,500 \text{ kJ/kg}$.

Please send your solutions to *Quantum*, 1840 Wilson Boulevard, Arlington VA 22201-3000 within a month of receipt of this issue. The best solutions will be noted in this space and their authors will receive special certificates from *Quantum*.

Mirror full of water

In the contest problem in the July/August issue of *Quantum*, we asked our readers to find the image produced by a concave mirror filled with water. The object is located along the optic axis at a distance $d = 3R/2$, where R is the radius of the mirror. We will model our solution after the very complete solution submitted by Prof. Anthony F. Behof of DePaul

University in Chicago, Illinois.

We begin by using the lens maker's formula to find the focal length f_w of the converging lens formed by the water:

$$\frac{1}{f_w} = \frac{n-1}{R} = \frac{4/3-1}{R} = \frac{1}{3R}.$$

Therefore, $f_w = 3R$.

Method A. We use the idea that the effective focal length f' of the combination of optical elements is the sum of the reciprocals of the individual elements. When we do this, we must use the focal length of the water lens twice because the light passes through the lens, strikes the mirror, and then passes through the lens again. The resulting formula is

$$\begin{aligned} \frac{1}{f'} &= \frac{1}{f_w} + \frac{1}{f_m} + \frac{1}{f_w} \\ &= \frac{1}{3R} + \frac{2}{R} + \frac{1}{3R} = \frac{8}{3R}. \end{aligned}$$

This effective focal length of $3R/8$ can be used in the mirror formula to find the image location:

$$\frac{1}{d'} = \frac{1}{f'} - \frac{1}{d} = \frac{8}{3R} - \frac{2}{3R} = \frac{2}{R}.$$

This tells us that the image is located at $d' = R/2$.

Thomas A. Davidson of Amarillo, Texas, points out that a mirror behaves the same whether it is immersed in air, water, or a vacuum. Introducing an air/water interface in front of the mirror effectively shortens the focal length of the mirror by the ratio of the indices of refraction. Thus, the effective focal length is $R/2n$, in agreement with the answer we obtained above.

Method B. An alternate method of finding the effective focal length of the water-mirror combination is to look at a ray parallel to the optic axis as shown in figure 2, where the angles and the thickness of the water have been exaggerated for clarity. Without the water this ray would pass through the focal point f of the mirror. However, because of the refraction at the surface, the ray

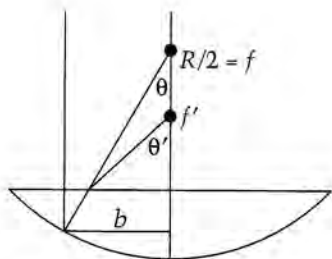


Figure 2

intersects the optic axis at the effective focal point f' . Snell's law tells us that

$$\frac{4}{3} \sin \theta = \sin \theta'.$$

Ignoring the thickness of the water, we also know that

$$\tan \theta \approx \frac{b}{f}$$

and

$$\tan \theta' \approx \frac{b}{f'}.$$

Because the angles are actually quite small, we use the small angle approximation that $\tan \theta \approx \sin \theta$ and $\tan \theta' \approx \sin \theta'$. Therefore, we have

$$\frac{4}{3} \frac{b}{f} \approx \frac{b}{f'},$$

or

$$f' \approx \frac{3}{4} f = \frac{3}{8} R.$$

Method C. The top surface of the water forms a virtual image at a distance

$$nd = \frac{4}{3} \frac{3R}{2} = 2R$$

above the surface of the water. This image acts like an object for the mirror without the water. Using the mirror formula with the focal length $f_m = R/2$ and $d = 2R$, we obtain

$$\frac{1}{d'} = \frac{1}{f} - \frac{1}{d} = \frac{2}{R} - \frac{1}{2R} = \frac{3}{2R},$$

or $d' = 2R/3$.

The light leaving the water gets bent once again to form an image at

$$\frac{d'}{n} = \frac{3}{4} \frac{2R}{3} = \frac{1}{2} R.$$

Method D. Finally we can treat the system as a combination of a water lens, a mirror, and a water lens by finding the image produced by each one and using the image as the object (sometimes imaginary) for the next one. Applying the lens formula to the water lens yields

$$\frac{1}{d'_1} = \frac{1}{f_w} - \frac{1}{d_1} = \frac{1}{3R} - \frac{2}{3R} = -\frac{1}{3R}.$$

Therefore, the image is virtual and located a distance $3R$ above the surface of the water.

This image acts as an object for the mirror. Ignoring the thickness of the water, the object distance for the mirror is $d_2 = 3R$. Using the mirror formula, we obtain

$$\frac{1}{d'_2} = \frac{1}{f_m} - \frac{1}{d_2} = \frac{2}{R} - \frac{1}{3R} = \frac{5}{3R},$$

and the image is located a distance $3R/5$ above the mirror.

Since this image is located on the "wrong" side of the water lens, the new object distance $d_3 = -3R/5$. Thus,

$$\frac{1}{d'_3} = \frac{1}{f_w} - \frac{1}{d_3} = \frac{1}{3R} - \frac{-5}{3R} = \frac{2}{R},$$

and the final image is located a distance $R/2$ above the surface as before.

Prof. Behof goes on to provide a fifth solution using matrix optics and a way of verifying the result experimentally. To see how to measure the effective focal length, let's first consider the case with no water and set the object and image distances equal to each other—that is, $d = d'$. We then find the well-known result that $d = 2f = R$. If we now add water, we find that $d = R/n = 3R/4$.

Now hold a lighted target above a mirror that has been filled with a few millimeters of water. Adjust the height of the target above the water until the target's image is in focus on a screen held at the same height. This height d is the effective radius of curvature of the combination and the effective focal length is $d/2$. ●

"HOW DO YOU FIGURE?" CONTINUED FROM PAGE 23

is $M = 7.5 \cdot 10^{22}$ kg (approximately equal to that of the Moon) and its radius $R = 1,750$ km. (D. Mogilevtsev)

P134

Frame in the B-field. A Π -shaped frame with equal sides made of thin wire is suspended freely from an articulated joint in a vertical magnetic field $\mathbf{B}$ (fig. 2). What is the maximum

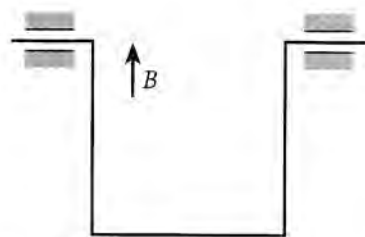


Figure 2

angle of deflection of the frame if a constant current I flows through it? The mass of a unit length of the wire is ρ . (A. Kipriyanov)

P135

Two lenses. The real image of a point source is formed at point A (fig. 3) by a thin lens located near the left-hand end of the line. After this

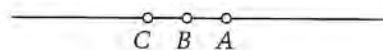


Figure 3

lens is replaced by another one positioned in the same place, one obtains an image at point B. Then the first lens is put very near the second one, and the image shifts to point C. Determine the location of the source geometrically. (A. Deshkovsky)

ANSWERS, HINTS & SOLUTIONS
ON PAGE 54

NSTA Executive Director Search

The National Science Teachers Association announces its search for an Executive Director to oversee the affairs of the largest organization of science teachers in the world.

The position requires

- ☐—Exercising the powers and duties of a secretary of a corporation.
- ☐—Administration of the national office of the Association and its staff, and proper disbursement of its funds including supervision of the budget and financial reports.
- ☐—Execution of business transactions on behalf of the Association and subject to the direction of the Board of Directors, contracts and agreements including notes, bonds, deeds, mortgages, leases, and other legal instruments without limitation.
- ☐—Maintenance of records of the official business, actions, and meetings of the Board of Directors and Executive Committee.
- ☐—Carrying out the policies and programs of the Board of Directors for presentment to the membership either through Association journals or at the annual convention or both.
- ☐—Action as public relations spokesperson, liaison, and representative for NSTA to other organizations, the press, business and industry groups, and government agencies.

Deadline for Applications: February 1, 1995

Qualifications: Science education background, teaching experience, management skills, good verbal and written communications skills and experience, financial planning abilities, negotiation skills, and sound leadership abilities. Knowledge of and experience working with the Washington, D.C., network, including government and private agencies, are extremely desirable.

Procedures: (1) Candidates submit applications for review by the NSTA Search Committee. (2) The Search Committee selects individuals for consideration and interviews the best potential candidates. (3) The top candidates resulting from these interviews are then

proposed to the NSTA Executive Committee for final interviews. (4) The NSTA Executive Committee will make the final selection and negotiate a contract with that individual for a term of five years subject to annual review. Interested candidates are invited to write for an application to NSTA Search Committee, 1840 Wilson Blvd., Arlington VA 22201-3000.

Timeline: Completed applications will be accepted until February 1, 1995. Final Candidates will be interviewed by the NSTA Executive Committee. Appointment date will be negotiated.

—Equal Opportunity Employer

The legacy of Norbert Wiener

Part II: Brownian motion and beyond

IN 1919 WIENER'S NOMADIC existence ended at last. He worked for a few months as a reporter for the *Boston Herald* and was fired. Finally, his father's friend Professor Osgood at Harvard interceded and obtained for Norbert an instructorship at MIT. In 1919, this was not a notable appointment. At the time, the mathematics department at MIT was purely a service department, valued only for its contribution to the engineering curriculum. Thus it is remarkable that MIT accommodated young Wiener, a man whose past experience did not recommend him as a teacher. In addition, even if MIT had sought prowess in mathematical research, Norbert Wiener in 1919 would not have been a strong candidate. He had published fifteen undistinguished articles on logic and nothing at all in traditional mathematics. But, whether MIT's decision to hire Wiener was guided by phenomenal insider information or was just a fortuitous product of the "old boy network," there can be no doubt that

Part I appeared in the November/December 1994 issue. Reprinted from the program booklet for The Legacy of Norbert Wiener: A Centennial Symposium in Honor of the 100th Anniversary of Norbert Wiener's Birth, October 8–14, 1994, prepared by the MIT Department of Mathematics with the assistance of Tony Rothman.

Wiener's appointment was a gamble that paid off for both parties. Wiener remained at MIT until his retirement in 1960, and during that period he not only put MIT on the map mathematically, he also played a profound part



Norbert Wiener in 1926.

in the creation of the culture to which MIT owes much of its present fame and prestige.

At MIT the prodigy bloomed. Perhaps his emergence was an expression of his having at last found in mathematics his true calling; maybe it was the sense of security and self-esteem that came with a steady job; or possibly it was simply that, at age 24, the ex-prodigy had caught up with himself and was ready to become a genius. In any case, during his first dozen years at MIT, Wiener made his most astounding contributions to pure mathematics: he constructed Brownian motion, laid a new foundation for potential

theory, and invented his generalized harmonic analysis.

The history of *Brownian motion* has taken some interesting twists and turns. The name honors the nineteenth-century botanist Robert Brown, who reported that pollen and many types of inorganic particles suspended in water perform a strange St. Vitus dance. Brown refuted some facile explanations of this motion, although debate still raged over whether the movement was of biological origin. It was Einstein's famous 1905 article on the subject that catapulted Brownian motion into twentieth-century physics. Einstein showed that a molecular (as opposed to a continuum) model of water

predicts the existence of the phenomenon that Brown observed. Interestingly, he predicted Brownian motion before learning about Brown's observations.¹

¹On page 17 of *Dynamical Theories of Brownian Motion* (Princeton University Press, 1967), Edward Nelson remarks, "It is sad to realize that despite all the hard work which had gone into the study of Brownian motion, Einstein was unaware of the existence of the phenomenon. He predicted it on theoretical grounds and formulated a correct quantitative theory of it." He quotes Einstein as saying, "My major aim . . . was to find facts which would guarantee as much as possible the existence of atoms of definite finite size."

Because it is virtually impossible to solve Newton's equations of motion for anything like the number of particles in a drop of water, Einstein adopted a statistical approach and showed that the evolution of the distribution of Brownian particles is governed by the heat equation. That is, the density of particles at each point follows the same physical law as the temperature at each point. Actually, from the physical point of view, this description of Einstein's paper throws out the baby with the wash. A physicist cannot talk about a one-size-fits-all *heat equation* any more than a one-size-fits-all *wave equation*; there are all-important constants that enter any physical equation. For the wave equation, the essential physical constant is the speed of light. In the case of the heat equation, there is the *diffusion constant*, and it was Einstein's formula for the diffusion constant that won his 1905 article its place in history. Namely, Einstein expressed the diffusion constant as the ratio of several physical quantities, one of which was Avogadro's number.² It turns out that, with the exception of Avogadro's number, all these quantities, including the diffusion constant itself, were either known or measurable experimentally. Thus, his formula led to the first accurate determination of Avogadro's number.

If one ignores physics and analyzes Einstein's model from a purely mathematical standpoint, what Einstein was saying is summarized by the following three assertions about the way in which Brownian particles move.

1. Brownian particles travel in such a way that the behavior over two different time intervals is independent. Thus, there is no way to predict future behavior from past behavior.

2. The particle is equally likely to move in any direction, and the

distance traversed by a Brownian particle during a time interval is on average proportional to the square root of the time.

3. The trajectories of Brownian particles are continuous.

With reasonably standard results from the modern theory of probability, one can deduce from Einstein's three assumptions the conclusion that the distribution of Brownian particles evolves according to a heat equation. (The all-important diffusion constant is determined by the proportionality constant in assertion 2.) Of course, in 1905, a mathematically satisfactory formulation of probability theory had yet to be given. Thus, Einstein's derivation was, mathematically speaking, rather primitive. Moreover, implicit in his model was an important mathematical challenge: *the verification that one can construct a distribution on the space of trajectories so that assertions 1, 2, and 3 are satisfied.*³

At the turn of the century, the French school of analysis was hard at work creating the subject that we now call *measure theory* (that is, the theory by which we assign volume to sets).⁴ The French school,

³Actually, Einstein's 1905 article was not the first one in which this problem appears. Five years earlier, H. Poincaré's brilliant student L. Bachelier came to the conclusion that the fluctuation of prices on the Paris Bourse follow trajectories whose distribution satisfies assertions 1, 2, and 3. It was not until the 1970s that the economics literature on this subject converged with the engineering and mathematical literature. The result is a much more sophisticated way to calculate risk in large financial markets, which has become an indispensable tool for loan, investment, and trading companies. Finally, one should remark that Bachelier, as distinguished from Einstein, really addressed the problem of computing the probability of nontrivial events that can be formulated only in the path-space context. The first physicist to address such problems was M. Smoluchowski, who used an approximation scheme based on random walks.

⁴Prior to their efforts, the only available theory was basically the one introduced by Archimedes, rediscovered by Fermat and Newton, and now forced on every calculus

especially E. Borel and H. Lebesgue, freed measure theory from its classical origins and made it possible to consider the problem of assigning *probabilities* to subsets of trajectories. However, in spite of their many magnificent achievements, neither Borel, Lebesgue, nor their disciples like P. Lévy, S. Banach, M. Fréchet, and A. N. Kolmogorov had been able to mathematically rationalize Einstein's model of Brownian motion. All of them were well aware of the essential problem, but none of them had been able to carry out the required construction. This was the problem that Wiener solved.

In hindsight, Wiener's strategy looks a little naïve. In particular, he completely circumvented the issues on which more experienced mathematicians had foundered. In a marvelous demonstration of the power of optimism, he supposed that the desired assignment of probabilities could be made and asked how this assignment would look in a cleverly chosen coordinate system. He then turned the problem around and showed that the coordinate description leads to the existence of the desired assignment. (This general line of reasoning is familiar to anyone who has ever solved a problem by saying "let x be the solution" and then found x as a consequence of the properties that it must have.) Wiener's Gordian-knot solution to the problem enhances its appeal, and the assignment of probabilities at which Wiener arrived in "Differential Space" has, ever since, borne his name. It is called *Wiener measure*.

The importance of Wiener measure is hard to exaggerate. It represents what we now dutifully call a *paradigm*. For one thing, its very

student. Of course, that theory had been tightened up by Cauchy, Riemann, and others, but it was still seriously deficient. For example, one could not show that the whole is the sum of its parts unless there were at most finitely many parts. In addition, although Riemann's theory served quite well in finite dimensional contexts, there was no theory at all for infinite dimensional spaces, like the space of all Brownian trajectories.

existence opened a floodgate and led Lévy, Kolmogorov, and others to create the theory of *stochastic processes*, thereby ushering in the modern theory of probability. In addition, Wiener measure is, in a sense that can be made very precise, as universal as the standard Gaussian (or normal) distribution on the real line: it is the distribution that arises whenever one carries out a central limit scaling procedure on path-space valued random variables.⁵ This is the underlying reason why Wiener measure arises as soon as one is studying a phenomenon that displays the properties 1, 2, and 3. It is also the reason why, again and again, Wiener measure comes up in models of situations in which one is observing the net effect of a huge number of tiny contributions from mutually independent sources—as in the motion of a pollen particle, the Dow Jones average, or, as Wiener himself observed, the distortions in a signal transmitted over a noisy line.

Although his construction of Brownian motion was Wiener's premiere achievement during the period, it was not his only one. In a sequence of articles from 1923 through 1925, Wiener also looked at a fundamental problem in the theory of electrostatics. The problem was to decide what shape electrical conductor can carry a fixed charge. Zaremba had shown that certain conductors in the shape of spikes are unable to carry charge—they discharge spontaneously at the tip. (The reverse of this phenomenon is what makes a lightning rod work.) On the other hand, Zaremba had shown that cone-shaped conductors do hold their charge. In the mathematical model spontaneous discharge corresponds to an abrupt change—a *discontinuity*—in the voltage across the interface between the conductor and the surrounding medium. The electrical field has a constant voltage on the conductor, and the equilibrium is stable (no

sparks) if the voltage is *continuous* across the interface.

Wiener described all shapes for which instability occurs and established a new framework for the entire subject of potential theory. In sharp contrast with many models in mathematical physics, he showed that the voltage in equilibrium is well defined mathematically, regardless of whether the conductor is stable or not. He then formulated a wholly original test, now known as the *Wiener criterion*, that determines at which points the voltage is discontinuous. A key step in Wiener's approach was to extend to arbitrary shapes a classical notion known as electrostatic capacity.⁶ He used a procedure that is analogous to, but more intricate than, the one invented by Lebesgue when he assigned a volume to regions for which there was no classical notion of volume. Indeed, Wiener's capacity is closely related to, but more subtle than, the measures used for fractals.⁷

Another topic that Wiener investigated during this period was what we now call distribution theory or the theory of generalized functions. Not long after Wiener arrived at MIT, Professor Jackson and other members of the electrical engineering department at MIT asked Wiener to develop a proper foundation for the Heaviside calculus—a calculus for solving differential equations by means of Fourier and Laplace transforms. Heaviside's calculus transforms a differential equation into an equation involving multiplication, as in $Ax = B$. To solve for x , we simply divide: $x = B/A$. The

⁶The electrostatic capacity of a conductor can be defined as the total charge carried by the conductor in equilibrium when the voltage difference between the conductor and its surroundings is fixed at, say, one hundred volts.

⁷There is an amusing irony associated with Wiener's investigations into potential theory. Namely, as S. Kakutani discovered in the early 1940s, potential theory is related to Brownian motion in deep and wonderful ways. Wiener completely missed this beautiful and useful connection with his previous work.

difficulty is that this easy formula for the solution then has to be transformed back into a meaningful statement about the solution to the original differential equation. This involves making sense of the inverse of the Fourier–Laplace transform. Wiener undertook the description of how multiplication and division correspond to the operations of differentiation and integration. Laurent Schwartz, the father of the theory of distributions, acknowledges that Wiener's treatment in 1926 anticipated all others by many years.

Just as the physics of Brownian motion had stimulated Wiener to profound new mathematics, so the practical problem of processing electrical signals led him to a deep extension of classical Fourier analysis. Fourier analysis consists of decomposing a periodic signal into a sum of pure sine waves. The fundamental formula of Fourier analysis—the Parseval formula—says that the total energy of the signal in each period is the sum of the energies of its pure waves. The collection of frequencies at which these amplitudes occur is known as the spectrum of the signal, and these come from a discrete list of values—the harmonics of a vibrating string. There is a similar fundamental formula due to Plancherel for the decomposition of nonperiodic waves that measures the total energy over all time. The spectrum of the signal is spread over the continuum of frequencies, and the formula measures the amount of energy of the signal concentrated in a given band of frequencies. The problem is that the signals that occur in practice in electrical systems do not fit into the frame of either of these theories. The signals are not periodic and the spectrum is not confined to a special list, so that Fourier series are inadequate. On the other hand, the total energy over an infinite time period is infinite, so that Plancherel's theory does not apply. Wiener overcame this difficulty with what he named generalized harmonic analysis. Wiener took as his starting place certain autocorrelation numbers, which compare the signal to the same signal with a time delay. These

⁵A full understanding of this universality came only in the 1950s and was provided by P. Lévy, R. H. Cameron, M. Donsker, P. Erdős, M. Kac, W. T. Martin, and I. E. Segal.

were precisely what could be measured in practice. Then, instead of dealing with total energy, Wiener considered the average energy of the signal over a long time interval. His theory was flexible enough to encompass both periodic signals and signals composed of a continuum of frequencies, such as "white noise."

One of the key ingredients in Wiener's generalized harmonic analysis was a new method to calculate limits of averages. His first step was to rephrase the problem so that it became one of determining when two different weighted averages are very close to each other. The recast problem fit into the general framework of so-called Tauberian theory—a theory to which Hardy and Littlewood had made several contributions. But instead of using some refinement of the techniques of his teachers, Wiener introduced a new approach that not only solved his own problem but revealed the fundamental mechanism of all previous problems

of this type.⁸ In his monograph on the subject, Wiener illustrates his ideas with an elegant proof of the Prime Number Theorem, one of the most beautiful applications of analysis to number theory.

With the publication of his work on generalized harmonic analysis and Tauberian theorems, Wiener's reputation was at last established. In 1932 he was promoted to Full Professor at MIT with a salary of \$6,000. The following year, he was elected to the National Academy of Sciences, and he won the Bôcher Prize, a prize given every five years for the best work in analysis in the United States.

The major works outlined above by no means exhaust Wiener's intellectual activity. Throughout the 1930s he continued to expand on harmonic analysis, with the same

⁸Wiener's work led to I. M. Gelfand's far-reaching formulation of a notion of spectrum that can be used to analyze multiplication and division in any algebraic system.

engineering applications clearly in view. He wrote an influential book with R. E. A. C. Paley and a seminal paper on integral equations with E. Hopf. He made excursions into quantum mechanics with Max Born and sortied into five-dimensional relativity (Kaluza-Klein theory) with Dirk Struik. In the late 1930s Wiener made a significant contribution to the mathematical foundations of statistical mechanics by extending G. D. Birkhoff's 1931 ergodic theorem. His 1938 paper "The Homogeneous Chaos," which undertakes to fathom nonlinear random phenomena, has descendants in constructive quantum field theory, under the name "Wick ordering." 

The concluding segment of this centenary essay will cover Wiener's work on the control of anti-aircraft fire during World War II and his most famous legacy—cybernetics.

TO BE CONTINUED
IN THE NEXT ISSUE

Norbert Wiener 1894–1964

by P.R. Masani

"...Masani has produced a valuable resource fitting for one of America's great figures."

—Historia Mathematica

"...an extraordinary mirror of the complex life of Norbert Wiener...excellent photos of those who worked with Wiener bring alive the excitement of groundbreaking collaborative discoveries."

—Journal of Interdisciplinary Studies

"...great book about a great mathematician, a second Leibniz, as the author calls him."

—Mathematical Reviews

A volume in the *Vita Mathematica* series

This biography traces Wiener's life, starting from his precocious childhood and training in mathematics, zoology, and philosophy and ending with his full maturity and death. His contributions to philosophy, science, and engineering, and his views on art and religion are outlined, as well as the effects of considerable interaction with other great minds of his time. Crucial parts of his correspondence with the Defense Department are published here for the first time. The biography is intended for the general reader as well as for the professional.

1990 416 Pages Hardcover

\$69.50 ISBN 0-8176-2246-2

For Orders and Information

Call 1 800 777-4643 or Fax (617) 876-1272
Write to Birkhäuser, Marketing Dept.
675 Massachusetts Ave
Cambridge, MA 02139

Birkhäuser

Boston • Basel • Berlin



Important components of learning components

You use vectors—but do you really understand them?

by Boris Korsunsky

FROM MY TEACHING EXPERIENCE (both in Russia and the United States) I strongly believe that many students have a hard time understanding the idea of vector physical quantities. In particular, the concept of components is especially hard for them. The worst of it is, many of these students sagely learn how to “follow the procedure” and are able to solve “standard” problems involving the idea of vector components without really understanding them. It’s funny—I have talked about this topic in my school with students taking Conceptual Physics, Intro Physics, and AP Physics C, and they all ask the same naïve questions! (Although the AP students are less aggressive—they rely on calculus . . .)

To prevent situations in which the teacher and the student are both convinced that the student actually *does*

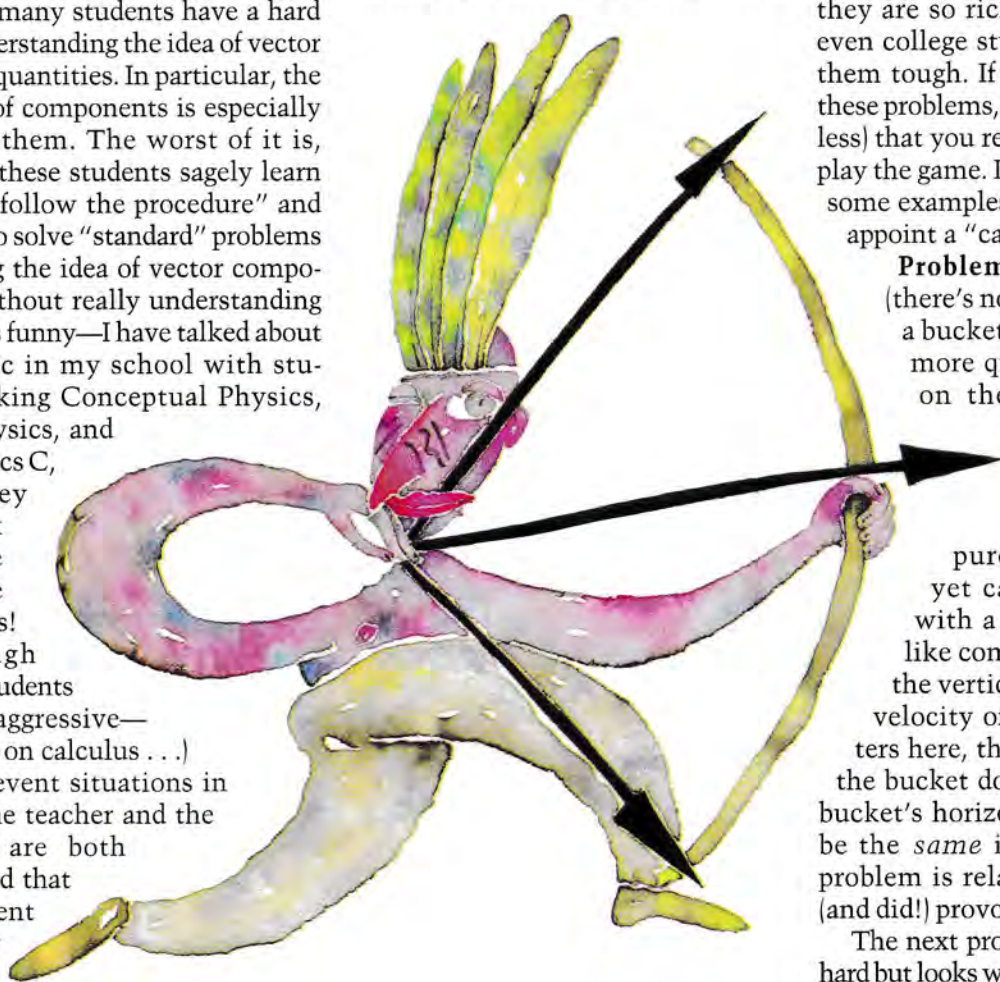
understand the idea while in fact he or she does *not*, I have used unusual, “tricky” problems that, as far as the

math is concerned, are accessible even for an introductory high school physics course. At the same time they are so rich conceptually that even college students find many of them tough. If you’re able to solve these problems, I can be sure (more or less) that you really do know how to play the game. In this article I’ll offer some examples that are sure to disappoint a “calculus person.”

Problem 1. It’s raining (there’s no wind, though). Will a bucket be filled with water more quickly if it’s resting on the ground or if it’s placed on a horizontally moving platform?

The question is purely qualitative and yet can be solved easily with a “quantitative” tool like components. Since only the vertical component of the velocity of the raindrops matters here, the time needed to fill the bucket doesn’t depend on the bucket’s horizontal speed and will be the *same* in both cases. This problem is relatively easy but can (and did!) provoke a nice discussion.

The next problem is also not that hard but looks weird to many students.



Art by Sergey Ivanov

Problem 2. A person is pulling a boat with a rope as shown in figure 1. At a certain moment the angle between the rope and the boat's velocity is θ . (You may say: "Well, first of all, that's impossible!" Is it?) The speed of the boat is v . Find the speed u at which the person must be pulling the rope at this moment.

Figure 1

This problem also makes use of the idea of components. The answer is $u = v \cos \theta$. If you caught the drift of the problem, you would say that the component of the boat's velocity along the rope equals the velocity of the rope (we assume the rope doesn't stretch—otherwise the problem would be pointless). Unfortunately, from my experience many students are totally convinced that in order to be able to deal with components in a particular problem, you *must* have two *perpendicular* coordinate axes. This problem clearly says, "No, you don't."

The next problem looks different, but it's actually quite similar to problem 2.

Problem 3. A bar propped against a wall begins to slide down (fig. 2).

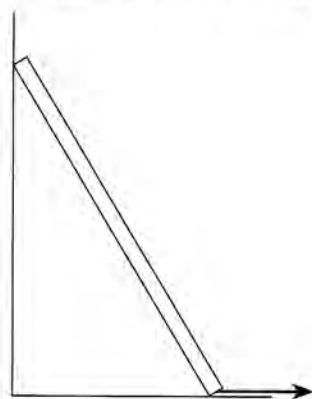


Figure 2

The velocity of the bottom end of the bar is given. Find the velocities of the top end of the bar and the middle of the bar graphically.

The way to solve this problem (which is admittedly a bit harder) is shown in figure 3. Since the bar is a rigid body, the components of the

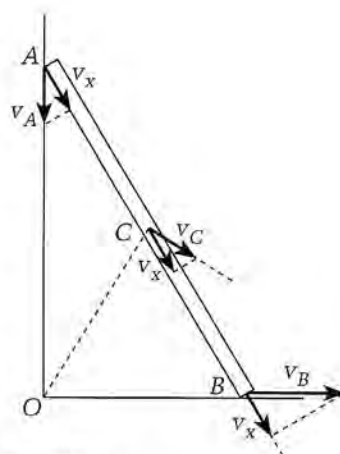


Figure 3

velocities of all points of the bar in the direction along the bar are the same. (Otherwise the distances between them would change.) If we know the component of a vector along a certain direction and the actual direction of the vector, we can easily find the vector itself. The direction of the velocity of the top end is obvious. How about the middle? Geometry tells us that as the bar slides down, the distance OC remains constant. This means that the midpoint C moves along an arc, and its velocity at all times is perpendicular to OC . Now we can "construct" the corresponding vector, as shown in figure 3.

The next problem is really tricky and I'm sure you'll enjoy it. (That is, unless you're *too good* at math, which might cause problems!)

Problem 4. Four ninja turtles are ready for battle, standing at points A, B, C, D forming a square, as shown in figure 4. At the same moment they start to chase one another: the velocity of turtle A (sorry—I can never manage to remember their wonderful names!) is directed at all times toward turtle B , whose velocity, in turn, is

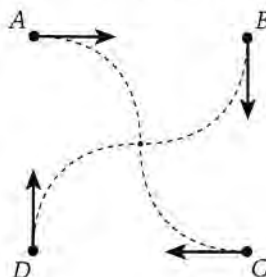


Figure 4

directed toward turtle C , who is chasing turtle D in the same manner. And turtle D is chasing turtle A , of course. They all have the same speeds, and it's pretty obvious that, moving in curved lines, they eventually come together at the center of the initial square $ABCD$. How long will it take if the side of the initial square is L and the speed of each ninja turtle is v ? (Bonus question: What's the point of such a contest?)

Isn't this a great problem? We certainly can't analyze these beyond-the-bounds-of-simple-math curves. Well, if you *can*, too bad—you'll miss all the fun! And the fun is to exploit the symmetry of the arrangement. At all times the turtles will form a square that decreases in size and simultaneously rotates. What a sophisticated motion! But the *center* of the square obviously does *not* move. And this is exactly where they meet—the point that interests us.

Now the components come into play. Although the direction of the velocity of each turtle changes continually, the component of the velocity of each turtle *directed toward the center* makes the same angle (45°) at all times with the velocity itself and, therefore, *retains its magnitude*, which is $v(\sqrt{2}/2)$. Now we get the answer right away. Isn't that great? (The answer is indeed L/v .)

Of course, components come in handy when we're faced with problems involving Newton's laws of motion. Here are a couple of nice examples.

Problem 5. The system shown in figure 5 is allowed to move freely from the state of rest with no friction. What will happen first: will block 1 hit the pulley, or will block 2 hit the wall?

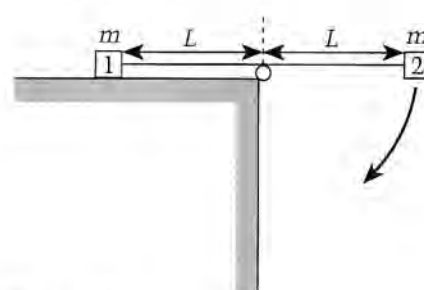


Figure 5

What can we do here? The direction and the magnitude of the force exerted on block 2 change continuously! Ready for some horrible integrating? Guess again!

This is one not-so-easy olympiad-style problem whose solution is amazingly short. Just consider the horizontal components. The force of tension of the string (which is certainly the same for both blocks) is the only one that contributes to the horizontal acceleration of both blocks. Of course, the horizontal component of this force is greater for block 1 *at all times*! Since both blocks have the same distance to go, block 1 will win the race. (The horizontal component of its velocity is *at all times* greater than that of block 2.)

The next (and last) problem brings in the idea of torques as well as components. (There's your hint!)

Problem 6. A uniform bar leans against a wall as shown in figure 6.

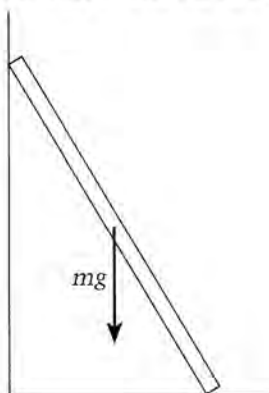


Figure 6

Given the fact that the wall is frictionless and given the vector representing the force of gravity acting on the bar, find the vector corresponding to the force of friction between the bar and the floor graphically. (Can the floor be frictionless, too?)

The solution is shown in figure 7. Two important ideas are involved. First, the net torque with respect to any point must be zero. Second, since the normal force of the wall and the force of gravity both "pass through" point A, the reactive force of the floor must also pass through the same point!

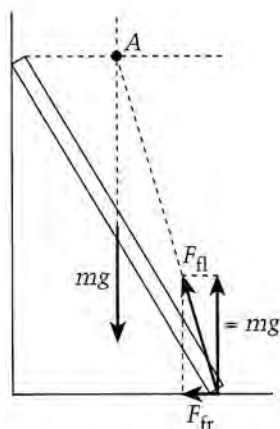


Figure 7

Now that we know the direction of this force, it's a good time to recall the fact that the vertical component of the floor's reactive force F_{fl} equals the force of gravity (which enables us to "construct" the vector corresponding to the floor's force). With this vector available to us, we can easily plot its horizontal component—which happens to be the unknown force of friction!

Tricky problems are a lot of fun and usually help us *really* understand a concept. I'll leave you with a few exercises. I'm sure you'll have a good time with them—eventually!

Exercises

1. A group of ants is pulling a small stick. At a certain moment the velocities of the ends A and B make the angles α and β , respectively, with the stick (fig. 8). The speed of end A is also given. Find the speed of end B.

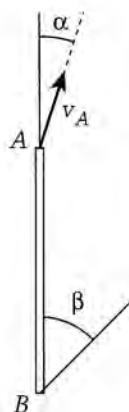


Figure 8

2. When the ants are done with the stick, they keep working hard. Now they are pulling a square piece of cardboard ABCD. At a certain moment it's known that the velocity of A equals v and is directed along AC. The velocity of C at this moment is

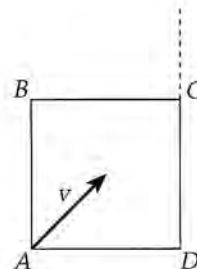


Figure 9

directed along CD (fig. 9). Find the velocities of B, C, and D.

3. When the ants finish this bit of work, they take a break. (You're welcome to do the same!)

After their siesta they pull a cardboard equilateral triangle ABC (fig. 10). It's known that at

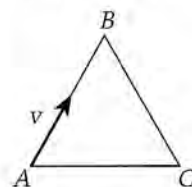


Figure 10

a certain moment the velocity of A is v and is directed along AB, whereas the velocity of C is directed along BC. Find the velocities of B and C.

4. Why is it easier to pull a nail out of a board if you turn it continuously while pulling? (Hint: consider the component of the force of friction, which acts against the force you exert in pulling.)

Boris Korsunsky teaches at Northfield Mount Hermon School in Northfield, Massachusetts.

ANSWERS, HINTS & SOLUTIONS
ON PAGE 59

If you have a question about your subscription...

If you would like to order back issues...

If you are interested in purchasing bulk issues for your classroom...

If you are moving...

Contact

Quantum

Springer-Verlag New York, Inc.
PO Box 2485

Secaucus NJ 07096-2485

Phone: 800 777-4643

(In New York, 201 348-4033)

Geometry in the pagoda

Classic problems of the great Japanese geometers

by George Berzsenyi

DURING ITS PERIOD OF ISOLATION, which included most of the 17th, 18th, and 19th centuries, Japan became a stronghold in the development of geometry. Its many skilled geometers recorded their findings as beautifully colored drawings on wooden tablets, which were presented in the shrines and temples as acts of devotion, to be hung under their roofs. Many of these have survived, while others are known only from their descriptions in handwritten books or from books printed from hand-carved wooden blocks prepared later. These often featured the solutions to the problems also, while the traditional tablets usually contain only the final results in an attractive visual form with the implicit challenge: "See if you can prove this!"

I first learned about the Japanese temple geometry problems (*san gaku*) in 1990, during the First International Congress of the World Federation of National Mathematics Competitions (WFNMC) at the University of Waterloo (Canada), where all participants were given a copy of a recently published book devoted to this topic. More recently, at the Second International Congress of the WFNMC (held in Potetz, Bulgaria, in 1994) I had the privilege of hearing a truly inspiring lecture by one of the authors of this book, Hidetosi Fukagawa. The problems posed below are based on his lecture.

Problem 1. Prove that in any $\triangle ABC$ and for any circle O tangent to AB and AC , the inscribed circle of $\triangle BFC$ and the inscribed circle of $\triangle ABC$ are tangent to side BC at the

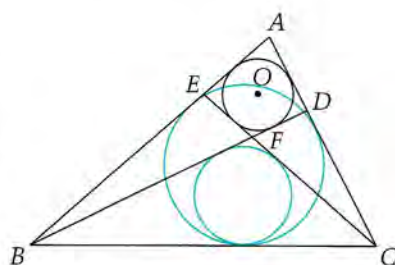


Figure 1

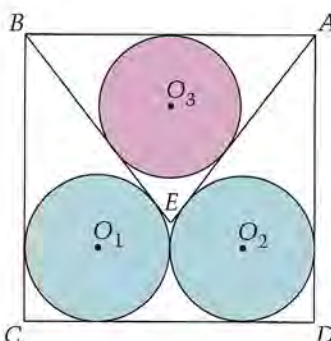


Figure 2

same point (fig. 1).

Problem 2. Prove that if circles O_1 and O_2 are equal and tangent to each other and to the sides of square $ABCD$, as shown in figure 2, then circle O_3 is of the same size.

Problem 3. Three circles, each of radius r , are inscribed in a square of base a , as shown in figure 3. Determine r in terms of a .

Problem 4. Four circles, each of radius r , are inscribed in an equilateral triangle of base a , as shown in figure 4. Determine r in terms of a .

In addition to clever arrangements of circles in squares, triangles, and lunar regions, some of the tablets also display ellipses and spheres in similarly

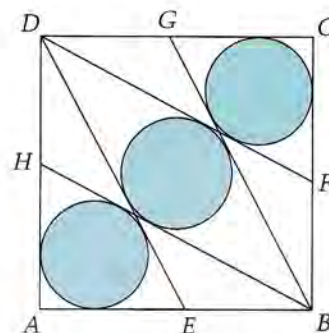


Figure 3

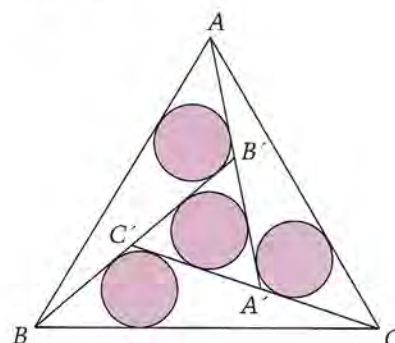


Figure 4

pleasing and challenging ways. The solutions presented in book form are also instructive, since inversion and other tools of modern geometry were not known to Japanese mathematicians in the 18th and 19th centuries.

For a more complete introduction to *san gaku*, the reader is referred to the English edition of Fukagawa's wonderful book *Japanese Temple Geometry Problems*, written with the cooperation of Dan Pedoe and published in 1984 by The Charles Babbage Research Centre (PO Box 272, St. Norbert Postal Station, Winnipeg R3V 1L6, Canada). ■

Programming challenges

Problems from the 1994 IOI

AS PROMISED IN THE LAST issue of *Quantum*, here are the problems from last summer's International Olympiad in Informatics, held in Stockholm, Sweden.

Day 1—problem 1

Figure 1 shows a number triangle. Write a program that calculates the highest sum of numbers passed on a route that starts at the top and ends somewhere on the base.

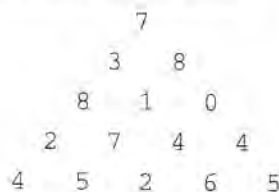


Figure 1

- Each step can go either diagonally down to the left or diagonally down to the right.

- The number of rows in the triangle is > 1 but ≤ 100 .

- The numbers in the triangle are all integers between 0 and 99 inclusive.

In the example above the route through 7, 3, 8, 7, 5 produces the highest sum 30.

Input data. Data about the number of rows in the triangle are first read from the INPUT.TXT file followed by the rows of the triangle. In our example, INPUT.TXT appears as follows:

```

5
7

```

```

3 8
8 1 0
2 7 4 4
4 5 2 6 5

```

Output data. The highest sum is written as an integer in the OUTPUT.TXT file. In our example this file would contain the number 30.

Day 1—problem 2

Figure 2 shows the map of a castle. Write a program that calculates

1. The number of rooms in the castle;

2. The size in modules of the largest room;

3. Which wall to remove from the castle (joining two existing rooms) to make as large a room as possible.

The castle is divided into a grid of square modules m rows by n columns ($m \leq 50$, $n \leq 50$). Each module can have from zero to four walls, inclusive.

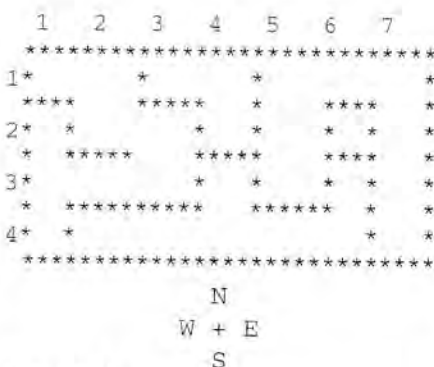


Figure 2

A string of *'s is a wall.

Input data. The map is stored in the INPUT.TXT file in the form of numbers, one for each module.

- The file starts with the number of modules in the north-south direction and the number of modules in the east-west direction.

- In the following lines each module is described by a number ($0 \leq p \leq 15$). This number is the sum of 1 (= wall to the west), 2 (= wall to the north), 4 (= wall to the east), 8 (= wall to the south). Inner walls are defined twice—a wall to the south in module 1,1 is also indicated as a wall to the north in module 2,1. The module at position 1,1 has a west, north, and south wall, so the sum is $1 + 2 + 8 = 11$.

- The castle always has at least two rooms.

INPUT.TXT for our example:

```

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

```

Output data. In the OUTPUT.TXT file, the following are written on three lines:

1. The number of rooms.

2. The area of the largest room counted in modules.

3. A suggestion of which wall to remove that will join two existing rooms to make as large a room as possible (first the row, then the column of the module next to the wall, and finally the compass direction that points to the wall). There may

be more than one but you need only choose one to display ("4 1 E" is one of several possibilities). In our example:

```
5
9
4 1 E
```

The last line refers to removing the wall pointed to below.

```

  1  2  3  4  5  6  7
*****
1*      *      *      *
****  *****  *      *****  *
2*  *      *      *      *      *
*  *****  *****  *****  *
3*      *      *      *      *
*  *****  *****  *****  *
4*->*      *      *      *
*****

```

Figure 3

Removing the east wall of module 4,1.

Day 1—problem 3

Figure 4 shows a square. Each row, each column, and the two diagonals can be read as a five-digit prime number. The rows are read from left to right. The columns are read from top to bottom. Both diagonals are read from left to right. Using the data in the INPUT.TXT file, write a program that constructs such squares.

```

1  1  3  5  1
3  3  2  0  3
3  0  3  2  3
1  4  0  3  3
3  3  3  1  1

```

Figure 4

- The prime numbers must have the same digit sum (11 in figure 4)
- The digit in the top left-hand corner of the square is predetermined (1 in figure 4).
- A prime number may be used more than once in the same square.
- If there are several solutions, all must be presented.
- A five-digit prime number cannot begin with zeros—that is, 00003 is not a five-digit prime number.

Input data. The program reads data from the INPUT.TXT file—first the digit sum of prime numbers and then the digit in the top left-hand

corner of the square. The file contains two lines. There will always be a solution to the given test data. In our example:

```
11
1
```

Output data. In the OUTPUT.TXT file, write five lines for each solution found, where each line in turn consists of a five-digit prime number. The above example has three solutions, which means that the OUTPUT.TXT file contains the following (the empty line between the different solutions is optional):

```

1 1 3 5 1
1 4 0 3 3
3 0 3 2 3
5 3 2 0 1
1 3 3 1 3

1 1 3 5 1
3 3 2 0 3
3 0 3 2 3
1 4 0 3 3
3 3 3 1 1

1 3 3 1 3
1 3 0 4 3
3 2 3 0 3
5 0 2 3 1
1 3 3 3 1

```

Day 2—problem 1

The nine numbers in the top portion of figure 5 represent the position of nine dials where each dial has one of four positions: north, or 12 o'clock (0); east, or 3 o'clock (1); south, or 6 o'clock (2); and west, or 9 o'clock (3) (as shown in the bottom portion of figure 5).

```

3  3  0
2  2  2
2  1  2

0
3--1
2

```

Figure 5

Dial positions.

There are nine different ways to turn the dials on the clocks. Each way is called a move. Each move is selected by a number from 1 to 9. That number will turn the dials marked by a "1" ninety degrees clockwise. Those marked with a "0" have no effect. The nine moves are displayed in figure 6.

1 1 0	1 1 1	0 1 1
1 1 0	0 0 0	0 1 1
0 0 0	0 0 0	0 0 0
Move 1	Move 2	Move 3
1 0 0	0 1 0	0 0 1
1 0 0	1 1 1	0 0 1
1 0 0	0 1 0	0 0 1
Move 4	Move 5	Move 6
0 0 0	0 0 0	0 0 0
1 1 0	0 0 0	0 1 1
1 1 0	1 1 1	0 1 1
Move 7	Move 8	Move 9

Figure 6

For example, the following sequence of moves has the corresponding affect on the dials (they all end up at 12 o'clock (0)):

```

3  3  0          3  0  0
2  2  2  Move 5 -> 3  3  3
2  1  2          2  2  2

          3  0  0
Move 8 -> 3  3  3
          3  3  3

          0  0  0
Move 4 -> 3  3  3
          0  3  3

          0  0  0
Move 9      0  0  0
          0  0  0

```

The problem is to write a program that will take any starting position for the clock and find the shortest sequence of moves that puts all the dials in the 12 o'clock (0) position.

Input data. Read nine numbers from

the INPUT.TXT file. The example above will have the input data file

```
3 3 0
2 2 2
2 1 2
```

Output data. Write to the OUTPUT.TXT file the shortest sequence of moves (numbers) that turns all the dials to the 0 (12 o'clock) position. In our example the OUTPUT.TXT file could look as follows:

```
5849
```

Only one solution is required.

Day 2—problem 2

A man arrives at a bus stop at 12:00. He remains there from 12:00 to 12:59. The bus stop is used by a number of bus routes. The man notes the times of arriving buses. The times when buses arrive are given with the following rules:

1. Buses on the same route arrive at regular intervals from 12:00 to 12:59.

2. Times are given in whole minutes from 1 to 59.

3. Each bus route has at least two buses arriving at the station between 12:00 and 12:59.

4. The number of bus routes in the test example will be ≤ 17 .

5. Buses from different routes may arrive at the same time.

6. Several bus routes can have the same time of first arrival and/or time interval. If two bus routes have the same starting time and interval, they are distinct and are both to be presented.

Find the fewest number of bus routes that must stop at the bus stop to satisfy the input data. For each bus route, output the starting time and the interval.

Input data. The input file INPUT.TXT contains a number n ($n \leq 300$) telling how many arriving buses have been noted, followed by the arrival times in ascending order.

Our example:

```
17
0 3 5 13 13 15 21 26 27
29 37 39 39 45 51 52 53
```

If two buses arrive at the same time, that time is listed twice.

Output data. Write a table to the OUTPUT.TXT file with one line for each bus route. Each line in the file give the time of arrival for the first bus and the time interval in minutes. The order of the bus routes does not matter. If there are several solutions, only one is required.

Our example gives

```
0 13
3 12
5 8
```

Day 2—problem 3

Consider the magic list of 5 numbers:

```
1 3 10 2 5
```

Any two number that are next to each other are considered neighbors. Also, the two end numbers 1 and 5 are neighbors, as if the list formed a circle of numbers. Starting with 2, we can form an unbroken sequence of integers from 2 to 21 using a single number in the list or by adding neighbors. Here is how the sequence is formed:

```
2, 3, 1 + 3 = 4, 5,
5 + 1 = 6, 2 + 5 = 7,
2 + 5 + 1 = 8,
5 + 1 + 3 = 9, 10,
2 + 5 + 1 + 3 = 11,
10 + 2 = 12, 3 + 10 = 13,
1 + 3 + 10 = 14,
3 + 10 + 2 = 15,
1 + 3 + 10 + 2 = 16,
10 + 5 + 2 = 17,
10 + 2 + 5 + 1 = 18,
5 + 1 + 3 + 10 = 19,
3 + 10 + 2 + 5 = 20,
1 + 3 + 10 + 2 + 5 = 21
```

You were given three numbers (n , m , and k) where

n = the length of the list of numbers,

m = the starting number,

k = the smallest possible value for a member of the list—that is, all numbers must be greater than or equal to k .

Your task is to choose n integers for the magic list where an unbroken sequence of all integers m , $m + 1$,

$m + 2$, ..., max can be generated, where max is as large as possible.

Input data. The INPUT.TXT file contains three integers (n , m , k). For our example, the file would be

```
5
2
1
```

Output data. The file OUTPUT.TXT must contain the following:

1. The highest number (max) that can be generated with the list of numbers.

2. All arrangements of numbers in a circle that produce a sequence from m to max (one per line). Each arrangement is a list of numbers starting with the smallest number (which is not necessarily unique).

Note: (2 10 3 1 5) is not a valid solution, since it does not start with the smallest number. (1 3 10 2 5) and (1 5 2 10 3) must both be included in the output. Note that (1 1 2 3), (1 3 2 1), (1 2 3 1) and (1 1 3 2) should all be output.

The output for our example would be

```
21
1 3 10 2 5
1 5 2 10 3
2 4 9 3 5
2 5 3 9 4
```

USACO: The 3rd Annual USA Computing Olympiad will be held from May 31 to June 6, 1995

IOI: The 7th International Olympiad in Informatics will be held in the Netherlands from June 26 to July 3, 1995

ICPSC: The 14th International Computer Problem Solving Challenge will be held on April 29, 1995

For more information about these programs, contact Donald T. Piele, USACO Director, University of Wisconsin-Parkside, Box 2000, Kenosha WI 53141-2000; e-mail: piele@cs.uwp.edu; phone: 414 595-2231 (O), 414 634-0868 (H).

Bulletin Board

A summer PROMYS in Boston

The annual Program in Mathematics for Young Scientists (PROMYS) will be held at Boston University from July 2 to August 12, 1995. PROMYS offers a lively mathematical environment in which ambitious high school students explore the creative world of mathematics. Through their intensive efforts to solve a large assortment of unusually challenging problems in number theory, the participants practice the art of mathematical discovery—numerical exploration, formulation and critique of conjectures, and techniques of proof and generalization. More experienced participants may also study abstract algebra, combinatorics, and the Riemann zeta function.

Problem sets are accompanied by daily lectures given by research mathematicians with extensive experience in Prof. Arnold Ross's long-standing Summer Mathematics Program at Ohio State University. In addition, a highly competent staff of 18 college-age counselors lives in the dormitories and is always available to discuss mathematics with students. Each participant belongs to a problem-solving group that meets with a professional mathematician three times a week. Special lectures by outside speakers offer a broad view of mathematics and its role in the sciences.

PROMYS is a residential program designed for 60 students entering grades 10 through 12. Admission decisions will be based on the following criteria: applicants' solutions to a set of challenging problems included in the application packet; teacher recommendations; high school transcripts; and student essays explaining their interest in the program. The estimated cost to par-

ticipants is \$1,300 for room and board. Books may cost an additional \$100. Financial aid is available. PROMYS is dedicated to the principle that no student will be unable to attend because of financial need.

PROMYS is directed by Prof. Glenn Stevens. Application materials can be obtained by writing to PROMYS, Department of Mathematics, Boston University, 111 Cummington St., Boston MA 02215, or by calling 617 353-2563. Applications will be accepted from March 1 until June 1, 1995.

The most personal computer

Many of us know more about the RAM, hard disks, and I/O devices in our desktop computer than we do about the neurons, blood vessels, and sensory mechanisms at work inside our own bodies. The same goes for the software that runs the two systems. If you're looking to optimize your mental performance, or if you're curious about the effects of certain chemical substances, or if you have trouble sleeping at night—in short, if you have any questions at all about the way your head functions, you will probably find an answer in *The Owner's Manual for the Brain* by Pierce J. Howard, Ph.D. Dr. Howard has distilled the results of current mind-brain research into a readable, well-organized compendium of tips and tidbits. And because many of the topics covered are still being studied, you will find the following tongue-in-cheek caveat scattered throughout the book: "Warning to Reader: Swallowing everything in this book hook, line, and sinker could be hazardous to your health."

The book begins with a primer on cognitive science, then goes on to explore such topics as sex-based differences, the aging process, nutrition, chemical agents, sleep, left- and right-

handedness, emotions, temperament, intelligence, motivation, ergonomics, the senses, memory, problem solving, creativity, and communication (with many subtopics in between). Each chapter concludes with a list of sources for further reading. The book is fun to browse in, and it also offers pointers to those who want to delve into a particular subject.

(*The Owner's Manual for the Brain: Everyday Applications from Mind-Brain Research* by Pierce J. Howard, Ph.D. 400 pp., \$19.95 pbk, \$29.95 hbk + \$2 S&H for first book, \$1 per book thereafter. Quantity discounts available. To order, call 800 945-3132, or write to Publication Services, 8870 Business Park Dr., Austin TX 78759.)

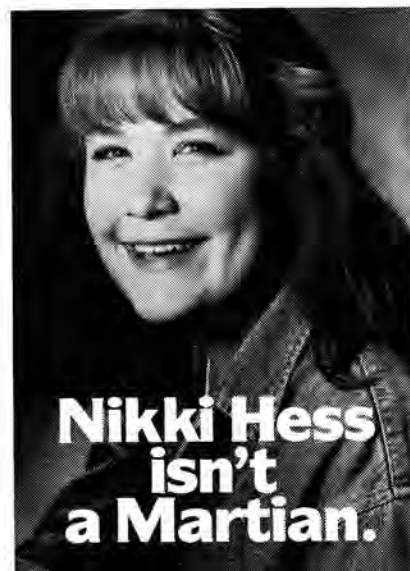


Photo by Charles Bush

But the way some kids treat her, she might as well be from another planet. Just because she has epilepsy.

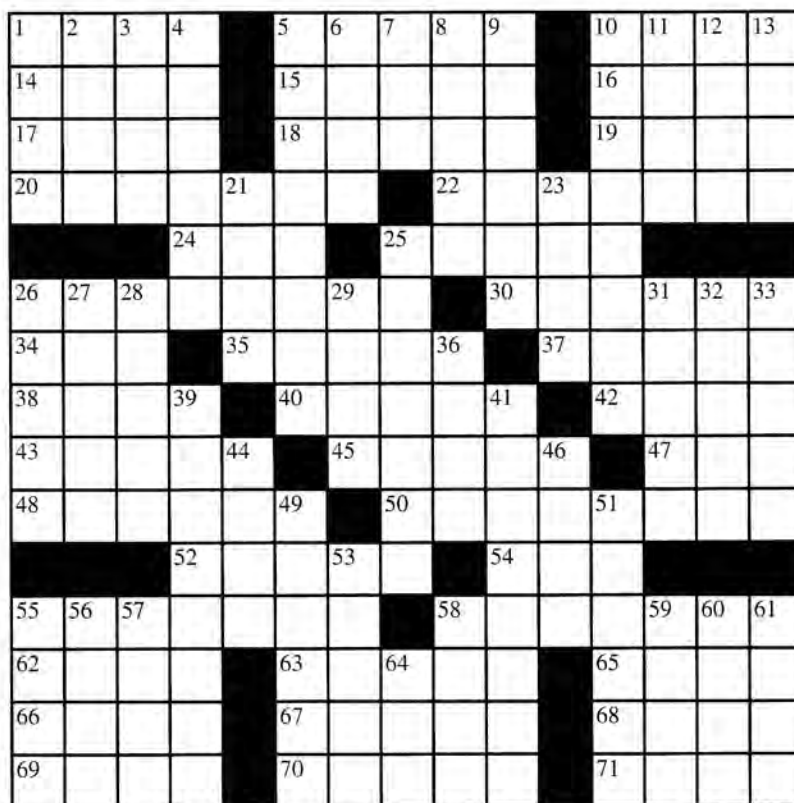
Epilepsy doesn't make her weird. It doesn't affect her abilities, her sense of humor, or her qualities as a friend.

Kids with epilepsy or any disability. Let's count 'em in.

Get the facts. Write or call The Epilepsy Foundation of America, 1-800-EFA-1000. Or contact your local EFA affiliate.

Crisscross science

by David R. Martin



Across

- 1 Remove ovaries
- 5 Scorches
- 10 60,362 (in base 16)
- 14 Exhaust
- 15 Fabry-___ interferometer
- 16 Verse segment
- 17 Singer James
- 18 Mournful song
- 19 Egyptian river
- 20 Fundamental science
- 22 ___ collision (K.E. conserved)
- 24 Infant
- 25 Build
- 26 10^{-10} meter
- 30 Afternoon nap
- 34 10^5 pascals
- 35 Golden
- 37 Indian official
- 38 Employee
- 40 Aluminum or copper, e.g.
- 42 One billionth: pref.
- 43 Periodic ___
- 45 Band
- 47 Poet's amongst

- 48 Shirt part
- 50 Hydrated sodium zirconium silicate
- 52 Length unit
- 54 Hawaiian wreath
- 55 ___ relativity
- 58 ___ cycle (thermodynamic cycle)
- 62 Birth control advocate
- ___ Guttmacher
- 63 Linear accelerator
- 65 Eye part
- 66 Fasting period
- 67 Atmosphere layer
- 68 Paper mulberry bark
- 69 Calculator display lights
- 70 Nostrils
- 71 Large plasma ball

Down

- 1 ___-down transformer
- 2 Medulla
- 3 Pretentiously creative
- 4 Various fungi

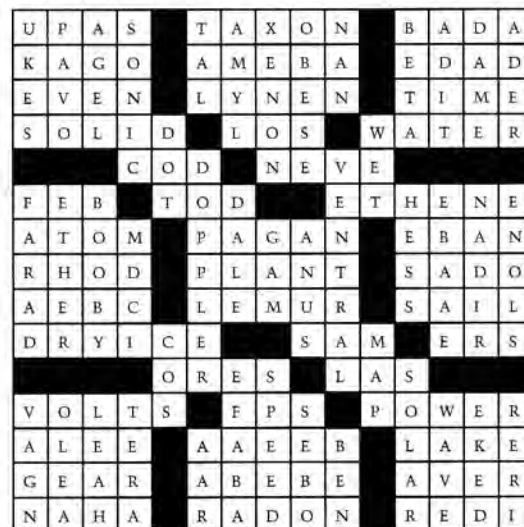
- 5 Range of frequencies
- 6 Conger and Moray, e.g.
- 7 Exist
- 8 Medical researcher ___ Guillemin
- 9 Pistil portions
- 10 Great physicist
- 11 ___-and-switch
- 12 951
- 13 44,012 (in base 16)
- 21 Theta follower
- 23 Electron pair acceptor
- 25 Transistor part
- 26 Borders upon
- 27 Of the nose
- 28 Diving bird
- 29 Unrefined metals
- 31 Hindu teacher
- 32 Phoenician goddess
- 33 Battery terminal
- 36 ___ Sagan
- 39 43A members
- 41 ___ equation (for electrostatic potential)
- 44 Always
- 46 Ship's landing place

- 49 Parallel mirrors
- 51 Number parts
- 53 Girl's name
- 55 ___ bladder
- 56 Gen. Robert ___
- 57 Logic operator
- 58 Rational

- 59 QED word
- 60 Palm tree
- 61 Russian ruler
- 64 Logic operator

SOLUTION IN THE
NEXT ISSUE

SOLUTION TO THE NOVEMBER/DECEMBER PUZZLE



ANSWERS, HINTS & SOLUTIONS

Math

M131

The answer is $x = 37$. Putting $f(x) = x^2 + ax + b$, $g(x) = x^2 + px + q$, from the condition we find that $111a + 3b = 111p + 3q$, or $37a + b = 37p + q$. But this means that 37 is the unique root of the (linear) equation $f(x) = g(x)$.

We could as well consider this problem for any three numbers x_1, x_2, x_3 instead of 1, 10, 100. Then the answer would be $x = (x_1 + x_2 + x_3)/3$ —the arithmetic mean of the given numbers. This answer becomes clearer if we notice that the problem on the whole is “linear” rather than “quadratic” because the squares cancel out in the equation we used to solve the problem, and even in the problem’s given condition.

M132

The answer is 11 points. First let’s prove that this score is sufficient to get into the final four. Suppose it is not—that is, the tournament may result in five teams scoring no less than 11 points each. Then the total score of these teams is no less than 55. On the other hand, their total score in the 10 games between themselves is 20, and their total score in the games with the other three teams is at most $5 \cdot 3 \cdot 2 = 30$. This is a contradiction: $20 + 30 < 55$.

It’s easy to design a tournament in which 10 points don’t guarantee entering the finals. For instance, five of the teams can all play to a draw among themselves and beat the other three teams. Then each of them scores $4 + 6 = 10$ points but may not get into the final four.

In the same way it can be shown that in a one-round tournament the

minimum score that ensures getting into the k best of n participants ($k < n$) equals $2n - k - 1$.

M133

If a gangster is chased by an infinite number of his “buddies,” we choose this infinite set of “buddies.” This satisfies our requirement.

Suppose each gangster is chased by a finite number of others. Take any gangster g_1 and send to prison all those who want to rub him out and also the one whom g_1 is chasing. In the remaining infinite set choose any gangster g_2 , send to prison those who want to kill him and the gangster he wants to kill (g_1 and g_2 will stay free), and proceed by induction: after n gangsters $g_1, \dots, g_n$ are chosen, we still have an infinite set of gangsters none of whom is chased by the first n , so we can choose g_{n+1} from them. So the process can be continued indefinitely to yield the required set.

M134

Consider a circle of length 99 partitioned into 99 unit arcs. Mark ten of the endpoints of these arcs such that the lengths of the 10 arcs defined by these points are equal in order to the 10 numbers around the first decagon. Similarly, construct a second circle representing the second decagon. Now lay the second circle over the first so as to fit the two initial fine partitions together.

Consider 99 rotations of the second circle through multiples of $360^\circ/99$. We will show that one of these rotations makes two pairs of marked points coincide. For if, after each of the rotations, at most one marked point on the second circle coincides with a marked point on the first circle, then there are no more than 99 coincidences in all.

But that’s impossible, because each of the 10 marked points on one circle must match each of the 10 marked points on the other circle exactly once. So the total number of coincidences must be 100. Therefore, there is a rotation that brings two pairs of marked points into coincidence. The arcs between these points on both circles are the same length, which means that the corresponding sums of successive numbers are equal.

M135

We’ll give only a solution to the more general problem (b). It will be based on the following useful fact that often helps prove the concurrency of straight lines.

CARNOT’S THEOREM. *Three perpendiculars drawn through points A, B, C to the sides B_0C_0 , C_0A_0 , A_0B_0 , respectively, of a triangle ABC meet at a point if and only if $(A_0C^2 - CB_0^2) + (B_0A^2 - AC_0^2) + (C_0B^2 - BA_0^2) = 0$.*

To prove this, notice that, by the Pythagorean theorem, for any point P the difference of squares $A_0P^2 - PB_0^2$ is equal to the similar difference $A_0P_1^2 - P_1B_0^2$ for the projection P_1 of P onto A_0B_0 (fig. 1). Since the position of point P_1 on line A_0B_0 is uniquely determined by the value of the second difference (the proof of this is left to the reader), it follows that the locus of points P with a constant value of $A_0P^2 - PB_0^2$ is a line through P_1 perpendicular to A_0B_0 . Now, if the perpendiculars in the

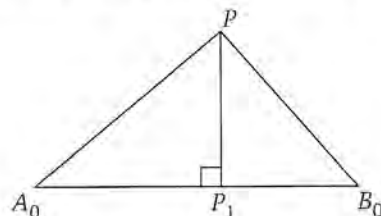


Figure 1

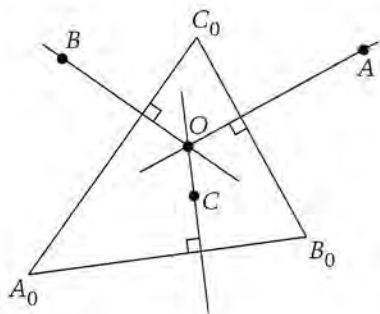


Figure 2

theorem do meet at a point O (fig. 2), we can replace all the points A, B, C in the equation with O without changing the value of its left side, after which all the terms there cancel out to yield zero. Conversely, if the equation is true and two of the perpendiculars—say, through A and B —meet at Q , then we can replace A and B with Q in the equation, which yields $A_0C^2 - CB_0^2 = A_0Q^2 - QB_0^2$. But this means that Q lies on the third perpendicular as well.

So our aim will be to find a triangle $A_0B_0C_0$ such that the lines in the problem are perpendicular to its sides and Carnot's equation is verifiable.

Construct point C_0 such that C_0AC_1 and C_0BC_1 are right angles and, similarly, points A_0 and B_0 (fig. 3). Let's prove that B_0C_0 is perpendicular to AA_2 . Extend C_1A by a segment $AD = C_1A$ (fig. 4). Since $\angle B_1AB_0 = \angle DAC_0 = 90^\circ$, the 90° rotation

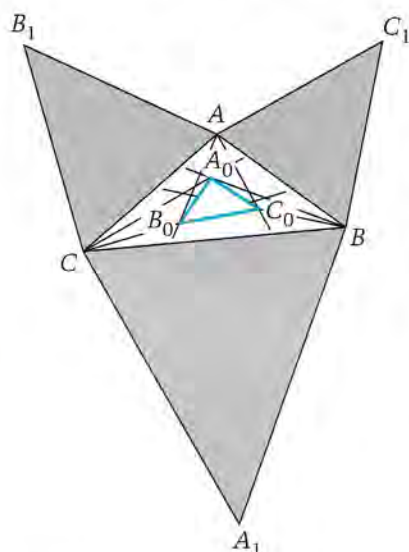


Figure 3

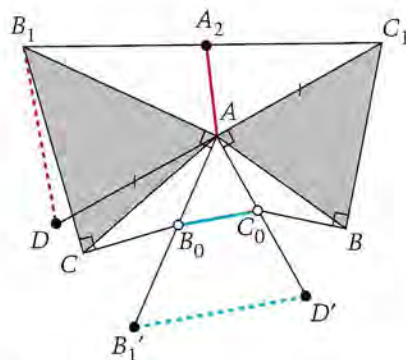


Figure 4

about A takes points B_1 and D into points B_1' and D' on AB_0 and AC_0 , respectively. Of course, $B_1D \perp B_1'D'$. But at the same time $B_1D \parallel AA_2$ (AA_2 connects two midpoints in triangle B_1C_1D), and $B_1'D' \parallel B_0C_0$ (because $AB_0/AB_1' = AB_0/AB_1 = AC_0/AC_1 = AC_0/AD = AC_0/AD'$). So $AA_2 \perp B_0C_0$. Similarly, BB_2 and CC_2 are perpendicular to the other two sides of triangle $A_0B_0C_0$.

Now it remains to check Carnot's condition, which is trivial, because points A_0, B_0, C_0 are equidistant from the endpoints of the corresponding sides of the given triangle:

$$(A_0B^2 - BC_0^2) + (C_0A^2 - AB_0^2) + (B_0C^2 - CA_0^2) = 0.$$

(V. Dubrovsky)

Physics

P131

Let's consider the problem in the reference frame where the car is at rest initially and point B moves with a constant velocity v . To reach point B as quickly as possible the car must move along a straight line with a uniform acceleration a . The direction of the line is determined by the condition that the meeting with point B occurs at a certain point D . For the triangle ABD (fig. 5), setting $AB = b$, we obtain

$$b^2 + (vt)^2 = \left(\frac{at^2}{2}\right)^2.$$

Solving the equation, we find

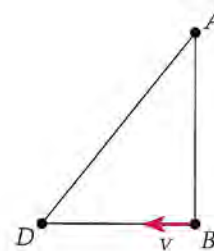


Figure 5

$$t = \sqrt{\frac{2v^2}{a^2} + \sqrt{\left(\frac{2v^2}{a^2}\right)^2 + \frac{4b^2}{a^2}}} = 50 \text{ s}.$$

Since the optimal motion of the car is uniformly accelerated, its trajectory relative to the Earth will be a parabola.

The strategy for reaching point C as quickly as possible is similar. However, to obtain the time necessary to reach point E (fig. 6), we must solve a fourth-degree equation of the general form

$$b^2 + (vt + c)^2 = \left(\frac{at^2}{2}\right)^2,$$

where $c = BC$, or

$$t = 4\sqrt{\frac{4}{a^2}(b^2 + (vt + c)^2)}. \quad (1)$$

We can solve this equation approximately by an iteration method. In essence, we must solve an equation like $f(t) = t$. As a first approximation we take an arbitrary value $t = t_0$ and then compose the sequence

$$t_1 = f(t_0), t_2 = f(t_1), \dots, t_n = f(t_{n-1}).$$

If this sequence has a limit S and the function $f(t)$ is continuous at the point S , then S is the root of equation (1).

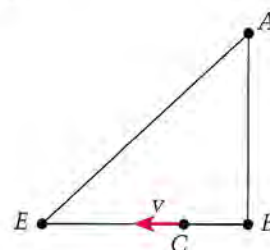


Figure 6

If we begin with $t_0 = 0$, we obtain the following sequence of values:

$$\begin{aligned} t_1 &= 41.2 \text{ s}, \\ t_2 &= 53.6 \text{ s}, \\ t_3 &= 57.5 \text{ s}, \\ t_4 &= 58.7 \text{ s}, \\ &\vdots \\ t_\infty &= 59.2 \text{ s}. \end{aligned}$$

Notice that a precision of 1% is achieved at $n = 4$, and that any two successive steps (that is, calculating the next two members in the sequence) improve the precision by a factor of two.

If the end point is located arbitrarily, there might be either a plus or minus sign in front of $c = BC$. To solve equation (1) in this case, the iteration method must be applied with caution as the result may be ambiguous and depend on the proper choice of the initial approximation.

And one more note. To drive along a parabolic trajectory with a constant vector acceleration surely demands a good bit of driving skill.

P132

According to Kepler's third law,

$$\frac{T^2}{a^3} = \text{constant},$$

and consequently, the smaller the radius of the spaceship's orbit, the shorter its period of revolution T around the Sun. The minimum period of revolution corresponds to the minimum orbital radius—that is, the radius of the Sun:

$$a_{\min} = R_S = \frac{\alpha}{2} R_{SE},$$

where R_{SE} is the distance between the Sun and the Earth. Now let's compare the motions of the spaceship and the Earth around the Sun:

$$\frac{T_{\min}^2}{a_{\min}^3} = \frac{T_E^2}{R_{SE}^3}.$$

So, remembering that the period of revolution of the Earth $T_E = 365.25$ days, we obtain

$$\begin{aligned} T_{\min} &= T_E \left(\frac{\alpha}{2} \right)^{3/2} \\ &\approx 0.116 \text{ day} \approx 2 \text{ hr } 47 \text{ min}. \end{aligned}$$

P133

If the thickness of the atmosphere is small relative to the planet's radius (we should check this is at the end of our calculations), we can use the following formula for the mass of oxygen:

$$m = \frac{4\pi R^2 P}{g}.$$

The acceleration due to gravity g on the surface of the planet is obtained from the law of universal gravitation:

$$g = \frac{GM}{R^2}.$$

Then

$$m = \frac{4\pi R^4 P}{GM} \approx 5 \cdot 10^{17} \text{ kg}.$$

This is approximately equal to $1.5 \cdot 10^{19}$ moles.

Decomposition of each molecule of carbonic acid yields one molecule of oxygen, which means that the process will take approximately $1.5 \cdot 10^{13} \text{ s} \approx 500,000$ years. Not so long . . .

For a rough estimate of the thickness of the atmosphere, let's find the density of the oxygen near the planet's surface:

$$\rho = \frac{PM}{RT} \approx 0.4 \text{ kg/m}^3.$$

With this density, the thickness of the homogeneous atmosphere will be

$$h = \frac{m}{4\pi R^2 \rho} \approx 30 \text{ km}.$$

The thickness is in fact several times greater (density decreases with altitude), but even the corrected value is still much less than the radius of the planet.

P134

The center of mass of the frame is located at a distance $2L/3$ from the

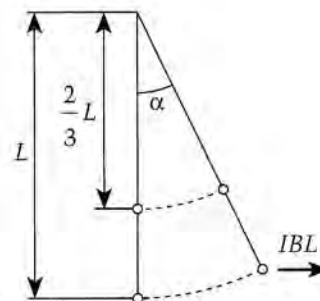


Figure 7

axis of revolution (fig. 7). Magnetic forces act on each conductor, but only the force applied to the horizontal conductor turns the frame (the other forces try to deform it). Let's label the angle of deflection α . Then the center of mass will be raised by $2L(1 - \cos \alpha)/3$. During the horizontal displacement $L \sin \alpha$, the magnetic force IBL performs the work $IBL^2 \sin \alpha$. It is this work that is equal to the change in the potential energy of the frame:

$$2\rho L^2 g(1 - \cos \alpha) = IBL^2 \sin \alpha.$$

From this the angle of deflection can be found, but we need to recall some trigonometry, particularly the half-angle formulas:

$$\frac{\sin \alpha}{1 - \cos \alpha} = \cot \frac{\alpha}{2} = \frac{2\rho g}{IB},$$

$$\alpha = 2 \arctan \frac{IB}{2\rho g}.$$

P135

Let a ray from the source strike the first lens at point D (fig. 8). After refracting it passes through point A , so we obtain the ray DA . Then we add the second lens to the system, and

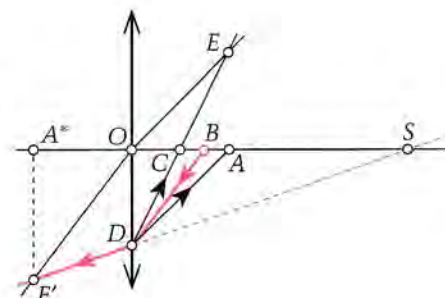


Figure 8

after refracting the ray DA becomes the ray DC . This enables us to construct the focal point of the second lens. We draw an auxiliary line passing through point O parallel to the ray DA —it intersects the extension of DC in the focal plane of the second lens, and hence point E lies in the focal plane (in figure 8 point E is directly above point A —it doesn't really matter in principle).

Now, using the focal point for the second lens, we can find the position of the source. Let's remember the reversibility of rays. We draw a ray from point B to point D —after refraction it must pass through point S (the source). It's possible that the source will be crossed not by the ray itself but by its extension.

Let's construct point A^* —the left-hand focal point of the second lens, which is symmetric to point A —and draw an auxiliary ray through point O parallel to the ray BD . This ray will pass through point E' belonging to the focal plane, and this point E enables us to construct the further path of the refracted ray—that is, DE' . The source is located on the extension of the refracted ray.

Figure 8 shows that this is an "imaginary" source—the lenses were struck by a beam of rays that converged at point S . If point C were farther from points A and B , we would obtain an ordinary "real" source.

Brainteasers

B131

The only possible letters are A (the number 1), B (2), I (11), J (12), S (21), and T (22). The only meaningful word corresponding to the given number is STATIST.

B132

The answer is $3,201 \times 3,201 = 10,246,401$. Since $FIVE \times V = 0$, $V = 0$. Since the products of FIVE by F, I, E have four digits each, $F \leq 3$, and $I, E \leq 2$. And since the product has eight digits, $FIVE > 3,200$. Therefore, $FIVE = 3,201$.

B133

Shadows on the Earth are less sharp because, as a rule, light scattered by the atmosphere enters into the shadows and makes them lighter. As for shadows on the Moon, which has no atmosphere, they can be illuminated only by rays that are reflected from other lunar objects. This relatively weak source of illumination has almost no effect on our visual perception of the objects.

B134

The answer is yes. Our grid consists of seven columns of three nodes each. Such a column can be colored in $2^3 = 8$ ways (using two colors). If two of the columns have the same coloring, take either of them. It contains at least two nodes of the same color, and these nodes together with the nodes on the same lines in the other column constitute the required four. If any two columns are colored differently, then only one of the eight possible colorings is not used, so in one of the columns all three nodes are the same color—say, red. Since all the colorings except one have been used, there is also a column with two red nodes. These nodes, along with the corresponding nodes in the first column, make up a one-color rectangle. (V. Dubrovsky)

B135

The perimeter of the rosette consists of arcs of the smaller circles. Consider one of these arcs—say, major arc AB in figure 9. Draw radii OA_1 and OB_1 of the big circle through A and B . The length of the arc A_1B_1 is

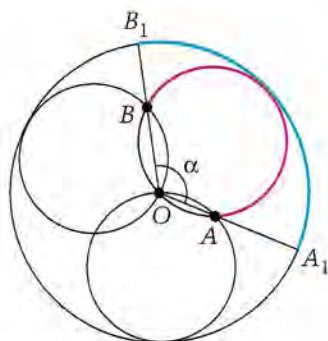


Figure 9

equal to $2r\alpha$, where $\alpha = \angle AOB$ and r is the radius of the smaller circles. The length of the arc AB is $2\alpha r$, because this arc subtends the inscribed angle AOB of measure α , and so the corresponding central angle is 2α . Therefore, the arcs AB and A_1B_1 are the same length. Adding up all such arcs, we see that the perimeter of the rosette is exactly equal to the big circumference.

Note that the result does not depend on how the smaller circles are arranged within the larger one (so long as they pass through point O) nor on how many there are.

Kaleidoscope

1. No, because in this case the electric and magnetic fields are constant and not interconnected with each other.

2. Usually a car radio receives the direct signal from a broadcast station, in which the electric field is vertically polarized. To obtain the maximum power from the input signal, the receiving antenna should also be vertical.

3. The short waves propagate long distances due to repeated reflections from the Earth's surface and the conducting layer of the atmosphere—the ionosphere. This results in areas where the signal can't be heard.

4. In the absence of direct solar radiation, the ionization of molecules in the ionosphere decreases. This enhances the reflectivity of the ionosphere and helps the radio waves propagate longer distances, thus increasing the transmission range of the radio stations.

5. Seawater strongly absorbs electromagnetic waves.

6. To determine the distance between the Moon and the Earth.

7. The ionosphere is transparent for the ultrashort waves used in TV broadcasts, and the waves do not diffract very well around objects on the ground.

8. Due to energy exchange resulting from thermal radiation.

9. Yes, it does.

10. Glass absorbs both infrared

and ultraviolet radiation.

11. The ultraviolet radiation of natural vegetation and that of camouflage are different, so they affect the photographic film differently.

12. The energy of X rays cannot exceed the binding energies of the electrons.

13. Yes, they can, because the wavelengths of gamma rays are even smaller.

Microexperiment. It reflects the infrared radiation emitted by the spiral filament.

Toy store

1. Any rearrangement of chips in a triangle of size n can be represented as a succession of swaps of adjacent chips. Two adjacent chips a and b can be swapped by an operation similar to the one shown in figures 1 and 2 in the article, but to apply it we must first create an isolated six-chip triangle containing chips a and b . This can be done by detaching triads from the given big triangle, one by one, until there is enough empty space for the required six-chip triangle with a and b to be detached (perhaps via the shift operation described in the article—see figure 2 there).

As to shifts, we know that a triad (the case $n = 2$) or a six-chip triangle (the case $n = 3$) can be shifted into any location on the grid. Figure 10a shows that for $n = 4$ a triangle of size n can be divided into two triads and a four-chip "diamond," which can be shifted by three spaces along any grid line as shown in figure 10b. A triangle of size $n = 3k + 2$ can be divided into k horizontal three-row layers (starting from its base) and a triad on top of them. Each of these layers can be divided into triads and, perhaps, a six-chip triangle (fig. 11). So we disassemble our big triangle into triangles of sizes 2 and 3, move these blocks apart, and then assemble them back on any desired new place. The same procedure works for $n = 3k$, except that the top block of the assembly will be a six-chip triangle rather than a triad. In the case $n = 3k + 1$ the top block is

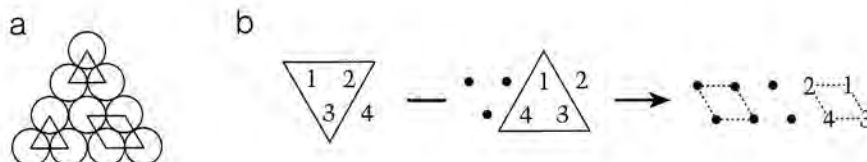


Figure 10

a triangle of size 4, and we've seen that it can be moved any number of spaces divisible by three; so this is true for the entire triangle as well. Other shifts in this case are impossible by the preservation of the invariants s_x and s_y (see the article).

2. For the triangle, $s_y = 1$; for the parallelogram, $s_y = 0$; so the answer is no.

3. A proof that a triangle of any of these sizes is indeed invertible can be based on figures 12–14. Figure 12 indicates how to invert a triangle



Figure 11

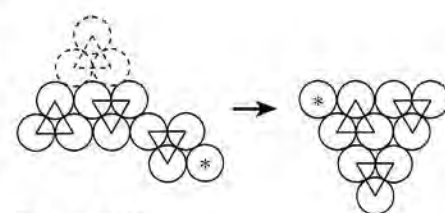


Figure 12

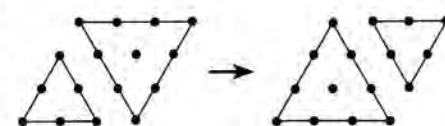


Figure 13

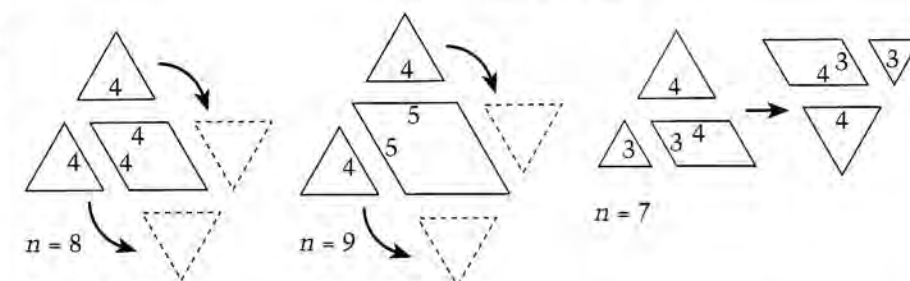


Figure 14

Inverting triangles with 8, 9, and 7 chips to a side. The numbers in the figure are the numbers of chips along the corresponding sides; in the case $n = 7$, the parallelogram is shifted during the rearrangement.

with four chips to a side (a "4-triangle"). In figure 13 we show how a 3-triangle can be inverted with the help of an auxiliary 4-triangle. A "rhombus" with four chips to a side can be split into a 3-triangle and a 4-triangle in two different ways, so that the orientations of the triangles of either size in these two partitions are opposite. The 4-triangle can be arbitrarily oriented and the 3-triangle can be arbitrarily shifted to join or part with the 4-triangle. Note that a single 3-triangle can't be inverted. Figure 14 indicates how we can divide a triangle into "parallelograms" and smaller invertible triangles.

4. The answer is no. This follows from comparing the invariants d_0 , d_1 , d_2 for the two configurations.

5. The operation given in the article consists of a swap of two balls ($1 \leftrightarrow 3$) and a 5-cycle ($4 \rightarrow 6 \rightarrow 9 \rightarrow 8 \rightarrow 5 \rightarrow 4$). If we repeat it five times, we get the swap without the 5-cycle. If we repeat it six times, we get the 5-cycle without the swap. We can similarly obtain other swaps of edge balls—say, $5 \leftrightarrow 8$. Then, to obtain a swap of two adjacent edge balls—for instance, $9 \leftrightarrow 8$ —we perform the above 5-cycle so as to bring balls 9 and 8 to the places 8 and 5, where we can swap them, after which we bring them back to places 9 and 8 (in the opposite order) using the inverse of our 5-cycle.

Bus and puddles

1. 10, 84.

$$2. \binom{m+n}{m} = \binom{m+n}{n}$$

$$3. \binom{9}{3} - \binom{4}{1}\binom{4}{1} = 84 - 16 = 68$$

4. Suppose, for example, that $x \in A_1 \cap A_2$. Then x is counted once each for $|A_1|$ and $|A_2|$ and is subtracted once for $|A_1 \cap A_2|$, so its presence is counted exactly once.

5. The number N of ways is

$$\begin{aligned} N &= |U| - \sum_i |A_i| + \sum_{i,j} |A_i \cap A_j| \\ &\quad - \sum_{i,j,k} |A_i \cap A_j \cap A_k| \\ &\quad + \sum_{i,j,k,l} |A_i \cap A_j \cap A_k \cap A_l| \\ &= \binom{10}{4} - \left[\binom{2}{1}\binom{7}{3} + \binom{4}{2}\binom{5}{4} + \binom{6}{2}\binom{3}{2} \right. \\ &\quad \left. + \binom{5}{4}\binom{4}{3} \right] \\ &\quad + \left[\binom{2}{1}\binom{5}{4} + \binom{2}{1}\binom{3}{1}\binom{3}{2} \right. \\ &\quad \left. + \binom{2}{1}\binom{4}{3} + 0 + 0 + 0 \right] \\ &\quad - 0 + 0 \\ &= 210 - 165 + 36 \\ &= 81. \end{aligned}$$

6. The ellipse looks like a hexagon (see figure 15).

7. Since $50 = 2 \cdot 5^2$, then

$$\begin{aligned} \phi(50) &= 50 - \left(\frac{50}{2} + \frac{50}{5} \right) + \frac{50}{10} \\ &= 20; \end{aligned}$$

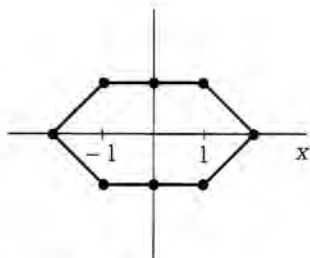


Figure 15

since $105 = 3 \cdot 5 \cdot 7$, then

$$\begin{aligned} \phi(105) &= 105 - \left(\frac{105}{3} + \frac{105}{5} + \frac{105}{7} \right) \\ &\quad + \left(\frac{105}{15} + \frac{105}{21} + \frac{105}{35} \right) - \frac{105}{105} \\ &= 48; \end{aligned}$$

since $210 = 2 \cdot 3 \cdot 5 \cdot 7$, then

$$\phi(210) = 210 - 247 + 101 - 17 + 1 = 48.$$

8. Consider the collection of the four sons—

$$S = \{\text{son 1, son 2, son 3, son 4}\}$$

—and let the universal set U represent the $4!$ permutations of the four sons. Thus,

$$U = \{(i, j, k, l): 1 \leq i, j, k, l \leq 4, i \neq j \neq k \neq l\}.$$

The ordering in each quadruple is important—for instance, (i, j, k, l) corresponds to son i 's card in the first envelope, son j 's card in the second, son k 's card in the third, and son l 's card in the fourth.

Next, we define four subsets A_i , $i = 1, 2, 3, 4$, of U , where A_i consists of all permutations of sons with son i 's card located in the i th envelope. Thus, for example, $A_1 = \{(1, j, k, l): 2 \leq j, k, l \leq 4, \text{ and } j, k, l \text{ are all different}\}$. We note that each set A_i contains $3! = 6$ elements, and each intersection of two different sets $A_i \cap A_j$ contains $2!$ elements.

The set of derangements that Grandma is interested in is precisely the set D_4 :

$$D_4 = U - (A_1 \cup A_2 \cup A_3 \cup A_4).$$

For if $(i, j, k, l) \in D_4$, then definitely $i \neq 1$, $j \neq 2$, $k \neq 3$, and $l \neq 4$. This

means that each son's card is not in its proper envelope. The question asked in exercise 8 concerns the cardinality of D_4 . Referring to the the inclusion-exclusion theorem, we get

$$\begin{aligned} |D_4| &= |U| - \sum_i |A_i| + \sum_{i,j} |A_i \cap A_j| \\ &\quad - \sum_{i,j,k} |A_i \cap A_j \cap A_k| \\ &\quad + \sum_{i,j,k,l} |A_i \cap A_j \cap A_k \cap A_l| \\ &= 4! - \binom{4}{1}3! + \binom{4}{2}2! - \binom{4}{3}1! + \binom{4}{4}0! \\ &= 9. \end{aligned}$$

Finally, note that the above expression

$$4! - \binom{4}{1}3! + \binom{4}{2}2! - \binom{4}{3}1! + \binom{4}{4}0!$$

can be rewritten as

$$4!(1 - 1/1! + 1/2! - 1/3! + 1/4!).$$

Extending this, we note that if Grandma Jones had had five sons, there would have been $5!(1 - 1/1! + 1/2! - 1/3! + 1/4! - 1/5!) = 44$ ways to replace the cards without a single card in its proper place.

Components

1. Since the stick is a rigid body, its length doesn't change. So the components of v_B and v_A along the stick must be equal. This immediately leads to the answer: $v_B = v_A \cos \alpha / \cos \beta$.

2. Since the square is a rigid body, the distances between any two points remain unchanged. The component of v_C along AC equals v_A .

Index of Advertisers

American University in Moscow	22
Birkhäuser	44
Cross Educational Software	28
NSTA Special Publications	29
U.S. Air Force Academy	Cover 4

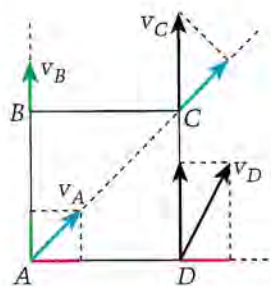


Figure 16

This fact, along with the given direction of v_C , enables us to reconstruct vector v_C (see figure 16).

Vector v_B is directed along AB (otherwise BC would change—look at v_C !). Its component along AB —that is, the vector itself—equals the component of v_A along AB . For v_D we know its components along CD (it's v_C) and along AD (it's the component of v_A along AD). Adding them as vectors leads to the answer: $v_B = v_A\sqrt{2}/2$; $v_C = v_A\sqrt{2}$; $v_D = v_A\sqrt{2.5}$.

Note: "adding them as vectors" usually works when you need to reconstruct a vector given its components. But it works only because the given components are usually perpendicular to one another. But we can find (and work with) a component of a given vector along any direction! And if we have components along directions that are not at right angles to one another, we need to use a more general (but not more difficult) method that would work for any case (including the "traditional" one).

3. The answer is $v_C = v_A$; $v_B = 2v_A$. It's easy to find v_C —its component along AC is the same as that of v_A . Finding v_B is more interesting. Although we know its component along AB (it equals v_A) and BC (it

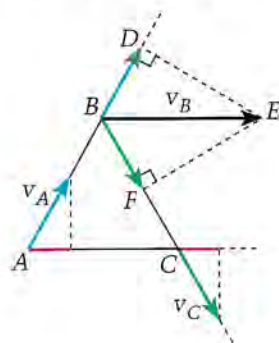


Figure 17

equals v_C), the reconstruction might be a problem. Look at figure 17: v_B is not a vector sum of its components. We draw DE perpendicular to AB , and FE perpendicular to BC (recall the definition of a component). Vector BE (whose length can easily be found, right?) represents v_B .

4. The force of friction is always directed opposite the velocity of an object. If we turn the nail while pulling it out, the velocities of each point of the nail are directed at a certain angle to the desired direction, and so the component of the force of friction acting against the direction of pulling is smaller (see figure 18).

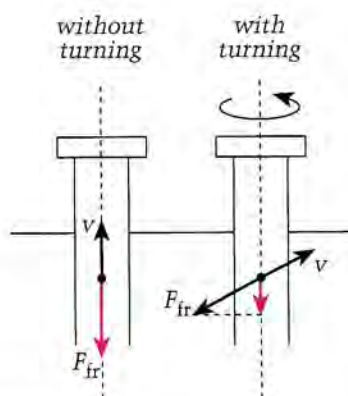


Figure 18

No calculus

(See "Look, Ma—No Calculus!" in the last issue)

1. Consider the (parallel) rays of sunlight as seen from Syene and Alexandria at noon on midsummer day (see figure 19). Since $\alpha = \beta$, we conclude that the ratio of AS to the circumference of the Earth is equal

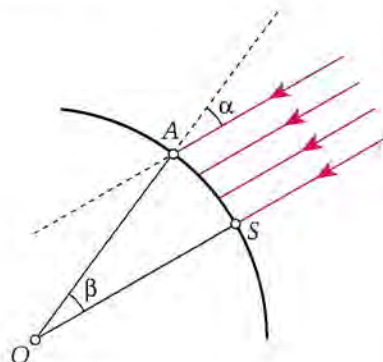


Figure 19

to $7.5/360$ —that is, $500/C = 7.5/360$, or $C = 48 \cdot 500 = 24,000$ miles. With $\pi \approx 3$, this leads to $R \approx 24,000/(2 \cdot 3) = 4,000$ miles.

2. Your spreadsheet calculations are of the form

$$N(t+1) - N(t) = 0.1N(t) - 0.0005N(t)^2 \\ = 0.0005N(t)[200 - N(t)].$$

This shows that $N(t)$ is increasing for $0 < N(t) < 200$ and decreasing for $N(t) > 200$.

IMO problems

(See the Happenings department in the last issue)

1. We can assume that $a_1 > a_2 > \dots > a_m$. We claim that $a_i + a_{m+1-i} \geq n+1$ for $1 \leq i \leq m$. If $a_i + a_{m+1-i} \leq n$ for some i , then $a_i < a_i + a_{m+1-i} \leq n$, $a_{m+1-i} < a_i + a_{m+1-i} \leq n$. Hence each of the i numbers $a_i, a_{m+1-i}, \dots, a_i + a_{m+1-i}$ is a different one from $a_1, a_2, \dots, a_{i-1}$. This is impossible. It follows that

$$2(a_1 + a_2 + \dots + a_m) \\ = (a_1 + a_m) + (a_2 + a_{m-1}) \\ + \dots + (a_m + a_1) \\ \geq m(n+1),$$

or

$$\frac{a_1 + a_2 + \dots + a_m}{m} \geq \frac{n+1}{2}.$$

2. First, assume that OQ is perpendicular to EF (fig. 20a). Now $OEBQ$ and $OCFQ$ are cyclic quadrilaterals. Hence $\angle OEQ = \angle OBQ = \angle OCQ = \angle OFQ$. It follows that $\triangle OEQ \cong \triangle OFQ$ and $QE = QF$. Now, suppose that $QE \neq QF$. Suppose the perpendicular through O to EF meets BC at $Q' \neq Q$ (fig. 20b). Draw the line through Q' parallel to EF such that it meets lines AB and AC at E' and F' , respectively. Then $Q'E' = Q'F'$ as before. Let AQ' meet EF at N . Then $N \neq Q$ and $NE = NF$, so that $QE \neq QF$. This is a contradiction, unless Q and Q' coincide—that is, OQ is perpendicular to EF .

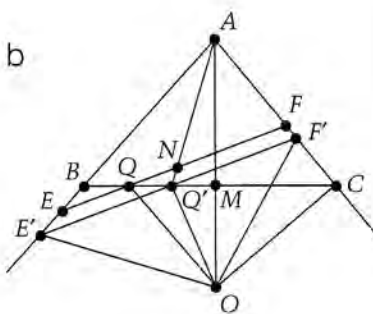
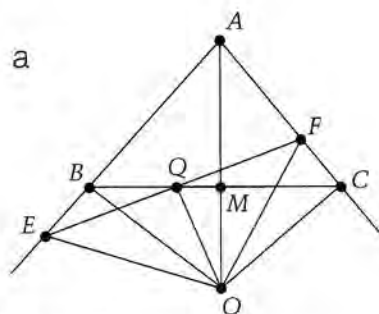


Figure 20

3. Note: the statement of the problem contained an error. In the first sentence, the last element in the set should be $2k$ (not $3k$, as printed).

(a) Let g_k denote the number of elements in the set $\{1, 2, \dots, k\}$ whose base 2 representation has precisely three 1's. Clearly, $f(k)$ and $g(k)$ are nondecreasing functions, and $f(k) = g(2k) - g(k)$. Hence

$$\begin{aligned} f(k+1) - f(k) &= g(2k+2) - g(k+1) - [g(2k) - g(k)] \\ &= g(2k+2) - g(2k) - [g(k+1) - g(k)]. \end{aligned}$$

Now the binary representation of $2k+2 = 2(k+1)$ is obtained from that of $k+1$ simply by appending a zero on the right side of the numeral. Hence either both $2k+2$ is counted in $g(2k+2)$ and $k+1$ is counted in $g(k+1)$, or neither is. It follows that $f(k+1) - f(k) = 1$ or 0 , depending on whether $2k+1$ is counted in $g(2k+2)$. In any case, $f(k)$ does not skip over any positive integers. Note that $g(2^n) = g(2^n - 1) = \binom{n}{3}$, so that $f(2^n) = g(2^{n+1}) - g(2^n) = \binom{n+1}{3} - \binom{n}{3} = \binom{n}{2}$. Hence $f(k)$ is not bounded from above. It follows that the range of $f(k)$ is the set of all non-negative integers. For any positive integer m , $f(k) = m$ has at least one solution.

(b) Suppose $f(k) = m$ has a unique solution. Then $f(k+1) - f(k) = 1 = f(k) - f(k-1)$. The former holds if and only if $2k+1$ is counted in $g(2k+2)$. Equivalently, there are precisely two 1's in the base 2 representation of k . The same holds for $k-1$. This is possible if and only if the last digit of $k-1$ is 1, the second-last digit is 0, and there is precisely one other digit

that is 1. In other words, $k = 2^n + 2$ for some integer $n \geq 2$. Now

$$\begin{aligned} f(2^n + 2) &= g(2^{n+1} + 4) - g(2^n + 2) \\ &= 1 + g(2^{n+1}) - g(2^n) \\ &= 1 + \binom{n}{2}. \end{aligned}$$

It follows that the set of positive integer m for which $f(k) = m$ has a unique solution is $\{1 + \binom{n}{2} \mid n \geq 2\}$.

4. Note that $mn - 1$ and m^3 are relatively prime. Then $mn - 1$ dividing $n^3 + 1$ is therefore equivalent to $mn - 1$ dividing $m^3(n^3 + 1) = m^3n^3 - 1 + m^3 + 1$, which in turn is equivalent to $mn - 1$ dividing $m^3 + 1$. If $m = n$, we have $(n^3 + 1)/(n^2 - 1) = n + 1/(n - 1)$. This is an integer if and only if $n = 2$. We now consider the case $m > n$. If $n = 1$, then $2/(m - 1)$ is an integer. This is so if and only if $m = 2$ or 3 . Suppose $n \geq 2$. Note that $n^3 + 1 \equiv 1 \pmod{n}$, while $mn - 1 \equiv -1 \pmod{n}$. Hence if $(n^3 + 1)/(mn - 1)$ is an integer, then it equals $kn - 1$ for some positive integer k . Now $kn - 1 < (n^3 + 1)/(n^2 - 1) = n + 1/(n - 1)$, or $(k - 1)n < 1 + 1/(n - 1)$. Hence $k = 1$, so that $n^3 + 1 = (mn - 1)(n - 1)$. This yields $m = (n^2 + 1)/(n - 1) = n + 1 + 2/(n - 1)$, which is an integer if and only if $n = 2$ or 3 . In either case, we have $m = 5$. In short, there are nine solutions—namely, $(2, 2)$, $(2, 1)$, $(3, 1)$, $(5, 2)$, $(5, 3)$, $(1, 2)$, $(1, 3)$, $(2, 5)$, and $(3, 5)$, the last four obtained by symmetry.

5. Condition (ii) implies that $f(x) = x$ has at most three solutions: one in $(-1, 0)$, one equal to 0, and one in $(0, \infty)$. Suppose $f(u) = u$ for some u in $(-1, 0)$. Setting $x = y = u$ in condition (i), we have $f(u^2 + 2u) = u^2 + 2u$. Since $0 < u + 1 < 1$, $u^2 + 2u = (u + 1)^2 - 1$ is also in $(-1, 0)$. Hence $u^2 + 2u = u$; but then u is not in $(-1, 0)$. Then $f(v) = v$ for some v in $(0, \infty)$

leads to a similar contradiction. However, $f(x + (1 + x)f(x)) = x + (1 + x)f(x)$ for all x in S . We must have $x + (1 + x)f(x) = 0$, so that $f(x) = -x/(1 + x)$. We now prove that this function does have the desired properties. Clearly, $f(x)/x = -1/(1 + x)$ is strictly increasing on S . For all x and y in S we have

$$y + (1 + y)f(x) = y - \frac{x(1 + y)}{1 + x}$$

$$= \frac{y - x}{1 + x},$$

and

$$f(x + (1 + x)f(y)) = f\left(\frac{x - y}{1 + y}\right)$$

$$= -\frac{\frac{x - y}{1 + y}}{1 + \frac{x - y}{1 + y}}$$

$$= \frac{y - x}{1 + x}.$$

6. Let A be the set of all positive integers of the form $q_1 q_2 \dots q_{q_1}$, where $q_1 < q_2 < \dots < q_{q_1}$ are primes. In other words,

$$\begin{aligned} A &= \{2 \times 3, 2 \times 5, 2 \times 7, \dots\} \\ &\cup \{3 \times 5 \times 7, 3 \times 5 \times 11, \dots\} \\ &\cup \{5 \times 7 \times 11 \times 13 \times 17, \dots\} \\ &\cup \dots \end{aligned}$$

For any infinite set $S = \{p_1, p_2, p_3, \dots\}$ of primes with $p_1 < p_2 < p_3 < \dots$, we can satisfy the requirement of the problem by taking $k = p_1$, $m = p_1 p_2 \dots p_k$, and $n = p_2 p_3 \dots p_{k+1}$.



The National Arbor Day Foundation

96,000 acres of irreplaceable rain forest are being burned every day. Join The National Arbor Day Foundation and support Rain Forest Rescue to help stop the destruction. Call now.

**Call Rain Forest Rescue NOW.
1-800-255-5500**

Triad and true

An intrepid trooper in facing down puzzling invariants

by Vladimir Dubrovsky

THIS ARTICLE WAS WRITTEN soon after the 1994 problem-solving workshop of the International Mathematics Tournament of the Towns, which, like a year before, again took place in Beloretsk, Russia, last August. This is a very special, perhaps unique event that combines features of a summer school, an olympiad, and a "research institute" for high school students. (To get a better idea of what happens there, read "A Tale of One City" in the May/June 1994 issue, which covers the previous workshop at Beloretsk.) I was invited there as a member of the jury and proposed an extended problem that is known to the readers of our Toy Store department as the *triads* puzzle (see the November/December 1993 and January/February 1994 issues). I expected that the problem would prove sufficiently attractive—indeed, even before I had time to finish the presentation, I noticed the swirl of coins modeling the puzzle on some desks in the class—but I didn't expect to receive such wonderful results. The first few warm-up questions (those posed in *Quantum*) were answered in practically no time (to be exact, in a

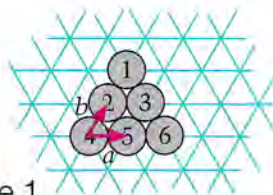


Figure 1

Basic triangle of six chips, their numbering, and the corresponding grid.

day). And even much more difficult and far-reaching extensions received exhaustive and beautiful, if not absolutely impeccable, solutions. So I'd like to share some of these remarkable findings with you.

Let me remind you that the game of triads, whose original version belongs to Sergey Grabarchuk, consists in rearranging six round chips that initially form a triangle with three chips to a side by moving them in small triangles of three chips touching each other ("triads"); a triad is only allowed to be slid along the board "parallel" to its initial position; after a triad arrives at its new place, new triads emerge, and the process is continued. We'll assume from the beginning that the chips are numbered 1, 2, ..., 6 as in figure 1, and consider only the arrangements in

which the centers of all chips fit on the nodes of the triangular grid in this figure. (This is convenient and, in fact, nonrestrictive.)

Permutations and shifts

One of the first warm-up questions about the triads proposed at the workshop was *whether all of the $6! = 720$ possible permutations of the chips can be created by moving triads*. The answer—which is yes—was already given in previous *Quantum* issues. It was even shown that we can obtain any permutation without shifting the big triangle as a whole. Another, much better solution was found by an American participant in the workshop, Joseph Shaeffer, a student at the Oak Tree School in Charlotte, North Carolina. His five-move operation, shown in figure 2, swaps a corner piece with an adjacent one, leaving the entire big triangle in its initial location. A proper rotation or reflection of this operation can swap any corner piece with any of its two neighbors. And it is not hard to show that these swaps suffice to create any permutation.

Along with a number of other participants, Joseph also found an

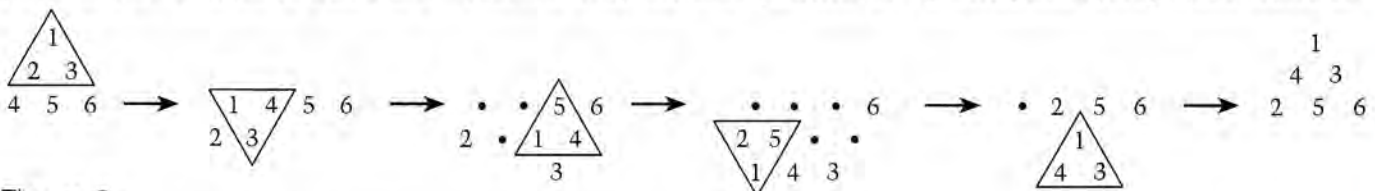


Figure 2

"Shaeffer's swap." Triangular frames indicate the triads we're about to move.

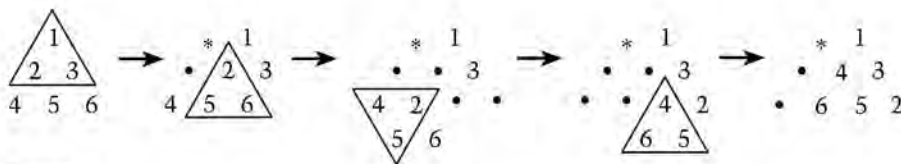


Figure 3

Unit shift.

operation (fig. 3) that moves the whole big triangle horizontally by one position—by the vector $\mathbf{a} = \overrightarrow{45}$ (fig. 1)—between two neighboring nodes of the grid. Reflecting this operation in the bisector of the angle 146 of our triangle, we obtain the shift by $\mathbf{b} = \overrightarrow{42}$. And, combining the two operations and their inverses, the big triangle can be shifted by any vector $n\mathbf{a} + m\mathbf{b}$ with integer n and m . This is an exhaustive answer to the question about shifts that was left open in the January/February 1994 Toy Store article. So the big triangle can be moved to any position on the grid, and even without changing the relative order of its chips, because any rearrangement accompanying $\mathbf{a}$ can be erased using the swaps described above.

An invariant

Since whatever rearrangements we wished to obtain so far turned out to be possible, the impression may develop that this will continue forever—that is, any two arrangements of equally many chips on our grid (each containing a triad, of course) can be transformed into one another. However, when we try to reproduce our results for a bigger triangle of chips (which is an obvious generalization), we run into an obstacle.

Exercise 1. Show that for a triangle with n chips to a side any permutation of chips is possible. But the set of possible shifts depends on n : for n of the form $3k$ or $3k + 2$ ($k = 1, 2, \dots$) the triangle can be moved by the vector $m\mathbf{a} + l\mathbf{b}$ with any integer m and l , and for $n = 3k + 1$ only by “triple” vectors $3m\mathbf{a} + 3l\mathbf{b}$.

That some shifts are impossible for some n seems to be a more difficult part of this problem. It requires a new approach—one not unfamiliar

to our readers, though. Such things can often be proved by means of *invariants*, the technique described in “Some Things Never Change” in the September/October 1993 issue. Let’s see what it means in this particular case.

Number the horizontal rows of the grid $\dots, -2, -1, 0, 1, 2, \dots$ as in figure 4. Consider an arbitrary set of chips on the grid and the sum of the numbers of the rows on which they sit. When a triad is moved, the three numbers under it clearly all change by the same amount, so the total sum changes by a multiple of three. This means that this sum modulo 3—that is, its remainder when divided by three—remains constant. Such unchanging values are called invariants. We’ll denote our invariant s_y , reserving the notation s_x for a similar invariant obtained from numbering another, slanting set of grid lines (fig. 4).

These invariants offer a clue to the question about shifts in exercise 1. If a set of N chips is moved horizontally—say, by a vector $\mathbf{a}$ —its s_y obviously doesn’t change, but its s_x changes by $N \pmod{3}$. So the $\mathbf{a}$ -shift is possible only for N divisible by 3; otherwise, we can do only the shifts by multiples of $3\mathbf{a}$. The same is true for slanting shifts. Now complete the solution yourself, and also try the following exercises.

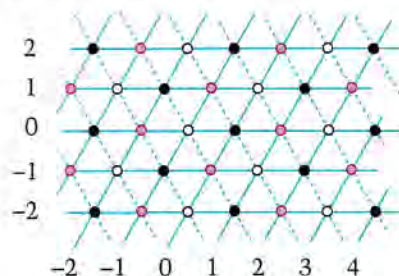


Figure 4

Numbering of grid lines and coloring of nodes used to define invariants.

STATEMENT OF OWNERSHIP, MANAGEMENT, AND CIRCULATION (Required by 39 U.S.C. 3685). (1) Publication title: Quantum. (2) Publication No. 008-544. (3) Filing Date: 10/1/94. (4) Issue Frequency: Bi-Monthly. (5) No. of Issues Published Annually: 6. (6) Annual Subscription Price: \$34.00. (7) Complete Mailing Address of Known Office of Publication: 175 Fifth Avenue, New York, NY 10010. (8) Complete Mailing Address of Headquarters or General Business Office of Publisher: 175 Fifth Avenue, New York, NY 10010. (9) Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor: Publisher: Bill G. Aldridge, National Science Teachers Association, 1840 Wilson Boulevard, Arlington, VA 22201 in cooperation with Springer-Verlag New York Inc., 175 Fifth Avenue, New York, NY 10010. Editors: Larry D. Kirkpatrick, Mark Saul, Constantine Bogdanov, Vladimir Dubrovsky, National Science Teachers Association, 1840 Wilson Boulevard, Arlington, VA 22201. Managing Editor: Timothy Weber, National Science Teachers Association, 1840 Wilson Boulevard, Arlington, VA 22201. (10) Owner: National Science Teachers Association. (11) Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities: None. (12) The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes: Has Not Changed During Preceding 12 Months. (13) Publication Name: Quantum. (14) Issue Date for Circulation Data Below: September 9, 1994. (15) Extent and Nature of Circulation: (a.) Total No. Copies (Net Press Run): Average No. Copies Each Issue During Preceding 12 Months, 11,333; Actual No. Copies of Single Issue Published Nearest to Filing Date, 10,800. (b.) Paid and/or Requested Circulation: (1) Sales Through Dealers and Carriers, Street Vendors, and Counter Sales: Average No. Copies Each Issue During Preceding 12 Months, 826; Actual No. Copies of Single Issue Published Nearest to Filing Date, 1003. (2) Paid or Requested Mail Subscriptions: Average No. Copies Each Issue During Preceding 12 Months, 7784; Actual No. Copies of Single Issue Published Nearest to Filing Date, 8004. (c.) Total Paid and/or Requested Circulation: Average No. Copies Each Issue During Preceding 12 Months, 8610; Actual No. Copies of Single Issue Published Nearest to Filing Date, 9007. (d.) Free Distribution by Mail: Average No. Copies Each Issue During Preceding 12 Months, 239; Actual No. Copies of Single Issue Published Nearest to Filing Date, 240. (e.) Free Distribution Outside the Mail: Average No. Copies Each Issue During Preceding 12 Months, 252; Actual No. Copies of Single Issue Published Nearest to Filing Date, 252. (f.) Total Free Distribution: Average No. Copies Each Issue During Preceding 12 Months, 491; Actual No. Copies of Single Issue Published Nearest to Filing Date, 492. (g.) Total Distribution: Average No. Copies Each Issue During Preceding 12 Months, 9101; Actual No. Copies of Single Issue Published Nearest to Filing Date, 9499. (h.) Copies Not Distributed: (1) Office Use, Leftovers, Spoiled: Average No. Copies Each Issue During Preceding 12 Months, 1407; Actual No. Copies of Single Issue Published Nearest to Filing Date, 299. (2) Return from News Agents: Average No. Copies Each Issue During Preceding 12 Months, 825; Actual No. Copies of Single Issue Published Nearest to Filing Date, 1002. (i.) (Sum of 15g, 15h(1), and 15h(2)): Average No. Copies Each Issue During Preceding 12 Months, 11,333; Actual No. Copies of Single Issue Published Nearest to Filing Date, 10,800. Percent Paid and/or Requested Circulation: Average No. Copies Each Issue During Preceding 12 Months, 94.60; Actual No. Copies of Single Issue Published Nearest to Filing Date, 94.82. (16) This Statement of Ownership will be printed in the January/February 1995 issue of this publication. I certify that all information furnished on this form is true and complete.

Craig Van Dyck

Craig Van Dyck
Vice President, Production

Exercises

2. Is it possible to transform a triangle of size 3 (three chips to a side) into the "parallelogram" in figure 5a?

3. For what values of n can a triangle of size n be inverted—that is, turned upside down?

4. Can a triangle of size 3 be transformed into the configuration in figure 5b?

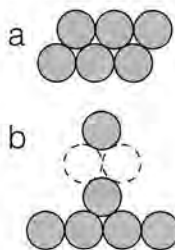


Figure 5
Can you obtain these shapes from a triangle?

Shapes

The last exercises illustrate a third type of question that can be asked about the triads: what shapes can a triangular set of six chips take after an arbitrary rearrangement by triads? (The first two were about permutations and shifts.) In fact, to solve this third question means to solve the ultimate problem, so to speak, about the triads—that is, to describe all possible arrangements of the six chips of our initial triangle that can be obtained within the rules of our game. Indeed, such an arrangement is determined by the shape and location of the set of the six nodes occupied by the chips and by the order of the chips on these nodes. But we know that any order is possible, because any permutation in the initial triangle is possible. We also know that any achievable shape can be arbitrarily moved over the grid with respect to the initial triangle, because this is true for the triangle itself. So the only thing that we don't know yet is what shapes are possible.

By the way, did you actually try to solve exercises 2–4? If you simply skipped them, do try. Computing the invariants s_x and s_y , you'll quickly find that the answer to exercise 2 is no. It will be a little more difficult to determine the values of n for which a triangle of size n can be inverted (exercise 3), and I guess it will take considerable time and effort to verify that the triangle is

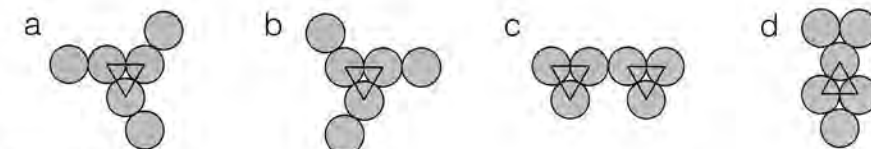


Figure 6
Shapes with a blocked triad.

really invertible for these values. As for exercise 4, our invariants do not distinguish between the six-chip triangle and the shape in figure 5b. Yet I'm sure you haven't found a desired transformation so far. But the fact that such a transformation is very hard to find may not mean it doesn't exist at all. It was hard in exercise 3 also, wasn't it? However, I have very good reasons to be sure. And maybe you found them, too: there are invariants other than s_x and s_y that prevent this transformation.

To describe them, color the nodes of the grid as shown in figure 4. This coloring can be defined algebraically. Using the numbering of lines in this figure, to each node we can assign a pair of integers (x, y) —the numbers of the slanting (x) and horizontal (y) lines that pass through this node ((x, y) are the coordinates of the node relative to the 0th lines as the axes). Then all the nodes with the same difference $x - y$ modulo 3 are colored the same: in figure 4 the nodes with $x - y = 0 \pmod{3}$ are black, those with $x - y = 1 \pmod{3}$ are red, and with $x - y = 2 \pmod{3}$ are white. A triad on the grid always covers one black, one red, and one white node, so when triads are moved, the numbers d_0, d_1 , and d_2 of black, red, and white nodes under all chips remain invariant. (Turn back now to exercise 4!)

So now we have five invariants: s_x, s_y, d_0, d_1, d_2 . They are not independent, though: clearly, $s_x = d_0 + d_1 + d_2 + s_y \pmod{3}$, so we can use, say, only s_y and forget s_x . Well, is that all now? Can we be sure that any arrangement of six chips whose four (and so, all five) invariants are the same as for the six-chip triangle is achievable starting from the triangle? Again no! Look at figure 6: all the configurations depicted there have the triangle's values of invariants—

that is, $d_0 = d_1 = d_2 = 2, s_y = 1$, but it's impossible to obtain the triangle from them. Indeed, the only triad in figure 6a (or 6b) simply can't move—it's blocked by the other three chips; and the two triads in figure 6c can only be shuffled all over the plane. The only way to create another triad is to fit them together as in figure 6d—and find out that this new triad is blocked as well. However, we'll soon see that these configurations (along with all other possible combinations of two "inverted" triads) are the only exceptions from the general rule according to which the equality of the four invariants ensures the mutual transformability of two sets of chips.

Necessary and sufficient

Now I want to formulate and prove the theorem on the triads discovered during the workshop by Hugh Robinson, who came all way down from Coventry in the United Kingdom to Beloretsk to win first prize for his investigation of this puzzle.

Let's slacken the rules of our game and allow a triad to be lifted off the plane when it's moved to another position. Then a configuration of six chips on our grid can be obtained from our original triangle if and only if it contains a triad and has the same values of the invariants d_0, d_1, d_2, s_y as the triangle ($d_0 = d_1 = d_2 = 2, s_y = 1$). The "only if" statement (necessity) follows from what was said above. A sketch of a proof of the "if" part is given in the next paragraph.

Figure 7 shows how we can remove a chip from a node and, at the same time, bring another chip onto an adjacent node of the same color: we simply attach a triad to the first chip to make a "diamond" of four chips, and then detach the other triad contained in this diamond from it.

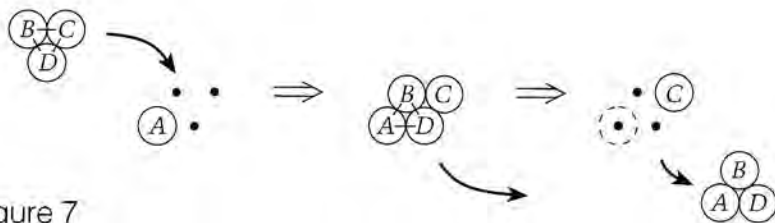


Figure 7

Jumping a chip.

The triad we attach can be “pointing up” or “pointing down,” and each of these two orientations give three of the six possible “jumps” from a node to one of its nearest neighbors of the same color. In each case, the triad is inverted after a jump. It’s not hard to see, from the values of our invariants, that any configuration that satisfies the condition of our theorem consists of a triad and three chips on nodes of different colors. We can use the triad to rearrange the “free” chips so as to form the base of the initial triangle—three chips in a row, because such a row always covers three differently colored nodes. After that we can complete the triangle by attaching the remaining triad to the base. To obtain the given configuration from the triangle we simply reverse the whole transformation.

I’ll confine myself to this outline of a proof, though it needs a certain refinement: we must make sure that all the steps of the transformation can really be performed. You can restore all the details using the equality of the invariants.

Note that the described method of transformation has to do only with the shape (and location) of the set of chips—that is, with the nodes they occupy. But this is sufficient, because, as we know, the chips in the triangle can always be permuted as desired. Also, it was essential that we could freely move the “working”

triads, which might not be the case if triads were only allowed to be slid. However, it’s not very difficult to examine all arrangements of a triad and three chips to make sure that the above proof can be conducted using sliding alone except for the cases shown in figure 6.

Thus, the transformations of the six-chip triangle are completely described. The same method can be applied to extend our theorem to any two sets of chips on the triangular grid. And not only that. If the initial set, besides a triad, contains at least two chips on nodes of different colors with respect to the coloration in figure 4, any permutation of its chips on the same set of nodes is possible, too. (To swap any two chips, we move them and three other chips so as to form a “trapezoid” like the one formed by chips 1–5 in figure 1, and then use the operation shown in this figure—chip 6 is unnecessary because it’s never moved there.)

Tetrads

I’ll end this story with a three-dimensional generalization of our puzzle. Consider a pyramid made of ten equal balls (fig. 8). We are allowed to detach a “tetrad”—a four-ball small pyramid—and move it without rotation into any new position, then detach and move a new tetrad, and so on. What configurations can be obtained from the initial pyramid?



Figure 8

Something to train your imagination.

First of all, we don’t have an underlying grid in this case: if **a**, **b**, and **c** are the vectors drawn from the center of one of the balls in a tetrad to the other three centers, then none of the vectors joining the same center to the center of the ball touching any three balls of the tetrad from the outside has the form $na + mb + kc$ with integer n, m, k . Not only that—the set of all possible positions of the balls’ centers that can arise in this game is everywhere dense in space. So we can’t define invariants the way we did on the plane, and it’s not clear how our investigation of shapes and shifts can be carried over to space.

But it’s possible to give a complete description of the permutations of the balls. Hugh Robinson and a few other participants in the conference found the simple three-move rearrangement shown in figure 9. It turns out that we don’t need anything else.

Exercise 5. Using this operation, show that we can swap any corner ball with any adjacent one.

Clearly, these swaps suffice to obtain any permutation of the balls in the pyramid, even without shifting it. ◼

ANSWERS, HINTS & SOLUTIONS
ON PAGE 58

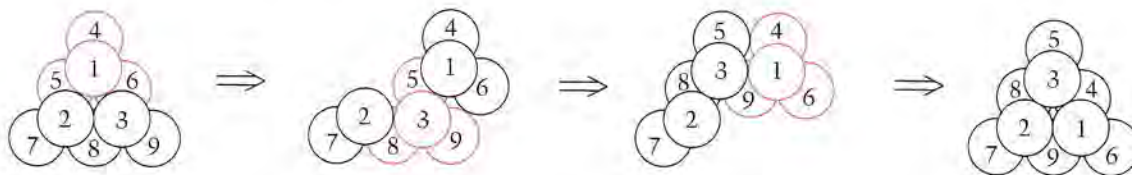


Figure 9

Key rearrangement in the tetrads puzzle, top view. (The top ball of the pyramid isn’t shown and isn’t moved.)



TAKE IT HIGHER

As one of the brightest and best math and science students in the nation, you're capable of reaching the top. And you want to attend a college or university that can help you do it. Consider the Air Force Academy. It's a college and more. It's a special place for students who seek excellence in all that they do.

At the Air Force Academy, you can "take it higher." The Academy offers a

full four-year scholarship, plus room and board. You'll graduate with a bachelor of science degree in one of 26 majors. Cadets who take honors courses build an excellent foundation for graduate studies.

Selection for the Air Force Academy is based on academic, athletic and extracurricular performance. In addition to the math and laboratory

science courses you've already taken, we recommend that you complete a solid college prep program, including four years of English, three years of social studies, two years of foreign language and one year of computer science.

In addition, we suggest that you develop your leadership abilities through school and community activities. You should also

prepare physically

by taking part in group and individual strength development and endurance programs.

The Academy's outstanding academic, athletic and leadership programs can prepare you to be an air and space leader in the 21st century.

For more details, call (719) 472-2520. Or write: HQ USAFA/RRS, 2304 Cadet Drive, Suite 200, USAF Academy, Colorado 80840-5025.

"The higher the climb, the broader the view."

American proverb

The
ACADEMY

