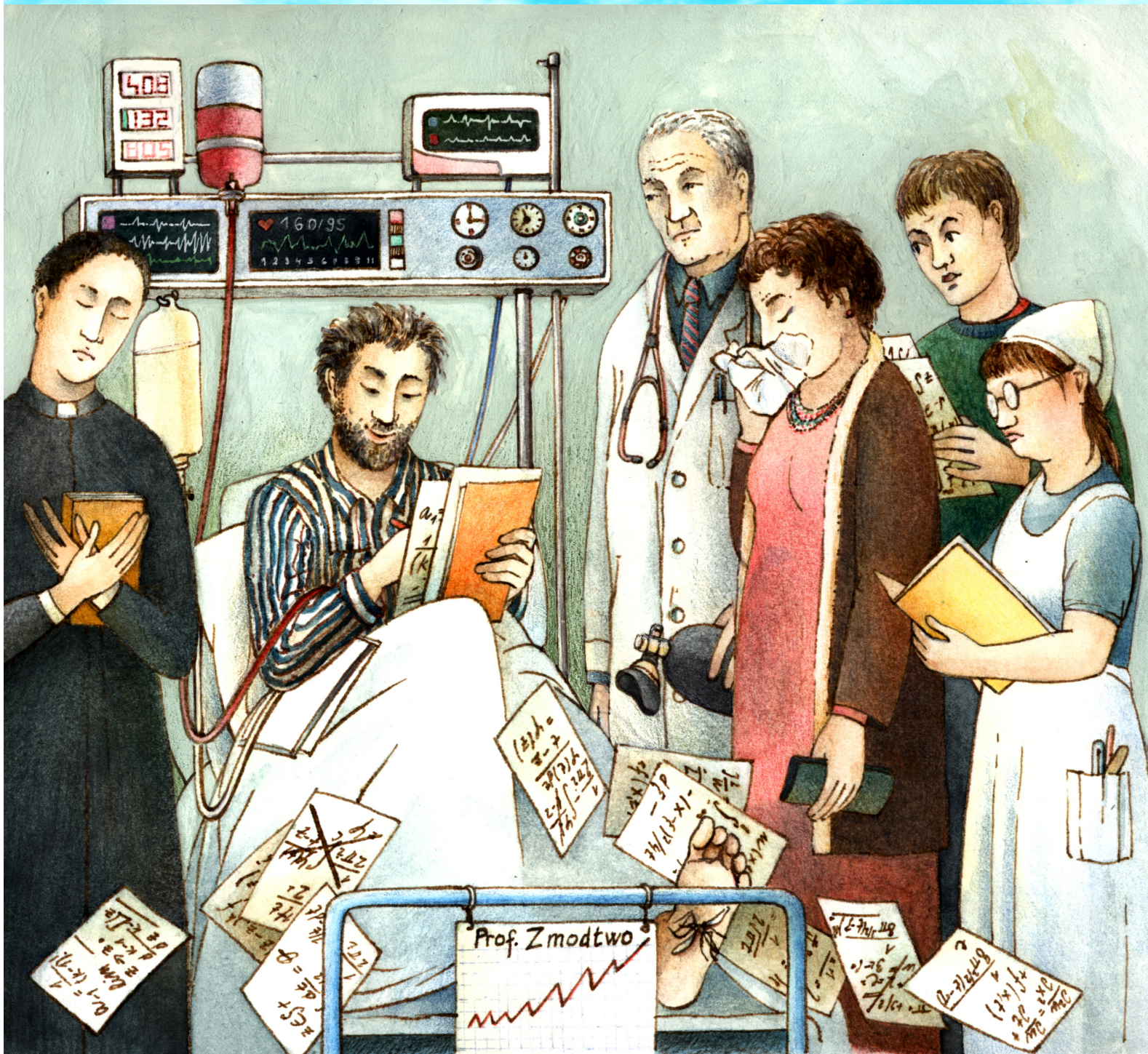


in the Sky

<http://www.pims.math.ca/pi>

Issue 8, December 2004





π *in the Sky* is a publication of the Pacific Institute for the Mathematical Sciences (PIMS). PIMS is supported by the Natural Sciences and Engineering Research Council of Canada, the Government of the Province of Alberta, the Government of the Province of British Columbia, Simon Fraser University, the University of Alberta, the University of British Columbia, the University of Calgary, the University of Victoria, the University of Washington, the University of Northern British Columbia, and the University of Lethbridge.

Significant funding for π *in the Sky* is provided by



π *in the Sky* magazine is primarily aimed at high-school students and teachers, with the main goal of providing a cultural context/landscape for mathematics. It has a natural extension to junior high school students and undergraduates, and articles may also put curriculum topics in a different perspective.

Contributions Welcome: π *in the Sky* accepts materials on any subject related to mathematics or its applications, including articles, problems, cartoons, statements, jokes, etc. Copyright of material submitted to the publisher and accepted for publication remains with the author, with the understanding that the publisher may reproduce it without royalty in print, electronic, and other forms. Submissions are subject to editorial review and revision. We also welcome **Letters to the Editor** from teachers, students, parents, and anybody interested in math education (be sure to include your full name and phone number).

Cover Page: This picture was created for π *in the Sky* by Czech artist Gabriela Novakova. The scene depicted was inspired by the article by Marjorie Wonham on “The Mathematics of Mosquitoes and West Nile Virus” appearing on page 5. Prof. Zmodtwo is again featured on the cover page, this time in a hospital bed.

Editor in Chief

Ivar Ekeland (PIMS & University of British Columbia)
Tel: (604) 822-3922, E-mail: director@pims.math.ca

Managing Editor

David Leeming (University of Victoria)
Tel: (250) 472-4271, E-mail: leeming@math.uvic.ca

Editorial Board

- Len Berggren (Simon Fraser University)
Tel: (604) 291-3335, E-mail: berggren@sfu.ca
- John Bowman (University of Alberta)
Tel: (780) 492-0532, E-mail: bowman@math.ualberta.ca
- John Campbell (Archbishop MacDonald High School, Edmonton)
Tel: (780) 441-6000, E-mail: jcdotcalm@shaw.ca
- Florin Diacu (University of Victoria)
Tel: (250) 721-6330, E-mail: diacu@math.uvic.ca
- Sharon Friesen (Galileo Educational Network, Calgary)
Tel: (403) 220-8942, E-mail: sfriesen@ucalgary.ca
- Dragos Hrimiuc (University of Alberta)
Tel: (780) 492-3532, E-mail: dhrimiuc@ualberta.ca
- Klaus Hoehsammann (University of British Columbia)
Tel: (604) 822-3782, E-mail: hoek@pims.math.ca
- Wieslaw Krawcewicz (University of Alberta)
Tel: (780) 492-7165, E-mail: wieslawk@shaw.ca
- Michael Lamoureux (University of Calgary)
Tel: (403) 220-8214, E-mail: mikel@math.ucalgary.ca
- Mark MacLean (University of British Columbia)
Tel: (604) 822-5552, E-mail: maclean@math.ubc.ca
- Alexander Melnikov (University of Alberta)
Tel: (780) 492-0568, E-mail: melnikov@ualberta.ca
- Volker Runde (University of Alberta)
Tel: (780) 492-3526, E-mail: vrunde@ualberta.ca
- Wendy Swonnell (Lambrick Park Secondary School, Victoria)
Tel: (250) 477-0181, E-mail: wswonnell@shaw.ca

Editorial Coordinator

Heather Jenkins (PIMS)
Tel: (604) 822-0402, E-mail: heather@pims.math.ca

Technical Assistant

Mande Leung (University of British Columbia)
Tel: (778) 898-7288, E-mail: mandel@ece.ubc.ca

Addresses:

<p>π <i>in the Sky</i> PIMS 449 Central Academic Bldg University of Alberta Edmonton, AB T6G 2G1, Canada</p> <p>Tel: (780) 492-4308 Fax: (780) 492-1361</p> <p>E-mail: pi@pims.math.ca</p>	<p>π <i>in the Sky</i> PIMS 1933 West Mall University of British Columbia Vancouver, BC V6T 1Z2, Canada</p> <p>Tel: (604) 822-3922 Fax: (604) 822-0883</p>
--	---

All issues of π *in the Sky* can be downloaded for free from <http://www.pims.math.ca/pi>

CONTENTS:

Editorial: On Being the Right Weight
Klaus Hoehsammann 3

Mathematical Models and Infectious Disease Dynamics
Mark Lewis 4

The Mathematics of Mosquitoes and West Nile Virus
Marjorie Wonham 5

What does Mathematics have to do with SARS?
Fred Brauer 10

Mathematical Modelling of Recurrent Epidemics
David J. D. Earn 14

Cid, Bru, One
Jeremy Tatum 18

Platonic Solids and the Golden Ratio
Klaus Hoehsammann 20

Kolmogorov, Turbulence, and British Columbia
Bob Stewart and Chris Garrett 22

Vedic Mathematics and the Calculations of Guru Tīrthajī
Jan Hogendijk 24

Tribonacci in the Sky: A Mathematical Mountain Walk
Achim Clausing 28

“Unstable” Rearrangements of n Points
Chris Garrett 32

Math Challenges 35



On Being the Right Weight

Outside genetics, where he was a pioneer, J.B.S. Haldane is now mostly remembered for a wonderful essay he wrote in 1928, called *On Being the Right Size*. If you have not read it yet, go find it on the Web—you are in for a treat. It opens by wondering about the different *sizes* of animals:

... for some reason the zoologists have paid singularly little attention to them. In a large textbook of zoology before me I find no indication that the eagle is larger than the sparrow, or the hippopotamus bigger than the hare, though some grudging admissions are made in the case of the mouse and the whale. But yet it is easy to show that a hare could not be as large as a hippopotamus, or a whale as small as a herring. For every type of animal there is a most convenient size...

He then takes us through a couple of delightful pages (e.g., “an elephant turning somersaults”) to teach us some very, very basic facts of life, with (here we go!) a mathematical background. For instance: an insect needs no lungs since the surface area of its body—say, it is shaped like an elf standing 17 mm high—would be about ten thousand times, but its volume a million times, smaller than yours. From the point of view of its innards, its outer skin is 100 times more capable than yours of supplying it with oxygen, etc. Likewise, if Nature *enlarged* you even just 10 times in all directions, you would (a) suffocate and (b) collapse, because each cm² of bone cross-section would have to carry 10 times its present load. To make sure that you can stand up straight, your weight should therefore be appropriately tuned to the *square* of your linear size.

This brings us to the Body Mass Index, whose odd formula

$$\text{BMI} = 703 \frac{w}{h^2},$$

where w is the weight in pounds and h the height in inches, triggered this editorial. A friend from Arkansas had informed us that public schools in his state must now account for the BMI of each of their students, and he was puzzled by the 3 in the 703. So were all of us at the PIMS office. Savvy as they are, Arkansas school nurses will, of course, toss off this calculation and record its result with 8-digit precision, but what if the calculator had broken down? If a guy was 70 inches tall and 140 pounds heavy, it would be so convenient to cancel 700×140 against the 70×70 for a BMI of 20, but the extra 3 ruined such simple-minded arithmetic. Why was it there?

Somebody suggested it had to do with the year 1937 (since $703 = 19 \times 37$), when US spinach growers erected a statue of Popeye the Sailor Man, whose BMI could be a benchmark. This was quickly rejected, because that sculpture was in Texas not Arkansas. Someone else noted that the BMI is measured in pounds per square inch, hence represents the *pressure* exerted by the body on a floor. Okay, we said, but 20 lb/in² would be an awful lot for anybody standing naked, even in high heels, and the 3-digit precision could only refer to a specific person on a specific day after a 48 hour fast. We seemed to be going nowhere, until our professional curmudgeon asked whether π *in the Sky* might not have a BMI-problem. “Maybe

it puts too much pressure on young brains,” he said. “Look at *your* eyes popping out because of 703.”

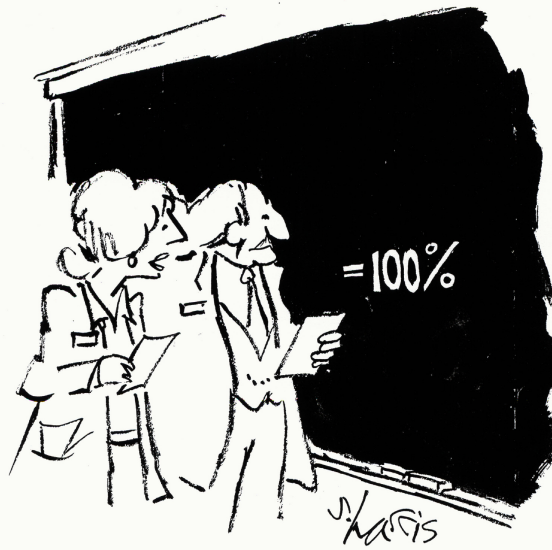
He loves to remind us of our editorial quandary: we wish to show you what our mathematics is doing out there in the world, but at the same time would like you to understand and enjoy it. But the *users* of mathematics—be they nurses in Arkansas or meteorologists in Manitoba—rarely look under the hood; they take the formulas or algorithms handed to them and *apply* them with heart-warming trust. Nevertheless, they do mould them to fit *their* needs—sometimes so thoroughly that outsiders can hardly recognize them. “In my days,” said Colin (the curmudgeon), “mathematicians would never publish anything they did not understand themselves.” And indeed, he has drawers full of papers with commentaries on anything mathematical that crosses his mind—including some mysterious passages of this magazine. He shoves them under our noses, but we are always too busy to look.

“This will now change,” spake our Editor-in-Chief after our defeat by the silly 703. “Supplementary explanations will be put on the Web, starting with the two items from the September, 2003 issue of π *in the Sky* that Colin just showed me.” And so it came to pass: you’ll find this first supplement at <http://www.pims.math.ca/pi/supp/7/>. Editing diligently, Colin will gradually empty his π -drawer, and the rest of us will chip in as best we can. Of course, we also welcome submissions from *you*, the readers. We need all the help we can get, as we try to catch up with the past and keep up with the present—not always an easy task.

You may have noticed that π *in the Sky* did not come down to earth last spring. It wanted to stay aloft, it said, “to watch all the crazy turmoil of life.” As a result, the present issue has an article on turbulence, and three on epidemics—also part of life. You’ll probably first be attracted by Mary Wonham’s mosquito story. The articles by Fred Brauer and David Earn have very similar subjects with different mixes of form and content—simple premise with rigorous treatment, or *vice versa*. They might well be read in parallel.

By the way, we *did* eventually find the clue to the 703. In the *metric* system, height is measured in centimetres and weight (actually: *mass*) in kilograms. The conversion factors $H = 2.54$ and $W = 0.4536$ yield $W/H^2 = 0.070308$. Hence the metric BMI factor is 10 000 instead of 703.

K.H.



"TERRIFIC SOLUTION. NOW IF WE COULD ONLY FIND A PROBLEM TO FIT IT."

©Copyright 2004
Sidney Harris



Mathematical Models and Infectious Disease Dynamics

Mark Lewis[†]

Mathematical models can be used to understand what factors govern infectious disease outbreaks including HIV/AIDS, West Nile virus, and even the bubonic plague! The purpose of the model is to take facts about the disease as inputs and to make predictions about the numbers of infected and uninfected people over time as outputs.

Factors that can go into the models include the length of time one is ill, the length of time one can infect others (often different than the total length of time one is ill), the level of contagiousness of the disease (i.e., the likelihood of infecting another individual if one comes into close contact), the number of uninfected (susceptible) individuals, and so forth. Mathematical modellers then feed these facts into a set of equations.

In the three accompanying articles (see pages 5–17), the equations used are referred to as differential equations. Differential equations involve derivatives of functions. For example, if the number of infected individuals at a given time t is represented by function $I(t)$, then the time derivative of I , denoted by $I'(t)$, says how quickly $I(t)$ is changing. If $I'(t) > 0$, the disease is increasing with time and if $I'(t) < 0$, the disease is decreasing with time.

Being able to make predictions about disease dynamics is really helpful for public health. If we know there will be an outbreak we can prepare for it. Alternatively, if we have a reliable model, we can study how to prevent an outbreak and save lives by changing the factors we can control using public health means. These factors include education, immunization, quarantine regulations, and health treatment strategies.

The application of mathematical models to infectious disease dynamics has been a real success story in 20th century science. Even though the dynamics of disease appear to be very complex, surprisingly simple mathematical models can be used to understand features governing the outbreak and persistence of infectious disease.

Early models, such as the one by Kermack–McKendrick (see Fred Brauer’s article on pages 10–13), were applied to understand the dynamics of historical diseases, such as bubonic plague. Amongst other things, the models could be used to predict the fraction of the population that would survive a major disease outbreak.

Here the form of the mathematical model is a system of differential equations (say, one equation to track the levels of each of the susceptible $S(t)$, infective $I(t)$, and previously infected but now recovered or dead $R(t)$ portion of the population). The equations are ‘coupled’ to each other, because

the growth of infectives relies on having new susceptibles to infect, the growth of removed (recovered or dead) relies on having had infected individuals, and so forth.

Such models can be easily put on the computer, using software such as Maple, to make predictions about changing levels of disease outbreak over time. Alternatively, as outlined in the article by Fred Brauer, mathematics can play a more central role. Analysis of the models yields useful statistics about the disease, such as the ‘basic reproduction number’—a measure of infectivity, expressed as the number of secondary infections arising directly from a single infective individual surrounded by susceptibles. Methods to control the disease are then summarized by the degree to which they are able to reduce the basic reproduction number to a number less than one.

Recurrent 20th century diseases, such as childhood measles, can be analyzed using a similar framework (see David Earn’s article on pages 14–17). Here models are extended to include birth and death of individuals and seasonal and other changes in the levels of contact, for example, between school-age children. Trends in birth rates and effects of immunization can also be incorporated into the models. With the analysis of these augmented models, new questions can be posed and answered, varying from: “why do diseases exhibit complex temporal patterns, ranging from cycles to erratic, seemingly chaotic fluctuations?” to “what is the optimal immunization strategy for a given disease, given finite resources and vaccine availability?”

Much of the recent mathematical work in disease modelling has focused on emerging diseases, ranging from HIV/AIDS, to SARS and West Nile virus (see Marjorie Wonham’s article on pages 5–10 and the last part of Fred Brauer’s article). Here recommendations regarding control methods are needed urgently. For most emerging diseases, modellers do not have the luxury of comprehensive data sets showing outbreak levels over time. Therefore the models must be developed based on detailed understanding of the components of the disease dynamics and from our experiences with modelling previous diseases. Here predictions and recommendations for control stem from the mathematical and numerical analysis of the models (see Marjorie Wonham’s article). The modelling of emerging disease is the current challenge for mathematical epidemiologists, and it is one that will be with us for some time to come.



©Copyright 2004
Wieslaw Krawcewicz

[†] Mark Lewis is the Canada Research Chair in Mathematical Biology and a Professor in the Department of Mathematical and Statistical Sciences at the University of Alberta. His e-mail address is mlewis@math.ualberta.ca.



The Mathematics of Mosquitoes and West Nile Virus

Marjorie Wonham[†]

Lying on my sleeping pad, I warily eye the mosquito perched above my head. I could reach up and squash it, but that would require extracting my arm from the warmth of my sleeping bag. So for now, it clings to the yellow nylon of my tent, unaware of its reprieve. Although I may triumph over this particular mosquito, I am all too aware of being vastly outnumbered outside my tent.

Until recently, my interest in mosquitoes was largely pragmatic: avoid, repel, or swat. Lately, though, I have developed a grudging curiosity about how they make a living. Fact: a female mosquito overwinters with fertilized eggs so the first thing she does in spring, before even feeding, is lay eggs. Fact: Even if she doesn't find a blood meal, she can survive by sucking plant juices. Fact: the combined meals of a mosquito horde can (and do) bleed a newborn calf to death. Fact: at best a mosquito bite simply itches; at worst, it means disease transmission—malaria, yellow fever, dengue fever, and West Nile virus are all mosquito-borne.

It is West Nile virus that has piqued my interest of late. First identified in Uganda in 1937, the virus is well established in its native Africa where it lives primarily in birds and is transmitted among them by mosquitoes. Only occasionally does a mosquito transmit the infection to a mammal. From time to time a West Nile virus outbreak occurs in Europe and Africa—in Israel in the 1950s and South Africa in 1974, and more recently in Romania, Morocco, Tunisia, Italy, France, and Russia. Just recently, West Nile virus made its first known, and headline grabbing, North American appearance.

In the summer of 1999, the birds of New York City began mysteriously to die, their bodies appearing conspicuously in the city zoo, parks, and backyards. At first the cause was unknown, but by December of the same year it had been identified, in two reports published in the same issue of Science magazine, as West Nile virus, a disease never before seen on this continent. In subsequent summers, West Nile virus spread west across the continent reaching Ontario in 2001, California and Washington in 2002, and Alberta in 2003.

Corvids—crows and jays—were the hardest hit among the birds; other passerines such as sparrows also carried the virus but were dying in smaller numbers. Among mammals, horses appeared especially vulnerable, with a mortality rate of approximately 40%. Human cases were less common and less likely to be fatal, but were a growing health concern nonetheless. By the end of 2003, the virus had been identified in 7 Canadian provinces and 46 U.S. states, in at least 10

mosquito, 150 bird, and 17 other vertebrate species, and in a total—for that year alone—of over 11,000 human cases in Canada and the U.S.

In Alberta, West Nile virus was first reported in 2003. The year before, I had moved to Edmonton to join a group of mathematical biologists at the University of Alberta. There, I was immersed in a world of mathematical modelling used to tackle biological questions. Two years later, here in my tent on a canoe trip in the Northwest Territories, I have a chance to reflect on the unpredicted collaboration that developed with my mathematical colleagues on the dynamics and control of West Nile virus.

At first glance, we made an unlikely trio for this project. Tomás was a graduate student and programming whiz who studied chamomile invasions on farmland. Mark was a mathematical biologist who had modelled the movement of birds and wolves. And I was a marine ecologist with a mathematical background largely limited to reading tide tables. But together we were galvanized by a question from a colleague in Ontario: “How come no one is modelling West Nile virus?” Hugh asked. And with that casual question began our Year of the Mosquito.

How could mathematical modelling help us understand West Nile virus dynamics? The virus was spreading, and control proposals were beginning to include spraying adult and larval mosquitoes, removing larval mosquito habitat, and even removing birds. Since all of these strategies would cause additional impacts on the environment, perhaps modelling could help maximize control effectiveness while minimizing unwanted effects? I had no idea where to begin, but luckily, I was in good mathematical hands.

Mark and Tomás introduced me to a class of disease models known by their acronym as SIR models (see accompanying articles by Fred Brauer and David Earn). These models were first extended to vector-borne diseases by R. Ross in the early 1900s and G. Macdonald in the 1950s, to combat malaria. Since then, a large associated body of mathematical theory, and an impressive history of contributing to disease control, have both evolved.

In an SIR model, the host population is divided into three groups: Susceptible (healthy uninfected individuals), Infectious (infected and capable of transmitting the disease), and Removed (immune, dead, or otherwise removed individuals). The rates at which an average individual moves from Susceptible to Infectious and from Infectious to Removed are determined, and the relevant birth and death rates are incorporated.

Once constructed, a key piece of information can be extracted from an SIR model, called the disease *reproduction number*, or \mathcal{R}_0 (“R-zero” or “R-naught”). \mathcal{R}_0 tells us the number of new infections that would result from the introduction of a single infectious individual into an entirely susceptible population. For example, if a student with chicken pox walked into a classroom of individuals with no previous exposure to the disease, how many new cases would be caused by direct contact with the initial infectious student? The answer is given by \mathcal{R}_0 . Or for West Nile virus, if an infectious bird arrived in a new city, how many other birds would be infected (via mosquito bites) by that original bird? Again, \mathcal{R}_0 tells us the answer.

The expression for \mathcal{R}_0 is constructed, according to a particular formula, from variables and parameters in the model. Reasonably enough, it takes into account factors such as how long the first individual remains infectious, the likelihood of contact between the infectious and susceptible individuals (either directly or via another species), and how often contact leads to disease transmission. Details of the mathematics underlying the calculation of \mathcal{R}_0 are given in the article by Fred Brauer.

[†] Marjorie Wonham is a Postdoctoral Fellow in the Centre for Mathematical Biology and the Department of Mathematical & Statistical Sciences at the University of Alberta. Her e-mail address is mwonham@ualberta.ca.



BY SWITCHING TERMS IN HER EQUATIONS, MARJORIE FOUND SHE COULD REVERSE THE WEST NILE VIRUS EPIDEMIC.

©Copyright 2004
Gabriela Novakova

For disease control, the value of \mathcal{R}_0 is key. $\mathcal{R}_0 < 1$ means that an infectious individual will, on average, generate fewer than one new infection, so the disease will die out even without control efforts. On the other hand, $\mathcal{R}_0 > 1$ means that an infectious individual will generate more than one new infection, so a disease outbreak will occur. In this case, the disease can be controlled by methods that alter one or more of the model components—such as mortality, contact, or transmission rates—to reduce \mathcal{R}_0 below one.

Armed with this mathematical background, we were ready to develop an SIR model for West Nile virus in North America, calculate \mathcal{R}_0 , and identify how and how much to control the disease to prevent an outbreak. First, we had to define the biological and geographic scope of our efforts. We had already learned that the virus persisted in transmission cycles between mosquitoes (vectors) and birds (reservoir hosts). Although it occasionally spread to other vertebrates (including humans), it seemed not to return to mosquitoes. In other words, although an infection might be deeply significant to the human in question, it would not influence the overall disease dynamics.

This biological fact helped us simplify the mathematical description: by viewing the virus outbreak level in mosquitoes and birds as a proxy for the human infection risk, we could limit our model to only the vector and the reservoir host. For mosquitoes, we focused on the species group that was emerging as the dominant North American vector (*Culex pipiens* spp.). For birds, we focused on the species with the best available infection and mortality data, the American crow (*Corvus brachyrhynchos*). And since the best virus prevalence data were available where West Nile had first appeared, we confined ourselves to modelling the New York City outbreak. Finally, since the disease outbreak at this latitude showed a marked seasonality, appearing in summer and disappearing in winter, we confined the model to a single season from spring through fall.

Soon we had the skeleton of a model. We defined three groups (S , I , and R) for both mosquitoes and birds. From the literature, we obtained estimates of the mosquito biting rate and the transmission probabilities that allowed a mosquito to

infect a bird (which was around 88%) and a bird to infect a mosquito (which was only around 16%). We had recovery and mortality rates for birds, but not for mosquitoes since they didn't seem to be affected by the disease. Since we had birth and death rates for mosquitoes but not for birds, we assumed the birds reproduced once in spring before the model began, and their background (natural) mortality rate would be negligible in the one summer.

To refine the model and assign numerical values to the parameters, we divided up the work according to our expertise. Tomás investigated how to solve and simulate these models on a computer, and Mark explored the mathematical theory underlying this approach. I searched the biological literature for parameter values for mosquitoes, birds, and virus transmission. I didn't envy Tomás and Mark, as my task seemed much the easiest; I was surprised to learn later that they felt the same way about their roles.

Nonetheless, tracking down the biological data took some sleuthing. Today, from the safety of my tent, I can make a guess as to the local mosquito abundance outside: it's very, very high. If I were to stick out a bare arm, it would be covered almost instantly; if I left it out, I could watch an individual probe repeatedly, biting several times before finding her blood meal. (I don't do this often.) As I canoe down the river, I sometimes see a cloud of mosquito larvae rising to the surface and hatching.

For our model, though, we needed observations like this to be quantified. Just how many mosquitoes were out there? How many eggs did one female lay, and how many larvae hatched? How many crows were there? How many mosquito bites per crow in a day? And how contagious was the virus? My quest was something of a scavenger hunt travelling back through biological history, with each paper leading me to an older one. At the same time, new reports about the virus were appearing almost daily in print and online. I divided my time between dusty library shelves and internet listserves.

New biological information, new equations, new parameter estimates, new model analyses, and new simulations surfaced. For several months, we worked to tailor the model as best we could to the biological information. In the interests of tractability, the model had to remain as simple as possible.



WE DO NOT RECOMMEND SWITCHING VARIABLES IN THE EQUATIONS!

©Copyright 2004
Wieslaw Krawcewicz

But the biological complexity seemed almost infinite. At times, I wondered if we were simplifying too much for the model to be informative, while Tomás and Mark wondered if the biology was making the model too complicated to be useful.

In the end, the biology dictated two substantial additions to the model. We learned that mosquitoes could spend quite a long time, up to 14 days, as aquatic larvae that don't bite birds and therefore don't transmit the disease. It can also take quite a long time, perhaps 10–12 days, for an infected mosquito to develop a viral load high enough to transmit the disease back to a bird. These two time periods could add up to almost half a mosquito's lifespan, so they could substantially alter the disease dynamics. We therefore added two new groups to the population: one for larval L and one for exposed E mosquitoes.

With these additions, it seemed we finally had a model we all felt was realistic and tractable. This is the model illustrated in Figure 1. Now that we had the model, it was time for a test: would it behave realistically? We had to know this before we could use \mathcal{R}_0 to make any predictions.

As a test, we chose the records of West Nile virus incidence in both mosquitoes and birds from New York City in 2000. We plugged our literature-based parameter estimates into the model, crossed our fingers, and ran the simulations. Sure

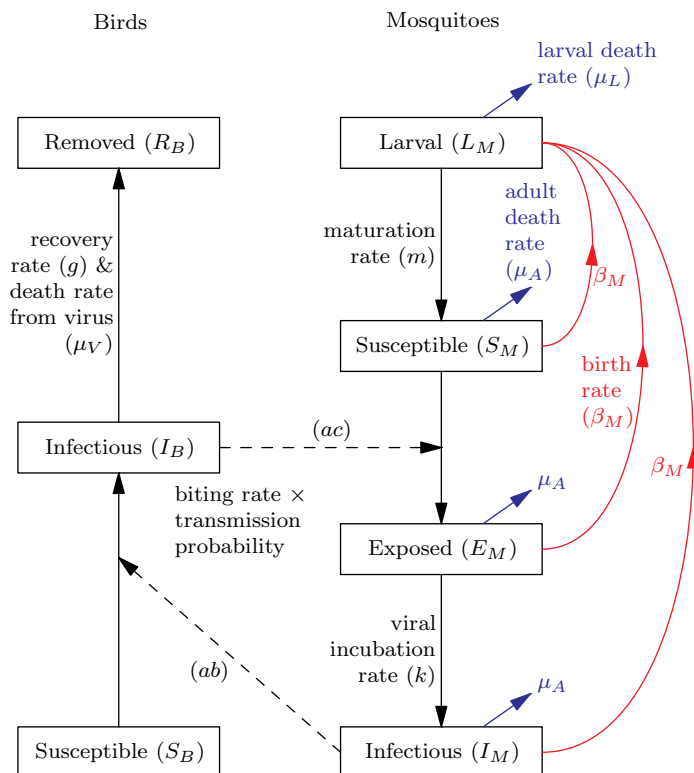


Figure 1: Cartoon illustration of SIR model for West Nile virus cross-infection between birds and mosquitoes. The boxes represent *variables*: the number (or proportion) of birds that are susceptible, infectious, and removed, and mosquitoes that are larval, susceptible, exposed, and infectious. The arrows represent *parameters*: the daily rates at which the value of each variable changes. Solid black lines show the rates (and reasons) for individual birds and mosquitoes moving from one category to another. The two dashed lines represent the rates at which mosquitoes infect birds and birds infect mosquitoes. The red and blue arrows respectively show the birth and death rates for mosquitoes (which can reproduce multiple times during a single season).

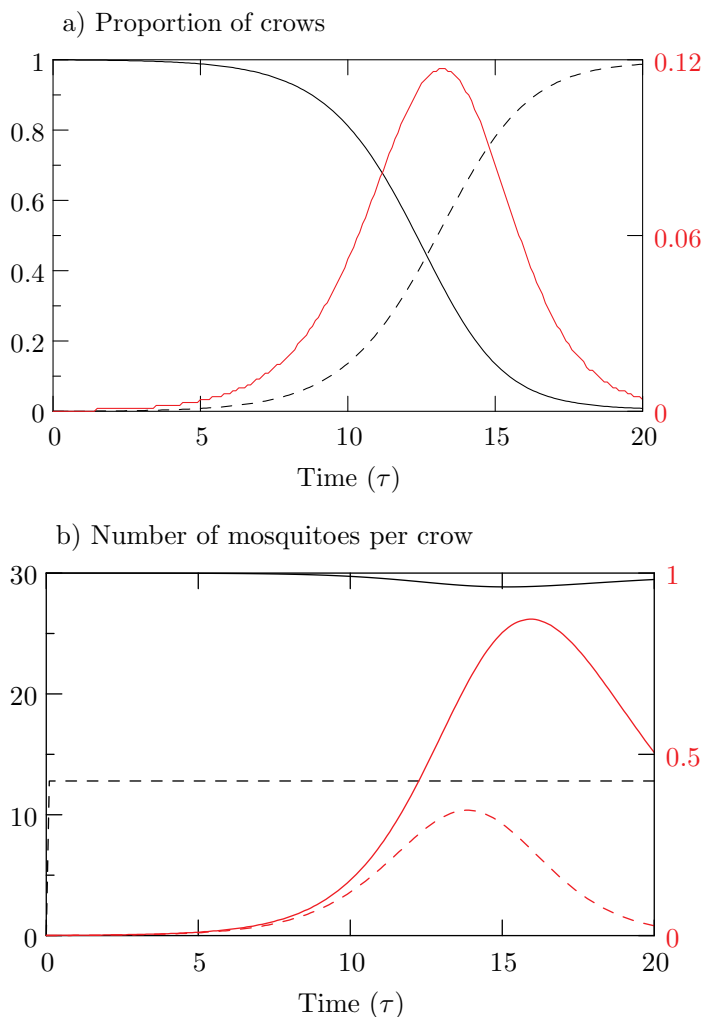


Figure 2: Numerical simulations of crow proportions and mosquito numbers predicted by the West Nile virus model for a steady-state mosquito population ($S_{M_0}/N_{B_0} = 30$) over a season lasting 14τ days, where τ is a time unit representing 9.4 days (this time unit may seem funny, but it is a result of manipulating the equations to a nondimensional form—a detail that is not shown here). a) the proportion of susceptible (solid line) and removed birds (dashed line) on the left axis and infectious birds (red line) on the right. b) the number of susceptible (solid line) and larval mosquitoes per bird (dashed line) on the left axis and exposed (dashed red line) and infectious mosquitoes per bird (solid red line) on the right.

enough, for a reasonable range of starting values the model predicted a range of mosquito and bird infection levels that matched the observed data (Figure 2). We were now in a position to use the model to investigate control strategies.

Following methods developed by other mathematicians, we calculated \mathcal{R}_0 for our model, which turned out to be made up of three elements:

$$\mathcal{R}_0 = \frac{ab}{\mu_A} \cdot \frac{ac}{\mu_V + g} \cdot \frac{S_{M_0}}{N_{B_0}} \left(\frac{k}{k + \mu_A} \right).$$

The parts work together like this. The first factor represents the number of new crow infections caused by an infectious mosquito. This is the number of bites per day per mosquito a multiplied by the probability of the virus being transmitted from an infectious mosquito to a crow b multiplied by the average number of days $1/\mu_A$ a mosquito would live.

The second factor is the mirror image: the number of new mosquito infections caused by an infectious crow. The mosquito biting rate a is multiplied by the transmission probability from crows to mosquitoes c and by the average number of days until the infectious crow either recovered ($1/g$) or died ($1/\mu_V$).

The third factor represents, generally speaking, the number of infectious mosquitoes per crow. Specifically, it is the number of initially susceptible mosquitoes S_{M_0} that survive the virus exposure period $k/(k + \mu_A)$ for every bird N_{B_0} . Together, these three elements give us the expression for the total disease \mathcal{R}_0 from birds to birds (via mosquitoes) or from mosquitoes to mosquitoes (via birds). Taking the square root of the right hand side of the expression for \mathcal{R}_0 is a common convention that gives the geometric mean \mathcal{R}_0 from bird to mosquito and *vice versa*.

(If you're looking at Figure 1 and wondering why the mosquito birth rate and larval death rate don't seem to show up in \mathcal{R}_0 , rest assured. They are accounted for by the equality $L_{M_0} = \beta_M S_{M_0}/(m + \mu_L)$, a simplifying assumption in the model that ensures a constant mosquito population.)

Our parameter values gave \mathcal{R}_0 greater than one, predicting a disease outbreak. We were now able to return to our original questions, namely, how could West Nile virus be controlled, and how much control was needed?

One possible answer was obvious: if every mosquito and bird were removed, the virus could not persist. But we were hoping to find a more palatable answer. Examining the expression for \mathcal{R}_0 , we found the ratio S_{M_0}/N_{B_0} in the numerator, telling us that reducing the number of mosquitoes S_{M_0} would reduce \mathcal{R}_0 , but reducing the number of birds N_{B_0} would not. In fact, reducing bird abundance would only make things worse by increasing the value of \mathcal{R}_0 . This was our first lesson: *reducing mosquitoes could help control the virus, but removing birds would only increase the chances of an outbreak.*

Would every mosquito have to be removed to prevent an outbreak? This was our second lesson. Plotting a graph of \mathcal{R}_0 vs. S_{M_0}/N_{B_0} showed us that *the virus could be controlled simply by reducing mosquito abundance, without requiring that every last individual be eliminated* (Figure 3).

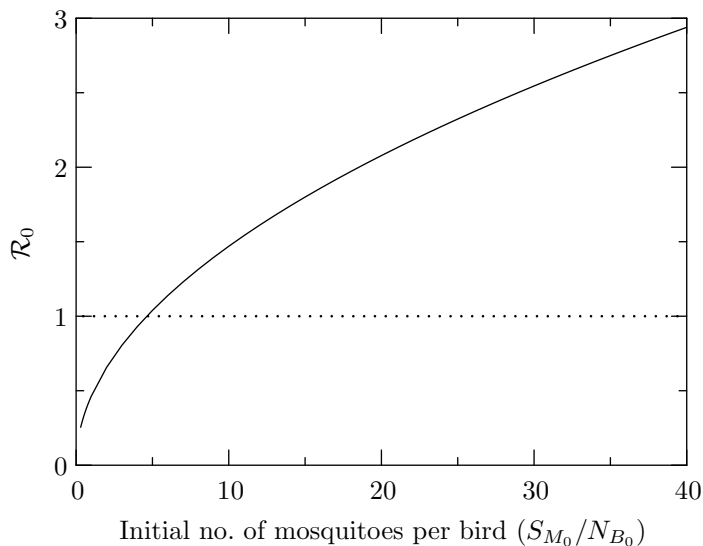


Figure 3: Plot of \mathcal{R}_0 vs. the initial number of mosquitoes per bird S_{M_0}/N_{B_0} , showing that the mosquito population can simply be reduced, and not completely eliminated, to bring \mathcal{R}_0 below one and therefore prevent a West Nile outbreak.

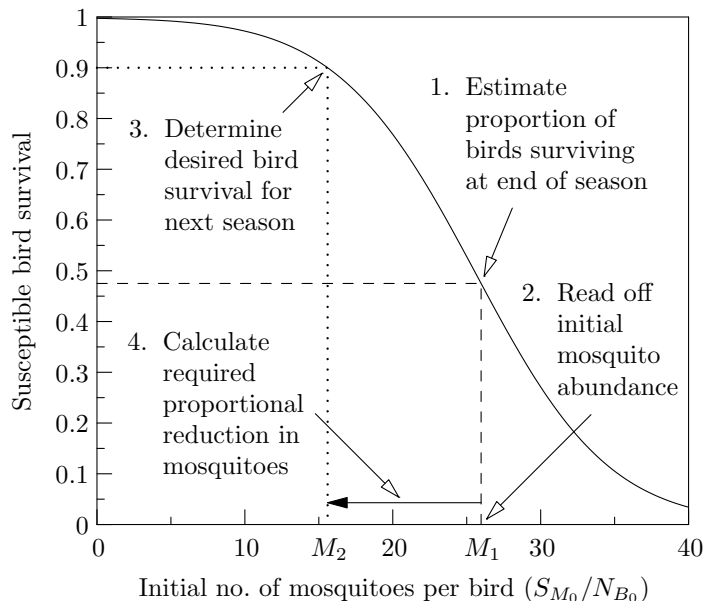


Figure 4: Plot of susceptible bird survival at the end of the season S_B vs. the initial number of mosquitoes per bird at the beginning of the season S_{M_0}/N_{B_0} , showing how future mosquito control efforts can be estimated from previous bird mortality data.

This begged the final question of exactly how much mosquito reduction was enough. It took us a while to work this out, but eventually we came up with an indirect route to the answer. It requires the following four steps, which are illustrated in Figure 4.

1. Obtain from public health authorities an estimate of the proportion of crows that survived the virus in the last season. (These data are becoming increasingly available.)
2. Determine how many mosquitoes per crow M_1 there must have been at the start of the last season to cause the observed bird mortality.
3. Determine the maximum allowable bird mortality for the following season, and the corresponding maximum allowable initial number of mosquitoes per crow M_2 .
4. The ratio of these two mosquito-per-crow estimates, M_2/M_1 , gives the proportional reduction in mosquitoes needed to control the virus in the next year.

This was our third lesson from the model, that *we could calculate the required amount of mosquito reduction*. In fact, we found that a 40–70% reduction in the New York City mosquito population would have prevented the 2000 outbreak. There is, of course, a catch with this approach: *the virus has to have already caused the initial mortalities that are used for subsequent control estimates*. However, as the virus continues to spread, bird mortality data are increasingly available. Perhaps the experience of cities and regions with the virus (and therefore with the data) can inform the control programs in areas where a virus outbreak has not yet occurred.

Like many research projects, this one raised more questions than it answered. How fast could West Nile virus spread across the continent and where would it go? How important was climate in determining the survival of birds, mosquitoes, and the virus? How would controlling the mosquitoes in an

urban area influence the surrounding rural area? How was our model similar or different compared to other models of similar diseases? Was there anything more general that could be learned about the epidemiology of vector-borne viruses? By working with some additional collaborators, we have begun to address these questions too.

In the meantime, I have an active interest in some very local mosquito control. Earlier this evening, I noticed that the spiders in the rocks behind my tent were making a killing, literally. Mosquitoes were landing in their webs so fast the spiders could hardly keep up. So now I'm starting to wonder how many spiders I would need to keep in my tent to control the mosquitoes. Or if not spiders, perhaps a bat would be more efficient? Clearly, I need another model. I'll have to consult with my collaborators.

Appendix

Bird equations:	
Susceptible	$\frac{dS_B}{dt} = -abI_M \frac{S_B}{N_B}$
Infectious	$\frac{dI_B}{dt} = abI_M \frac{S_B}{N_B} - \mu_V I_B - gI_B$
Removed	$\frac{dR_B}{dt} = (g + \mu_V) I_B$
Mosquito equations:	
Larval	$\frac{dL_M}{dt} = \beta_M(S_M + E_M + I_M) - mL_M - \mu_L L_M$
Susceptible	$\frac{dS_M}{dt} = -acS_M \frac{I_B}{N_B} + mL_M - \mu_A S_M$
Exposed	$\frac{dE_M}{dt} = acS_M \frac{I_B}{N_B} - kE_M - \mu_A E_M$
Infectious	$\frac{dI_M}{dt} = kE_M - \mu_A I_M$

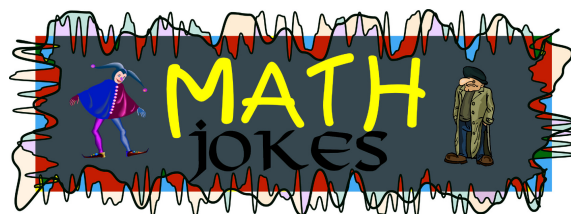
Table 1: Ordinary differential equations (ODEs) representing the transmission of West Nile virus between birds and mosquitoes. Each equation keeps track of the change in a variable: the number of susceptible S_B , infected I_B , and removed R_B birds, and the number of larval L_M , susceptible S_M , exposed E_M , and infected I_M female mosquitoes per bird. Change in these variables is represented by parameters, which are probabilities and rates. (Note that all variables are indicated by capital letters, and all parameters by lower-case letters.) These equations match up with the boxes and arrows in Figure 1. To create your own numerical simulation of this model, you will need to choose the initial conditions, or values, for each variable. For a West Nile virus invasion, you might start with all variables equal to zero, except for I_B or I_M , which could be a very small positive number. Then you can set up the ODEs to change the values of the variables in each time step. What initial value of I_B or I_M do you need in order for the infection to increase? How do your results change if you allow birds to recover (i.e., set $g > 0$)?

Parameter	Mean (range)	Biological meaning
a	0.09 (0.03–0.16)	Mosquito daily per capita biting rate on birds
b	0.88 (0.80–1.00)	Probability per bite of West Nile virus being transmitted by an infectious mosquito biting a bird
c	0.16 (0.02–0.24)	Probability per bite of West Nile virus being transmitted by a mosquito biting an infectious bird
β_M	0.537 (0.036–42.5)	Mosquito daily per capita birth rate
m	0.068 (0.051–0.093)	Mosquito larva daily per capita maturation rate to susceptible adults
μ_A	0.029 (0.016–0.070)	Adult mosquito daily per capita mortality rate
μ_L	1.191 (0.213–16.9)	Larval mosquito daily per capita mortality rate
k	0.106 (0.087–0.125)	Mosquito daily per capita transition rate from exposed to infectious
μ_V	0.143 (0.125–0.200)	Bird daily per capita mortality rate from West Nile virus
g	0	Bird daily per capita recovery rate from West Nile virus

Table 2: Sample parameter values (and ranges) estimated from the literature for the West Nile virus model shown in Figure 1 and Table 1. The phrase *per capita* is a common biological term meaning per individual.

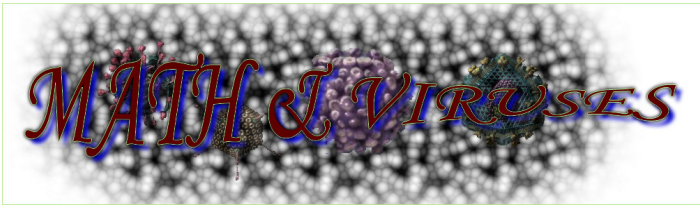
Further Information

- [1] This article is based on the paper: Wonham, M., T. de-Camino-Beck and M. Lewis, 2004: An epidemiological model for West Nile virus: Invasion analysis and control applications. *Proceedings of the Royal Society of London B*, **271**, 501–507.
- [2] Read more about this and other research at the University of Alberta's Centre for Mathematical Biology: <http://www.math.ualberta.ca/~mathbio/>.



Q: How does a mathematician induce good behaviour in his/her children?
A: "If I've told you n times, I've told you $n + 1$ times..."

Q: What keeps a square from moving?
A: Square roots.



What does Mathematics Have to do with SARS?

Fred Brauer[†]

At least since the beginning of recorded history there have been epidemics. One of the plagues that Moses brought down upon Egypt described in the Book of Exodus was murrain, an infectious cattle disease, and there are many other biblical descriptions of epidemic outbreaks. The Black Death (thought to be bubonic plague) spread from Asia through Europe in several waves beginning in 1346, causing the death of one-third of the population of Europe between 1346 and 1350 and recurring regularly in Europe for more than 300 years, notably as the Great Plague in London of 1665–1666. Recurring invasions of cholera killed millions in India in the 19th century. The influenza epidemic of 1918–19 killed 20 million people overall and more than half a million in the United States. More recently, the SARS epidemic of 2002–3 caused worldwide concern and even more recently several strains of avian flu have forced the killing of millions of birds and worries about spread to humans.

Diseases that are *endemic* (always present), especially in less developed countries, have effects that are probably less widely known but may be of even more importance. Every year millions of people die of measles, respiratory infections, diarrhoea and other diseases that are easily treated and not considered dangerous in the Western world. Diseases such as malaria, typhus, cholera, schistosomiasis, and sleeping sickness are endemic in many parts of the world. The effects of high disease mortality on mean life span and of disease debilitation and mortality on the economy in afflicted countries are considerable. There are many useful practical conclusions that have been drawn from models for endemic diseases. One example was the possibility of eliminating smallpox worldwide by vaccination; this was successfully achieved in 1977. However, to keep the mathematics relatively simple in this article, we shall discuss only models for epidemic diseases. We hope that the reader who finds this introduction to the modelling of epidemics interesting may be motivated to study enough additional mathematics to learn about models for endemic diseases.

Perhaps the first epidemic to be examined from a modelling point of view was the Great Plague in London. The Great Plague killed about one-sixth of the population of London. The village of Eyam near Sheffield, England suffered an outbreak of bubonic plague in 1665–1666 whose source is believed to be the Great Plague. The Eyam plague was survived by only 83 of an initial population of 350 persons. There were actually two epidemics in Eyam and the first phase was survived by 261 persons. As detailed records were preserved and as the community was persuaded to quarantine itself to try to prevent the spread of disease to other communities, the second phase of the epidemic in Eyam has been used as a case study for modelling. The actual data for the Eyam epidemic are remarkably close to the predictions from the simple model as described below.

If there is no vaccine or treatment available for a disease, the only control strategies available are isolation of individuals diagnosed with the disease and quarantine of suspected infectives. The rates of isolation and quarantine may be varied, depending on decisions about the amount of effort to invest in these strategies. It is rarely possible to compare possible control strategies such as the division of efforts into isolation and quarantine during an actual epidemic. For this reason, mathematical modelling of epidemics is a promising tool for comparison of possible strategies. If and when a vaccine is developed, models might indicate whether it is more urgent to concentrate on vaccination or isolation.

One of the early triumphs of mathematical epidemiology was the formulation in 1927 of a simple model by a public health physician, W. O. Kermack, and a biochemist, A. G. McKendrick, predicting behaviour similar to that observed in countless epidemics, namely that diseases develop suddenly and then disappear just as suddenly without infecting the entire community. Kermack and McKendrick considered the class S of individuals susceptible to the disease, that is, not yet infected, and the class I of infected individuals, assumed infective and able to spread the disease by contact with susceptibles. In their model, individuals who have been infected and then removed from the possibility of being re-infected or of spreading infection are ignored. Removal is carried out through isolation from the rest of the population, immunization against infection, recovery from the disease with immunity against reinfection, or through death caused by the disease. These characterizations of removed members are quite different from an epidemiological perspective and of course also from a human point of view, but are equivalent from a modelling point of view that takes into account only the state of an individual with respect to the disease.



Angry SARS virus

[†] Fred Brauer is a professor in the Department of Mathematics, University of British Columbia. His e-mail address is brauer@math.ubc.ca.

The Kermack–McKendrick epidemic model makes very simple assumptions about the rates of disease transmission and removal. One of the assumptions is that the disease is transmitted from one individual of a population to another by direct contact. Thus it is not applicable to diseases that are transmitted by a *vector*, that is, diseases transmitted back and forth between two populations such as mosquitoes and birds, as in West Nile virus. However, the ideas that go into the formulation of the Kermack–McKendrick model are also useful for the formulation of more complicated epidemic models.

The model contains only two parameters (the values of which are to be determined from observed data) and could be applied to many diseases transmitted by direct contact. While a more detailed model might be a better description of a specific disease, it would require more parameters. Since data are often incomplete and inaccurate because of under-reporting and mis-diagnosis at the beginning of an epidemic, a simple model may give better predictions.

The Kermack–McKendrick model, which is a deterministic compartmental model, is formulated in terms of the rates of flow of members of the population between compartments. There is a flow from S to I representing the rate of new infections and a flow out of I representing the rate of recovery or disease death. Mathematically, these rates of change are described as *derivatives* with respect to time t . We will use dS/dt to denote the derivative of S and dI/dt the derivative of I , thinking of S and I as functions of time t . There are techniques for calculating the derivative of a given function, but our situation is that we want to set up equations for the derivatives of the functions S and I and draw some conclusions about the behaviour of the functions.

The specific assumptions about the flow rates are as follows:

- (i) An average infective member of the population makes contact sufficient to transmit infection with βN others per unit time, where N represents total population size.
- (ii) A fraction γ of infectives leave the infective class per unit time.
- (iii) There is no entry into or departure from the population, except possibly through death from the disease.

According to (i), since the fraction of contacts by an infective with a susceptible, who can then transmit infection, is S/N , the number of new infections in unit time per infective is $(\beta N)(S/N)$, giving a rate of new infections $(\beta N)(S/N)I = \beta SI$. Fortunately, we need not give an algebraic expression for N since it cancels out of the final model. The hypothesis (ii) says that the infective periods are exponentially distributed with mean infective period $1/\gamma$. The hypothesis (iii) really says that the time scale of the disease is much faster than the time scale of births and deaths, so that demographic effects on the population may be ignored.

When these assumptions are translated into mathematical statements of the transition rates between classes, the result is a pair of equations, called differential equations, for the derivatives dS/dt and dI/dt . These equations are

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI, \\ \frac{dI}{dt} &= (\beta S - \gamma)I.\end{aligned}$$

In words, there is a rate of flow βSI of new infections out of S and into I , and a rate of flow γI out of I .

Let us think of a population of initial size N into which a small number of infectives is introduced, so that $S(0) \approx N$, $I(0) \approx 0$. If $S(0) \approx N < \gamma/\beta$, then I decreases to zero

(no epidemic), while if $N > \gamma/\beta$, then I first increases to a maximum attained when $S = \gamma/\beta$ and then decreases to zero (epidemic). The quantity $\beta N/\gamma$ is a threshold quantity, called the *basic reproduction number* and denoted by \mathcal{R}_0 , which determines whether there is an epidemic or not.

The definition of the basic reproduction number \mathcal{R}_0 is that the basic reproduction number is the number of secondary infections caused by a single infective introduced into a wholly susceptible population of size N over the course of the infection of this single infective. In this situation, an infective makes βN contacts in unit time, all of which are with susceptibles and thus produce new infections, and the mean infective period is $1/\gamma$; thus the basic reproduction number is $\beta N/\gamma$.

Initially, the number of infectives grows if and only if $\mathcal{R}_0 > 1$ because the equation for I may be approximated by

$$\frac{dI}{dt} = (\beta N - \gamma)I$$

and thus dI/dt is positive when $t = 0$ if $\mathcal{R}_0 > 1$ and negative if $\mathcal{R}_0 < 1$. If we could solve the above pair of differential equations for the functions S and I , we would have a prediction of the numbers of susceptibles and infectives as functions of time. Unfortunately, we cannot solve them analytically, but there is some information that we can deduce directly from the differential equations. Since $dS/dt < 0$, the function $S(t)$ is always decreasing. Since $dI/dt < 0$ if $\beta S < \gamma$, the function $I(t)$ decreases after S reaches the value $S = \gamma/\beta$. Thus, eventually I decreases to zero. If $\mathcal{R}_0 < 1$, then $I(t)$ is always decreasing, which means that the initial infection does not develop into an epidemic. On the other hand, if $\mathcal{R}_0 > 1$, then $I(t)$ increases initially until $S(t)$ decreases to the value $S = \gamma/\beta$ and then decreases to zero, an epidemic.

There is a useful mathematical trick that will give us some additional information about the behaviour of solutions. The trick is that instead of thinking of S as a function of the time t , we think of t as a function of S . Then I being a function of t , which in turn is a function of S , makes I indirectly a function of S . Calculus tells us that the derivative of I with respect to S , which we denote by dI/dS , is given by the rule

$$\frac{dI}{dS} = \frac{\frac{dI}{dt}}{\frac{dS}{dt}},$$

which allows us to calculate

$$\begin{aligned}\frac{dI}{dS} &= \frac{(\beta S - \gamma)I}{-\beta SI} \\ &= -1 + \frac{\gamma}{\beta S}.\end{aligned}$$

More calculus rules (techniques of integration) give I as a function of S , namely

$$I = -S + \frac{\gamma}{\beta} \ln S + c,$$

where c is a constant that we still need to determine. We use the initial state of the population to do this. The constant c is determined by the initial values $S(0)$, $I(0)$ of S and I , respectively. With $S(0) + I(0) = N$ we substitute the value $t = 0$ into the solution to give

$$c = N - \frac{\gamma}{\beta} \ln S(0).$$

Thus

$$I = -S + \frac{\gamma}{\beta} \ln S + N - \frac{\gamma}{\beta} \ln S(0).$$

If we use the fact that $I(t) \rightarrow 0$ as $t \rightarrow \infty$ and let S_∞ be the limiting value of $S(t)$ as $t \rightarrow \infty$, we obtain

$$N - \frac{\gamma}{\beta} \ln S(0) = S_\infty - \frac{\gamma}{\beta} \ln S_\infty,$$

the *final size equation*. In particular, this equation tells us that $S_\infty > 0$, so that some members of the population escape the epidemic. This has frequently been observed in real life. After an epidemic passes there are always some members of the population that do not have disease antibodies, which means that they are not immune and have not been infected.

It is convenient to depict the epidemic model by showing the *phase portrait*, the orbit shown as a curve in the (S, I) plane in Figure 1. Since S is a decreasing function of time, the orbit is traversed from right to left. The vertical line in Figure 1 has the equation $S = \gamma/\beta$, so that $\mathcal{R}_0 > 1$ if $S(0)$ is to the right of this line and $\mathcal{R}_0 < 1$ if $S(0)$ is to the left of this line. This shows graphically that if $\mathcal{R}_0 > 1$, there is an epidemic with the number of infectives first increasing to a maximum I_{\max} and then decreasing to zero, while if $\mathcal{R}_0 < 1$, the infection dies out without spreading. The graph in Figure 1 was actually drawn using a computer algebra system (Maple) to solve the pair of differential equations numerically with the parameter values

$$\beta = 0.0178, \quad \gamma = 2.73, \quad S(0) = 254, \quad I(0) = 7, \quad N = 261,$$

which correspond to the Eyam plague epidemic data. It is worth noting that we may use these data to calculate $I_{\max} = 30.26$ from the relation between I and S with $S = \gamma/\beta$. We may also use the final size equation to estimate numerically that $S_\infty = 76.05$, again using a computer algebra system (because the final size equation gives S_∞ implicitly). We should remember that since the actual observed results have integer values, we should interpret the model as predicting $S_\infty = 76$, $I_{\max} = 30$, compared with the observed values $S_\infty = 83$, $I_{\max} = 29$.

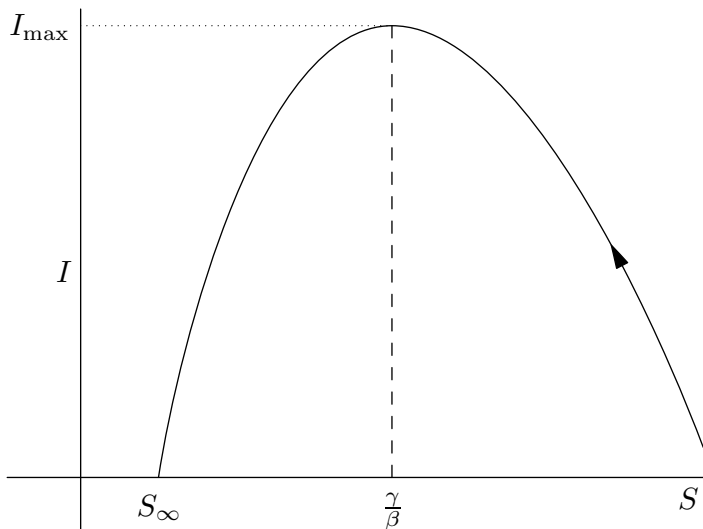


Figure 1: The phase portrait.

The actual data for the Eyam epidemic are remarkably close to the predictions of this very simple model. However, the model is really too good to be true. Our model assumes that infection is transmitted directly between people. While this is possible, bubonic plague is transmitted mainly by rat fleas. When an infected rat is bitten by a flea, the flea becomes extremely hungry and bites the host rat repeatedly, spreading the infection in the rat. When the host rat dies

its fleas move on to other rats, spreading the disease further. As the number of available rats decreases, the fleas move to human hosts; this is how plague starts in a human population (although the second phase of the epidemic may have been the pneumonic form of bubonic plague, which can be spread from person to person.) One of the main reasons for the spread of plague from Asia into Europe was the passage of many trading ships; in medieval times ships were invariably infested with rats. An accurate model of plague transmission would have to include flea and rat populations, as well as movement in space. Such a model would be extremely complicated and its predictions might well not be any closer to observations than our simple unrealistic model.

In the village of Eyam the rector persuaded the entire community to quarantine itself to prevent the spread of disease to other communities. One effect of this policy was to increase the infection rate in the village by keeping fleas, rats, and people in close contact with one another, and the mortality rate from bubonic plague was much higher in Eyam than in London. Further, the quarantine could do nothing to prevent the travel of rats and thus did little to prevent the spread of disease to other communities. One message this suggests to mathematical modellers is that control strategies based on false models may be harmful, and it is essential to distinguish between assumptions that simplify but do not alter the predicted effects substantially, and wrong assumptions that make an important difference, even if they seem to fit observed data. To model an epidemic we must know the transmission route, even if we can not yet identify a virus or a method of transmission.

A real epidemic differs considerably from an idealized model. One difference is that when it is realized that an epidemic has begun, some individuals will modify their behaviour by avoiding crowds to reduce contacts and by being more careful about hygiene to reduce the risk that a contact will produce infection. Diagnosed infectives may be hospitalized, both for treatment and to isolate them from the rest of the population. Contact tracing of diagnosed infectives may identify people at risk of becoming infective, and these may be quarantined (instructed to remain at home and avoid contacts) and monitored so that they may be isolated immediately if and when they become infective. Isolation may be imperfect and there may be infections transmitted in hospitals.

Soon after the beginning of the SARS epidemic of 2002–3 scientists began looking at the formulation of models to describe this epidemic. Usually, this work was carried out by groups of scientists, including mathematicians, statisticians, epidemiologists, and microbiologists. For example, a network was set up by MITACS (Mathematics of Information Technology and Complex Systems) including scientists from Health Canada and universities across Canada. Some early observations were that the SARS epidemic was an example of a general class of models, not a completely separate situation, and that it was possible to formulate more general and more inclusive models of Kermack–McKendrick type. For this reason, the MITACS group, which was originally set up to study SARS, became a group studying the modelling of infectious diseases in general. Models have been developed that incorporate the additional features described above and have been used to try to answer such questions as whether control of an epidemic should concentrate on isolation of diagnosed infectives or quarantine of contacts of infectives, or a combination of both. For example, it appears that in an epidemic that can be controlled by isolation, the additional gains from a quarantine are slight. Of course, it is important to remember that a model is only a model and will not be an accurate description of all aspects of an epidemic. Nevertheless, although mathematics has not yet cured any diseases (except

possibly math anxiety), it may help in controlling future epidemics. In the event of future outbreaks, you can expect that epidemiologists and mathematical modellers will collaborate to suggest the best control strategy. This kind of mathematical modelling requires some knowledge of calculus, differential equations, and linear algebra. Further discussion of the above epidemic model and more realistic models is contained in [1, Chapter 7] and in [2, Section 6.6].

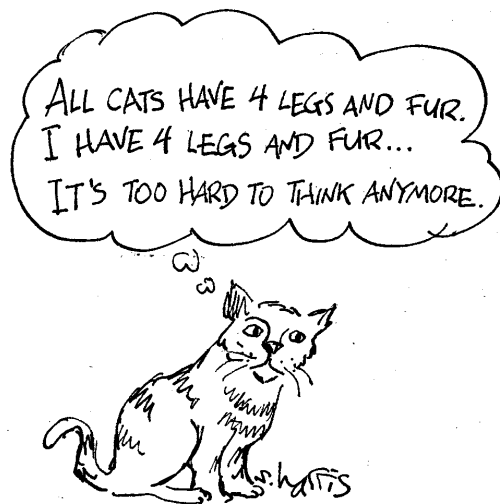
Acknowledgements

I would like to acknowledge David Earn and Pauline van den Driessche who made some very helpful suggestions in the preparation of this article.

References

[1] Brauer, F. and C. Castillo-Chavez, 2001: Mathematical Models in Population Biology and Epidemiology. *Texts in Applied Mathematics*, 40, Springer-Verlag, New York.

[2] Keshet, L., 2004: Mathematical Models in Population Biology. *SIAM Classics in Applied Mathematics*, 46, SIAM.

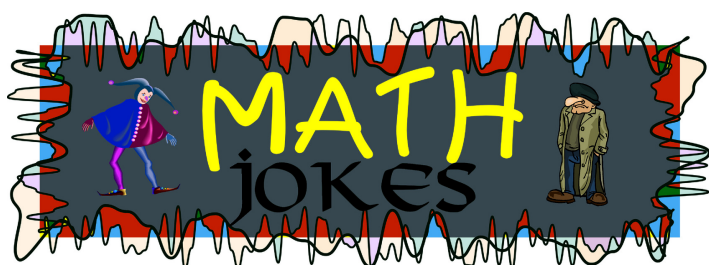


©Copyright 2004
Sidney Harris

A vector walks into two bars. . . and everyone yells, "Norm!"

If a math presentation was:

- Understood by everybody in the audience: *it was a worthless bunch of triviality.*
 - Only some people were able to follow: *it was definitely NOT my area.*
 - Nobody understood even the first definition: *it was a great talk—serious research leading to important results.*
-



There were three medieval kingdoms on the shores of a lake. There was an island in the middle of the lake, over which the kingdoms had been fighting for years. Finally, the three kings decided that they would send their knights out to do battle, and the winner would take the island.

The night before the battle, the knights and their squires pitched camp and readied themselves for the fight. The first kingdom had 12 knights, and each knight had five squires, all of whom were busily polishing armor, brushing horses, and cooking food. The second kingdom had twenty knights, and each knight had 10 squires. Everyone at that camp was also busy preparing for battle. At the camp of the third kingdom, there was only one knight, with his squire. This squire took a large pot and hung it from a looped rope in a tall tree. He busied himself preparing the meal, while the knight polished his own armor.

When the hour of the battle came, the three kingdoms sent their squires out to fight (this was too trivial a matter for the knights to join in). The battle raged, and when the dust had cleared, the only person left was the lone squire from the third kingdom, having defeated the squires from the other two kingdoms, thus proving that the squire of the high pot and noose is equal to the sum of the squires of the other two sides.



©Copyright 2004
Sidney Harris



Mathematical Modelling of Recurrent Epidemics

David J. D. Earn[†]

One of the most famous examples of an epidemic of an infectious disease in a human population is the Great Plague of London, which took place in 1665–1666. We know quite a lot about the progression of the Great Plague because weekly bills of mortality from that time have been retained. A photograph of such a bill is shown in Figure 1. Note that the report indicates that the number of deaths from plague (5533) was more than 37 times the number of births (146) in the week in question, and that wasn't the worst week! (As Fred Brauer notes in his article in this issue, an even worse plague occurred in the 14th century, but no detailed records of that epidemic are available.)

The Diseases and Casualties this Week	
Born	1
Christened	146
Males	68
Females	78
In all	146
Plague	5533
Aged	50
Consumption	99
Dropsie	22
Frighted	1
Gowt	1
Grief	1
Gripping in the Guts	3
Jaundies	35
Impothame	2
Infants	8
Kingevill	9
Meagrome	2
Purples	2
Rickets	5533
Rising of the Lights	10
Rupture	13
Scurvy	1
Spotted Fever	5
Stilborn	65
Stone	10
Stopping of the Stomach	6
Suddenly	1
Surfeit	36
Teeth	113
Thrush	3
Tifick	5
Vomiting	4
Winde	1
Wormes	12
Found dead in the Fields at St. Mary Hllington	1
Feaver	268
Fiflula	2
Flox and Small-pox	4
Flux	1
Cough	1
Convulsion	63
Apoplexie	42
Childbed	11
Chriftomes	1
Cold	1
Aged	50
Bortive	6

Decreased in the Burials this Week	1837	Parishes clear of the Plague	7	Parishes Infected	123
Buried	6460	Plague	5533		
Males	3212				
Females	3248				
In all	6460				

The Assize of Bread set fourth by Order of the Lord Mayor and Courts of Aldermen, A penny Wheaten Loaf to contain Nine Ounces and a half, and three half-penny White Loaves the like weight.

Figure 1: A photograph of a bill of mortality for the city of London, England, for the week of 26 September to 3 October 1665.

[†] David J. D. Earn is a professor in the Department of Mathematics & Statistics at McMaster University. His web site is <http://www.math.mcmaster.ca/earn>.

Putting together the weekly counts of plague deaths from all the relevant mortality bills, we can obtain the *epidemic curve* for the Great Plague, which I've plotted in the top left panel of Figure 2. The characteristic exponential rise, turnover and decline is precisely the pattern predicted by the classic susceptible-infective-recovered (SIR) model of Kermack and McKendrick [1] that I describe below (and Fred Brauer also discusses in his article). While this encourages us to think that mathematical modelling can help us understand epidemics, some detailed features of the epidemic curve are not predicted by the simple SIR model. For example, the model does not explain the jagged features in the plotted curve (and there would be many more small ups and downs if we had a record of daily rather than weekly deaths). However, with some considerable mathematical effort, these “fine details” can be accounted for by replacing the differential equations of Kermack and McKendrick with equations that include stochastic (i.e., random) processes [2]. We can then congratulate ourselves for our modelling success... until we look at more data.

The bottom left panel of Figure 2 shows weekly mortality from plague in London over a period of 70 years. The Great Plague is the rightmost (and highest) peak in the plot. You can see that on a longer timescale, there was a complex pattern of plague epidemics, including extinctions and re-emergences. This cannot be explained by the basic SIR model (even if we reformulate it using stochastic processes). The trouble is likely that we have left out a key biological fact: there is a reservoir of plague in rodents, so it can persist for years, unnoticed by humans, and then re-emerge suddenly and explosively. By including the rodents and aspects of spatial spread in a mathematical model, it has recently been possible to make sense of the pattern of 17th century plague epidemics in London [3]. Nevertheless, some debate continues as to whether all those plagues were really caused by the same pathogenic organism.

A less contentious example is given by epidemics of measles, which are definitely caused by a well-known virus that infects the respiratory tract in humans and is transmitted by airborne particles. Measles gives rise to characteristic red spots that are easily identifiable by physicians who have seen many cases, and parents are very likely to take their children to a doctor when such spots are noticed. Consequently, the majority of measles cases in developed countries end up in the office of a doctor (who, in many countries, is required to report observed measles cases to a central body). The result is that the quality of reported measles case data is unusually good, and it has therefore stimulated a lot of work in mathematical modelling of epidemics.

An epidemic curve for measles in New York City in 1962 is shown in the top right panel of Figure 2. The period shown is 17 months, exactly the same length of time shown for the Great Plague of London in the top left panel. The 1962 measles epidemic in New York took off more slowly and lasted longer than the Great Plague of 1665. Can mathematical models help us understand what might have caused these differences?

Using the same notation as Fred Brauer uses in his article in this issue, the basic SIR model is

$$\frac{dS}{dt} = -\beta SI, \quad (1)$$

$$\frac{dI}{dt} = \beta SI - \gamma I. \quad (2)$$

Here, S and I denote the numbers of individuals that are susceptible and infectious, respectively. The derivatives dS/dt and dI/dt denote the rates of change of S and I with

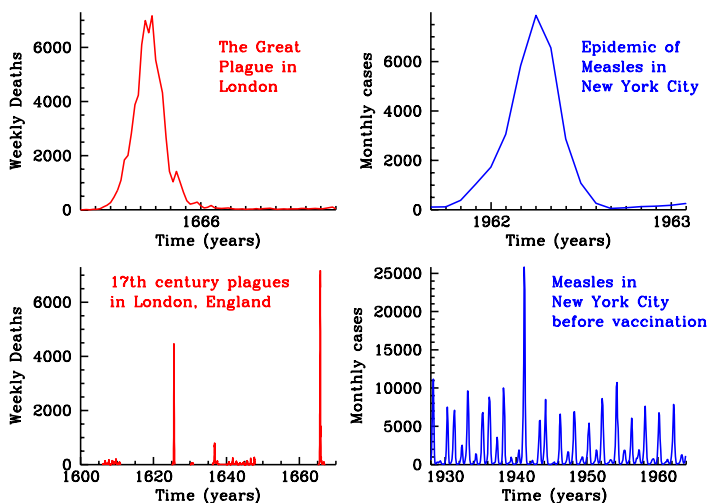


Figure 2: Epidemic curves for plague in London (left panels) and measles in New York City (right panels). For plague, the (red) curves show the number of deaths reported each week. For measles, the (blue) curves show the number of cases reported each month. In the top panels, the small ticks on the time axis occur at monthly intervals.

respect to time. The *mean transmission rate* is β and the *mean recovery rate* is γ (so the *mean infectious period* is $1/\gamma$). As Fred Brauer discusses, if the total population size is N , and everyone is initially susceptible ($S(0) = N$), then a newly introduced infected individual can be expected to infect $\mathcal{R}_0 = \beta N/\gamma$ individuals (the *basic reproduction number* \mathcal{R}_0 is also discussed at length by Marjorie Wonham in her article in this issue). You can find a discussion of the SIR model together with the mathematical ideas it is based on in some introductory calculus textbooks (see, for example, [4]).

As Fred Brauer notes, we cannot solve the SIR equations and obtain formulae for the functions $S(t)$ and $I(t)$. Yet the epidemic curves that we are trying to explain are essentially given by $I(t)$, so it is hard to proceed without it! Fortunately, computers come to our rescue. Rather than seeking an explicit formula for $I(t)$, we can instead obtain a numerical approximation of the solution. One simple approach is Euler’s method, which we can implement as follows (using a spreadsheet or any standard programming language).

The derivative dS/dt is defined as the ratio of the change in S in a given short time interval dt , divided by that time interval, in the limit that dt approaches zero. Dealing with that limit is tricky, but at any time t we can approximate the derivative by writing $dS = S(t + dt) - S(t)$ and solving for the number of susceptibles at a time $t + dt$ in the future,

$$S(t + dt) = S(t) - \beta S(t)I(t)dt. \quad (3)$$

Similarly, we can approximate the number of infectives at time $t + dt$ as

$$I(t + dt) = I(t) + \beta S(t)I(t)dt - \gamma I(t)dt. \quad (4)$$

Equations (3) and (4) together provide a scheme for approximating solutions of the basic SIR model. To implement this scheme on a computer, you need to decide on a suitable small time interval dt . If you want to try this, I’d suggest taking dt to be one tenth of a day. I should point out that I am being extremely cavalier in suggesting the above method. Do try this at home, but be forewarned that you can easily generate garbage using this simple approach if you’re not careful. (To avoid potential confusion, include a line in your program

that checks that $S(t) \geq 0$ and $I(t) \geq 0$ at all times. Another important check is to repeat your calculations using a much smaller dt and make sure your results don’t change.)

In order for your computer to carry out the calculations specified by equations (3) and (4), you need to tell it the parameter values (β and γ , or \mathcal{R}_0 , N and γ) and initial conditions ($S(0)$ and $I(0)$). For measles, estimates that are independent of the case report data that we’re trying to explain indicate that the mean infectious period is $1/\gamma \sim 5$ days and the basic reproduction number is $\mathcal{R}_0 \sim 18$ [5]. The population of New York City in 1960 was $N = 7\,781\,984$. If we now assume one infectious individual came to New York before the epidemic of 1962 ($I(0) = 1$), and that everyone in the city was susceptible ($S(0) = N$), then we have enough information to let the computer calculate $I(t)$. Doing so yields the epidemic curve shown in the top panel of Figure 3, which does *not* look much like the real data for the 1962 epidemic in New York. So is there something wrong with our model?

No, but there is something very wrong with our initial conditions. The bottom right panel of Figure 2 shows reported measles cases in New York City for a 36 year period, the end of which includes the 1962 epidemic. Evidently, measles epidemics had been occurring in New York for decades with no sign of extinction of the virus. In late 1961, most of New York’s population had already had measles and was already immune, and the epidemic certainly didn’t start because one infectious individual came to the city. The assumptions that $I(0) = 1$ and $S(0) = N$ are ridiculous. If, instead, we take $I(0) = 123 \cdot (5/30)$ (the number of reported cases in September 1961 times the infectious period as a proportion of the length of the month) and $S(0) = 0.065N$, then we obtain the epidemic curve plotted in the middle panel of Figure 3, which is much more like the observed epidemic curve of Figure 2 (top right panel). This is progress—we have a model that can explain a single measles epidemic in New York City—but the model cannot explain the recurrent epidemics observed in the bottom right panel of Figure 2. This is not because we still don’t have exactly the right parameter values and initial conditions: no parameter values or initial conditions lead to recurrent epidemics in this simple model. So, it would seem, there must be some essential biological mechanism that we have not included in our model. What might that be?

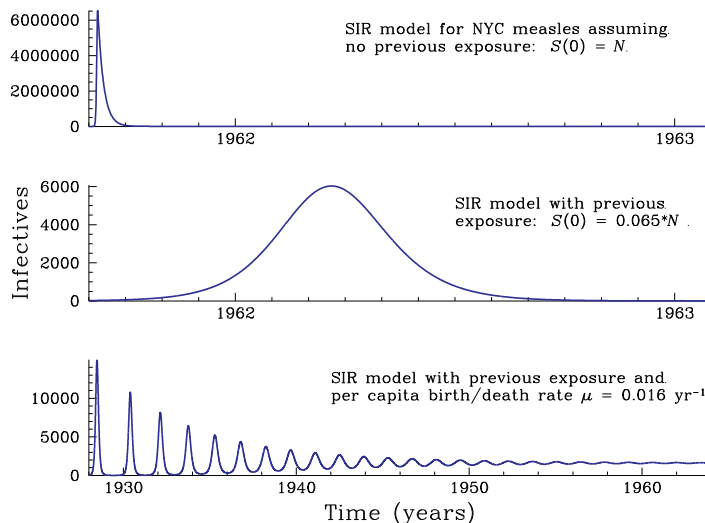


Figure 3: Epidemic curves for measles in New York City, generated by the basic SIR model. The curves show the number of infectives $I(t)$ at time t . In the top two panels, the small ticks on the time axis occur at monthly intervals.

Let's think about why a second epidemic cannot occur in the model we've discussed so far. The characteristic turnover and decline of an epidemic curve occurs because the pathogen is running out of susceptible individuals to infect. To stimulate a second epidemic, there must be a source of susceptible individuals. For measles, that source cannot be previously infected people, because recovered individuals retain lifelong immunity to the virus. Newborns typically acquire immunity from their mothers, but this wanes after a few months. So births can provide the source we're looking for.

If we expand the SIR model to include B births per unit time and a natural mortality rate μ (per capita), then our equations become

$$\frac{dS}{dt} = B - \beta SI - \mu S, \quad (5)$$

$$\frac{dI}{dt} = \beta SI - \gamma I - \mu I. \quad (6)$$

The timescale for substantial changes in birth rates (decades) is generally much longer than a measles epidemic (a few months), so we'll assume that the population size is constant (thus $B = \mu N$, so there is really only one new parameter in the above equations; we can take it to be B). As before, we can use Euler's trick to convert the equations above into a scheme that enables a computer to generate approximate solutions. An example is shown in the bottom panel of Figure 3, where I have taken the birth rate to be $B = 126\,372$ per year (the number of births in New York City in 1928, the first year for which we have data). The rest of the parameters and initial conditions are as in the middle panel of the figure.

Again we seem to be making progress. We are now getting recurrent epidemics, but the oscillations in the numbers of cases over time damp out, eventually reaching an equilibrium. While the graph is just an approximate solution for a single set of initial conditions, it can actually be proved that all initial conditions with $I(0) > 0$ yield solutions that converge onto this equilibrium. So we still don't have a model that can explain the real oscillations in measles incidence from 1928 to 1964, which showed no evidence of damping out. Back to the drawing board?

Don't give up. We've nearly cracked it. So far, we have been assuming implicitly that the transmission rate β (or, equivalently, the basic reproduction number \mathcal{R}_0) is simply a constant and, in particular, that it does not change in time. Let's think about that assumption. The transmission rate is really the product of the rate of contact among individuals and the probability that a susceptible individual who is contacted by an infectious individual will become infected. But the contact rate is *not* constant throughout the year. To see that, consider the fact that in the absence of vaccination, the average age at which a person is infected with measles is about five years [5]; hence most susceptibles are children. Children are in closer contact when school is in session, so the transmission rate varies seasonally. A crude approximation of this seasonality is to assume that β varies sinusoidally,

$$\beta(t) = \beta_0(1 + \alpha \cos 2\pi t). \quad (7)$$

Here, β_0 is the mean transmission rate, α is the *amplitude* of seasonal variation and the time t is assumed to be measured in years. If, as above, β is assumed to be a periodic function (with a period of one year) then the SIR model is said to be *seasonally forced*. We can still use Euler's trick to solve the equations approximately, and I encourage you to do that using a computer for various values of the seasonal amplitude α (you must have $0 \leq \alpha \leq 1$: why?).

You might think that seasonal forcing is just a minor tweak of the model, but in fact this forcing has an enormous impact on the epidemic dynamics that the model predicts. If you've taken Physics and studied the forced pendulum, then you might already have some intuition for this. A pendulum with some friction will exhibit damped oscillations and settle down to an equilibrium. But if you tap the pendulum with a hammer periodically then it will never settle down and it can exhibit quite an exotic range of behaviours including chaotic dynamics [6] (oscillations that look random). Similarly complex dynamics can occur in the seasonally forced SIR model.

Most importantly, with seasonal forcing, the SIR model displays *undamped* oscillations similar to the patterns seen in the real measles case reports. But we are left with another puzzle. If you look carefully at the New York City measles reports in the bottom right panel of Figure 2 you'll see that before about 1945 the epidemics were fairly irregular, whereas after 1945 they followed an almost perfect two-year cycle. While the SIR model can generate both irregular dynamics and two-year cycles, this happens for different parameter values, not for a single solution of the equations. How can we explain *changes over time* in the *pattern* of measles epidemics?

Once again, the missing ingredient in the model is a changing parameter value. This time it is the birth rate B , which is not really constant. Birth rates fluctuate seasonally, but to such a small extent that this effect is negligible. What turns out to be more important is the much slower changes that occur in the average birth rate over decades. For example, in New York City the birth rate was much lower during the 1930s (the "Great Depression") than after 1945 (the "baby boom") and this difference accounts for the very different patterns of measles epidemics in New York City during these two time periods [7].

A little more analysis of the SIR model is very useful. It is possible to prove that changes in the birth rate have exactly the same effect on disease dynamics as changes of the same relative magnitude in the transmission rate or the proportion of the population that is vaccinated [7]. This equivalence makes it possible to explain historical case report data for a variety of infectious diseases in many different cities [8].

One thing that you may have picked up from this article is that successful mathematical modelling of biological systems tends to proceed in steps. We begin with the simplest sensible model and try to discover everything we can about it. If the simplest model cannot explain the phenomenon we're trying to understand, then we add more biological detail to the model, and it's best to do this in steps because we are then more likely to be able to determine which biological features have the greatest impact on the behaviour of the model.

In the particular case of mathematical epidemiology, we are lucky that medical and public health personnel have painstakingly conducted surveillance of infectious diseases for centuries. This has created an enormous wealth of valuable data with which to test hypotheses about disease spread using mathematical models, making this a very exciting subject for research in applied mathematics.

Acknowledgements

It is a pleasure to thank Sigal Balshine, Will Guest, and the anonymous referee for helpful comments. The photograph in Figure 1 was taken by Claire Lees at the Guildhall in London, England. The weekly plague data plotted in Figure 2 were digitized by Seth Earn.

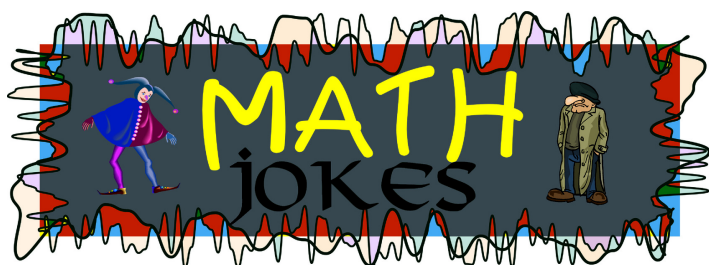
References

- [1] Kermack, W. O. and A. G. McKendrick, 1927: A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Series A*, **115**, 700–721.
- [2] Bailey, N. T. J., 1975: *The mathematical theory of infectious diseases and its application*. Griffin, London, 2nd edition.
- [3] Keeling, M. J. and C. A. Gilligan, 2000: Metapopulation dynamics of bubonic plague. *Nature*, **407**, 903–906.
- [4] Hughes-Hallett D. et al, 2002: *Applied Calculus*. John Wiley & Sons, Toronto, 2nd edition.
- [5] Anderson, R. A. and R. M. May, 1991: *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford.
- [6] Gleick, J., 1987: *Chaos*. Abacus, London.
- [7] Earn, D. J. D., P. Rohani, B. M. Bolker and B. T. Grenfell, 2000: A simple model for complex dynamical transitions in epidemics. *Science*, **287**, 667–670.
- [8] Bauch, C. T. and D. J. D. Earn, 2003: Transients and attractors in epidemics. *Proceedings of the Royal Society of London, Series B*, **270**, 1573–1578.

Practical Further Reading Suggestions from the Editors

The SIR Model for Spread of Disease web site at <http://www.math.duke.edu/education/ccp/materials/diffcalc/sir/contents.html> includes Maple, Mathematica and Matlab files.

A spreadsheet program that is easy to use is available from <http://ugrad.math.ubc.ca:8099/mathsheet/index.html>.



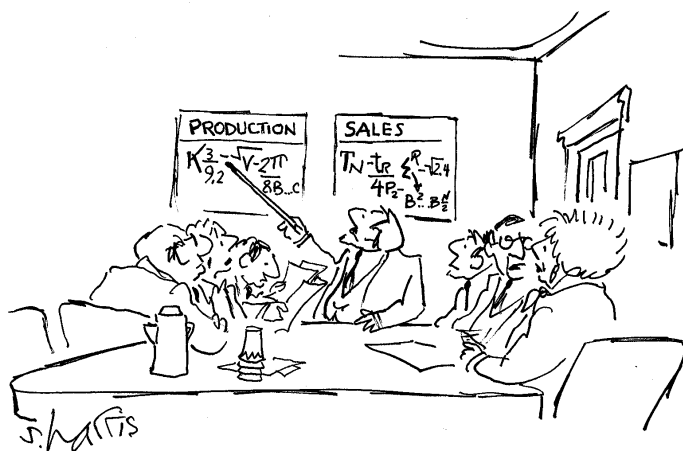
Q: What is a mathematician's pick when faced with the choice between poutine and eternal bliss in the afterlife?
 A: Poutine! Because nothing is better than eternal bliss in the afterlife, and poutine is better than nothing.

Q: What is a topologist?
 A: A person who cannot tell a doughnut from a coffee mug.

Q: Why did the mathematician have trouble computing $A^{-1}A$?
 A: Because he was having an identity crisis.

Absurdity of zero: there is no such a thing as nothing.

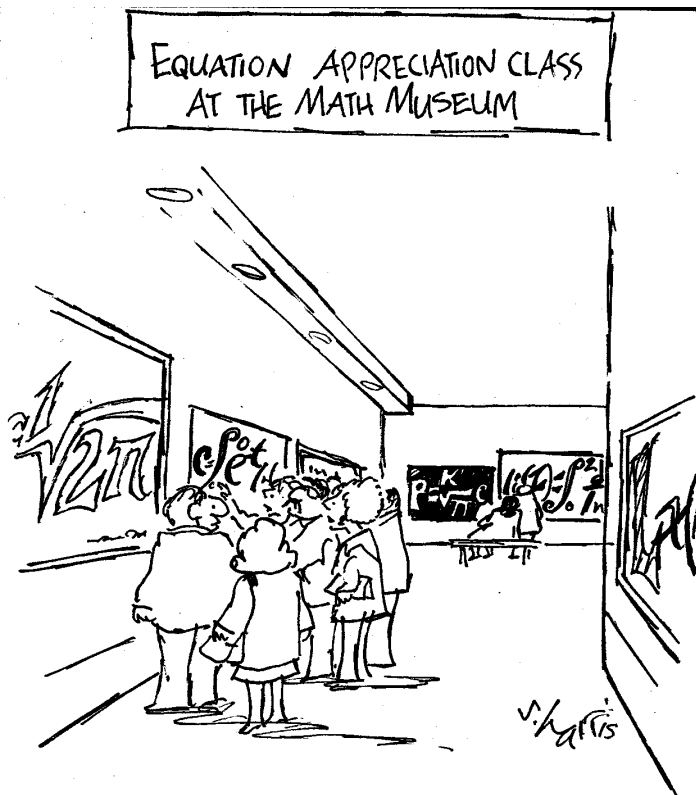
Q: What is yellow, sour, and equivalent to the axiom of choice?
 A: Zorn's lemon....



"HE NEVER LETS US FORGET THAT HE WAS A MATH MAJOR."

©Copyright 2004
 Sidney Harris

Q: What is normed, complete, and yellow?
 A: A Bananach space....



©Copyright 2004
 Sidney Harris



Cid, Bru, One

Jeremy Tatum[†]

One of the “Math Jokes” in the September, 2003 issue declared that there are 10 kinds of mathematicians: those who can think in binary and those who can’t.

At one time I was very definitely in the latter category—I couldn’t add one and one. Looking back, I realize that one of my difficulties was that I didn’t know of any easy *names* for numbers when written in binary form. For example, how do you pronounce the binary number 10 in the joke above? If you pronounced it “ten,” you missed the point of the joke!

So—what *are* binary numbers? We are all familiar with the decimal system, in which we have ten single-digit numbers: 0, 1, 2, . . . 9. If we want to write numbers greater than 9, we have to use two digits. Thus: 10. This enables us to write numbers as high as 99. If we want to write numbers greater than 99, we have to use three digits. Thus: 100. The number 100 is 10^2 and the number 1 000 is 10^3 , and so on. We are very familiar with all of that.

In the *binary* system (which is what computers use, even if we are unaware of it), we have only two “digits”—or “bits” as they are called in the binary system. These are 0 and 1. If we want to write numbers greater than 1, we have to use two bits, namely 10. This number is the binary equivalent of “two.” So now you can go back and read the joke, and perhaps have a good chuckle over it. The binary number 11 is equivalent to “three”—and, if we want to write numbers greater than this, we have to use three bits. Thus 100 is the binary equivalent of “four.” Here are the binary representations of the first eleven numbers:

zero	0
one	1
two	10
three	11
four	100
five	101
six	110
seven	111
eight	1000
nine	1001
ten	1010

You will see that 100 means 2^2 , 1000 means 2^3 , and so on.

Now I am going to ask you to add together five binary numbers, as follows:

```

101101
111011
110110
110011
100111

```

[†] **Jeremy Tatum** is a former professor in the Department of Physics and Astronomy of the University of Victoria. His e-mail address is universe@uvvm.uvic.ca.

If you are at all like me when I first encountered an apparently simple problem like this, you may find this surprisingly difficult. Perhaps you will at first try to convert each number into decimal form. The binary number 101 101, for example, is 45 when written in decimal. You might add them all up in decimal form, and then try to convert the decimal number back into binary form. Or you might try to add the numbers as binary numbers column by column in the usual way for adding numbers. Thus 1 + 1 is 10; and 10 and 1 is 11; and 11 and 1 is—oh, dear, I’ve lost count and I’m not quite sure what to do next. In case you want to persist, the answer is 11 111 000.

Many years ago I read an article (I’m afraid I cannot remember where or by whom—if anyone knows, please tell me) in which the author proposed names for numbers written in binary notation. To some it will sound like risible gibberish—yet I have found it surprisingly helpful for doing calculations in binary. Perhaps it is well known and in widespread use—though I have never met anyone who has heard of it. The difficulty I always had in thinking in binary was that I had no *names* for binary numbers such as 10, 100, 1000. The forgotten author supplied the key by giving *names* to these binary numbers.

As I recall, the powers of two were given names as follows:

1	one
10	bru
100	cid
1000	dag
10000	bru dag
100000	cid dag
1000000	hi

Examples of some other numbers:

11	bru one
101	cid one
1110	dag, cid bru
10101	bru dag, cid one
110011	cid bru dag, bru one

In addition to these names, I have also found it helpful to typeset long numbers in groups of three, with a small gap between each group, as is also standard typesetting practice with large decimal numbers.

Now let’s see how it works for adding the above five numbers. Although I don’t need to, I have, by way of further example, “pronounced” their “names” to the right of each.

101101	cid one dag, cid one
111011	cid bru one dag, bru one
110110	cid bru dag, cid bru
110011	cid bru dag, bru one
100111	cid dag, cid bru one
<u>11111000</u>	bru hi, cid bru dag

The only way to see how this works is to mumble things like “one, bru, cid, dag” under your breath while doing the addition. I’m going to try and transcribe my mumbblings to paper as I start by adding up the numbers *in the right-hand (least significant) column*, and then move on to the next column to the left. It goes something like this: “One, bru, bru one, cid; that’s zero, and bru to carry to the next column. Now the next column. Bru, bru one, cid, cid one, cid bru; that’s zero, and bru one to carry to the next column.”

. . .And so on and so on!

It's hard to describe the process in writing, but try a bit of mumbling yourself for a while and it will either become clear and quick very soon, or you'll collapse into helpless laughter.

How does *Multiplication* work? For example, try $11\ 011 \times 1\ 101$. That is, what is bru one dag, bru one times dag, cid one? In fact the multiplication is the easy part. You can just set it out as an ordinary long multiplication:

$$\begin{array}{r}
 11\ 011 \\
 \times 1\ 101 \\
 \hline
 11\ 011 \\
 1\ 101\ 100 \\
 11\ 011\ 000 \\
 \hline
 \end{array}$$

That part was easy. Now you have to add the last three binary numbers—and I'll leave you to do that in the same way as before. The answer is $101\ 011\ 111$.

Indeed the one, bru, cid nomenclature is really only useful in the *addition of several or many binary numbers*. In a *subtraction* problem, you just have two numbers to deal with—the number you start with, and the subtrahend (the number you subtract); bru, cid and dag are then of little help. I expect that you can subtract $1\ 001$ from $10\ 101$ to get $1\ 010$ with little difficulty.

To appreciate how helpful bru, cid and dag can be, there is no substitute for setting yourself several long addition problems in binary numbers and adding them when mumbling bru, cid dag under your breath. At first it will seem ridiculous—but after a short while you will soon find yourself adding long strings of binary numbers accurately and with great speed and confidence.

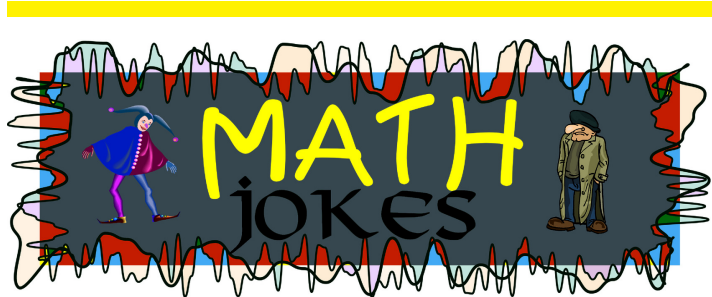
You may remember that Alice had some difficulty with Addition:

“Can you do Addition?” the White Queen asked. “What’s one and one and one and one and one and one and one and one and one and one?”

“I don’t know,” said Alice. “I lost count.”

“She can’t do Addition,” the Red Queen interrupted.¹

Alice, of course, should promptly have answered: *“Dag bru.”*



Differential equation of human life: equal laws for everybody, but better initial conditions for some of us.

Q: What is the code name of a spy mathematician?
A: $S\pi$

¹ Quoted from *Alice’s Adventures in Wonderland and Through the Looking Glass* by Lewis Carroll.

The chef instructs his apprentice: “You take two thirds of water, one third of cream, and one third of broth. . .”

The apprentice interrupts: “But that makes four thirds already!”

“Well—then you just take a larger pot!”



MOZART WRITING THE DIGITAL VERSION OF HIS SYMPHONY NO.38 IN D MAJOR

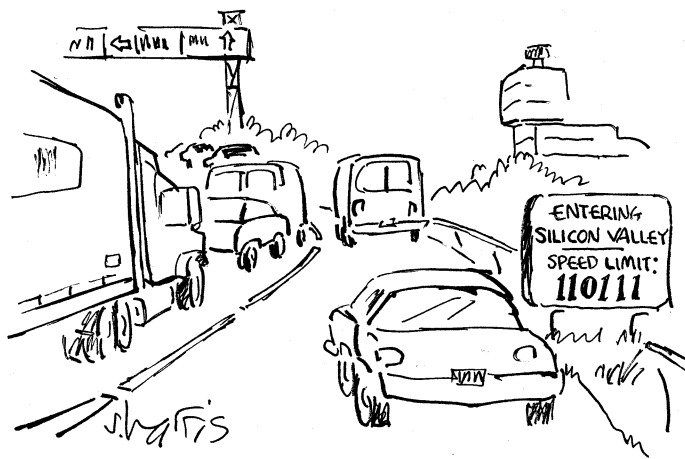
©Copyright 2004
Sidney Harris

Q: Why didn’t the permutation ever get hired for any jobs in the big city?

A: Because he didn’t commute.

Math is free but you need to download it yourself.

Marriage Math: one plus one forever.



©Copyright 2004
Sidney Harris



Little Bits of Geometry

Platonic Solids and the Golden Ratio*

Klaus Hoechsmann†

The best known of the Platonic Solids is certainly the *cube*: all its faces are the same (namely congruent squares) and at every “vertex,” i.e., corner-point, the same number of them meet (namely three). In the customary xyz -coordinate system, the vertices could be the eight points $(\pm w, \pm w, \pm w)$, and the cube then has side-length $2w$.

The six midpoints of the faces are $(\pm w, 0, 0)$, $(0, \pm w, 0)$, $(0, 0, \pm w)$. If we connect each of them to its two neighbours—assuming we are inside the cube to do this—we get another Platonic Solid, called the regular *octahedron*, because it has eight faces, which obviously form equilateral triangles. If we were curious enough to connect the mid-points of these triangular faces, we’d find—guess what?—another (smaller) cube. This is why cube and octahedron are said to be *dual* to each other.

The simplest Platonic Solid is the regular *tetrahedron*. It answers the question sometimes found on IQ tests: “how can you arrange six matches so as to form four equilateral triangles?” If we try to play the duality game with it, we just get a smaller tetrahedron pointing the other way. Thus, in some sense, the tetrahedron is self-dual.

The most handsome Platonic Solids are the regular *icosahedron*, with twenty triangular faces (five around each of the twelve vertices) and its dual, the regular *dodecahedron*. These are the ones we shall study. They and their connection to the Golden Ratio are the smashing finale of Euclid’s famous “Elements,” and have fascinated countless people before and since. Leonardo da Vinci, for instance, took the trouble to produce attractive drawings of them for Luca Pacioli’s book “*De divina proportione*.” Luca was his friend and mathematical tutor.

Luca’s “divine proportion” is, of course the Golden Ratio. Numerically it can be best described as the elusive limit of successive quotients in the Fibonacci sequence. Geometrically, a $u \times v$ rectangle is golden, if you can chop off a $v \times v$ square ($v < u$) and wind up with a residual rectangle similar to the original. If you think about it, this says:

$$\frac{v}{u} = \frac{u}{v} - 1.$$

Even if you don’t have time to think about it, we’ll take that as the definition of the Golden Ratio. Whether this illustrious title should go to u/v or v/u is a matter of taste.

To get the most out of this story, you will need some cardboard and string to make the model shown on the opposite page. The three rectangles should measure 2×5 toes, where “toe” is a newly invented unit that can be 1 inch, 2 or 3 centimetres, or anything else—depending on the thickness of

your cardboard (for the rigidity of the model). We imagine a z axis and a y axis going lengthwise through the middle of the vertical and the horizontal rectangles, respectively. The x axis will be pointing right at you through the middle of the third rectangle. Each rectangle has a 2 toe slit to allow another rectangle to be stuck through. One of them needs an extra cut from its edge to its slit; otherwise, you won’t be able to assemble them as shown.

The triangles ABC and CDE are equilateral. To see this, imagine the rectangles getting slimmer and slimmer, v shrinking toward zero and C wandering toward the z axis: at the end, ABC and CDE have morphed into the faces of a regular octahedron. As v expands again, their sides shrink in perfect unison, thus always remaining equal. For brevity, let these equilateral triangles be called “trigons.” There are eight of them in all. Can you see them?

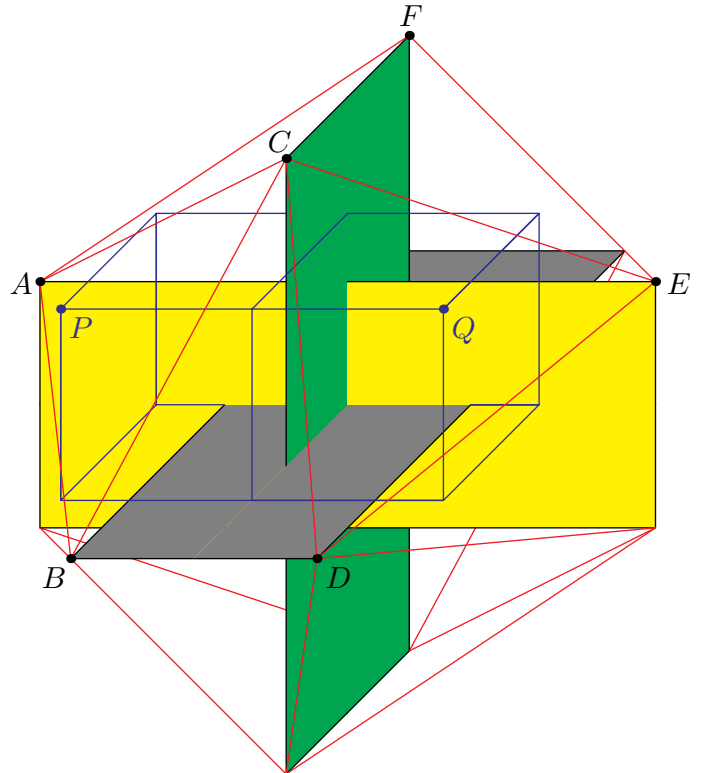


Figure 1: Cardboard-and-string skeletons for various attractive solids.

For the time being, let us forget them and concentrate on triangles like BCD , which we shall call “darts” for short. Each of our twelve vertices is the tip of exactly one dart, which is isosceles—sharing its legs with two neighbouring trigons, and sporting an acute vertex angle because u is greater than v (remember?). Altogether we have a polyhedron with twenty faces: twelve darts plus eight trigons.

Our first plan of action is to widen the darts until they, too, are equilateral; obviously there must be a stage where this occurs—we just need to find it. When we do, the only way to distinguish darts and trigons is indirectly, for instance, by saying that the midpoints of the latter always lie on space diagonals.

The Regular Icosahedron

To make the darts equilateral, we only have to tune u and v so as to make BD and DC equal in length. By symmetry, all the other lengths will then become equal as well. Now, BD has length $2v$, but for DC we have to do more work.

* This article originally appeared in Vector (the Journal of the BC Association of Math Teachers) and it is reprinted with permission.

† Klaus Hoechsmann is a Professor Emeritus at the University of British Columbia in Vancouver. You can find more information about the author and other interesting articles at <http://www.math.ubc.ca/~hoek/Teaching/teaching.html>.

The vector pushing C to D has the coordinates $[u - v, v, u]$, hence the square of its length is $(u - v)^2 + v^2 + u^2$, using the Pythagorean Theorem twice (that's how you find distances in three dimensions). Equating this to $4v^2$, the length of BD squared, we get $2v^2 = 2u^2 - 2uv$. Now divide by $2uv$ and stand back:

$$\frac{v}{u} = \frac{u}{v} - 1,$$

the signature of the Golden Ratio. In other words, *our (adjusted) model yields a regular icosahedron, if and only if it is made up of golden rectangles.*

For later use, let us square the golden signature $u/v - v/u = 1$, to conclude $u^2/v^2 + v^2/u^2 = 3$. Also notice the five (now) equilateral triangles CFA , CAB , CBD , CDE , and CEF surrounding the point C . Cuts made through their midpoints, parallel to the side opposite C , would show C as the tip of a pentagonal pyramid with equilateral sides (do you see?). Hence those midpoints, too, form a regular planar pentagon. All twelve of our vertices are ringed by three such pentagons: *these are the faces of the dual dodecahedron.*

We have deliberately kept our initial model asymmetric in order to be able to distinguish darts and trigons, but if you have enough energy (or help), you may wish to rebuild it with approximately golden rectangles, say 144 by 89 millimetres, or (if cardboard is in short supply) 89 by 55. These numbers are borrowed from the Fibonacci sequence: the exact Golden Ratio exists only in our minds (that's why we say it is "irrational").

The Circumscribed Sphere

Before we get into the handicrafts, let us try to picture what we are doing. For this phase, it would be good to have a geographical globe, showing the oceans and countries of our planet. We would orient it in xyz -space so that the positive z axis passes through the North Pole, the xy plane contains the Equator, and the positive x axis pierces the meridian of zero longitude (defined by Greenwich, England). This forces the negative y axis to pass through the Galapagos Islands (90° West, on the Equator).

Now we imagine our three rectangles inside that globe, their twelve corners ("vertices") just touching its surface from the inside. The coordinates of these vertices are $(0, \pm u, \pm v)$, $(\pm v, 0, \pm u)$, $(\pm u, \pm v, 0)$, where u and v will eventually be allowed to vary, but always with u greater than v . Pythagoras says $u^2 + v^2 = r^2$, where r is the radius of the globe.

To get oriented, let us begin looking at the five points visible from a satellite hovering high above Greenwich. We see $A = (0, -u, v)$ near Campeche, Mexico, $B = (u, -v, 0)$ about half-way between Brazil and Ghana, $C = (v, 0, u)$ near the Norwegian island of Jan Mayen, $D = (u, v, 0)$ just east of where the Congo River crosses the Equator, and $E = (0, u, v)$ in the Gulf of Bengal, near Calcutta. These five points form a kind of distorted W whose sides slope very gently while the central peak is unusually tall.

For our next adventure, we introduce the eight new points $(\pm u, \pm u, \pm v)$, vertices of a cube just touching the inside of the globe. Pythagoras reminds us that $3w^2 = u^2 + v^2$. We are most interested in $P = (w, -w, w)$ and $Q = (w, w, w)$ because they are the closest to the dart BCD containing Greenwich. In geographical terms, P is about half-way between Florida and the Azores, while Q lies near Baghdad. After this quick survey, we are ready to go about varying u and v again so as to make an even more amazing model. To include the new points in it, glue eight small paper cubes into the corners of the original 3-D "cross" made of 2×5 toe rectangles. Pythagoras says: the side-lengths of these cubes should be just under 1.6 toes (how does he know?).

The Regular Dodecahedron

We release u and v from their golden bondage and let them vary again. As promised, P and Q now step into the game, and we consider the five points B, P, C, Q, D . Can we arrange for

- 1) the five lengths BP, PC, CQ, QD, DB to be equal, and
- 2) the five points B, P, C, Q, D to lie in the same plane?

If we manage to adjust u and v so that $QD = DB (= 2v)$, then $BP = BD$ follows by symmetry. Moreover, QC bears the same relation to CF that QD bears to DB , and therefore will also have the length $2v$. Likewise for PC, CF , and PD . In other words, (1) will follow from the single equality $2v = QD$.

Again squaring both sides, we get $4v^2 = (w - u)^2 + (w - v)^2 + w^2$. Remembering that $3w^2 = u^2 + v^2 (= r^2)$, and dividing by 2, our equation becomes

$$2v^2 = u^2 + v^2 - uw - vw,$$

or

$$(u + v)w = u^2 - v^2.$$

Now we can divide by $(u + v)$ to get $w = u - v$, and on squaring, we breathlessly arrive at

$$w^2 = u^2 + v^2 - 2uv = 3w^2 - 2uv$$

and finally $uv = w^2$. Now what?

Now we divide $u^2 + v^2 = 3w^2$ by our latest result ($uv = w^2$) and stand back in awe:

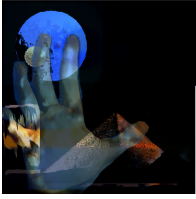
$$\frac{v}{u} + \frac{u}{v} = 3,$$

the signature of the of the *Square of the Golden Ratio*, as we observed near the end of our icosahedral story (remember that these are not the same values for u and v).

This clears up Question (1), but what about (2)? It follows from the fact that regular dodecahedra, with flat pentagonal faces, are the duals of regular icosahedra. We enlarge such a dual to fit just inside the globe, and label the vertices of one of its faces B', P', C', Q', D' (not using the unadorned B, P, C, Q, D in order to avoid unwarranted conclusions). Then we put B' and D' on the Equator so that the Greenwich Meridian passes right through the middle between them, and C' lies in the Northern Hemisphere. As we have seen in connection with the icosahedron, the point Q' then lies above the midpoint of a trigon, hence on a space diagonal. The lengths $B'D'$ and $D'Q'$ are given as equal, and our calculations show that there is only one way this can happen: the points $BPCQD$ must coincide with their "primed" counterparts.

To make an actual cardboard model, we could again steal numbers from the Fibonacci sequence, going two steps down this time, say, using 144 and 55 (do you see why?). But now, the paper cubes are essential, and we connect their outward vertices with the bases of the former darts.

Seeing the golden rectangles, Fibonacci numbers, and pentagons, regular readers of this magazine will suspect (correctly) that this is yet another digression from the story of the Rose and the Nautilus. While the last one was caused by people hollering for ratios, the present one tries to refute charges that we are deliberately avoiding algebra. In the next issue, the main story will be back on track, we hope.



Math Discovery

Kolmogorov, Turbulence, and British Columbia

Bob Stewart[†] and Chris Garrett*

We were delighted to read Alexander Melnikov’s splendid account, in the last issue of *π in the Sky*, of the life of the great Russian mathematician Alexander Nikolaevich Kolmogorov. The main purpose of this short article is to describe how a study in B.C. waters provided the first and strongest evidence in support of Kolmogorov’s famous 1941 “Two-Thirds Law.”

As nicely summarised by Melnikov, this law predicts that in every turbulent flow (such as we see in rushing rivers or vigorously boiling pans of water, for example), “the mean square difference of the velocities at two points a distance r apart is proportional to $r^{2/3}$ (Figure 1).” Quite apart from the elegance of this result, it provides a key to understanding and predicting how pollutants and other substances are dispersed.

Explaining where this prediction comes from provides an opportunity to introduce the powerful technique of *dimensional analysis*. This really just says that we’d better get the units right. In the present case, we suppose that the mean square velocity difference depends on the separation r and something characterizing the strength of the turbulence. Now in turbulent flow the factor that matters most is the rate of transfer of energy from large eddies, via eddies of ever-decreasing size, down to very small eddies that dissipate the energy through the action of viscous forces. This energy transfer rate per unit mass, called the *energy dissipation rate*, has units m^2s^{-3} (though we could use other units than a metre for length and a second for time). Denoting the energy dissipation rate by ϵ , and arguing that the mean square velocity difference depends only on this and the separation r , then the mean square velocity difference must be proportional to $\epsilon^{2/3}r^{2/3}$ in order to have units of m^2s^{-2} .

A closely associated result, also derived by dimensional reasoning, is that the kinetic energy spectrum, i.e. the distribution of energy over different scales, is proportional to $\epsilon^{2/3}k^{-5/3}$ [5]. Here k is something called the wavenumber, and is just proportional to the inverse of the eddy scale. The result comes from assuming that the energy density per unit wavenumber, with dimension m^3s^{-2} , depends only on ϵ and k . The assumption is likely to be valid in what is known as the “inertial subrange” separating the scale at which energy is injected into large eddies from the much smaller scale at which viscous dissipation occurs.

This amazing prediction of what is now known as the “Kolmogorov spectrum” needed to be verified in a situation with as large an inertial subrange as possible so that the $-5/3$ power law could be confirmed with confidence. Interestingly, the exceptionally strong tidal currents near Seymour Narrows north of Campbell River on Vancouver Island (with a peak speed of more than 6 m s^{-1}) offered one of the best opportunities on earth, with a range of nearly 10^5 between the energy input scale at about 100 metres and the dissipation at a scale of a millimetre or so. Measurements obtained there in the

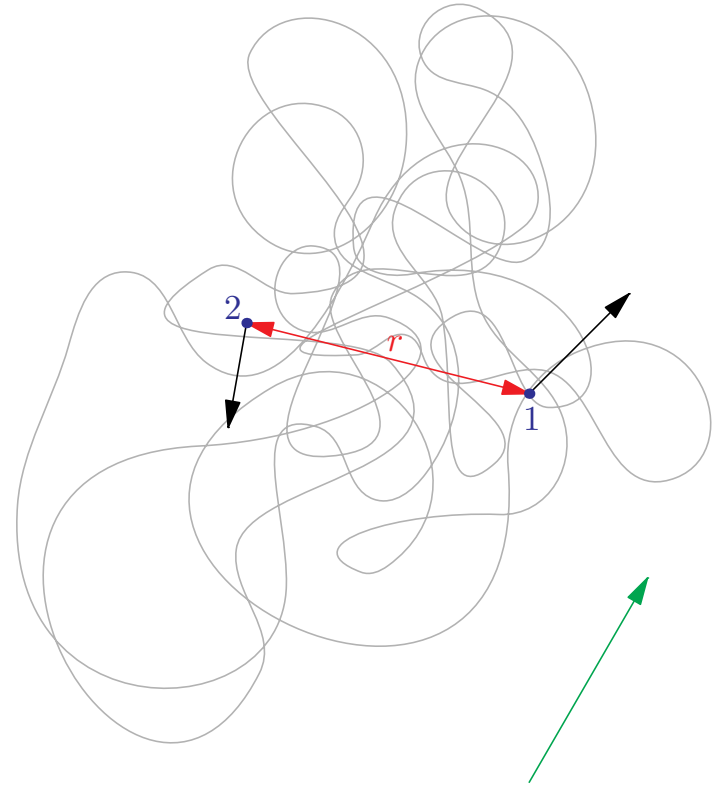


Figure 1: Two particles a distance r apart in a turbulent flow have different velocities, indicated by the black arrows at positions 1 and 2. The green arrow in the lower half of the figure depicts the velocity at position 1 minus the velocity at position 2, which can be thought of as the velocity with which a particle at position 1 is moving away from a particle at position 2. We are interested in the average of the square of this velocity difference.

late 1950s by a research group from the Pacific Naval Laboratory in Esquimalt, B.C., did indeed provide dramatic confirmation of Kolmogorov’s predicted spectral shape, and also determined the value of the constant that multiplies the formula $\epsilon^{2/3}k^{-5/3}$ [3]. The sensitive probes that measured the turbulent velocity fluctuations in the water were mounted at the front of a converted minesweeping paravane that could be towed behind a ship and well below the sea surface. A modern towed instrument, with improved sensors for velocity and many other things, is shown in Figure 2.

A graph reproduced from the original paper by Grant *et al.* [3] is shown in Figure 3. The straight line with slope $-5/3$ fits the data well on a log–log plot; if the spectrum is ϕ and the wavenumber k , then

$$\log \phi = -\frac{5}{3} \log k + \text{constant}$$

means that $\phi \propto k^{-5/3}$, as predicted. The $-5/3$ slope seen in Figure 3 actually continues to much smaller values of k .

This being science, the story doesn’t stop there. The senior author of this letter, and co-author of the B.C. paper [3], remembers presenting the results at a turbulence meeting in Marseilles in 1961 at which Kolmogorov himself expressed strong reservations about his own theory, on the grounds that turbulence is too intermittent to be described by simple power laws [6]. The audience was presumably puzzled by the simultaneous experimental confirmation and withdrawal of a theory! In fact, it seems that the original arguments are valid for second-order statistics, such as the $2/3$ power

[†] **Bob Stewart** is former Director of the Institute of Ocean Sciences at Sidney, British Columbia.

* **Chris Garrett** is Lansdowne Professor of Ocean Physics at the University of Victoria. His e-mail address is garrett@phys.uvic.ca.



Figure 2: The Towed Ocean Microstructure Instrument (TOMI) developed by Rolf Lueck at the University of Victoria. Here it is being recovered from the water in Knight Inlet after making simultaneous measurements of the turbulent small-scale fluctuations of velocity and temperature, as well as measuring the backscatter of high-frequency sound from biological organisms and the turbulence. The masts at right angles to the main body of the instrument contain other sensors and become vertical in the water, with the orange one on top.

law for the mean square of the velocity difference, or the energy spectrum, but do need modification for higher order statistics such as the mean cube velocity difference (e.g. [4]).

A similar situation occurred more recently in the research of the junior author of this article. A dimensional argument akin to that of Kolmogorov led to the prediction that air bubbles in the turbulent region beneath breaking ocean waves should be distributed over different sizes with a “size spectrum” proportional to (bubble radius)^{-10/3} [2], so that there are many more small bubbles than large ones. The argument assumed that air entrained in breaking waves is broken up into ever smaller bubbles by the action of turbulence. But herein lay the problem: whereas energy is cascaded to ever smaller scales by turbulence and eventually removed by viscosity, bubble break-up is eventually limited by the strong influence of surface tension for small bubbles, but the bubbles don’t then disappear, at least not quickly. We thus predicted and withdrew our $-10/3$ power law all in the same paper, and were then surprised when, a couple of years later, the $-10/3$ power law was confirmed observationally [1]!

There is clearly more to be understood in both of these situations. While dimensional reasoning is very powerful and leads to useful predictions in physics (much like Darwin’s theory of natural selection in biology!), a full understanding of turbulent flows is still not available. According to an apocryphal story, Werner Heisenberg said on his deathbed “When I meet God, I am going to ask him two questions: Why relativity? And why turbulence? I really believe he will have an answer for the first.” The general problem of fluid flow is, in fact, one of the seven Millennium Prize Problems for which the Clay Institute of Mathematics is offering \$1 million (http://www.claymath.org/Millennium_Prize_Problems/Navier-Stokes_Equations/).

References

- [1] Dean, G. B. and M. D. Stokes, 2002: Scale dependence of bubble creation in breaking waves. *Nature*, **418**, 839–844.
- [2] Garrett, C., M. Li and D. Farmer, 2000: The connection between bubble size spectra and energy dissipation rates in the upper ocean. *J. Phys. Oceanogr.*, **30**, 2163–2171.
- [3] Grant, H. L., R. W. Stewart and A. Moilliet, 1961: Turbulence spectra from a tidal channel. *J. Fluid Mech.*, **12**, 241–263.
- [4] Jiménez, J., 2000: Turbulence. In *Perspectives in Fluid Dynamics*, Ed. Batchelor, G. K., H. K. Moffatt and M. G. Worster, C. U. P., 231–288.
- [5] Kolmogorov, A. N., 1941: The local structure of turbulence in incompressible viscous fluids at very large Reynolds numbers. *Dokl. Akad. Nauk. SSSR*, **30**, 301–305. Reprinted in 1991: *Proc. R. Soc. Lond. A*, **434**, 9–13.
- [6] Kolmogorov, A. N., 1962: A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high Reynolds number. *J. Fluid Mech.*, **13**, 82–85.

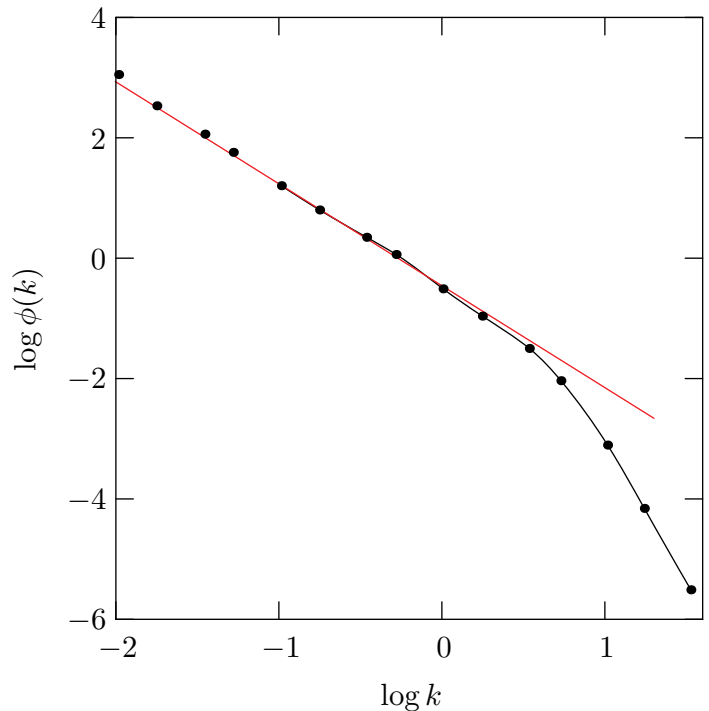


Figure 3: The spectrum of turbulent velocity fluctuations shows how the turbulent energy is distributed over various scales, with their size being the inverse of the wavenumber k . Here, experimental data obtained in 1959 are displayed in a log–log plot, along with a straight line with a slope of $-5/3$. The wavenumber k has been scaled to have a value of 1, so that $\log k = 0$, for the length scale of a few millimetres at which viscous forces start to matter. The roll-off of the data for $\log k$ greater than zero is caused by viscosity.



Vedic Mathematics and the Calculations of Guru Tīrthajī*

Jan Hogendijk†

Introduction

This article was written as a reaction to several publications in the Netherlands on methods of mental calculation, which have been presented as “Vedic mathematics.” The adjective “Vedic” suggests that these methods date back to the ancient Vedic period before 500 BCE.¹ In this article I will first give two examples of genuine Vedic mathematics. After a brief intermezzo on the decimal position system and decimal fractions, the so-called “Vedic” methods of mental calculation will be identified as the inventions of a twentieth-century Guru. More information on the subject can be found in [3].

The March 2003 issue of *π in the Sky* also includes an article on Vedic Mathematics. It was written by Jeganathan Sriskandaraiah from Madison Area Technical College.

The Vedas

The Vedas are ancient religious and philosophical scriptures that originated in India in the Vedic period. These texts were composed in an Indo-European language that was the predecessor of Sanskrit. Initially the Vedas were transmitted orally, but later on they were recorded in writing. The word Veda has the same root as the Dutch word “weten,” the German “wissen” (meaning: to know), and the English “witness,” “unwitting” and “to wit.” The Vedic texts refer to rituals and fire-altars. Rules for the precise construction of these altars are given in the *Sulba-sūtra’s* (rules for the measuring cord), which also belong to the authentic Vedic literature. Some of these sutras deal with the geometric knowledge that is necessary for the construction of brick altars of different shapes and sizes.

Examples of Vedic Mathematics

The following four sutras are taken from the Baudhāyanaśulbasūtra, which was composed before 500 BCE. The text and translation are adapted from the edition of Sen and Bag, which is based on old Indian manuscripts in Benares, Munich, London and Ujjain. In this work, the sutras are stated without further explanation and without proof. I will not provide

any commentary either, so as to leave the reader the pleasure of pondering the meaning of these sutras. (See [5, pp. 18, 19, 78–80, 150–154, 164–169]; the numbering from [5] is used here.)

(1.9) samacaturaśrasyākṣṇayārajjurdvistāvātīm bhūmiṃ karoti

The diagonal of a square produces double the area.

(1.12) dīrghacaturaśrasyākṣṇayārajjūḥ pārśvamānī tiryānmānī yatpṛthagbhūte kurutastadubhayaṃ karoti

The areas produced separately by the length and the breadth of a rectangle together equal the area produced by the diagonal.

(1.13) tāsām trikacatuskayordvādaśikapāñcikayoḥ pañcadaśikaṣṭikayoḥ saptikacaturviṃśikayor dvādaśikapāñcatriṃśikayoḥ pañcadaśikaṣaṭtriṃśikayorityetāsūpalabdhiḥ

This is observed in rectangles having sides three and four, twelve and five, fifteen and eight, seven and twenty-four, twelve and thirty-five, fifteen and thirty-six.

(2.12) pramāṇaṃ tṛtīyena vardhayettacca caturthenātmacatuṣṭriṃśonena savīṣeṣaḥ

The measure is to be increased by its third and this again by its own fourth less the thirty-fourth part [of the latter (Ed.)]; this is the diagonal of a square.

The last sutra has been interpreted as a close approximation to $\sqrt{2}$. Historians generally assume that this approximation was discovered not by accident but by some clever mathematical method.

Historic Intermezzo: Discovery of the Decimal System

In the Vedic sutras, numbers are expressed in words. Indo-European roots of number words can be recognized in the preceding quotations: tri = three, catur = quattuor in Latin (cf. quadrilateral), pañca = pente in Greek (cf. pentagon), dva = two, etc.

From the third century BCE onwards, that is to say, after the Vedic period, number symbols were used in North India for the numbers one through nine. These symbols are the predecessors of the modern 1, 2, 3, . . . , 9. Other symbols were used for ten, twenty. . . , hundred, two hundred. . . , and a symbol for zero had not yet been invented.

The oldest known symbols for “zero” occur in astronomical computations in the sexagesimal place-value system, on Babylonian clay tablets written in last three centuries BCE. The zero was used as a place holder, as in “29 degrees, 0 minutes, and 43 seconds.” The sexagesimal system, including the zero, was adapted by the Greek astronomers in the second century BCE, and are still to be found today in angles (degrees, minutes, and seconds) and units of time (hours, minutes and seconds). From 300 BCE onwards, astronomical methods from Babylon and ancient Greece, including the sexagesimal system, were transmitted to India as well. Probably around 300 CE (six centuries later), the Indian astronomers developed the modern decimal place-value system for integers. This system was transmitted into Arabic in the eighth and early ninth century, but it never became popular in the medieval Islamic world. In the eleventh and twelfth centuries, the Indian number symbols were introduced in Christian Europe. The final victory of the system began in the Renaissance as a result of the increase of trade and commercial transactions. Decimal fractions were introduced by a few medieval Islamic mathematicians in order to write the end results of computations, but they were hardly used in the computations themselves (which were usually done in the sexagesimal

* This article was translated and revised by the author from an original version published in Dutch in the *Nieuwe Wiskrant* vol. 23 no. 3 (March 2004), pp. 49–52.

† Jan Hogendijk teaches mathematics in Utrecht (Netherlands) and history of mathematics in Leiden (Netherlands) and Dhahran (Saudi Arabia). His research interests focus on the history of mathematical sciences in medieval Islamic civilization. See www.math.uu.nl/people/hogendijk for his publications. His e-mail address is hogend@math.uu.nl.

¹The notation BCE stands for Before Common Era and CE stands for Common Era. The BCE/CE system is synonymous with the much more common BC/AD system (Ed.).

system). In Europe, decimal fractions were independently discovered by the Flemish mathematician Simon Stevin, who published his discovery in 1585 in Leiden in Holland. Decimal fractions probably arrived in India in the seventeenth century, when French astronomical tables were studied by astronomers in and around Delhi. During the English colonial regime, decimal fractions were of course taught in schools.

The “Vedic” Mathematics of Guru Tirthajī

The “Vedic” methods of mental calculations in the decimal system are all based on the book *Vedic Mathematics* by Jagadguru (world guru) Swami (monk) Śrī (reverend) Bhāratī Kṛṣṇa Tīrthajī Maharaja, which appeared in 1965 and which has been reprinted many times [4].

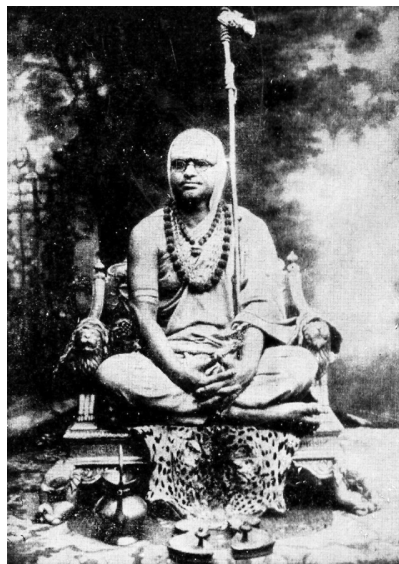


Figure 1: Guru Tīrthajī

This Guru was born in 1884 near Madras as Venkatraman Shastri. He studied English, mathematics, Sanskrit, and other subjects at various schools and colleges in India. In 1911 he quit his teaching job in order to devote himself completely to study and meditation in an ashram near Mysore. In 1919 he was initiated as a sannyasi (monk) and received the name Bhāratī Kṛṣṇa Tīrtha. Shortly afterwards, he became the leader of ashrams. He travelled extensively through India, and, at the invitation of the Self-Realization Fellowship of Yogananda, also to the United States in 1958.

Bhāratī Kṛṣṇa Tīrtha died in 1960 and his main work *Vedic Mathematics* appeared posthumously.

The book contains sixteen brief sutras that can be used for mental calculations in the decimal place-value system. An example is the sutra *ekādhikena purvena*, meaning: *by one more than the previous one*. The Guru explains that this sutra can for example be used in the mental computation of the period of a recurring decimal fraction such as $1/19 = 0.052631578947368421$, as follows:

The last decimal of the period is 1. We now apply the sutra *by one more than the previous one* to the number 19. The *previous one* is the penultimate decimal in the decimal representation of 19, so *one more than the previous one* is 2. We multiply the last decimal of the period by 2 and get the penultimate decimal of the period, and so on. Thus we find the final decimals $\dots 8421$. We now compute $8 \times 2 = 16$, write the 6 before the 8, and remember the 1. We add this 1 to the next product 6×2 to obtain $\dots 368421$, and so on.²

²The Guru does not give a proof of why the method works. Here is a simple number-theoretic proof: let n_k be the k -th decimal in the decimal expansion of $1/19$ (thus $n_1 = 0, n_2 = 5, n_3 = 2 \dots$). Then for $k = 1, 2, \dots$ we have $10^k = 19(10^{k-1}n_1 + 10^{k-2}n_2 + \dots + 10n_{k-1} + n_k) + m_k$ where m_k is an integer such that $0 < m_k < 19$; thus $m_1 = 10, m_2 = 5, m_3 = 12 \dots$. Because $19n_k + m_k$ must be a multiple of 10, which implies that $-n_k + m_k$ must be a multiple of 10, we see that n_k is equal to the last digit of m_k , which is the remainder of 10^k after division by 19. Because $10^{18} \equiv 1 \pmod{19}$, we have $n_{18} = m_{18} = 1$. Since $2 \cdot 10^{k+1} = 20 \cdot 10^k \equiv 10^k \pmod{19}$, we have $2m_{k+1} \equiv m_k \pmod{19}$, hence $n_{17} = m_{17} = 2, n_{16} = m_{16} = 4, n_{15} = m_{15} = 8, m_{14} = 16, \text{ and } n_{14} = 6$. Then

The same method can be used for $1/29$, but now we have to take 3 as *one more than the previous one*.

The word “Vedic” in the title of the book suggests that these calculations are authentic Vedic mathematics. The question now arises how the Vedic mathematicians were able to write the recurrent decimal fraction of $1/19$, while decimal fractions were unknown in India before the seventeenth century. We will first investigate the origin of the sixteen sutras. We cite the Guru himself [4, pp. xxxviii–xxix]:

“And the contemptuous or, at best, patronizing attitude adopted by some so-called orientalists, indologists, antiquarians, research-scholars etc. who condemned, or light heartedly, nay irresponsibly, frivolously and flippantly dismissed, several abstruse-looking and recondite parts of the Vedas as ‘sheer nonsense’ or as ‘infant-humanity’s prattle,’ and so on, ... further confirmed and strengthened our resolute determination to unravel the too-long hidden mysteries of philosophy and science contained in ancient India’s Vedic lore, with the consequence that, after eight years of concentrated contemplation in forest-solitude, we were at long last able to recover the long lost keys which alone could unlock the portals thereof.

“And we were agreeably astonished and intensely gratified to find that exceedingly tough mathematical problems (which the mathematically most advanced present day Western scientific world had spent huge lots of time, energy and money on and which even now it solves with the utmost difficulty and after vast labour involving large numbers of difficult, tedious and cumbersome ‘steps’ of working) can be easily and readily solved with the help of these ultra-easy Vedic Sūtras (or mathematical aphorisms) contained in the Paṛiśīṣṭa (the Appendix-portion) of the Atharvaveda in a few simple steps and by methods which can be conscientiously described as mere ‘mental arithmetic.’ ”

The Guru adds [4, p. xl]:

“The Sūtras (aphorisms) apply to and cover each and every part of each and every chapter of each and every branch of mathematics (including arithmetic, algebra, geometry—plane and solid, trigonometry—plane and spherical, conics—geometrical and analytical, astronomy, calculus—differential and integral etc.). In fact, there is no part of mathematics, pure or applied, which is beyond their jurisdiction.”

Concerning the applicability of the sixteen sutras to all mathematics, we can consult the Foreword to *Vedic Mathematics* written by Swami Pratyagātmananda Saraswati. This Swami states that one of the sixteen sutras reads *Calanakalana*, which can be translated as *Becoming*. The Guru himself translates the sutra in question as “differential calculus” [4, p. 186]. Using this “translation” the sutra indeed promises applicability to a large area in mathematics; but the sutra is of no help in differentiating or integrating a given function such as $f(x) = 1/\sin x$.

Sceptics have tried to locate the sutras in the extant *Paṛiśīṣṭa*’s (appendices) of the *Atharva-Veda*, one of the four Vedas. However, the sutras have never been found in authentic texts of the Vedic period. It turns out that the Guru had “seen” the sutras by himself, just as the authentic Vedas were, according to tradition, “seen” by the great Rishi’s or seers of ancient India. The Guru told his devotees that he had “reconstructed” his sixteen sutras from the *Atharva-Veda* in the eight years in which he lived in the forest and spent his time on contemplation and ascetic practices. The book *Vedic Mathematics* is introduced by a General Editor’s Note [4, p. vii], in which the following is stated about the sixteen sutras: “[the] style of language also points to their discovery by Śrī Swāmijī (the Guru) himself.”

$m_{13} \equiv 2 \cdot 16 \equiv 2 \cdot (10 + 6) \equiv 20 + 2 \cdot 6 \equiv 1 + 2 \cdot 6 \pmod{19}$. Thus the Guru correctly adds the penultimate digit 1 of the number 16 to two times the last digit 6. The same argument works for all other $m_i \geq 10$.

The Guru finishes his book on *Vedic Mathematics* with a (supposedly ancient) Indian verse that is a hymn to Krishna or Shankara. The Guru states that the verse also contains the value of $\pi/10$ expressed in what he calls an *Alphabetical Code-Language* [4, pp. 360]. We finish this article with a brief analysis of this verse, derived from [1]. In the verse, the decimals of π are encoded according to the Indian Kaṭapayādi system. This system was used by Indian astronomers in order to encode complicated numbers (mathematical and astronomical constants) into words and verses that were easy to remember. The system was developed in Kerala in South-India, probably in the ninth century CE. This dating already presents a problem in connection with the supposed “Vedic” origin of the verse. In order to explain an even more serious problem with the verse, we first give some details about the Kaṭapayādi system and then present the verse itself.

The Kaṭapayādi system works as follows. To each digit between 0 and 9, a small group of consonants of the Sanskrit alphabet is assigned (see the list below). This Kaṭapayādi system can be compared to, but is more complicated than the modern telephone keypad system 1 = ABC, 2 = DEF, and so on. To encode a digit between 0 and 9, we choose any Sanskrit consonant from the group belonging to this digit. For the digit 1, for example, we can choose among the four possibilities k, ṭ, p, and y (this is why the system is called Kaṭapayādi). Of course, each consonant can encode at most one digit. In order to dispel all illusions about the Guru’s verse for π , it is necessary to list the possible encodings in detail (abbreviations such as ḍh represent a single Sanskrit consonant):

- 1 = k, ṭ, p, y; 2 = kh, th, ph, r; 3 = g, ḍ, b, l;
 4 = gh, ḍh, bh, v; 5 = ṅ, ṇ, m, ś; 6 = c, t, ṣ;
 7 = ch, th, s; 8 = j, d, h; 9 = jh, dh; 0 = ṅ, n.

To make words and verses, the following rules are added. If two successive consonants (such as ‘vr’) are placed in a word only the last consonant r counts, and the v is ignored. A double consonant such as jj counts only once. Vowels and consonants that do not occur in the list do not have a numerical meaning, so they can be inserted anywhere without changing the numerical value of the word or verse. Thus the system was very flexible, and it was not difficult to compose words or verses for complicated (series of) numbers. The digits were encoded in reverse order.

A fifteenth century example from Kerala in South-India is the verse:

*vidvāms
 tunnabalah
 kavīśanicayah
 sarvārthasūlasthiro
 nirviddhānganarendraru*
 [2, p. 53].

The verse means:

The wise ruler whose army has been struck down gathers together the best of advisers and remains firm in his conduct in all matters; then he shatters the king whose army has not been destroyed [2, pp. 57–58].

This verse consists of five Sanskrit words, and the letters that carry numerical values are vv=44, tnl=6033, kvśncy=145061, sv(th)śl(th)r=7475372, nv(dh)gnrrr=04930222. Because the digits were encoded in reverse order, the five words encode the sexagesimal numbers

$$\begin{aligned} &44/60^3, \\ &33/60^2 + 06/60^3, \\ &16/60 + 05/60^2 + 41/60^3, \\ &273/60 + 57/60^2 + 47/60^3, \\ &2220/60 + 39/60^2 + 40/60^3, \end{aligned}$$

which astronomers wanted to memorize because they occur in an expression used for computation of the Indian sine, equivalent to a modern Taylor series. These numbers are written in terms of the first three powers of $1/60$, and they are rounded values of (in modern terms) $\frac{90}{11!}(\frac{\pi}{2})^{10}$, $\frac{90}{9!}(\frac{\pi}{2})^8$, $\frac{90}{7!}(\frac{\pi}{2})^6$, $\frac{90}{5!}(\frac{\pi}{2})^4$, and $\frac{90}{3!}(\frac{\pi}{2})^2$, respectively.

Now we know enough about the authentic Katapayādi system to identify the origin of the Guru’s verse about $\pi/10$. Here is the verse: (it should be noted that the abbreviation ṛ represents a vowel in Sanskrit):

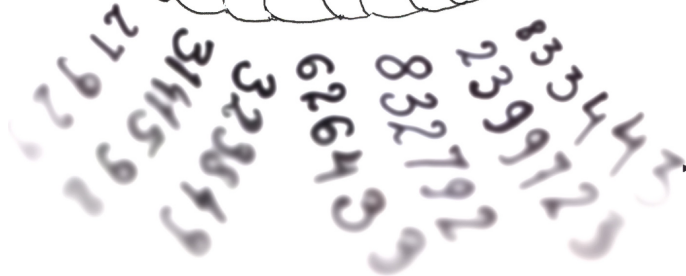
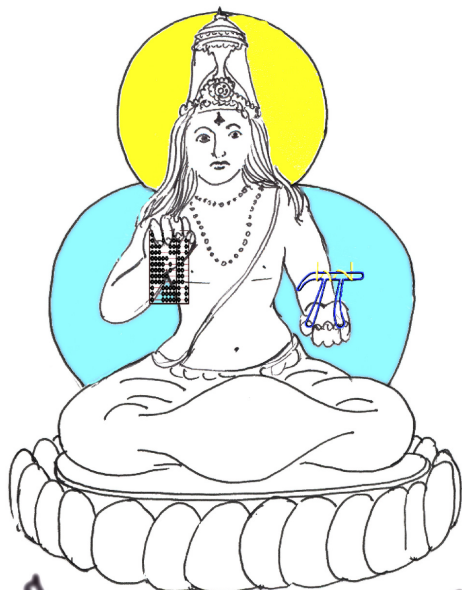
*gopī bhāgya madhuvrata
 śṛṅgiśo dadhī sandhiga
 khala jīvita khātāva
 gala hālā rasandhara.*

According to the Guru, decoding the verse produces the following number:

31415 92653 58979 32384 62643 38327 92.

In this number we recognize the first 31 decimals of π (the 32-th decimal of π is 5).

In the authentic Kaṭapayādi system, the decimals are encoded in reverse order. So according to the authentic system, the verse is decoded as



©Copyright 2004
 Wieslaw Krawcewicz

We conclude that the verse is not medieval, and certainly not Vedic. In all likelihood, the Guru is the author of the verse.

Conclusion

There is nothing intrinsically wrong with easy methods of mental calculations and mnemonic verses for π . However, it was a miscalculation on the part of the Guru to present his work as ancient Vedic lore. Many experts in India know that the relations between the Guru's methods and the Vedas are faked. In 1991 the supposed "Vedic" methods of mental calculation were introduced in schools in some cities, perhaps in the context of the political program of saffronisation, which emphasizes Hindu religious elements in society (named after the saffron garments of Hindu Swamis). After many protests, the "Vedic" methods were omitted from the programs, only to be reintroduced a few years later. In 2001, a group of intellectuals in India published a statement against the introduction of the Guru's "Vedic" mathematics in primary schools in India.

Of course, there are plenty of real highlights in the ancient and medieval mathematical tradition of India. Examples are the real Vedic sutras that we have quoted in the beginning of this paper; the decimal place-value system for integers; the concept of sine; the cyclic method for finding integer solutions x, y of the "equation of Pell" in the form $px^2 + 1 = y^2$ (for p a given integer); approximation methods for the sine and arctangents equivalent to modern Taylor series expansions; and so on. Compared to these genuine contributions, the Guru's mental calculation are of very little interest. In the same way, the Indian philosophical tradition has a very high intrinsic value, which does not need to be "proved" by the so-called applications invented by Guru Tīrthajī.

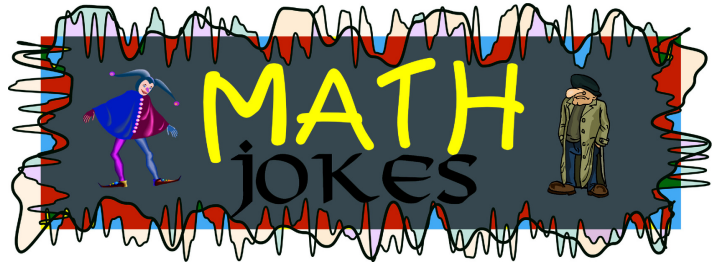
Acknowledgements

I wish to thank Prof. Dr. Michio Yano, Kyoto Sangho University, for sending me the article by K. Chandra Hari, and Dr. Len Berggren, Simon Fraser University, Canada, for correcting my translation; Dr. Kim Plofker, Utrecht and an anonymous referee for their comments on an earlier version of this paper.

References

- [1] Chandra Hari, K., 1999: A critical study of Vedic mathematics of Śankaracārya Śrī Bhāratī Kṛṣṇa Tīrthajī Mahārāj. *Indian Journal of History of Science* **34**, 1–17.
- [2] Gold, D. and D. Pingree, 1991: A hitherto unknown Sanskrit work concerning Mādhava's derivation of the power series for sine and cosine. *Historia Scientiarum*, **42**, 49–65.
- [3] Gupta, R. C., 1994: Six types of Vedic Mathematics. *Gaṇita Bhāratī* **16**, 5–15.
- [4] Jagadguru Swāmī Śrī Bhāratī Kṛṣṇa Tīrthajī Mahārāja, 1992: *Vedic Mathematics*. Delhi: Motilal Banarsidas, revised edition.
- [5] Sen, S. N. and A. K. Bag, 1983: *The Śulbasūtras*. New Delhi: Indian National Science Academy.
- [6] Interesting web site on Vedic ritual: <http://www.jyotistoma.nl>.

- [7] Pamphlet "Neither Vedic Nor Mathematics," against the introduction of "Vedic" method of computations in school teaching: <http://www.sacw.net/DC/CommunalismCollection/ArticlesArchive/NoVedic.html>.



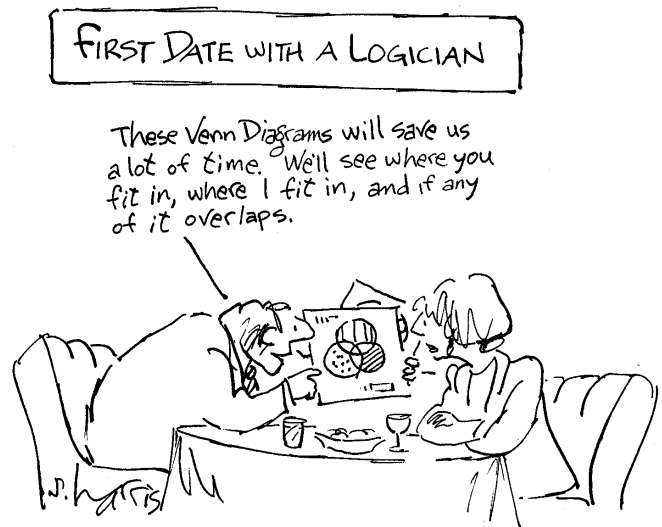
Q: "How does one insult a mathematician?"
A: "Your brain is smaller than any $\epsilon > 0!$ "

Two mathematicians are studying a convergent series.
The first one says: "Do you realize that the series converges even when all the terms are made positive?"
The second one asks: "Are you sure?"
"Absolutely!"

Q: What does a mathematician present to his fiancée when he plans to propose to her?
A: A polynomial ring!

Q: Why don't mathematicians ever spend time at the beach?
A: Because they don't need the sun to get a tan, they have sin and cos.

Q: Why is it that derivatives can never finish telling a story?
A: Because they are always going off on tangents.



©Copyright 2004
Sidney Harris



Tribonacci in the Sky: A Mathematical Mountain Walk

Achim Clausing[†]

Going Downhill

Here is a game that you can find in some elementary school math books:

“Every child in the class chooses four numbers. Below each number, he or she writes the difference between this number and its neighbour to the right. (The last number gets the first one as its right-hand neighbour.) This calculation is called a *downward step*. Now the children repeat stepping down until they arrive at four equal numbers. The winner is the child (or the children) requiring the *most* steps to arrive there.”

Here is an example:

3	1	7	14
2	6	7	11
4	1	4	9
3	3	5	5
0	2	0	2
2	2	2	2

We say that the quadruple (3, 1, 7, 14) has *height* 5 since it takes 5 steps down to reach flat terrain. Thus, a step down is one application of the map

$$\delta(a, b, c, d) = (|b - a|, |c - b|, |d - c|, |a - d|).$$

The point of this game is not that the children should be kept busy calculating for quite a while. Triples or five-tuples of numbers would do well for that purpose, but with quadruples it is amazing how fast one usually arrives at a constant tuple.

Look at these three quadruples with their “trails” whose values I have chosen at random in $\{0 \dots 99\}$. They are rather typical (rather than trust me, repeat the experiment for yourself):

25	86	92	23	57	34	90	42	37	18	15	3
61	6	69	2	23	56	48	15	19	3	12	34
55	63	67	59	33	8	33	8	16	9	22	15
8	4	8	4	25	25	25	25	7	13	7	1
4	4	4	4					6	6	6	6

[†] Achim Clausing is a professor of computing science at the Department of Mathematics and Computing Science of the University of Münster (Germany). His e-mail address is achim.clausing@uni-muenster.de.

With the aid of a computer, it is easy to simulate a really large class playing this game. A statistic of the heights of 10,000 random quadruples with values in $\{0 \dots 99\}$ is shown in Figure 1.

Almost 90% of the quadruples reach a constant tuple within three to five steps!

The average height of a random quadruple in this experiment is ≈ 3.88 . If the random numbers were chosen from a different range, e. g. from $\{0 \dots 10^9\}$, we would get essentially the same picture (including the striking local maximum in the frequency of tuples of height 3).

But, in case you didn't notice it from the graph: there were indeed 3 quadruples of height 10 and 2 of height 11 (thus two of the 10,000 children in this class would be the winners).

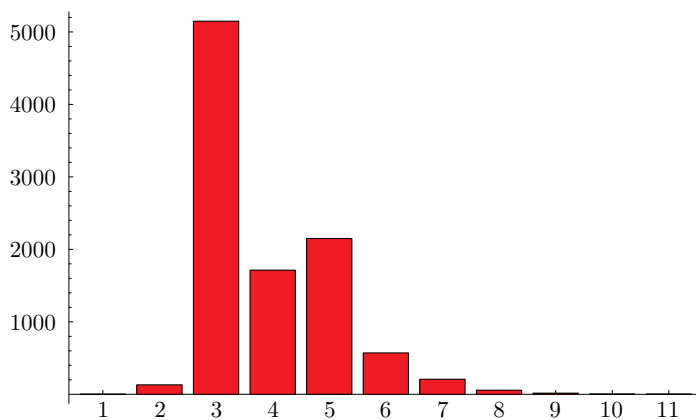


Figure 1: The heights of 10,000 random quadruples with values in $\{0 \dots 99\}$.

On Returning Safely

High mountains, that is, quadruples with a large height, are not easy to conquer. But before we venture to attack them, we should, as a safety measure, make sure that there is a way home from the larger heights. Simply walking downhill might not always lead into the plains, as every mountain walker knows. It is all too easy to get stuck somewhere.

That is, we should ask ourselves whether there is, for every quadruple (a, b, c, d) , an integer n and a value e such that $\delta^n(a, b, c, d) = (e, e, e, e)$ (and then, $\delta^{n+1}(a, b, c, d) = (0, 0, 0, 0)$).

The question is by no means trivial. Even for very simple maps f of some set X into itself, it has been observed that repeated application of f always leads to the same element, but nobody is able to *prove* this. One of the more prominent examples for this phenomenon is the *Syracuse* function f on the positive integers (named after the Syracuse University, where L. Collatz found it in the thirties). Let

$$f(n) = \begin{cases} 3n + 1 & \text{if } n \text{ is odd,} \\ \frac{n}{2} & \text{if } n \text{ is even.} \end{cases}$$

Regardless of the number n you start with, repeated application of f always leads down to 1. Although this has been verified experimentally for values of n up to truly astronomical size, no one has yet been able to show that it is true for all n .

Fortunately, it is easy to see that the stepping game always ends at a constant tuple. If we start with an arbitrary quadruple (a, b, c, d) of non-negative integers, then each step decreases the maximum component in the quadruple, except when the right neighbour of the maximum is zero. Let us assume that the maximum is a and that $b = 0$.

Then $\delta(a, 0, c, d) = (a, c, |d - c|, a - d)$. This tuple also has a as its maximum, hence either $c > 0$ and the next step again reduces the maximum, or $c = 0$ and the next step is $\delta(a, 0, d, a - d) = (a, d, a - 2d, d)$. One more time: Either $d > 0$ and the maximum gets smaller, or $d = 0$ and the next step is $\delta(a, 0, a, 0) = (a, a, a, a)$.

Thus, after four steps, either the maximum of the sequence has got smaller or we have reached a constant tuple. This proves that in the long run, we always get safely down to a constant tuple.

As an alternative, it is easy to show that with four steps we either will reach a constant tuple or all components of the tuple are multiples of 2. This again implies that the four number game eventually will reach a constant quadruple.

Climbing

Climbing means doing one step upwards after the other. Sounds simple. But occasionally, one comes to a point where it is not possible to continue upwards. This can happen here, too.

If $\delta(x, y, z, u) = (a, b, c, d)$, we call (x, y, z, u) a *predecessor* of (a, b, c, d) . It is one step higher than (a, b, c, d) .

The definition of the map δ tells us that there are signs s_i ($i = 0..3$) such that $y - x = s_0a$, $z - y = s_1b$, $u - z = s_2c$ and $x - u = s_3d$. Therefore,

$$(x, y, z, u) = (x, x + s_0a, x + s_0a + s_1b, x + s_0a + s_1b + s_2c),$$

and from $x = u + s_3d$ we see that

$$s_0a + s_1b + s_2c + s_3d = 0$$

holds. Furthermore, if we can find signs s_i such that this equation holds, then clearly (a, b, c, d) has a predecessor.

For example, the tuple $(3, 1, 7, 14)$ considered above obviously has no such signs since $3 + 1 + 7 < 14$. From this quadruple, it is not possible to do an upward step. On the other hand, if $a, b, c, d > 0$, and

$$a + b + c = d, \quad (*)$$

we can chose $s_0 = s_1 = s_2 = 1$ and $s_3 = -1$, and find the predecessor

$$(x, y, z, u) = (x, x + a, x + a + b, x + a + b + c),$$

where $x \geq 0$ can be chosen arbitrarily. How about a predecessor of this tuple? We can choose $x \geq 0$ such that

$$x + y + z = u$$

holds. This means

$$x + (x + a) + (x + a + b) = x + a + b + c,$$

or

$$x = (c - a)/2.$$

Thus, if for quadruples (a, b, c, d) with $a + b + c = d$ and $a \leq c$ we define the upwards-map γ by

$$\begin{aligned} \gamma(a, b, c, d) &= \left(\frac{c-a}{2}, \frac{c-a}{2} + a, \frac{c-a}{2} + a + b, \frac{c-a}{2} + a + b + c \right) \\ &= \frac{1}{2}(-a + c, a + c, a + 2b + c, a + 2b + 3c) \\ &= (x, y, z, u), \end{aligned}$$

then $\delta(x, y, z, u) = (a, b, c, d)$, $x \leq z$ and $x + y + z = u$. We can choose the signs s_i as before and conclude that (x, y, z, u) again has a predecessor.

Now we can climb one step after the other! Stated in somewhat less sportive terms, the tuples

$$(a_n, b_n, c_n, d_n) = \gamma^n(a, b, c, d) \quad (n \geq 0)$$

have the heights $h_n = n + h$, where h is the height of (a, b, c, d) .

Don't forget to start from a tuple where $a + b + c = d$ and $a \leq c$. Let us, for example, take a mountain walk of 15 steps starting from the tuple $(0, 0, 1, 1)$ with the height $h = 2$:

0	0	1	1	$\frac{3}{4}$	$\frac{11}{8}$	$\frac{5}{2}$	$\frac{37}{8}$
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{7}{8}$	$\frac{13}{8}$	3	$\frac{11}{2}$
0	$\frac{1}{2}$	1	$\frac{3}{2}$	$\frac{17}{16}$	$\frac{31}{16}$	$\frac{57}{16}$	$\frac{105}{16}$
$\frac{1}{2}$	$\frac{1}{2}$	1	2	$\frac{5}{4}$	$\frac{37}{16}$	$\frac{17}{4}$	$\frac{125}{16}$
$\frac{1}{4}$	$\frac{3}{4}$	$\frac{5}{4}$	$\frac{9}{4}$	$\frac{3}{2}$	$\frac{11}{4}$	$\frac{81}{16}$	$\frac{149}{16}$
$\frac{1}{2}$	$\frac{3}{4}$	$\frac{3}{2}$	$\frac{11}{4}$	$\frac{57}{32}$	$\frac{105}{32}$	$\frac{193}{32}$	$\frac{355}{32}$
$\frac{1}{2}$	1	$\frac{7}{4}$	$\frac{13}{4}$	$\frac{17}{8}$	$\frac{125}{32}$	$\frac{115}{16}$	$\frac{423}{32}$
$\frac{5}{8}$	$\frac{9}{8}$	$\frac{17}{8}$	$\frac{31}{8}$	$\frac{81}{32}$	$\frac{149}{32}$	$\frac{137}{16}$	$\frac{63}{4}$

There is no need to worry about the fact that these quadruples are not integers. Just multiply them by the GCD of their denominators; this clearly does not alter their heights.

If, say, we multiply the last line $(\frac{81}{32}, \frac{149}{32}, \frac{137}{16}, \frac{63}{4})$ of our mountain hike by 32, we obtain the integer quadruple $(81, 149, 274, 504)$, which is 17 steps high.

Tuples of this height are rare; you don't find them by a quick random search. But once you have one, you can easily produce others of the same height from it by multiplying it with a nonzero constant, by adding the same constant to every element of the tuple, by rotating it or by reversing its order. None of these operations affects the height of the tuple.

We call tuples *equivalent*, if they can be converted into each other by one of these operations. In particular, for every tuple $(a, b, c, d) \neq (0, 0, 0, 0)$ there is an equivalent *normalized* one, denoted by $[a, b, c, d]$, whose smallest and largest element are 0 and 1. It is obtained by subtracting the minimal component from each of the components and dividing the result by its maximum component.

In our example we find $[81, 149, 274, 504] = (0, \frac{68}{423}, \frac{193}{423}, 1)$.

The Sky is the Limit

Now that we have seen that by repeated down stepping one reaches the plains from every quadruple of *rational* numbers, it is only natural to ask where we get from a non-rational starting quadruple. One might conjecture that the "trails" of such tuples would in general converge to a constant without actually reaching it in finitely many steps.

As an experiment, we try the tuple $(1, e, \sqrt{7}, \pi)$. The result is a surprise:

$$\begin{array}{cccc} 1 & e & \sqrt{7} & \pi \\ e-1 & e-\sqrt{7} & \pi-\sqrt{7} & \pi-1 \\ \sqrt{7}-1 & \pi-e & \sqrt{7}-1 & \pi-e \\ \sqrt{7}+e-\pi-1 & \sqrt{7}+e-\pi-1 & \sqrt{7}+e-\pi-1 & \sqrt{7}+e-\pi-1 \end{array}$$

It took us only three steps to get down! In fact, we should not be surprised that much. The rational approximation $(1, 2.7, 2.6, 3.1)$ to our starting tuple has height 3, as have

better approximations, and there is simply no reason to suspect that in the limit anything should change.

Or is there? What, if we had a *converging* sequence of rational quadruples of ever-increasing height? Then the limiting quadruple is perhaps one of the more interesting viewpoints on our mountain hike.

The question is not an academic one since in fact we already have such a sequence of quadruples! Look at the normalized version $[a_n, b_n, c_n, d_n]$ ($n \geq 0$) of the sequence that we constructed in the preceding section. The normalized quadruples of our 15-step mountain walk are, numerically,

0	0	1	1	0	0.16129	0.451613	1
0	0	0	1	0	0.162162	0.459459	1
0	0.333333	0.666667	1	0	0.159091	0.454545	1
0	0	0.333333	1	0	0.161905	0.457143	1
0	0.25	0.5	1	0	0.16	0.456	1
0	0.111111	0.444444	1	0	0.161074	0.456376	1
0	0.181818	0.454545	1	0	0.160563	0.456338	1
0	0.153846	0.461538	1	0	0.160757	0.456265	1

It is not hard to believe that this sequence converges to $(0, 0.160\dots, 0.456\dots, 1)$.

Say, the limit is (a, b, c, d) . What properties could it have? Since $(a_{n+1}, b_{n+1}, c_{n+1}, d_{n+1}) = \gamma(a_n, b_n, c_n, d_n)$, it is not unreasonable to hope that

$$[\gamma(a, b, c, d)] = [a, b, c, d] = [\delta(a, b, c, d)]$$

holds. In accordance with the elements of the sequence, the limit should satisfy the relation (\star) , and since we are free to multiply with a constant, we put $(a, b, c, d) = (1, x, y, 1+x+y)$ with unknown x, y about which we assume $1 \leq x \leq y$. (We can assume whatever we like, this is just a heuristic guess about what the limit is.)

Then $\delta(1, x, y, 1+x+y) = (x-1, y-x, x+1, x+y)$. An equivalent tuple is $(1, \frac{y-x}{x-1}, \frac{x+1}{x-1}, \frac{1+x+y}{x-1})$. Hence we try to solve

$$(1, x, y, 1+x+y) = (1, \frac{y-x}{x-1}, \frac{x+1}{x-1}, \frac{1+x+y}{x-1})$$

and find by comparing the second components

$$x = \frac{y-x}{x-1}$$

that $y = x^2$, and from

$$y = x^2 = \frac{x+1}{x-1}$$

that x must be a (real) solution of $1+x+x^2 = x^3$. This equation has exactly one real solution, which is

$$x = \frac{1}{3} + \frac{(19-3\sqrt{33})^{\frac{1}{3}}}{3} + \frac{(19+3\sqrt{33})^{\frac{1}{3}}}{3} = 1.83929\dots$$

The quadruple $(1, x, x^2, 1+x+x^2) = (1, x, x^2, x^3)$ indeed satisfies

$$\begin{aligned} [1, x, x^2, x^3] &= (0, \frac{x-1}{x^3-1}, \frac{x^2-1}{x^3-1}, 1) \\ &= (0, \frac{1}{x^3}, \frac{1}{x^3} + \frac{1}{x^2}, 1) \\ &= (0, 0.160713\dots, 0.456311\dots, 1). \end{aligned}$$

It is, by construction, an eigenvector of the linear map $(a, b, c, d) \mapsto (b-a, c-b, d-c, d-a)$:

$$\begin{aligned} \delta(1, x, x^2, x^3) &= (x-1, x^2-x, x^3-x^2, x^3-1) \\ &= (x-1) \cdot (1, x, x^2, \frac{x^3-1}{x-1}) \\ &= (x-1) \cdot (1, x, x^2, 1+x+x^2) \\ &= (x-1) \cdot (1, x, x^2, x^3) \\ &= 0.83929\dots \cdot (1, x, x^2, x^3). \end{aligned}$$

By down stepping from $(1, x, x^2, x^3)$, we obtain an equivalent tuple, albeit with slightly smaller components. We have reached a place from where there is no way back to flat earth: There is not just π *in the Sky*, but some unexpected quadruples, too!

A quadruple from which one cannot reach a constant tuple by repeated application of δ does not have a finite height. It is natural to say that it has height ∞ . (We probably should at this point define the *sky* as the set of all quadruples of height ∞ . But by this we would throw π out of the paradise, an unforgivable sin in this magazine.)

Tribonacci in the Sky

The sky is where the gods are, and they live on a high mountain. The old Greeks knew that (and some mathematics as well). Now the mathematical sky is not just inhabited by Greek gods (say, Pythagoras, Euclid, or the immortal number π) but also by many lesser celebrities. One of them is a certain Mr. Tribonacci*, a cousin of the more famous mathematician Leonardo of Pisa, known as Fibonacci.

We all know and love the Fibonacci numbers $0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, \dots$, where each element is the sum of the two preceding ones: $f_n = f_{n-1} + f_{n-2}$. The ratios f_n/f_{n-1} of two consecutive Fibonacci numbers converge to the Golden Ratio $\frac{1}{2}(1 + \sqrt{5}) = 1.61803\dots$, which is the positive root of the equation $1+x = x^2$.

In the *Tribonacci* number sequence $0, 1, 1, 2, 4, 7, 13, 24, 44, 81, 149, 274, 504, \dots$, each element is the sum of the preceding *three* elements: $t_n = t_{n-1} + t_{n-2} + t_{n-3}$. The ratios t_n/t_{n-1} of two consecutive Tribonacci numbers converge to the positive root of the equation $1+x+x^2 = x^3$, which we happen to know as $x = 1.83929\dots$

This sequence, though less common than the Fibonacci sequence, appears in many contexts (it has number A75 in Sloane's On-Line Encyclopedia of Integer Sequences, the Fibonacci sequence being A45). It is closely connected with our stepping game.

Did you notice that the four numbers $81, 149, 274, 504$ in the Tribonacci sequence form the quadruple of height 17 that we constructed in our 15-step-mountain walk? And that the Tribonacci recursion is just the condition (\star) for a quadruple to have a predecessor?

The Tribonacci sequence is another "stairway to heaven," in the technical sense of this article: The heights $h_n = \text{height}(t_n, t_{n+1}, t_{n+2}, t_{n+3})$ of the tuples in this sequence are $5, 5, 8, 8, 11, 11, 14, 14, 17, 17, \dots$, that is, they increase by 3 every two steps. This is easy to prove. A similar fact holds for Tribonacci-like sequences starting with triples other than $\{0, 1, 1\}$.

The Tribonacci numbers also show up in the answer to the following question: What is the smallest integer a such that $\{0 \dots a\}^4$ contains a quadruple of height n ?

Let us denote this integer by a_n . The first elements of the sequence (a_1, a_2, \dots) are

$$(1, 1, 1, 3, 3, 4, 9, 11, 13, 31, 37, 44, 105, 125, 149, 355, \dots).$$

*The "Mr." stands for the Italian title "Messer," poor Mr. Tribonacci not having any first name.

As an example, $a_7 = 9$ since $(0, 1, 4, 9)$ has height 7 and there is no integer tuple of height 7 whose elements are all less than 9. Can you design a tuple of height 12 whose largest element is 44?

This sequence is A045794 in the On-Line Encyclopedia of Integer Sequences, a less prominent rank. But the encyclopedia knows that it can be expressed using the Tribonacci numbers as

$$(t_0 + t_2, t_1 + t_2, t_3, t_2 + t_4, t_3 + t_4, t_5, t_4 + t_6, \dots),$$

and it also tells the relation of this sequence to the four number game.

Viewing the Scenery

The question remains why the average height of a random quadruple should only be about 4. Why are the higher tuples so scarce?

To see the reason, we divide the set of all quadruples of non-negative numbers into two disjoint parts, one part being those tuples in which the minimum and maximum elements are neighbours (in the cyclic sense), the other part consisting of the tuples where the maximal and minimal elements are not neighbours.

In this latter part, all quadruples have a height that is at most 5. To see this, we first observe that a tuple with two equal components that are not neighbours, say (a, b, c, b) , has a height $h \leq 3$: $\delta(a, b, c, b) = (u, v, v, u)$ with $u = |b - a|$, $v = |c - b|$, $\delta(u, v, v, u) = (w, 0, w, 0)$ with $w = |v - u|$, and $\delta(w, 0, w, 0) = (w, w, w, w)$.

Now, for a tuple whose maximal and minimal elements are not neighbouring, there is always an equivalent one of the form $(0, x, 1, y)$ with $0 \leq x \leq y \leq 1$ (do you see why?). Since $\delta(0, x, 1, y) = (x, 1 - x, 1 - y, y)$, and $\delta(x, 1 - x, 1 - y, y) = (a, b, c, b)$ with $a = |1 - 2x|$, $b = |y - x|$ and $c = |1 - 2y|$, we see that $(0, x, 1, y)$ has height at most 5.

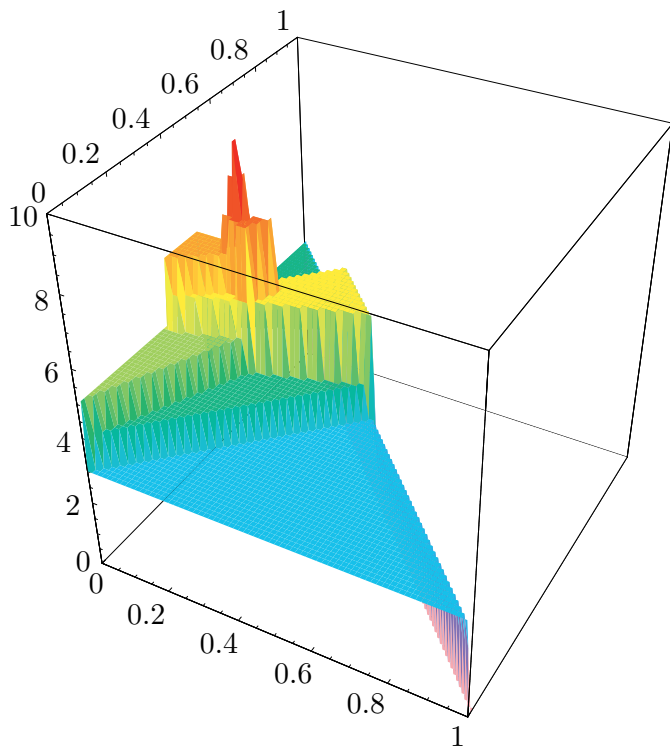


Figure 2: An approximation of the graph of the function $f(x, y) = \text{height}(0, x, y, 1)$ in the domain $\{(x, y) : 0 \leq x, y, x + y \leq 1\}$.

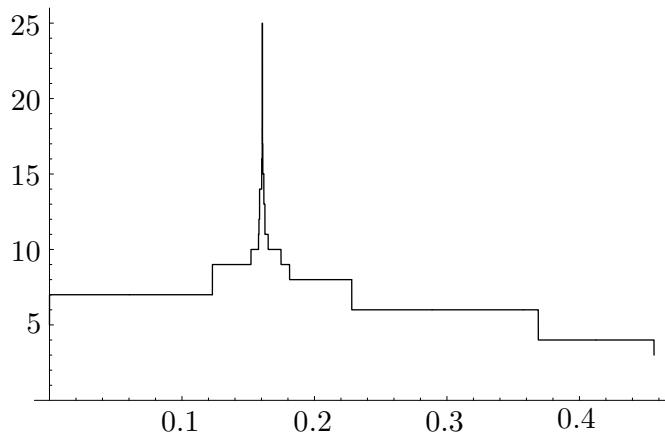


Figure 3: The graph of $g(x) = f(x, 0.456)$.

This simple argument shows that indeed roughly half of all quadruples have a “low” height. What about the others, that is, those with cyclically adjacent minimum and maximum components?

By normalizing and possibly reversing the order of such a tuple, we can always obtain an equivalent tuple of the form $(0, x, y, 1)$ with $x, y \in [0, 1]$. We may also assume that $x + y \leq 1$ since otherwise the tuple $(0, x, y, 1)$ can be replaced by the equivalent one $(0, 1 - y, 1 - x, 1)$.

Figure 2 shows (a rather crude approximation of) the graph of the function $f(x, y) = \text{height}(0, x, y, 1)$ in the domain $\{(x, y) : 0 \leq x, y, x + y \leq 1\}$.

It is apparent that in the part where $x > y$, the value is everywhere 3. (This is again easy to verify.) In that sense, at least three quarters of all quadruples have height $h \leq 5$.

Only in the remaining part exist points (x, y) —that is, quadruples $(0, x, y, 1)$ —with larger heights.

The red peak in the preceding figure is centered around the normalized point $(0, 0.160\dots, 0.456\dots, 1)$. It is the mountain on whose summit the quadruple of height ∞ resides. Only here[†] we can find quadruples of large height. The probability that (the normalized version of) a randomly chosen quadruple lies close to this point is apparently small. This explains why it is difficult to find high quadruples by a random search.

Just how difficult, is illustrated by a section through the graph of f at the critical point. Figure 3 shows the graph of $g(x) = f(x, 0.456)$.

As you can see, the peak around $x = 0.160\dots$ (where $g(x) = \infty$) is really sharp, a true challenge for every mountaineer!

References

- [1] Ciamberlini, C. and A. Marengoni, 1937: Su una interessante curiosità numerica. *Periodiche di Matematiche*, **17**, 25–30.
- [2] Berlekamp, E. R., 1975: The design of slowly shrinking labelled squares. *Math. Comp.*, **29**, 25–27.
- [3] Greenwell, R., 1989: The game of Diffy. *Math. Gazette*, **73**, 222–225.
- [4] McLean, K. R., 1999: Playing Diffy with real sequences. *Math. Gazette*, **83**, 58–68.
- [5] Sloane, N. J. A., 2004: The On-Line Encyclopedia of Integer Sequences. <http://www.research.att.com/~njas/sequences/index.html>.

[†]And, of course, near all equivalent points, in particular, near its mirror point on the other side of the line $x + y = 1$.



“Unstable” Rearrangements of n Points

Chris Garrett[†]

Mathematics students learn early that there are $n!$ ways of arranging n different points; the first point can be chosen in n different ways, the second in just $n - 1$ ways, and so on. If this doesn't seem familiar, have a look at the nice article by Byron Schmuland in the last issue of *π in the Sky* [3]. An interesting variation on $n!$ cropped up recently in a study of ocean mixing conducted by Kate Stansfield, Richard Dewey and me [4].

We were analyzing profiles of water density obtained in Juan de Fuca Strait using an instrument called a CTD. This records electrical conductivity (C), temperature (T) and pressure, hence depth (D). The conductivity depends on the salinity (S) as well as the temperature, so that C and T give S. In turn, T and S give the density of the water.

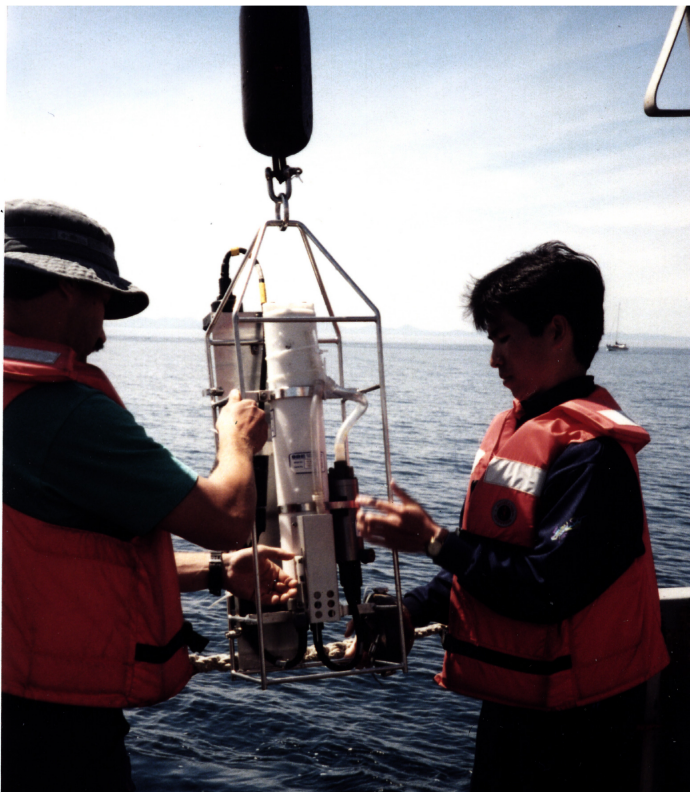


Figure 1: A CTD about to be lowered into the sea. This model records internally and is suitable for use from a small boat (in this case the University of Victoria’s 16m John Strickland). Other CTDs send signals up a conducting cable for logging on board ship.

[†] Chris Garrett is Lansdowne Professor of Ocean Physics at the University of Victoria. His e-mail address is garrett@phys.uvic.ca.

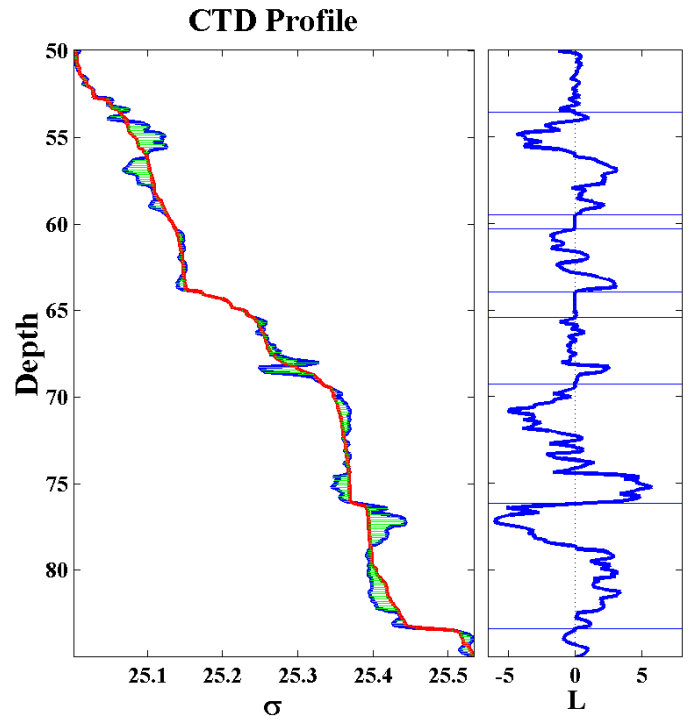


Figure 2: The blue line in the left-hand panel is the vertical profile of a quantity σ , indicating density, plotted against the depth in metres. The red line is the profile after sorting it so that σ increases with depth. The right-hand panel shows the vertical displacement L , in metres, of each data point during this sorting, and shows several distinct overturning regions. The Thorpe scale L_T is the root mean square value of L .

If the density increases with depth, the water is hydrostatically stable. If, on the other hand, there are sections of the profile with dense water on top of light water (Figure 2), these are hydrostatically unstable and cannot persist. They must have been produced by some sort of overturning motion (perhaps an internal equivalent of a breaking wave at the sea surface) and will subsequently collapse. These overturns are an indication that mixing is occurring, with possibly profound influences on the physics and biology of the sea.

With some handwaving about the physics, it seems that the vertical mixing rate (which is what we were interested in) can be related to the distances that water parcels have to be moved vertically to produce a reordered stable profile in which the density increases steadily downwards (Figure 2). The root mean square (rms) distance that the water parcels are moved in this reordering is denoted L_T and known as the *Thorpe scale* after Steve Thorpe, who first used this approach while studying mixing in the fresh water of Loch Ness (where the water density depends only on temperature). The turbulent mixing rate actually has the units of diffusivity (say $\text{m}^2 \text{s}^{-1}$), with the so-called *eddy diffusivity* found empirically to be given approximately by $0.1NL_T^2$. Here N is a measure of the strength of the stratification of the reordered density profile, with the units of frequency (s^{-1}). It is the frequency with which a water parcel would oscillate if displaced vertically in the reordered profile. The collapse time of the unstable regions observed before reordering is proportional to N^{-1} .

So much for the physics background. We wanted to go beyond just looking at the rms displacements of water parcels and look at the probability distribution of individual displacements, i.e. did the rms value come from lots of small displacements or just a few big ones? We determined this from our data, but wanted some theoretical ideas for comparison. An

obvious one was to see what would happen if all redistributions were equally likely. For a start on this, we needed to know how many reorderings we had to consider. This is less than the simple answer of $n!$ as some of the profiles were never unstable to start with, at least not in the sense that one has to reorder the whole set. For example, with just two points, labelled 1 and 2, with density increasing with the numerical value, then profile $\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ is unstable, but $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ is not. Thus we have only 1 case to consider instead of $2! = 2$. With 3 points, $\begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}$, $\begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$ qualify as “complete” overturns in which the whole array is involved in a reordering to the stable profile $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$, even though 2 stays put in the

second case. On the other hand, the profile $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ itself obviously does not need any reordering, and the profiles $\begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$ and $\begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}$ can be reordered without involving the first and last points respectively. Thus we have 3 cases to consider instead of $3! = 6$.

With 4 points there are 24 cases to evaluate in the same way. To give a couple of examples, $\begin{pmatrix} 4 \\ 2 \\ 1 \\ 3 \end{pmatrix}$ qualifies as a com-

plete overturn, even though 2 is not moved, but $\begin{pmatrix} 2 \\ 1 \\ 4 \\ 3 \end{pmatrix}$ does

not qualify, as reordering involves just the reversal of the top and bottom pairs; it is thus a superposition of two second-order overturns rather than a fourth-order one. The reader may check that for 4 points we end up with 13 cases instead of 24 and for 5 we have 71 instead of 120, but this is getting tedious, and a simple computer program needs to be written to go much further. Running even this becomes time-consuming as n gets large; already for $n = 8$ we have to evaluate $8! = 40\,320$ cases.

Thus our little problem was to determine the number of cases, denoted by $F(n)$, say, as a function of n . Maybe the mathematical readers of this newsletter can immediately write down the answer, and I could pretend that we did. The truth is that we cheated by generating the series up to $n = 8$ on a computer and then checking the wonderful compilation by Sloane and Pouffe [2]. There turned out to be more than one candidate solution all the way up to $n = 7$, but uniqueness seemed to emerge by $n = 8$. The answer is the recurrence relation

$$F(n) = n! - \sum_{k=1}^{n-1} k!F(n-k).$$

This actually arose in a completely different problem [1], but led us to slap our foreheads when we saw it. We had spent a little time trying for a closed form for $F(n)$, whereas the recurrence formula is really rather obvious: presented with

the $n!$ rearrangements of n points we have to exclude the ones that start with the unaltered point 1 followed by the $F(n-1)$ complete overturns of the remaining $n-1$ points, also the $2!$ arrangements of the first two points times the $F(n-2)$ complete overturns of the remaining $n-2$ points, and so on. I’d add an exclamation mark at this point were it not for the risk of the reader confusing it with the factorial sign! (Oops.) The only detail remaining was that we had to choose $F(1) = 1$ to get this recurrence relation to start properly and give $F(2) = 1$, even though we could not really define a reordering of a single point. The difference between $F(n)$ and $n!$ is actually a small fraction of $n!$ as n becomes large, and maybe a clever reader can work out a good approximate formula (known as an “asymptotic” approximation) for $F(n)$ for large n .

Anyway, after establishing how many cases we needed to consider, we could return to the first problem we had set for ourselves. This was to find the probability distribution (which is just the likelihood of occurrence) of displacements of a given magnitude within any overturn of, say, n points, assuming that each permitted rearrangement is equally likely. First considering all $n!$ possible rearrangements (even though we have ruled out some of them), we see that each point can move any number of positions from 0 to whatever takes it to get to the farthest end of the overturn. A zero displacement only occurs for 1 of these n possibilities, regardless of where the point starts, and so has a probability of $1/n$. A move of 1 position can be in either direction, unless the point is already at the top or bottom of the overturn, and so on. I leave the reader to establish that a move of m has a probability of $2(n-m)/n^2$ for m from 1 to n . Overall, then the probability is as shown in Figure 3, with small displacements more likely than large ones.

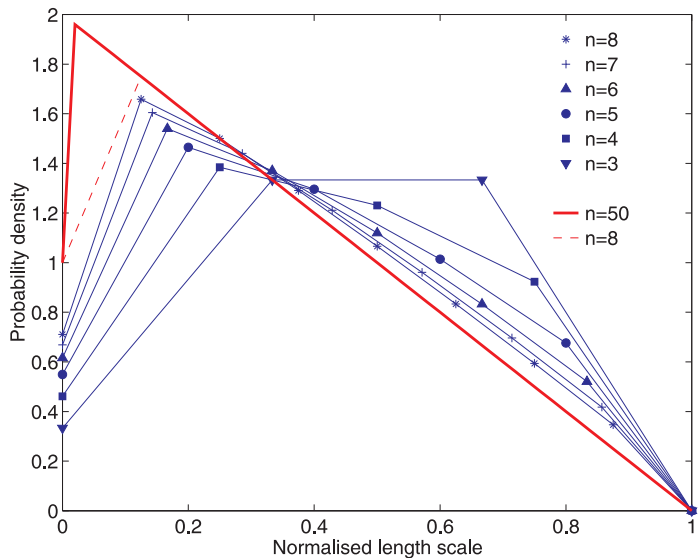


Figure 3: The lines joining the points give the probability of a particle displacement as a function of the fractional displacement (the normalised length scale) m/n . The solid red line is for $n = 50$, considering all $n!$ cases. For other values of n the solid red line still applies for a fractional displacement of $1/n$ and larger, but connects to 1 for $m = 0$, as shown by the dashed red line for $n = 8$. The blue lines connect the values for $n = 3$ to 8, now only allowing for complete overturns. Note that the probability has also been multiplied by n , so that the probability values times $1/n$ add up to 1.

We now, as before, exclude the cases that do not constitute complete overturns. This changes the picture somewhat, increasing the likelihood of larger displacements (Figure 3), though the difference decreases as n increases. Details of the mathematics are described in our paper [4].

Actual data (Figure 4) showed a greater probability of small displacements than given by our model, for reasons that may be associated with molecular diffusion, blurring extreme values. There is more to be done, but this is straying back into physical oceanography.

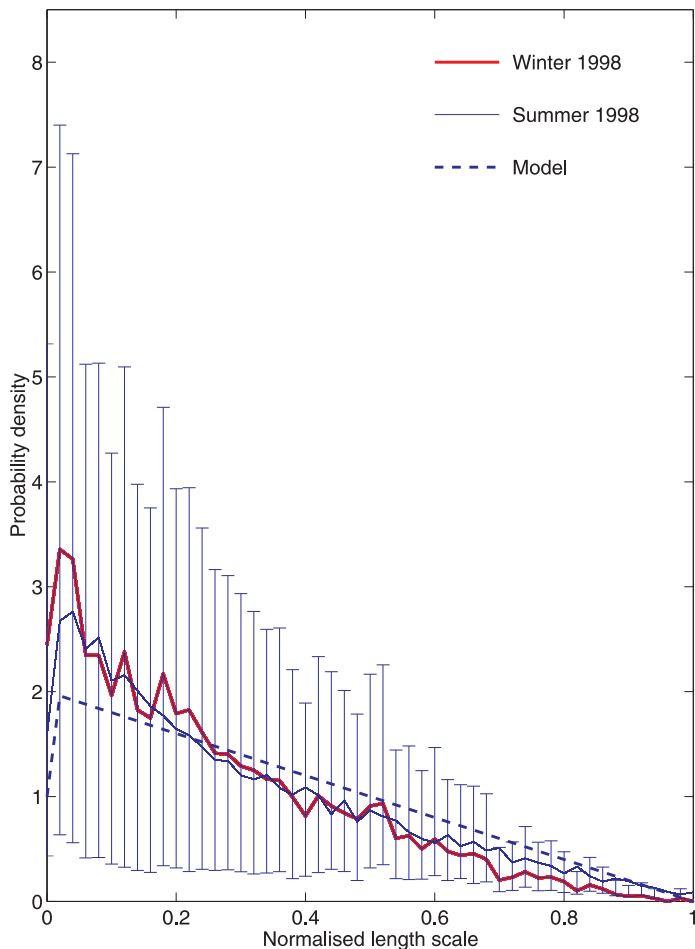
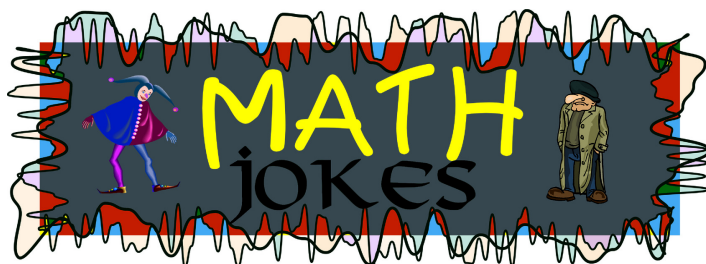


Figure 4: The solid red and blue lines show the probability distribution of actual displacements from a number of CTD profiles taken in winter (red) and summer (blue). The error bars on each value (vertical lines) for each value are large, but the trends in the results are clear. The dashed blue line shows, for comparison, the expectations from our model for $n = 50$ (a typical value). This is for the unrestricted case (with all $n!$ possible rearrangements considered), but there is little difference from the restricted case for large values of n .

So what are the messages of this article? One is that even those of us who have drifted from physics and mathematics into their application in the real world still love problems like the one described here when they crop up in our work. Another is that Monsieur Comtet surely didn't anticipate oceanographic applications of his series, where as Sloane and Plouffe might be less surprised at the range of users of their encyclopaedia. My main message to any student reader, though, is that if you enjoy mathematics and physics, and want to apply them to a research field that is fun, consider ocean physics!

References

- [1] Comtet, L., 1972: Sur les coefficients de l'inverse de la série formelle $\sum n!t^n$. *C. R. Acad. Sci. Paris, Série A*, **275**, 569–572.
- [2] Sloane, N. J. A. and S. Plouffe, 1995: *The Encyclopedia of Integer Sequences*. Academic Press.
- [3] Schmuland, B., 2003: Shouting factorials! *Pi in the Sky*, September issue.
- [4] Stansfield, K., C. Garrett, and R. Dewey, 2001: The probability distribution of the Thorpe displacement within overturns in Juan de Fuca Strait. *J. Phys. Oceanogr.*, **31**, 3421–3434.



Q: Why do mathematicians, after a dinner at a Chinese restaurant, always insist on taking the leftovers home?

A: Because they know the Chinese remainder theorem!

“That math prof.’s marriage, I heard, is falling apart.”

“That doesn't surprise me: he's into scientific computing, and she's incalculable....”

Q: What is polite and works for the phone company?

A: A deferential operator....

Trigonometry for farmers: swine and coswine....



“BOYS, THE TEAM STATISTICIAN WILL NOW EXPLAIN WHY OUR RECORD OF 23 WINS AND 59 LOSSES IS VERY, VERY GOOD.”

©Copyright 2004
Sidney Harris



Math Challenges

Problem 1. Two hospitals H_1 and H_2 treat exactly the same number of patients during a year. Each patient suffers from exactly one of two diseases, D_1 or D_2 . Hospital H_1 cures a greater percentage of patients than H_2 . However, H_2 cures a greater percentage of D_1 patients and also a greater percentage of D_2 patients. Is this possible?

Problem 2. The numbers $1, \dots, n$ are written in a row. It is permitted to transpose any two neighbouring numbers. If 2005 operations are performed, is it possible that the final arrangement of numbers coincides with the original arrangement?

Problem 3. Prove that at least two of 100 distinct positive integers ≤ 294 have a difference of 2, 3, or 5.

Problem 4. Find the real solutions of the system $(x + y)^5 = z$, $(y + z)^5 = x$, $(z + x)^5 = y$.

Problem 5. Let A_1, \dots, A_n be points on a circle of radius one. Prove that there is a point P on the circle such that $\sum_k^n PA_k > n$.

Send your solutions to π in the Sky: [Math Challenges](#).

Solutions to the Problems Published in the September, 2003 Issue of π in the Sky:

Problem 1. Let $g : (0, \infty) \rightarrow (0, \infty)$, $g(x) = x + \sqrt{x}$. Note that g is increasing and bijective, and

$$f(x + \sqrt{x}) \leq x \iff f(g(x)) \leq x.$$

On letting $y = g(x) \Rightarrow x = g^{-1}(y)$, we may write this condition as

$$f(y) \leq g^{-1}(y). \quad (1)$$

On the other hand,

$$x \leq f(x) + \sqrt{f(x)} \iff x \leq g(f(x)).$$

Since g^{-1} also is increasing, we have

$$g^{-1}(x) \leq g^{-1}(g(f(x)));$$

that is,

$$g^{-1}(x) \leq f(x). \quad (2)$$

From (1) and (2) we get

$$f(x) = g^{-1}(x) = \frac{1}{4}(-1 + \sqrt{1 + 4x})^2.$$

Problem 2. (Solution given by Robert Bilinski, Montréal) The given equation

$$x^2 - xy + y^2 = x + y$$

is equivalent to

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

Since

$$\Delta_x = (-y - 1)^2 - 4(1)(y^2 - y) = -3y^2 + 6y + 1$$

needs to be positive or zero for x to be integral, we need to solve

$$-3y^2 + 6y + 1 \geq 0.$$

Since

$$\Delta_y = 6^2 - 4(-3)(1) = 48 > 0$$

and $a = -3$, $\Delta_x \geq 0$ in between the roots. In other words, $\Delta_x \geq 0$ if

$$y \in \left[\frac{3 - 2\sqrt{3}}{3}, \frac{3 + 2\sqrt{3}}{3} \right] = [-0.1547\dots, 2.1547\dots].$$

But we want integer solutions, so that $y \in \{0, 1, 2\}$. Substituting in the original equation and solving for x , we find that

$$(x, y) \in \{(0, 0), (0, 1), (1, 0), (1, 2), (2, 1), (2, 2)\}.$$

Problem 3. There are subsets of $\{1, 2, \dots, 2003\}$ with 668 elements satisfying the required property. Indeed,

$$A = \{3k + 1 | k = 0, 1, \dots, 667\}$$

has 668 elements and satisfies the required property: for every $x, y \in A$,

$$x - y \equiv 0 \pmod{3}$$

and

$$x + y \equiv 2 \pmod{3};$$

thus $x - y$ does not divide $x + y$. If there is a subset

$$B \subset \{1, 2, \dots, 2003\}$$

with more than 668 elements, then there are two of them with a difference of less than 3. Hence B contains two elements a and b such that $a - b = 1$ or $a - b = 2$. If $a - b = 1$ then $a - b$ divides $a + b$. If $a - b = 2$ then a and b are of the same parity and $a - b$ divides $a + b$. Therefore the required set is A .

Problem 4. Let us solve a more general problem: If x_1, \dots, x_n are positive real numbers such that

$$\sum_{i=1}^n \frac{1}{n-1+x_i} > 1,$$

then

$$x_1 x_2 \dots x_n < 1.$$

Solution: Let

$$x_i = \frac{(n-1)y_i}{1-y_i}.$$

Then

$$\begin{aligned} \sum_{i=1}^n \frac{1}{n-1+x_i} > 1 &\iff \sum_{i=1}^n \frac{1-y_i}{1-y_i+y_i} > n-1 \\ &\iff n - \sum_{i=1}^n y_i > n-1 \iff \sum_{i=1}^n y_i < 1. \end{aligned}$$

Now

$$\begin{aligned} x_1 x_2 \dots x_n &= (n-1)^n \frac{y_1 \dots y_n}{(1-y_1) \dots (1-y_n)} \\ &< (n-1)^n \frac{y_1 \dots y_n}{(y_2 + \dots + y_n) \dots (y_1 + \dots + y_{n-1})} \\ &\leq \frac{(n-1)^n}{(n-1)^n} \frac{y_1 \dots y_n}{\sqrt[n-1]{(y_1 \dots y_n)^{n-1}}} = 1. \end{aligned}$$

(In the last inequality we have applied the AGM inequality: $\frac{y_2 + \dots + y_n}{n-1} \geq \sqrt[n-1]{y_2 \dots y_n}$ etc.)

Problem 5. Let A, B, C , and D be the given points. The triangle made with any three of these points is acute. Indeed, if consider $\triangle ABC$ then

$$\cos \hat{A} = \frac{AB^2 + AC^2 - BC^2}{2AB \cdot AC} \geq \frac{2+2-4}{2AB \cdot AC} = 0.$$

Therefore

$$\hat{A} \leq 90^\circ.$$

We conclude that $ABCD$ is a convex quadrilateral and also $B \leq 90^\circ, C \leq 90^\circ$, and $D \leq 90^\circ$. As $\hat{A} + \hat{B} + \hat{C} + \hat{D} = 360^\circ$, the only consistent situation is $A = B = C = D = 90^\circ$. Hence $ABCD$ must be a rectangle. If we assume that $AB = a, BC = b, a \neq b$, and $a, b \in [\sqrt{2}, 2]$, then

$$AC = \sqrt{a^2 + b^2} > \sqrt{2+2} = 2,$$

which is a contradiction. Therefore $a = b$ and $ABCD$ is a square.

Problem 6. Let O be the center of the circles and A, B, C be the vertices of the triangle with the maximum area. The point O must belong to the interior of the triangle. On the other hand, it must be the orthocenter of the triangle ABC . Indeed, if O is not the orthocenter of the triangle then AO is not perpendicular to BC . Take a line through A parallel to BC . This line is not tangent to the circle and hence intersects the circle at a point D (see Figure 1). Now if we choose a point M in the arc AD the area of the $\triangle BMC$ is greater then the area of $\triangle BAC$, a contradiction. In $\triangle AOC$ (see Figure 2)

$$AC^2 = 1 + 16 + 8 \cos B = 17 + 8 \cos B.$$

If R is the radius of the circumcircle of $\triangle ABC$, then

$$AC^2 = 4R^2 \sin^2 B.$$

Hence

$$4R^2 \sin^2 B = 17 + 8 \cos B. \quad (3)$$

In $\triangle BOA'$ and $\triangle BAA'$

$$BA' = OB \cos(90^\circ - C) = \sqrt{7} \sin C$$

and

$$BA' = AB \cos B = 2R \cos B \sin C.$$

Hence

$$2R \cos B = \sqrt{7}. \quad (4)$$

Eliminating R from (3) and (4) we obtain

$$8 \cos^3 B + 24 \cos^2 B - 7 = 0;$$

hence

$$(2 \cos B - 1)(4 \cos^2 B + 14 \cos B + 7) = 0.$$

The only acceptable solution is $\cos B = 1/2$ and hence $B = 60^\circ$. We immediately obtain $\widehat{BAA'} = 30^\circ, OC' = 1/2$ and hence $CC' = 9/2$. Also $\sin \widehat{ABB'} = 1/(2\sqrt{7}), BC' = 3\sqrt{3}/2, AC = \sqrt{21}, AC' = \sqrt{3}/2$ and $AB = 2\sqrt{3}$. Thus $\text{Area}(\triangle ABC) = 9\sqrt{3}/2$ is the required maximum area.

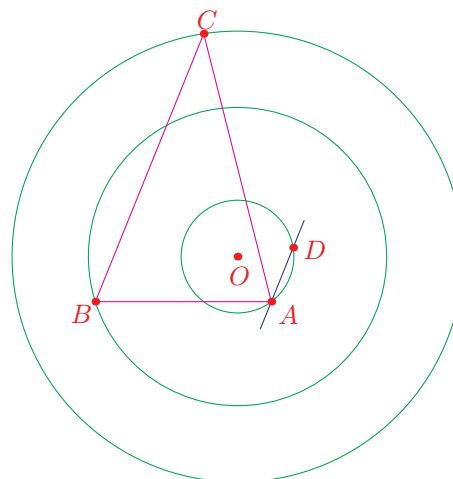


Figure 1

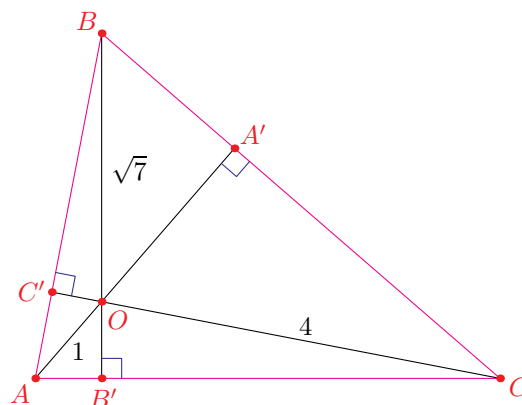


Figure 2



The figures on pages 7, 8, 12, 20, 22, 23, and 36 were drawn with Asymptote, a powerful new descriptive vector graphics language for technical drawing developed at the University of Alberta. The term *vector graphics* refers to a method for producing figures that retain their high quality even at arbitrarily large magnifications.

The authors of Asymptote (Andy Hammerlindl, John C. Bowman, and Tom Prince) would like to thank the Natural Sciences and Engineering Research Council of Canada, the Pacific Institute for Mathematical Sciences, and the University of Alberta Faculty of Science for their generous financial support. Asymptote is freely available, under the GNU General Public License, from the web site <http://asymptote.sourceforge.net>, which includes a gallery of example Asymptote code and output.