The Steiner Ratio Conjecture

Ian Stewart

16/09/2008

Mathematics Institute
University of Warwick
Coventry CV4 7AL
UK

"But the shortest paths between points are *straight lines!*" shouted Miles Spanning in exasperation.

"I'm not disputing that," said Horatio Steiner quietly. "I'm just suggesting that we don't have to use existing towns as points."

"Then where do you propose to put all these extra points?"

"That's precisely the question I'm asking."

Peak District Cable, a new company with big ambitions but a budget more in line with these recessionary times, was considering laying a network of cables that would connect together the three towns of Loughborough, Stoke-on-Trent, and Rotherham into an informatic superhighway. Its two senior managers were having trouble agreeing on strategy. Miles was arguing that the best idea was to pick one town and link it to the others by two straight cables.

"Which town?" asked Horatio.

"That depends upon the distances between them. But it's simple enough to decide. Look at the triangle formed by the towns and discard its longest side: lay cable along the other two."

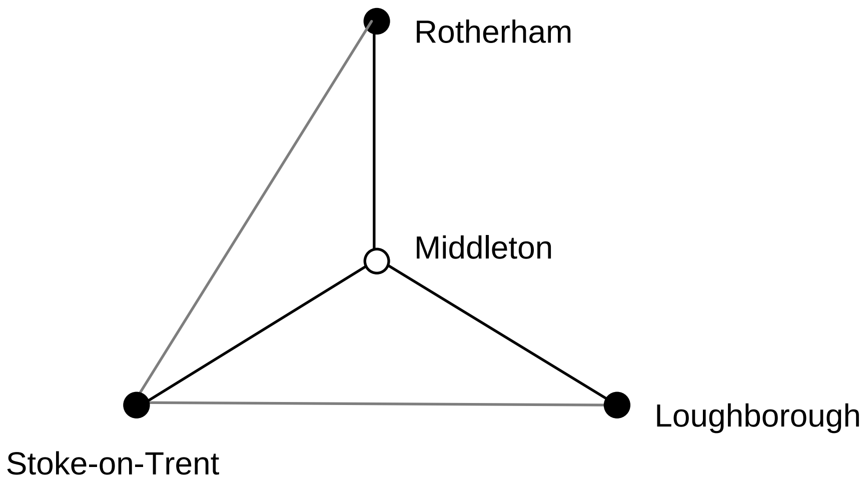
"Yeah, *sounds* sensible. But..." Horatio had got it into his head that adding an extra town to the network might actually make it *shorter*. It may seem unlikely that he could be right — surely extra towns need extra cable? — but if he *is*, where should the extra town be located to save the most cable, and how much cable does it save?

As it happens, Rotherham, Loughborough and Stoke-on-Trent are all the same distance apart, 75 km. Miles's network will be 150 km long, no matter which of the three towns is used as the central link. What about Horatio's hare-brained suggestion? Nestling in the Derbyshire dales near Matlock, roughly in the middle of the triangle formed by the three towns, is a village called — quite coincidentally — Middleton. It is roughly 44 km from each town.

"Look, Miles: if we locate a node of the network at Middleton, and run cables from there to each of the three towns, I'm sure we'll save cable."

"Rubbish. All right, smartypants: if we save cable, how much?"

"Well, if we run one link from Middleton there to each of the three towns," Horatio mused, "then the total will be 3_44 = 132 km instead of 150 km. That's a saving of 18km, or about 12%." (See **Fig.1.**)



How to save cable by adding a town. The dotted network is 150km long, the solid one only 132 km.

"What?" yelled Miles. He scratched his head, then laughed. "Well, I'll be hornswoggled — you're right. It does save cable, doesn't it."

"Well, it certainly can't waste it," Horatio pointed out. "Nobody is *forcing* us to create new nodes, Miles."

"True," said Spanning. The fax machine bleeped but they both ignored it.

"Of course," Horatio continued, "we could ask the same questions for any collection of towns that you like. What's the shortest network that links them all, with nodes — connections — only at the towns? And what's the shortest network if you add extra nodes elsewhere?"

"Yes — and how much better can the network with extra nodes be?" said Miles. "*That's* the crucial question."

"I'll go and check up on the database," said Steiner.

"Right. I'll get Maggie to sort out these faxes." But their secretary was nowhere to be seen.

Several hours later Horatio returned. "We're not the first to ask those questions, Miles. There's a huge literature. In particular, in 1968 Edgar Gilbert and Henry Pollak of AT&T's Bell Laboratories conjectured that no matter how the towns are initially located, the maximum saving in cable that can be obained by adding extra towns is 13.34%. This conjecture has become known as the *Steiner Ratio Conjecture*."

"Why not the Gilbert-Pollak Conjecture?"

"Because of the usual rules for mathematical attribution."

"Which are?"

"Name the idea after someone historical who wasn't responsible for it but is vaguely connected with the problem, Miles. Anyway, it now has to be called the Steiner Ratio *Theorem* —"

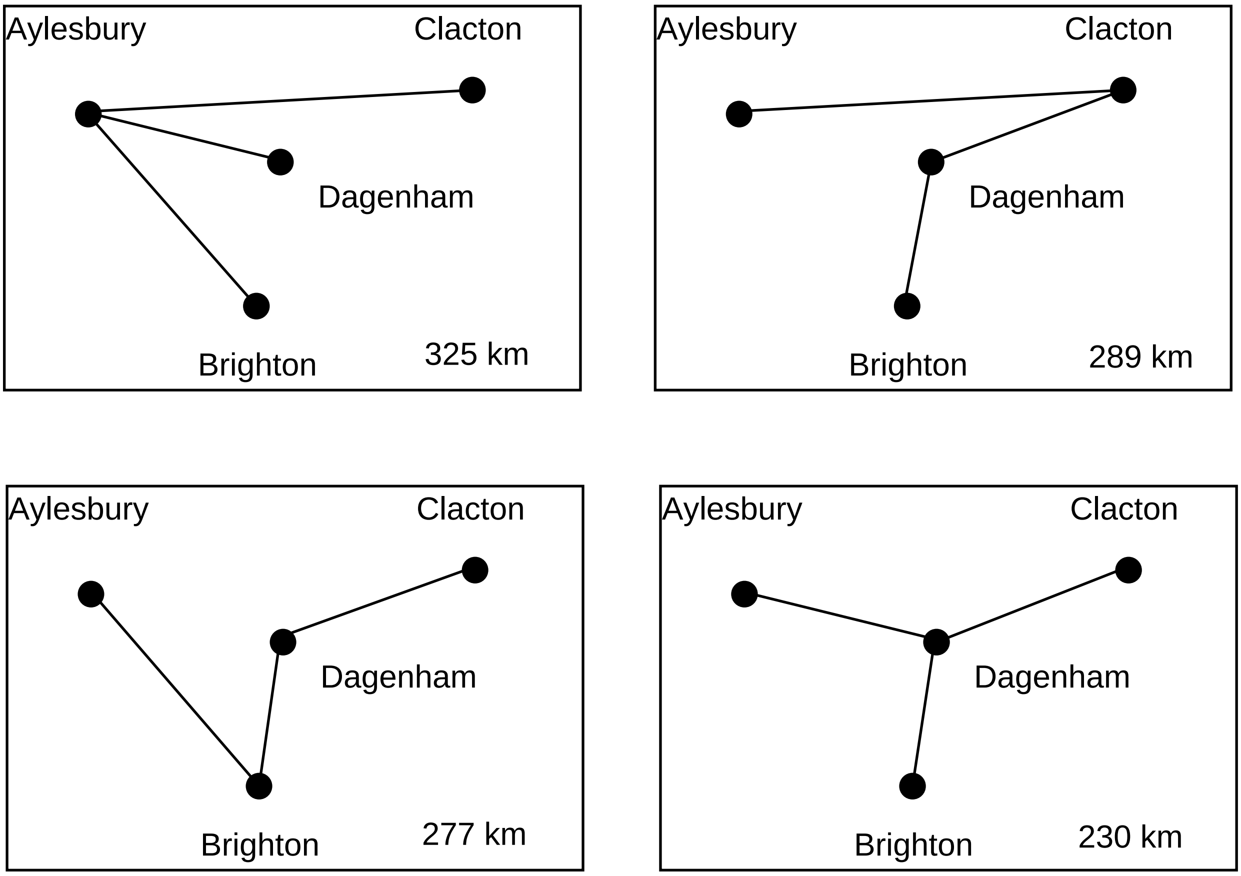
"Gilbert-Pollak Theorem," said Miles firmly.

"Actually, Du-Hwang Theorem would be fairest. Because in 1991, after 23 years of unrelenting but largely unrewarded effort, it was proved by Ding Zhu Du at Princeton University and Frank Hwang at Bell Labs."

Miles wandered over to the fax machine and glanced idly at the messages that once more were piling up. He stuck a plastic cup into a coffee machine and was rewarded by some brownish liquid. "I can see why Bell might be interested, what with AT&T being a telephone company."

"Yes. There's an interesting story about that, but I'll leave it for a moment until I've formulated the problem more mathematically. In a formal setting the towns to be connected are represented by points in the plane, and the cables linking them are straight lines. Whether or not we invent new towns, it's clear that the links must form a *tree* — a network without any loops. Loops just waste cable, joining up towns that are *already* joined by some other route. If no new towns are involved, then such a network is called a *spanning tree*. There are a lot of spanning trees to choose from, but in principle you can just list them all and see which is the shortest."

For example, suppose there are four towns: Aylesbury, Brighton, Clacton, and Dagenham. **Fig.2** shows some of the possible spanning trees and their lengths.



Four of the 16 possible spanning trees for four particular towns. The one at lower right is the shortest among all 16.

The shortest one has a 'branch point' at Dagenham, from which three links run to the other three towns. On the other hand, if the towns are Ashburton, Bristol, Cheltenham, and Daventry, which are roughly in a straight line, then you can easily convince yourself that the shortest spanning tree joins them in that order and has no 'branch points'.

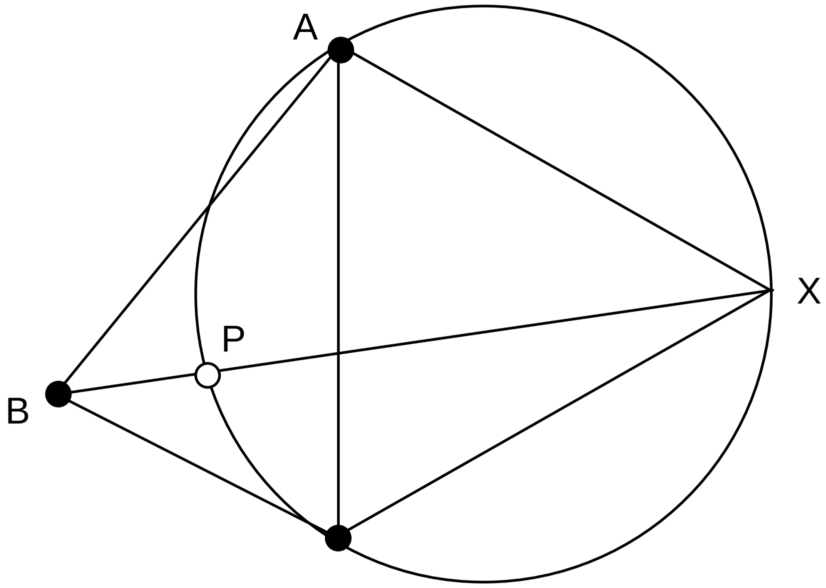
The problem is far more subtle if links are allowed to meet out in the countryside as well. For example, if there are three towns at the vertices of an equilateral triangle, as in the opening example (Rotherham, Loughborough, Stoke), then the shortest network joins all three to the centre of the triangle (Middleton — or more precisely a nondescript place somewhere in a field not far from it). The shortest such network must also be a tree: it is called a *Steiner tree*.

"Who was Steiner, anyway?" asked Miles, when Horatio got this far in his explanation. "And what does he have to do with the problem?"

"Not much," replied Horatio. "Jacob Steiner was a 19th Century Swiss mathematician, who solved the problem for three towns in 1837. His name was attached to the problem by Richard Courant and Herbert Robbins in 1941, in their classic popularization *What is Mathematics?* They, and Steiner, seem not to have known that he was beaten to the punch by Evangelista Torricelli and Francesco Cavalieri around 1640."

"Wasn't Torricelli the chap who invented the barometer?"

"The very same. And Cavalieri was one of the forefathers of the calculus. Torricelli and Cavalieri broke the problem down into two different cases. If the triangle has an angle of 120_ or more, then the shortest network consists of just two links, joining that vertex to the other two. But if the triangle formed by the three towns has all of its angles less than 120_, then the shortest network consists of three links that lead from the towns to the *Steiner Point*, the unique place where all three roads meet at angles of 120_ (**Fig.3**).



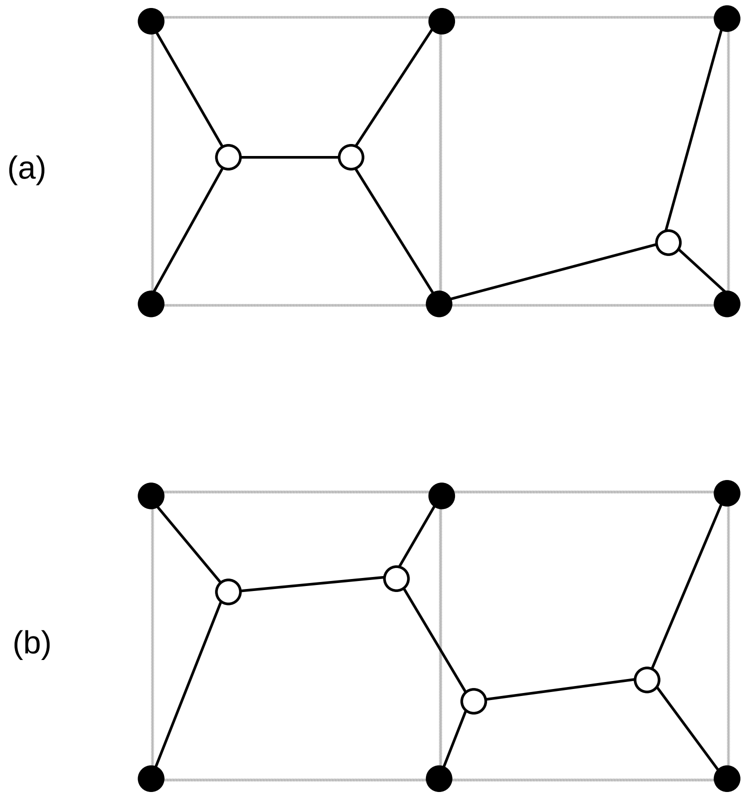
Constructing the Steiner point of triangle ABC. Draw an equilateral triangle ACX. Its circumcircle cuts BX at the Steiner point P.

"Steiner also proved that when there are several towns the edges of any Steiner tree must meet at 120_ at each new town added, a simple consequence of the solution for three towns. He was less forthcoming on how to find such trees. The problem of using a Steiner tree to join a larger number of towns than three was first investigated seriously by Milos Kössler and Vojtech Jarník in 1934.

"You can probably see why it's not easy to find the shortest Steiner tree in any given case," Horatio concluded.

"Yeah," said Miles. "It's a much more complicated calculation than finding the shortest spanning tree, because ever such a lot of possible new Steiner points have to be considered."

"Right. Let's try a simple example. Suppose there are six towns arranged at the corners of two adjacent squares (**Fig.4**)



(a) Combining Steiner trees for a square and an isosceles right triangle.

(b) A shorter Steiner tree for the same set of towns.

One possible Steiner tree is shown in Fig.4a: it is found by solving the problem for a square of four towns first, and then linking in the two remaining towns via their Steiner point with one that is already hooked in."

"Is that the answer?"

"No. *This* (Fig.4b) is the shortest Steiner tree."

"Curious."

"Yes. It shows that you can't build up shortest Steiner trees piecemeal."

"I see that," said Miles. "I wonder where Maggie's got to? These faxes really ought to be dealt with... oh, what the heck. What was Gilbert and Pollak's contribution?"

"They asked whether the two versions of the problem might be related. Call the length of the shortest spanning tree the *spanning length* of the set of towns, and that of the shortest Steiner tree the *Steiner length*. Now, every spanning tree is also a Steiner tree (either invent no new towns, or put them on top of the existing links). So, for any set of towns, the spanning length is always greater than or equal to the Steiner length. The question is: how much greater can it be?"

"Well," said Miles, "for an equilateral triangle of unit side, the spanning length is 2 and the Steiner length is œ3."

"Precisely," said Horatio. "So in that case the ratio between the Steiner length and the spanning length is œ3/2 = 0.866. Which means that —"

"The saving in length obtained by using the shortest Steiner tree rather than the shortest spanning tree is about 13.34%."

"No need to look so smug. Now, Gilbert and Pollak's Steiner Ratio Conjecture states that you can *never* do better than this. That is, for any number of towns arranged in any possible manner, the ratio of the Steiner length to the spanning length is always greater than or equal to œ3/2."

"You mean that the saving in length obtained by using the shortest Steiner tree instead of the shortest spanning tree is never more than 13.34%?"

"Exactly, Miles."

"So we shouldn't spend more money trying to compute the optimal solution than 13.34% of the total cost of laying the amount of cable in the shortest spanning tree?"

"You got it. Unless we could do the calculation cheaply, it would save money to use a quick-and-dirty solution with a spanning tree. The cost of getting the answer is crucial."

Similar — though in practice much more complicated — problems arise whenever a company is planning how to route telephone lines, gas pipes, cable TV networks, buses, trains, or aeroplanes. A closely related question turns up in one approach to the evolution of living organisms. The genetic material of living creatures is DNA, which 'encodes' developmental information as a sequence of four types of base (adenine, thymine, cytosine, guanine). In this (grossly simplified) picture, genetic information is specified by a long sequence of bases such as AATTCGCTCA... . In the application of Steiner trees, the 'towns' are sequences of DNA in different organisms, and the 'distance' is some measure of the similarity between different sequences, such as what proportion of corresponding bases are equal. Steiner points correspond to 'most plausible common ancestors'. There is of course no guarantee that this common ancestor actually existed; but the method provides interesting clues to how the DNA molecule might have changed and how the organisms are related genetically.

There are also many other problems analogous to that of Steiner trees. The most important practical ones occur in the design of electronic circguits. Here the connections are generally laid out on a rectangular grid, running only horizontally or vertically; but the same kinds of questions can be asked, and similar methods may help with their solutions.

The Steiner ratio conjecture is important for the 'economics' of all such networks, because — as we shall see — it is much easier to find the shortest spanning tree than the shortest Steiner tree. So it may therefore be worth sacrificing the 13.34% error to save computational effort.

"You know," said Horatio, "that until quite recently AT&T used the spanning length to charge customers who wanted to connect their offices together. But it was worried that customers might discover that they could make substantial savings on their bills by inventing imaginary offices located at appropriate Steiner points! The conjecture limits such savings to 13.34%, which is not too embarrassing."

"Why didn't AT&T save itself the worry and use the Steiner length itself?" asked Miles.

"It couldn't."

"Why ever not?"

Horatio sighed. "Finding the spanning length is a simple computation, even for a huge number of towns. You can solve it using the *greedy algorithm*: start with the shortest link you can find, and at each stage thereafter add on the shortest remaining link that doesn't complete a closed loop, until every town is included in the tree."

"What's an algorithm, Horatio?"

"A specific process that a computer can carry out, with a guarantee that it will get the answer."

"We could do with an algorithm for finding where Maggie's gone, then — we've got about twenty kilometres of faxes here. Oh well. Why is it called the 'greedy' algorithm when it picks the *shortest* link at each stage?"

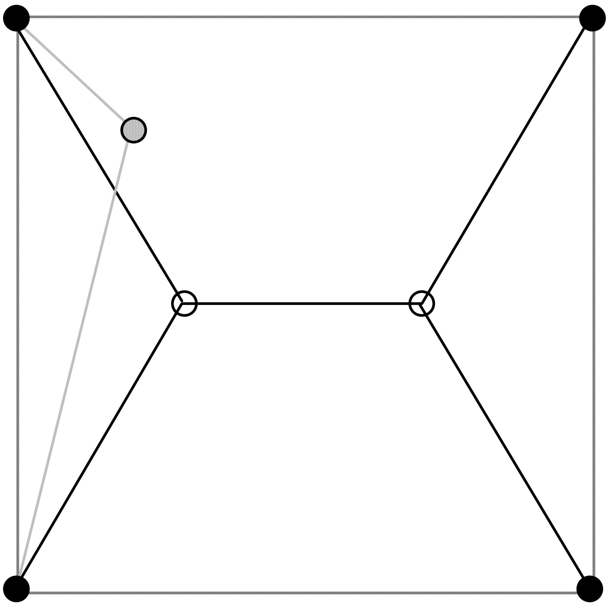
"I think because it resembles other algorithms which solve problems by always choosing the biggest thing available, which have become known as 'greedy'. It minimizes things instead of maximizing them, but the idea is really the same, so the same name is used."

"Ought to be called the 'frugal algorithm'," muttered Miles. "Anyway, what about an algorithm for finding the shortest Steiner tree, then?"

"That's the problem," said Horatio, looking glum. "It's very hard to do that."

"Can't you just take all possible triples of towns, find their Steiner points, and then look for the shortest tree that joins the towns together and meets either at towns or at Steiner points?"

"No. The correct generalization of 'Steiner point' to a whole set of many towns is any point at which a system of links can meet at 120_. For as simple an example as four towns at the vertices of a square, such points are *not* Steiner points of any subset of three towns." (See **Fig.5**.).



Steiner points (white) for four towns in a square (black) are different from the Steiner points of a subset of three towns (grey).

"I see. And there are infinitely many points in the plane..."

"Indeed. Even though most of them are probably irrelevant, it's not obvious that finite algorithms exist at all!"

Miles looked worried. "Do they?"

"Yes — the first was invented by Z.A.Melzak of the University of British Columbia. But his method isn't really practical, it becomes unwieldy even for moderate numbers of towns. It has since been improved, but not dramatically."

Miles stared at the ever-growing fax mountain, then shook his head like a dog emerging from a pond. "What makes it so hard?"

Horatio leaned back in his chair and made an expansive gesture. "There are very good reasons why these solutions are inefficient, Miles. The growing use of computers has led to the development of a new branch of mathematics, Complexity Theory. This studies not just algorithms — methods for solving problems — but how efficient those algorithms are."

"I don't follow you."

"Well, given a problem involving some number n of objects (here towns), how fast does the running time of the solution algorithm grow as n grows? If the running time grows no faster than a constant multiple of a fixed power of n, such as $5n^2$ or $1066n^4$, then the algorithm is said to run in *polynomial time*, and the problem is considered to be 'easy'. Usually this means that the algorithm is practical (unless the constant is absolutely huge). If the running time grows non-polynomially — faster than any constant multiple of powers of n, for instance exponentially, like 2^n or 10^n — then the problem has non-polynomial running time and is 'hard'. Usually this means that the algorithm is totally impractical."

"Give me some examples."

"OK. Adding two n-digit numbers requires at most $2n$ one-digit additions, including carries, so the time taken is bounded by a constant multiple (namely 2) of the first power of n. Long multiplication of two such numbers involves about n^2 one-digit multiplications and no more than $2n^2$ additions, or $3n^2$ operations on digits, so now the bound involves only the second power of n. The opinions of schoolchildren notwithstanding, these problems are therefore 'easy'."

"Got it."

"Good. Now, in contrast, consider the Travelling Salesman Problem: find the shortest route that takes a salesman through a given set of cities. If there are n cities then the number of routes that we have to consider is $n! = n(n-1)(n-2)...3.2.1$ which grows faster than any power of n. So case-by-case enumeration is hopelessly inefficient."

"Right," said Miles. A thought struck him. "What about in between polynomial time and exponential time?"

"That's a no man's land of 'fairly easy' or 'moderately hard' problems, where practicality is more a matter of experience."

"I see."

"Oddly enough," continued Horatio, "the big problem in Complexity Theory is to prove that the subject actually exists. That is, to prove that some 'interesting' problem really is hard."

"How about answering all these faxes? That's getting exponentially harder by the minute."

"Be serious. The difficulty is that it is easy to prove a problem is easy, but hard to prove that it is hard! To show a problem is easy, you just exhibit *one* algorithm that solves it in polynomial time. It doesn't have to be the best, or the cleverest: any will do. But to prove that a problem is hard, it is not enough to exhibit some algorithm with non-polynomial running time. Maybe you've chosen a silly one, maybe there's a better one which *does* run in polynomial time. You have to find some mathematical way to consider *all possible algorithms* for the problem, and show that *none* of them runs in polynomial time."

"Yeah, I can see that might be difficult. Any good guesses?"

"Well, there are lots of plausible *candidates* for hard problems — the travelling salesman problem, the bin-packing problem (how can you best fit a set of items of given sizes into a set of sacks of given sizes?), and the knapsack problem (given a fixed size sack, and lots of objects, does any set of objects fill the bag exactly?). But so far nobody's managed to *prove* that any of them really are hard. However, in 1971 Stephen Cook of the University of Toronto showed that if you can prove that any one problem in this candidate group really is hard, then they all are. Roughly speaking, you can 'code' any one of them to become a special case of one of the others: they sink or swim together. These problems are called *NP–complete*, where NP stands for 'non-polynomial'. Everyone *believes* they really are hard."

"Why?"

"Probably on the psychological grounds that you can't expect to get all good things at once. Anyway, Ron Graham, Michael Garey and David Johnson of AT&T have proved that the problem of the Steiner length is NP-complete. An efficient algorithm to find the precise Steiner length for any set of towns would automatically lead to efficient solutions to all sorts of computational problems that are widely believed not to possess such solutions."

"So everything we'd like to know about algorithms reduces to a specific problem about communication networks," said Miles in amazement.

"Right. Or to any of a thousand other specific problems. Solve one, you solve 'em all. The Steiner ratio conjecture is also important because it shows you can replace a hard problem by an easy one without losing very much. Which means that the methods used to prove it are also important."

"And what are they?"

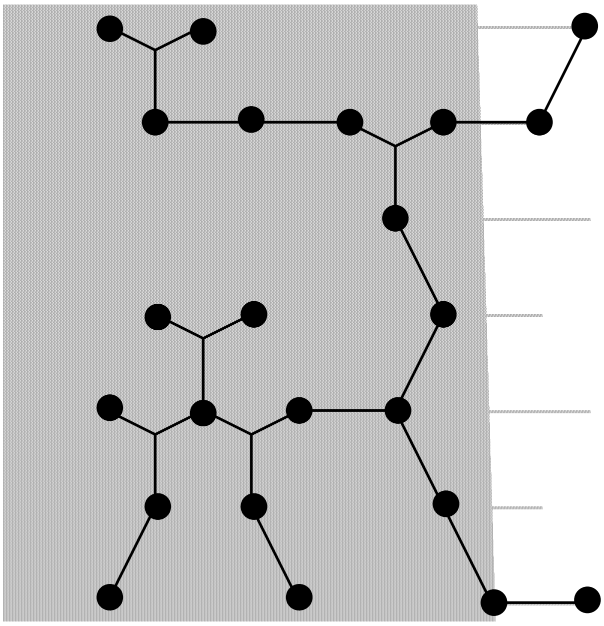
"Well, the equilateral triangle example that started the whole business is very natural. It suggests that there must be a simple proof of the conjecture. However, its simplicity may be deceptive. If there is a simple proof, nobody has ever found it. Even the Du-Hwang proof is quite tricky. And, leaving their proof aside, direct attacks for small numbers of towns lead to vast and messy calculations. Gilbert and Pollak had quite a lot of evidence for their conjecture; and in particular they could show that something along those lines must be true: they proved that the ratio is always at least 0.5. By 1990 various people had performed heroic calculations to verify the conjecture completely for networks of 4, 5, and 6 towns. For general arrangements of as many towns as you like, they also pushed up the limits on the ratio from 0.5 to 0.57, 0.74, and 0.8. A few years ago Graham and Fang Chung at Bell Communications Research raised it to 0.824, in a computation that they describe as 'really horrible — it was clear it was the wrong approach'."

"That," said Miles with heavy sarcasm, "is a really helpful thing to say."

"More helpful than you migth think. It focusses attention on more abstract and conceptual approaches. To make further progress possible, the horrible calculations have to be simplified. Du and Hwang found an approach that is so much better that it does away with the horrible calculations completely. The basic question is how to get equilateral triangles in on the act. There's a big gap between the triangle example, which sets up the bound on the ratio, and a general system of towns, which is supposed to obey the same bound. How can this no man's land be crossed?"

"Come to that, when will these faxes be answered?."

"Shhh, this bit's really fascinating. There's a kind of halfway house. Imagine the plane tiled with identical equilateral triangles, in a triangular lattice (**Fig.6**).



A Steiner tree for towns that lie on a triangular lattice has a much more rigid and regular structure than that for general towns. Du and Hwang reduced the Steiner ratio conjecture to the same problem for lattice trees.

Put towns only at the corners of the tiles. It turns out that the only Steiner points that need be considered are the centres of the tiles. In short, you have a lot of control, not just on computations, but on theoretical analyses."

"Brilliant. But not every set of towns conveniently lies on a triangular lattice."

"Of course not. Du and Hwang's insight is that the ones that really matter do. Suppose the conjecture is false. Then there must exist a *counterexample*: some set of towns for which the ratio is *less* than œ3/2. They show that if a counterexample to the conjecture exists, then there must be a counterexample in which all the towns lie on a triangular lattice. This introduces an element of regularity into the problem, and it is then relatively simple to polish it off."

Miles considered this for a moment. "Fine, but how do they prove this lattice property?"

"It's a wonderful exercise in lateral thinking. First they reformulate the conjecture as a minimax problem. Such problems arise in game theory, where players compete and try to limit (*mini*mize the *max*imum of) the gains (payoff) made by their opponents. Game theory was invented by John von Neumann and Oskar Morgenstern in their classic *Theory of Games and Economic Behavior* of 1947. In the Du-Hwang version of the Steiner ratio conjecture, one player selects the general 'shape' of the Steiner tree, and the other picks the shortest one of that shape that they can find. Du and Hwang deduce the existence of a lattice counterexample by observing that the payoff for their game has a special 'convexity' property. This elegant new method neatly disposes of a question that previously looked totally intractable, and cuts through a mass of tangled calculations and case-by-case investigations to give a 'clean', conceptually simple solution. The Du-Hwang method is not easy, and it does require a certain amount of mathematical technique; but it is such a dramatic improvement that it knocks all previous approaches on the head. More importantly, it provides a paradigm for investigating analogous questions. Formulate the problem in game-theoretic terms, prove a suitable convexity property, reduce it to a combinatorially 'rigid' question with many fewer possibilities — and then solve that, by whatever method you can. The motto 'think first, calculate later' should be engraved on every mathematician's heart."

"Quite," said Miles. He walked across to the ever-growing pile of faxes. "Do you think you could try to formulate the problem of locating Maggie in game-theoretic terms?" He began pawing through the twisted heap of paper. "Oh, bugger!"

"What?"

"One of these unanswered faxes is from Loughborough Town Council. They've decided not to award us the contract for the informatic superhighway linking them to Stoke-on-Trent and Rotherham."

Horatio's face fell. "Why not?"

Miles glared at him. "AT&T underbid us by 13.34%."

FURTHER READING

Marshall W. Bern and Ronald L.Graham, The shortest-network problem, *Scientific American*, January 1989, 66-71.

Barry A. Cipra, Euclidean geometry alive and well in the computer age, *SIAM News* **24**, January 1991, 1, 16-17, 19.

Richard Courant and Herbert Robbins, *What is Mathematics?*, Oxford University Press, Oxford 1961.

Martin Gardner, *Mathematical Magic Show*, Penguin Books, Harmondsworth 1985.

E.N.Gilbert and H.O.Pallak, Steiner minimal trees, *SIAM Journal of Applied Mathematics* **16** (1968) 1-29.

Z.A.Melzak, *Companion to Concrete Mathematics*, Wiley, New York 1973.

Hugo Steinhaus, *Mathematical Snapshots,* Oxford University Press, Oxford 1960.

Pawel Winter, Steiner problems in networks: a survey, *Networks* **17** (1987) 129-167.