DAN KALMAN

When Existence is Enough

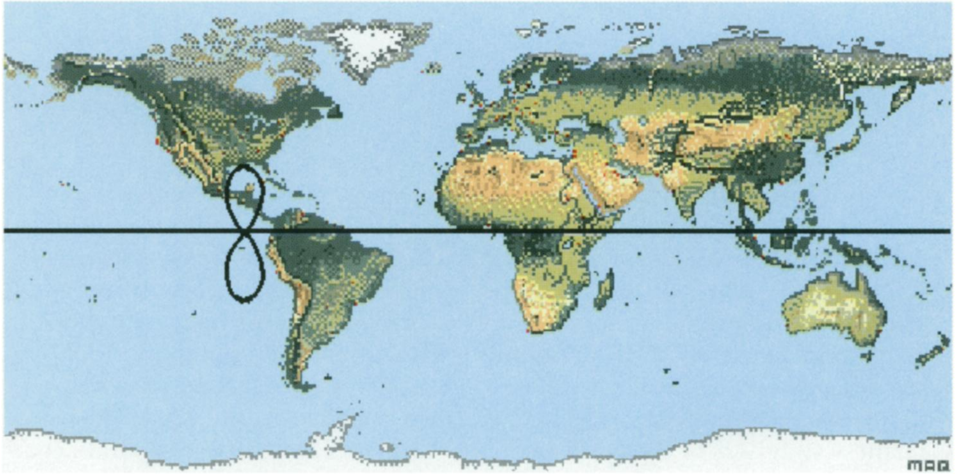

Figure 1

A mathematician, a chemist, and an engineer shared a house. One night a fire broke out. The mathematician woke up, smelled the smoke, and deduced that there was a fire. After careful consideration, he declared "There exists a solution!" and went back to sleep. The chemist, too, woke up and smelled the smoke. She noted that there were a number of surprising chemical compounds in the smoke, and started trying to figure out what reactions were taking place. Meanwhile, the engineer also woke up and smelled the smoke. Like his house mates, he deduced that there was a fire. So he jumped out of bed, grabbed a fire extinguisher, and put the fire out.

This story is one of several that play on differences between the methods and interests stereotypically attributed to specialists in different disciplines. Some of the stories poke fun at one kind of specialist, and in this example, it is the engineers who are having a laugh at the expense of chemists, and especially, mathematicians. It is true that sometimes mathematicians are content to prove that a problem has a solution, even if there is no obvious way to find it. And sometimes this attitude is criticized as being theoretical to the point of absolute uselessness. But there really are situations where knowing about existence is enough, even in quite practical problems. Let me tell you about one that I worked on in the aerospace industry.

DAN KALMAN is the Associate Executive Director, Director of Member Services and Programs for the Mathematical Association of America.

Satellite Communication Systems

The problem I will describe came up in connection with the design of a satellite communication system. As you can imagine, satellites are expensive to manufacture and put in orbit, so a lot of preliminary planning goes into the design. Everyone wants to make sure that the system will provide the greatest amount of utility at the least cost. For communications systems, modeling when it is possible for a particular station on the ground to exchange data with a particular satellite is of fundamental concern. There are usually constraints imposed by the electronics that can be approximated by simple geometric rules. For example, if a line drawn from the station to the satellite is too close to the horizon, it may not be possible for the station and satellite to communicate. That can determine a rule of this sort: to communicate, the satellite must be at least three degrees above the horizon, as viewed from the station. The simplest rule is simply that the satellite must be in view from the station. In that case, we say that the satellite and the station can see each other.

It has been known from the time of Newton how to work out equations describing the motion of satellites. And the motion of the Earth itself is certainly well understood. Given some information about the orbit of a satellite, this allows you to predict when the satellite will be able to see a particular point on the ground. That in turn can be used to make predictions about the amount of time available for communication between a satellite and a ground station.

In developing a design for a communications system, one important consideration centers on the orbits that the satellites will be placed in. There are a wide variety of orbits, with different characteristics. Some remain over a fixed spot on the equator at all times. Others trace out a ground track that repeats a figure 8 shape, as in Figure 1 or trace a succession of shifted figure eights (Figure 2). There are other possibilities, as well.

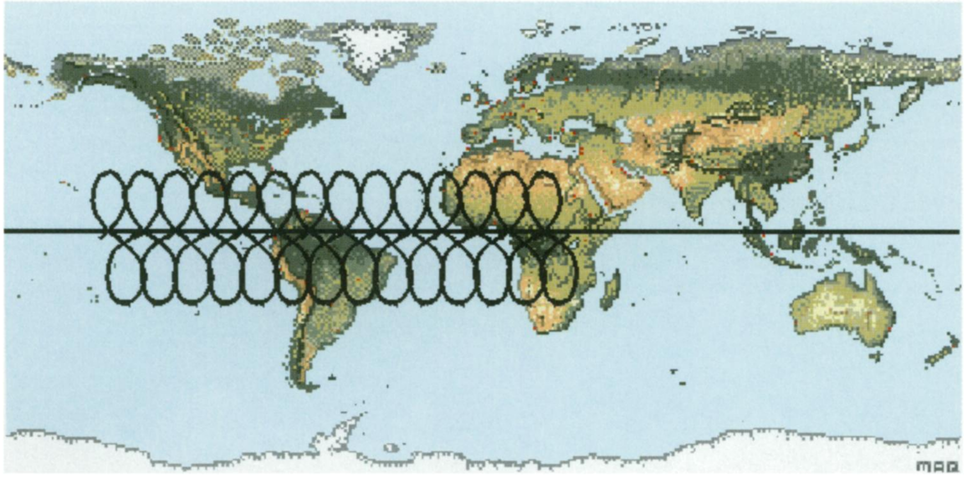

Figure 2

How high up the satellite is determines how fast it completes an orbit, and how much of the Earth's surface is visible at any one time. Some orbits are circular, others are in the shape of an ellipse. In an elliptical orbit, the satellite moves fastest when it is closest to the earth and most slowly when it is furthest away. All of these variations have an effect on how long a satellite will be visible from a ground station. So there is a tremendous number of variables to consider in deciding how many satellites to use and what orbits to put them in.

With that as a background, I can now describe a simulation that was used in comparing different designs for the communication system. For each design, there were a number of satellites, from as few as 6 to as many as 20. There was also an established network of ground stations. For each ground station, a certain amount of time was supposed to be available each day to send information up to the network of satellites. For example, there might be stations at Los Angeles, Denver, Chicago, and Boston. For the Los Angeles station, perhaps 60 minutes of satellite contact per day are required. For New York the required amount of contact time could be 105 minutes. The numbers are not particularly important in this discussion. I just want to get across an idea of the overall situation. For each city, some number of minutes of contact time will be a given in the problem.

What I have referred to as a simulation was actually a computer program. The input data for the program specified the orbits for each satellite, as well as the location of the satellite at the start time for the simulation. The locations of the ground stations were also included in the input. The program would perform a separate computation for each minute of the simulated period. So if the program was supposed to simulate the operation of the communication system over a 24 hour period, which means $24 \times 60 = 1440$ minutes, then there would be a separate calculation for each of the 1440 minutes.

What was computed for each minute? The answer has several parts. First, the position in space of each satellite was computed. Similarly, the location of each ground station was determined. And then, for every possible pairing of a satellite with a ground station, the program computed whether or not communication was possible. At the end of the simulation, the output data specified on a minute by minute basis, which satellites could see which ground stations.

The Assignment Problem

Now we are coming to the tricky part of the problem. Each satellite can only communicate with one ground station at a time. At any given minute, a satellite might be visible to many different ground stations. And from each ground station, there might be several different satellites in view. If you were actually running this system, you would have to come up with an assignment schedule. At each minute, you would have to specify which satellite would be talking to which ground station. And deciding how to make those assignments is not an easy job.

Suppose you simply make the assignments at random. Go down the list of satellites, and at each minute, assign the satellite to talk with one ground station (which must be visible to the satellite). So, at the first minute, satellite 5 might be talking to Los Angeles. At the second minute, you could assign satellite 3 to talk to Los Angeles. Perhaps at the third minute no satellite would be assigned to Los Angeles. Without any particular plan, you just assign satellites to talk to stations.

But wait a minute. Los Angeles wants to have access to the communications network for at least 60 minutes. If you do the assignments at random, Los Angeles might get the desired 60 minutes of connect time, and it might not. The trick is to do the assignments so every ground station has access to the network for the required number of minutes per day.

That is a huge problem. If there are 20 satellites and 10 ground stations, and 1440 time steps, the number of possible combinations is astronomical. You might fiddle around with assignments for a year and never find one that provided each ground station with its required allocation of air time. In fact, how could you even know whether the required assignments were possible? The problem is to assign each satellite to some ground station at each minute in such a way that each ground station accumulates its predetermined quota of air time. How do we know that a solution exists?

This is where the existence question comes into play. This situation can be analyzed using graph theory. And there is a theorem of graph theory that predicts when a solution exists. In essence, that theorem says nothing about how to find a solution. On a deeper level, if you examine the proof of the theorem, it does suggest a method of searching for the solution, but it is not a practical method—it amounts to an exhaustive search of all possible assignments, and remember that there are an astronomical number of possibilities. The ingenious part of the proof is that it provides a way to recognize when the exhaustive search will be successful, without ever performing the search. Instead, you make some computations based on the characteristics of the graphs involved, and in a very short amount of time you can work out whether the problem is solvable.

Existence Proofs

The situation I have described is the classical setting for an existence result. A problem is to be solved. A direct assault on computing a solution is not feasible. But something else about the problem is computable, something that predicts whether or not a solution exists. If you feel that there is something unsatisfying about this kind of proof, you are not alone. In fact, it is a relatively recent arrival on the mathematical scene. It was in 1888 that David Hilbert, who would go on to become one of the preeminent mathematicians of the turn of the century, stunned the mathematical world by publishing a completely original and unexpected solution to a problem of long standing: Gordan's problem. Gordan's problem concerned invariants of algebraic equations. It is not important here to explain just what an invariant is. You can think of it as an identity that remains unchanged when certain kinds of transformations are performed. It is enough here to know that Gordan's problem asked whether there was a finite set of invariants that could be used to construct the complete set of invariants. Before Hilbert, all of the work in the area had approached this problem by setting out to construct the desired finite set subject to some additional assumptions. But Hilbert took an entirely different approach. He gave an existence proof. It gave no clue how to find the desired finite set of invariants. But it showed that such a set was forced to exist as a logical necessity.

Hilbert's proof was initially quite controversial. The lack of a concrete construction of the solution to Gordan's problem left many mathematicians with an uneasy feeling. Gordan himself, in whose honor the problem was named, declared of Hilbert's work:

This is not mathematics. This is theology.

Others were quick to recognize the power of Hilbert's method. One of the leading mathematicians of the day, Felix Klein, called it "wholly simple and, therefore, logically compelling." Klein's view soon prevailed, Hilbert was recognized for his accomplishment, and the existence proof was firmly established as an important method of mathematical research. And Gordan? In the lyrical account of Constance Reid, his pronouncement "has echoed in mathematics long after his own mathematical work has fallen silent." Not recognizing the value and significance of the existence proof is what Gordan is best remembered for.

> *"wholly simple and, therefore, logically compelling"*
>
> —Felix Klein

Today, the typical course of study in college level mathematics acquaints students with existence arguments in a variety of settings. One of the most familiar arises in connection with differential equations. There, existence and uniqueness theorems provide a means for predicting when a differential equation has a solution. The conditions that must be met to invoke these theorems have to do with properties of the equations: differentiability, continuity, that sort of thing. And the conclusion simply says that there exists a function that adheres to the conditions of a differential equation. The theorems do not provide a method for finding the solutions.

A second example concerns transcendental numbers. You are probably familiar with the hierarchy of successively more complicated number systems. Simplest are the positive integers. Next comes the full set of integers, positive and negative, then rational numbers. There are irrational numbers, like $\sqrt{2}$, that arise as solutions to polynomial equations. The fact that no rational number satisfies $x^2 - 2 = 0$ implies that $\sqrt{2}$ is irrational.

There are infinitely many irrational numbers that come about as roots of polynomial equations. Indeed, if you go around defining polynomials at random, the likelihood is that your polynomials will have irrational roots. But are there any other numbers besides the ones I have named so far? Are there any numbers that are not integers, not rationals, not even roots of polynomials with coefficients that are rationals? The answer is yes. These are called transcendental numbers, and there is a simple proof of their existence based on counting. As first developed by Cantor, a contemporary of Hilbert, there are different sizes of infinite sets. There are infinitely many integers, it is true, and there are infinitely many real numbers. But the infinity that describes how many real numbers exist is far larger than the one that describes the integers. Surprisingly, one can show that the rational numbers form an infinite set with the same size of infinity as the integers. Likewise, there are no more roots of rational coefficient polynomials than there are rational numbers themselves. So the much larger number of reals must be made up of irrational numbers that are not roots of rational coefficient polynomials, that is, transcendental numbers. There are simply not enough roots to fill up all the reals. That is an existence proof. It is meant to convince that transcendental numbers exist, but does not give any indication of how to find one.

But how did the existence proof from graph theory solve any real problems about designing satellite systems? As you will recall, that theorem from graph theory provided a method to determine whether or not the problem of assigning satellites to stations was solvable. In fact, it reduced the issue to a feasible computation. The data describing which satellites could be seen by which stations at each minute of the simulation needed to be tabulated in a certain way, to generate a set of statistics. These statistics were entered into a series of inequalities. If all the inequalities were satisfied, that guaranteed the existence of a solution; if even one of the inequalities was not satisfied, no solution could exist.

This was all built into the computer program that performed the simulation. After running the program, you obtained a printout that either confirmed that the desired assignment schedule was possible, or stated definitively that no such schedule existed.

You can probably appreciate already why this information might have been useful in the context of designing a communication system. For a given proposed set of satellite orbits, the simulation told us whether it would be possible to provide the required amount of communications time. But one of my colleagues suggested taking the analysis a step further. He pointed out that the amount of air time required at each ground station was determined by the rate of data transmission. If you doubled the transmission rate, that would cut the required amount of time in half. So, he suggested a procedure for finding the minimal data transmission rate at which a proposed satellite system could provide the required service. Start with a first guess at the transmission rate. Run the simulation. If the result was that no solution existed for the assignment problem, then increase the transmission rate a little bit. That would proportionally lower each ground station's requirements for connect time. Or, if the simulation said that the assignment problem did have a solution, we would lower the transmission rate a little, thereby driving up the required amount of connection time. In either case, each time the transmission rate was adjusted, the conditions in the existence theorem would be recomputed, leading to a new round of adjustments. After a few repetitions, this method led to a pretty good estimate of the minimal feasible transmission rate.

Ultimately, this method was used to compute the minimal feasible transmission rate for a large number of competing satellite designs. The results provided one way to compare different designs. In general, the higher the required transmission rate, the more expensive the system becomes. There are many other considerations that also enter into the cost and effectiveness of a communications system. The minimal feasible transmission rate was just one of many criteria used to evaluate and compare the different systems. But it definitely played a role in reaching a solution of a real design problem. And it did so using a theoretical result concerned only with the existence of solutions. At that preliminary design phase, it really was not ever necessary to figure out how to assign the satellites to ground stations to achieve the optimal use of resources. Instead, for this problem, demonstrating the existence of a solution was enough. ■

Footnotes

1. Constance Reid, *Hilbert*, Springer Verlag, New York, 1970