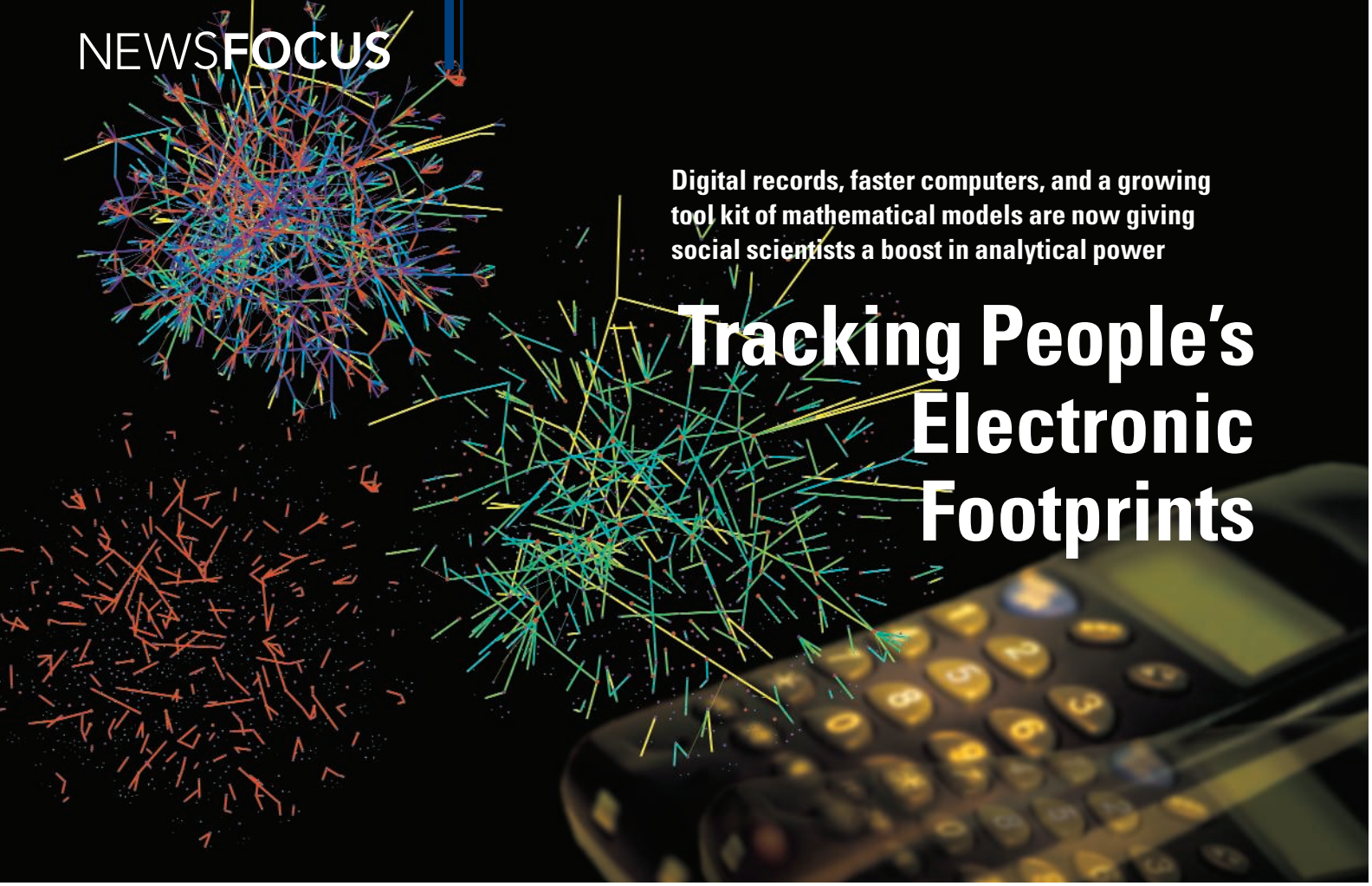


Digital records, faster computers, and a growing tool kit of mathematical models are now giving social scientists a boost in analytical power

Tracking People's Electronic Footprints

OXFORD, UNITED KINGDOM—The audience perked up noticeably when physicist Jukka-Pekka Onnela clicked to the slide showing his results—something like a big, colorful hairball. The average viewer might not be impressed. But it caused a buzz among the scientists meeting here recently to talk about complex networks.* The vast flurry of points and lines represents relationships between people in a communication network. What makes it remarkable is that it is no simulation: The data are from actual telephone calls among 7 million real people over an 18-month period.

The data set was given to Onnela and his team at Helsinki University of Technology and the University of Oxford by a mobile telephone company, after replacing phone numbers with codes. "I felt a little surge of jealousy," admitted Marco van der Leij, an economist at Erasmus University in Rotterdam, the Netherlands. Social scientists have dreamed for decades of getting their hands on such a global lode of data.

The mobile phone data set was one of a variety of new collections on display at the meeting—many of them based on the captured digital signatures of human interactions such as communication, travel, voting, and shopping. These interactions have long been the bread and butter of the social sciences. But researchers have been frustrated by the size and complexity of the phenomena they study. Electronic footprints, faster computers, and a growing tool kit of mathematical models are now giving researchers a boost in analytical power.

* European Conference on Complex Systems, Saïd Business School, University of Oxford, U.K., 25–29 September 2006.

Up close and personal

Some of the new data sets are downright intimate. Take for example a study by Oxford sociologist Peter Hedström of the records of the 3 million people above the age of 16 who lived in Stockholm from 1990 to 2003. After an ethics panel granted approval, the Swedish government gave Hedström data covering everything from workplace absenteeism and divorces to taxes, school grades, and criminal records. (Names and addresses were replaced by codes.)

Hedström's goal is to see how the decisions of individuals add up to large-scale patterns such as unemployment, crime, and gender bias. "We often resort to hand waving" in trying to make the connection between individual behavior and social phenomena, he says. Having data for individuals in an entire society allows questions to be asked that "traditional social scientists simply could not address."

For example: Are suicides contagious? The traditional method of studying the social causes of suicide "has been either to do small case studies or try to include some questions in larger surveys about the very local networks individuals are embedded in," says Hedström. But he notes that this approach will never capture a complete web of social interactions. Hedström's team is trying to track the ripple effect caused by each of the 2621 recorded suicides in Stockholm over a decade by looking for the social connections that link them. Although the results are "preliminary," he says, they indicate that the chance that exposure to a suicide will tip an already unstable person into taking his or her own life is related to the strength of the social ties. "Not surprisingly," he says, "the suicide of a family member has the strongest effect on an individual's suicide risk." But a suicide in a school or workplace exposes far more people, so although the individual effect may be smaller, "the public health effect is large."

Others, such as Onnela, are studying the architecture of social webs. His team is interested in how information flows through soci-

ADAPTED FROM NETWORK DATA: WWW.ARXIV.ORG/ABS/PHYSICS/0610104; (PHOTO) PHOTOS.COM

Published by AAAS

Intimate links. Researchers are probing a data set of real calls made by 7 million telephone users in an unnamed European country.

ety, and how the network imposes "constraints," he says. His data set of 7 million people represents 20% of the population of a European country where 90% have mobile phones. (The team agreed to keep the country's identity secret.) Aside from the very young and old, says Onnela, "this is a good representation of the entire society." Because the phone records contain no personal information, the researchers characterized relationships by weighing the "intimacy" of the links based on the number and duration of phone conversations. Because the data only include calls between mobile phones, most business calls are excluded, says Onnela, because most businesses use landline phones. "We think this is a reasonable proxy" for intimacy, says Onnela.

To examine patterns of diffusion, Onnela's team "infected" a single individual in a simulated version of the real network with a piece of information and watched it spread, with the chance of it passing between two people determined by the intimacy of their relationship. The result suggests that a classic idea in network theory—that large, complex networks tend to maximize flow efficiency—does not apply. The information tended to become trapped within tightly knit communities rather than spreading freely across the society.

Probing the network further, Onnela's team blocked the phone connections between people in different categories, starting with the most intimate relationships. In another case, they started from the opposite end, severing the least intimate relationships. The difference is dramatic. Although losing 20% of the most intimate connections causes individual communities to break down, society's interconnections hold together, and information still flows from one end to the other. But after the same fraction of the weakest links are cut, the system shatters into islands (see figure on p. 914). Van der Leij calls this the first large-scale, empirical confirmation of a theory, first proposed in 1973 by Mark Granovetter, a sociologist at Stanford University in California, that "for keeping society connected, acquaintances are more important than close friends."

The big picture

On the macro end of the scale, the search is on for fundamental rules that may undergird collective behavior. This work is aided by recent progress on the mathematics of networks (*Science*, 4 August, p. 604). But "getting our

Google's Hidden Wealth

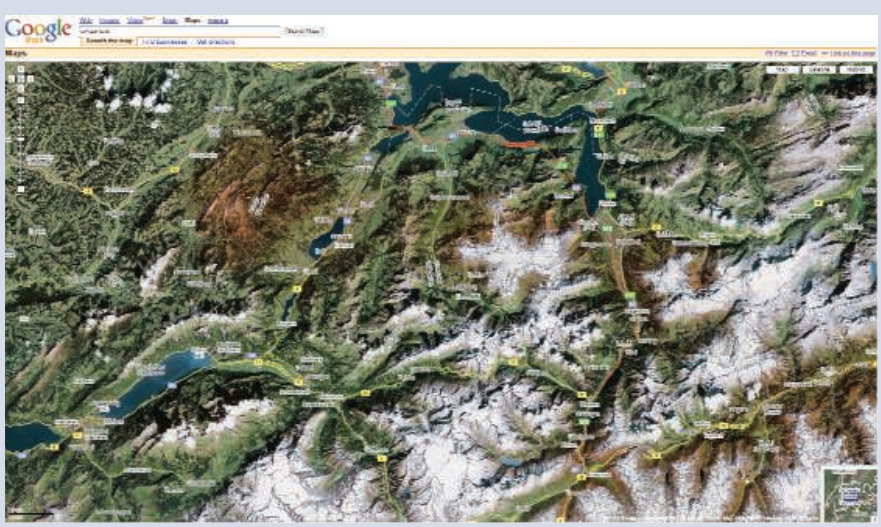
Type the word "science" into the Google search engine, and a list of one-and-a-half million Web pages appears in a fraction of a second. Behind this service lies an enormous reservoir of data that researchers would like to harness for science of their own, in fields from social psychology to global economics. But although some computer-based companies such as Microsoft have eagerly embraced scientific collaboration, Google so far has not. "Google has a reputation … for being very negative to letting researchers in," says Richard Swedberg, a sociologist at Cornell University. This could soon change, a Google spokesperson has told *Science*.

Google's data are a potential social science gold mine, "both for observing social interactions in real time and also for measuring their consequences for individual and collective behavior," says Duncan Watts, a sociologist at Columbia University. The key is the electronic "cookie." As you browse the Internet, many Web sites such as Google's record a string of text—the cookie—representing the identity of your computer. And when you use Google, its servers keep track not only of what you search for but also where you go next. People add new entries to this record at the rate of 200 million Web searches per day. This electronic record is key to Google's business model: Most of its $1 billion annual revenue comes from Internet advertising targeted to individuals.

Google expanded its reach in 2001 when it acquired the largest group of Internet-based communities, or "chat groups," known as Usenet and rechristened as Google Groups, including Usenet's records of topic-specific conversations between 25 million people going back to 1981, all of it searchable. And Google is amassing other treasures, such as its regularly updated satellite-based map of Earth. Users can instantly retrieve many kinds of sociological data such as local crime rates from that map. Thousands of people are voluntarily developing new (but not peer-reviewed or verified) layers of data with so-called mash-ups that are freely available on the Internet.

Google has been cautious about scientific collaboration because "we don't want to give users the impression that we're free and easy with their data," says Rachel Whetstone, a London-based Google spokesperson, "especially in light of what happened with AOL." In August, the Internet company American Online (AOL) released a record of Internet searches done by 650,000 people. A furor erupted when it was discovered that people's identities were easily reconstructed from the data. AOL removed the data from the Internet 3 days later, but the file had already been downloaded and replicated worldwide. In what may be Google's first invitation, the company's public relations department said in an e-mail to *Science* that "Google wants to support scientific endeavors" and would consider providing data for "legitimate scientific research … so long as we could ensure that it included no personally-identifiable information."

Some academics are urging caution. There is "significant potential for abuse, given the ease of transporting computerized data," says Frank Miller, a bioethicist at the National Institutes of Health in Bethesda, Maryland. "Ethics review committees will need to scrutinize research using such data very carefully to ensure that adequate protections are in place." Requiring people's consent will be difficult, he says, and "investigators might resist this move, as it could narrow the pool of subjects." **–J.B.**



Added value. Users are adding their own data overlays, or "mash-ups," to Google Earth.

CREDIT: GOOGLE EARTH

hands on real and sufficiently detailed empirical data is what is truly exciting and new," says Felix Reed-Tsochas, a theoretical physicist who now does network research at Oxford's Saïd Business School.

In an effort to understand how social networks survive stress, Reed-Tsochas, Serguei Saavedra, an engineer at Oxford, and Brian Uzzi, a sociologist at Northwestern University in Evanston, Illinois, are studying the New York City garment industry. In a complex web of collaborations, clothing is designed, manufactured from raw materials, distributed, and finally sold in retail stores. New York's industry shrank over 2 decades as garment production shifted to Asia, declining from 300,000 workers in 3000 firms during the 1980s to 190 firms today.

In spite of this big shrink, the network has held together and continued to function throughout. That robustness is a mystery, says Uzzi, because "there is no master planner," and "the individual actors are not even aware of the system beyond their local part of the network." When the team modeled the same contraction based on what is known about network dynamics, the garment industry quickly fell apart, he says.

Luckily for science, a New York garment workers union has kept a digital record since 1985 of 700,000 financial transactions among the firms and gave Uzzi access. Nearly all of the research on network dynamics has been based on periods of expansion, says Reed-Tsochas, but "this is the first well-characterized example of a network undergoing sustained contraction."

The researchers have created an evolving map of the flow of money. As companies went bankrupt, relocated, and cut budgets, the remaining ones were forced to decide which relations to sever and which to keep. The study is at an early stage, but some ingredients of the network's robustness are becoming clear, says Uzzi. The contraction looks like a movie of the expansion "played backwards in time," says Reed-Tsochas. The team has devised a model that, they say, can explain how robustness is an unintended consequence of individuals following their own self-interest based on local information. It will debut in a journal soon.

Reed-Tsochas and his colleagues built their model from a wealth of data. Social scientists studying the collective behavior of terrorist groups don't have that luxury: Members of such groups don't keep detailed records. But their deadly attacks are chronicled. To see what can be gleaned from such data, a pair of Oxford physicists, Neil Johnson and Sean Gourley, have teamed up with social scientists at the Conflict Analysis Resource Center (CERAC), based in Bogotá, Columbia. Researchers at CERAC have so far amassed a record of more than 55,000 attacks going back to the 1960s, compiled from other studies; they have also sifted information on events around the world from media and government reports, ranging in size from a single death to the 3000 killed at the World Trade Center.

Shrinkable. A study in New York City's garment district found that social networks remained strong during a period of attrition.

A striking pattern has emerged. When the researchers graphed all the attacks within a given conflict, with the number of attacks plotted against the number killed in each, it produces a fat-tailed exponential curve. And the exponent of the function, which determines the curve's shape, is nearly always the same. "Terrorism and guerrilla warfare everywhere in the world has a signature of about 2.5," says Gourley. Plotting the distribution of these events over time produces another, distinctive signature.

Johnson and Gourley have been building computer models of terrorism to see what kind of social networks can fit the patterns. Only one does the job, says Gourley, and it's a surprisingly simple model of human gregariousness. "All you need is to have people forming cohesive groups that share information, technology, and supplies," he says. Using this simplified social network model, they are drawing conclusions about the Iraq insurgency that are extremely difficult to assess from the ground. For example, "the bursty distribution of attacks over time shows that terrorists don't rely on a hierarchical organization to pass along orders, nor do they attack at random," says Gourley. Instead, "they must be coordinating by proxy," such as by reading the very same media reports of each other's attacks.

Johnson and Gourley also believe they can infer how many different factions are involved throughout Iraq. "In the first 180 days of the war, there were 15 to 35 groups," he says, and "after day 540, our model estimates there to be 100 to 130 different groups." The model assumes that each group is capable of no more than one attack per day, he adds, so that number could be lower if some groups are capable of multiple daily attacks.

The fact that all the conflicts around the world they have analyzed share these patterns "is extraordinary," says Gourley, "when you consider how different they are, involving actors with very different motives and goals, operating in very different environments." They must be following rules without being aware of them, he says: "There seem to be only a limited number of ways for people to form networks and coordinate activities."

Whether laws governing social groups can be found is an open question. But many social scientists are optimistic that such sets of real-world data will lead the way, and they are hungrily eyeing new sources (see sidebar on p. 915). "Great science can potentially come out of these efforts," says James Moody, a sociologist at Duke University in Durham, North Carolina. But he and others agree that it will take more than "just mining the data" to learn what drives social phenomena. What's needed is an exponential boost in the power of social science theory and analysis. And this, says Granovetter, "is a very tall order."

–JOHN BOHANNON

CREDIT: GAIL MOONEY/CORBIS

Published by AAAS