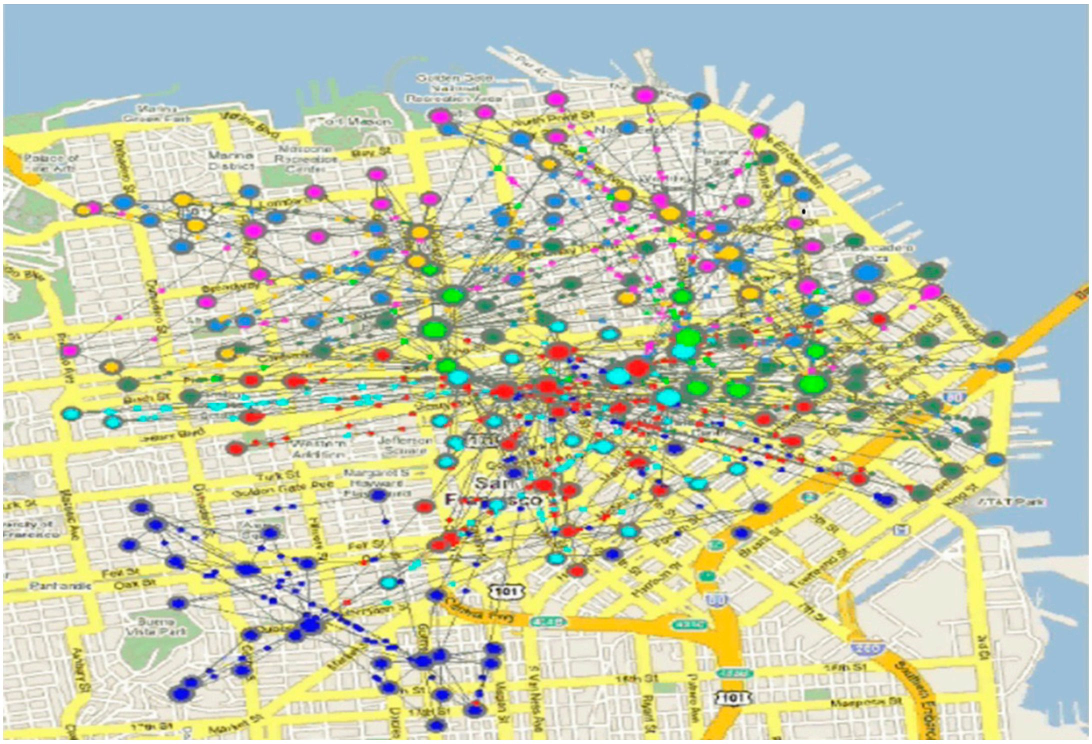
Computational social science

Adam Mann, *Science Writer*

Cell phone tower data predicts which parts of London can expect a spike in crime (1). Google searches for polling place information on the day of an election reveal the consequences of different voter registration laws (2). Mathematical models explain how interactions among financial investors produce better yields, and even how they generate economic bubbles (3).

These are just a few examples of how a suite of technologies is helping bring sociology, political science, and economics into the digital age. Such social science fields have historically relied on interviews and survey data, as well as censuses and other government databases, to answer important questions about human behavior. These tools often produce results based on individuals—showing, for example, that a wealthy, well-educated, white person is statistically more likely to vote (4)—but struggle to deal with complex situations involving the interactions of many different people.

A growing field called "computational social science" is now using digital tools to analyze the rich and interactive lives we lead. The discipline uses powerful computer simulations of networks, data collected from cell phones and online social networks, and online experiments involving hundreds of thousands of individuals to answer questions that were previously impossible to investigate. Humans are fundamentally social creatures and these new tools and huge datasets are giving social scientists insights into exactly how connections among people create societal trends or heretofore undetected patterns, related to everything from crime to economic fortunes to political persuasions. Although the field provides powerful ways to study the world, it's an ongoing challenge to ensure



Using cell-phone and taxi GPS data, researchers classified people in San Francisco into "tribal networks," clustering them according to their behavioral patterns. Student's, tourists, and businesspeople all travel through the city in various ways, congregating and socializing in different neighborhoods. Image courtesy of Alex Pentland (Massachusetts Institute of Technology, Cambridge, MA).

www.pnas.org/cgi/doi/10.1073/pnas.1524881113

that researchers collect and store the requisite information safely, and that they and others use that information ethically.

Society in High Resolution

Although it builds on traditional methods, computational social science is a young discipline. In February 2009, 15 researchers published a paper in *Science* announcing the emergence of the field (5). Computer scientist Alex Pentland of the Massachusetts Institute of Technology, one of the paper's coauthors, admits that declaring the birth of a new field was "a bit cheeky." But the article made a splash and has since been cited more than 500 times, according to the Web of Science.

New technology has made possible the types of observations driving the field's growth. A social scientist in the 1930s had to go door to door asking people how much money they spent last year. Today, researchers can follow transactions across an entire city, on millisecond timescales, through credit card data. This incredible abundance of data is allowing computational social science practitioners to tease out much more subtle, high-resolution results than older methods could have ever provided. "It's like having an electron microscope versus a light microscope," says sociologist Michael Macy of Cornell University in Ithaca, New York.

Powerful computer simulations have been a particular boon to the field. Starting in the mid- to late-2000s, researchers showed that as cities add more residents, many of their traits—from gross domestic products to patents per head to crime and sexually transmitted disease transmission rates—increase exponentially. For example, a doubling in population led to an average 130% increase in economic productivity. But nobody could figure out exactly why this should be.

Pentland and a team of colleagues investigated this phenomenon with a computer model that simulated social ties in virtual cities of from 10 thousand to 10 million residents (6). They found that, as the population density grew, the number of interactions each individual could have increased by an exponential factor. From their model, they derived a mathematical curve that almost perfectly predicted the observational data from cities around the globe.

The work (6) suggested possible ways to improve real-world cities that didn't seem to be living up to their potential; for example, in third-world countries the exponential increase in productivity didn't materialize, despite increasing populations. The team believes it's because the transportation networks in these places are usually underdeveloped, meaning that people can't get around and interact with one another easily. "So if you want to make a richer city, make transportation better," says Pentland.

Mining the Social Network

As digital means have come to dominate how we communicate, social scientists have also discovered a great deal more about our real-life interactions. Every day we share links on Facebook, publish pictures on Instagram, and listen to music on Spotify. "Each time we express our views, send an email, or post something online, we generate breadcrumbs of behavior," notes political scientist Solomon Messing of Stanford University in California.

Cell phone data in particular has become a valuable computational social science tool. Research from David Lazer and his colleagues has shown how mobile phones can lead to better predictions of unemployment rates (7). On the surface, the two don't seem to have anything in common. But cell towers provide a proxy for people's movements, and the employed have different movement patterns than the unemployed. The most obvious disparity: the employed tend to regularly travel back and forth between two points on weekdays.

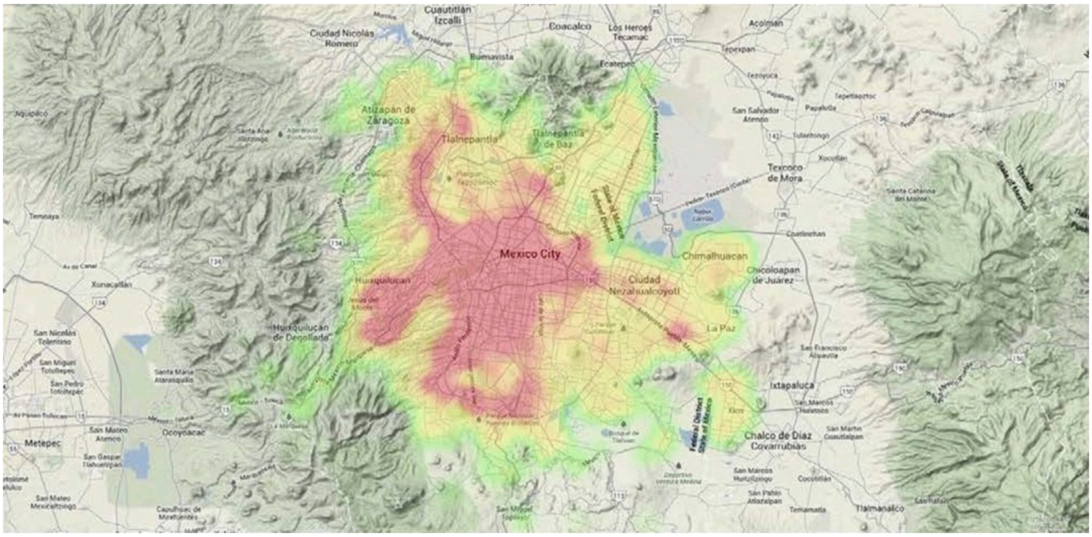
"The thing you have to remember about unemployment statistics is they're very slow and noisy," says Lazer, who teaches political science and communication at Northeastern University in Boston. It takes months to collect and publish such conventional unemployment data, which can contain errors resulting simply from the fact that sometimes people don't immediately admit that they're unemployed. Cell phone towers provided Lazer's team with information that was both more up-to-date and fine-grained than that which had been gathered via traditional means. With these data in hand, they were able to accurately forecast unemployment rates up to four months before the release of official reports.

Big Experiments, Big Pitfalls

Perhaps the most controversial thread in computational social science is randomized online experiments, similar to drug trials in which a single variable is manipulated to see its effects. Traditional behavioral science often relies on the responses of 30–40 volunteer psychology students, limiting their application to the general population. But social networks, such as Facebook, Reddit, or Wikipedia, have thousands or millions of users, providing ideal living laboratories to conduct such research.

One Facebook study showed that, during the 2010 election, users were more inclined to seek out polling place information and vote if they were presented with a message when they logged in telling them that close friends and family members had voted (8). In a later, controversial study on emotional contagion, Facebook preferentially displayed status updates with either positive or negative words to different users (9). Those who saw the positive output were slightly more likely to then post messages with more positive content, whereas those receiving negative updates did the opposite.

Although both studies received some condemnation, the second was widely decried by the popular press and Facebook users, who felt the network was manipulating them without their consent. And yet, notes Macy, variations of such experiments are far from rare. Online companies, such as Twitter and Amazon, he says, are constantly customizing what users see and running experiments to improve the



Where people hail from in the Mexico City area, here indicated by different colors, feeds into a crime-prediction model devised by Alex Pentland and colleagues (6). Image courtesy of Alex Pentland (Massachusetts Institute of Technology, Cambridge, MA).

user experience. “They don’t get our permission but it happens all the time,” Macy says.

The public outcry shows how careful social science researchers need to be when treading in these new waters. Even anonymous data have been shown, under some circumstances, to lead back to individuals (10). Computational social scientists need to work hard to secure their databases and make sure that hackers don’t steal private information. And although traditional social science measurements have been validated over many decades of study, researchers are still learning the true limits of these new techniques. A Facebook user could, in principle, be lying when he or she clicks the “I Voted” button, complicating the results of any work that uses this as a proxy for actual voter behavior.

Computational social science can seem like something straight out of the future, evoking Isaac Asimov’s fictional field of psychohistory from the *Foundation* series, in which the future can be perfectly predicted from the aggregate behavior of individuals. But the real practice is “much more grounded in reality than that,” says Messing. “There’s nothing magic or sci-fi. It’s just a lot of grunt work and math.”

1 Bogmolov A, et al. (2014) Once upon a crime: Towards crime prediction from demographics and mobile data. *Proceedings of the 16th International Conference on Multimodal Interaction* (ACM, New York), pp 427–434.

2 Street A, et al. (2015) Estimating voter registration deadline effects with web search data. *Polit Anal* 23(2):225–241.

3 Pan W, et al. (2012) Decoding social influence and the wisdom of the crowd in financial trading network, SOCIALCOM-PASSAT ’12. *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust* (Computer Society, Washington, DC), pp 203–209.

4 File T (2015) *Who Votes? Congressional Elections and the American Electorate: 1978–2014*. Available at https://www.census.gov/content/dam/Census/library/publications/2015/demo/p20-577.pdf. Accessed December 28, 2015.

5 Lazer D, et al. (2009) Social science. Computational social science. *Science* 323(5915):721–723.

6 Pan W, Ghoshal G, Krumme C, Cebrian M, Pentland A (2013) Urban characteristics attributable to density-driven tie formation. *Nature Commun* 4:1961.

7 Toole JL, et al. (2015) Tracking employment shocks using mobile phone data. *J R Soc Interface* 12(107):20150185.

8 Bond RM, et al. (2012) A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415):295–298.

9 Kramer AD, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. *Proc Natl Acad Sci USA* 111(24):8788–8790.

10 Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y (2013) Identifying personal genomes by surname inference. *Science* 339(6117):321–324.