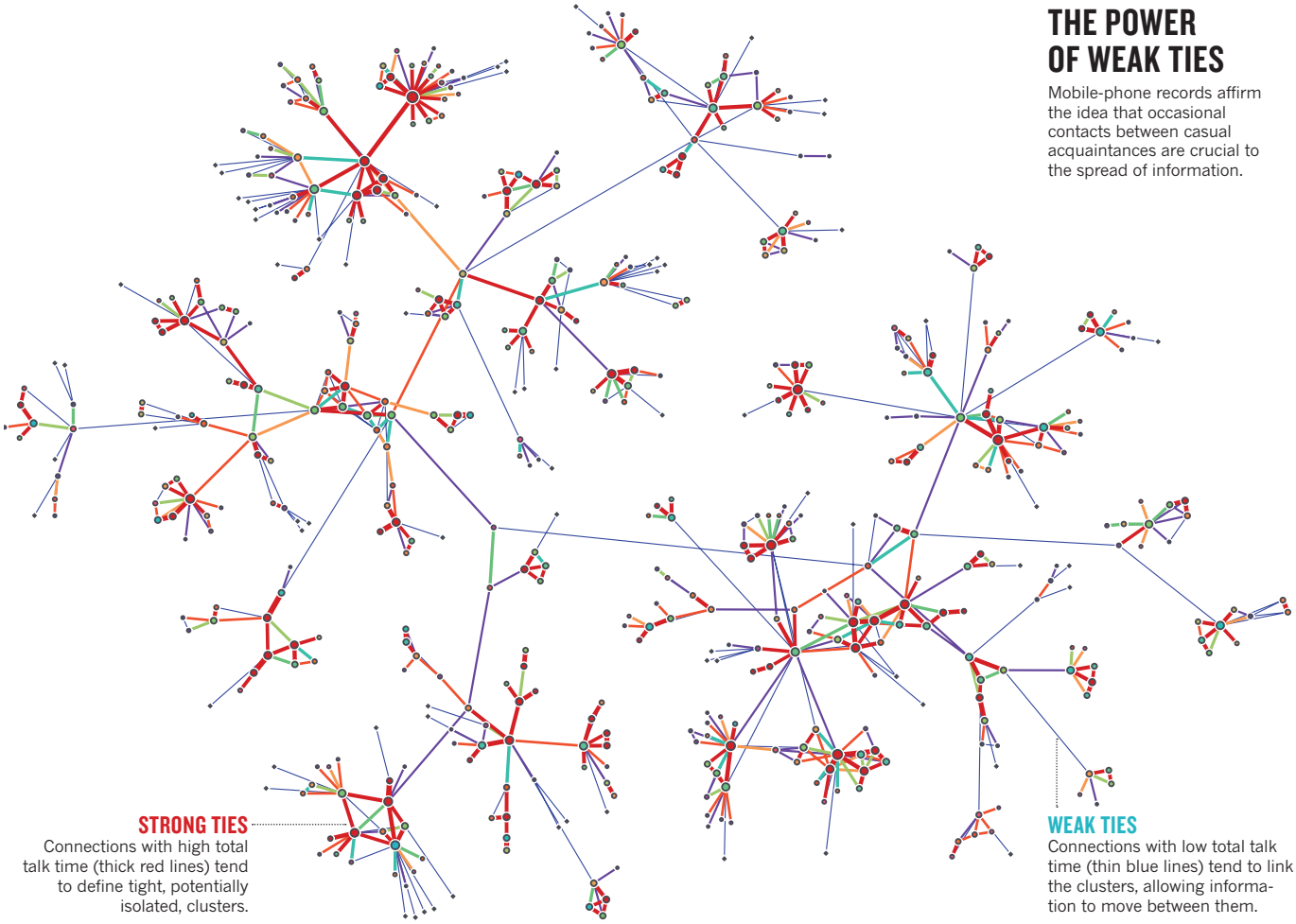


MAKING THE LINKS

FROM E-MAILS TO SOCIAL NETWORKS, THE DIGITAL TRACES LEFT BY LIFE IN THE MODERN WORLD ARE TRANSFORMING SOCIAL SCIENCE.

BY JIM GILES

Jon Kleinberg's early work was not for the mathematically faint of heart. His first publication[1], in 1992, was a computer-science paper with contents as dense as its title: 'On dynamic Voronoi diagrams and the minimum Hausdorff distance for point sets under Euclidean motion in the plane'.

That was before the World-Wide Web exploded across the planet, driven by millions of individual users making independent decisions about who and what to link to. And it was before Kleinberg began to study the vast array of digital by-products generated by life in the modern world, from e-mails, mobile phone calls and credit-card purchases to Internet searches and social networks. Today, as a computer scientist at Cornell University in Ithaca, New York, Kleinberg uses these data to write papers such as 'How bad is forming your own opinion?'[2] and 'You had me at hello: how phrasing affects memorability'[3] — titles that would be at home in a social-science journal.

"I realized that computer science is not just about technology," he explains. "It is also a human topic."

Kleinberg is not alone. The emerging field of computational social science is attracting mathematically inclined scientists in ever-increasing numbers. This, in turn, is spurring the creation of academic departments

and prompting companies such as the social-network giant Facebook, based in Menlo Park, California, to establish research teams to understand the structure of their networks and how information spreads across them.

"It's been really transformative," says Michael Macy, a social scientist at Cornell and one of 15 co-authors of a 2009 manifesto[4] seeking to raise the profile of the new discipline. "We were limited before to surveys, which are retrospective, and lab experiments, which are almost always done on small numbers of college sophomores." Now, he says, the digital data-streams promise a portrait of individual and group behaviour at unprecedented scales and levels of detail. They also offer plenty of challenges — notably privacy issues, and the problem that the data sets may not truly be reflective of the population at large.

Nonetheless, says Macy, "I liken the opportunities to the changes in physics brought about by the particle accelerator, and in neuroscience by functional magnetic resonance imaging".

SOCIAL CALLS

An early example of large-scale digital data being used on a social-science issue was a study in 2002 by Kleinberg and David Liben-Nowell, a computer scientist at Carleton College in Northfield, Minnesota. They looked at a mechanism that social scientists believed helped drive the formation of personal relationships: people tend to become friends with the friends of their friends. Although well established, the idea had never been tested on networks of more than a few tens or hundreds of people.

Kleinberg and Liben-Nowell studied the relationships formed in scientific collaborations. They looked at the thousands of physicists who uploaded papers to the arXiv preprint server during 1994–96. By writing software to automatically extract names from the papers, the pair built up a digital network several orders of magnitude larger than any that had been examined before, with each link representing two researchers who had collaborated. By following how the network changed over time, the researchers identified several measures of closeness among the researchers that could be used to forecast future collaborations[5].

As expected, the results showed that new collaborations tended to spring from researchers whose spheres of existing collaborators overlapped — the research analogue of 'friends of friends'. But the mathematical sophistication of the predictions has allowed them to be used on even larger networks. Kleinberg's former PhD student, Lars Backstrom, also worked on the connection-prediction problem — experience that he has put to good use now that he works at Facebook, where he designed the social network's current friend-recommendation system.

Another long-standing social-science idea affirmed by computational researchers is the importance of 'weak ties' — relationships with distant acquaintances who are encountered relatively rarely. In 1973, Mark Granovetter, a social scientist now at Stanford University in Stanford, California, argued that weak ties form bridges between social cliques and so are important to the spread of information and to economic mobility[6]. In the pre-digital era it was almost impossible to verify his ideas at scale. But in 2007, a team led by Jukka-Pekka Onnela, a network scientist now at Harvard University in Cambridge, Massachusetts, used data on 4 million mobile-phone users to confirm that weak ties do indeed act as societal bridges[7] (see 'The power of weak ties').

In 2010, a second group, which included Macy, showed that Granovetter was also right about the connection between economic mobility and weak ties. Using data from 65 million landlines and mobile phones in the United Kingdom, together with national census data, they revealed a powerful correlation between the diversity of individuals' relationships and economic development: the richer and more varied their connections, the richer their communities[8] (see 'The economic link'). "We didn't imagine in the 1970s that we could work with data on this scale," says Granovetter.

INFECTIOUS IDEAS

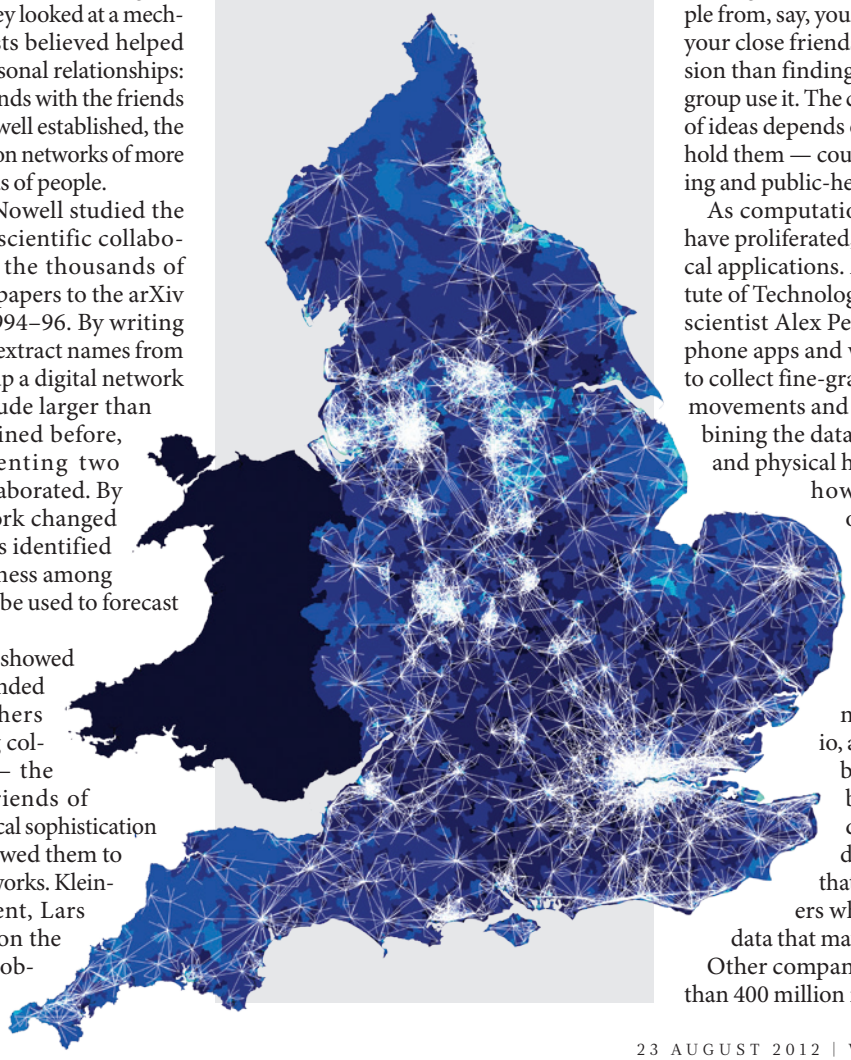
In some instances, big data have showed that long-standing ideas are wrong. This year, Kleinberg and his colleagues used data from the roughly 900 million users of Facebook to study contagion in social networks — a process that describes the spread of ideas such as fads, political opinions, new technologies and financial decisions. Almost all theories had assumed that the process mirrors viral contagion: the chance of a person adopting a new idea increases with the number of believers to which he or she is exposed.

Kleinberg's student Johan Ugander found that there is more to it than that: people's decision to join Facebook varies not with the total number of friends who are already using the site, but with the number of distinct social groups those friends occupy[9]. In other words, finding that Facebook is being used by people from, say, your work, your sports club and your close friends makes more of an impression than finding that friends from only one group use it. The conclusion — that the spread of ideas depends on the variety of people that hold them — could be important for marketing and public-health campaigns.

As computational social-science studies have proliferated, so have ideas about practical applications. At the Massachusetts Institute of Technology in Cambridge, computer scientist Alex Pentland's group uses smartphone apps and wearable recording devices to collect fine-grained data on subjects' daily movements and communications. By combining the data with surveys of emotional and physical health, the team has learned how to spot the emergence of health problems such as depression[10]. "We see groups that never call out," says Pentland. "Being able to see isolation is really important when it comes to reaching people who need to be reached." Ginger.io, a spin-off company in Cambridge, Massachusetts, led by Pentland's former student Anmol Madan, is now developing a smartphone app that notifies health-care providers when it spots a pattern in the data that may indicate a health problem. Other companies are exploiting the more than 400 million messages that are sent every

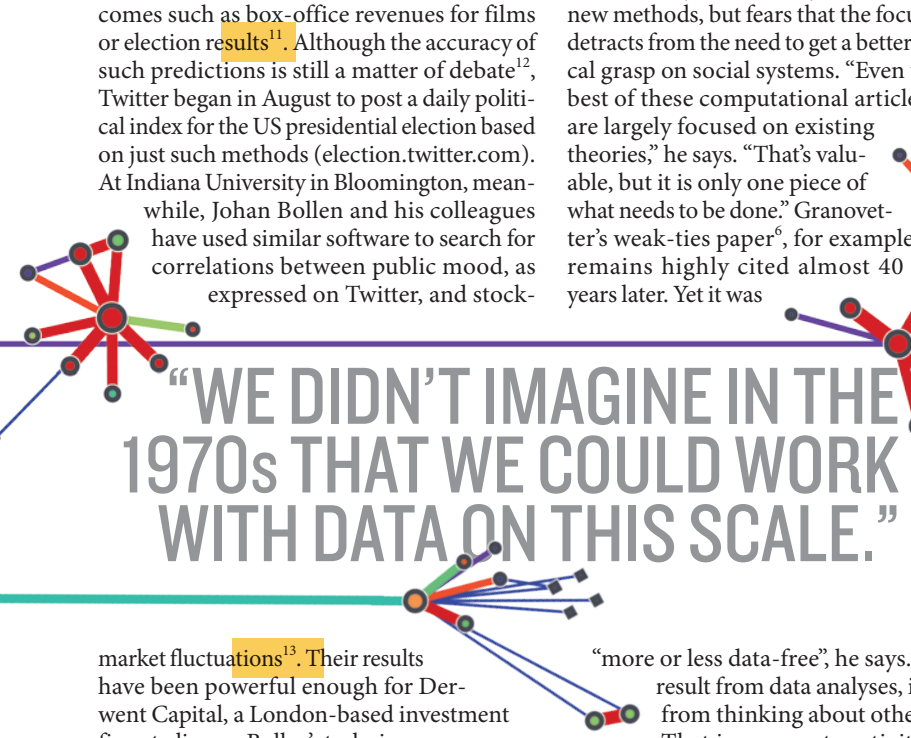
THE ECONOMIC LINK

British telephone records show that England's communication diversity (white links) correlates strongly with higher economic prosperity (light blue).



R. CLAXTON/BT GROUP/REF. 8

day on Twitter. Several research groups have developed software to analyse the sentiments expressed in tweets to predict real-world outcomes such as box-office revenues for films or election results[11]. Although the accuracy of such predictions is still a matter of debate[12], Twitter began in August to post a daily political index for the US presidential election based on just such methods (election.twitter.com). At Indiana University in Bloomington, meanwhile, Johan Bollen and his colleagues have used similar software to search for correlations between public mood, as expressed on Twitter, and stock-market fluctuations[13]. Their results have been powerful enough for Derwent Capital, a London-based investment firm, to license Bollen's techniques.



> "WE DIDN'T IMAGINE IN THE 1970s THAT WE COULD WORK WITH DATA ON THIS SCALE."

MESSAGE RECEIVED

When such Twitter-based polls began to appear around two years ago, critics wondered whether the service's relative popularity among specific demographic groups, such as young people, would skew the results. A similar debate revolves around all of the new data sets. Facebook, for example, now has close to a billion users, yet young people are still overrepresented among them. There are also differences between online and real-world communication, and it is not clear whether results from one sphere will apply in the other. "We often extrapolate from how one technology is used by one group to how humans in general interact," notes Samuel Arbesman, a network scientist at Harvard University. But that, he says, "might not necessarily be reasonable".

Proponents counter that these are not new problems. Almost all survey data contain some amount of demographic skew, and social scientists have developed a variety of weighting methods to redress the balance. If the bias in a particular data set, such as an excess of one group or another on Facebook, is understood, the results can be adjusted to account for it.

Services such as Facebook and Twitter are also becoming increasingly widely used, reducing the bias. And even if the bias remains, it is arguably less severe than that in other data sets such as those for psychology and human behaviour, where most work is done on university students from Western, educated, industrialized, rich and democratic societies (often denoted WEIRD).

Granovetter has a more philosophical reservation about the influx of big data into his field. He says he is "very interested" in the new methods, but fears that the focus on data detracts from the need to get a better theoretical grasp on social systems. "Even the very best of these computational articles are largely focused on existing theories," he says. "That's valuable, but it is only one piece of what needs to be done." Granovetter's weak-ties paper[6], for example, remains highly cited almost 40 years later. Yet it was "more or less data-free", he says. "It didn't result from data analyses, it resulted from thinking about other studies. That is a separate activity and we need to have people doing that."

The new breed of social scientists are also wrestling with the issue of data access. "Many of the emerging 'big data' come from private sources that are inaccessible to other researchers," Bernardo Huberman, a computer scientist at HP Labs in Palo Alto, wrote in February[14]. "The data source may be hidden, compounding problems of verification, as well as concerns about the generality of the results."

A prime example is Facebook's in-house research team, which routinely uses data about the interactions among the network's 900 million users for its own studies, including a re-evaluation of the famous claim that any two people on Earth are just six introductions apart. (It puts the figure at five[15].) But the group publishes only the conclusions, not the raw data, in part because of privacy concerns. In July, Facebook announced that it was exploring a plan that would give external researchers the chance to check the in-house group's published conclusions against aggregated, anonymized data — but only for a limited time, and only if the outsiders first travelled to Facebook headquarters[16].

In the short term, computational social scientists are more concerned about cultural problems in their discipline. Several institutions, including Harvard, have created programmes in the new field, but the power of academic boundaries is such that there is often little traffic between different departments. At Columbia University in New York, social scientist and network theorist Duncan Watts recalls a recent scheduling error that forced him to combine meetings with graduate students in computer science and sociology. "It was abundantly clear that these two groups could really use each other: the computer-science students had much better methodological chops than their sociology counterparts, but the sociologists had much more interesting questions," he says. "And yet they'd never heard of each other, nor had it ever occurred to any of them to walk over to the other's department."

Many researchers remain unaware of the power of the new data, agrees David Lazer, a social scientist at Northeastern University in Boston, Massachusetts, and lead author on the 2009 manifesto. Little data-driven work is making it into top social-science journals. And computer-science conferences that focus on social issues, such as the Conference on Weblogs and Social Media, held in Dublin in June, attract few social scientists.

Nonetheless, says Lazer, with landmark papers appearing in leading journals and data sets on societal-wide behaviours available for the first time, those barriers are steadily breaking down. "The changes are more in front of us than behind us," he says.

Certainly that is Kleinberg's perception. "I think of myself as a computer scientist who is interested in social questions," he says. "But these boundaries are becoming hard to discern." ■

Jim Giles *is a freelance writer in San Francisco.*

1. Huttenlocher, D. P., Kedem, K. & Kleinberg, J. M. *Proc. 8th Annu. Symp. on Computational Geometry* 110–119 (1992).
2. Bindel, D., Kleinberg, J. & Oren, S. *Proc. IEEE 52nd Annu. Symp. Foundations of Computer Science* 57–66 (2011).
3. Danescu-Niculescu-Mizil, C., Cheng, J., Kleinberg, J. & Lee, L. *Proc. 50th Annu. Meeting Assoc. Computational Linguistics* (in the press); Preprint at http://arxiv.org/abs/1203.6360.
4. Lazer, D. *et al. Science* **323,** 721–723 (2009).
5. Liben-Nowell, D. & Kleinberg, J. *J. Am. Soc. Inf. Sci. Technol.* **58,** 1019–1031 (2007).
6. Granovetter, M. S. *Am. J. Sociol.* **78,** 1360–1380 (1973).
7. Onnela, J.-P. *et al. Proc. Natl Acad. Sci. USA* **104,** 7332–7336 (2007).
8. Eagle, N., Macy, M. & Claxton, R. *Science* **328,** 1029–1031 (2010).
9. Ugander, J., Backstrom, L., Marlow, C. & Kleinberg, J. *Proc. Natl Acad. Sci. USA* **109,** 5962–5966 (2012).
10. Madan, A., Cebrian, M., Moturu, S., Farrahi, K. & Pentland, S. *IEEE Pervasive Computing* http://dx.doi.org/10.1109/MPRV.2011.79 (2011).
11. Asur, S. & Huberman, B. A. *Proc. 2010 IEEE/WIC/ACM Int. Conf. Web Intelligence and Intelligent Agent Technology* Vol. 1, 492–499 (2010).
12. Gayo-Avello, D., Metaxas, P. T. & Mustafaraj, E. *Proc. Fifth Int. AAAI Conf. on Weblogs and Social Media* 490–493 (2011).
13. Bollen, J. & Mao, H. *IEEE Computer* **44**(10), 91–94 (2011).
14. Huberman, B. A. *Nature* **482,** 308 (2012).
15. Backstrom, L., Boldi, P., Rosa, M., Ugander, J. & Vigna S. Preprint at http://arxiv.org/abs/1111.4570 (2011).
16. Corbyn, Z. *Nature* http://dx.doi.org/10.1038/nature.2012.11064 (2012).