

# Data sharing threatens privacy

When web provider AOL's research division published an analysis of search behaviour on the Internet last year<sup>1</sup>, it had what it thought was a bright idea: it would reach out to academics by making an anonymized version of the data freely available for download from its website. But within hours, it had to pull the site, after bloggers managed to infer many identities from the data and view the associated search histories.

AOL's mistake highlights a growing dilemma for the social sciences. The hottest growth area in the field is computational social science. This is often based on privileged access to electronic data sets such as e-mail records, mobile-phone call logs and web-search histories of millions of individuals.

Such studies are ushering in a revolution in the social sciences, specialists say. But there is a trade-off between the scientific interest in working with such data and concerns about privacy. "It's a huge issue," says David Lazer, a researcher at the John F. Kennedy School of Government at Harvard University.

Lazer is co-author of a social-network analysis based on the largest mobile-phone data set ever studied by academics — 18 weeks of details of who called whom, when and for how long, among 7 million users, representing 20% of the population of an unidentified European country, supplied by an unnamed operator<sup>2</sup>. A month after it appeared, Microsoft researchers published an even larger study<sup>3</sup>, this time of instant messages, featuring 30 billion conversations among 240 million people worldwide.

Until now, social science has struggled to obtain tools that do more than scratch the surface of some of its questions. These range

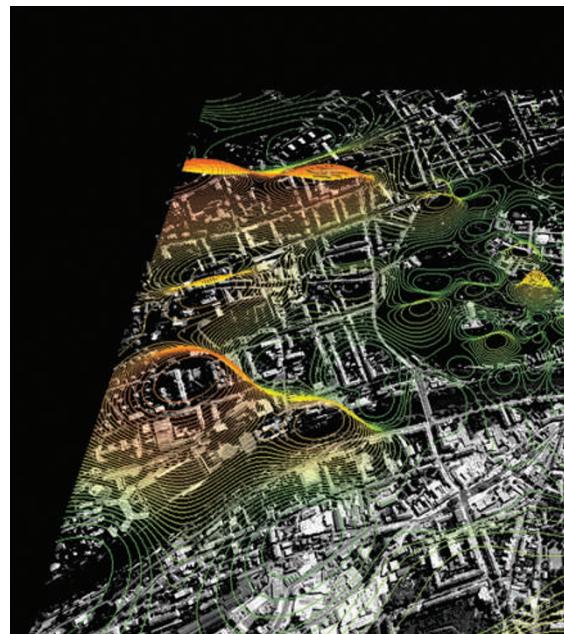
from identifying the driving forces behind violence, to the factors influencing how ideas, attitudes and prejudices spread through human populations. The available tools have largely remained in a time warp, consisting of analyses of national censuses, small-scale surveys, or lone researchers with a notebook observing interactions within small groups.

Being able to automatically and remotely obtain such massive amounts of continuous data opens up unprecedented opportunities for social scientists to study organizations and entire communities or populations, says Marshall Van Alstyne, a researcher at Boston University. He is conducting research on the dynamics and productivity of organizations by analysing network patterns of e-mails among volunteers.

"There is enormous potential here for lines of research that shed new light on basic social-science questions," says Jon Kleinberg, a specialist in network analysis at Cornell University in Ithaca, New York.

But the privacy issue looms large. Repetition of the AOL gaffe by other researchers might create a damaging public backlash, warns Myron Gutmann, director of the Inter-university Consortium for Political and Social Research, based in Ann Arbor, Michigan. The consortium acts as a clearinghouse for secure access to conventional sensitive social-science data sets, such as the raw data of the US national census.

"So far, researchers have been careful, and successful at avoiding serious problems," says Kleinberg. "But as the number of these types of study increases, the community is clearly going to need to engage in deeper discussions about the right way to safeguard privacy in working with these kinds of data."



Lazer says the mobile-phone study was possible only after taking many precautions, including anonymizing data before researchers gained access to them, confidentiality agreements between the phone company and Harvard University, and approval and stipulations on access controls by the university's internal review board.

But those ad hoc arrangements can go only so far, says Lazer, and a larger institutional framework is needed to set best practices, especially for safe sharing of data among academics. The social scientists pioneering such research often have computer-science backgrounds. As the software tools developed go mainstream, they get taken up by academics with less expertise in protecting data from abuse.

Although AOL had anonymized the data it released — 20 million web queries from around

# Doctors not to blame over HIV infection by tainted blood

A. WYLD/CANADIAN PRESS/AP



Acquitted: Roger Perrault leaves court.

A Canadian court has acquitted four doctors and a US blood products company of criminal negligence in the case of four haemophiliacs who were infected with HIV after receiving transfusions of tainted blood in the 1980s.

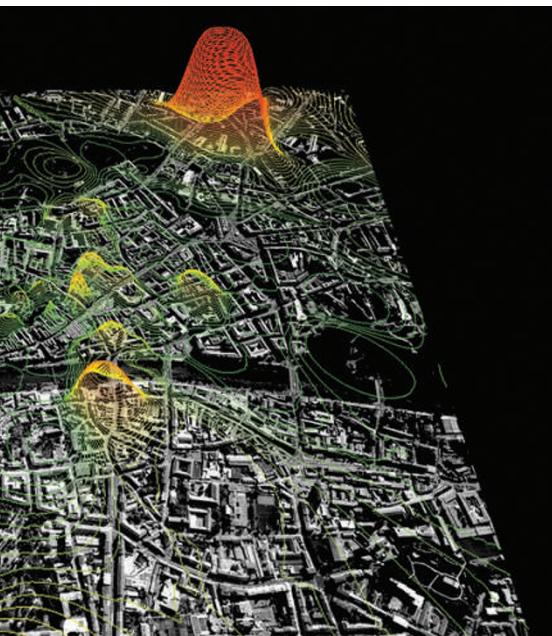
After a five-year police investigation and a lengthy trial that involved more than 100 witnesses and 1,000 exhibits, Judge Mary Lou Benotto of the Ontario Superior Court of

Justice in Toronto effectively said that the doctors and the company were not only acquitted but fully exonerated.

"The allegations of criminal conduct on the part of these men and this corporation were not only unsupported by the evidence, they were disproved," Benotto wrote in her 1 October decision. "The events here were tragic. However, to assign blame where none exists is to compound the tragedy."

The case is the latest of numerous global investigations into the circumstances in which thousands of patients were given infected blood even after it became known, in autumn 1984, that heat treatment killed HIV in blood products. More than 1,000 Canadians, about 700 of them with haemophilia, were infected by HIV from transfusions, almost all of them before mid-1985.

The Canadian court case



MIT SENSEABLE CITY LAB.

The data behind analyses, such as this map of mobile-phone use in Rome, present privacy issues.

650,000 users over 3 months — the company overlooked the fact that people's search queries often include vanity searches on their own name. This fact, combined with context from other search queries, often allowed complete search records to be narrowed down to individuals.

Using algorithms similar to those for cracking cryptography keys, Kleinberg also showed in an May paper<sup>4</sup> that more sophisticated attacks can be used to identify people in completely anonymized data. In the model, individuals in the network, and simulated false users, were able first to identify themselves in the network, and subsequently obtain the supposedly confidential data of all those connected to them. "It's hard to guarantee in practice that

any data set, however well anonymized, is truly safe from privacy breaches," says Kleinberg.

This work reinforces the need for a systematic, institutional approach to improving the privacy rights of those whose data are used, says Van Alstyne. That echoes the conclusions of a May study by the US National Academies<sup>5</sup>, which said that safeguarding privacy cannot safely be left to individual researchers. It stated that: "Institutional solutions involve establishing tiers of risk and access, and developing data-sharing protocols that match the level of access to the risks and benefits of the planned research."

But Gutmann and other social scientists also stress that the risks should be kept in perspective. Scientists must meet strict rules on any research on human subjects. In contrast, private firms are largely free from such constraints, and already have wide latitude to snoop on, and data mine, their employees' work habits.

Specialized firms, not to mention governments, are also building vast cross-referenced databases of information on every aspect of individuals' lives, including their web search behaviour, interests and personal preferences. An excessive focus on research would be "fixing the leaky faucet when the bath-tub's overflowing", says Van Alstyne. "Businesses seem more prone to misuse private data than scientists of any stripe." ■

**Declan Butler**

- Pass, G., Chowdhury, A. & Torgeson, C. *Proc. 1st Int. Conf. Scalable Information Sys.* doi:10.1145/1146847.1146848 (ACM Press, New York, 2006).
- Onnela, J.-P. et al. *Proc. Natl Acad. Sci. USA* **104**, 7332–7336 (2007).
- Leskovec, J. & Horvitz, E. *Microsoft Res. Tech. Rep. MSR-TR-2006-186* (2007).
- Bachstrom, L., Dwork, C. & Kleinberg, J. *Proc. 16th Int. World Wide Web Conf.* doi:10.1145/1242572.124259 (2007).
- Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data* www.nap.edu/catalog/11865.html (Natl Acad. Sci., Washington DC, 2007).

**See Editorial, page 637.**

addressed the infection of 4 people in British Columbia and Alberta in 1986 and 1987. They received an HIV-infected blood-clotting product made by Armour Pharmaceutical, then a maker of blood products based in New Jersey.

The now-elderly doctors charged were Roger Perrault, then the national director of the Canadian Red Cross blood transfusion service; John Furesz, then director of Health Canada's Bureau of Biologics (BDB); Wark Boucher, who headed the BDB's blood

product division; and Michael Rodell, then a vice-president of Armour Pharmaceutical.

The Canadian Hemophilia Society said it was "surprised and disappointed" by the outcome. "This verdict sends the wrong message to those responsible for the health of the public," Pam Wilton, its president, said in statement.

But David Scott, an Ottawa lawyer who defended Furesz, said: "These charges were undoubtedly politically motivated. If those in charge of the criminal process

succumb to political pressure to lay criminal charges in inappropriate cases, they will unwittingly undermine the public confidence in the administration of justice."

Lawyers for the Crown didn't say whether they plan to appeal within the 30-day time frame allowed. "I cannot imagine them doing so," says Edward Greenspan, the Toronto lawyer who represented Perrault. "But then again I couldn't imagine them spending 17 months in a court of law proving nothing." ■

Meredith Wadman

## ZOO NEWS

### Redhead goes for blondes

Sibu the orang-utan is driving staff at Apenheul Primate Park in the Netherlands crazy by refusing potential mates. Staff report that the 31-year-old male seems to be interested only in tattooed blonde women, having taken quite a shine to one of his previous keepers.

## SCORECARD



### Peaceful sleep

Partners of heavy snorers can breathe a sigh of relief — a new computerized pillow senses the sound of snoring and inflates or deflates to alter the angle of the slumberer's head to allow easier breathing.



### Concussion

Also new to the technology market is a computerized helmet for American football, which analyses impacts and creates a profile for each player, tallying up the inevitable carnage and warning of impending injury.

## SHOWBIZ NEWS

### Star (Trek) in the sky

George Takei (pictured), perhaps better known as USS Enterprise helmsman Mr Sulu, is the latest sci-fi icon to have an asteroid named in his honour by the International Astronomical Union. The rock, now called 7307 Takei, joins asteroids named after Star Trek creator Gene Roddenberry and Nichelle Nichols, aka Lieutenant Uhura.

Sources: Associated Press, Reuters, Scientific American



E. FIORE/REX FEATURES

## SIDELINES