The Economist

Data privacy

We’ll see you, anon

Can big databases be kept both anonymous and useful?

Aug 15th 2015 | From the print edition

FREQUENT visitors to the Hustler Club, a gentlemen’s entertainment venue in New York, could not have known that they would become part of a debate about anonymity in the era of “big data”. But when, for sport, a data scientist called Anthony Tockar mined a database of taxi-ride details to see what fell out of it, it became clear that, even though the data concerned included no direct identification of the customer, there were some intriguingly clustered drop-off points at private addresses for journeys that began at the club. Stir voter-registration records into the mix to identify who lives at those addresses (which Mr Tockar did not do) and you might end up creating some rather unhappy marriages.

The anonymisation of a data record typically means the removal from it of personally identifiable information. Names, obviously. But also phone numbers, addresses and various intimate details like dates of birth. Such a record is then deemed safe for release to researchers, and even to the public, to make of it what they will. Many people volunteer information, for example to medical trials, on the understanding that this will happen.

But the ability to compare databases threatens to make a mockery of such protections. Participants in genomics projects, promised anonymity in exchange for their DNA, have been identified by simple comparison with electoral rolls and other publicly available information. The health records of a governor of

Massachusetts were plucked from a database, again supposedly anonymous, of state-employee hospital visits using the same trick. Reporters sifting through a public database of web searches were able to correlate them in order to track down one, rather embarrassed, woman who had been idly searching for single men. And so on.

Each of these headline-generating stories creates a demand for more controls. But that, in turn, deals a blow to the idea of open data—that the electronic "data exhaust" people exhale more or less every time they do anything in the modern world is actually useful stuff which, were it freely available for analysis, might make that world a better place.

Of cake, and eating it

Modern cars, for example, record in their computers much about how, when and where the vehicle has been used. Comparing the records of many vehicles, says Viktor Mayer-Schönberger of the Oxford Internet Institute, could provide a solid basis for, say, spotting dangerous stretches of road. Similarly, an opening of health records, particularly in a country like Britain, which has a national health service, and cross-fertilising them with other personal data, might help reveal the multifarious causes of diseases like Alzheimer's.

This is a true dilemma. People want both perfect privacy and all the benefits of openness. But they cannot have both. The stripping of a few details as the only means of assuring anonymity, in a world choked with data exhaust, cannot work. Poorly anonymised data are only part of the problem. What may be worse is that there is no standard for anonymisation. Every American state, for example, has its own prescription for what constitutes an adequate standard.

Worse still, devising a comprehensive standard may be impossible. Paul Ohm of Georgetown University, in Washington, DC, thinks that this is partly because the availability of new data constantly shifts the goalposts. "If we could pick an industry standard today, it would be obsolete in short order," he says. Some data, such as those about medical conditions, are more sensitive than others. Some data sets provide great precision in time or place, others merely a year or a postcode. Each set presents its own dangers and requirements.

Fortunately, there are a few easy fixes. Thanks in part to the headlines, many now agree that public release of anonymised data is a bad move. Data could instead be

released piecemeal, or kept in-house and accessible by researchers through a question-and-answer mechanism. Or some users could be granted access to raw data, but only in strictly controlled conditions.

All these approaches, though, are anathema to the open-data movement, because they limit the scope of studies. "If we're making it so hard to share that only a few have access," says Tim Althoff, a data scientist at Stanford University, "that has profound implications for science, for people being able to replicate and advance your work."

Purely legal approaches might mitigate that. Data might come with what have been called "downstream contractual obligations", outlining what can be done with a given data set and holding any onward recipients to the same standards. One perhaps draconian idea, suggested by Daniel Barth-Jones, an epidemiologist at Columbia University, in New York, is to make it illegal even to attempt re-identification.

While some level of anonymisation will remain part of any resolution of the dilemma, mathematics may change the overall equation. One approach that would shift the balance to the good is homomorphic encryption, whereby queries on an encrypted data set are themselves encrypted. The result of any inquiry is the same as the one that would have been obtained using a standard query on the unencrypted database, but the questioner never sets eyes on the data. Or there is secure multiparty computation, in which a database is divided among several repositories. Queries are thus divvied up so that no one need have access to the whole database.

These approaches are, on paper, absolute in their protections. But putting them to work on messy, real-world data is proving tricky. Another set of techniques called differential privacy seems further ahead. The idea behind it is to ensure results derived from a database would look the same whether a given individual's data were in it or not. It works by adding a bit of noise to the data in a way that does not similarly fuzz out the statistical results.

Hot fuzz

America's Census Bureau has used differential privacy in the past for gathering commuters' data. Google is employing it at the moment as part of a project in which a browser plug-in gathers lots of data about a user's software, all the while

guaranteeing anonymity. Cynthia Dwork, a differential-privacy pioneer at Microsoft Research, suggests a more high-profile proving ground would be data sets—such as some of those involving automobile data or genomes—that have remained locked up because of privacy concerns.

For now, differential privacy’s difficult mathematical underpinnings make it tricky to implement more broadly. That needs to change, according to Salil Vadhan, of the Centre for Research on Computation and Society at Harvard. “The ball is in our court to not just to write papers, but to produce general-purpose tools,” he says.

Public education is also needed. Data science could well lead to safer roads and long-sought cures, but people have to understand the trade-offs. In July researchers at Britain’s Office for National Statistics (ONS), whose releases of data underpin billions of pounds of public spending, began to consult members of the public about their comfort with different types of data disclosure. There is always some risk to anonymity, says Jane Naylor of the ONS. But “there’s also a risk of not making the best use of data.”

From the print edition: Science and technology