PRIVACY

Credit card study blows holes in anonymity

Attack suggests need for new data safeguards

By **John Bohannon**

For social scientists, the age of big data carries big promises: a chance to mine demographic, financial, medical, and other vast data sets in fine detail to learn how we lead our lives. For privacy advocates, however, the prospect is alarming. They worry that the people represented in such data may not stay anonymous for long. A study of credit card data in this week's issue of *Science* (p. 536) bears out those fears, showing that it takes only a tiny amount of personal information to de-anonymize people.

The result, coming on top of earlier demonstrations that personal identities are easy to pry from anonymized data sets, indicates that such troves need new safeguards. "In light of the results, data custodians should carefully limit access to data," says Arvind Narayanan, a computer scientist at Princeton University who was not involved with the study. Or as the study's lead author, Yves-Alexandre de Montjoye, an applied mathematician at the Massachusetts Institute of Technology (MIT) in Cambridge, puts it: When it comes to sensitive personal information, "the open sharing of raw data sets is not the future."

De Montjoye's team analyzed 3 months of credit card transactions, chronicling the spending of 1.1 million people in 10,000 shops in a single country. (The team is tightlipped about the data's source—a "major bank," de Montjoye says—and it has not disclosed which country.) The bank stripped away names, credit card numbers, shop addresses, and even the exact times of the transactions. All that remained were the metadata: amounts spent, shop type—restaurant, gym, or grocery store, for example—and a code representing each person.

But because each individual's spending pattern is unique, the data have a very high "unicity." That makes them ripe for what de Montjoye calls a "correlation attack." To reveal a person's identity, you just need to correlate the metadata with information about the person from an outside source.

One correlation attack became famous last year when the New York City Taxi and Limousine Commission released a data set of the times, routes, and cab fares for 173 million rides. Passenger names were not included. But armed with time-stamped photos of celebrities getting in and out of taxis—there are websites devoted to celebrity spotting—bloggers, after deciphering taxi driver medallion numbers, easily figured out which celebrities paid which fares.

"The open sharing of raw data sets is not the future."

Yves-Alexandre de Montjoye, MIT

Stealing a page from the taxi data hack, de Montjoye's team simulated a correlation attack on the credit card metadata. They armed their computers with a collection of random observations about each individual in the data: information equivalent to a single time-stamped photo. (These clues were simulated, but people generate the real-world equivalent of this information day in and day out, for example through geolocated tweets or mobile phone apps that log location.) The computer used those clues to identify some of the anonymous spenders. The researchers then fed a different piece of outside information into the algorithm and tried again, and so on until every person was de-anonymized.

Just knowing an individual's location on four occasions was enough to fingerprint 90% of the spenders. And knowing the amount spent on those occasions—the equivalent of a few receipts from someone's trash—made it possible to de-anonymize nearly everyone and trace their entire transaction history with just three pieces of information per person. The findings echo the results of a 2013 *Scientific Reports* study in which de Montjoye and colleagues started with a trove of mobile phone metadata on subscribers' movements and showed that knowing a person's location on four occasions was enough to identify them.

One way to protect against correlation attacks is to blur the data by binning certain variables. For example, rather than revealing the exact day or price of a transaction, the public version of the data set might reveal only the week in which it occurred or a price range within which it fell. Binning did not thwart de Montjoye's correlation attack; instead, it only increased the amount of information needed to de-anonymize each person to the equivalent of a dozen receipts.

These studies needn't be the death knell for social science research using big data. "We need to bring the computation to the data, not the other way around," de Montjoye says. Big data with sensitive information could live "in the cloud," protected by gatekeeper software, he says. The gatekeeper would not allow access to individual records, thwarting correlation attacks, but would still let researchers ask statistical questions about the data.

The mathematics needed to run such a system, a set of standards and algorithms known as differential privacy, is one of the hottest topics in data science. "It works best when you have a large amount of data," says Cynthia Dwork, a computer scientist at Microsoft Research in Mountain View, California, who is one of the pioneers of the technique. She admits that it is a stark departure from the traditional academic practice of open data sharing, and many scientists are resistant.

But without such safeguards, rich databases could remain off limits. Take, for example, the data MIT has accumulated from its massive open online courses. It's an information trove that education researchers dream of having: a record of the entire arc of the learning process for millions of students, says Salil Vadhan, a computer scientist at Harvard University. But the data are under lock and key, partly out of fears of a prospective privacy breach. "If we can provide data for research without endangering privacy," Vadhan says, "it will do a lot of good." ■

PHOTO: ADAM GAULT/GETTY IMAGES