The New York Times

Bits
Business, Innovation, Technology, Society
PRIVACY

With a Few Bits of Data, Researchers Identify 'Anonymous' People

By Natasha Singer January 29, 2015 2:01 pm

Even when real names and other personal information are stripped from big data sets, it is often possible to use just a few pieces of the information to identify a specific person, according to a study to be published Friday in the journal Science.

In the study, titled "Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata," a group of data scientists analyzed credit card transactions made by 1.1 million people in 10,000 stores over a three-month period. The data set contained details including the date of each transaction, amount charged and name of the store.

Although the information had been "anonymized" by removing personal details like names and account numbers, the uniqueness of people's behavior made it easy to single them out.

In fact, knowing just four random pieces of information was enough to reidentify 90 percent of the shoppers as unique individuals and to uncover their records, researchers calculated. And that uniqueness of behavior — or "unicity," as the researchers termed it — combined with publicly available information, like Instagram or Twitter posts, could make it possible to reidentify people's records by name.

"The message is that we ought to rethink and reformulate the way we think about data protection," said Yves-Alexandre de Montjoye, a graduate student in computational privacy at the M.I.T. Media Lab who was the lead author of the study. "The old model of anonymity doesn't seem to be the right model when we are talking about large-scale metadata."

The analysis of large data sets containing details on people's behavior holds great potential to improve public health, city planning and education.

But the study calls into question the standard methods many companies, hospitals and government agencies currently use to anonymize their records. It may also give ammunition to some technologists and privacy advocates who have challenged the consumer-tracking processes used by advertising software and analytics companies to tailor ads to so-called anonymous users online.

This is hardly the first research effort to identify weaknesses in standard methods of de-identifying sensitive information about people.

In a study in 2008, two computer scientists, Arvind Narayanan and Vitaly Shmatikov, reported that they had been able to reidentify some Netflix users in a database of nameless customer records the company had made available for researchers competing to improve the company's recommendation engine.

In a study in 2013, Latanya Sweeney, a computer scientist at Harvard, demonstrated that researchers were able to reidentify patients by name in a supposedly anonymized hospitalization data set made publicly available by Washington State.

And last fall, a reporter at Gawker was able to reidentify Kourtney Kardashian, Ashlee Simpson and other celebrities in an "anonymized" database of taxi ride records made public by New York City's Taxi and Limousine Commission.

If companies or institutions are to continue to make these kinds of data sets widely available, they should quantitatively attest to the risks of reidentification, the researchers wrote in the study in Science.

"A data set's lack of names, home addresses, phone numbers or other obvious identifiers," they wrote, "does not make it anonymous nor safe to release to the public and to third parties."

A version of this article appears in print on 02/02/2015, on page B5 of the NewYork edition with the headline: With Little Data, Study Identifies the u2018Anonymousu2019.