May 22, 2019 | Unknown author Gina Kolata July 23 2019

Your Data Were ‘Anonymized’? These Scientists Can Still Identify You

Your medical records might be used for scientific research. But don’t worry, you’re told — personally identifying data were removed.

Information about you gathered by the Census Bureau might be made public. But don’t worry — it, too, has been “anonymized.”

On Tuesday, scientists showed that all this information may not be as anonymous as promised. The investigators developed a method to re-identify individuals from just bits of what were supposed to be anonymous data.

In most of the world, anonymous data are not considered personal data — the information can be shared and sold without violating privacy laws. Market researchers are willing to pay brokers for a huge array of data, from dating preferences to political leanings, household purchases to streaming favorites.

Even anonymized data sets often include scores of so-called attributes — characteristics about an individual or household. Anonymized consumer data sold by Experian, the credit bureau, to Alteryx, a marketing firm, included 120 million Americans and 248 attributes per household.

Scientists at Imperial College London and Université Catholique de Louvain, in Belgium, reported in the journal Nature Communications that they had devised a computer algorithm that can identify 99.98 percent of Americans from almost any available data set with as few as 15 attributes, such as gender, ZIP code or marital status.

Even more surprising, the scientists posted their software code online for anyone to use. That decision was difficult, said Yves-Alexandre de Montjoye, a computer scientist at Imperial College London and lead author of the new paper.

Ordinarily, when scientists discover a security flaw, they alert the vendor or government agency hosting the data. But there are mountains of anonymized data circulating worldwide, all of it at risk, Dr. de Montjoye said.

So the choice was whether to keep mum, he said, or to publish the method so that data vendors can secure future data sets and prevent individuals from being re-identified.

"This is very hard," Dr. de Montjoye said. "You have to cross your fingers that you did it properly, because once it is out there, you are never going to get it back."

Some experts agreed with the tactic. "It's always a dilemma," said Yaniv Erlich, chief scientific officer at MyHeritage, a consumer genealogy service, and a well-known data privacy researcher.

"Should we publish or not? The consensus so far is to disclose. That is how you advance the field: Publish the code, publish the finding."

This not the first time that anonymized data has been shown to be not so anonymous after all. In 2016, individuals were identified from the web-browsing histories of three million Germans, data that had been purchased from a vendor. Geneticists have shown that individuals can be identified in supposedly anonymous DNA databases.

The usual ways of protecting privacy include "de-identifying" individuals by removing attributes or substituting fake values, or by releasing only fractions of an anonymized data set.

But the gathering evidence shows that all of the methods are inadequate, said Dr. de Montjoye. "We need to move beyond de-identification," he said. "Anonymity is not a property of a data set, but is a property of how you use it."

The balance is tricky: Information that becomes completely anonymous also becomes less useful, particularly to scientists trying to reproduce the results of other studies. But every small bit that is retained in a database makes identification of individuals more possible.

"Very quickly, with a few bits of information, everyone is unique," said Dr. Erlich.

One possible solution is to control access. Those who want to use sensitive data — medical records, for example — would have to access them in a secure room. The data can be used but not copied, and whatever is done with the information must be recorded.

Researchers also can get to the information remotely, but "there are very strict requirements for the room where the access point is installed," said Kamel Gadouche, chief executive of a research data center in France, C.A.S.D., which relies on these methods.

The center holds information on 66 million individuals, including tax and medical data, provided by governments and universities. "We are not restricting access," Mr. Gadouche said. "We are controlling access."

But there is a drawback to restricted access. If a scientist submits a research paper to a journal, for example, others might want to confirm the results by using the data — a challenge if the data were not freely available.

Other ideas include something called "secure multiparty computation."

"It's a cryptographic trick," Dr. Erlich said. "Suppose you want to compute the average salary for both or us. I don't want to tell you my salary and you don't want to tell me yours."

So, he said, encrypted information is exchanged that is unscrambled by a computer.

"In theory, it works great," said Dr. Erlich. But for scientific research, the method has limits. If the end result seems wrong, "you cannot debug it, because everything is so secure you can't see the raw data."

The records gathered on all of us will never be completely private, he added: "You cannot reduce risk to zero."

Dr. de Montjoye worries that people do not yet appreciate the problem.

Two years ago, when he moved from Boston to London, he had to register with a general practitioner. The doctor's office gave him a form to sign saying that his medical data would be shared with other hospitals he might go to, and with a system that might distribute his

information to universities, private companies and other government departments.

The form added that the although the data are anonymized, “there are those who believe a person can be identified through this information.”

“That was really scary,” Dr. de Montjoye said. “We are at a point where we know a risk exists and count on people saying they don’t care about privacy. It’s insane.”