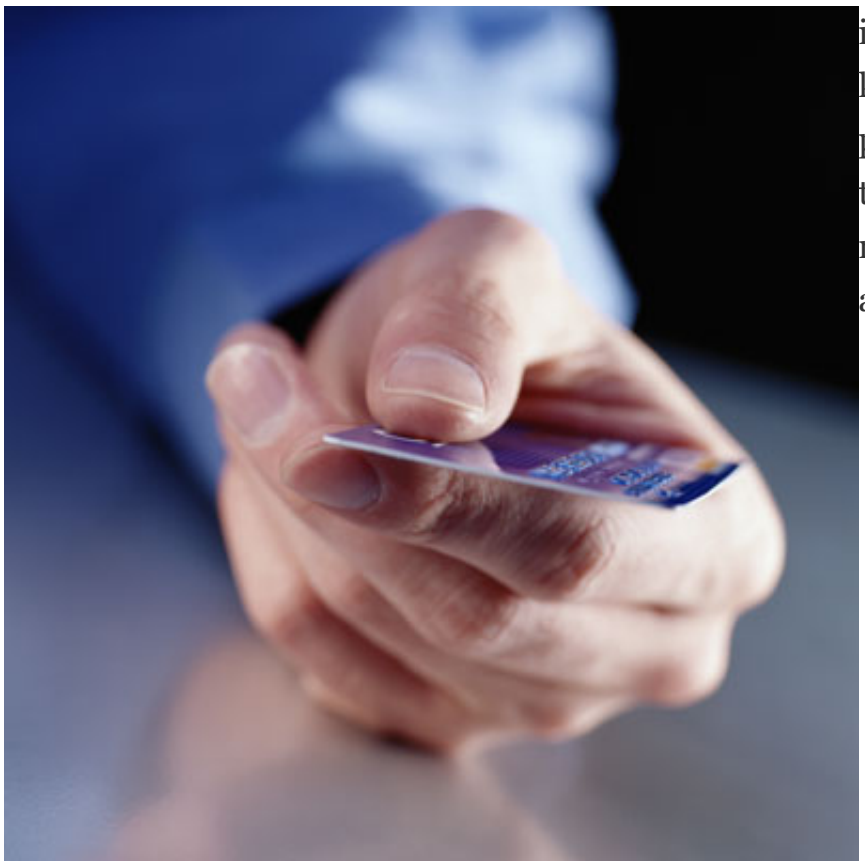
Observations

Shopping Habits Reveal Personal Details in "Anonymized" Data

Details about where and when you use your credit card could help reveal your identity to data thieves—even if they don't know your name, address and other personal information.

By Larry Greenemeier on January 29, 2015

Details about where and when you use your credit card could help reveal your



identity to data thieves—even if they don't know your name, address and other personal information. That's according to the latest study to poke holes in the notion that anonymous data records are an effective way to protect privacy.

Credit/Source: PhotoDisc/ Getty Images

Businesses, medical facilities and government agencies often claim that sanitizing the data they store can maintain customer, patient or constituent confidentiality in the event that information is lost or stolen. Using an approach sometimes referred to as "de-identification," they cleanse the data fields that could reveal one's identity—IP addresses, usernames and Social Security numbers, for example. They do this any number of ways, including encryption and the exclusion of certain data from records when they are shared.

Often those storing sensitive personal data rely on anonymization as a way of avoiding legal obligations to inform the public when there is a database leak or breach. However, a study appearing in this week's *Science*—part of the journal's "Privacy in a Data-Driven World" special issue—claims that researchers at the Massachusetts Institute of Technology (M.I.T.), Aarhus University (Denmark) and Rutgers were able to "re-identify" individuals in an anonymized

database of credit card records.

The researchers analyzed three months of “simply anonymized” credit card records—without names, account numbers or other obvious identifiers—for 1.1 million people shopping at 10,000 stores spread throughout an unidentified country. This information was provided by a bank operating in that country. Included in the dataset they analyzed were the names and locations of the shops where the purchases took place, the days on which they took place, and the purchase amounts.

Using both the credit card and transaction information the researchers identified 90 percent of the individuals in the data set. When they added the exact prices of transactions to the mix, they increased their ability to re-identify anonymous records by 22 percent. The researchers found that they could identify people even if they knew only their general location or a time frame during which the people shopped.

One of the more significant conclusions of the study “is that we have to think harder and reform how we approach data protection and go beyond anonymity, which is very difficult to achieve given the trail of information we all leave digitally,” says Yves-Alexandre de Montjoye, a senior PhD student in computational privacy at M.I.T.’s Media Lab. “You want there to be uncertainty over someone’s identity even when you have lots of data about that person.”

Women were easier to re-identify than men, as were those who earned a lot of money compared to those with low incomes, according to the study. De Montjoye and his colleagues, including Media Lab director and data scientist Alex "Sandy" Pentland, don’t delve into the reasons for this disparity. De Montjoye points out, however, that this could be an indication that different people might be easier to identify based on their buying behavior, such as the number of shops they visit and amount of time spent there.

More troubling than the results of this study is the fact that so many laws in the U.S. encourage anonymization as a means of privacy protection. The Health Insurance Portability and

Accountability Act of 1996 (HIPAA), for example, treats 18 different categories of data as protected health information—including name, Social Security number and birth date—that must be "de-identified." Yet HIPPAA does not require the same for information about patient visits—such as the year of the visit, patient age or diagnosis—an oversight that undermines patient anonymity, according to Paul Ohm, an associate professor and associate dean for academic affairs at the University of Colorado Law School, in a 2009 article for the *UCLA Law Review*.

Anonymization's flaws are particularly relevant now as the U.S. Congress maps out federal data breach notification legislation in response to President Obama's call for a Personal Data Notification and Protection Act. The law would create a national standard to replace 47 different state laws that currently use different standards to determine when businesses have to publically report lost or stolen customer data.

ABOUT THE AUTHOR(S)

Larry Greenemeier

Larry Greenemeier is the associate editor of technology for *Scientific American*, covering a variety of tech-related topics, including biotech, computers, military tech, nanotech and robots.

Credit: Nick Higgins

Recent Articles

How Hackers Take Down Web Sites [Video]

Is Artificial Intelligence Being Oversold?

What Could Criminals Do with 5.6 Million Fingerprint Files?