SEPTEMBER 11, 2009 | BY SETH SCHOEN

What Information is "Personally Identifiable"?

Mr. X lives in ZIP code 02138 and was born July 31, 1945.

These facts about him were included in an anonymized medical record released to the public. Sounds like Mr. X is pretty anonymous, right?

Not if you're Latanya Sweeney, a Carnegie Mellon University computer science professor who showed in 1997 that this information was enough to pin down Mr. X's more familiar identity -- William Weld, the governor of Massachusetts throughout the 1990s.

Gender, ZIP code, and birth date *feel* anonymous, but Prof. Sweeney was able to identify Governor Weld through them for two reasons. First, each of these facts about an individual (or other kinds of facts we might not usually think of as identifying) independently narrows down the population, so much so that the combination of (gender, ZIP code, birthdate) was unique for about 87% of the U.S. population. If you live in the United States, there's an 87% chance that you don't share all three of these attributes with any other U.S. resident. Second, there may be particular data sources available (Sweeney used a Massachusetts voter registration database) that let people do searches to bootstrap what they know about someone in order to learn more -- including traditional identifiers like name and address. In a very concrete sense, "anonymized" or "merely demographic" information about people may be neither. (And a web site that asks "anonymous" users for seemingly trivial information about themselves may be able to use that information to make a unique profile for an individual, or even look up that individual in other databases.)

Many contemporary privacy rules and debates center on the notion of "personally identifiable information" (PII). The PII concept is used by several legal regimes and many organizations' privacy policies; generally, information that identifies a particular person is considered much more sensitive than information that does not. For instance,

- Federal telecommunications privacy laws use "individually identifiable information" (about a subscriber) as a basis for the category of protected information called Customer Proprietary Network Information (CPNI);
- Federal health privacy regulations use "individually identifiable health information" (about a patient) as a basis for the category called Protected Health Information (PHI);
- Federal financial privacy laws, the EU Data Protection Directive, and state privacy laws all employ similar terms and concepts;

and, in each case, facts deemed "personally identifiable" or "individually identifiable" may receive dramatically higher protections under these laws and regulations.

But research by Prof. Sweeney and other experts has demonstrated that surprisingly many facts, including those that seem quite innocuous, neutral, or "common", could potentially identify an individual. Privacy law, mainly clinging to a traditional intuitive notion of identifiability, has largely not kept up with the technical reality.

A recent paper by Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization", provides a thorough introduction and a useful perspective on this issue. Prof. Ohm's paper is important reading for anyone interested in personal privacy, because it shows how deanonymization results achieved by researchers like Latanya Sweeney and Arvind Narayanan seriously undermine traditional privacy assumptions. In particular, the binary distinction between "personally-identifiable information" and "non-personally-identifiable information" is increasingly difficult to sustain. Our intuition that certain information is "anonymous" is often wrong. Given the proper circumstances and insight, almost any kind of information might tend to identify an individual; information about people is more identifying than has been assumed, and in the long run the whole

enterprise of classifying facts as "PII" or "not PII" is questionable.

Statistical inference and clever use of databases has resulted in impressive examples of deanonymization of supposedly anonymous data, the kinds of data that most organizations have not regarded as PII. Apart from combinations of demographic data, some of the sorts of things that may well uniquely identify you include your search terms; your purchase habits; your preferences or opinions about music, books, or movies; and even the *structure* of your social networks -- in a purely abstract sense, even when shorn of the identities of your friends and contacts. Deanonymization is effective, and it's dramatically easier than our intuitions suggest. Given the number of variables that potentially distinguish us, we are much more different from each other than we expect, and there are more sources of data than we realize that may be used to narrow down exactly who a particular record refers to.

Many of these papers were meant as proofs of concept: they show that people can *potentially* be re-identified by these kinds of data, not that everyone will be. Not *everyone*'s medical records were as easy to put a name to as Governor Weld's. And Narayanan and Shmatikov's research definitively identified only two Netflix users from their movie ratings -- not *every* user whose ratings were published by Netflix. Still, many of these research results deliberately do not use all the data available about individuals because their goal is to show the effectiveness of mathematical techniques, not to violate individuals' privacy. Real-world attacks will use many more kinds of available information simultaneously to narrow in on people's identities. As Bruce Schneier has observed, such attacks only get better over time; they never get worse.

Ohm argues that it's more appropriate to think of identifiability as a continuum. The notion of "anonymized" or "sanitized" data is then problematic; researchers habitually share, or even publish, data sets which assign code numbers to individuals. There have already been conspicuous problems with this practice, like when AOL published "anonymized" search logs which turned out to identify some individuals from the content of their search terms alone.

We hope "Broken Promises of Privacy" encourages people who work with personal data to think more critically about their retention and sharing practices and the effectiveness of the anonymization or pseudonymization techniques they're using. We also hope it finds a broad audience and helps start a wider discussion among researchers, technologists, and lawyers about what "privacy protection" should mean in the era of deanonymization.

Privacy Anonymity

Global Chokepoints
HTTPS Everywhere
Manila Principles
Medical Privacy Project
Open Wireless Movement
Patent Busting
Privacy Badger
Student Activism
Surveillance Self-Defense
Takedown Hall of Shame
Teaching Copyright
Transparency Project
Trolling Effects
Ways To Help

MORE DEEPLINKS POSTS LIKE THIS

JULY 2007
Ask.com Takes the Lead on Log Retention; Microsoft and Yahoo! Follow

NOVEMBER 2011
EFF Joins Advocacy Organizations in Criticizing Secure Communities

MARCH 2009
Last.fm and the Diabolical Power of Data Mining

OCTOBER 2010
New FOIA Documents Reveal DHS Social Media Monitoring During Obama Inauguration

MAY 2010
Facebook Violates Privacy Promises, Leaks User Info to Advertisers

RECENT DEEPLINKS POSTS

NOV 20, 2015
EFF Joins Broad Coalition of Groups to Protest the TPP in Washington D.C.

NOV 20, 2015
Unintended Consequences, European-Style: How the New EU Data Protection Regulation will be Misused to Censor Speech

NOV 20, 2015
New Report Rates Peruvian ISPs: Who Defends Your Data?

NOV 19, 2015
Nuevo reporte muestra qué ISPs peruanas resguardan la privacidad de usuarios

NOV 19, 2015
YouTube Backs Its Users With New Fair Use Protection Program

DEEPLINKS TOPICS

Fair Use and Intellectual Property: Defending the Balance
Free Speech
Innovation
International
DRM
E-Voting Rights
EFF Europe
Encrypting the Web
Export Controls
Patents
PATRIOT Act
Pen Trap
Policy Analysis
Printers