POLICY —

"Anonymized" data really isn't—and here's why not

Companies continue to store and sometimes release vast databases of " ...

NATE ANDERSON - 9/8/2009, 2:25 PM

The Massachusetts Group Insurance Commission had a bright idea back in the mid-1990s—it decided to release "anonymized" data on state employees that showed every single hospital visit. The goal was to help researchers, and the state spent time removing all obvious identifiers such as name, address, and Social Security number. But a graduate student in computer science saw a chance to make a point about the limits of anonymization.

Latanya Sweeney requested a copy of the data and went to work on her "reidentification" quest. It didn't prove difficult. Law professor Paul Ohm describes Sweeney's work:

> At the time GIC released the data, William Weld, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting identifiers. In response, then-graduate student Sweeney started hunting for the Governor’s hospital records in the GIC data. She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes. For twenty dollars, she purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter. By combining this data with the GIC records, Sweeney found Governor Weld with ease. Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code. In a theatrical flourish, Dr. Sweeney sent the Governor’s health records (which included diagnoses and prescriptions) to his office.

Boom! But it was only an early mile marker in Sweeney's career; in 2000, she showed that 87 percent of all Americans could be uniquely identified using only three bits of information: ZIP code, birthdate, and sex.

Such work by computer scientists over the last fifteen years has shown a serious flaw in the basic idea behind "personal information": almost all information can be "personal" when combined with enough other relevant bits of data.

That's the claim advanced by Ohm in his lengthy new paper on "the surprising failure of anonymization." As increasing amounts of information on all of us are collected and disseminated online, scrubbing data just isn't enough to keep our individual "databases of ruin" out of the hands of the police, political enemies, nosy neighbors, friends, and spies.

If that doesn't sound scary, just think about your own secrets, large and small—those films you watched, those items you searched for, those pills you took, those forum posts you made. The power of reidentifiation brings them closer to public exposure every day. So, in a world where the PII concept is dying, how *should* we start thinking about data privacy and security?

Don't ruin me

> For almost every person on earth, there is at least one fact about them stored in a computer database that an adversary could use to blackmail, discriminate against, harass, or steal the identity of him or her. I mean more than mere embarrassment or inconvenience; I mean legally cognizable harm.

Examples of the anonymization failures aren't hard to find.

When AOL researchers released a massive dataset of search queries, they first "anonymized" the data by scrubbing user IDs and IP addresses. When Netflix made a huge database of movie recommendations available for study, it spent time doing the same thing. Despite scrubbing the obviously identifiable information from the data, computer scientists were able to identify individual users in both datasets. (The Netflix team then moved on to Twitter users.)

In AOL's case, the problem was that user IDs were scrubbed but were replaced with a number that uniquely identified each user. This seemed like a good idea at the time, since it allowed researchers using the data to see the complete list of a person's search queries, but it also created problems; those complete lists of search queries were so thorough that individuals could be tracked down simply based on what they had searched for. As Ohm notes, this illustrates a central reality of data collection: "data can either be useful or perfectly anonymous but never both."

The Netflix case illustrates another principle, which is that the data itself might seem anonymous, but when paired with other existing data, reidentification becomes possible. A pair of computer scientists famously proved this point by combing movie recommendations found on the Internet Movie Database with the Netflix data, and they learned that people could quite easily be picked from the Netflix data.

Such results are obviously problematic in a world where Google retains data for years, "anonymizing" it after a certain amount of time but showing reticence to fully delete it. "Reidentification science disrupts the privacy policy landscape by undermining the faith that we have placed in anonymization," Ohm writes. "This is no small faith, for technologists rely on it to justify sharing data indiscriminately and storing data perpetually, all while promising their users (and the world) that they are protecting privacy. Advances in reidentification expose these promises as too often illusory."

For users, the prospect of some secret leaking to the public grows as databases proliferate. Here is Ohm's nightmare scenario: "For almost every person on earth, there is at least one fact about them stored in a computer database that an adversary could use to blackmail, discriminate against, harass, or steal the identity of him or her. I mean more than mere embarrassment or inconvenience; I mean legally cognizable harm. Perhaps it is a fact about past conduct, health, or family shame. For almost every one of us, then, we can assume a hypothetical 'database of ruin,' the one containing this fact but until now splintered across dozens of databases on computers around the world, and thus disconnected from our identity. Reidentification has formed the database of ruin and given access to it to our worst enemies."

Because most data privacy laws focus on restricting personally identifiable information (PII), most data privacy laws need to be rethought. And there won't be any magic bullet; the measures that are taken will increase privacy or reduce the utility of data, but there will be no way to guarantee maximal usefulness and maximal privacy at the same time.

There are approaches that can reduce problems. Instead of releasing these huge anonymized databases, for instance, make them interactive, or have them report most results in the aggregate. (But such techniques sharply limit the usefulness of the data.)

Ohm's alternative is an admittedly messier system, one that can't be covered with simple blanket laws against recording Social Security numbers or releasing people's name and addresses. Such an approach has failed, and now looks like playing "Whac-A-Mole" with personal data. "The trouble is that PII is an ever-expanding category, writes Ohm. "Ten years ago, almost nobody would have categorized movie ratings and search queries as PII, and as a result, no law or regulation did either."

Expanding privacy rules each time some new reidentification technique emerges would be unworkable.

Instead, regulators will need to exercise more judgment, weighing harm against benefits, and the rules may turn out to be different for crucial systems like healthcare. At the same time, the US needs comprehensive legislation on data privacy to set a minimum threshold for all databases, since Netflix, AOL, and others have made clear that we have no real idea in advance which pieces of seemingly harmless data will turn out to identify us and our secrets.

READER COMMENTS 41 SHARE THIS STORY

NATE ANDERSON

Nate is the deputy editor at Ars Technica, where he oversees long-form feature content and writes about technology law and policy. He is the author of *The Internet Police: How Crime Went Online, and the Cops Followed*.

EMAIL nate@arstechnica.com // **TWITTER** @natexanderson

SITREP: F-16 replacement search a signal of F-35 fail?

WATCH

SITREP: F-16 replacement search a...

Sitrep: Boeing 707

The F-35's next tech upgrade

More videos

← PREVIOUS STORY

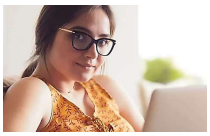
NEXT STORY →

Related Stories

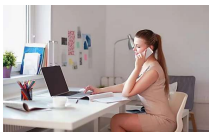
Sponsored Stories

Recommended by

Wall Street legend's Surprising...
Visionary Profit



Online Jobs from Kenya. Salaries May Surprise You
Work From Home | ...



Do you speak English? You Might Be Able T...
Online Jobs | Spons...

Parkland school shooting sentencing trial...
Americas

Finding a Job in the USA from Kenya Might be...
Job in the USA | Sear...

These liposuction patches are winning the...
Well-being-review.com

Today on Ars

NEWSLETTER SIGNUP

Join the Ars Orbital Transmission mailing list to get weekly updates delivered to your inbox.

SIGN ME UP →