De-identification, re-identification, and the definition of PII

IAPP Global Privacy Summit
8-9 March 2012
Omer Tene
omer@omertene.com
The Age of Big Data

The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT
Dangerous side effect of common drug combination discovered by data mining

BY KRISTA CONGER

A widely used combination of two common medications may cause unexpected increases in blood glucose levels, according to a study conducted at the Stanford University School of Medicine, Vanderbilt University and Harvard Medical School. Researchers were surprised at the finding because neither of the two drugs — one, an antidepressant marketed as Paxil, and the other, a cholesterol-lowering medication called Pravachol — has a similar effect alone.

The increase is more pronounced in people who are diabetic, and in whom the control of blood sugar levels is particularly important. It’s also apparent in pre-diabetic laboratory mice exposed to both drugs. The researchers speculate that between 500,000 and 1 million people in this country may be taking the two medications simultaneously.

The researchers’ study relied on an adverse-event reporting database maintained by the U.S.
Big Data: Benefits

Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National

States | Cities (Experimental)
First there was de-identification
Then there was re-identification.

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in...
Latanya Sweeney, 2000

HMO data
- ethnicity
- visit date
- diagnosis
- procedure
- medication
- total charge

Voter registration data
- ZIP
- birth date
- sex
- name
- address
- date reg.
- party
- affiliation
- last voted
Computer scientists have recently undermined our faith in the privacy-protecting power of anonymization, the name for techniques that protect the privacy of individuals in large databases by deleting information like names and social security numbers. These scientists have demonstrated that they can often “reidentify” or “deanonymize” individuals hidden in anonymized data with astonishing ease. By understanding this research, we realize we have made a mistake, labored beneath a fundamental misunderstanding, which has assured us much less privacy than we have assumed. This mistake pervades nearly every information privacy law, regulation, and debate, yet regulators and legal scholars have paid it scant attention. We must respond to the surprising failure of anonymization, and this Article provides the tools to do so.
Privacy in the Age of Big Data

A Time for Big Decisions

by Omer Tene & Jules Polonetsky

Omer Tene is an Associate Professor at the College of Management Haim Striks School of Law. Jules Polonetsky is a Co-Chair and Director of the Future of Privacy Forum.
Additional material

- Narayanan & Shmatikov, Robust De-anonymization of Large Sparse Datasets, available [here](#).
- El Amam et al, A Systematic Review of Re-Identification Attacks on Health Data, available [here](#).

[www.omertene.com](http://www.omertene.com)
[omer@omertene.com](mailto:omer@omertene.com)