Bruce Schneier

Crypto-Gram Newsletter

January 15, 2008

by Bruce Schneier
Founder and CTO
BT Counterpane
schneier@schneier.com
http://www.schneier.com
http://www.counterpane.com

A free monthly newsletter providing summaries, analyses, insights, and commentaries on security: computer and otherwise.

For back issues, or to subscribe, visit <http://www.schneier.com/crypto-gram.html>.

You can read this issue on the web at <http://www.schneier.com/crypto-gram-0801.html>. These same essays appear in the "Schneier on Security" blog: <http://www.schneier.com/blog>. An RSS feed is available.

In this issue:

Anonymity and the Netflix Dataset

Last year, Netflix published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using. The data was anonymized by removing personal details and replacing names with random numbers, to protect the privacy of the recommenders.

Arvind Narayanan and Vitaly Shmatikov, researchers at the University of Texas at Austin, de-anonymized some of the Netflix data by comparing rankings and timestamps with public information in the Internet Movie Database, or IMDb.

Their research illustrates some inherent security problems with anonymous data, but first it's important to explain what they did and did not do.

They did *not* reverse the anonymity of the entire Netflix dataset. What they did was reverse the anonymity of the Netflix dataset for those sampled users who also entered some movie rankings, under their own names, in the IMDb. (While IMDb's records are public, crawling the site to get them is against the IMDb's terms of service, so the researchers used a representative few to prove their algorithm.)

The point of the research was to demonstrate how little information is required to de-anonymize information in the Netflix dataset.

On one hand, isn't that sort of obvious? The risks of anonymous databases have been written about before, such as in this 2001 paper published in an IEEE journal. The researchers working with the anonymous Netflix data didn't painstakingly figure out people's identities -- as others did with the AOL search database last year -- they just compared it with an already identified subset of similar data: a standard data-mining technique.

But as opportunities for this kind of analysis pop up more frequently, lots of anonymous data could end up at risk.

Someone with access to an anonymous dataset of telephone records, for example, might partially de-anonymize it by correlating it with a catalog merchants' telephone order database. Or Amazon's online book reviews could be the key to partially de-anonymizing a public database of credit card purchases, or a larger database of anonymous book reviews.

Google, with its database of users' internet searches, could easily de-anonymize a public database of internet purchases, or zero in on searches of medical terms to de-anonymize a public health database. Merchants who maintain detailed customer and purchase information could use their data to partially de-anonymize any large search engine's data, if it were released in an anonymized form. A data broker holding databases of several companies might be able to de-anonymize most of the records in those databases.

What the University of Texas researchers demonstrate is that this process isn't hard, and doesn't require a lot of data. It turns out that if you eliminate the top 100 movies everyone watches, our movie-watching habits are all pretty individual. This would certainly hold true for our book reading habits, our internet shopping habits, our telephone habits and our web searching habits.

The obvious countermeasures for this are, sadly, inadequate. Netflix could have randomized its dataset by removing a subset of the data, changing the timestamps or adding deliberate errors into the unique ID numbers it used to replace the names. It turns out, though, that this only makes the problem slightly harder. Narayanan's and Shmatikov's de-anonymization algorithm is surprisingly robust, and works with partial data, data that has been perturbed, even data with errors in it.

With only eight movie ratings (of which two may be completely wrong), and dates that may be up to two weeks in error, they can uniquely identify 99 percent of the records in the dataset. After that, all they need is a little bit of identifiable data: from the IMDb, from your blog, from anywhere. The moral is that it takes only a small named database for someone to pry the anonymity off a much larger anonymous database.

Other research reaches the same conclusion. Using public anonymous data from the 1990 census, Latanya Sweeney found that 87 percent of the population in the United States, 216 million of 248 million, could likely be uniquely identified by their five-digit ZIP code, combined with their gender and date of birth. About half of the U.S. population is likely identifiable by gender, date of birth and the city, town or municipality in which the person resides. Expanding the geographic scope to an entire county reduces that to a still-significant 18 percent. "In general," the researchers wrote, "few characteristics are needed to uniquely identify a person."

Stanford University researchers reported similar results using 2000 census data. It turns out that date of birth, which (unlike birthday month and day alone) sorts people into thousands of different buckets, is incredibly valuable in disambiguating people.

This has profound implications for releasing anonymous data. On one hand, anonymous data is an enormous boon for researchers -- AOL did a good thing when it released its anonymous dataset for research purposes, and it's sad that the CTO resigned and an entire research team was fired after the public outcry. Large anonymous databases of medical data are enormously valuable to society: for large-scale pharmacology studies, long-term follow-up studies and so on. Even anonymous telephone data makes for fascinating research.

On the other hand, in the age of wholesale surveillance, where everyone collects data on us all the time, anonymization is very fragile and riskier than it initially seems.

Like everything else in security, anonymity systems shouldn't be fielded before being subjected to adversarial attacks. We all know that it's folly to implement a cryptographic system before it's rigorously attacked; why should we expect anonymity systems to be any different? And, like everything else in security, anonymity is a trade-off. There are benefits, and there are corresponding risks.

Narayanan and Shmatikov are currently working on developing algorithms and techniques that enable the secure release of anonymous datasets like Netflix's. That's a research result we can all benefit from.

http://www.cs.utexas.edu/~shmat/...
http://www.cs.utexas.edu/~shmat/netflix-faq.html
http://www.securityfocus.com/news/11497
http://arxivblog.com/?p=142

2001 IEEE paper:
http://people.cs.vt.edu/~naren/papers/ppp.pdf

De-anonymizing the AOL data: