Austin American-Statesman
statesman.com

PRINTTHIS

Click to Print

SAVE THIS | EMAIL THIS | Close

COMMENTARY

Seringhaus & Gerstein: Putting too much information online can erode individual privacy

Michael Seringhaus and Mark Gerstein, SPECIAL TO THE HARTFORD COURANT

Friday, June 05, 2009

When it comes to online privacy, we all appreciate the risk of publicizing juicy information such as incriminating photos or credit-card numbers. But few of us realize a subtler threat: In abundance, innocuous, everyday data can divulge sensitive information as well.

Some questions shouldn't be asked. Employers, for instance, generally are not allowed to discriminate based on marital status, sexual orientation and so on. But our growing digital footprint threatens our ability to dodge inappropriate inquiries. Through data mining, employers, insurers, advertisers and others can infer the answers to private questions without even asking.

They need two things: a heap of personal data, and the techniques to crunch it. Both are readily available.

People generate and share more information than ever. Besides consciously generated Web content such as blogs, Facebook profiles and YouTube videos, a steady stream of data is exchanged in the background. Companies track our searches, browsing and shopping behavior. Personal electronic devices can silently disclose our location while we post status updates and photos to the Web. All that seems innocent enough — and the more others do it, the safer we all feel. After all, what's one more Twitter update among millions?

But the crowd doesn't hide you. Instead, it's the key to coaxing value from your information. Data mining relies on the principle that certain information — though useless in isolation — can take on new meaning when viewed en masse, or combined with other data. Scientists already use this technique.

There are two main approaches. First, data integration involves combining different types of data to learn something new. Consider a photograph of a bicycle: Alone, it's an abstract representation. But tag the photo with your home location and a time stamp — and a public listing identifying the bike as stolen — and suddenly it becomes very meaningful.

A second approach is data aggregation. Gather enough of a certain type of data, and trends emerge. For instance, a cell phone's location can be determined by tracking its signal. By aggregating enough location data from a single cell phone, we derive an increasingly reliable map of one person's regular routes of travel. From this, we can estimate where the phone's owner is likely to be at a given time and perhaps even guess his home location, income and so forth.

Fusing these approaches is even more powerful: that is, combining and mining multiple data sets, each very large. Google did this last year, pairing aggregate Web search queries with location and timing data to predict which regions would next come down with the flu. It outperformed the Centers for Disease Control and Prevention.

Bringing data mining out of research labs and applying it to personal data is surprisingly straightforward.

Suppose a researcher wants to guess something about you — say, your political party affiliation. This becomes the target variable, the hidden question. Almost any other individual feature — smoking, hair color, preference in breakfast cereal — can be correlated with this target. To do this, the researcher needs to do some background crunching on a "gold standard": a small, representative and reliable group for which the relationship between the target and a given feature is known. Individually, most such correlations are weak, and don't matter much. Some may not matter at all.

But even if no one variable correlates well with the target, a group of them together may. So the researcher characterizes many different combinations of variables and correlates them to a target variable, such as political affiliation. The correlations are often impossible to spot by eye, but computers excel in sniffing them out.

And given several modest correlations, a stronger prediction can emerge. This is the power of data mining: combining weak correlations to generate a powerful statistical predictor of the target variable. These predictions are rarely 100 percent correct, but they don't need to be. And the more correlations are known, the more an answer-seeker can rely on clusters of shadow attributes, innocent bits of freely available data that correlate strongly with the target.

In a world of unlimited data where all statistical correlations have been computed, it becomes unnecessary to ask a forbidden question. The answer can be approximated quietly elsewhere, everywhere, through the web of interconnected data that describes each of us. Today a job candidate can remove her wedding band or otherwise demur when questioned; tomorrow her reply may be impossible to conceal — or withhold.

The larger our personal data trails grow, the more severe this threat becomes. So if we care about protecting individual privacy in a meaningful way, it is no longer enough just to forbid taboo questions. We must also prevent parties from harvesting and crunching data in a manner that circumvents the need to ask them at all.

Seringhaus is entering his third year at Yale Law School. Gerstein is a professor of biomedical informatics at Yale.

Vote for this story!

Find this article at:
http://www.statesman.com/opinion/content/editorial/stories/06/05/0605seringhaus_edit.html

Click to Print

Check the box to include the list of links referenced in the article.