How Privacy Vanishes Online

By STEVE LOHRMARCH 16, 2010

If a stranger came up to you on the street, would you give him your name,Social Security number and e-mail address?

Probably not.

Yet people often dole out all kinds of personal information on the Internet that allows such identifying data to be deduced. Services like Facebook, Twitter and Flickr are oceans of personal minutiae — birthday greetings sent and received, school and work gossip, photos of family vacations, and movies watched.

Computer scientists and policy experts say that such seemingly innocuous bits of self-revelation can increasingly be collected and reassembled by computers to help create a picture of a person's identity, sometimes down to the Social Security number.

"Technology has rendered the conventional definition of personally identifiable information obsolete," said Maneesha Mithal, associate director of the Federal Trade Commission's privacy division. "You can find out who an individual is without it."

In a class project at the Massachusetts Institute of Technology that received some attention last year, Carter Jernigan and Behram Mistree analyzed more than 4,000 Facebook profiles of students, including links to friends who said they were gay. The pair was able to predict, with 78 percent accuracy, whether a profile belonged to a gay male.

So far, this type of powerful data mining, which relies on sophisticated statistical correlations, is mostly in the realm of university researchers, not identity thieves and marketers.

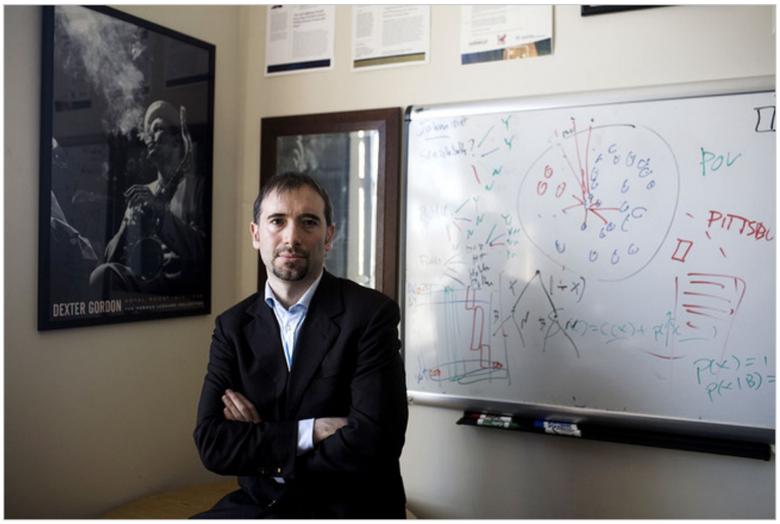
But the F.T.C. is worried that rules to protect privacy have not kept up with technology. The agency is convening on Wednesday the third of three workshops on the issue.

Its concerns are hardly far-fetched. Last fall, Netflix awarded $1 million to a team of statisticians and computer scientists who won a three-year contest to analyze the movie rental history of 500,000 subscribers and improve the predictive accuracy of Netflix's recommendation software by at least 10 percent.

On Friday, Netflix said that it was shelving plans for a second contest — bowing to privacy concerns raised by the F.T.C. and a private litigant. In 2008, a pair of researchers at the University of Texas showed that the customer data released for that first contest, despite being stripped of names and other direct identifying information, could often be “de-anonymized” by statistically analyzing an individual’s distinctive pattern of movie ratings and recommendations.

In social networks, people can increase their defenses against identification by adopting tight privacy controls on information in personal profiles. Yet an individual’s actions, researchers say, are rarely enough to protect privacy in the interconnected world of the Internet.

You may not disclose personal information, but your online friends and colleagues may do it for you, referring to your school or employer, gender, location and interests. Patterns of social communication, researchers say, are revealing.



Alessandro Acquisti mined Web data to successfully predict Social Security numbers.CreditRoss Mantle for The New York Times

“Personal privacy is no longer an individual thing,” said Harold Abelson, the computer science professor at M.I.T. “In today’s online world, what your mother told you is true, only more so: people really can judge you by your friends.”

Collected together, the pool of information about each individual can form a distinctive “social signature,” researchers say.

The power of computers to identify people from social patterns alone was demonstrated last year in a study by the same pair of researchers that cracked Netflix’s anonymous database: Vitaly Shmatikov, an associate professor of computer science at the University of Texas, and Arvind Narayanan, now a researcher at Stanford University.

By examining correlations between various online accounts, the scientistsshowed that they could identify more than 30 percent of the users of both Twitter, the microblogging service, and Flickr, an online photo-sharing service, even though the accounts had been stripped of identifying information like account names and e-mail addresses.

“When you link these large data sets together, a small slice of our behavior and the structure of our social networks can be identifying,” Mr. Shmatikov said.

Even more unnerving to privacy advocates is the work of two researchers from Carnegie Mellon University. In a paper published last year, Alessandro Acquisti and Ralph Gross reported that they could accurately predict the full, nine-digit Social Security numbers for 8.5 percent of the people born in the United States between 1989 and 2003 — nearly five million individuals.

Social Security numbers are prized by identity thieves because they are used both as identifiers and to authenticate banking, credit card and other transactions.

The Carnegie Mellon researchers used publicly available information from many sources, including profiles on social networks, to narrow their search for two pieces of data crucial to identifying people — birthdates and city or state of birth.

That helped them figure out the first three digits of each Social Security number, which the government had assigned by location. The remaining six digits had been assigned through methods the government didn't disclose, although they were related to when the person applied for the number. The researchers used projections about those applications as well as other public data, like the Social Security numbers of dead people, and then ran repeated cycles of statistical correlation and inference to partly re-engineer the government's number-assignment system.

To be sure, the work by Mr. Acquisti and Mr. Gross suggests a potential, not actual, risk. But unpublished research by them explores how criminals could use similar techniques for large-scale identity-theft schemes.

More generally, privacy advocates worry that the new frontiers of data collection, brokering and mining, are largely unregulated. They fear "online redlining," where products and services are offered to some consumers and not others based on statistical inferences and predictions about individuals and their behavior.

The F.T.C. and Congress are weighing steps like tighter industry requirements and the creation of a "do not track" list, similar to the federal "do not call" list, to stop online monitoring.

But Jon Kleinberg, a professor of computer science at Cornell University who studies social networks, is skeptical that rules will have much impact. His advice: "When you're doing stuff online, you should behave as if you're doing it in public — because increasingly, it is."