Just four bits of credit card data can identify most anyone (Update)

January 29 2015

Credit: Yves-Alexandre de Montjoye

In this week's issue of the journal *Science*, MIT researchers report that just four fairly vague pieces of information—the dates and locations of four purchases—are enough to identify 90 percent of the people in a data set recording three months of credit-card transactions by 1.1 million users.

When the researchers also considered coarse-grained information about the prices of purchases, just three data points were enough to identify an even larger percentage of people in the data set. That means that someone with copies of just three of your recent receipts—or one receipt, one Instagram photo of you having coffee with friends, and one tweet about the phone you just bought—would have a 94 percent chance of extracting your credit card records from those of a million other people. This is true, the researchers say, even in cases where no one in the data set is identified by name, address, credit card number, or anything else that we typically think of as personal information.

The paper comes roughly two years after an earlier analysis of mobile-phone records that yielded very similar results.

"If we show it with a couple of data sets, then it's more likely to be true in general," says Yves-Alexandre de Montjoye, an MIT graduate student in media arts and sciences who is first author on both papers. "Honestly, I could imagine reasons why credit-card metadata would differ or would be equivalent to mobility data."

De Montjoye is joined on the new paper by his advisor, Alex "Sandy" Pentland, the Toshiba Professor of Media Arts and Science; Vivek Singh, a former postdoc in Pentland's group who is now an assistant professor at Rutgers University; and Laura Radaelli, a postdoc at Tel Aviv University.

The data set the researchers analyzed included the names and locations of the shops at which purchases took place, the days on which they took place, and the purchase amounts. Purchases made with the same credit card were all tagged with the same random identification number.

For each identification number—each customer in the data set—the researchers selected purchases at random, then determined how many

other customers' purchase histories contained the same data points. In separate analyses, the researchers varied the number of data points per customer from two to five. Without price information, two data points were still sufficient to identify more than 40 percent of the people in the data set. At the other extreme, five points with price information was enough to identify almost everyone.

The researchers characterized price very coarsely, treating all prices that fell within a few fixed ranges as functionally equivalent. So, for instance, a purchase of $20 at some store on some day in one person's history would count as a match with a purchase of $40 by someone else at the same store on the same day, since both purchases fell within the range $16 to $49. This was an attempt to represent the uncertainty of someone estimating purchase amounts from secondary information, such as an Instagram photo of the food on someone's plate. The limits of each range were based on a fixed percentage of its median value: The range $16 to $49, for instance, is the median value of purchases ($32.50) plus or minus 50 percent, rounded to the nearest dollar.

Preserving anonymity in large data sets is a pressing concern because public and private entities alike see aggregated digital data as a source of novel insights. Retailers studying anonymized credit-card histories could certainly learn something about the tastes of their customers, but economists might also learn something about the relationship of, say, inflation or consumer spending to other economic factors.

So the MIT researchers also examined the effects of coarsening the data—intentionally making it less precise, in the hope of preserving privacy while still enabling useful analysis. That makes identifying individuals more difficult, but not at a very encouraging rate. Even if the data set characterized each purchase as having taken place sometime in the span of a week at one of 150 stores in the same general areas, four purchases (with 50 percent uncertainty about price) would still be

enough to identify more than 70 percent of users.

Nonetheless, de Montjoye and Pentland remain adamant that socially beneficial uses of big data should be pursued. "Sandy and I do really believe that this data has great potential and should be used," de Montjoye says. "We, however, need to be aware and account for the risks of re-identification."

In separate work, de Montjoye, Pentland, and other members of Pentland's group have begun developing a system that would enable people to store the data generated by their mobile devices on secure servers of their own choosing. Researchers looking for useful patterns in aggregate data would send queries through the system, which would return only the pertinent data—such as, for instance, the average amount spent on gasoline during different time periods.

More information: "Unique in the shopping mall: On the reidentifiability of credit card metadata," by Y.-A. de Montjoye et al. *Science*, www.sciencemag.org/lookup/doi/ … 1126/science.1256297

Provided by Massachusetts Institute of Technology