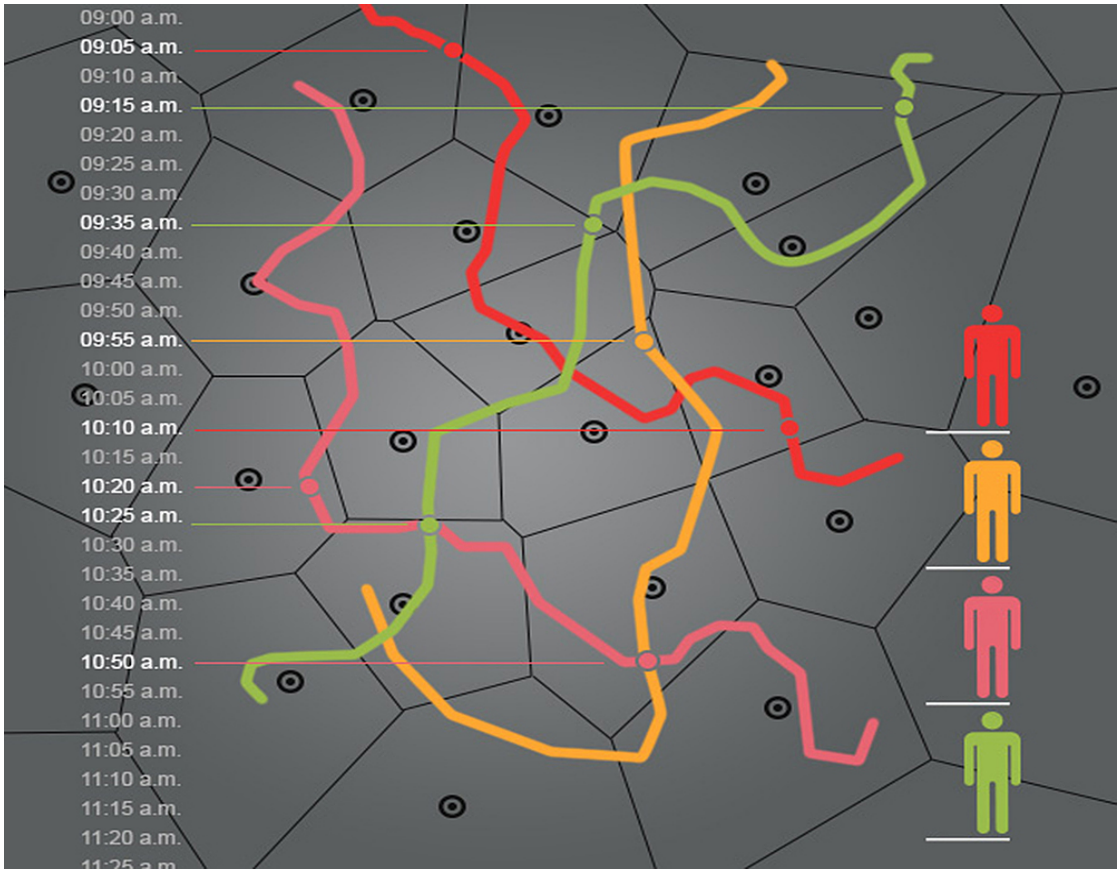
How hard is it to 'de-anonymize' cellphone data?

MIT



A new formula that characterizes the privacy afforded by large, aggregate data sets may be discouraging, but could help sharpen policy discussion.

The proliferation of sensor-studded cellphones could lead to a wealth of data with socially useful applications — in urban planning, epidemiology, operations research and emergency preparedness, among other things. Of course, before being released to researchers, the data would have to be stripped of identifying information. But how hard could it be to protect the identity of one unnamed cellphone user in a data set of hundreds of thousands or even millions?

According to a paper appearing this week in Scientific Reports, harder than you might think. Researchers at MIT and the Université Catholique de Louvain, in Belgium, analyzed data on 1.5 million cellphone users in a small European country over a span of 15 months and found that just four points of reference, with fairly low spatial and temporal resolution, was enough to uniquely identify 95 percent of them.

In other words, to extract the complete location information for a single person from an “anonymized” data set of more than a million people, all you would need to do is place him or her within a couple of hundred yards of a cellphone transmitter,

sometime over the course of an hour, four times in one year. A few Twitter posts would probably provide all the information you needed, if they contained specific information about the person's whereabouts.

The first author on the paper is Yves-Alexandre de Montjoye, a graduate student in the research group of Toshiba Professor of Media Arts and Science Sandy Pentland. He's joined by César Hidalgo, an assistant professor of media arts and science; Vincent Blondel, a visiting professor at MIT and a professor of applied mathematics at Université Catholique; and Michel Verleysen, a professor of electrical engineering at Université Catholique.

Focusing the debate

Hidalgo's group specializes in applying the tools of statistical physics to a wide range of subjects, from communications networks to genetics to economics. In this case, he and de Montjoye were able to use those tools to uncover a simple mathematical relationship between the resolution of spatiotemporal data and the likelihood of identifying a member of a data set.

According to their formula, the probability of identifying someone goes down if the resolution of the measurements decreases, but less than you might think. Reporting the time of each measurement as imprecisely as sometime within a 15-hour span, or location as imprecisely as somewhere amid 15 adjacent cell towers, would still enable the unique identification of half the people in the sample data set.

But while its initial application may be discouraging, de Montjoye and Hidalgo hope that their formula will provide a way for researchers and policy analysts to reason more rigorously about the privacy safeguards that need to be put in place when they're working with aggregated location data.

"Both César and I deeply believe that we all have a lot to gain from this data being used," de Montjoye says. "This formula is something that could be useful to help the debate and decide, OK, how do we balance things out, and how do we make it a fair deal for everyone to use this data?"

Everybody's different

In the data set that the researchers analyzed, the location of a cellphone was inferred solely from that of the cell tower it was connected to, and the time of the connection was given as falling within a one-hour interval. Each cellphone had a unique, randomly generated identifying number, so that its movement could be traced over time. But there was no information connecting that number to the phone's owner.

The researchers randomly selected a representative sampling from the set of 1.5 million cellphone traces and, for each trace, began choosing points at random. For 95 percent of the traces, just four randomly selected points was enough to distinguish them from all other traces in the database. In the worst (or, from another perspective, best) case, 11 measurements were necessary.

“There’s a concern with this data, to what extent can we preserve anonymity,” says Luis Bettencourt, a professor at the Santa Fe Institute who studies social systems. “What they are showing here, quite clearly, is that it’s very hard to preserve anonymity.”

But for Bettencourt, the uniqueness of people’s trajectories through cities is itself precisely the type of information that analysis of cellphone data is meant to uncover. “This is interesting, from a scientific point of view, to understand how people use urban space,” Bettencourt says. “It shows what kind of social systems cities are.”

The researchers suspect that similar relationships might hold for other types of data. “I would not be surprised if a similar result — maybe requiring more points — would, for example, extend to web browsing,” Hidalgo says. “The space of potential combinations is really large. When a person is, in some sense, being expressed in a space in which the total number of combinations is huge, the probability that two people would have the same exact trajectory — whether it’s walking or browsing — is almost nil.”

Source: http://web.mit.edu/newsoffice/2013/de-anonymize-cellphone-data-0327.html [1]

Links:
[1] http://web.mit.edu/newsoffice/2013/de-anonymize-cellphone-data-0327.html