The privacy risks of compiling mobility data

7 December 2018, by Rob Matheson

MIT researchers find that the growing practice of compiling massive datasets about people's movement patterns for urban planning and development research may, in fact, put people's private data at risk — even if that data is anonymized. Credit: Massachusetts Institute of Technology

A new study by MIT researchers finds that the growing practice of compiling massive, anonymized datasets about people's movement patterns is a double-edged sword: While it can provide deep insights into human behavior for research, it could also put people's private data at risk.

Companies, researchers, and other entities are beginning to collect, store, and process anonymized data that contains "location stamps" (geographical coordinates and time stamps) of users. Data can be grabbed from mobile phone records, credit card transactions, public transportation smart cards, Twitter accounts, and mobile apps. Merging those datasets could provide rich information about how humans travel, for instance, to optimize transportation and urban planning, among other things.

But with big data come big privacy issues: Location stamps are extremely specific to individuals and can be used for nefarious purposes. Recent research has shown that, given only a few randomly selected points in mobility datasets, someone could identify and learn sensitive information about individuals. With merged mobility datasets, this becomes even easier: An agent could potentially match users trajectories in anonymized data from one dataset, with deanonymized data in another, to unmask the anonymized data.

In a paper published today in *IEEE Transactions on Big Data*, the MIT researchers show how this can happen in the first-ever analysis of so-called user "matchability" in two large-scale datasets from Singapore, one from a mobile network operator and one from a local transportation system.

The researchers use a statistical model that tracks location stamps of users in both datasets and provides a probability that data points in both sets come from the same person. In experiments, the researchers found the model could match around 17 percent of individuals in one week's worth of data, and more than 55 percent of individuals after one month of collected data. The work demonstrates an efficient, scalable way to match mobility trajectories in datasets, which can be a boon for research. But, the researchers warn, such processes can increase the possibility of deanonymizing real user data.

"As researchers, we believe that working with large-scale datasets can allow discovering unprecedented insights about human society and mobility, allowing us to plan cities better. Nevertheless, it is important to show if identification is possible, so people can be aware of potential risks of sharing mobility data," says Daniel Kondor, a postdoc in the Future Urban Mobility Group at the Singapore-MIT Alliance for Research and Technology.

"In publishing the results—and, in particular, the consequences of deanonymizing data—we felt a bit like 'white hat' or 'ethical' hackers," adds co-author Carlo Ratti, a professor of the practice in MIT's

Department of Urban Studies and Planning and director of MIT's Senseable City Lab. "We felt that it was important to warn people about these new possibilities [of data merging] and [to consider] how we might regulate it."

The co-authors of the study are Behrooz Hashemian, a postdoc at the Senseable City Lab, and Yves-Alexandre de Mondjoye of the Department of Computing and Data Science Institute of Imperial College London.

Eliminating false positives

To understand how matching location stamps and potential deanonymization works, consider this scenario: "I was at Sentosa Island in Singapore two days ago, came to the Dubai airport yesterday, and am on Jumeirah Beach in Dubai today. It's highly unlikely another person's trajectory looks exactly the same. In short, if someone has my anonymized credit card information, and perhaps my open location data from Twitter, they could then deanonymize my credit card data," Ratti says.

Similar models exist to evaluate deanonymization in data. But those use computationally intensive approaches for re-identification, meaning to merge anonymous data with public data to identify specific individuals. These models have only worked on limited datasets. The MIT researchers instead used a simpler statistical approach—measuring the probability of false positives—to efficiently predict matchability among scores of users in massive datasets.

In their work, the researchers compiled two anonymized "low-density" datasets—a few records per day—about mobile phone use and personal transportation in Singapore, recorded over one week in 2011. The mobile data came from a large mobile network operator and comprised timestamps and geographic coordinates in more than 485 million records from over 2 million users. The transportation data contained over 70 million records with timestamps for individuals moving through the city.

The probability that a given user has records in both datasets will increase along with the size of the merged datasets, but so will the probability of false positives. The researchers' model selects a user from one dataset and finds a user from the other dataset with a high number of matching location stamps. Simply put, as the number of matching points increases, the probability of a false-positive match decreases. After matching a certain number of points along a trajectory, the model rules out the possibility of the match being a false positive.

Focusing on typical users, they estimated a matchability success rate of 17 percent over a week of compiled data, and about 55 percent for four weeks. That estimate jumps to about 95 percent with data compiled over 11 weeks.

The researchers also estimated how much activity is needed to match most users over a week. Looking at users with between 30 and 49 personal transportation records, and around 1,000 mobile records, they estimated more than 90 percent success with a week of compiled data. Additionally, by combining the two datasets with GPS traces—regularly collected actively and passively by smartphone apps—the researchers estimated they could match 95 percent of individual trajectories, using less than one week of data.

Better privacy

With their study, the researchers hope to increase public awareness and promote tighter regulations for sharing consumer data. "All data with location stamps (which is most of today's collected data) is potentially very sensitive and we should all make more informed decisions on who we share it with," Ratti says. "We need to keep thinking about the challenges in processing large-scale data, about individuals, and the right way to provide adequate guarantees to preserve privacy."

To that end, Ratti, Kondor, and other researchers have been working extensively on the ethical and moral issues of big data. In 2013, the Senseable City Lab at MIT launched an initiative called "Engaging Data," which involves leaders from government, privacy rights groups, academia, and business, who study how mobility data can and should be used by today's data-collecting firms.

"The world today is awash with big data," Kondor says. "In 2015, mankind produced as much information as was created in all previous years of human civilization. Although data means a better knowledge of the urban environment, currently much of this wealth of information is held by just a few companies and public institutions that know a lot about us, while we know so little about them. We need to take care to avoid data monopolies and misuse."

More information: Daniel Kondor et al. Towards matching user mobility traces in large-scale datasets, *IEEE Transactions on Big Data* (2018). DOI: 10.1109/TBDATA.2018.2871693

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology