Sorry, your data can still be identified by Google even if its anonymized

Urban planners and researchers at MIT found that it's shockingly easy to "reidentify" the anonymous data that people generate all day, every day in cities.

*By Kelsey Campbell-Dollaghan*2 minute Read

Thanks to the near-complete saturation of the city with sensors and smartphones, we humans are now walking, talking data factories. Passing through a subway turnstile, sending a text, even just carrying a phone in your pocket: we generate location-tagged data on an hourly basis. All that data can be a boon for urban planners and designers who want to understand cities–and, of course, for tech companies and advertisers who want to understand the people in them. Questions about data privacy are frequently met with a chorus of, *It's anonymized! Any identifying features are scrubbed from the data!*

The reality, a group of MIT scientists and urban planners show in a new study, is that it's fairly simple to figure out who is who anyway. In other words, anonymized data can be deanonymized pretty quickly when you're working with multiple datasets within a city.

Carlo Ratti, the MIT Senseable City Lab founder who co-authored the study in *IEEE Transactions on Big Data,* says that the research process made them feel “a bit like ‘white hat’ or ‘ethical’ hackers” in a news release. First, they combined two anonymized datasets of people in Singapore, one of mobile phone logs and the other of transit trips, each containing “location stamps” detailing just the time and place of each data point. Then they used an algorithm to match users whose data overlapped closely between each set–in other words, they had phone logs and transit logs with similar time and location stamps–and tracked how closely those stamps matched up over time, eliminating false positives as they went. In the end, it took a week to match up 17% of the users and 11 weeks to get to a 95% rate of accuracy. (With the added GPS data from smartphones, it took less than a week to hit that number.)

While the MIT group wasn’t trying to unmask specific users in this dataset, they proved that someone acting in bad faith could merge such anonymized datasets with personal ones using the same process, easily pinning the timestamps together to figure out who was who.

The takeaway is not just that a malicious actor or company could use this process to surveil citizens. It’s that urban planners and designers who stand to learn so much from these big urban datasets–for instance, Ratti’s own lab recently used such data for a project on reducing parking, while other groups use it to study everything from urban poverty to accessibility–need to be careful about whether all that data could be combined to deanonymize it.

“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,” observed Daniel Kondor of MIT’s Future Urban Mobility Group in the release. “Nevertheless, it is important to show if identification is possible, so people can be aware of potential risks of sharing mobility data,” adding, “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.”

In other words, as urban planners, tech companies, and governments collect and share data, we now know that “it’s anonymized” is never a guarantee of privacy. And as they dig deep into the data we generate, cities and citizens need to demand that this data can never be reidentified.