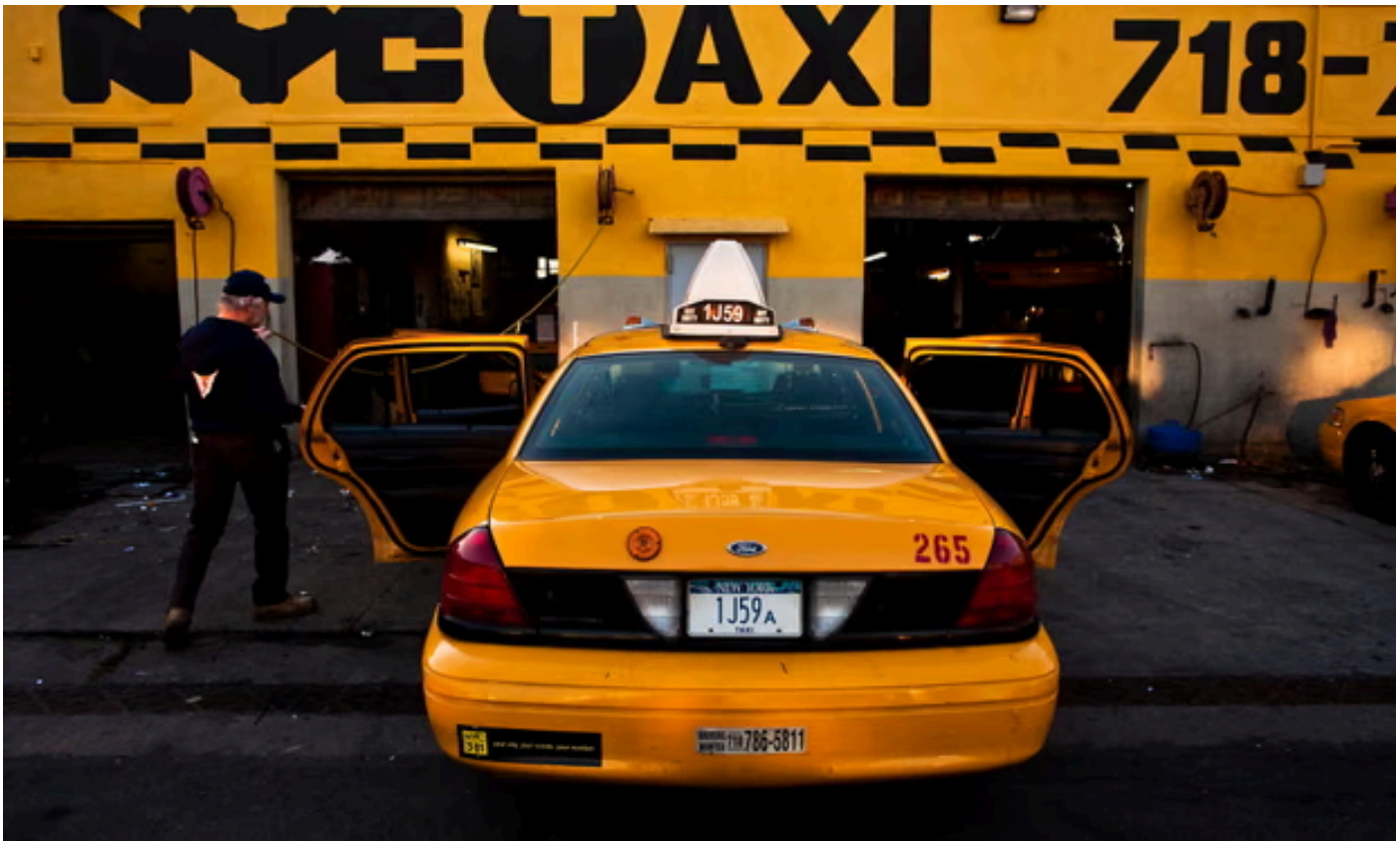
Technology

This article is more than **11 years old**

New York taxi details can be extracted from anonymised data, researchers say

FoI request reveals data on 173m individual trips in US city - but could yield more details, such as drivers' addresses and income



Data about New York city taxi drivers and rides could be de-anonymised, researchers warn. Photograph: Jan Johannessen/Getty Images Photograph: Jan Johannessen/Getty Images

Alex Hern

The trove of information comes from a Freedom of Information request filed by open data activist Chris Whong. In the original release, each record includes the time and location of the pickup and drop off, as well as an anonymised licence number and medallion number, which identifies the driver and taxi respectively. But Vijay Pandurangan, founder of secure password manager Mitro, discovered that the anonymous data was easy to restore to its original, personally identifiable format.

"These data are a veritable trove for people who love cities, transit, and data visualisation," Pandurangan wrote. "But there's a big problem: the personally identifiable information (the driver's licence number and taxi number) hasn't been anonymised properly—what's worse, it's trivial to undo, and with other publicly available data, one can even figure out which person drove each trip."

Panduragan realised that the medallion and licence numbers both have a very specific format. Medallions only take one of three formats - either 5X55, XX555 or XXX555 - while licences are all six-digit or seven-digit numbers starting with a five. That means that there are only 2m possible license numbers, and 22m possible medallion numbers.

That let Panduragan reverse-engineer the anonymised data to find out which trips were carried out by which drivers, and in which taxis. The data had been anonymised by hashing, a cryptographic function which is supposed to be "one-way": it's very easy to find the hash of a given piece of data, and very hard - mathematically impossible, in theory - to find the piece of data which resulted in a given hash (for instance, the MD5 hash, the particular type used by NYC, of the data "Alex" is a08372b70196c21a9229cf04db6b7ceb). As the same piece of data always results in the same hash, such functions are frequently used to anonymise just this sort of data.

But once Panduragan had narrowed the possible entries down to 24m different numbers, it was the matter of only minutes to determine which numbers were associated with which pieces of anonymised data.

"Modern computers are fast: so fast that computing the 24m hashes took less than two minutes," he said. "It took a while longer to de-anonymise the entire dataset, but... [I] had it done within an hour.

"There's a ton of resources on NYC Taxi and Limousine commission, including a mapping from licence number to driver name, and a way to look up owners of medallions. I haven't linked them here but it's easy to find using a quick Google search... This anonymisation is so poor that anyone could, with less than two hours work, figure which driver drove every single

trip in this entire dataset. It would even be easy to calculate drivers' gross income or infer where they live."

Paduragan points out that there are a number of ways that the city could have more successfully anonymised the data. The first is if they hadn't tried to be so smart: rather than going through the effort of hashing the data, if they had simply assigned random numbers to each licence plate, it would have been much more difficult to work backwards. New York's Taxi and Limousine Commission was asked for comment, but didn't respond by publication time.

But in an update to his initial post on Medium, he says that there may be little the city could have done to fully guard against some data being de-anonymised. Citing a similar case where Netflix released a large quantity of anonymised data, which was then de-anonymised, he says that "there are a number of other ways in which [personally identifiable information] may be reconstructed ... NYC is dense enough that it may be much more challenging to target specific passengers using these data, however.

"Anonymising data is really hard."

In Netflix's case, in 2006 the company released the movie ratings of 500,000 customers in an effort to encourage better recommendation algorithms. Unlike the NYC data, there was no way to reverse engineer identifiable information from the dataset itself. But what two researchers realised is that some users were likely to rate movies similarly on Netflix and IMDB. What's more, they are likely to rate the same movies at the same time, which made it possible to link an anonymised entry in the Netflix dataset with a very identifiable user on IMDB - even if the IMDB ratings don't include more sensitive movies which the user may not have wanted to make public.

For instance, the researchers said, a user's "political orientation may be revealed by his strong opinions about Power and Terror: Noam Chomsky in Our Times and Fahrenheit 9/11, and his religious views by his ratings on Jesus of Nazareth and The Gospel of John." Eventually, the revelations led to a 2009 lawsuit from an in-the-closet lesbian mother, who sued Netflix for privacy violation.

The Netflix experience raises a further question: how much more information can be deanonymised from the taxi data set? "I think there's a much bigger privacy issue here than what the author focuses on," says one commenter on Hacker News.

"Couldn't you deduce many passenger identities based on addresses? There's a lot of scenarios where passenger identities could be effectively de-anonymised, just based on GPS data. You could then use this data set to analyse their comings and goings ... can you imagine someone just plotting

all the trips from a single gay bar? Listing off all the connected residential addresses? And not only that, any subsequent trips home from those addresses the next morning?"

Uber: the smartphone app that is driving London cabbies to distraction

Most viewed