Anonymizing personal data 'not enough to protect privacy,' shows new study

23 July 2019

Credit: CC0 Public Domain

With the first large fines for breaching EU General Data Protection Regulation (GDPR) regulations upon us, and the UK government about to review GDPR guidelines, researchers have shown how even anonymised datasets can be traced back to individuals using machine learning.

The researchers say their paper, published today in *Nature Communications*, demonstrates that allowing data to be used—to train AI algorithms, for example—while preserving people's privacy, requires much more than simply adding noise, sampling datasets, and other de-identification techniques.

They have also published a demonstration tool that allows people to understand just how likely they are to be traced, even if the dataset they are in is anonymised and just a small fraction of it shared.

They say their findings should be a wake-up call for policymakers on the need to tighten the rules for what constitutes truly anonymous data.

Companies and governments both routinely collect and use our personal data. Our data and the way it's used is protected under relevant laws like GDPR or the US's California Consumer Privacy Act (CCPA).

Data is 'sampled' and anonymised, which includes stripping the data of identifying characteristics like names and email addresses, so that individuals cannot, in theory, be identified. After this process, the data's no longer subject to data protection regulations, so it can be freely used and sold to third parties like advertising companies and data brokers.

The new research shows that once bought, the data can often be reverse engineered using machine learning to re-identify individuals, despite the anonymisation techniques.

This could expose sensitive information about personally identified individuals, and allow buyers to build increasingly comprehensive personal profiles of individuals.

The research demonstrates for the first time how easily and accurately this can be done—even with incomplete datasets.

In the research, 99.98 per cent of Americans were correctly re-identified in any available 'anonymised' dataset by using just 15 characteristics, including age, gender, and marital status.

First author Dr. Luc Rocher of UCLouvain said: "While there might be a lot of people who are in their thirties, male, and living in New York City, far fewer of them were also born on 5 January, are driving a red sports car, and live with two kids (both girls) and one dog."

To demonstrate this, the researchers developed a machine learning model to evaluate the likelihood for an individual's characteristics to be precise

enough to describe only one person in a population of billions.

They also developed an online tool, which doesn't save data and is for demonstration purposes only, to help people see which characteristics make them unique in datasets.

The tool first asks you put in the first part of their post (UK) or ZIP (US) code, gender, and date of birth, before giving them a probability that their profile could be re-identified in any anonymised dataset.

It then asks your marital status, number of vehicles, house ownership status, and employment status, before recalculating. By adding more characteristics, the likelihood of a match to be correct dramatically increases.

Senior author Dr. Yves-Alexandre de Montjoye, of Imperial's Department of Computing, and Data Science Institute, said: "This is pretty standard information for companies to ask for. Although they are bound by GDPR guidelines, they're free to sell the data to anyone once it's anonymised. Our research shows just how easily—and how accurately—individuals can be traced once this happens.

He added: "Companies and governments have downplayed the risk of re-identification by arguing that the datasets they sell are always incomplete.

"Our findings contradict this and demonstrate that an attacker could easily and accurately estimate the likelihood that the record they found belongs to the person they are looking for."

Re-identifying anonymised data is how journalists exposed Donald Trump's 1985-94 tax returns in May 2019.

Co-author Dr. Julien Hendrickx from UCLouvain said: "We're often assured that anonymisation will keep our personal information safe. Our paper shows that de-identification is nowhere near enough to protect the privacy of people's data."

The researchers say policymakers must do more to protect individuals from such attacks, which could have serious ramifications for careers as well as personal and financial lives.

Dr. Hendrickx added: "It is essential for anonymisation standards to be robust and account for new threats like the one demonstrated in this paper."

Dr. de Montjoye said: "The goal of anonymisation is so we can use data to benefit society. This is extremely important but should not and does not have to happen at the expense of people's privacy."

More information: Luc Rocher et al. Estimating the success of re-identifications in incomplete datasets using generative models, *Nature Communications* (2019). DOI: 10.1038/s41467-019-10933-3

Provided by Imperial College London