Researchers spotlight the lie of ‘anonymous’ data

techcrunch.com/2019/07/24/researchers-spotlight-the-lie-of-anonymous-data/

Natasha Lomas

Natasha Lomas @riptari / 2 years

Researchers from two universities in Europe have published a method they say is able to correctly re-identify 99.98% of individuals in anonymized data sets with just 15 demographic attributes.

Their model suggests complex data sets of personal information cannot be protected against re-identification by current methods of “anonymizing” data — such as releasing samples (subsets) of the information.

Indeed, the suggestion is that no “anonymized” and released big data set can be considered safe from re-identification — not without strict access controls.

“Our results suggest that even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR [Europe’s General Data Protection Regulation] and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model,” the researchers from Imperial College London and Belgium’s Université Catholique de Louvain write in the abstract to their paper, which has been published in the journal Nature Communications.

It’s of course by no means the first time data anonymization has been shown to be reversible. One of the researchers behind the paper, Imperial College’s Yves-Alexandre de Montjoye, has demonstrated in previous studies looking at credit card metadata that just four random pieces of information were enough to re-identify 90% of the shoppers as unique individuals, for example.

In another study, which de Montjoye co-authored, that investigated the privacy erosion of smartphone location data, researchers were able to uniquely identify 95% of the individuals in a data set with just four spatio-temporal points.

At the same time, despite such studies that show how easy it can be to pick individuals out of a data soup, “anonymized” consumer data sets such as those traded by brokers for marketing purposes can contain orders of magnitude more attributes per person.

The researchers cite data broker Experian selling Alteryx access to a de-identified data set containing 248 attributes per household for 120 million Americans, for example.

By their models’ measure, essentially *none* of those households are safe from being re-identified. Yet massive data sets continue being traded, greased with the emollient claim of “anonymity”…

(If you want to be further creeped out by how extensively personal data is traded for commercial purposes the disgraced, and now defunct, political data company, Cambridge Analytica, said last year — at the height of the Facebook data misuse scandal — that its foundational data set for clandestine U.S. voter targeting efforts had been licensed from well-known data brokers such as Acxiom, Experian and Infogroup. Specifically it claimed to have legally obtained “millions of data points on American individuals” from “very large reputable data aggregators and data vendors.”)

While research has shown for years how frighteningly easy it is to re-identify individuals within anonymous data sets, the novel bit here is the researchers have built a statistical model that estimates how easy it would be to do so to any data set.

They do that by computing the probability that a potential match is correct — so essentially they’re evaluating match uniqueness. They also found small sampling fractions failed to protect data from being re-identified.

“We validated our approach on 210 datasets from demographic and survey data and showed that even extremely small sampling fractions are not sufficient to prevent re-identification and protect your data,” they write. “Our method obtains AUC accuracy scores ranging from 0.84 to 0.97 for predicting individual uniqueness with low false-discovery rate. We showed that 99.98% of Americans were correctly re-identified in any available ‘anonymised’ dataset by using just 15 characteristics, including age, gender, and marital status.”

They have taken the perhaps unusual step of releasing the code they built for the experiments so that others can reproduce their findings. They have also created a web interface where anyone can play around with inputting attributes to obtain a score of how likely it would be for them to be re-identifiable in a data set based on those particular data points.

In one test based on inputting three random attributes (gender, data of birth, ZIP code) into this interface, the chance of re-identification of the theoretical individual scored by the model went from 54% to a full 95% by adding just one more attribute (marital status) — which underlines that data sets with far fewer attributes than 15 can still pose a massive privacy risk to most people.

The rule of thumb is the more attributes in a data set, the more likely a match is to be correct and therefore the less likely the data can be protected by “anonymization.”

This offers a lot of food for thought when, for example, Google-owned AI company DeepMind has been given access to one million “anonymized” eye scans as part of a research partnership with the U.K.’s National Health Service.

Biometric data is of course chock-full of unique data points by its nature. So the notion that any eye scan — which contains more than (literally) a few pixels of visual data — could really be considered “anonymous” just isn’t plausible.

Europe’s current data protection framework does allow for truly anonymous data to be freely used and shared — versus the stringent regulatory requirements the law imposes for processing and using personal data.

Though the framework is also careful to recognize the risk of re-identification — and uses the categorization of pseudonymized data rather than anonymous data (with the former very much remaining personal data and subject to the same protections). Only if a data set is stripped of sufficient elements to ensure individuals can no longer be identified can it be considered “anonymous” under GDPR.

The research underlines how difficult it is for any data set to meet that standard of being truly, robustly anonymous — given how the risk of re-identification demonstrably steps up with even just a few attributes available.

“Our results reject the claims that, first, re-identification is not a practical risk and, second, sampling or releasing partial datasets provide plausible deniability,” the researchers assert.

“Our results, first, show that few attributes are often sufficient to re-identify with high confidence individuals in heavily incomplete datasets and, second, reject the claim that sampling or releasing partial datasets, e.g., from one hospital network or a single online service, provide plausible deniability. Finally, they show that, third, even if population uniqueness is low—an argument often used to justify that data are sufficiently de-identified to be considered anonymous —, many individuals are still at risk of being successfully re-identified by an attacker using our model."

They go on to call for regulators and lawmakers to recognize the threat posed by data reidentification, and to pay legal attention to "provable privacy-enhancing systems and security measures" which they say can allow for data to be processed in a privacy-preserving way — including in their citations a 2015 paper which discusses methods such as encrypted search and privacy preserving computations; granular access control mechanisms; policy enforcement and accountability; and data provenance.

"As standards for anonymization are being redefined, incl. by national and regional data protection authorities in the EU, it is essential for them to be robust and account for new threats like the one we present in this paper. They need to take into account the individual risk of re-identification and the lack of plausible deniability—even if the dataset is incomplete —, as well as legally recognize the broad range of provable privacy-enhancing systems and security measures that would allow data to be used while effectively preserving people's privacy," they add.

"Moving forward, they question whether current de-identification practices satisfy the anonymization standards of modern data protection laws such as GDPR and CCPA [California's Consumer Privacy Act] and emphasize the need to move, from a legal and regulatory perspective, beyond the de-identification release-and-forget model."