New kind of attack called 'downcoding' demonstrates flaws in anonymizing data

October 10 2022, by Rob Mitchum

Credit: Matthew Ansley via Unsplash

When datasets containing personal information are shared for research or used by companies, researchers try to disguise data—removing the final one or two digits of a zip code, for example—while still preserving its utility for insight.

But while deidentification is often intended to satisfy legal requirements

for data privacy, the most commonly used methods stand on shaky technical ground.

University of Chicago computer scientist Aloni Cohen deals the latest decisive blow against the most popular deidentification techniques in a new paper.

By describing a new kind of attack called "downcoding," and demonstrating the vulnerability of a deidentified data set from an online education platform, Cohen sends a warning that these data transformations should not be considered sufficient to protect individuals' privacy.

"Even by the regulatory standards, there's a problem here," said Cohen, an assistant professor of computer science.

Sounding the alarm

For years, computer science security and privacy researchers have sounded the alarm about the methods most often used to deidentify data, finding new attacks that can reidentify seemingly anonymized data points and proposing fixes. Yet these methods remain in common use, and held up as legally sufficient for fulfilling privacy-protection regulations such as HIPAA and GDPR.

"Policymakers care about real world risks instead of hypothetical risks," Cohen said. "So people have argued that the risks security and privacy researchers pointed out were hypothetical or very contrived."

While pursuing his Ph.D. at MIT, Cohen set out to disprove this argument. The most common deidentification methods stem from an approach called k-anonymity, which transforms data just enough to make each individual indistinguishable from a certain number of other

individuals in the data set. Cohen's idea was that the very target of this deidentification method left it open to attack.

"The goal when you're doing that sort of technique is to redact as little as you need to guarantee a target level of anonymity," Cohen said. "But if you achieve that goal of redacting just as little as you need, then the fact that that's the minimum might tell you something about what was redacted."

Deidentification works by redacting quasi-identifiers—information that can be put together with data from a second source to de-anonymize a data subject. Failing to account for all possible quasi-identifiers can lead to disclosures. In one famous example, researchers took deidentified Netflix viewing data and combined it with data from the online movie review site IMDB, identifying users in the first data set by when they logged reviews of the movies they had recently watched.

Since these discoveries in the 2000s, policymakers have relied on experts to determine which aspects of a dataset are quasi-identifiers or not, to establish the bar for anonymity. Cohen tested the extreme—if every attribute is considered a quasi-identifier, do k-anonymity and its derivative techniques still work?

"If deidentification works at all, it should work when everything is quasi-identifying," Cohen said. "That's part of what makes this work powerful. It also means that the attacks work against almost all the techniques related to k-anonymity instead of any one specifically. The Netflix attack showed that it's hard to say what is and isn't a quasi-identifier. The downcoding attacks shows that, in certain settings, it doesn't much matter."

'Not a magic wand'

The paper describes two theoretical attacks and one real-world example that undermine the argument for these protections. The first, downcoding, reverse-engineers the transformations performed on the data, such as the zip code example mentioned earlier. The second attack uses downcoding for a predicate singling-out (PSO) attack, a specific type of attack against data anonymization standards under the European Union's privacy law GDPR. That proof was important to show policymakers that k-anonymity is not sufficient for "publish-and-forget" anonymization under GDPR, Cohen said.

"The argument we're making is against the idea that any of those techniques are sufficient to meet the legal bar of anonymization," Cohen said. "We're directly pushing back on that claim. Even by the regulatory standards, there's a problem here."

Cohen illustrated this insufficiency with a separate real-world demonstration on deidentified data from edX, the popular massively open online course (MOOC) platform. By combining the dataset with data scraped from resumes posted to LinkedIn—information that would be trivially available to potential employers—Cohen could identify people who started but failed to complete edX courses, a potential violation of FERPA, the Family Educational Rights and Privacy Act. (edX was alerted to the flaw and has changed its data protections.)

The takeaway message, Cohen said, is that these deidentification methods are not a magic wand for waving away privacy concerns when sharing potentially sensitive data. He hopes that regulators will realize that a layered approach will be much more effective to achieve their goals.

"If what you want to do is take data, sanitize it, and then forget about it—put it on the web or give it to some outside researchers and decide that all your privacy obligations are done—you can't do that using these

techniques," Cohen said. "They should not free you of your obligations to think about and protect the privacy of that data."

More information: Aloni Cohen, Attacks on Deidentification's Defenses. www.usenix.org/conference/usen … 2/presentation/cohen

Provided by University of Chicago