New tool simplifies data sharing, preserves privacy

October 29 2020, by Daniel Tkacik

Credit: Unsplash/CC0 Public Domain

Meet Company X. Company X makes a popular product that lots of people—millions, in fact—use on a daily basis. One day, Company X decides it would like to improve some of the hardware in its product, which is manufactured by Vendor Y. To make these improvements, the company would need to share data with Vendor Y about how its customers use the product.

Unfortunately, that data may contain personal information about Company X's customers, so sharing it would be an invasion of their privacy. Company X doesn't want to do that, so they abandon the improvement opportunity.

According to a new study authored by researchers in Carnegie Mellon University's CyLab and IBM, a new tool can help circumvent this privacy issue in data sharing. Companies, organizations, and governments alike have to deal with this issue in today's world of Big Data. The study is being presented at this week's ACM Internet Measurement Conference, where it has been named a finalist in the conference's Best Paper Award.

One approach that has been used to avoid breaching privacy is to synthesize new data that mimic the original dataset while leaving the sensitive information out. This, however, is easier said than done.

The team of researchers created a new tool—dubbed "DoppelGANger"—that utilizes generative adversarial networks, or GANs, which employ machine learning techniques to synthesize datasets that have the same statistics as the original "training" data.

On the datasets they evaluated, models trained with DoppelGANger-produced synthetic data had up to 43 percent higher accuracy than models trained synthetic data from competing tools.

Most tools today require expertise in complex mathematical modeling, which creates a barrier for data sharing across different levels of expertise. However, DoppelGANger requires little to no prior knowledge of the dataset and its configurations due to the fact that GANs themselves are able to generalize across different datasets and use cases. This makes the tool highly flexible, the researchers say, and that flexibility is key to data sharing in cybersecurity situations.

"We believe that future organizations will need to flexibly utilize all available data to be able to react to an increasingly data-driven and automated attack landscape," says CyLab's Vyas Sekar, a professor in ECE and Lin's co-advisor. "In that sense, any tools that facilitate data sharing are going to be essential."

CyLab's Giulia Fanti, a professor in ECE and Lin's Ph.D. co-advisor, also sees the tool as being beneficial to security engineers.

"Synthetic network data can be used to help create realistic training testbeds for network security engineers without exposing real, sensitive data," says Fanti.

The team's next step is to expand to tool's capabilities, because despite its remarkable performance, it's limited to relatively simple datasets.

"Many networking datasets require significantly more complexity than DoppelGANger is currently able to handle," Lin says.

For those interested in using the tool, DoppelGANger is open-sourced on Github. The research was sponsored in part by the National Science Foundation and the Army Research Laboratory.

More information: Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions, arXiv:1909.13403 [cs.LG] arxiv.org/abs/1909.13403

Provided by Carnegie Mellon University