Altered data sets can still provide statistical integrity and preserve privacy

17 February 2019

Synthetic networks may increase the availability of some data while still protecting individual or institutional privacy, according to a Penn State statistician.

"My key interest is in developing methodology that would enable broader sharing of confidential data in a way that can aid in scientific discovery," said Aleksandra Slavkovic, professor of statistics and associate dean for graduate education, Eberly College of Science, Penn State. "Being able to share confidential data with minimal quantifiable risk for discovery of sensitive information and still ensure statistical accuracy and integrity, is the goal."

Slavkovic has found solutions to this data privacy problem through interdisciplinary collaborations, especially with computer and social scientists. Her research focuses on various data, including network data that capture relationship information between entities such as individuals or institutions. She reported her approaches to providing synthetic networks that satisfy a notion of differential privacy today (Feb 16) during the 2019 annual meeting of the American Association for the Advancement of Science in Washington, D.C.

Differential privacy provides a mathematically provable guarantee of the level of privacy loss to individuals.

Scientists want access to data collected by others for their research, but such access could also compromise personal privacy, even after removal of so-called personally identifiable data.

"An abundance of auxiliary data is the main culprit," said Slavkovic. "With methodological and technological advances in data collection and record linkage, easier access to variety of data sources that could be linked with a dataset in hand, and funding agencies requirements to share data, the risks to data privacy are increasing. But, finding good solutions for managing privacy loss are essential for enabling sound scientific discovery."

Publicly available information from a drug trial on an HIV drug, for example, would indicate who was in the treatment group and who was in the control group. The treatment group would contain only people diagnosed with HIV and even though the data owners withheld personal particulars from that data set, some identifying information would remain. Because so much information is today available online in social media and in other datasets, it is possible to connect the dots and identify people, potentially revealing their HIV status.

"Techniques to link two data sets, say voter records and health insurance data, have greatly improved," said Slavkovic. "In one of the earliest findings, Latanya Sweeny (now at Harvard) showed that by linking these type of data, you can identify 87 percent of the people in the U.S. Census from 1990 based on their date of birth, gender and 5-digit zip code. More recently, researchers used tweets and associated Twitter metadata to show that they can identify users with 96.7 percent accuracy."

Slavkovic notes that it is not just people or institutions whose data are contained in the databases, but that people outside the database can also suffer from invasion of privacy, directly or by association. Linkages between information in a dataset and information on social media might lead to a serious privacy breech—something like HIV status or sexual orientation could have severe repercussions if revealed.

While privacy is important, collected datasets make up an essential source of information for researchers. Currently, in some cases when the data are exceptionally sensitive, researchers must physically go to the data repositories to do their research, making research more difficult and expensive.

Slavkovic is interested in network data. Information that shows the interconnectedness of people or institutions—the nodes—and the connections between nodes. Her approach is to create slightly altered, mirrored network datasets with a few of the nodes moved, connections shifted or edges altered.

"The aim is to create new networks that satisfy the rigorous differential privacy requirements and at the same time capture most of the statistical features from the original network," said Slavkovic.

These synthetic datasets might be sufficient for some researchers to satisfy their research needs. For others, it would be sufficient to test their approaches and hypothesis before having to go to the data storage site. Researchers could test code, do exploratory research and perhaps basic analysis while waiting for permission to use the original data in its repository site.

"We can't satisfy demands for all statistical analysis with the same type of altered data," said Slavkovic. "Some people will need the original data, but others might go a long way with synthetic data such as synthetic networks."

Provided by Pennsylvania State University