Social network analysis privacy tackled

February 14 2015

Protecting people's privacy in an age of online big data is difficult, but doing so when using visual representations of such things as social network data may present unique challenges, according to a Penn State computer scientist.

"Our goal is to be able to release information without making personal or sensitive data available and still be accurate," said Sofya Raskhodnikova, associate professor of computer science and engineering. "But we have to figure out what 'privacy' means and develop a rigorous mathematical foundation for protecting privacy."

In a way, the privacy issue is one of balancing protection of personal data against the benefits of global statistics as a public good. Large databases like those of the U.S. Census Bureau contain data that, when

aggregated and analyzed, can illuminate social, economic and health problems and solutions. However, protecting privacy requires more than simply removing identifying data from files before publishing databases and analytic results. With multiple open public databases available, data can easily be correlated between databases to pull together bits and pieces of deleted data and recover the identifying information.

Raskhodnikova gives the example of a database of movies from a paid service that included, in some cases, movies that people might prefer others did not know they viewed. The database owners wanted to be able to predict what people wanted to view, so they set up a competition to develop the best learning algorithms for real data. They thought that removing identifying information made it safe to publish the database. However, when viewing patterns from the paid service were compared to viewing patterns on a public, self-reporting movie-viewing database, there was enough information to identify an overwhelming majority of the individuals involved.

"Similar problems occur with search engine histories and other databases," said Raskhodnikova. "It isn't enough to just strip the obviously identifying information."

She suggests that "differential privacy" is needed to maximize accuracy of analysis while preventing identification of individual records. Differential privacy restricts the types of analyses that can be performed to those for which the presence or absence of one person does not matter much. It guarantees that an analysis performed on two databases that differ in only one record will return nearly the same result.

"One approach for achieving differential privacy is adding a small amount of noise to the actual statistics before publishing them," said Raskhodnikova. "One problem is figuring out how much noise—it depends on how sensitive the statistic is—and how to do it so that we can

still retain accuracy of results. There are other more sophisticated approaches for achieving differential privacy."

The notion of differential privacy may be especially important to protection of graph data.

Publishing a network of romantic relationships between students in one high school using only gender as an identifier would seem innocuous, for example, but in reality is not. A person with knowledge of the school under study could easily correlate the observed behavior of an outlier - a student with a minority sexual preference, for example - to identify that student's node within the network. Once one node is linked to a particular person, only a small amount of additional information is required to identify the rest of the students included in the network.

As long as the class and students are not identified, researchers once thought that protecting this type of information was not important. However, they now know that using other databases that might be open and available online and various clues from the graph such as the total number of students and the ratio of boys to girls, it might be possible to identify the class and then individual students.

"What the researchers tried to do is strip identifiers—name, etc., and publish, but this is not enough," said Raskhodnikova. "Other information is available that can be correlated with the network information."

Researchers have found differentially private methods for releasing many graph statistics. Examples include the number of occurrences of specified small patterns in a graph and the degree distribution. The degree distribution of a social network specifies what fraction of people have no friends, one friend, two friends, etc. However, some information, for example, the exact number of connections a specific person has, is inherently too sensitive to be released with differential

privacy.

Data that classifies as an outlier also poses a problem. If only one person in a salary database makes over $1 million a year, it becomes fairly easy to identify that person. Some types of information inherently pose a threat to privacy, said Raskhodnikova, and should not be published.

Provided by Pennsylvania State University