Theoretical computer science provides answers to data privacy problem

October 7 2015

Privacy Tools for Sharing Research Data: A National Science Foundation Secure and Trustworthy Cyberspace Project is a multidisciplinary effort to help enable the collection, analysis, and sharing of personal data for research in social science and other fields while providing privacy for individual subjects. In particular, the researchers aim to build an array of computational, statistical, legal, and policy tools that can be incorporated into data repositories to make privacy-protective data-sharing easier for lay researchers. These tools will integrate with the Dataverse Network software, which is already used to host data repositories around the world. Credit: Theoretical computer science research community and the Computing Community Consortium

The promise of big data lies in researchers' ability to mine massive datasets for insights that can save lives, improve services and inform our understanding of the world.

These data may be generated by surfing the web, interacting with medical devices or passing sensors. Some data may be trivial, but in many cases, data are deeply personal. They can even influence our insurance premiums or the price we pay for a product online.

When planning a study, data scientists need to balance their desire to uncover new knowledge with the privacy of the people whom the data represent.

"The science of understanding human behavior, health, and interactions is being transformed by the ability of researchers to collect, analyze, and share data about individuals on a wide scale," a team of Harvard University researchers wrote in a July 2014 paper, "Integrating Approaches to Privacy across the Research Lifecycle."

However, the paper continued, "a major challenge for realizing the full potential of such data science is ensuring the privacy of human subjects."

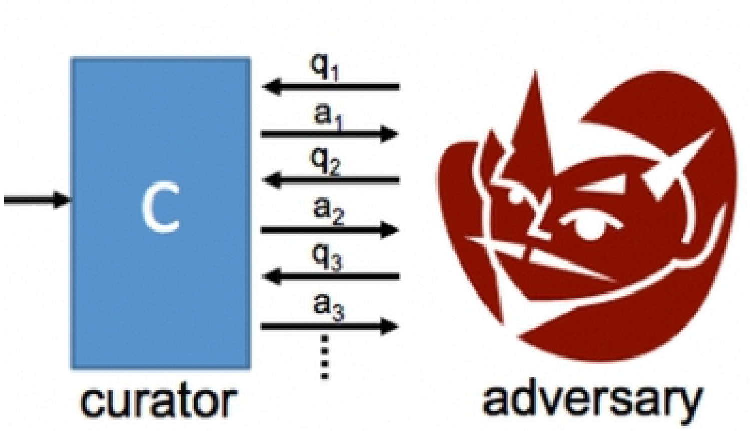
Initially, researchers believed that anonymizing data—erasing the names and replacing them with arbitrary identifiers—was enough to protect the identities and personal information of those who had agreed (knowingly or unknowingly) to contribute information. However, in a well-known study published in 2000, Latanya Sweeney led a team that uncovered the identities of patients, including then Massachusetts governor William Weld, by correlating anonymized data with other publicly available data.

In a more recent case, researchers Arvind Narayanan and Vitaly Shmatikov from The University of Texas at Austin partially de-anonymized a Netflix dataset containing half a million movie reviews. By cross-referencing that dataset with information in the Internet Movie Database, the researchers showed that attackers could potentially identify known users, compromising their data.

As cases of re-identification and de-anonymization emerge, researchers are exploring new, more robust approaches to privacy protection.

Salil Vadhan, a professor of computer science at Harvard University and former director of the Center of Research on Computation and Society, is among the researchers exploring an approach known as "differential privacy" that allows one to investigate data without revealing confidential information about participants. Initially introduced by Cynthia Dwork, Frank McSherry , Kobbi Nissim and Adam Smith, among others, in the mid-2000s, researchers continue to develop the concept today to apply it for real-world problems.

As the lead researcher for the National Science Foundation (NSF) supported "Privacy Tools for Sharing Research Data," Vadhan and his team at Harvard are developing a new computer system that acts as a trusted curator—and identity protector—of sensitive, valuable, data. (The Sloan Foundation and Google, Inc. are providing the project with additional support.)



Through a differentially private interface for data analysis, it is impossible for an adversary to extract information that is specific to one individual, no matter how much other information or computing power it has at its disposal. Credit: Salil

Vadhan, Harvard University

The system works like this: Researchers ask the virtual curator questions based on the data—for instance, "What percentage of individuals who have Type B blood are also HIV positive?" The computer returns an answer that is approximately accurate, but that includes just enough "noise" that no matter how hard someone tries, they cannot find out anything specific to any individual in the database.

"Even if an adversary tries to target an individual in the dataset, the adversary should not be able to tell the difference between the world as it is and one where that individual's data is entirely removed from the dataset," Vadhan said. "Randomization turns out to be very powerful."

If the system is implemented simply, the level of privacy degrades with multiple queries, so one could keep asking questions until the point where identifying people in the database becomes possible. However, by judiciously increasing the amount of noise and carefully correlating it across queries, the system can maintain privacy protection, even in the face of very large number of questions.

Differential privacy has become a hot topic in recent years. A 2015 Science magazine article referred to differential privacy as one of the most promising technical solutions for protecting the data of students enrolled in Massive Open Online Courses (MOOCs). Projects including OnTheMap, used for U.S. census data, RAPPOR, a new product from Google, apply forms of differential privacy for data sharing.

Speaking at the NSF in early 2015, Vadhan explained how ideas from theoretical computer science inspired the development of differential privacy algorithms, which have are now entering the research ecosystem.

Harvard's Institute for Quantitative Social Science is planning to use differential privacy techniques to enable more researchers to share, retain control of, and credit for their data contributions as part of the Dataverse Network, a project that guarantees the long-term preservation of critical datasets.

Unlocked Scientific Potential

Dataverse is the largest public general-purpose research data repository in the world. However, the scientific community could access far more datasets that are currently not publicly available, if differential privacy's promise is fulfilled, according to Gary King, Albert J. Weatherhead III University Professor at Harvard University and Director of the Institute for Quantitative Social Science.

"That's why we're so thrilled to be working on this project," King said. "The social sciences are finally getting to the point in human history where we have enough information to move from studying problems to actually solving them. As we make progress on the privacy problem, we will be able to unlock more and more of the potential of this new information."

The differential privacy tool Vadhan and his team are developing will allow the inclusion of datasets that were previously withheld because the information was too sensitive and privacy was uncertain.

"Currently, Dataverse is not equipped to handle datasets with privacy concerns associated with them," Vadhan said. "If a researcher says that a dataset has identifiable personal information, it is not made available for download."

Differential privacy doesn't work for every type of research question. Vadhan pointed to regression, machine learning, and social network analysis as areas where there are very promising theoretical results, but challenges remain to making differential privacy work well in practice.

Differential privacy also doesn't help when you're looking for identity of a specific individual: as in the case of identifying terrorists or a match for a kidney donor. But that's the point: each individual should be "hidden" even as they contribute to the greater good of any given study.

"This project could significantly enhance the state-of-the-art in privacy," said Nina Amla, a program director at NSF who oversees the award. "They take a highly interdisciplinary approach which brings together deep expertise in computer science, social science, statistics, and law."

According to Vadhan, differential privacy has rich connections with other parts of computer science theory and mathematics.

"It turned out not to just be an island in itself but to be deeply intertwined with other theoretical questions," Vadhan said. "And we're seeing interest from many communities, such as privacy law, medical informatics and social science, to see whether differential privacy can address the privacy problems they think about."

The team hopes to release a preliminary version of their tool for public exploration and feedback this fall and have published their work in the Annals of the American Academy of Political and Social Science and will present their research on differential privacy at many major conferences, including the upcoming 2015 IEEE Symposium on Foundations of Computer Science.

"Our goal in the project is to enable the wider sharing of data while protecting privacy," Vadhan said, "and to make sharing easier for a non-expert researcher without experience in computer science, law or statistics."

Provided by National Science Foundation

Citation: Theoretical computer science provides answers to data privacy problem (2015, October 7) retrieved 18 April 2024 from https://phys.org/news/2015-10-theoretical-science-privacy-problem.html