Data Privacy Detective Podcast – Episode 55: Differential Privacy and Academic Research

Oct 26, 2020

Categories:

Blogs
Podcast
Technology Blog

Authors:

Joseph J. Dehner

Science and knowledge advance through information gathered, organized, and analyzed. It is only through databases about people that social scientists, public health experts and academics can study matters important to us all. As never before, vast pools of personal data exist in data lakes controlled by Facebook, Google, Amazon, Acxiom, and other companies. Our personal data becomes information held by others. To what extent can we trust those who hold our personal information not to misuse it or share it in a way that we don't want it shared? And what will lead us to trust our information to be shared for database purposes that could improve the lives of this and future generations, and not for undesirable and harmful purposes?

Dr. Cody Buntain, Assistant Professor at the New Jersey Institute of Technology's College of Computing and an affiliate of New York University's Center for Social Media and Politics discusses in this podcast how privacy and academic research intersect.

Facebook, Google, and other holders of vast stores of personal information face daunting privacy challenges. They must guard against unintended consequences of sharing data. They will not generally share with and will not sell to academic researchers access to databases. However, they will consider and approve collaborative agreements with researchers that result in providing academics access to information for study purposes. This access can aim to limit access to identifying individuals through various techniques, including encryption, anonymization, pseudonymization, and "noise" (efforts to block users from being able to identify individuals who contributed to a database).

"Differential privacy" is an approach to the issues of assuring privacy protection and database access for legitimate purposes. It is described by Wikipedia as "a system for publicly sharing information about a dataset by describing the patterns of groups

within the dataset while withholding information about individuals in the dataset." The concept is based on the point that it is the group's information that is being measured and analyzed, and any one individual's particular circumstances are irrelevant to the study. By eliminating the need for access to each individual's identity, the provider of data through differential privacy seeks to assure data contributors that their privacy is respected, while providing to the researcher a statistically valid sample of a population. Differentially private databases and algorithms are designed to resist attacks aimed at tracing data back to individuals. While not foolproof, these efforts aim to reassure those who contribute their personal information to such sources that their private information will only be used for legitimate study purposes and not to identify them personally and thus risk exposure of information the individuals prefer to keep private.

"Data donation" is an alternative. This provides a way for individuals to provide their own data to researchers for analysis. Some success has been achieved by paying persons to provide their data or allowing an entity gathering data for research to collect what it obtains by agreement with a group of persons. Both solutions have their limits of protection, and each can result in selection bias. Someone active in an illicit or unsavory activity will be reluctant to share information with any third party.

We leave "data traces" through our daily activity and use of digital technology. Information about us becomes 0's and 1's that are beyond erasure. There can be false positives and negatives. Algorithms can create mismatches, for example a mistaken report from Twitter and Reddit identifying someone as a Russian disinformation agent.

Academia has developed internal review boards (IRB's) that develop and promulgate ethical and other standards for databases. This reduces the risks of unethical use of personal information from third party access. But safeguards are not absolute. And ironically, such standards do not exist for general company use of personal information to the same extent as for use by academics. Individuals will continue to be skeptical about sharing their personal data for the general benefit unless there is a greater belief that their

personal information will not be misused against them. This is true particularly in authoritarian countries, where citizens do not wish to become targets of their government by expressing ideas or comments antagonistic to the public authorities or contrary to local law. Differential privacy is not an antidote in such circumstances.

France's *Systéme national des données de santé* is a successful effort to encourage French residents to post their healthcare information online. This allows those who contribute to have online access to their own health records and for use when needed by others (e.g., at an emergency room). But it also allows for open uses of anonymized data and approved access to pseudonymized data for research and public interest work. This sensitivity to protecting privacy has encouraged substantial sign-up for the program and provides researchers with access to data that can improve public health.

FBT Gibbons