A data acquisition mechanism that maximizes platforms' utility while compensating privacy-sensitive users

December 5 2023

Credit: Pixabay/CC0 Public Domain

The rise of artificial intelligence and machine learning is increasing the demand for data from app users, device owners, firms, consumers, and even patients. As data-hungry technologies are getting more and more

efficient, the key question is how to incentivize data-sharing while protecting users' privacy, said Ali Makhdoumi, an associate professor of decision sciences at Duke University's Fuqua School of Business.

In a new paper to appear in the journal *Operations Research*, Makhdoumi and co-authors Alireza Fallah of University of California, Berkeley, Azarakhsh Malekian of University of Toronto, and Asuman Ozdaglar of Massachusetts Institute of Technology argue that the solution may be in designing a mechanism that measures the privacy sensitivity of users and compensates them for relinquishing personal data. The paper is also available on the pre-print server *arXiv*.

"In many machine learning applications, we treat data as given," Makhdoumi said. "However, typically data is provided by individuals who have their own privacy concerns. So, the question is: how should we compensate these privacy-concerned users? And before answering this question, we need to understand how one can guarantee privacy."'

In their research, Makhdoumi and colleagues use "differential privacy," an approach widely adopted in the tech industry.

This approach involves adding noise to the data, a level of randomization that will make the information less revealing of the person who is sharing it, he said.

He made an example of a company querying hospital records to determine the percentage of individuals within a zip code who have a certain medical condition that may be sensitive.

"Let's say that the dataset shows that 20% of individuals have that condition," Makhdoumi said. "To make it private, the hospital adds noise to the true average so that the response to the query may show that the percentage of individuals with that condition is some random number

between 10% and 30%."'

He also said that privacy is currently delivered either locally or centrally. In the local setting, the data is randomized directly on the user's device, before being shared with the entity that will process it. An example of the localized approach is Apple's privacy settings, which the company amplified in the campaign "What happens on your iPhone stays on your iPhone." In the centralized system, users share the raw data with the companies, which will then add noise to the results.

The local system produces less accurate statistical estimations in business or other types of analysis, Makhdoumi said, because the data is randomized before being analyzed.

"The data is provided by individuals," he said. "So the natural question is how should I compensate those individuals for the data and for their loss of privacy?"

People have different levels of concern with digital privacy, he said, and they value differently the utility they derive from the service the platforms provide.

For example, in a medical setting, users might decide that the societal benefits of scientific research may justify some loss of privacy, Makhdoumi said. In different settings, for example social media, or even in governmental studies, people may have different intrinsic privacy concerns, he said.

In their research, Makhdoumi and co-authors designed a new data acquisition mechanism that considers users' privacy sensitivity, assigns a value to it, and determines the optimal incentive mechanism for data-dependent platforms.

By considering both a price for users' privacy loss and the utility they derive from the service, the mechanism provides the optimal way for a company to collect the data while compensating users for sharing their personal information.

"There are companies already paying users for their data," he said.

The research also shows it is most efficient for platforms to collect data centrally, a setting that ensures the most precise results for business analysis.

Addressing privacy concerns

"Technologies such as AI and machine learning have penetrated society before policymakers and the public could even think about how these tools deal with private information," Makhdoumi said.

As platforms learn details about users' preferences and characteristics, the risk is that they use those data points to enable price discrimination and other manipulations useful for them and harmful for users, he said.

"For instance, if they learn the financial situation of individuals, they might initially offer loans at a low interest rate and then raise it later," he said.

He also said empirical studies showed that some companies are able to predict when consumers are most vulnerable, and then target certain products at that time, when they may feel most attractive to them.

"Depending on your status, they offer you something that, for example, looks glossy, but that in the long run is going to hurt you," he said.

Makhdoumi said this research is a first step in tackling the issue of data

market and privacy, and the benefits and losses of different data architectures for platforms, users, and society as a whole.

"We still have a long way to go in understanding the societal harms of data collection."

More information: Alireza Fallah et al, Optimal and Differentially Private Data Acquisition: Central and Local Mechanisms, *arXiv* (2022). DOI: 10.48550/arxiv.2201.03968

Provided by Duke University