Researchers at the Advanced Digital Sciences Center have dramatically reduced the amount of noise that must be added to sensitive data before it can be published with a provable guarantee of privacy. Their novel approach, called the compressive mechanism (CM), fits into the general framework of differential privacy, which traditionally involves adding noise chosen randomly from a Laplace distribution. Compared to previous approaches to differential privacy, CM achieves the same privacy guarantee with much less noise, making the published data much more accurate than thought possible previously. Further, CM can be computed quickly, so it is appropriate for time-critical situations.

Detailed data is the lifeblood of researchers and practitioners in fields as diverse as urban planning, medical research, and computational advertising. However, privacy concerns prevent the publication of sensitive data such as medical records and web browsing history. One way to address this issue is to “anonymize” the data by removing personal identifiers, such as names, ID numbers, and addresses. But this method is not sufficient protection against attackers with background knowledge or powerful inference tools. For example, in 2006, AOL published 20M Internet searches from 650K users over a 3-month period, with all personal identifiers removed. Reporters discovered some users’ identities by analyzing their search terms, such as person names (e.g., “Arnold”) and addresses (“Lilburn, GA”). A public outcry followed, and the CTO of AOL resigned soon afterwards. Similarly, in 2009 Netflix released 100M anonymized ratings of 17K movies by 500K users, as part of a contest to improve its movie recommendation system. Researchers identified individuals in this dataset by comparing it to users’ reviews in the IMDB online movie database. After a lawsuit was filed and the Federal Trade Commission raised concerns, Netflix cancelled its plans for future ratings releases. On the other hand, researchers still heavily use these valuable data sets to test proposed new search methodologies and recommendation algorithms.

Differential privacy addresses these problems by providing an ironclad bound on an attacker’s chance of correctly guessing whether any particular person’s data was included in the dataset, no matter how many other data records are already known to the attacker. The data publisher controls the level of privacy by setting a parameter $\epsilon$: smaller values mean the attacker has a lower chance of guessing correctly, due to more noise in the published data. Before ADSC invented CM, it was believed that certain lower bounds held for the scale of the noise that had to be added to achieve a given value of $\epsilon$, which limited the practicality of publishing differentially private data. (The scale of a Laplace distribution is proportional to its standard deviation. As the scale increases, so does the magnitude of the noise added to a data record, on average.) CM circumvents these lower bounds with a sophisticated solution based on the recently-proposed theory of compressive sensing, which has found widespread applications in signal processing. Instead of adding noise to the original data, CM first encodes the data as in compressive sensing; then, CM adds noise to the encoded data, decodes the result as in compressive sensing, and publishes it. Because the transformed data are highly compressed, they require much less noise to achieve differential privacy. For many practical applications requiring precise analysis, such an accuracy boost is critical, and CM is currently the only viable candidate for providing differential privacy.

This work is part of ADSC’s A*STAR-funded project on enabling biomedical research with differential privacy. More information is available at adsc.illinois.edu.

About Us
The Advanced Digital Sciences Center (ADSC), established in 2009 and located in Singapore, is a wholly-owned subsidiary of the University of Illinois at Urbana-Champaign.