Cornell University

Department of
Computer Science
Vitaly Shmatikov
Professor
Cornell Tech
111 8th Avenue #302
New York, NY 10011

shmat@cs.cornell.edu

May 17, 2016

Statement for the National Committee on Vital and Health Statistics hearing:

The problem of data privacy cannot be understood and solved in isolation. It is inextricably linked to all the ways in which information about individuals is collected, analyzed, and used in the digital society. The most pressing issue we are facing today is how to protect individuals' privacy and dignity while enabling all the useful services, science, and research made possible by large-scale data analysis – in short, how to balance data privacy and data utility. This challenge with only grow in importance with the emergence of powerful computational methods based on machine learning that are capable of extracting rich information from large volumes of data.

Unfortunately, existing de-identification standards are not informed by the expected uses of the data, nor do they take into account the full spectrum of threats to data privacy. These standards tend to follow a simplistic, "one size fits all" approach, instead of adapting protection to the contemplated use of the data. De-identification is very brittle and provides little, if any protection against sophisticated re-identification and de-anonymization threats, such as inference, correlation, and adversarial machine learning based on the rich sources of fine-grained information about individuals that became publicly available in recent years (in particular, social media).

Experience shows that there is no single set of individual attributes that can be defined once and for all as "sensitive" or "identifying." Any combination of attributes or features can become a quasi-identifier or at least a basis for inferring sensitive information about the individual when combined with public datasets and other external sources of information. For many types of data – including biomedical, clinical, and especially genomic data – the problem is exacerbated by the fact that we don't yet fully understand which parts of the data are sensitive from the privacy perspective, yet the value of the data is likely to grow over time (in contrast to the usual data protection scenarios) and the data of one individual has linkages and correlations to all of his or her blood relatives, presenting a difficult challenge to standard access control.

Effective solutions to data privacy will involve an integrated combination of law and technology. They should also balance restrictions on data collection and access with restrictions on data use. Some inspiration can be drawn from the existing laws in other domains, such as the Fair Credit Reporting Act, that restrict uses of consumer data. In general, privacy protections should focus on preventing harm to individuals and deterring privacy violations, as opposed to attempting to prevent disclosure. At the very least, privacy protection standards must be revised to explicitly acknowledge the existence of public sources of information about individuals and the feasibility of re-identification, algorithmic inference, and indirect leakage of sensitive information.

Design and implementation of privacy protection technologies should be integrated with the design and implementation of data analysis techniques. It is unlikely that a single solution such as de-identification can protect all kinds and types of data while keeping them useful for all conceivable uses. Instead, privacy engineering should become a standard part of the R&D cycle, with appropriate protections developed and/or adapted for each class of big-data technologies, including, in particular, new and existing machine learning systems. Fortunately, recent progress in computer science has yielded promising technologies for certain data uses. For example, some modern privacy protection techniques allow computation of pooled analytics, general statistics, machine-learning models, and other "big-data" tasks in a way that does not reveal too much information about individual data records. A different type of technology operates on personalized "small data" in a protected computation environment, providing useful services to the individual but not leaking his or her information to the outside.

Law, policy, and regulation can help solve privacy problems that are not addressed by technology, such as holding data collectors liable for inappropriate uses of data. Transparency and clear restrictions on data use are important from the consumer protection perspective. Informed consent has a role to play, but care must be taken lest it turns into a meaningless waiver. Furthermore, as companies and organizations become custodians of vast stores of personal data, it is essential that they deploy state-of-art access control and storage protection technologies to defend this data from computer security threats.

In general, technology moves faster than laws and policies. It is important for the regulators to keep abreast of the latest developments in data analytics – such as the emergence of large-scale machine learning – and the concurrent developments in digital privacy, both offensive (new threats and new adversarial techniques) and defensive (such as domain-specific protections for health data analytics and privacy-preserving machine learning). We should not be looking for silver bullets that will solve all data privacy problems now and forever. Instead, we should expect a continuous evolutionary process in which new privacy protection technologies are developed concurrently and in harmony with the new ways of using the data that enhance human well-being.

Sincerely,

[redacted]

Vitaly Shmatikov
Professor
Cornell University

About the writer: *Vitaly Shmatikov is a Professor of Computer Science at Cornell University and Cornell Tech, Cornell's new applied sciences campus in New York City. Professor Shmatikov's research area is computer security and digital privacy. Before joining Cornell, he was a faculty member at the University of Texas at Austin and computer scientist at SRI International. Professor Shmatikov received the PET Award for Outstanding Research in Privacy Enhancing Technologies twice, in 2008 and 2014, and was a runner-up in 2013. Professor Shmatikov's research group won the Best Practical Paper or Best Student Paper Awards at the 2012, 2013, and 2014 IEEE Symposiums on Security and Privacy, as well as the NYU-Poly AT&T Best Applied Security Paper and the Test-of-Time Award from the ACM Conference on Computer and Communications Security. Professor Shmatikov earned his PhD from Stanford University.*