Initial Thoughts for *Accelerating Responsible Sharing of Federal Data*

David Evans, Professor of Computer Science, University of Virginia

16 October 2019

Before discussing the roundtable questions, I want to emphasize the obvious, but important, fact that the best way to preserve privacy is to not collect data in the first place. Similarly to the way the Paperwork Reduction Act requires government collection of data to carefully consider the time cost burden data collection forms impose on the public,[1] the most effective privacy mechanism would limit data collection by requiring all data collection and distribution activities to carefully consider the risks of that data being exposed. Those risks include both the potential for that data to be compromised by adversaries, and what might be exposed through the intended use of that data.

If sensitive data must be collected, the next lowest risk is to limit its exposure as much as possible. A breach like the 2014 Office of Personnel Management incident which leaked records of over 20 million Americans should not be possible for many reasons. The simplest one is that this much sensitive but no longer needed data should never be stored on an on-line system in the first place. Data retention and isolation policies should be in place that limit on-line data to that which is frequently needed, to purge data that is no longer needed, and to move infrequently accessed sensitive data to isolated, off-line systems that can only be accessed via human actions by appropriate authorities.

Roundtable Question: *Which technical approaches have the most promise for near-term use in addressing privacy and security issues for data sharing, and why?*

Computing on Encrypted Data. Several methods have been developed that enable functions to be computed on encrypted data, without ever needing to decrypt that data. At the end of the computation, the encrypted result can be decoded and revealed, but no information about the inputs or any intermediate results of the computation is leaked. There are two main approaches that use encrypted computation to provide data privacy: *outsourced encrypted computation* and *secure multi-party computation* [3].

Outsourced encrypted computation enables a data owner to encrypt its data and send the encrypted data to an external server. Then, the external server can carry out computations on that encrypted data without learning anything to produce encrypted results, which can be sent back to the data owner who can decrypt them to learn the actual values. Fully homomorphic encryption (FHE) is an emerging technology that allows any function to be computed this way. Although there has been rapid progress in FHE over the past decade, it is still several breakthroughs away from being practical for large problems. Instead, scalable outsourced computation solutions use encryption systems that are not semantically secure. This means some information about the data leaks to the external server or adversaries that can observe the network traffic. For example, most encrypted databases use order-preserving encryption which supports efficient search queries by generating ciphertexts that preserve the order of the original data. These methods may be useful for satisfying regulatory requirements, but they leak a lot of information to motivated adversaries (e.g., [7]).

Secure multi-party computation (MPC) enables a group of independent data owners to jointly compute a function on all of their private inputs, without leaking any information about those private inputs. MPC differs from outsourced computation in that all of the protocol participants are data owners who participate together to execute a protocol. MPC is a relatively mature technology, with protocols developed over the past 30 years and numerous systems developed for building MPC applications. Although there have been

[1] In contract, the Privacy Act is nearly all about intentional disclosure of data and mechanisms to protect collected data, but seems to have little to say about the privacy risks of collecting data in the first place.

a handful of interesting deployments of MPC involving multiple data owners, the most successful deployments aim to enhance security for a single data owner. For this use case, the data is split using secret sharing to be stored on multiple hosts. The secret sharing ensures that the information available on any single host is not enough to learn anything about the sensitive data. Using MPC, computations can still be performed on the data by combining each host's share within the encrypted computation. This model can be very efficient, and provides a high level of security when the hosts are selected to be diverse and separately administered since the data is protected unless all of the hosts are simultaneously compromised.

Differential Privacy. Encrypted computation techniques protect data during a computation, but do not ensure privacy since the released output may reveal sensitive data. To know if the output of a function computed on sensitive data can be safely released, we need to understand what an adversary might be able to infer about the sensitive data from that output. Differential privacy is an intuitive and mathematically powerful definition of privacy that gives a probabilistic bound on the amount the output can change based on a change in the input [2]. When used carefully, differentially private mechanisms provide individuals with a bound on the risk of exposure they assume by providing their data. Differential privacy can be applied either locally, adding noise to protect data before it is collected, or centrally, adding noise to outputs computed on centralized data before they are released. Differential privacy can provide strong guarantees of individual privacy for simple data analysis problems, but for complex problems (such as machine learning) there are huge gaps between theory and practice, and between the noise needed to achieve differential privacy guarantees and the noise that can be added without harming accuracy [6].

Roundtable Question: *Which technical challenges are more likely to have to promising near-term solutions?*

The costs of generic secure computation are prohibitive for most problems, but custom solutions for particular problems are often practical and cost-effective. In particular, hybrid approaches that combine partially homomorphic encryption with multi-party computation often enable efficient custom solutions to important problems such as private set intersection functionalities, genomics (e.g., [5]), and regressions (e.g., [4]). There are also opportunities to use MPC to provide strong security for centrally-owned sensitive data, where data is needed infrequently in computation but in ways that are too complex and unpredictable to allow for offline storage solutions.

Roundtable Question: *What are viable approaches the Federal government can take to test or pilot the most promising near-term approaches?*

MPC and several differential privacy mechanisms are mature enough to be suitable for near-term deployment. Two recent exemplars that point to opportunities for future pilots: the Boston wage equity study [1], which used MPC techniques to collect sensitive data from employers to measure gender pay equity; and the plans for using differential privacy for the 2020 Census. The first is a non-federal example of designing a study to collect data using privacy techniques that could otherwise not be collected. I believe there are opportunities for similar approaches to be used for education and medical data, but designing and executing such a study requires deep understanding of the privacy implications and buy-in from all the stakeholders. The second is an example of using privacy techniques to protect what is released about data that is already collected in ways that carefully balance tradeoffs between privacy and data value. For simple data analytics problems, there are good solutions that provide differential privacy guarantees. For more complex problems, finding the right tradeoff between useful data release and sufficient privacy is challenging.

References

[1] Azer Bestavros, Andrei Lapets, and Mayank Varia. *User-Centric Distributed Solutions for Privacy-Preserving Analytics*. Communications of the ACM, 60(2):37-39, 2017.

> This short essay summarizes some of the challenges, but technical and non-technical, in deploying MPC for the Boston wage equity study, and how they were able to use MPC to securely collect and analyze sensitive data.

[2] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. *Differential Privacy — A Primer for the Perplexed*. In Joint UNECE/Eurostat work session on statistical data confidentiality, 2011.

> This is a brief discussion of differential privacy, targeted to dispelling the most common misunderstandings about what differential privacy is and the privacy implications of a system providing differential privacy.

[3] David Evans, Vladimir Kolesnikov and Mike Rosulek. *A Pragmatic Introduction to Secure Multi-Party Computation*. NOW Publishers, December 2018. (available from https://www.securecomputation.org).

> This short book describes the main protocols for secure multi-part computation and the techniques that have been developed to make those protocols efficient enough to be used in practice.

[4] Adrià Gascón, Phillipp Schoppmann, Borja Balle, Mariana Raykova, Jack Doerner, Samee Zahur and David Evans. *Privacy-Preserving Distributed Linear Regression on High-Dimensional Data*. In *Proceedings on Privacy Enhancing Technologies* (PETS). July 2017.

> This is an example of a hybrid protocol, where techniques from homomorphic encryption and multi-party computation are combined to solve a special-purpose problem efficiently.

[5] Karthik A. Jagadeesh, David J. Wu, Johannes A. Birgmeier, Dan Boneh, and Gill Bejerano. *Deriving Genomic Diagnoses without Revealing Patient Genomes*. Science, Vol 357, Issue 6352. 18 August 2017.

> Describes a system using multi-party computation to enable patients to receive diagnoses that depend on their genome, without revealing their genome. In additional to clinical uses like this, there have been methods developed to perform genomics research on encrypted genomic data.

[5] Bargav Jayaraman and David Evans. *Evaluating Differentially Private Machine Learning in Practice*. In *USENIX Security Symposium*. August 2019.

> This research paper describes the differential privacy mechanisms used in machine learning, and experimentally measures the current gaps between what can be learned with reasonable accuracy and what privacy mechanisms can guarantee, and between differential privacy guarantees and what current inference attacks can learn about sensitive data in realistic settings.

[7] Muhammad Naveed, Seny Kamara, Charles V. Wright. *Inference Attacks on Property-preserving Encrypted Databases*. In *22nd ACM SIGSAC Conference on Computer and Communications Security*. 2015.

> This paper demonstrates the vulnerabilities of encrypted computation solutions that do not use semantically secure encryption, showing how much information can be revealed by encrypted databases using semantics-preserving encryption methods like order-preserving encryption.