DOI:10.1145/3029603

Azer Bestavros, Andrei Lapets, and Mayank Varia

Privacy and Security

User-Centric Distributed Solutions for Privacy-Preserving Analytics

How can cryptography empower users with sensitive data to access large-scale computing platforms in a privacy-preserving manner?

For over a year, a high-profile initiative spearheaded by the City of Boston and the Boston Women's Workforce Council (BWWC) strived to identify salary inequities across various employee gender and ethnic demographics at different levels of employment, from executive to entry-level positions.[11] While the effort was supported by a diverse set of more than 100 employer organizations in the city—including major corporations, small businesses, and public/nonprofit organizations—it was stalled by concerns about the confidentiality of the data to be collected in order to calculate aggregate metrics.[2]

A key enabling technology that allowed this effort to move forward was a Web-based application (which can be seen at 100talent.org) that we designed and implemented at Boston University to support the aggregation of sensitive salary data using secure multi-party computation (MPC).[8] This application was used in a first-of-its-kind collaborative effort to compute aggregate payroll analytics without revealing the individual datasets of contributing organizations. This deployment of MPC, which received significant media attention,[2,15] finally enabled the BWWC to conduct their analysis and produce a report presenting their findings.[4]

MPC privately shards users' sensitive data across multiple servers in such a way that analytics may be jointly computed and released while ensuring that (small collections of) servers cannot learn any user's data. Theoretical constructs for MPC have been known for 35 years, with several existing software frameworks designed over the past 10 years.[7,9]

MPC techniques can possess substantial social value: they enable society to benefit from collective data aggregation and analysis in contexts where the raw data is encumbered by legal and corporate policy restrictions on data sharing. Other examples of deploying MPC for social good include tax fraud detection[3] and disease surveillance.[5] Additionally, because MPC decouples computing and networking resources from data, users can leverage the benefits of large data centers without ceding control over their sensitive data.

However, MPC's social benefits cannot be realized unless we empower participating organizations (that is, their executives, directors, and legal advisors) with a clear, confident understanding of exactly how MPC protects their sensitive data and mathematically guarantees compliance with data sharing restrictions. The design and implementation of our own unique MPC platform was informed by nearly two years' worth of discussions with non-technical personnel (including CIOs, CTOs, HR executives, and lawyers from key participating organizations), social scientists, and members of the city council that commissioned the study.[13] These discussions had to take place in meetings and teleconferences where the only aids were whiteboards and slideshows; they involved both describing secret sharing in a concrete, hands-on way as well as providing details of the implementation and how it realized the capabilities and guarantees of this technique. Ultimately, these exchanges were necessary to demystify MPC for decision makers and, more generally, to help us understand and mitigate what we have come to realize are the hurdles that face real-world MPC deployments.

The systems community has grappled recently with the realization that its significant body of work on scalable platforms did not adequately consider the question of what minimum distributed computing configuration outperforms a single thread (COST).[10] Analogously, in this column we argue that the extensive body of MPC research to date has not adequately considered the needs and circumstances of the ultimate users of MPC. Our own experience echoes and confirms thoughts expressed by other researchers in the community:[16] "Secure computation is a general scheme; in reality one has to choose an application, starting from a very real business need, and build the solution from the problem itself choosing the right tools, tuning protocol ideas into a reasonable solution, balancing security and privacy needs vs. other constraints: legal, system setting, etc." We draw from our experience to advocate for the design of platforms that address concerns along Usability, Scalability, Entrustment, and Risk (USER) dimensions.

Usability

To meet the needs of our users, we rejected the most algorithmically expressive MPC solutions available in the literature.[7] Instead, we found that what we needed was the simplest of protocols: just expressive enough for the application at hand while being comprehensible enough to fuel adoption among corporate officers, legal representatives, and rank-and-file employees. We also found that participants' software platform and IT infrastructure inflexibilities and limitations (legacy systems, restrictive policies, firewalls, and so on) required the most lightweight solution: a simple browser-based application that could accommodate the familiar look and feel of a spreadsheet, with transparent open source code to enable outside auditing. Finally, our MPC protocol needed to accept contributors' data asynchronously to simplify coordination and idempotently to allow contributors to fix errors.

Usable MPC is an enabling technology with substantial potential for social good, but only if enough participants are willing to contribute toward the analysis. In the pay equity scenario, the usability of both the protocol and its implementation helped decision makers—after only a few conversations—gain confidence in their understanding of the technology, appreciate that it would impose no significant burdens on their staff and infrastructure, and assured that features such as idempotence and asynchrony would make deployment logistically feasible and likely to produce meaningful results. This, in turn, increased the willingness of participants to contribute their sensitive data.

Usability also extends to the specification of policies governing proper uses of data. Existing MPC frameworks neglect to address privacy policies, in part because the policy may not be expressible by either the original data contributor (who may lack expertise in privacy-related matters) or the analyst (who doesn't know the users' preferences or other uses of the data). Existing techniques from the programming languages research and formal methods communities such as policy-agnostic programming (in which the policies that govern inputs are specified independently from the dataflows and logic of the algorithm), as well as static analysis (to automatically derive policies from algorithms and compare them to user-specified policies) can play a significant role in validating whether an analytic is compatible with a specified privacy policy.

Scalability

Typically, MPC frameworks are evaluated based on their computational efficiency for simple analytics over relatively small datasets. This is a situation in which all modern frameworks perform rather well (that is, seconds to minutes).[1]

However, human time dominates computing time in scenarios involving small-scale data such as the pay equity effort, in which a window spanning multiple days may be required to collect salary data from a large number of contributors operating according to incompatible schedules, rendering the computing time negligible by comparison. In this case, MPC frameworks should prioritize software development and IT infrastructure design over the speed of computing the analytic. At the other extreme, when aggregating large-scale datasets, an MPC framework should optimize the computation that can be performed locally so as to minimize the costs incurred due to MPC.

To resolve both challenges, we have integrated existing MPC frameworks into the Musketeer big data workflow manager.[6] Whereas prior MPC frameworks require that software engineers design analytics in a domain-specific language, we permit rapid development in the well-known SQL and MapReduce paradigms, with automated generation of code to execute in existing back-end distributed frameworks like Hadoop, Spark, or Naiad so that developers and administrators can "focus on the what rather than the how of security."[12] Additionally, our framework automatically infers when sensitive data crosses trust boundaries in order to minimize usage of MPC. We tested this system to compute a market concentration metric over 160GB of public NYC taxi trips' fare information with just 8.3% overhead over the corresponding insecure computation.[14]

Entrustment

At its heart, MPC permits a federation of trust among several computing entities such that each user only needs to trust that any one of them (or a small fraction) is honest. Most existing MPC research papers and software frameworks envision homogeneous entities. By contrast, we design a more flexible MPC framework that allows contributors to entrust entities with different responsibilities.

Along these lines, we provide a taxonomy of roles for entities that participate in MPC: a large, potentially a priori unknown number of *contributors* with private data; an *analyzer* who specifies an analytic; a publicly accessible *service provider* who collects encoded data from the contributors without requiring them to be online simultaneously and who also participates in the distributed computation; additional servers who participate in the distributed computation; one or more repositories that host the secure computing software; and the recipients of the analysis. Behind the scenes, there may also be privacy experts and software engineers who assemble one or more of the components in this ecosystem. In practice, parties using MPC may take on several of these roles simultaneously.[a] MPC provides the recipients with the results of the analytic over the contributors' data, and it provably guarantees that nobody learns anything else.

Just as each entity has different assignments, so too might they have different levels of trust in one another. For brevity, we focus here on the service provider, who must connect to all oth-

[a] Some readers may be familiar with a related technology: fully homomorphic encryption (FHE). Abstractly, FHE can be viewed as a specialization of MPC to the two-party outsourcing setting in which the contributor, analyzer, and recipient are the same party and in which the service provider's computation does not require interaction.[1,7]

The empowering and enabling aspects of MPC will make substantial contributions to data-driven analysis and policymaking.

er entities and may require immense computing power. When both of these characteristics simultaneously apply, the service provider has a large attack surface and is well suited to being run within a cloud computing datacenter.

Our pay equity software enables the most powerful computing entity also to be the *least trusted*. Our service provider runs on Amazon Web Services to collect and store encoded data; however, contributors can choose instead to entrust the BWWC to protect the confidentiality of their data. We envision a future in which cloud providers offer 'secure computing-as-a-service' deployments of MPC that decouple control over data from computing power.

Risk

MPC research studies four types of adversaries: semi-honest entities who execute software as provided but may attempt to glean information along the way, covert adversaries who cheat only if they are unlikely to be caught, rational adversaries who cheat as long as the expected payout is larger than the expected penalty if caught, and fully malicious entities who perform any action necessary to breach the confidentiality or integrity of honest users.

We advocate for the MPC community to match cryptographic models of adversarial behavior with the economic (for example, reputation-based) and legal incentives that real-world users face. A more accurate and fine-grained characterization of risks can result in a faster, simpler MPC protocol that satisfies users' needs. Our pay equity project exposed delicate economic and legal concerns whose impact upon risk models should be explored further.

First, the existing risk models fail to capture the subtlety of reputation-based economic incentives. In the pay equity scenario, the analyzer and repository have the capacity to alter the software to leak secrets; however, they should not execute this capability due to the long-term damage to their reputation and economic viability. Analogously to the differences between the oneshot and iterated prisoner's dilemma games, the rational model of MPC provides an incomplete view because it focuses on a single execution.

Second, MPC has a complex interconnection with the law. In our pay equity scenario, even if the BWWC could somehow learn the contributors' data by cheating, it has a strong legal incentive not to acquire this data because it could then be exposed to lawsuits. Indeed, one of the major hurdles that faced BWWC prior to their use of our solution was the unwillingness of any single entity (including a major local university, originally enlisted to perform the study) to assume the liability in case of leakage or loss of data entrusted to them. Moreover, following MPC honestly may provide BWWC legal protections afforded by following best practices or by restricting data sharing. Hence, the BWWC has a strong legal incentive to act in a semi-honest manner. Conversely, appropriately written legal contracts can enshrine MPC's constraints (for example, operating in the best interest of another entity, or forbidding collusion between entities) with enforceable civil penalties. We propose a greater examination of the implications of the law upon MPC and vice versa.

Conclusion

We are convinced that the empowering and enabling aspects of MPC will make substantial contributions to data-driven analysis and policymaking by enabling individuals and organizations at all levels to derive insights about their collective data without requiring that they share that data, but only if the technology is accessible both conceptually and technologically to a broad audience. In this column, we proposed a four-pronged research agenda to make MPC more usable along a variety of dimensions, increase its scalability for humans and computers alike, assign responsibilities that align with existing trust relationships, and systematically understand the legal and economic risks when trust is violated. These recommendations are informed by our prior work deploying MPC to aggregate wage data and compute pay equity metrics—work that is, in the words of BWWC co-chair Evelyn Murphy, "beginning to show how to use sophisticated computer science research for public programs."[15]

References

1. Archer, D.W. Maturity and performance of programmable secure computation. *IACR Cryptology ePrint Archive*, (1039), 2015.
2. Barlow, R. Computational thinking breaks a Logjam: Hariri Institute helps address Boston's male-female pay gap. (Apr. 27, 2015); *BU Today*.
3. Bogdanov, D. *How the Estonian Tax and Customs Board Evaluated a Tax Fraud Detection System Based on Secure Multi-party Computation.* Springer, Berlin, Heidelberg, 2015, 227–234.
4. Boston Women's Workforce Council Report 2016; http://bit.ly/2iR2KhW
5. El Emam, K. A secure protocol for protecting the identity of providers when disclosing data for disease surveillance. *Journal of the American Medical Informatics Association 18*, 3 (May 2011), 212–217.
6. Gog, I. Musketeer: All for one, one for all in data processing systems. In *Proceedings of the Tenth European Conference on Computer Systems* (EuroSys), (2015), 2:1–2:16.
7. Hamlin, A. Cryptography for big data security. In Fei Hu, Ed., *Big Data: Storage, Sharing, and Security*. CRC Press, May 2016.
8. Lapets, A. *Secure Multi-Party Computation for Analytics Deployed as a Lightweight Web Application.* Technical Report BUCS-TR-2016-008, CS Dept., Boston University, July 2016.
9. Lindell, Y. and Pinkas, B. Secure multiparty computation for privacy-preserving data mining. *The Journal of Privacy and Confidentiality 1* (2009), 59–98.
10. McSherry, F. Scalability! But at what COST? In *Proceedings of the 15th Workshop on Hot Topics in Operating Systems (HotOS XV)*, Kartause Ittingen, Switzerland (May 2015). USENIX Association.
11. 100% Talent: The Boston Women's Compact; http://bit.ly/YWryu2
12. Shen, E. Cryptographically secure computation. *IEEE Computer 48*, 4 (2015), 78–81.
13. Signers of 100% Talent: The Boston Women's Compact; http://bit.ly/2bN48QJ
14. Volgushev, N. DEMO: Integrating MPC in big data workflows. In *Proceedings of CCS 2016: The 23rd ACM SIGSAC Conference on Computer and Communications Security*, 2016.
15. Will Data Help Close The Gender Pay Gap?; http://wbur.fm/2hdbjXa
16. Yung, M. From mental poker to core business: Why and how to deploy secure computation protocols? In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS '15*, ACM, 2015.

Azer Bestavros (best@bu.edu) is Professor of Computer Science and Founding Director of the Hariri Institute for Computing at Boston University.

Andrei Lapets (lapets@bu.edu) is a Research Scientist and Director of Research Development at the Hariri Institute for Computing at Boston University.

Mayank Varia (varia@bu.edu) is Research Scientist and Co-Director of the Center for Reliable Information Systems and Cyber Security at the Hariri Institute for Computing at Boston University.

This material is based upon work supported by the National Science Foundation under Grants No. 1414119 and 1430145.