Design and Deployment of Usable, Scalable MPC

Azer Bestavros, Andrei Lapets, Frederick Jansen, Mayank Varia, and Nikolaj Volgushev (Boston University)
Malte Schwarzkopf (MIT)

I. Introduction

Secure multi-party computation (MPC) can possess substantial social value: it enables companies, government agencies, and other organizations to benefit from collective data aggregation and analysis in contexts where the raw data are encumbered by legal and corporate policy restrictions on data sharing. Theoretical constructs for MPC have been known for decades [1]–[3], and the past few years have seen several successful deployments of MPC [4]–[6] along with the development of a number of software frameworks [7]–[9] that aim to deliver MPC's benefits to end-users.

However, MPC's social benefits cannot be realized unless we design MPC systems whose security and functionality can be comprehended by the executives, directors, and legal advisors of participating organizations, and we automate the secure compilation of legacy applications that were developed in well-known programming paradigms. This article summarizes two efforts undertaken by our research team to improve the usability and scalability of MPC.

II. Usable MPC for End-Users

We describe here a 2-year effort to design and deploy an accessible MPC system computing aggregate pay equity metrics by gender and ethnicity for the Boston Women's Workforce Council (BWWC), based upon Equal Employment Opportunity Commission wage data reporting from 40-70 companies in the Greater Boston Area. Compensation data must to be collected from privately held companies in order to calculate an aggregate statistic (sum) over the data. Each company submits employee earnings aggregated by gender and job category. BWWC may view the aggregate totals across *all* companies, but individual company numbers must remain private for both privacy and legal reasons.

MPC Deployment: We implemented and deployed an MPC protocol as a web-based service so that no new software needs to be installed within corporate environments. The user interface provides a familiar spreadsheet that can be filled with data manually or via copy-paste. We successfully deployed this service twice (in 2015 and 2016) to analyze compensation data from a collection of 40–70 employer organizations [10]. The client application can be viewed at http://100talent.org. Practical deployment difficulties included browser and OS compatibility, minimizing human error that might skew the analytics, and scheduling data collection within a 1-2 week time window.

Roles: We consider three roles in the deployed protocol: (1) an unknown quantity of *contributors* who contribute private data for the calculation; (2) an automated, publicly-accessible *service provider* that sees only encrypted data and connects all other participants without requiring them to maintain servers (or even to be online simultaneously); and (3) one or more *analyzers* who receive the output of the analytic. Several *security* and *usability* considerations drove protocol design and implementation.

Security: We rely on MPC with passive (semi-honest) security and without collusion [11]. This suffices in our scenario because the service provider and analyzer lack incentives to falsify the results or to learn private inputs: completing the study successfully is directly beneficial to BWWC (as the study initiator) and to BU (as an institution reliant upon a reputation of integrity). Additionally, obtaining any private contributor data (by colluding or actively deviating from the protocol) creates a liability risk for the service provider and analyzer. The semi-honest model protects service providers from the legal risks of processing sensitive data so long as the protocol is followed.

Usability: A secure MPC protocol only has value if multiple parties trust it and use it. The pay equity scenario involves individuals with a wide range of technical backgrounds utilizing computing resources that are outside of our control and governed by a variety of organizational constraints. Thus, our protocol and web service must satisfy many usability goals: comprehensibility (to drive adoption); transparency (open-source code); easy deployability (no specialized software, hardware, synchronization, or continuous network access); idempotent resubmission; input validation in the client interface; simplicity to describe the security guarantees to decision-makers (Fig. 1 depicts the picture that we showed to managers, lawyers, and human resources employees); and others [12].

III. Automated MPC for Software Engineers

Big data analytics enable companies and regulatory agencies to draw vital insights, especially when they are executed across data sets from multiple sources. However, justified privacy concerns related to data sharing inhibit computing analytics across multiple competing organizations, even if the result would serve a common interest.

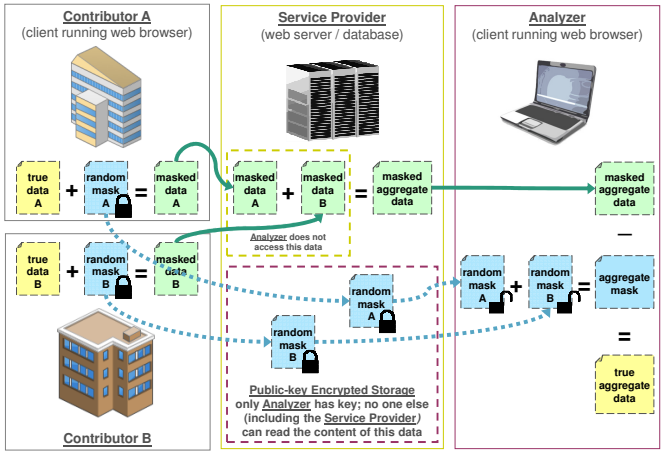

Fig. 1. Diagram of protocol deployment for two contributors used to explain the protocol to potential participants.

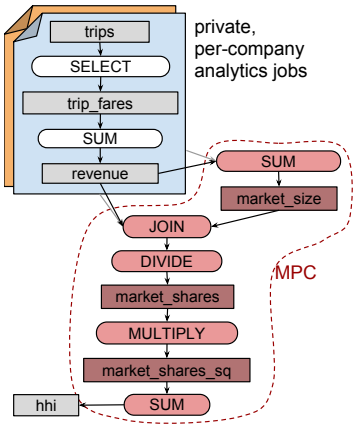

Fig. 2. Vehicle-for-hire market concentration workflow. Gray boxes are tables, rounded nodes are operators. The red, shaded operations happen in MPC, and arrows crossing the MPC boundary correspond to private inputs.

MPC can resolve these privacy concerns at a cost. Note that many real-world use-cases only necessitate MPC for a few crucial operations as part of a larger workflow. While some MPC frameworks support a *mixed-mode*[1] operation that combines local computation with secure, distributed MPC steps [14], existing MPC frameworks still face three key challenges in real-world use: poor integration with widely-used analytics workflows and data processing systems, significant expert knowledge required to implement and execute analytics, and poor scalability to large data sets because they do not support efficient data-parallel processing outside MPC.

Conclave: Our Conclave system addresses these three challenges by judiciously augmenting legacy software with MPC. By doing so, MPC can be made viable for societally important use cases that involve large data sets, such as bank stress tests and early detection of market oligopolies.

In more detail, Conclave adds support for MPC to the Musketeer big data workflow manager [15], whose key premise is to decouple the specification of data-parallel workflows in a high-level frontend language (such as SQL or MapReduce) from their execution in a parallel backend execution engine (such as Hadoop, Spark, and Naiad). With our extensions, Musketeer generates MPC code automatically from programs specified in SQL, requiring no expert knowledge. It also automatically embeds the MPC into larger workflows that involve private processing steps on multiple organizations' heterogeneous data analytics clusters. Even if organizations use different data processing stacks, we automatically generate both the preprocessing code and the "glue code" for embedding MPC in the workflow.

[1]The "mixed-mode" term is overloaded: sometimes, it is used to mean a combination of different *types* of MPC (e.g., arithmetic MPC based on secret sharing [13] and boolean MPC based on garbled circuits [1]). Our system can also support the latter.

Automation: Conclave provides usability by automating three features: *code generation* so that participating parties need no in-house MPC implementation or deployment expertise, *portability* so that different companies can map a high-level joint computation to their individual existing data analytics stacks (e.g., Hadoop, Spark, Naiad) and have these systems automatically feed data into the MPC, and *framework choice* to pick between different MPC backends (e.g., VIFF, Sharemind) to use in a particular application based on Musketeer's performance models of each relational operator in these frameworks.

Prototype use case: We deployed Conclave to compute the Herfindahl-Hirschman Index (HHI) [16], a standard measure of market concentration, over the market shares of several vehicle-for-hire (VFH) companies. The HHI is the sum of squared market shares. MPC allows this computation to be performed without market participants having to disclose their revenue composition. This computation, for example, might allow a regulator to assess the long-term impact of a changing market environment – such as the emergence of "ride-sharing" services such as Uber and Lyft – on market concentration. The workflow proceeds as shown in Fig. 2. We stress that Conclave determines the boundaries of MPC automatically.

Our evaluation uses six years of public NYC taxi trips' fare information [17] as input data, dividing the data across five imaginary taxi companies with 50%, 20%, 10%, 10%, and 10% market shares. We compare the runtime of our market share computation to two extremes. First, Conclave takes only 8.3% longer (17.5 minutes versus 16.2 minutes) than an insecure baseline in which the entire computation is executed at a trusted third party. Second, if the entire computation were executed in the VIFF MPC framework, it would not have finished after two hours, and it required substantial MPC expertise to implement. Conclave is open-source software and available at https://github.com/hicsail/Musketeer.

Acknowledgments

We would like to acknowledge all the members of the Boston Women's Workforce Council, and to thank in particular MaryRose Mazzola, Christina M. Knowles, and Katie A. Johnston, who led the efforts to organize participants and deploy the protocol as part of the 100% Talent: The Boston Women's Compact [18], [19] data collections. We also thank the Boston University Initiative on Cities (IOC), and in particular Executive Director Katherine Lusk, who brought this potential application of secure multi-party computation to our attention. The BWWC, the IOC, and several sponsors contributed funding to complete this work. Support was also provided in part by Smart-city Cloud-based Open Platform and Ecosystem (SCOPE), an NSF Division of Industrial Innovation and Partnerships PFI:BIC project under award #1430145, and by Modular Approach to Cloud Security (MACS), an NSF CISE CNS SaTC Frontier project under award #1414119. Finally, we must thank the many representatives from the participating employer organizations for their interest, diligence, and feedback throughout the orientation, testing, and deployment processes.

References

[1] A. C. Yao, "Protocols for secure computations," in *Proceedings of the 23rd Annual Symposium on Foundations of Computer Science*, ser. SFCS '82. Washington, DC, USA: IEEE Computer Society, 1982, pp. 160–164. [Online]. Available: http://dx.doi.org/10.1109/SFCS.1982.88

[2] O. Goldreich, S. Micali, and A. Wigderson, "How to play any mental game or A completeness theorem for protocols with honest majority," in *Proceedings of the 19th Annual ACM Symposium on Theory of Computing, 1987, New York, New York, USA*. ACM, 1987, pp. 218–229.

[3] M. Ben-Or, S. Goldwasser, and A. Wigderson, "Completeness theorems for non-cryptographic fault-tolerant distributed computation (extended abstract)," in *Proceedings of the 20th Annual ACM Symposium on Theory of Computing, May 2-4, 1988, Chicago, Illinois, USA*. ACM, 1988, pp. 1–10.

[4] P. Bogetoft, D. L. Christensen, I. Damgård, M. Geisler, T. Jakobsen, M. Krøigaard, J. D. Nielsen, J. B. Nielsen, K. Nielsen, J. Pagter, M. Schwartzbach, and T. Toft, "Financial cryptography and data security," R. Dingledine and P. Golle, Eds. Berlin, Heidelberg: Springer-Verlag, 2009, ch. Secure Multiparty Computation Goes Live, pp. 325–343. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-03549-4_20

[5] I. Damg?rd, K. Damg?rd, K. Nielsen, P. S. Nordholt, and T. Toft, "Confidential benchmarking based on multiparty computation," Cryptology ePrint Archive, Report 2015/1006, 2015, http://eprint.iacr.org/.

[6] D. Bogdanov, L. Kamm, B. Kubo, R. Rebane, V. Sokk, and R. Talviste, "Students and Taxes: a Privacy-Preserving Study Using Secure Computation," *PoPETs*, vol. 2016, no. 3, p. 117?135, 2016. [Online]. Available: http://www.degruyter.com/view/j/popets.2016.2016.issue-3/popets-2015-0019/popets-2016-0019.xml

[7] "VIFF, the Virtual Ideal Functionality Framework," http://viff.dk/, [Accessed: August 15, 2015].

[8] D. Bogdanov, S. Laur, and J. Willemson, "Sharemind: A Framework for Fast Privacy-Preserving Computations," in *Proceedings of the 13th European Symposium on Research in Computer Security - ESORICS'08*, ser. Lecture Notes in Computer Science, S. Jajodia and J. Lopez, Eds., vol. 5283. Springer Berlin / Heidelberg, 2008, pp. 192–206.

[9] M. Keller, P. Scholl, and N. P. Smart, "An architecture for practical actively secure mpc with dishonest majority," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, ser. CCS '13. New York, NY, USA: ACM, 2013, pp. 549–560. [Online]. Available: http://doi.acm.org/10.1145/2508859.2516744

[10] R. Barlow, "Computational Thinking Breaks a Logjam," http://www.bu.edu/today/2015/computational-thinking-breaks-a-logjam/, [Accessed: August 15, 2015].

[11] O. Goldreich, *The Foundations of Cryptography - Volume 2, Basic Applications*. Cambridge University Press, 2004.

[12] A. Lapets, N. Volgushev, A. Bestavros, F. Jansen, and M. Varia, "Secure Multi-Party Computation for Analytics Deployed as a Lightweight Web Application," CS Dept., Boston University, Tech. Rep. BUCS-TR-2016-008, July 2016. [Online]. Available: http://www.cs.bu.edu/techreports/pdf/2016-008-mpc-lightweight-web-app.pdf

[13] A. Shamir, "How to share a secret," *Communications of the ACM*, vol. 22, no. 11, pp. 612–613, 1979.

[14] A. Rastogi, M. A. Hammer, and M. Hicks, "Wysteria: A programming language for generic, mixed-mode multiparty computations," in *Proceedings of the 2014 IEEE Symposium on Security and Privacy*, ser. SP '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 655–670. [Online]. Available: http://dx.doi.org/10.1109/SP.2014.48

[15] I. Gog, M. Schwarzkopf, N. Crooks, M. P. Grosvenor, A. Clement, and S. Hand, "Musketeer: all for one, one for all in data processing systems," in *Proceedings of the 10th ACM European Conference on Computer Systems (EuroSys)*, Apr. 2015.

[16] A. O. Hirschman, "The paternity of an index," *The American Economic Review*, vol. 54, no. 5, pp. 761–762, 1964. [Online]. Available: http://www.jstor.org/stable/1818582

[17] T. W. Schneider, "NYC taxi trip data," https://github.com/toddwschneider/nyc-taxi-data, accessed 03/08/2016.

[18] "100% Talent: The Boston Women's Compact," http://www.cityofboston.gov/women/workforce/compact.asp, [Accessed: August 15, 2015].

[19] "Boston: Closing the Wage Gap," http://www.cityofboston.gov/images_documents/Boston_Closing%20the%20Wage%20Gap_Interventions%20Report_tcm3-41353.pdf, [Accessed: August 15, 2015].