Federal Register Notice 87 FR 35250, https://www.federalregister.gov/documents/2022/06/09/2022-12432/request-for-information-on-advancing-privacy-enhancing-technologies, June 9, 2022

Request for Information (RFI) on

Advancing Privacy Enhancing Technologies

Peisert, Sean

Privacy-Preserving Data Sharing for Science and Public Policy

Dr. Sean Peisert
Senior Scientist,
Lawrence Berkeley National Laboratory

July 14, 2022

1 Introduction

Data useful to science, public policy, healthcare, and other vital functions in the national interest is not shared as much as it should or could be, particularly when that data contains sensitivities of some kind. We advocate the use of hardware *trusted execution environments (TEEs)* and *differential privacy* as means to significantly change approaches to and trust relationships involved in secure data management [1]. There are many reasons why data may not be shared, including laws and regulations related to personal privacy or national security, or because data is considered a proprietary trade secret. Examples of this include electronic health records, containing protected health information (PHI); IP addresses or data representing the locations or movements of individuals, containing personally identifiable information (PII); the properties of chemicals or materials, and more. Two drivers for this reluctance to share, which are duals of each other, are concerns of data owners about the risks of sharing sensitive data, and concerns of providers of computing systems about the risks of hosting such data. As barriers to data sharing are imposed, data-driven results are hindered, because data isn't made available and used in ways that maximize its value.

And yet, as emphasized widely in scientific communities [2, 3], by the National Academies, and via the U.S. Government's initiatives for "responsible liberation of Federal data," finding ways to make sensitive data available is vital for advancing scientific discovery, public policy, and other important functions. When data isn't shared, certain research may be prevented entirely, be significantly more costly, take much longer, or might simply not be as accurate because it is based on smaller, potentially more biased datasets.

Computing systems used for data analysis today include institutional computing resources and commercial clouds, and, for certain functions, supercomputers such as those present in high-performance computing (HPC) centers sponsored by U.S. Department of Energy's Office of Science and the U.S. National Science Foundation. Not all data analysis is large, but at the largest scale, it can be characterized by massive datasets and distributed, international efforts to analyze that data. However, when sensitive data is used, computing options available are much more limited in computing scale and access [4].

2 Limitations of Current Privacy Approaches

2.1 Current Secure Computing Environments

Today, where remote access to sensitive data is permitted at all, significant technical and procedural constraints may be put in place, such as instituting ingress / egress "airlocks," requiring "two-person" rules to move software and data in or out, and requiring the use "remote desktop" systems. Architectures like this are becoming more and more common as means for computing involving sensitive data [4]. However, even with these security protections, traditional enclaves still require implicitly trusting system administrators and anyone with physical access to the system containing the sensitive data, thereby increasing the risk to and liability of an institution for accepting responsibility for hosting data. This security limitation can

significantly weaken the trust relationships involved in sharing data, particularly when groups are large and distributed. These concerns can be partially mitigated by requiring data analysts to be physically present in a facility owned by the data provider in order to access data. However, in all these cases, analysis is hindered for communities – such as scientific communities — whose abilities and tools are optimized for working in open, collaborative, and distributed environments. Further, consider the pandemic in which a requirement of physical presence in a particular facility for analysis would be a significant public health risk at various times.

2.2 Reducing Data Sensitivity Using "Anonymization" Techniques

Sometimes attempts are made to avoid security requirements by making data less sensitive by applying "anonymization" processes in which data is masked or made more general. Examples of this approach remove distinctive elements from datasets such as birthdates, geographical locations, or IP network addresses. Indeed, removing 18 specific identifiers from electronic health records satisfies the HIPAA Privacy Rule's "Safe Harbor" provisions to provide legal de-identification. However, on a technical level, these techniques have repeatedly been shown to fail to preserve privacy, typically by merging external information containing identifiable information with quasi-identifiers in the dataset to re-identify "anonymized" records [5]. Therefore, de-identification doesn't necessarily address the risk and trust issues involved in data sharing because re-identification attacks can still result in significant embarrassment, if not legal sanctions. In addition, the same masking used in these processes also removes data that is critical to the analysis [5]. Consider public health research for which the last two digits of a zip code, or the two least significant figures of a geographic coordinate are vital to tracking viral spread.

3 Confidential Computing

Hardware TEEs can form the basis for platforms that provide strong security benefits while maintaining computational performance [6]. TEEs are portions of certain modern microprocessors that enforce strong separation from other processes on the CPU, and some can even encrypt memory and computation. Common commercial TEEs today include ARM's TrustZone, Intel's Secure Guard Extensions (SGX), and AMD's Secure Encrypted Virtualization (SEV). All three vendors take extremely different approaches and have extremely different strengths, weaknesses, use cases, and threat models.

TEEs can be used to maintain or even increase security over traditional enclaves, at minimal cost to performance in comparison to computing over plaintext. TEEs can isolate computation, preventing even system administrators of the machine in which the computation is running from observing the computation or data being used or generated in the computation, including even from certain "physical attacks" against the computing system. They can implement similar functionality as software-based homomorphic and mutiparty computation [7] approaches, but without the usability issues and with dramatically smaller performance penalties.

The use of TEEs to protect against untrustworthy data centers is not a novel idea, as seen by the creation of the Linux Foundation's Confidential Computing Consortium [8] and Google's recent "Move to Secure the Cloud From Itself." [9]. Google has comparing the importance of the use of TEEs in its cloud platform to the invention of email [10]. However, TEEs have not yet seen broad interest and adoption in data analysis, although they are now present in Amazon, Google, and Microsoft's cloud computing platforms.

The approach we envision is to leverage TEEs when data processing environments are out of the direct control of the data owner, such as in third-party (including DOE or NSF) HPC facilities or commercial cloud environments, in order to prevent exposure of sensitive data to other users of those systems or even the administrators of those systems. Data providers can specify the configuration of the system, even if they are not directly the hosts of the computing environment, to specify access control policies, a permitted list of software or analyses that can be performed, and output policies to prevent data exfiltration by the user. The notion of being able to leverage community HPC and cloud environments also enables the use

of data from multiple providers simultaneously while protecting the raw data from all simultaneously, each potentially with their own distinct policies.

Researchers at the Berkeley Lab and UC Davis empirically evaluated Intel SGX and AMD SEV TEEs for their performance under typical HPC workloads including both traditional modeling and simulation benchmarks, ML/AI benchmarks, and real-world ML/AI applications. Our results [11] show that AMD's SEV generally imposes minimal performance degradation for single-node computation and represents a performant solution for high-performance computing — including large-scale data applications — with lower ratios of communication to computation. Importantly, the major commercial clouds, as well as modern HPC centers, such as the the DOE's NERSC-9, contain AMD processors that support the SEV TEE, and thus it is our hope that our results will provide some of the evidence needed to justify the use of TEEs in large-scale, data-driven computing.

4 Research Opportunities to Advance Privacy-Enhancing Technologies

4.1 Trusted Execution Environments

Although numerous commercial TEEs exist, no TEEs yet exist in processors other than CPUs, such as in GPUs and accelerators, although NVIDIA has indicated plans to expand TEEs to GPUs, and Google has indicated plans to expand TEEs to GPUs, TPUs, and FPGAs. There are also issues with low-latency communication between TEEs, and also the cost of virtualization, that must be addressed to enable secure data analysis and machine learning scale [11]. In addition, promising RISC-V efforts such as Keystone [12] exist that carry both the promise of broadening the scope of processors that contain TEEs, while also being open source and possible to formally verify. However, RISC-V based TEEs have not yet been developed that target algorithms that center around data, such as large graph workloads and machine learning. Most likely, an entirely new TEE architecture tailored for scientific computing and data analysis applications will be needed, which is a focus of Berkeley Lab's efforts in this space [13, 14].

4.2 Differential Privacy

Output policies are another area that deserve investigation. While TEEs protect against untrusted computing providers, and can provide certain measures of protection from malicious users, output policies determine what data is returned to the user. Differential privacy [15] is a particularly interesting approach to providing strong privacy protection of data output. Differential privacy is a statistical technique that can guarantee the bounds on the amount of information about a dataset that can be leaked to a data analyst as a result of a query or computation by adding "noise" and enforcing a "privacy budget" that bounds information leakage. It is now a mainstream solution, with production use by Apple, Google, and the U.S. Census Bureau, the existence of several open source distributions, and successful application to a diverse range of data types. However, differential privacy is not appropriate everywhere, and applying it is currently challenging, requiring a high degree of expertise and effort. Thus, differential privacy is highly useful today, albeit in a limited set of situations for datasets that have sufficiently wide use to justify the time and expense required. Work is needed to advance the usability of differential privacy so it can more easily be broadly leveraged. Energy data and mobility studies are two areas that Berkeley Lab has demonstrated success in applications of differential privacy [16, 17].

5 Summary and Next Steps

In contrast to traditional secure enclaves, TEEs enable sensitive data to be leveraged without having to trust system administrators and computing providers. However, while the application of TEEs have now been widely heralded in cloud environments, they have not advanced to be performant for large-scale data

analysis, despite the significant concerns frequently expressed by both data providers and computing facilities about hosting sensitive data. But while improvements are needed to truly harness TEEs for large-scale data applications, the current generation of TEEs is here, those TEEs are available, and until we start making use of them in scientific computing, data is not shared as much as it should or could be by leveraging TEEs to address the trust issues underlying current limits on data sharing.

What is missing is a connection to the particular infrastructure used in large-scale data-driven computing, including I/O subsystems, custom workflows, highly specialized instruments, community data repositories, and so on. Therefore what is needed is a conversation between processor manufacturers, system vendors, and cloud and scientific computing operators regarding enabling the TEE functionality already present in the AMD EPYC processors — and presumably in other, future processors — into scientific computing environments, while simultaneously developing and preparing for the next generation of TEEs. However, the path forward is not solely technical. It requires the community to build infrastructure around TEE technology and integrate that infrastructure into scientific computing facilities and workflows, and into the mindset of operators of such facilities.

Acknowledgements

This response is being provided provided by Dr. Sean Peisert, a Senior Scientist at Lawrence Berkeley National Laboratory, a Federally Funded Research and Development Center (FFRDC) operated by the University of California for the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author.

References

[1] Sean Peisert. Trustworthy Scientific Computing. *Communications of the ACM (CACM)*, 64(5):18–21, May 2021.

[2] Justine S. Hastings, Mark Howison, Ted Lawless, John Ucles, and Preston White. Unlocking Data to Improve Public Policy. *Communications of the ACM*, 62(10):48–53, September 2019.

[3] Jane Macfarlane. When Apps Rule the Road: The Proliferation of Navigation Apps is Causing Traffic Chaos. It's Time to Restore Order. *IEEE Spectrum*, 56(10):22–27, 2019.

[4] Sean Peisert. An Examination and Survey of Data Confidentiality Issues and Solutions in Academic Research Computing. Trusted CI Report — `https://escholarship.org/uc/item/7cz7m1ws`, September 2020.

[5] Arvind Narayanan and Edward W. Felten. No Silver Bullet: De-identification Still Doesn't Work. `http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf`, July 9, 2014.

[6] Mark Russinovich, Manuel Costa, Cédric Fournet, David Chisnall, Antoine Delignat-Lavaud, Sylvan Clebsch, Kapil Vaswani, and Vikas Bhatia. Toward Confidential Cloud Computing: Extending Hardware-Enforced Cryptographic Protection to Data While in Use. *Queue*, 19(1):49–76, February 2021.

[7] Joseph I Choi and Kevin RB Butler. Secure Multiparty Computation and Trusted Hardware: Examining Adoption Challenges and Opportunities. *Security and Communication Networks*, 2019(1368905), 2019.

[8] Fahmida Y. Rashid. The Rise of Confidential Computing. *IEEE Spectrum*, 57(6):8–9, 2020.

[9] Lily Hay Newman. Google Moves to Secure the Cloud From Itself. *Wired*, July 14, 2020.

[10] Sunil Potti and Eyal Manor. Expanding Google Cloud's Confidential Computing portfolio. `https://cloud.google.com/blog/products/identity-security/expanding-google-clouds-confidential-computing-portfolio`, September 8, 2020.

[11] Ayaz Akram, Anna Giannakou, Venkatesh Akella, Jason Lowe-Power, and Sean Peisert. Performance Analysis of Scientific Computing Workloads on General Purpose TEEs. In *Proceedings of the 35th IEEE International Parallel & Distributed Processing Symposium*, 2021.

[12] Dayeol Lee, David Kohlbrenner, Shweta Shinde, Krste Asanović, and Dawn Song. Keystone: An Open Framework for Architecting Trusted Execution Environments. In *Proceedings of the Fifteenth European Conference on Computer Systems (EuroSys)*, 2020.

[13] Ayaz Akram, Venkatesh Akella, Sean Peisert, and Jason Lowe-Power. Enabling Design Space Exploration for RISC-V Secure Compute Environments. In *Proceedings of the Fifth Workshop on Computer Architecture Research with RISC-V (CARRV), (co-located with ISCA 2021)*, June 17, 2021.

[14] Ayaz Akram, Venkatesh Akella, Sean Peisert, and Jason Lowe-Power. Simulating Trusted Execution Environments in gem5 (extended abstract). In *Proceedings of the Workshop on Modeling & Simulation of Systems and Applications (ModSim)*, October 6–8, 2021.

[15] Cynthia Dwork. Differential Privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, part II (ICALP)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12, July 2006.

[16] Nikhil Ravi, Anna Scaglione, Sachin Kadam, Reinhard Gentz, Sean Peisert, Brent Lunghino, Emmanuel Levijarvi, and Aram Shumavon. Differentially Private K-means Clustering Applied to Meter Data Analysis and Synthesis. *IEEE Transactions on Smart Grid*, 2022.

[17] Ammar Haydari, Michael Zhang, Chen-Nee Chuah, Jane Macfarlane, and Sean Peisert. Adaptive Differential Privacy Mechanism for Aggregated Mobility Dataset. arXiv preprint 2112.08487, 2021.