Initial Thoughts on Cybersecurity And Reproducibility

Ewa Deelman
deelman@isi.edu
Information Sciences Institute
Los Angeles, California

Victoria Stodden
vcs@illinois.edu
School of Information Sciences, University of Illinois at Urbana-Champaign
Urbana, Illinois

Michela Taufer
taufer@utk.edu
The University of Tennessee Knoxville
Knoxville, Tennessee

Von Welch
vwelch@iu.edu
Indiana University
Bloomington, Indiana

ABSTRACT

Cybersecurity, which serves to protect computer systems and data from malicious and accidental abuse and changes, both supports and challenges the reproducibility of computational science. This position paper explores a research agenda by enumerating a set of two types of challenges that emerge at the intersection of cybersecurity and reproducibility: challenges that cybersecurity has in supporting the reproducibility of computational science, and challenges cybersecurity creates for reproducibility of computational science.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**.

KEYWORDS

computer security, computational science reproducibility, data integrity

ACM Reference Format:
Ewa Deelman, Victoria Stodden, Michela Taufer, and Von Welch. 2019. Initial Thoughts on Cybersecurity And Reproducibility. In *2nd International Workshop on Practical Reproducible Evaluation of Computer Systems (P-RECS'19), June 24, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3322790.3330593

1 INTRODUCTION

Computing has become nearly ubiquitous in the scientific research process. Unfortunately, attempts to access and misuse computing systems by unauthorized and malicious entities are common, regardless of the intended use of a computing system. This includes computer systems used for scientific research and housed on campuses and national laboratories (e.g., [3, 5, 9, 11, 12, 18, 19]).

P-RECS'19, June 24, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6756-1/19/06…$15.00
https://doi.org/10.1145/3322790.3330593

In the context of computational science, a common aspect of the mission of scientific projects, research laboratories, etc. is reproducible science. Thus, this mission often implies a certain level of openness and the sharing of data, results, and resources as part of the scientific process (for example sharing within the lab or research group, sharing within a field, or even open public availability of research artifacts). However, no matter how open a particular computational science project may be, there is a need to prevent the computer systems it uses from misuse. Further, there is often a need to protect intellectual property, assure the privacy of sensitive information, and provide data integrity. These needs call for a level of cybersecurity that can manage risks to an entity's mission caused by unauthorized access or attacks to its computer systems.

The goal of this position paper is to explore a research agenda for how the field of cybersecurity best relates to the field of reproducibility in computational science, an area we believe has not been formally explored by either the cybersecurity or reproducibility communities. Hence, the key aim of this position paper is to start a conversation between researchers and practitioners in the cybersecurity and reproducibility fields, and to take an initial step towards defining a research agenda at the intersection of these fields.

We note our goal is distinct from work that seeks to improve the reproducibility of cybersecurity experiments (e.g. [8]). In contrast, we are exploring the role cybersecurity has in both supporting the reproducible computational science (e.g., ensuring computational infrastructure is free from malicious interference with unpredictable consequences) and conflicts that arise between cybersecurity and reproducibility in practice (e.g., the common cybersecurity practice of patching systems with possible implications on repeatably).

2 DEFINITIONS: CYBERSECURITY AND REPRODUCIBILITY

While a common definition of cybersecurity is focused on unauthorized access, e.g. Merriam-Webster defines cybersecurity as the "measures taken to protect a computer or computer system (as on the Internet) against unauthorized access or attack" [20], we take a slightly broader definition and include the related areas of data integrity and privacy. Data integrity failures are unauthorized or unintended changes to data caused by either malicious actors or random information technology system failures. Privacy, in the context of this paper, is the maintenance of the confidentiality of

research data such as personal health information or other data whose publication would be considered harmful to research subjects or other entities.

We define reproducibility in the computational sense: providing digital scholarly objects associated with the computational findings that would allow a reader to understand and regenerate the results. This includes any data, codes or scripts, inputs, and other relevant information, and made available in an open way if possible. [14, 15].

3 THE ROLE OF CYBERSECURITY IN SUPPORTING REPRODUCIBILITY

Before discussing the challenges that cybersecurity creates for reproducibility, we argue that cybersecurity plays is integral to enabling reproducibility. A key role of cybersecurity is to prevent, detect, and recover from unauthorized access and modifications to a computer system, software, and data. Such modifications may compromise the intended behavior of the computer system, and may remain undetected or detected at a much later date. A computer system with poor cybersecurity will be susceptible to unauthorized changes and hence one cannot have confidence in the behavior of that system. Hence, cybersecurity is necessary for reproducibility and the challenges in the remainder of this paper cannot be overcome by neglecting cybersecurity and its related challenges.

4 CYBERSECURITY AND REPRODUCIBILITY CHALLENGES

In this section we enumerate a set of challenges that emerge at the intersection of cybersecurity and reproducibility in two ways: challenges that cybersecurity has in supporting the reproducibility of computational science, and challenges cybersecurity creates for reproducibility of computational science.

4.1 Impact of Unauthorized Access on Reproducibility

As we describe earlier in this paper, a key role of cybersecurity is preventing unauthorized access to a computer system. Unfortunately, cybersecurity is not perfect and unauthorized access to computing systems occurs with regularity. When an unauthorized access occurs to a computing system involved in computational science, one loses some amount of confidence that the computer system is behaving as it is intended. Analysis of logs, both on the computing system and elsewhere, can retroactively establish to varying extent the actions of the unauthorized actors, but in practice one is never completely certain their actions are fully understood. How does this loss of confidence impact reproducibility?

4.2 Impact of Patching on Reproducibility

From the perspective of reproducibility, computer systems and software used for research would ideally be static, helping to ensure they would repeatedly produce the same results. However, a common cybersecurity practice is the updating of software (or firmware or even hardware) to mitigate vulnerabilities that could be exploited by unauthorized parties [6] - a practice commonly referred to as "patching." Ideally, patching would mitigate the vulnerability without otherwise impacting the functionality of the computer system. Unfortunately, this is not always the case and a patch may have undesirable impacts on functionality, and hence impact reproducibility. For example, the Spectre and Meltdown patches had significant impacts on system performance [1]. How does one determine the impacts of system software changes on reproducibility? For example, a code may no longer execute after system software has been patched. If running this code is necessary for the regeneration of scientific findings, reproducibility may be affected.

4.3 Impact of Imperfect Data Integrity on Reproducibility

A concern of unauthorized access would be the changing of data, whether intentional or a byproduct of misbehavior. There is also concern that integrity errors may occur in computer systems due to various implementation errors and statistical aberrations [17]. A number of cybersecurity controls exist to protect against accidental or malicious changes to data. For examples, hashes are commonly used to detect changes to data both at rest and in transit. However, as with any cybersecurity mechanisms, these are not perfect nor are they always carried out.

What is unclear is to what extent integrity errors impact reproducibility. Does a one-bit random error in a petabyte dataset invalidate any science based on that dataset? It seems reasonable to argue that not all bits are equal across all domains of science - many sciences sensing physical events deal with significant noise in their data while other science uses data that seems highly sensitive to any perturbation. So the question is: what types of data modifications are significant enough to impact reproducibility for different types of computational science?

4.4 Confidentiality of Data and Software

Cybersecurity and privacy concerns often call for data and software to be kept confidential. For example, data containing personally identifiable individual information is commonly restricted, often due to legal restrictions (e.g. [4, 6, 7]). There is also a debate in the cybersecurity community about keeping software, particularly source code, confidential in order to limit exposing any vulnerabilities in that software that could be used by attackers to compromise computer systems on which the software is installed.

Confidentiality in research artifacts challenges a principle tenet of reproducibility: exposure of the underlying data, code, and computational steps taken to produce scientific findings. Hence, the question that arises regarding whether any reproducibility is possible in such contexts. Arguments have been made to maximize the reproducibility possible (for example perhaps some steps can be exposed) and to use mitigating procedures such as differential privacy when data contain personally identifiable information [16].

We also note that data privacy is also a challenge faced by cybersecurity research itself (e.g. [13]), which often relies on data from computer systems and networks that can contain private information about users of the system (e.g. web browsing habits)[2].

4.5 Cybersecurity as an Ethical Issue

There is a human element to trust in scientific findings. What role does cybersecurity, or the perception of cybersecurity, play in the

human trust of computational output? Reproducibility of findings helps increase trust, but arguably, a more secure computational system will also bolster trust in output. Are there metrics or certifications of "cybertrustworthiness" of systems that would lend greater credibility to research output from these systems [15]?

4.6 Costs and Efficiency: Trading off Reproducibility and Productivity

A scientific result may be reproducible, and perhaps even deemed reliable and accepted by the relevant domain researchers, but what if this reliability was obtained on a system that is highly secure but onerous to use? The cost of increased security may be increased run times, greater complexity of scientific or system applications, an increased learning curve for domain scientists using the system, and possibly larger teams of researchers. This trade off may be "worth it" in terms of the increased reliability of the findings, but how do we assess the risk and the potential loss of accuracy in the scientific output due to investments in greater cybersecurity?

5 CONCLUSIONS AND NEXT STEPS

This paper attempts to initiate a discussion regarding challenges that emerge at the intersection of cybersecurity and reproducibility. We expect a next step to be a discussion among cybersecurity and reproducibility researchers and practitioners to articulate, refine, and prioritize these challenges, and to evolve them into a research agenda. An initial task could be the identification and documentation of a number of examples of the issues described in this position position paper, perhaps as Grand Challenges, and then made available as a community resource so that the broader computational science community can better understand how cybersecurity breaches impact computational reproducibility. Ultimately we hope the resulting research will produce concrete guidance to cybersecurity and reproducibility practitioners to handle these challenges. The National Science Foundation (NSF) Cybersecurity Center of Excellence [10], with its goal of supporting trustworthy science across the NSF portfolio of funded research, can potentially serve as a facilitator of these conversations.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under Grants 1547272, 1642070, 1642053, 1642090, 1813537, and 1823385. The views expressed do not necessarily reflect the views of the National Science Foundation or any other organization.

REFERENCES

[1] Peter Bright. 2018. Here's how, and why, the Spectre and Meltdown patches will hurt performance. https://arstechnica.com/gadgets/2018/01/heres-how-and-why-the-spectre-and-meltdown-patches-will-hurt-performance/. Accessed: 2019-4-8.

[2] L Camp, Lorrie Cranor, Nick Feamster, Joan Feigenbaum, Stephanie Forrest, Dave Kotz, Wenke Lee, Patrick Lincoln, Vern Paxson, Mike Reiter, and others . 2009. Data for Cybersecurity Research: Process and Wish List. (01 2009).

[3] Adam Clark Estes. 2015. Anonymous: Still Trolling After All These Years. https://gizmodo.com/anonymous-still-trolling-after-all-these-years-1700374189. Accessed: 2018-4-2.

[4] Jeffrey Mervis. 2019. Can a set of equations keep U.S. census data private? https://www.sciencemag.org/news/2019/01/can-set-equations-keep-us-census-data-private. Accessed: 2019-4-6.

[5] Ellen Nakashima. 2018. Research firm releases new details on alleged Iranian hacking campaign targeting 300 universities. *The Washington Post* (March 2018).

[6] Engineering National Academies of Sciences and Medicine. 2019. *Reproducibility and Replicability in Science.* The National Academies Press, Washington, DC. https://doi.org/10.17226/25303

[7] U.S. Department of Health and Human Services. 2015. Health Information Privacy. https://www.hhs.gov/hipaa/index.html. Accessed: 2019-4-8.

[8] Sean Peisert and Matt Bishop. 2007. How to Design Computer Security Experiments. In *Fifth World Conference on Information Security Education*, Lynn Futcher and Ronald Dodge (Eds.). Springer US, Boston, MA, 141–148.

[9] Nicolas Perpitch. 2017. How cyber attackers almost stole a unique chance from Australian astrophysicists. *ABC News* (Oct. 2017).

[10] Trusted CI Project. 2019. Trusted CI: the NSF Cybersecurity Center of Excellence. http://trustedci.org. Accessed: 2019-4-6.

[11] Suan Ramsey. 2015. Anatomy of a Breach: Lessons Learned. 2015 NSF Cybersecurity Summit.

[12] Kathleen Ricker, James Barlow, and Craig Adams. 2008. *FBI Major Case 216: A Case Study.* Technical Report NCDIR-TR-2008-01. National Center for Digital Intrusion Response.

[13] Darren Shou. 2012. Ethical Considerations of Sharing Data for Cybersecurity Research. In *Financial Cryptography and Data Security*, George Danezis, Sven Dietrich, and Kazue Sako (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 169–177.

[14] Victoria Stodden. [n. d.]. The Legal Framework for Reproducible Scientific Research. *IEEE Computing in Science and Engineering* 11, 1 ([n. d.]), 35–40. https://doi.org/10.1109/MCSE.2009.19

[15] Victoria Stodden. 2013. Resolving Irreproducibility in Empirical and Computational Research Âń IMS Bulletin. http://bulletin.imstat.org/2013/11/resolving-irreproducibility-in-empirical-and-computational-research/. Accessed: 2019-4-6.

[16] Victoria Stodden, Friedrich Leisch, and Roger D. Peng. 2014. *Implementing reproducible research.* CRC Press/Taylor and Francis, Boca Raton, FL.

[17] Jonathan Stone and Craig Partridge. 2000. When the CRC and TCP Checksum Disagree. *SIGCOMM Comput. Commun. Rev.* 30, 4 (Aug. 2000), 309–319. https://doi.org/10.1145/347057.347561

[18] CERN Computer Security Team. 2017. Computer Security: Virtual Misconduct – Real Consequences. https://home.cern/news/news/computing/computer-security-virtual-misconduct-real-consequences. Accessed: 2019-4-6.

[19] Tiffany Trader. 2014. US Researcher Caught Mining for Bitcoins on NSF Iron. https://www.hpcwire.com/2014/06/09/us-researcher-caught-mining-bitcoins-nsf-iron/. Accessed: 2018-6-21.

[20] A Walls, E Perkins, and J Weiss. 2013. Definition: Cybersecurity. *Retrieved from Gartner. com website: https://www. gartner. com/doc/2510116/def initioncybersecurity* (2013).