University Data Can Fuel AI/ML Cybersecurity Research

A Vision with an OmniSOC Proofpoint

Ryan Kiser
Center for Applied Cybersecurity Research
Indiana University
Bloomington IN USA
rlkiser@iu.edu

Von Welch
Office of the Vice President for IT
Indiana University
Bloomington IN USA
vwelch@iu.edu

Brad Wheeler
Kelley School of Business
Indiana University
Bloomington IN USA
bwheeler@iu.edu

ABSTRACT

Research in AI/ML for cybersecurity requires high quality data showing both normal and abnormal events. Obtaining such data at meaningful scale has shown itself to be a challenge. Higher education institutions own large amounts of such data that they generate from their own IT and cybersecurity operations. This paper presents a vision for managed sharing of that data for research, acknowledging the challenges with implementing that vision and providing means for mitigating those challenges. A real-world example of such data sharing is described through a collaboration between a major, multi-institutional security operations center, a AI/ML cybersecurity researcher, and an applied research center.

CCS CONCEPTS

Security and privacy~Intrusion/anomaly detection and malware mitigation~Intrusion detection systems

KEYWORDS

Data Sharing, intrusion detection, prototype

1 Introduction

Research into the application of AI/ML to cybersecurity is increasingly critical in order to keep up with the growing complexity, sophistication, speed, and number of attacks [1]. However this research is challenged by the availability of real-world data [2], [3]. There are a number of reasons for this lack of availability, including privacy concerns, operational security concerns, reproducibility, and the lack of a clear financial model to provide the data in the face of these challenges.

College and university IT operations collect extensive and essential operational data as they manage vast campus networks, cybersecurity, and other systems. Such data and its appropriate uses are governed by campus policies, and most policies are necessarily restrictive in what constitutes appropriate uses. If managed properly, this operational data can provide a rare and extremely rich resource to advance the research mission of an institution in support of faculty research and development efforts. While the match for institutional data sources and research uses is self-evident, many factors have long impeded effective and policy-compliant collaborations between researchers and IT operations. These include vastly different cultures between critical operations staff and researchers; lack of mutual trust; policy (or interpretation of policy) impediments; provisioning the data in secure and useful ways; fear of campus community perceptions of operational data used for research; and others.

Our goal in this paper is to put forth the vision of how universities can foster research of AI/ML into cybersecurity using operational data they already generate, the challenges faced to make that a broadly achievable reality and provide a real-world example of achieving this vision the authors enabled at Indiana University utilizing a collaborative, multi-institutional security operations center, the OmniSOC.

2. The OmniSOC

In 2017, a group of research university CIOs and CISOs created the OmniSOC as a shared, cyber security operations center [4]. Its mission is to (1) provide highly efficient, real-time, and scaled cybersecurity services to support campus security operations; (2) to support policy-compliant research by faculty and students; and (3) enable workforce development through staff training and internships. The OmniSOC was viewed as a way to bridge beyond the obstacles that had so long impeded appropriate uses of institutional data to support the research mission.

In the period between July of 2019 to the end of April 2021, the OmniSOC has ingested 4.1 petabytes of data, representing 6.56 trillion events, and generated 704 alerts from that data. Since its inception, it has expanded its scope from the founding members in the Big Ten Academic Alliance to serve a diverse community that includes a smaller private university and National Science Foundation Major Facilities.

In its first years the OmniSOC surmounted many common impediments to support both information security risk mitigation needs of its members and provide institutional data for their research missions. Members agree to work together in good faith to foster research through managed access to data. While the data access process is to-date ad hoc, depending on the exact nature of the research involved and the data needed, it does represent a real effort to tackle the challenges, described in the subsequent Section 3, which has already resulted in a collaboration and published paper with follow-up research and software improvements, as described in Section 4.

3. Challenges

Using the operational data from a university for any form of research invokes many challenges, the types of operational data needed for cybersecurity research especially so. In this Section we briefly describe those challenges, initial efforts taken to overcome them for our initial research collaboration, and our thoughts on long-term and scalable solutions.

3.1. Privacy and Confidentiality

Sharing operational data of interest to AI/ML cybersecurity researchers bears two risks:

1. It may reveal operational details that compromise the security of the institutions. For example, network topology, passwords that are inadvertently logged (for example, if someone types their password into a username field), metadata about communications that indicate confidential initiatives, etc. This data leakage may be directly from the shared data, indirectly through the resulting research, or by novel means which aren't readily anticipated (e.g. membership inference attacks [5]).

2. It may reveal attributes or behavior of members of the institution's community that are embarrassing, sensitive, or private. For example, traffic to websites that imply health concerns, sexuality, family planning, etc.

The extent of these risks varies greatly on the type of data and hence so do the solutions. In practice, a review of the research by an IRB or IRB-like body, who understands the technical issues involved is needed to understand the risks and set a protocol for their mitigation. Such protocols can include anonymization of the data, signed agreement by the researchers, and/or controlled access to the data - in situ analysis where an algorithm is run on behalf of the researcher with their having access to the data [6].

3.2. Data Formats, Semantics, and Access Standards

There is the technical challenge that researchers, and their analysis software, need to be able to access, process, and interpret the data. Standards exist for logs (e.g. [7]), but there is no ubiquitous standard researchers can expect. Additionally, AI/ML researchers often need data well labeled with distinctions about the data provided for training. These factors require significant work by both the data providers and consuming researchers to utilize data.

3.3. Market Creation

A common opening conversation between data owners and researchers is the data owner asking what data the researcher wants and the researcher asking what the data owner can provide, the owner providing a list, and the researcher asking for examples. This exchange reflects the lack of well-defined data products that allow the provider and researcher to easily communicate the demand and supply such that they can come to easily come to a shared understanding of what would be useful.

3.4. Reproducibility

Assuming other challengers are overcome, research is accomplished and published, there is the question of whether it can be reproduced. Is the data provider obligated to maintain a copy of the data used for the research? For how long? If the data is of significant size, this contributes to the subsequently described challenge of developing a sustainable financial model [8].

The adversarial nature of the cybersecurity domain itself contributes substantially to reproducibility challenges. The tactics, techniques, and procedures used by attackers are continuously changing in order to overcome defensive measures, and in response defensive measures must change as well. Because of this, reality may rapidly diverge from data sets collected at a point in time and provided for research purposes.

3.5 Sustainable Financial Model

Providing operational data for research is not without costs. As described in previous challenges, risks must be reviewed and mitigated, data may need to be anonymized, researchers supported in data access or in running their algorithms, and data may need to be sustained for reproducibility.

Having researchers pay for access to the data is one possible model but may present a barrier to adoption until the value of the data has been proven. Institutions may also choose to fund this in order to provide their researchers an advantage (and secondarily the institution through research competitiveness). And funding itself can include support for researcher access, as done for example by the ResearchSOC [9], which builds on the OmniSOC.

Financial models that incentivize university investments in ongoing, necessary, and valuable cyber risk mitigation are more likely to be sustainable than one-off research funding efforts. Thus, both can be achieved when the research data is

purposefully engineered to be a secondary product of operational cybersecurity investments.

4. A real-world Example: ASSERT

In August 2019, a collaboration between researchers at the Rochester Institute of Technology (RIT), the OmniSOC, and Indiana University's Center for Applied Cybersecurity Research (CACR) identified ASSERT [10] as a candidate research project for analysis of live security alert data from Indiana University provided by OmniSOC. ASSERT is an unsupervised learning system developed by Yang, et al. which categorizes related attacker behaviors derived from alerts and other information into descriptive models. These models can then be used to help analysts and other security practitioners to better understand attacks. CACR established a testbed which could be used to evaluate ASSERT and worked with OmniSOC to design and establish an appropriate connection between OmniSOC's infrastructure and the CACR testbed to feed alert data to a software prototype. CACR staff maintained the testbed and provided access to the research team from Rochester Institute of Technology led by Dr. Yang to install and configure the prototype.

CACR developed a data usage proposal to make use of IU security alert data from OmniSOC. The chosen data consisted of live alerts generated by Suricata alerting software from sensors on IU's network which are fed to OmniSOC's infrastructure for analysis. Identifying data was removed from the alerting feed with the exception of IP addresses which were hashed and provided as identifiers for source and destination machines in individual alerts. ASN numbers were added to enable us to identify inbound and outbound traffic flows in alerts. Indiana University's information security and privacy offices determined that this configuration for the data feed represented a policy-compliant, acceptable risk and granted approval for the feed to be started. In addition, the data gathered during operation of the testbed was provided to the research team to enable them to do follow-up work and make additional improvements to the ASSERT prototype. The testbed was operational until October 2020 and the research team continues to analyze the outputs of this work to identify and develop improvements to ASSERT.

The duration of this evaluation period allowed the research team to engage directly with security practitioners and to develop key insights which enabled further development of the software. The initial prototype was designed to analyze Splunk alerts and present a web frontend to analyze results. During this project, the research team added the necessary functionality to consume information from anonymized Suricata alerts and the CACR team integrated the prototype's outputs with Elasticsearch so that analysts could interact with the results using Kibana. These efforts have resulted in significant improvements to the software and identification of new use cases [11].

One key example of new features is the ability to consume ASN numbers, which adds another dimension to the attack models that can be used to analyze network traffic flows associated with models without identifying specific systems involved. This may enable the use of ASSERT in environments where there is not a well established trust relationship. In addition, the software now uses an exponentially weighted moving average approach to process new alerts so that the software can continuously consume alert streams from operational sources that do not have a defined end. This functionality enables the software to consume data streams from production network monitoring systems where the stream may continue indefinitely. These features were developed to address challenges identified by working within the context of a large-scale operational cybersecurity environment which may not have been identified without collaborating with an organization such as OmniSOC.

5. Conclusion

Institutions of higher education have a great need for rapid advances in AI/ML to help secure their large and critical IT operations. Likewise, they have a research mission with faculty and students who need access to vast and real data to advance their research.

We have laid out a vision for how institutions can leverage their operational data to foster AI/ML research for cybersecurity, discussed challenges with making this vision a reality and some ways to overcome those challenges, and described a real-world and ongoing multi-institutional example where their challenges were overcome to produce published research. Our hope is by articulating this vision and direction, we encourage higher education institutions to follow this approach, bolstering both AI/ML cybersecurity research and ultimately operational cybersecurity in higher education.

ACKNOWLEDGMENTS

The OmniSOC was co-founded by the CIOs and CISOs at Indiana University, Northwestern University, Purdue University, Rutgers University, and the University of Nebraska in 2017 [FYI, concept in July 2016, 5 founders signed and invested by June 2017, public announcement March 2018].

Work described in this paper was funded under NSF Grants 1840034, 1920430, and 2016431. We gratefully acknowledge support from the Indiana University Office of the Vice President for Information Technology and the Vice President for Research. The views expressed do not necessarily reflect the views of the National Science Foundation or any other organization.

We would like to thank Dr. S. Jay Yang, Dr. Ahmet Okutan, Gordon Werner, Ayush Goel, and Ren Chauret from Rochester Institute of Technology for their work developing ASSERT and

their collaborative efforts with the Indiana University Center for Applied Cybersecurity Research.

REFERENCES

[1] S. Samtani, M. Kantarcioglu, and H. Chen, "Trailblazing the artificial Intelligence for cybersecurity discipline," ACM Trans. Manag. Inf. Syst., vol. 11, no. 4, pp. 1–19, Dec. 2020, doi: 10.1145/3430360. [Online]. Available: https://dl.acm.org/doi/10.1145/3430360

[2] L. Jean Camp, L. Cranor, N. Feamster, and J. Feigenbaum, "Data for Cybersecurity Research: Process and Wish List," Jan. 2009 [Online]. Available: https://www.researchgate.net/publication/255960171_Data_for_Cybersecurity_Research_Process_and_Wish_List. [Accessed: 11-May-2018]

[3] J. Mirkovic et al., "Cybersecurity Datasets: A Mirage." [Online]. Available: https://www.caida.org/workshops/wombir/2101/slides/wombir2021-paper23.pdf. [Accessed: 07-May-2021]

[4] "First-of-its-kind higher education joint cyber security operations center launches." [Online]. Available: https://itnews.iu.edu/articles/2018/first-of-its-kind-higher-education-joint-cyber-security-operations-center-launches.php. [Accessed: 07-May-2021]

[5] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Towards Demystifying Membership Inference Attacks," arXiv [cs.CR], 28-Jun-2018 [Online]. Available: http://arxiv.org/abs/1807.09173

[6] J. Zeng, G. Ruan, A. Crowell, A. Prakash, and B. Plale, "Cloud computing data capsules for non-consumptiveuse of texts," in Proceedings of the 5th ACM workshop on Scientific cloud computing, 2014, pp. 9–16, doi: 10.1145/2608029.2608031 [Online]. Available: https://dl.acm.org/citation.cfm?doid=2608029.2608031. [Accessed: 31-May-2018]

[7] "Elastic Common Schema (ECS) Reference." [Online]. Available: https://www.elastic.co/guide/en/ecs/current/index.html. [Accessed: 06-May-2021]

[8] E. Deelman, V. Stodden, M. Taufer, and V. Welch, "Initial Thoughts on Cybersecurity And Reproducibility," in Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems, Phoenix, AZ, USA, 2019, pp. 13–15, doi: 10.1145/3322790.3330593 [Online]. Available: https://doi.org/10.1145/3322790.3330593. [Accessed: 07-May-2021]

[9] "NSF Award Search: Award # 1840034 - CICI: CSRC: Research Security Operations Center (ResearchSOC)." [Online]. Available: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1840034&HistoricalAwards=false. [Accessed: 07-May-2021]

[10] A. Okutan and S. J. Yang, "ASSERT: attack synthesis and separation with entropy redistribution towards predictive cyber defense," Cybersecurity, vol. 2, no. 1, pp. 1–18, May 2019, doi: 10.1186/s42400-019-0032-0. [Online]. Available: https://cybersecurity.springeropen.com/articles/10.1186/s42400-019-0032-0. [Accessed: 06-May-2021]

[11] S. J. Yang, A. Okutan, G. Werner, S.-H. Su, A. Goel, and N. D. Cahill, "Near Real-time Learning and Extraction of Attack Models from Intrusion Alerts," arXiv [cs.CR], 25-Mar-2021 [Online]. Available: http://arxiv.org/abs/2103.13902