Data Sharing for Energy Delivery Systems

Website: http://cred-c.org/researchactivity/datasharing

Researchers (Illinois): Qi Wang, Karan Ganju, Carl Gunter and Nikita Borisov

Industry Collaboration:

- Currently seeking collaborators from industry, power utilities, or national labs who would like to collaborate with us on data sharing for energy delivery systems. Contact Carl Gunter or Nikita Borisov to discuss how we can exchange ideas or to collaborate with our team.

Description of research activity: Data and information sharing is an effective strategy that can help reduce cybersecurity threats in Industrial Control Systems (ICS). When an attack is detected, a signature can be developed and shared with others to prevent the spread of the attack and enable proactive deployment of defenses. Energy Delivery Systems (EDS) can also benefit from such threat intelligence sharing schemes. To this end, the DoE has created E-ISAC (Electricity Information Sharing and Analysis Center) which encourages participants to share cybersecurity-related information such as indicators of compromise and audit logs to help other participants identify threats early and respond quickly. However, by using these logs a clever adversary (a competing vendor or an attacker compromising the shared log database) may be able to extract sensitive information pertaining to the internal infrastructure and topology of a service provider or enterprise and use this information for negative publicity, competitive advantage, or targeted attacks. Hence, participants may be reluctant to share their data via these platforms. Additionally, it is not clear how threat intelligence generated at a particular EDS can be generalized and applied to other EDSs where the infrastructure could be different and the attack manifests itself differently.

In light of these challenges, a promising solution is to share machine learning classifiers instead of sharing entire logs and other threat intelligence data. Participating vendors would simply query the classifier and determine if their infrastructure is vulnerable to the threat modeled by the classifier in question without having access to any sensitive information. With sufficient training data, the classifier would be able to generalize the underlying attack patterns and respond to queries from different EDSs with different topologies and infrastructure. However, one challenge is that classifiers have been shown to leak information about training data via repeated querying. The extent of the leakage is governed by the underlying nature of the training data and the types of models being shared. Countermeasures such as learning with differential privacy [10] have been proposed in the literature, but they are mostly focused on individual privacy not enterprise confidentiality. A more detailed investigation of the space is needed to provide meaningful security guarantees for EDSs.

This activity focuses on investigating and developing secure data and information sharing schemes for EDSs based on classifier sharing. The focus will be on understanding the information leaked by shared classifiers and developing models that preserve the confidentiality of the party that trained it. To train the models, we propose using audit logs from EDSs and supplement EDS data with traditional network audit logs wherever EDS data is missing (given the similarities between the two, the classifier might be able to generalize its internal models). As evident, this is a large space involving the study of secure machine learning models, understanding confidentiality in the context of EDSs and development of protection mechanisms to reduce any associated risk to the confidentiality of the training data.

The phase-wise systematic breakdown of the research activity is mentioned below:

Phase 1: Collect audit data such as network data generated in EDS. Augment that with traditional network logs (such as from [11] and [12]) where needed.

Phase 2: Investigate machine learning models feasible for EDSs and train classifiers on the collected data. Explore the type of models that are common in enterprise networks and ICSs and leverage that information where needed.

Phase 3: Investigate potential confidential information that can be inferred from the machine learning models.

Phase 4: Develop requirements for EDS confidentiality and come up with standards to categorize data and information into various levels based on their sensitivity.

Phase 5: Develop algorithms to assess confidential information inference risk.

Phase 6: Develop protection measures to eliminate or reduce information leakage.

The ultimate goal is to create mechanisms for protecting enterprise confidentiality. This includes developing machine learning models that protect training data and cannot be used to infer information pertaining to the enterprise or EDS that generated the classifier. The results of the activity would be algorithms and systems that would help achieve these goals. The tools that will be developed should be part of the data sharing approach, which is the focus of the E-ISAC and the DoE. For example, one organization could use the tools to assess the inference risks of sharing its machine learning models. The organization could then use the tools to reduce the inference risks before publishing or otherwise sharing the models. The activity will also explore the possibility of extending STIX/TAXII using its customized objects and properties to facilitate the sharing of machine learning classifiers.

The activity will allow us to better understand and answer challenges pertaining to secure data sharing in EDSs. The main questions are: What information can be extracted by reviewing the logs of an enterprise such as firewall and server access logs? What information is leaked by a classifier trained on such data? Is that leak meaningful/damaging to the EDS and a cause for concern? Can we develop metrics and standards to quantify the leaky nature of a classifier? How can we improve the confidentiality assurances of such classifiers without excessively reducing their effectiveness? Do these classifiers actually detect the types of attacks specific to EDS? Can we build more flexible data sharing schemes such as generating synthetic data that can be used to train classifiers?

How does this research activity address the Roadmap to Achieve Energy Delivery Systems Cybersecurity?
The proposed research ties into the roadmap via various ways. We highlight some below:

Build a Culture of Security: With our focus on enterprise confidentiality and secure data sharing, we are enabling a data sharing approach that would encourage vendors to protect data and enterprise confidentiality while still being able to share threat and other beneficial information with peering vendors in an effective and meaningful way. Participants will develop an understanding of how data can leak from seemingly harmless logs or classifiers and what are the best practices and techniques to prevent this unwanted leakage. This will create raised awareness amongst admins and inculcate a culture rooted in security when it comes to data sharing.

Assessing and Monitoring Risks: One of the goals of this activity would be developing metrics and standards that would allow us to quantify and measure the EDS-specific confidentiality risks associated with various data sharing schemes. As mentioned previously, classifiers have been shown to leak information and despite the fact that the leakage is less severe as opposed to sharing a full network or access log, we currently lack mathematically sound mechanisms to quantify and compare the degree of leakage. Once metrics have been developed that would allow us to compute the "strength" of various approaches and machine learning models, a more informed decision can be taken as to the best choice for secure data sharing given the current context.

Manage Incidents: A classifier that models a particular attack will allow vendors to check the vulnerability of their own infrastructure against the vectors specific to the threat. If a vendor determines that it is vulnerable to the threat it can proactively make changes and deploy defenses before the attack even begins. At times, however, the classifier will predict that the attack will succeed and upon further inspection a vendor might find that its infrastructure has already been compromised (attackers in the Ukraine attack broke in and stayed dormant for 6 months before making a move). In response, the vendor could quickly flush out the attackers, patch any vulnerabilities and put up defenses. Additionally, they could change a few parameters pertaining to the infrastructure and query the classifier again to determine if the modified infrastructure is still vulnerable to the attack. If the classifier outputs a reasonable degree of security against the attack, the vendor could actually make those changes in their physical infrastructure. The approach allows vendors to better manage current incidents and allowing them to check how different changes to the underlying infrastructure and topology will help prevent future incidents.

Sustain security improvements: One notable near-term goal specified in the EDS cybersecurity Roadmap is timely sharing of threat information. The fundamental need is to quickly disseminate information pertaining to an attack to all stakeholders in order to minimize damage and contain the spread. Our classifier-sharing framework is closely aligned with this goal. We are aiming to develop technologies and tools that will together enable easy and hassle-free sharing of such information in a timely manner. Additionally, the approach minimizes the duplication of technology development efforts as participating vendors can simply update a published classifier by training it on their own incident data for the benefit of the remaining community.

Summary of EDS gap analysis: Cyber security-related information such as indicators of compromise, forensics artifacts/samples, and incident reports are currently being shared among stakeholders in EDS via platforms such E-ISAC. However, the manner in which this information is shared currently has two primary disadvantages. First, the shared information may compromise enterprise confidentiality by revealing sensitive information (such as the infrastructure) of the distributor, which discourages participants from sharing data. Second, given that attacks evolve continuously and may exhibit different patterns for different topologies and infrastructure, each recipient of the threat intelligence still requires considerable effort to first map the problem onto their particular infrastructure and then develop detection/mitigation tools.

As machine learning applications are widely used in EDS to thwart evolving attacks, there is room to address these problems by sharing the machine learning classifiers and models instead of actual logs and reports. This will could minimize leakage while providing a classifier that can be used for diverse topologies and infrastructures without mapping or specialized tuning.

Full EDS gap analysis: Cyber Threat Intelligence (CTI) is a collection of information that details the current and emerging security threats, which enables an organization to determine its responses at the strategic, operational and tactical levels. CTI has been aggressively collected and increasingly exchanged across organizations in EDS to help the participants defend against cyber threats, often in the form of Indicators of Compromise (IOC), which are forensic artifacts (e.g., attack domains and MD5 of malware) that are used as signs that a system has been compromised by an attack or that it has been infected with a particular malicious software. As the volume and velocity of the information generated becomes hard to manage, E-ISAC encourages the use of threat intelligence sharing standards such as STIX/TAXII for automated information sharing with small signatures. However, this approach has problems. First, even though automatic extraction of IOC using different sources has been studied before [1] [2], it is difficult or sometimes impossible to create simple IOC for complex and novel attacks. Second, organizations still require a sizeable effort to comprehend how the shared IOCs can be used in their particular context as attacks typically employ multiple penetration vectors and what fails on one type of infrastructure might succeed on the other and vice versa. Simply sharing IOCs is not enough and further steps are needed in order to correctly parse the shared data and apply it to each vendor's specialized infrastructure. This automatically implies that defenses might also be different for each vendor, which further makes data sharing less beneficial. Third, the entire process needs to be done in a timely manner while simultaneously preserving the confidentiality of various stakeholders' sensitive data. Hence, there is a need to develop generic sharing schemes that can assist all participants without requiring modifications or complex parsing/mapping operations.

Machine learning techniques have gained widespread acceptance and success at effectively thwarting rapidly evolving attacks even at scale [3] [4]. It may be advantageous to share the machine learning models/classifiers so other participants can use the models to identify the attacks timely without spending time on developing their own signatures and parsing other organizations' logs. Additionally, this preserves the confidentiality of the participating vendors as training data is "absorbed" by the classifier and never leaves the organization (only the classifier is shared). To the best of our knowledge, sharing machine learning models is not being used in EDS currently.

However, the implications of sharing machine learning models is the focus of numerous recent studies. Research has shown that valuable or sensitive information about the training sets can be extracted from the models by selective and repeated querying of the model. For example, [5] and [6] study the membership inference attacks against machine learning models, which can be used to infer whether a record was used in the training set. [7], [8] and [9] study model

inversion attacks to extract context-specific meaningful information from the classifiers. Nevertheless, all these past attempts have focused on consumer data such as Kaggle's purchase dataset [13] and studied individual information inference and it is not clear how their findings translate to the context of organization data. The focus of this activity is on audit data, such as network packet traces (e.g., DARPA intrusion detection dataset [14]) and network flow traces (e.g., the CIDDS dataset [15]), which are currently available in EDSs. The goal is to study whether enterprise confidentiality can be compromised by launching attacks against shared classifiers. Additionally, as part of this study we also plan on developing metrics to help assess information leakage risks associated with classifiers and how best to protect classifiers from leaking any information not intended to be public. The TAXII protocol is designed to share threat information in the STIX format. Currently, the STIX specification mainly focuses on IOCs. But TAXII is not limited to STIX and can be used to share data in other formats. We will investigate how to extend STIX/TAXII using its customized objects and properties to facilitate the sharing of machine learning classifiers. The end goal is a data-sharing toolchain for the typical enterprise/vendor in the EDS ecosystem.

Bibliography:

1. Liao, Xiaojing, Kan Yuan, XiaoFeng Wang, Zhou Li, Luyi Xing, and Raheem Beyah. "Acing the IOC game: Toward automatic discovery and analysis of open-source cyber threat intelligence." In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 755-766. ACM, 2016.
2. Catakoglu, Onur, Marco Balduzzi, and Davide Balzarotti. "Automatic extraction of indicators of compromise for web applications." In Proceedings of the 25th International Conference on World Wide Web, pp. 333-343. International World Wide Web Conferences Steering Committee, 2016.
3. Zamani, Mahdi, and Mahnush Movahedi. "Machine learning techniques for intrusion detection." arXiv preprint arXiv:1312.2177 (2013).
4. Gilmore, Colin, and Jason Haydaman. "Anomaly Detection and Machine Learning Methods for Network Intrusion Detection: an Industrially Focused Literature Review." In Proceedings of the International Conference on Security and Management (SAM), p. 292. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2016.
5. Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership inference attacks against machine learning models." In Security and Privacy (SP), 2017 IEEE Symposium on, pp. 3-18. IEEE, 2017.
6. Hayes, Jamie, Luca Melis, George Danezis, and Emiliano De Cristofaro. "LOGAN: Evaluating Privacy Leakage of Generative Models Using Generative Adversarial Networks." arXiv preprint arXiv:1705.07663 (2017).
7. Fredrikson, Matthew, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. "Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing." In USENIX Security Symposium, pp. 17-32. 2014.
8. Ateniese, Giuseppe, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers." International Journal of Security and Networks 10, no. 3 (2015): 137-150.
9. Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures." In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1322-1333. ACM, 2015.
10. Abadi, Martín, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. "Deep learning with differential privacy." In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308-318. ACM, 2016.
11. Datahub, https://datahub.io/dataset?q=network, 2017.
12. SecRepo, http://www.secrepo.com/#network, 2017.
13. Kaggle. "Acquire Valued Shoppers Challenge." https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data, 2013.

14. MIT 1998 DARPA Intrusion Detection Evaluation Data Set, https://ll.mit.edu/ideval/data/1998data.html, 1998.
15. The CIDDS-001 data set, https://goo.gl/hXH4Ro, 2017.