Data-Driven Assessment of Network-Based Anomaly Detection Systems Protecting Cyber-Physical Systems

Robert Gillen

Importance and Relevance

According to the National Institute for Standards and Technology (NIST), “Cyber-Physical Systems (CPS) comprise interacting digital, analog, physical, and human components engineered for function through integrated physics and logic. These systems will provide the foundation of our critical infrastructure, form the basis of emerging and future smart services, and improve our quality of life in many areas.”[1] We see real-world examples of these systems in nearly every aspect of daily life: traffic flow management systems; protection systems on our electrical power grid; pump, flow, and chemical controls within a water treatment plant to name just a few. Even a cursory review of current news and reporting reveals a deep-rooted concern for the security and safety of these systems due to the dramatic effect they can have on our daily lives[2][3][4][5].

As a response to the concern for these systems, recent years have seen both government and industry make both significant investments and progress in developing approaches for the security and reliable operation of cyber-physical systems. These investments span a wide range of research including hardware, software, modeling and simulation, and empirical experiments. Many of these efforts have been funded on the premise that improved use of artificial intelligence-derived data analytics (machine learning, deep learning, anomaly detection, etc.) is key to properly securing the cyber-physical systems which comprise our critical infrastructure.

Many of these efforts suffer from a common flaw. While much effort is exerted in developing the algorithms and techniques to support a given defensive mechanism, little effort is expended in attempting to defeat said approach. This “honeymoon period” is both expected and valuable as new research areas need time to mature before being attacked. The time has now come to develop a scientifically based critical eye when looking at these defensive techniques and to establish a capability to challenge their assertions in real-world scenarios. Such a capability should be both measured and disciplined in its approach and target the assumptions of both science and implementation.

Research Problem and Goals of Proposed Work

Making the previous assertion is the easy part - actually doing it is much harder. Questions such as “how does one assess the suitability of such a system?” logically follow and yet remain largely unanswered. Many papers treat anomaly detection systems as traffic classifiers and then utilize common f-score approaches to quantify their successes.[6][7] Others[8] take an information theoretic approach which, while valid in their own right, struggle to account for the impact of real-world deployment considerations and configuration settings.

I propose to take a data-driven approach to attack the defenses of cyber-physical systems and more specifically, industrial control system networks. Existing research[9] has established that, if one can know when an anomaly detection system is being trained, one can poison the training data and thereby affect the definition of “normal” - allowing attacks that would otherwise be caught to succeed. My research aims to measure the degree to which a given system is susceptible to these types of attacks and standard attacks can be made to succeed. Further, I aim to externally measure (or estimate) the bounds of the system’s definition of normal to ascertain if a patient attacker (using his own anomaly detection system for the purposes of discovering the likely bounds of the deployed system) can craft attacks such that they are accepted as normal by the protection platform. The output of this effort will be a scoring or measurement system which conveys the degree to which a system is susceptible to these types of attacks and can be utilized to inform the defensive posture of such.

In a fashion not unlike cryptographic systems, it is often discovered that deployed systems exhibit behaviors different than the theoretical models and it is the assumptions or compromises made during the engineering and development that provide the most fertile ground for attack. As such, this research will focus on specific instantiations of the protection methodologies rather than evaluating theoretical or model-based designs. These systems have many “knobs” (configuration settings) which may significantly impact the realized effectiveness of the defense. A simple example is how the administrator adjusts the minimum anomaly score after which alerts are generated. There is an incentive to set this high enough to reduce false-positives yet low enough to not miss actual incidents of concern. One objective of this work is to help quantify the ramifications of changing such a setting. It is important to understand, however, that due to the assessment methodology (black box), the assessment will be unable to differentiate between the ramifications of an administrator-controlled configuration setting and a system-designer algorithm selection. An attacker would not generally be privy to this information and, as such, neither will I.

Based on the initial study of the domain, the primary research problems to be addressed are 1) effective manipulation of network traffic to assess the bounds of the system’s definition of “normal” and 2) alteration of existing attacks to attempt to conform to this definition. Work on the first challenge will be rooted in the assumption that the attacker has no visibility into how the protection system is configured (e.g. features used, algorithms applied, etc.) but is positioned on the network such that he can observe alerts that are generated (black-box testing). One can view this stage as a multi-dimensional parameter sweep that attempts to reverse engineer both the features and the associated weights used in the model. The result does not need to be perfect, but sufficient to inform the second phase.

Work on the second problem will focus on modifying the attacks in such a way as to comply (if possible) with the parameters derived during the first step. This will require an understanding of the nuance and intent of the attack and will model the position of a sophisticated attacker. Assuming the attack remains successful, an attempt will be made to quantify the relation between the “slack” available via the realized definition of “normal” and the effectiveness of the attack. For example, a configuration that allows a successful attack that slowed the attack from 5 seconds to 5 minutes would receive a worse score than a configuration that required the same attack to now take three days, or reduced the likelihood of success from 90% to 30%.

I do not intend to focus heavily on the numeric methods of expressing this relationship. The approach is certainly important and will be discussed, however, the intent is to express a relative relationship between two configurations within an otherwise identical environment and not to establish a globally-relevant score for a particular system or methodological approach.

Existing Work

There is a large body of work applying anomaly detection techniques to the problem of detecting intrusions or attacks on network traffic (CITE MANY). While some of these adopt a “purist” approach and simply return a measure of “weirdness” relative to the idyllic model of “normal”, most support some fashion of single-class classification either explicitly or implicitly in practice (trigger an alert or not).

Talk about others doing work in this space. Attacks, movements, etc.

Talk about adversarial ML that results in vector-based responses.

Talk about work attempting to defeat/detect adversarial attacks.

Approach and Methodology

In order to accomplish the stated research objectives, the following tasks will be performed:

- Determine relevant categorizations of network-based anomaly detection systems and obtain multiple deployable instantiations of each
- Develop means of modifying attack examples to approximate the bounds of "normal" for each deployed anomaly detection system
- Develop an approach to manipulating attacks such that they may cause their desired effect while still be considered "normal"
- Develop means of quantifying the effects to the attack (e.g. longer to execute, less effective, etc.) of the requisite changes
- Publish resulting methodology

This research has been in the proposal and planning stages for the past six months but actual work in on this project has only just recently commenced. As such, many of the specifics (e.g. test beds, test equipment, etc.) are still in flux. I intend to focus on energy infrastructure and nuclear science test beds and network-based anomaly detection systems (vs. host-based).

In collaboration with others at ORNL working on cyber-physical systems, I expect to publish the results of this work in one or more of the following: the International Journal of Infrastructure Protection[10], the IET Cyber-Physical Systems: Theory and Applications[11] journal, and possibly the Department of Homeland Security (DHS) newsletter for Industrial Control Systems[12].

Relevant Publications

[1] N. I. of Standards and Technology, "Cyber-physical systems." [Online]. Available: https://www.nist.gov/el/cyber-physical-systems.

[2] T. W. House, "Presidential executive order on strengthening the cybersecurity of federal networks and critical infrastructure." [Online]. Available: https://www.whitehouse.gov/presidential-actions/presidential-executive-order-strengthening-cybersecurity-federal-networks-critical-infrastructure/.

[3] D. of E. Office of Electricity, "Reducing cyber risk to critical infrastructure: NIST framework." [Online]. Available: https://www.energy.gov/oe/cybersecurity-critical-energy-infrastructure/reducing-cyber-risk-critical-infrastructure-nist.

[4] D. of Homeland Security, "Protecting critical infrastructure." [Online]. Available: https://www.dhs.gov/topic/protecting-critical-infrastructure.

[5] T. Koppel, *Lights out: A cyberattack, a nation unprepared, surviving the aftermath.* Penguin Random House, 2015.

[6] R. G. Goss and G. S. Nitschke, "Automated pattern identification and classification: Anomaly detection case study," in *Proceedings of the genetic and evolutionary computation conference companion*, 2017, pp. 59–60.

[7] S. Zanero, "ULISSE, a network intrusion detection system," in *Proceedings of the 4th annual workshop on cyber security and information intelligence research: Developing strategies to meet the cyber security and information intelligence challenges ahead*, 2008, pp. 20:1–20:4.

[8] G. Gu, P. Fogla, D. Dagon, W. Lee, and B. Skorić, "Measuring intrusion detection capability: An information-theoretic approach," in *Proceedings of the 2006 acm symposium on information, computer and communications security*, 2006, pp. 90–101.

[9] M. Kloft and P. Laskov, "Online anomaly detection under adversarial impact." *Journal of Machine Learning Research - Proceedings Track*, vol. 9, pp. 405–412, Jan. 2010.

[10] Elsevier, “International journal of infrastructure protection.” [Online]. Available: https://www.journals.elsevier.com/international-journal-of-critical-infrastructure-protection/.

[11] I. of Engineering Technology, “IET cyber-physical systems: Theory & applications.” [Online]. Available: https://ieeexplore.ieee.org/xpl/aboutJournal.jsp?punumber=7805360.

[12] D. of Homeland Security, “DHS newsletter for industrial control systems.” [Online]. Available: https://ics-cert.us-cert.gov/monitors.