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. Given the pervasive nature of these systems, logic would dictate that these are some of the most well-protected against attack and misuse. Unfortunately, the reality is often quite the opposite. This condition stems, at least in part, from the fact that many of these systems were designed to operate as independent or air-gapped systems yet are increasingly connected to corporate and other networks.[2] While security professionals have been raising concerns for years, events such as Stuxnet (2010), the compromise of the dam in Rye Brook, New York (2013) and attacks on the Ukrainian power grid (2015) have awakened both the general public and governmental awareness to this issue. 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[3][4][5][6].

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.

Unfortunately, the majority 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 - the execution of such is much harder. Questions such as "how should one actually assess the suitability of such a system?" logically follow yet remain largely unanswered. Many papers treat anomaly detection systems as traffic classifiers and then utilize common f-score approaches to quantify their successes.[7][8] Others[9] 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[10] 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, feature weights, hyper-parameters, 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. This stage attempts to determine the location of the threshold as illustrated in Figure 1. It is expected that the level of effort to develop this interrogation capability will be significant as some algorithms actively adjust and only trigger if the rate of change exceeds a certain threshold relative to the baseline[11] (rather than simple deltas from the norm).

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. Referencing once again Figure 1, the detected anomalies in this example are almost laughably obvious. The real question that should be asked, is how well the system can protect against attacks that are *just barely* different than the norm. 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.

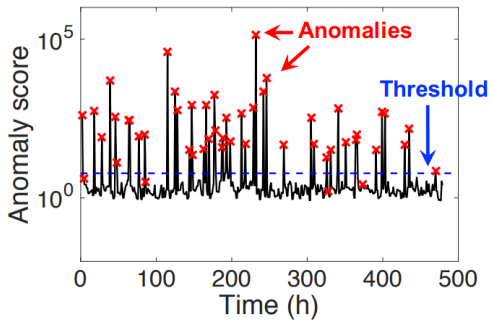

Figure 1: Example of anomalies relative to a threshold. The focus of the first stage of this research is to see if an attacker can ascertain the position of the threshold line. The second stage is to see if the standard attacks can be modified in a fashion as to live below that threshold. Image credit [12].

Existing Work

There is a large body of work applying anomaly detection techniques to the problem of detecting intrusions or attacks on network traffic[13][14][15][16][17]. 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). It is this delta that I intend to establish can be determined (or approximated) by an attacker for his benefit.

There exists a similarly large number of attacks on anomaly detection systems. In [10] Kloft and Laskov demonstrate different levels of success they can have attacking systems with different levels of access to the system (gray-box, white-box, training data, etc.). Ling et al. discuss a number of attacks against different classes of anomaly detection systems (PCA-based, spherical, etc.) in [18].

In [19] Hays and Danezis describe their attack on black-box deep learning models. They assume that they will have access to the data used to train the original model and that they will have access to the answers from the original model (resultant class weights). They then use this data to run the data through (train) their own model with the goal of mutating the input images in a visibly imperceptible fashion so as to cause the original model to misclassify the input data. While this is a novel approach and would likely work well for images, I do not expect this (or similar derivative style training approaches) to work in the context of network packets (though it may). My initial assumption is that the parameters of the attacks will need to be hand-modified to fit within the definition of "normal". It is hoped, however, that this work will lead to a prescriptive approach to addressing this issue.

There are a number of papers [20][21][22] that propose various methods of detecting the attempt to poison training data via adversarial examples (attacker attempting to change classification results), however, I do not believe many of these directly apply due to the fact that I am not specifically attacking the learning process. It is possible that some systems will employ an active learning approach, thereby allowing for poisoning-style attacks, but this is not the primary objective of my work. Further, I am uncertain that the behavior will differ significantly from systems that utilize a "governor" of sorts on the rate of change of the baseline (e.g. CuSUM).

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 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[23], the IET Cyber-Physical Systems: Theory and Applications[24] journal, and possibly the Department of Homeland Security (DHS) newsletter for Industrial Control Systems[25].

Relevant Publications

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

[2] G. L. Fuehring, "Graphic analysis and planning of electrical distribution systems," in *Proceedings of the 7th annual conference on computer graphics and interactive techniques*, 1980, pp. 204–210.

[3] 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/.

[4] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] J. Goh, S. Adepu, M. Tan, and Z. S. Lee, "Anomaly detection in cyber physical systems using recurrent neural networks," in *2017 ieee 18th international symposium on high assurance systems engineering (hase)*, 2017, pp. 140–145.

[12] H. S. B. Hooi D. Eswaran and C. Faloutsos, "GridWatch: Sensor placement and anomaly detection in the electrical grid," in *European conference on machine learning and principles and practice of knowledge discovery in databases*, 2018.

[13] S. S. Kim and A. L. N. Reddy, "Statistical techniques for detecting traffic anomalies through packet header data," *IEEE/ACM Trans. Netw.*, vol. 16, no. 3, pp. 562–575, Jun. 2008.

[14] J. Zhang, Y. Tong, and T. Qin, "Traffic features extraction and clustering analysis for abnormal behavior detection," in *Proceedings of the 2016 international conference on intelligent information processing*, 2016, pp. 25:1–25:6.

[15] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *Proceedings of the 2004 conference on applications, technologies, architectures, and protocols for computer communications*, 2004, pp. 219–230.

[16] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "An effective unsupervised network anomaly detection method," in *Proceedings of the international conference on advances in computing, communications and informatics*, 2012, pp. 533–539.

[17] A. Marnerides, D. P. Pezaros, and D. Hutchison, "Detection and mitigation of abnormal traffic behaviour in autonomic networked environments," in *Proceedings of the 2008 acm conext conference*, 2008, pp. 51:1–51:2.

[18] L. Huang, A. D. Joseph, B. Nelson, and B. I. P. Rubinstein, "Adversarial Machine Learning," no. October, pp. 43–57, 2011.

[19] J. Hayes and G. Danezis, "Learning Black-Box Adversarial Examples."

[20] E. Wong and J. Z. Kolter, "Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope," 2018.

[21] A. Ilyas, U. T. Austin, C. Daskalakis, and A. G. Dimakis, "The Robust Manifold Defense : Adversarial Training using Generative Models," pp. 1–25, 2017.

[22] A. Madry, L. Schmidt, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," pp. 1–22.

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

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

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