Title

A new metric to compare anomaly detection algorithms in cyber-physical systems

ISBN

9781450371476

Authors

Giraldo, Jairo

Cardenas, Alvaro A

Publication Date

2019-04-01

DOI

10.1145/3314058.3318166

Peer reviewed

A New Metric to Compare Anomaly Detection Algorithms in Cyber-Physical Systems

Jairo Giraldo
University of Texas at Dallas
Richardson, TX
jairo.giraldo@utdallas.edu

Alvaro A. Cardenas
University of California Santa Cruz
Santa Cruz, CA
alvaro.cardenas@ucsc.edu

ABSTRACT

The performance of different anomaly detection algorithms is typically compared using metrics that depend on the true positive rate (TPR) and the false positive rate (FPR). However, to obtain the TPR it is necessary to generate attacks that will be detected, which is useless to evaluate detection strategies against more realistic adversaries that can adapt their attacks to remain undetected. On the other hand, the FPR can be misleading and hard to interpret in practical applications since the amount of time a process is observed is not fixed. In this poster, we present a novel metric that is based on the maximum impact an adversary can cause while remaining stealthy, and on the expected time between false alarms. Our metric is useful for the evaluation and comparison of anomaly detection strategies in CPS.

ACM Reference Format:
Jairo Giraldo and Alvaro A. Cardenas. 2019. A New Metric to Compare Anomaly Detection Algorithms in Cyber-Physical Systems. In *Hot Topics in the Science of Security Symposium (HotSoS), April 1–3, 2019, Nashville, TN, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3314058.3318166

1 INTRODUCTION

One of the differences between detecting attacks in control systems when compared to detecting attacks in general IT systems is that researchers do not have readily available data from attacks in the wild. Even if we test our algorithms on the few known examples (like Stuxnet), they are domain specific and it is not clear they will give insights into the evaluation other than to show that we can detect Stuxnet (which can be easily detected *ex post*). For that reason, the question we would like to address in this poster is how to create attacks that are general enough to be applicable across multiple industrial control domains but that will also allow us to define an evaluation metric that is fair (and that is not biased to detect the specific attacks from the researchers).

To motivate the need of a new metric, we now discuss the challenges and limitations of common metrics in literature.

Measuring the True Positive Rate is Misleading. To obtain the true positive rate of a detection algorithm we need to generate an attack that will be detected. Publications using the true positive rate [9] generate their attacks as random signals (e.g., a sensor reporting fake random values). This type of non-strategic random failure is precisely what the fault-detection community has been working on for over 40 years; with those attacks we are not advancing the state of the art on attack-detection, but rather reinforcing the fact that fault-detection works when sensor or control signals fail in a non-malicious way.

Stealthy Attacks and ROC Curves. If we evaluate our anomaly detection algorithm against using a traditional intrusion detection metric like Receiver Operating Characteristic (ROC) curves, and the attacker is able to generate stealthy attacks, we would have had a 0% detection rate; that is, our ROC curve would be a flat line along the x-axis with a 0% value in the y-axis [2].

Summary. A *classification accuracy* metric of an anomaly detection algorithm $\mathcal{A}$ needs to capture two things: (1) the ability of $\mathcal{A}$ to detect attacks (we call this a *security metric*), and (2) the ability of $\mathcal{A}$ to label correctly *normal* events so that it does not raise too many false alarms (we call this a *usability metric*). The *security metric* and the *usability metric* represent a trade-off that needs to be balanced (lower false alarm rates typically means lower ability to detect attacks).

In this poster, we present the trade-off curve introduced in [8] that includes both (the security metric and the usability metric) and that is useful to evaluate and compare anomaly detection algorithms for cyber-physical systems, where the main goal of adversaries is to disrupt the physical process as much as possible while remaining undetected (e.g., stuxnet).

2 NEW EVALUATION METRIC

We assume an attacker that has compromised a sensor (e.g. pH level in a water treatment plant) or an actuator (e.g. pump or valve) in our system. We assume that the adversary has complete system knowledge, i.e. she knows the physical model we use, the statistical test we use, and the thresholds we select to raise alerts. Given this knowledge, she generates a stealthy attack, where the detection statistic will always remain below the selected threshold. Given this strong assumptions for the adversary, we can compute our proposed metric.

Computing Y-axis (Security). We consider a strong adversary model where the attacker knows all details about our anomaly detection test, and thus can remain undetected, even if we use *active* monitoring Given an anomaly detection threshold τ we want to evaluate how much "damage" the attacker can do without raising an alarm. The adversary wants to drive the system to the worst possible condition it can without being detected, where "worst:"

HotSoS, April 1–3, 2019, Nashville, TN, USA

ACM ISBN 978-1-4503-7147-6/19/04.
https://doi.org/10.1145/3314058.3318166

refers to *the maximum deviation of a signal from its true value that the attacker can obtain* (without raising an alarm, and given a fixed-period of time, otherwise given infinite time, the attacker might be able to grow this deviation without bound). Let y_k denote a sensor reading at an instant k. The adversary wants to maximize the deviation of a variable of interest y_k (per time unit) without being detected. The true value of this variable is $y_k, y_{k+1}, \ldots, y_N$, and the attack starts at time k, resulting in a new observed time series $y^a_k, y^a_{k+1}, \ldots, y^a_N$. The goal of the attacker is to maximize the distance $\max_i ||y_i - y^a_i||$.

An optimal greedy-attack (y^{a*}) at time $k \in [\kappa, \kappa_f]$ (where κ and κ_f are the initial and final attack times, respectively), satisfies the equation: $y^{a*}_{k+1} = \arg\max_{y^a_{k+1}} f(y^a_{k+1})$ (where $f(y^a_{k+1})$ is defined by the designer of the detection method to quantify the attack impact) subject to not raising an alert (instead of max it can be min).

Notice that while we have defined a specific impact for undetected attacks in our y-axis for clarity, we believe that designers who want to evaluate their system using our metric should define an appropriate *worst case undetected attack* optimization problem specifically for their system. In particular, the y-axis can be a representation of a cost function f of interest to the designer. There are a variety of metrics (optimization objectives) that can be measured such as the product degradation from undetected attacks, or the historical deviation of the system under attack $\sum_i |y_i - \hat{y}^a_i|$ or the deviation at the end of the attack $|y_N - \hat{y}^a_N|$, etc.

Computing X-axis (Usability). While the y-axis of our proposed metric is completely different to ROC curves, the x-axis is similar, but instead of using the false alarm rate, we use instead the expected time between false alarms $E[T_{fa}]$. This value has a couple of advantages over the false alarm rate: (1) it addresses the deceptive nature of low false alarm rates due to the base-rate fallacy [1], and (2) it addresses the problem that some anomaly detection statistics make a decision ("alarm" or "normal behavior") at non-constant time-intervals. We argue that telling security analysts that e.g., they should expect a false alarm every hour is a more direct and intuitive metric rather than giving them a probability of false alarm number over a decision period that will be variable depending of the anomaly detection tests.

The usability metric for each evaluated detection mechanism is obtained by counting the number of false alarms nFA for an experiment with a duration T_E under normal operation (without attack), so for each threshold τ we calculate the estimated time for a false alarm by $E[T_{fa}] \approx T_E/nFA$.

3 CASE STUDY

To illustrate the use of our metric, we will implement two detection strategies in a real testbed, the bad-data detection (stateless) and the CUSUM (stateful). We will consider the evolution of the water level in a tank in the Secure Water Treatment Testbed (SWaT) at the Singapore University of Technology and Design.

The goal of the attacker is to deviate the water level in a tank as much as possible until the tank overflows without being detected. A **successful attack** occurs if, when the PLC receives from the sensor a *High* water-level message (the point when the PLC sends a command to close the inlet, which corresponds to 0.8 m), then the real water level has already overflowed (the real level of water reaches 1.1 m). Therefore, the impact of the attack (the Y axis) is given by $\Delta = h^{real} - 0.8$, where h^{real} is the real level of water when the PLC closes the inlet valve. The value of Δ is computed for different τ for each detection strategy. The usability metric (X axis) is calculated for $T_E = 8$ h, which is the time of the experiment without attacks. Figure 1 illustrates our proposed trade-off curve. Clearly, when the stateless detection is being used, the attacker has enough room to launch a stealthy attack that will cause an overflow (i.e., $\Delta = 0.3$). On the other hand, the CUSUM algorithm is able to limit the impact of the adversary. Our proposed curve allows us to find an adequate threshold that will lead to a large enough expected time between false alarms, but at the same time a reasonable maximum impact.

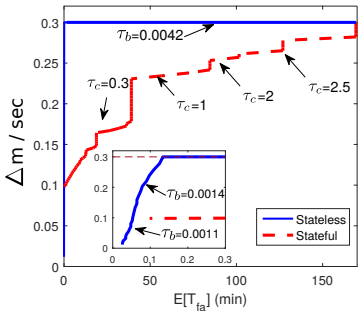

Figure 1: Comparison of stateful and stateless detection. Stateless tests are not good for this use case.

It is worth to mention that our metric has been used by several research groups in [3–7].

4 ACKNOWLEDGEMENTS

This work is partially supported by NSF CNS-1553683 and NIST 70NANB17H282.

REFERENCES

[1] Stefan Axelsson. 2000. The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security (TISSEC)* 3, 3 (2000), 186–205.
[2] Alvaro A Cárdenas, John S Baras, and Karl Seamon. 2006. A framework for the evaluation of intrusion detection systems. In *Proceedings of Symposium on Security and Privacy*. IEEE, 15–pp.
[3] Luis Francisco Combita, Jairo Alonso Giraldo, Alvaro A. Cardenas, and Nicanor Quijano. 2018. *DDDAS for Attack Detection and Isolation of Control Systems*. Springer International Publishing, Cham, 407–422.
[4] Amin Ghafouri. 2018. *Resilient Anomaly Detection in Cyber-Physical Systems*. Ph.D. Dissertation. Vanderbilt University.
[5] Jezdimir Milošević, David Umsonst, Henrik Sandberg, and Karl Henrik Johansson. 2018. Quantifying the impact of cyber-attack strategies for control systems equipped with an anomaly detector. In *Proceedings of the 2018 European Control Conference*. IEEE, 331–337.
[6] Kaveh Paridari, Niamh O'Mahony, Alie El-Din Mady, Rohan Chabukswar, Menouer Boubekeur, and Henrik Sandberg. 2018. A framework for attack-resilient industrial control systems: Attack detection and controller reconfiguration. *Proc. IEEE* 106, 1 (2018), 113–128.
[7] David Umsonst and Henrik Sandberg. 2018. Anomaly detector metrics for sensor data attacks in control systems. In *Proceedings of the 2018 American Control Conference (ACC)*. IEEE, 153–158.
[8] David I Urbina, Jairo A Giraldo, Alvaro A Cardenas, Nils Ole Tippenhauer, Junia Valente, Mustafa Faisal, Justin Ruths, Richard Candell, and Henrik Sandberg. 2016. Limiting the impact of stealthy attacks on industrial control systems. In *Proceedings of the Conference on Computer and Communications Security (CCS)*. ACM, 1092–1105.
[9] Yong Wang, Zhaoyan Xu, Jialong Zhang, Lei Xu, Haopei Wang, and Guofei Gu. 2014. SRID: State Relation Based Intrusion Detection for False Data Injection Attacks in SCADA. In *Proceedings of European Symposium on Research in Computer Security (ESORICS)*. Springer, 401–418.