Evaluating the Security of Anomaly Detection in CPS

Alvaro A. Cárdenas, Ovidiu Daescu
University of Texas at Dallas

The integration of Information Technology (IT) systems (computations and communications–the cyber world) with sensor and actuation data (the physical world), can introduce new, and fundamentally different approaches to security research in the growing field of Cyber-Physical Systems (CPS), when compared to other purely-cyber systems. In our earlier work [3, 1, 2, 5], we have shown that because of the automation and real-time requirements of many control actions, traditional security mechanisms are not enough for protecting CPS, and we require *resilient control and estimation* algorithms for true CPS defense-in-depth.

This has led to a lot of interest in exploring anomaly detection schemes for cyber-physical systems by using data collected from sensors. In the general setting, data obtained from *normal behavior* of the system is used to create a model and then any outlier is considered an anomaly and a potential failure or attack.

This line of research is actually very similar to the safety mechanisms that have been deployed in control systems for decades. In particular, the protection of control systems has traditionally been enforced by several safety mechanisms, which include bad data detection, protective relays, safety shutdowns, interlock systems, etc.

These reliability algorithms $\mathcal{R}$ receive inputs from various sensors $y \in \mathbb{R}^m$ and take a protective control action if an internal threshold τ is met. Most cases can be modeled by assuming that the protective control action is enabled whenever $\mathcal{R}(y) > \tau$.

Most of these algorithms were designed under the assumption of random faults, or typically well-known failure conditions (e.g., a frozen sensor). Under normal conditions we can model that measurement y comes from *null Hypothesis* H_0: a model of the system working properly, and under a random failure we can assume that y is a random variable following a distribution under the *alternate hypothesis* H_1. Under these assumptions we can measure the survivability of the system to these random failures by $\Pr[\mathcal{R}(y) > \tau | H_1]$; in other words, if we have a high probability that $\mathcal{R}$ is greater than the threshold τ, then we know the failure will be detected with high probability and corrective actions will be taken.

In practical terms, typical anomaly detection schemes are evaluated by the method described in Figure 1.

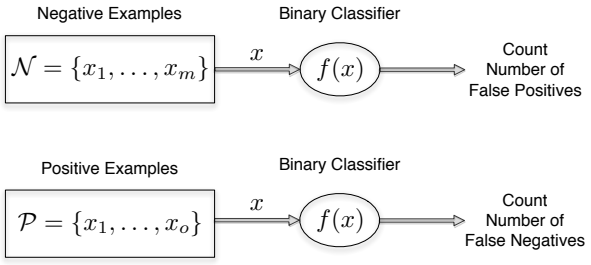

Figure 1: *Vanilla* testing is good to detect failures, but it is not a good way to evaluate systems when they face attacks.

As illustrated recently [4, 6], these traditional reliability mechanisms do not work against sophisticated attackers, because the attack will not be random, but rather a sophisticated control or sensor signal that will prevent $\mathcal{R}(y)$ from ever reaching the threshold τ.

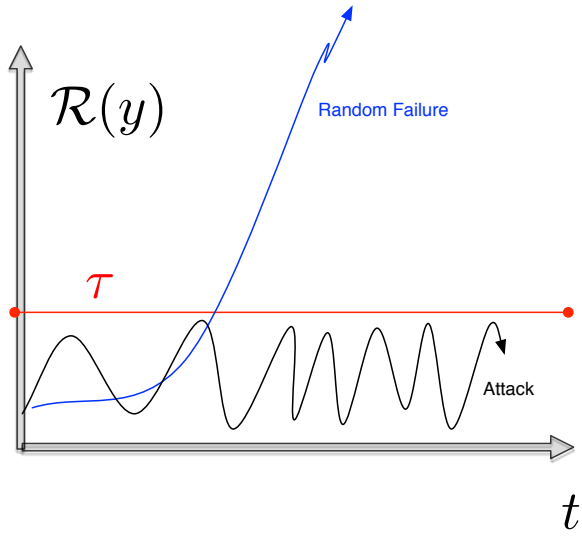

Figure 2: Difference between failures and attacks: sophisticated attackers will avoid crossing the anomaly detection threshold τ.

Our main argument is that instead of measuring the reliability of a control system under heuristically-created attacks (which will result in the blue curve of Figure 2) we need to find what is the **worst-possible undetected attack** (an undetected attack will generate something similar to the black curve in Figure 2). This has been the line of research we have pursued in recent work [2, 5].

In particular, instead of evaluating algorithms based on the trade-off between false positives (probability of false alarm) and true positives (probability of detection), our new proposed trade-off is between the false positives (probability of false alarm) and the **cost of undetected attacks**, as illustrated in Figure 3.

Figure 4 shows a simulation of attacks against a chemical reactor [2]; in particular, it shows how the statistic of the anomaly detector under under attack (red) can be controlled by the attacker to prevent being detected for a long time (red statistic can be maintained below the blue threshold).

With this new approach, we can build trade-

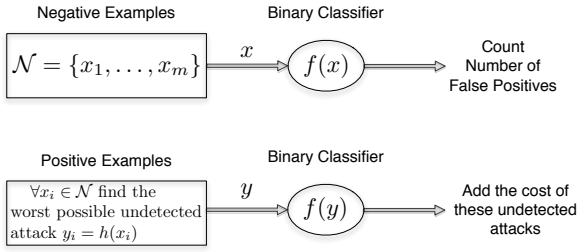

Figure 3: Our proposed attack evaluation metrics consider an attacker that creates a function $h(x)$ to find the worst-possible undetected attack. This is different from the "Vanilla testing" shown in Figure 1.

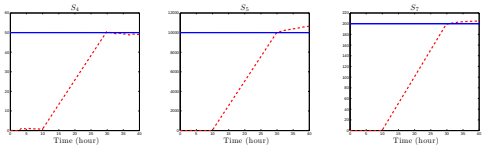

Figure 4: The attack statistics will remain below the threshold.

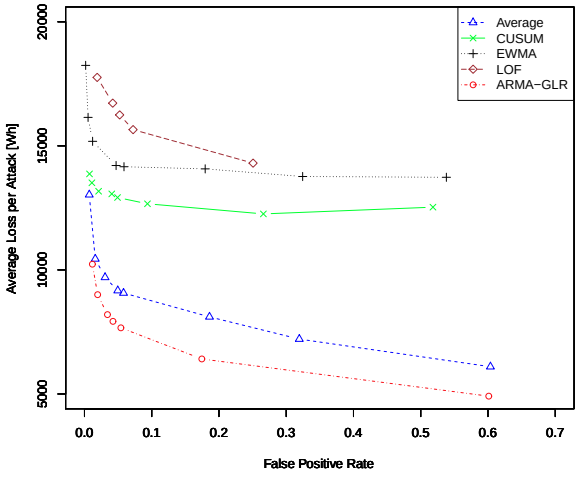

Figure 5: Trade-off curve between usability and security.

off curves different from Receiver Operator Characteristic (ROC) curves, and focus on new attack curves, where the y-axis is the cost of an attack. For example, in recent work [5] we evaluated the performance of electricity-theft detectors subject to attacks that will steal the most electricity without being detected (See Figure 5).

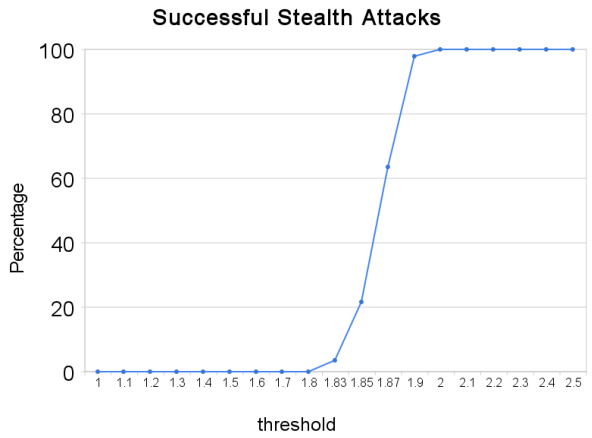

Figure 6: We can change the threshold of the anomaly detector in such a way that stealthy attacks are not detected by the anomaly detector.

We can also identify the anomaly detection classifier threshold in order to guarantee that stealth attacks will not cause catastrophic damages (Figure 6). By implementing a threshold below 1.8 we were able prevent undetected attacks from elevating the pressure of a chemical reactor beyond safety levels [2] (see Figure 7).

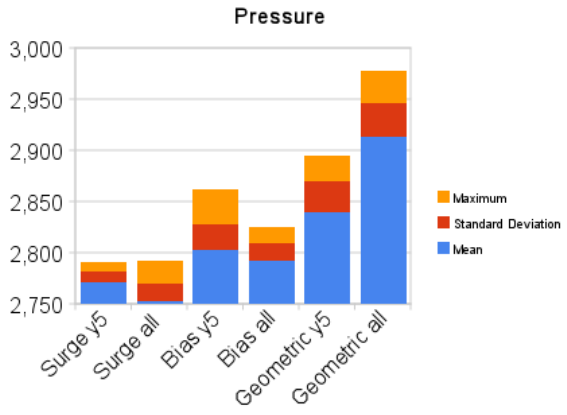

Figure 7: With appropriate thresholds, we can prevent the pressure in the tank to reach 3,000kPa with undetected attacks.

References

[1] A.A. Cardenas, S. Amin, and S. Sastry. Research Challenges for the Security of Control Systems. In *3rd USENIX workshop on Hot Topics in Security (HotSec '08). Associated with the 17th USENIX Security Symposium.*, July 2008.

[2] Alvaro A Cárdenas, Saurabh Amin, Zong-Syun Lin, Yu-Lun Huang, Chi-Yen Huang, and Shankar Sastry. Attacks against process control systems: risk assessment, detection, and response. In *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, pages 355–366. ACM, 2011.

[3] Alvaro A. Cardenas, Saurabh Amin, and Shankar Sastry. Secure control: Towards survivable cyber-physical systems. In *Proceedings of the First International Workshop on Cyber-Physical Systems.*, June 2008.

[4] Yao Liu, Peng Ning, and Michael K Reiter. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security (TISSEC)*, 14(1):13, 2011.

[5] Daisuke Mashima and Alvaro A Cárdenas. Evaluating electricity theft detectors in smart grid networks. In *Research in Attacks, Intrusions, and Defenses*, pages 210–229. Springer, 2012.

[6] Mark Zeller. Myth or reality—does the aurora vulnerability pose a risk to my generator? In *Protective Relay Engineers, 2011 64th Annual Conference for*, pages 130–136. IEEE, 2011.