Multistage Downstream Attack Detection in a Cyber Physical System

Rizwan Qadeer[1], Carlos Murguia[1], Chuadhry Mujeeb Ahmed[1], and Justin Ruths[2]

[1] Singapore University of Technology and Design, ⋆
[2] University of Texas Dallas, USA
{rizwan_qadeer,murguia_rendon}@sutd.edu.sg,
chuadhry@mymail.sutd.edu.sg,jruths@utdallas.edu

Abstract. We present an attack detection scheme for a water treatment system. We leverage the connectivity of two stages of the process to detect attacks downstream from the point of attack. Based on a mathematical model of the process, carefully crafted and executed attacks, are detected by deploying CUSUM and Bad-Data detectors. Extensive experiments are carried out and the results show the performance of the proposed scheme.

Keywords: CPS, CPS Security, ICS Security, Water Treatment Systems

1 Introduction

Cyber Physical Systems (CPS) are the integration of computing elements with the physical world [8]. The incorporation of communication networking technologies with legacy industrial control systems have exposed these to outside world. The secure operation of such systems requires novel security solutions as the threat models are different from cyber only systems [16]. Developing new theory to detect these and other attacks has been the focus of research in computer science, systems and control engineering, and other fields [2–5, 7, 9–11, 13, 14].

In this manuscript, we look into security threats in a water treatment testbed. These plants are spread over vast geographical areas, where the physical process is controlled based on remote sensor readings received over the communication networks. However, an attacker might change those sensor measurements which could lead to an undesired control. Several attacks have been reported on water systems in ICS-CERT report [6]. Most of the control theoretic approaches on secure CPS are based on the dynamic system model of the physical process. A residual signal is obtained by subtracting the sensor measurements from the

⋆ This work was supported in part by the National Research Foundation (NRF), Prime Ministers Office, Singapore, under its National Cybersecurity R&D Programme (Award No. NRF2014NCR-NCR001-040) and administered by the National Cybersecurity R&D Directorate.

sensor estimates (obtained using the system model). An anomaly is detected based on the statistical properties of this residual signal. A large proportion of the literature considers attacks that are executed and attempted to be detected on the same portion of the system, however, many CPS systems are large-scale multistage processes in which the whole process is subdivided into several interconnected stages. Typically each stage is dependent on the previous stage of the plant, thus it is interesting to model systems at the multistage level [1, 7, 9, 13]. For this case study, we work on a 6 stage water treatment testbed as explained in section 4. In the work presented here, we show the ability to detect attacks that occur in previous stages of the plant, thereby exploiting the coupling between the stages through the physical process. By extensive experimentation on a real testbed, we have shown that due to multistage combined estimation, the detectors on either stage would detect the executed attacks on another different stage. Two major contribution of our work are, (a): Proposed a multistage attack detection scheme, (b): Implementation of the proposed scheme on a real world testbed.

2 Background and System Model

We consider a Linear Time Invariant (LTI) stochastic process of the form:

$$\begin{cases} x(t_{k+1}) = Fx(t_k) + Gu(t_k) + v(t_k), \\ \quad y(t_k) = Cx(t_k) + \eta(t_k), \end{cases} \tag{1}$$

with sampling time-instants t_k, $k \in \mathbb{N}$, state $x \in \mathbb{R}^{\mathrm{n}}$, measured output $y \in \mathbb{R}^{\mathrm{m}}$, control input $u \in \mathbb{R}^{\mathrm{l}}$, matrices F, G, and C of appropriate dimensions, and i.i.d. multivariate zero-mean Gaussian noises $v \in \mathbb{R}^{\mathrm{n}}$ and $\eta \in \mathbb{R}^{\mathrm{m}}$ with covariance matrices $R_1 \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$, $R_1 \geq 0$ and $R_2 \in \mathbb{R}^{\mathrm{m}\times\mathrm{m}}$, $R_2 \geq 0$, respectively. The initial state $x(t_1)$ is assumed to be a zero-mean Gaussian random vector with covariance matrix $R_0 \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$, $R_0 \geq 0$. The processes $v(t_k)$, $k \in N$ and $\eta(t_k)$, $k \in N$ and the initial condition $x(t_1)$ are mutually independent. At the time-instants t_k, $k \in N$, the output of the process $y(t_k)$ is sampled and transmitted over a communication channel. In this paper, we focus on attacks on sensor measurements by spoofing the signals coming from the sensors to the controller. After each transmission and reception, the attacked output $\bar{y}$ takes the form:

$$\bar{y}(t_k) := y(t_k) + \delta(t_k) = Cx(t_k) + \eta(t_k) + \delta(t_k), \tag{2}$$

where $\delta(t_k) \in \mathbb{R}^{\mathrm{m}}$ denotes additive sensor attacks. Define $x_k := x(t_k)$, $u_k := u(t_k)$, $v_k := v(t_k)$, $y_k := y(t_k)$, $\eta k := \eta(t_k)$, and $\delta_k := \delta(t_k)$.

Residual-based detection mechanisms require an estimator of the system state; here we use the steady state Kalman Filter:

$$\hat{x}_{k+1} = F\hat{x}_k + Gu_k + L(\bar{y}_k - C\hat{x}_k), \tag{3}$$

with estimated state $\hat{x}_k \in \mathbb{R}^{\mathrm{n}}$, $\hat{x}_1 = E[x(t_1)]$, where $E[\cdot]$ denotes expectation, and gain matrix $L \in \mathbb{R}^{\mathrm{n}\times\mathrm{m}}$. The estimation error, $e_k := x_k - \hat{x}_k$, is governed by the following difference equation

$$e_{k+1} = \big(F - LC\big)e_k + v_k - L\eta_k - L\delta_k. \tag{4}$$

If pair (F, C) is detectable, the observer gain L can be selected such that $(F-LC)$ is Schur. Moreover, under detectability of (F, C), the covariance matrix $P_k := E[e_k e_k^T]$ converges to steady state (in the absence of attacks) in the sense that $\lim_{k\to\infty} P_k = P$ exists. For $\delta_k = \mathbf{0}$ and given L (such that $(F - LC)$ is Schur), it can be verified that the asymptotic covariance matrix $P = \lim_{k\to\infty} P_k$ is given by the solution P of the following Lyapunov equation: $(F - LC)P(F - LC)^T - P + R_1 + LR_2L^T = \mathbf{0}$ where $\mathbf{0}$ denotes the zero matrix of appropriate dimensions. It is assumed that the system has reached steady state before an attack occurs.

The estimator predictions are compared with sensor measurements $\bar{y}_k$ which potentially include attacks. If the difference between what is measured and the estimation is larger than expected, there may be a fault in or attack on the system. Define the residual random sequence r_k, $k \in \mathbb{N}$ as

$$r_k := \hat{y}_k - C\hat{x}_k = Ce_k + \eta_k + \delta_k, \tag{5}$$

For this residual, we formulate a one-sided hypothesis where we either accept or reject the null hypothesis that there are no attacks, in which case, the distribution of the residual is zero mean with the attack-free variance.

2.1 Detection Methods

We consider a dedicated detector on each sensor. Throughout the rest of this paper we will reserve the index i to denote the sensor/detector, $i \in \mathcal{I} := \{1, 2, \ldots, m\}$. With C_i being the i-th row of C and $\eta_{k,i}$ and $\delta_{k,i}$ denoting the i-th entries of η_k and δ_k, respectively. We propose the absolute value of the entries of the residual sequence as distance measures:

$$z_{k,i} := |r_{k,i}| = |y_{k,i} - C_i x_{k,i} + \delta_{k,i}| = |C_i e_k + \eta_{k,i} + \delta_{k,i}|. \tag{6}$$

Note that, if there are no attacks, $|r_{k,i}|$ follows a *half-normal distribution* [15].

CUSUM Detector: $S_{k,i} = 0,\ i \in l := \{1, 2, ..., m\}$,

$$\begin{cases} S_{k,i} = \max(0, S_{k-1,i} + |r_{k,i}| - b_i), & \text{if } S_{k-1,i} \leq \tau_i, \\ S_{k,i} = 0 \text{ and } \bar{k}_i = k - 1, & \text{if } S_{k-1,i} > \tau_i, \end{cases} \tag{7}$$

with bias $b_i \in \mathbb{R}_{>0}$, detection threshold $\tau_i \in \mathbb{R}_{>0}$, and alarm time(s) $\bar{k}_i$. The idea is that the test sequence $S_{k,i}$ accumulates $|r_{k,i}|$ and alarms are triggered when $S_{k,i}$ exceeds the threshold τ_i. Once the the bias is chosen, the threshold τ_i must be selected to fulfill a desired false alarm rate A_i^* [12].

Bad-Data Detector:

$$\text{If } \ |r_{k,i}| > \alpha_i, \qquad \bar{k_i} = k, i \in I. \tag{8}$$

where $\alpha_i \in \mathbb{R}_{>0}$ is the detection threshold and $\bar{k}_i$ are the alarms time(s). In this case, the idea is that alarms are triggered if $|r_{k,i}|$ exceeds the threshold α_i. Similar to the CUSUM procedure, the parameter α_i is selected to satisfy a required false alarm rate A_i^*.

3 Attacker Model

In this section, we introduce the attacks launched on the system. A usual attacker model for CPSs encompasses the *intentions and goals* of the attacker [16]. Attacker's intentions may vary from damaging components to changing a system property or performance degradation. It is assumed that the attacker has access to real-time sensor measurements. It also has perfect knowledge of the system dynamics, the control inputs, and the implemented detection procedures. We launch attacks on the two tanks (Tank-A, Tank-B) subsystem of the SWaT as shown in Figure 1. We consider a man-in-the-middle (*MitM*) attacker profile [17]. This attacker is able to get access to level sensor readings from Tank-A and Tank-B in real-time and inject signals. Three attacks (corresponding to three different injected signals) are considered and implemented on Tank-A of the real water treatment facility:

Constant Bias Injection Attack: In such an attack, the attacker adds constant offsets to true sensor measurements, i.e., $\delta_{k,i} = \bar{\delta}_i \in \mathbb{R}$. Thus, the controller receives an attacked sensor measurement of the form $\bar{y}_{k,i} = y_{k,i} + \bar{\delta}_i$, where $\bar{\delta}_i$ denotes the false data injected by the attacker to sensor i. As we will see later in results section, the constant bias attack is easily detected using the proposed detection methods.

Zero-Alarm Attack for Bad-Data Detector: This attack is designed to stay undetected by the Bad-Data detectors. Because the attacker knows the system dynamics, has access to sensor readings, and knows the detector parameters, it is able to inject false data into real-time measurements and stay undetected. Consider the Bad-Data procedure and write (8) in terms of the estimated state $\hat{x}_k$:

$$|r_{k,i}| = |y_{k,i} - C_i\hat{x}_{k,i} + \delta_{k,i}| \leq \alpha_i, \quad i \in \mathcal{I}. \tag{9}$$

By assumption, the attacker has access to $y_{k,i} = C_i y_k + \eta_{k,i}$. Moreover, given its perfect knowledge of the observer, the opponent can compute the estimated output $C_i\hat{x}_k$ and then construct $y_{k,i} - C_i\hat{x}_{k,i}$. It follows that

$$\delta_{k,i} = C_i\hat{x}_{k,i} - y_{k,i} + \alpha_i - \epsilon_i,\, (\alpha_i > \epsilon_i) \rightarrow |r_{k,i}| = \alpha_i - \epsilon_i, \quad i \in \mathcal{I}, \tag{10}$$

is a feasible attack sequence given the capabilities of the attacker. The constant $\epsilon_i > 0$ is a small positive constant introduced to account for numerical precision. These attacks maximize the damage to the CPS by immediately saturating and maintaining $|r_{k,i}|$ at the constant $\alpha_i - \epsilon_i$. Therefore, for this attack, the sensor measurements received by the controller take the form:

$$\bar{y}_{k,i} = C_i\hat{x}_{k,i} + \alpha_i - \epsilon_i. \tag{11}$$

Zero-Alarm Attack for CUSUM Detector: This attack is designed to stay undetected by the CUSUM detectors. Consider the CUSUM procedure and write (7) in terms of the estimated state $\hat{x}_k$:

$$S_{k,i} = \max(0, S_{k-1,i} + |y_i - C_i\hat{x}_k + \delta_{k,i}| - b_i), \tag{12}$$

if $S_{k-1,i} \leq \tau_i$ and $S_{k,i} = 0$ if $S_{k-1,i} > \tau_i$. As with the Bad-Data procedure, we look for attack sequences that immediately saturate and then maintain the CUSUM statistic at $S_{k,i} = \tau_i - \epsilon_i$ where ϵ_i ($\min(\tau_i, b_i) > \epsilon_i > 0$) is a small positive constant introduced to account for numerical precision. Assume that the attack starts at some $k = k^* \geq 1$ and $S_{k^*-1,i} \leq \tau_i$, i.e., the attack does not start immediately after a false alarm. Consider the attack:

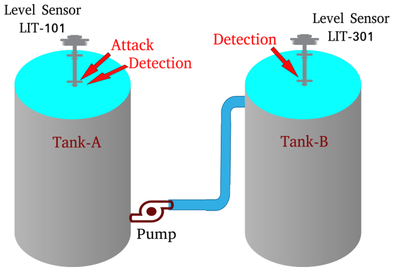
$$\delta_{k,i} = \begin{cases} \tau_i - \epsilon_i + b_i - y_i + C_i\hat{x}_k - S_{k-1,i}, & k = k^*, \\ b_i - y_i + C_i\hat{x}_k, & k > k^*. \end{cases} \tag{13}$$

This attack accomplishes $S_{k,i} = \tau_i - \epsilon_i$ for all $k \geq k^*$ (thus zero alarms). Note that the attacker can only induce this sequence by exactly knowing $S_{k^*-1,i}$, i.e., the value of the CUSUM sequence one step before the attack. This is a strong assumption since it represents a real-time quantity that is not communicated over the communication network. Even if the opponent has access to the parameters of the CUSUM, (b_i, τ_i), given the stochastic nature of the residuals, the attacker would need to know the complete history of observations (from when the CUSUM was started) to be able to reconstruct $S_{k^*-1,i}$ from data. This is an inherent security advantage in favor of the CUSUM over static detectors like the Bad-Data or Chi-Squared. Nevertheless, for evaluating the worst case scenario, we assume that the attacker has access to $S_{k^*-1,i}$. Therefore, for this attack, the sensor measurements received by the controller take the form:

$$\bar{y}_{k,i} = \begin{cases} C_i\hat{x}_{k,i} + \tau_i - \epsilon_i + b_i - S_{k-1,i} - \epsilon_i, & k = k^*, \\ C_i\hat{x}_{k,i} + b_i, & k > k^*. \end{cases} \tag{14}$$

4 Experimentation Setup

Majority of work on attack detection has considered a single stage for attack and detection (e.g., see [18]). Here, we evaluate the situation of using multiple detectors throughout the process while carrying out a spoofing attack on only one point. In this case we setup the attack on LIT-101 and then we implement a detection mechanism on this tank (LIT-101 at Tank-101) and also on the second tank (LIT-301 of Tank-301). The challenge in using a process-wide detector is that we require a model that captures not only each stage individually, but also the physical coupling caused by their interconnection. This experiment considers possibly the most obvious of this sort of interconnection and dependency between stages, in the sense that the water out-flow from Tank-101 (Tank-A) should equal the water in-flow to Tank-301 (Tank-B). We can see an illustration of this



Parameter	Tank-A	Tank-B
α	4.61×10^{-4}	4.59×10^{-4}
τ	1.60×10^{-4}	1.48×10^{-4}
bias b	3.27×10^{-4}	3.26×10^{-4}
$\mathcal{A}^*$	0.025	0.04

Fig. 1. Two-Tank Illustration: Tank-101(A), Tank-301(B). The adjoining table reports the parameters for both detectors.

scenario in Fig. 1. We model the water level and sensor measurements of the two tanks using the following difference equations:

$$\begin{cases} x_{k+1,1} = x_{k,1} + u_{k,1} - u_{k,2}, \\ x_{k+1,2} = x_{k,2} + u_{k,2} - u_{k,3}, \end{cases} \qquad \begin{cases} y_{k,1} = x_{k,1} + \eta_{k,1}, \\ y_{k,2} = x_{k,2} + \eta_{k,2}, \end{cases} \tag{15}$$

where $x_{k,j}$, $j = 1,2$ is the water level at tank j, $u_{k,1}$ and $u_{k,2}$ denote water flowing in and out of tank one, respectively, $u_{k,3}$ is the water flowing out of tank two, and $\eta_{k,j}$ denotes sensor noise. Then, the model of the coupled tanks is of the form (1) with matrices:

$$\left\{ F = R_2 = C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, G = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix}, R_1 = R_0 = \mathbf{0}. \right. \tag{16}$$

Having the system model, we can construct a Luenberger observer of the form (3) to estimate the state of the system. The observer matrix L is selected such that the matrix $F - LC$ (with F and C as in (16)) is Schur and its eigenvalues are at 0.5:

$$L = \begin{pmatrix} 0.35 & 0.15 \\ -0.15 & 0.65 \end{pmatrix}.$$

For both detectors, the thresholds (and biases for the CUSUM) have to be selected to satisfy desired false alarm rates $\mathcal{A}_j^*$. These parameters are selected according to the results in [12] to satisfy a false alarm rate of approximately $\mathcal{A}_1^* = \mathcal{A}_2^* = 0.025$ for both detectors. For our combined detector, we test the obtained detector parameters (shown in Table 1) for both the Bad-Data and the CUSUM procedures. We verify by experimenting that for the given parameters, the alarms raised by the detectors converge to $\mathcal{A}^* = 0.04$ (approximately) in the absence of attacks.

5 Performance of Proposed Detectors

We executed the three types of attacks introduced in Section 3 with a combined detection procedure running on Tank-A and Tank-B simultaneously.

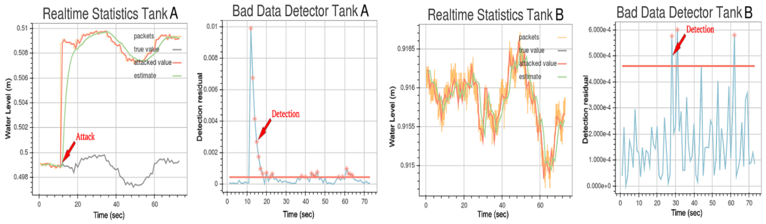


Fig. 2. Constant bias attack detection by combined Bad-Data detector

Constant Bias Attack Detection Figure 2 shows the water level at the tanks when the system is under a constant bias attack of $\bar{\delta}_1 = 0.01$m. The PLC received this attacked measurement value with Bad-Data detectors running on both tanks. The true value (plotted in gray) of the level at Tank-A is about 0.5m. This true level remains constant throughout the attack and the inlet pump and valve are switched OFF. The attack is launched at $k = 11$s (time instant in plot) and the Bad-Data detector monitoring Tank-A detects it immediately. The Bad-Data detector monitoring Tank-B detects the attack at $k = 28$s. This proves that the combined detection procedure for Bad-Data detector works well. Furthermore this attack was also detected by the CUSUM detectors running at Tank-A and Tank-B.

Zero-Alarm Attack for Bad-Data Detector We now test a zero-alarm Bad-Data attack on Tank-A. Both the detector types (i.e., Bad-Data detector and CUSUM detector) monitor Tank-A and Tank-B. In our proposed scheme, when an attack is launched at a single stage, it can be detected by a detector running on another stage. Here we launch a zero-alarm attack against the Bad-Data detector at Tank-A, and found that it can be detected only by CUSUM detector at Tank-A and by both detectors (CUSUM and Bad-Data) running at Tank-B. For The attacker to remain undetected at Tank-A he have to spoof the sensor value according to section 3.

Zero-Alarm Attack for Bad-Data and CUSUM Detector The last attack type which we executed is the zero-alarm attack for Bad-Data and CUSUM detector. Since this attack is designed to raise no alarms for the Bad-Data or the CUSUM detectors, neither detector on Tank-A detects the attack. The attacker has the complete knowledge of the detectors running on Tank-A, so he can deviate the level of the tank in such a way that Bad-Data detector and CUSUM detector at Tank-A would not be able to detect it, but the combined estimate and detection of the two-tank multistage process makes it possible to detect this attack by the detectors at Tank-B. The result is shown in Fig. 3.

6 Conclusion

In this paper we have provided a real-life experimental case-study about how a multistage detection procedure could be very useful. We showed that proper

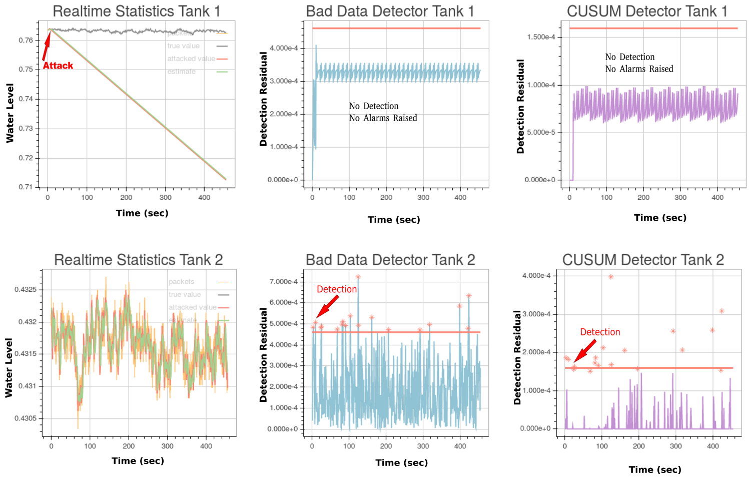


Fig. 3. Zero-alarm Bad-Data/CUSUM detection by combined detectors at Tank-B.

modeling of the system and the selection of right parameters for detection threshold are very important. Our study points out the limitations of statistical anomaly detectors towards stealthy attacks which are intelligently designed to raise no alarms (zero-alarm attacks). However, we can still use these detection methods if physics of the system is properly integrated in the model for system dynamics. Due to state inter-dependencies, an attacker can hide itself in one stage but it's effects can be seen in the following stages. Our results show that it is possible to detect zero-alarm attacks using the proposed scheme.

References

1. Ahmed, C.M., A.Sridhar, Aditya, M.: Limitations of state estimation based cyber attack detection schemes in industrial control systems. In: IEEE Smart City Security and Privacy Workshop, CPSWeek (2016)
2. Ahmed, C.M., Murguia, C., Ruths, J.: Model-based attack detection scheme for smart water distribution networks. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. pp. 101–113. ASIA CCS '17, ACM, New York, NY, USA (2017), http://doi.acm.org/10.1145/3052973.3053011
3. Bai, C.Z., Gupta, V.: On kalman filtering in the presence of a compromised sensor: Fundamental performance bounds. In: 2014 American Control Conference. pp. 3029–3034. IEEE (2014)
4. Bai, C.Z., Pasqualetti, F., Gupta, V.: Security in stochastic control systems: Fundamental limitations and performance bounds. In: 2015 American Control Conference (ACC). pp. 195–200. IEEE (2015)
5. Cardenas, A.A., Amin, S., Lin, Z.S., Huang, Y.L., Huang, C.Y., Sastry, S.: Attacks against process control systems: risk assessment, detection, and response. In: Proceedings of the 6th ACM symposium on information, computer and communications security. pp. 355–366. ACM (2011)

6. ICS-CERT: Ics-mm201408: May-august 2014. Report no., U.S. Department of Homeland Security-Industrial Control Systems-Cyber Emergency Response Team, Washington, D.C. Available online at https://ics-cert.us-cert.gov. (2014)
7. Kwon, C., Liu, W., Hwang, I.: Security analysis for cyber-physical systems against stealthy deception attacks. In: American Control Conference (ACC). pp. 3344–3349 (2013)
8. Lee, E.A.: Cyber physical systems: Design challenges. In: EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2008-8. http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS- 2008-8.html (Jan 2008)
9. Miao, F., Zhu, Q., Pajic, M., Pappas, G.J.: Coding sensor outputs for injection attacks detection. In: IEEE conference in Decision and Control (CDC). pp. 5776–5781 (2014)
10. Mo, Y., Garone, E., Casavola, A., Sinopoli, B.: False data injection attacks against state estimation in wireless sensor networks. In: 49th IEEE Conference on Decision and Control (CDC). pp. 5967–5972. IEEE (2010)
11. Mujeeb, C.A., Mathur, A.P.: Hardware identification via sensor fingerprinting in a cyber physical system. In: Software Quality, Reliability and Security Companion (QRS-C), 2017 IEEE International Conference on. pp. 517–524. IEEE (2017)
12. Murguia, C., Ruths, J.: Characterization of a cusum model-based sensor attack detector. In: 55th IEEE Conference on Decision and Control Conference (CDC) (2016)
13. Murguia, C., Ruths, J.: Cusum and chi-squared attack detection of compromised sensors. In: 2016 IEEE Conference on Control Applications (CCA). pp. 474–480 (Sept 2016)
14. Pasqualetti, F., Dörfler, F., Bullo, F.: Attack detection and identification in cyber-physical systems. IEEE Transactions on Automatic Control 58(11), 2715–2729 (2013)
15. Ross, M.: Introduction to Probability Models, Ninth Edition. Academic Press, Inc., Orlando, FL, USA (2006)
16. Sridhar, A., Aditya, M.: Generalized attacker and attack models for cyber physical systems. In: 40th IEEE COMPSAC (2016)
17. Urbina, D., Giraldo, J., Tippenhauer, N.O., Cardenas, A.: Attacking Fieldbus Communications in ICS : Applications to the SWaT Testbed. Singapore Cyber-Security Conference (SG-CRC) 14, 75–89 (2016)
18. Urbina, D.I., Giraldo, J., Cardenas, A.A., Tippenhauer, N.O., Valente, J., Faisal, M., Ruths, J., Candell, R., Sandberg, H.: Limiting the Impact of Stealthy Attacks on Industrial Control Systems. Acm Ccs 3(iii) (2016)