2018 European Control Conference (ECC)
June 12-15, 2018. Limassol, Cyprus

Quantifying the Impact of Cyber-Attack Strategies for Control Systems Equipped with an Anomaly Detector

Jezdimir Milošević, David Umsonst, Henrik Sandberg, and Karl Henrik Johansson[1]

Abstract— **Risk assessment is an inevitable step in the implementation of cost-effective security strategies for control systems. One of the difficulties of risk assessment is to estimate the impact cyber-attacks may have. This paper proposes a framework to estimate the impact of several cyber-attack strategies against a dynamical control system equipped with an anomaly detector. In particular, we consider denial of service, sign alternation, rerouting, replay, false data injection, and bias injection attack strategies. The anomaly detectors we consider are stateless, cumulative sum, and multivariate exponentially weighted moving average detectors. As a measure of the attack impact, we adopt the infinity norm of critical states after a fixed number of time steps. For this measure and the aforementioned anomaly detectors, we prove that the attack impact for all of the attack strategies can be reduced to the problem of solving a set of convex minimization problems. Therefore, the exact value of the attack impact can be obtained easily. We demonstrate how our modeling framework can be used for risk assessment on a numerical example.**

I. Introduction

The necessity of improved cyber-security of industrial control systems has been demonstrated by a number of high-profile cyber-attacks [1]–[3], as well as by numerous research studies [4]–[8]. The overall recommendation for improving the cyber-security of these systems is to implement the so called defense-in-depth strategies, which consist of several layers of security measures [9]. Unfortunately, the large amount of legacy equipment within many industrial control systems, combined with their complexity and real-time requirements, can make the deployment and maintenance of security measures costly.

In order to implement defense-in-depth strategies in a cost-effective manner, it is crucial to conduct risk assessment prior to deployment of security measures [9]. The first step of risk assessment is to identify security vulnerabilities. Subsequently, the likelihood of each vulnerability being exploited, and the impact that may occur in that case are estimated. Once the risk assessment is completed, one can prioritize which vulnerabilities should be treated first based on estimates of the impact and likelihood. Given that cyber-attacks against control systems may endanger the physical world, it is natural to use models of physical dynamics to estimate the attack impact.

In this paper, we focus on the problem of estimating the impact of cyber-attacks in control systems equipped with various types of anomaly detectors. Aspects of this problem were earlier investigated in studies [10]–[12]. Cárdenas *et al.* [10] considered a control system equipped with a *cumulative sum* (CUSUM) anomaly detector, and proposed several attack strategies that can be used to quantify the attack impact. Ahmed *et al.* [11] investigated the performance of *stateless* and CUSUM anomaly detectors in presence of false data injection and zero alarm attacks. In order to compare different types of anomaly detectors, Urbina *et al.* [12] introduced a novel metric and an attack model that can be used for that purpose.

We identify two main directions in which these studies can be extended. Firstly, the aforementioned works mostly considered an attacker that is very resourceful. For instance, the attacker possesses full model knowledge, controls a considerable number of components within the system, and is able to inject arbitrary signals to sensors and actuators it controls. Simpler attack strategies, which are also more likely to happen, were not considered. Some of these strategies include denial of service [13], [14], rerouting [15], [16], sign alternation [17], [18], and replay [19] attacks. Secondly, it is often unclear from the literature how the worst case attack impact is calculated. The problem of estimating the attack impact usually represents a constrained maximization problem, and algorithms that return the *exact* solution of these problems are rarely available. However, in previous work [20], the authors introduced infinity norm of the states as a measure of impact, and formulated the problem of finding the attack impact as an optimization problem that yields the exact solution. Thus, we adopt this metric to quantify the impact of the attack strategies we consider in this paper.

The contributions of this paper are as follows. Firstly, we propose a unified framework for quantifying the attack impact in control systems equipped with an anomaly detector. Our framework is flexible, and can be used to quantify the impact of both simple attack strategies such as denial of service, sign alternation, rerouting, replay, and bias injection, but also more complex false data injection attacks with full model knowledge. Secondly, our analysis is valid for both stateless and CUSUM anomaly detectors observed in previous work, but we also extend our analysis to the multivariate exponentially weighted moving average (MEWMA) detector [21]. Thirdly, for the impact measure introduced in [20], we prove that the impact for all attack strategies and all considered detectors can be obtained by solving a

*This work was supported by the Swedish Civil Contingencies Agency through the CERCES project, the Swedish Research Council, Knut and Alice Wallenberg Foundation, and the Swedish Foundation for Strategic Research.

[1] The authors are with the Department of Automatic Control, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden. Emails: {jezdimir, umsonst, hsan, kallej} @kth.se.

978-3-9524-2699-9 ©2018 EUCA

set of convex minimization problems (Propositions 1–4). This implies that the exact value of the attack impact can easily be obtained, since the algorithms for solving convex minimization problems are well known. Finally, we illustrate on a numerical example of a chemical process how our framework can be used for risk assessment.

The remainder of the paper is organized as follows. In Section II, we introduce a model of the control system under attack, and models for anomaly detectors. In Section III, we introduce several attack strategies, and prove that the impact for these strategies can be obtained by solving a set of convex minimization programs. In Section IV, we introduce an illustrative example that demonstrates how the proposed framework can be used for assessing the impact of cyber-attacks. In Section V, we conclude the paper and outline some directions for future work.

II. Model Setup

We adopt the modeling framework introduced in [22], where the control system was modeled as an interconnection of the plant, the controller, the anomaly detector, and the attacker. In what follows, we provide more detailed models of each of these blocks.

A. Plant, Controller, and Anomaly Detector

The physical plant is modeled as

$$\mathcal{P}\colon \begin{cases} x(k+1) = A_p x(k) + B_p \tilde{u}(k) \\ \quad y(k) = C_p x(k) \end{cases} \tag{1}$$

where $x(k) \in \mathbb{R}^{n_x}$ is the state of the plant at time step k, $y(k) \in \mathbb{R}^{n_y}$ is the vector of sensor measurements, and $\tilde{u}(k) \in \mathbb{R}^{n_u}$ is the control input applied to the process.

The plant is controlled with a controller of the form

$$\mathcal{F}\colon \begin{cases} z(k+1) = A_f z(k) + B_f \tilde{y}(k) + K_f y_r(k) \\ \quad u(k) = C_f z(k) + D_f \tilde{y}(k) + E_f y_r(k) \end{cases} \tag{2}$$

where $z(k) \in \mathbb{R}^{n_z}$ is the state of the controller, $\tilde{y}(k) \in \mathbb{R}^{n_y}$ is the vector of sensor measurements used by the controller to calculate the control signal $u(k) \in \mathbb{R}^{n_u}$, and $y_r(k) \in \mathbb{R}^{n_{y_r}}$ is the bounded reference signal. In particular, we assume

$$-\delta_{y_r} \leq y_r(k) \leq \delta_{y_r} \tag{3}$$

where $\delta_{y_r} \in \mathbb{R}_+^{n_{y_r}}$ is the predefined bound. The assumption is that the controller is designed so that stability and acceptable performances are achieved in the absence of anomalies.

During the nominal operation, the signals $\tilde{y}(k)$ and $\tilde{u}(k)$ are equal to $y(k)$ and $u(k)$, respectively. However, because of an attack or a fault in the system, these values may differ. In order to detect these anomalies, an anomaly detector is used. The first step of the detection procedure is to calculate the so called residual signal. We consider a residual-generating filter of the form

$$\mathcal{D}\colon \begin{cases} s(k+1) = A_d s(k) + B_d u(k) + K_d \tilde{y}(k) \\ \quad r(k) = C_d s(k) + D_d u(k) + E_d \tilde{y}(k) \end{cases} \tag{4}$$

where $s(k) \in \mathbb{R}^{n_s}$ is the state of the filter, and $r(k) \in \mathbb{R}^{n_r}$ is the residual signal evaluated to detect potential anomalies. We assume that the filter is designed such that the following properties are satisfied:

1) the value of the residual $r(k)$ converges asymptotically to zero in absence of anomalies;
2) the residual $r(k)$ is sensitive to attacks and anomalies, and in case when $\tilde{u}(k) \neq u(k)$ and/or $\tilde{y}(k) \neq y(k)$, $r(k)$ is different from zero except in pathological cases such as zero dynamic attacks (see [22]).

These are standard assumptions adopted from the fault-diagnosis literature [23].

The second step of the detection procedure is to process the residual signal $r(k)$ to obtain a security metric $S(k+1)$. When this metric exceeds a certain threshold $\delta_r > 0$, an alarm is raised. How $S(k+1)$ is determined depends on the detector used. In this paper, we are focused on the following three anomaly detectors.

1) Stateless Detector: A stateless detector is defined as

$$S(k+1) = ||Q_r r(k)||_p^2$$

where $Q_r \in \mathbb{R}^{n_r \times n_r}$ represents a scaling matrix, and $||(.)||_p$ represents the p-norm. The common values for p used in the literature are 2 or ∞.

2) CUSUM Detector: The CUSUM detector is a stateful detector, which in its non-parametric form is defined as

$$S(k+1) = \max\{S(k) + ||Q_r r(k)||_p^2 - \delta, 0\}$$

where $\delta > 0$ is the forgetting factor. The metric S is reset to zero once an alarm occurs, that is, when $S(k+1) > \delta_r$.

3) MEWMA Detector: The MEWMA detector is another stateful detector, which is defined as

$$\begin{aligned} \tilde{S}(k+1) &= \beta Q_r r(k) + (1-\beta)\tilde{S}(k) \\ S(k+1) &= \frac{2-\beta}{\beta}||\tilde{S}(k+1)||_2^2 \end{aligned}$$

where $\beta \in (0, 1]$ is the forgetting factor. As for the CUSUM detector, $\tilde{S}$ is reset to zero if an alarm occurs.

B. System Under Attack

By exploiting some security vulnerability, the attacker is able to manipulate the subsets of sensors $\mathcal{V}_y \subseteq \{1, 2, \ldots, n_y\}$ and actuators $\mathcal{V}_u \subseteq \{1, 2, \ldots, n_u\}$. The influence of the attack on the signals $y(k)$ and $u(k)$ is modeled as

$$\tilde{y}(k) = y(k) + D_y a_y(k) \qquad \tilde{u}(k) = u(k) + D_u a_u(k) \tag{5}$$

where $a_y(k) \in \mathbb{R}^{n_{a_y}}$ represents the attack against sensors, $a_u(k) \in \mathbb{R}^{n_{a_u}}$ represents the attack against actuators, and the matrices $D_y \in \mathbb{R}^{n_y \times n_{a_y}}$ and $D_u \in \mathbb{R}^{n_u \times n_{a_u}}$ model the influence of attacks on actuators and sensors, respectively. We remark that the matrices D_y and D_u depend on the sets $\mathcal{V}_y$ and $\mathcal{V}_u$. If the attacker is able to manipulate the sensors measurements $\mathcal{V}_y = \{j_1, j_2, \ldots, j_{n_{a_y}}\}$, then the elements $(j_1, 1), (j_2, 2), \ldots, (j_{n_{a_y}}, n_{a_y})$ of the matrix D_y are equal to one, and the remaining elements are equal to zero. The matrix D_u is defined in an analogous way.

In order to formulate some of the attack strategies in a more compact form, we introduce the augmented vectors

$$x_e(k) = [x^T(k)\ z^T(k)\ s^T(k)]^T \qquad a(k) = [a_u^T(k)\ a_y^T(k)]^T$$

which represent the extended state of the system and extended attack vector, respectively. We denote the dimension of the vector $x_e(k)$ by $n_e = n_x + n_z + n_s$, and the dimension of the vector $a(k)$ by $n_a = n_{a_u} + n_{a_y}$. By combining (1), (2), (4), and (5), the dynamics of the system under attack can be expressed as

$$\begin{aligned} x_e(k+1) &= A_e x_e(k) + B_e a(k) + K_e y_r(k) \\ r(k) &= C_e x_e(k) + D_e a(k) + E_e y_r(k) \end{aligned} \tag{6}$$

where[1]

$$\begin{aligned} A_e &= \begin{bmatrix} A_p + B_p D_f C_p & B_p C_f & \mathbf{0}_{n_x \times n_s} \\ B_f C_p & A_f & \mathbf{0}_{n_z \times n_s} \\ (K_d + B_d D_f) C_p & B_d C_f & A_d \end{bmatrix} \\ B_e &= \begin{bmatrix} B_p D_u & B_p D_f D_y \\ \mathbf{0}_{n_z \times n_{a_u}} & B_f D_y \\ \mathbf{0}_{n_s \times n_{a_u}} & (B_d D_f + K_d) D_y \end{bmatrix} \quad K_e = \begin{bmatrix} B_p E_f \\ K_f \\ B_d E_f \end{bmatrix} \\ C_e &= \begin{bmatrix} (D_d D_f + E_d) C_p & D_d C_f & C_d \end{bmatrix} \\ D_e &= \begin{bmatrix} \mathbf{0}_{n_r \times n_{a_u}} & (D_d D_f + E_d) D_y \end{bmatrix} \qquad E_e = D_d E_f. \end{aligned}$$

III. Quantifying Attack Impact

The main goal of this paper is to estimate the impact that can occur once the attacker exploits some security vulnerability. To do that, we first introduce the criterion based on which we characterize the impact of cyber-attacks. We then introduce several attack strategies, and prove that the impact of the attack in all cases can be obtained by solving a set of convex minimization problems.

A. Criterion for Characterizing the Attack Impact

In order to estimate the impact of possible attacks, we use the concept of critical states. Let $Q_c \in \mathbb{R}^{n_c \times n_e}$ be a matrix that maps the extended state vector to a subset of critical states

$$x_c(k) = Q_c x_e(k).$$

These critical states may model levels in tanks with hazardous materials that must not be overflown, or pressures that should not exceed some safety limit. From the perspective of the defender, we want to prevent the attacker in driving any of these states far from the steady state. Therefore, one way to estimate the impact would be to check if the attacker can drive the critical states far from the steady state during some time interval. For simplicity, we assume that the attack starts at $k = 0$, and we observe how far the attacker can drive the critical states in the time interval $[0, N]$. The attack impact $I(\mathcal{V}_y, \mathcal{V}_u)$ can then be defined as $I(\mathcal{V}_y, \mathcal{V}_u) = ||x_c(N)||_\infty$.

Besides driving the critical states as far as possible from the steady state, we are interested to check if the attack can stay undetected by an anomaly detector. The assumption is that if we are able to detect the attack, we can start safety procedures in order to prevent the attacker from causing large damage to the system. Therefore, we also want to check if the attack can be conducted without triggering an alarm. Hence, we impose the constraints $S(k+1) \leq \delta_r, k = 0, \ldots, N$, where $S(k+1)$ is calculated by using one of the detectors we introduced in the previous section.

From (6), the system has two input signals–the reference signal $y_r(k)$ and the attack signal $a(k)$. Thus, during the attack, the system trajectory depends on both of these signals. Given that the reference signal is often a constant signal, we adopt the following standing assumption.

Assumption 1: The reference signal is constant and equal to the reference prior to attack $y_r(k) = y_r$, $k = 0, 1, \ldots, N$. The system has reached steady state before the attack happens, which implies $S(0) = 0, r(0) = 0, x_e(0) = Q_{ss} y_r$, where $Q_{ss} \in \mathbb{R}^{n_e \times n_{y_r}}$ represents the steady state gain of the transfer function from the reference signal $y_r(k)$ to the extended state $x_e(k)$ of the system. ■

In what follows, we are performing off-line analysis of the attack impact. Thus, the exact value of the reference signal from the interval (3) at the beginning of the attack is unknown to us. For this reason, throughout the paper we identify the worst possible value of the reference y_r when estimating the attack impact.

B. Attack Strategies

The attacker can use different attack strategies in order to conduct an attack. In this paper, we observe denial of service, rerouting, sign alternation, replay, false data injection, and bias injection attack strategies. We show that for all of these attack strategies and for all of the anomaly detectors we consider, the impact can be obtained by solving an optimization problem of the following form.

Problem 1:

$$\begin{aligned} &\underset{d}{\text{maximize}}\ ||Td||_\infty \\ &\text{subject to}\ \ f_i(d) \leq \delta_i \quad i = 1, \ldots, n_i \end{aligned}$$

where $d \in \mathbb{R}^{n_d}$ is the decision variable, T is a matrix from $\mathbb{R}^{n_c \times n_d}$, and $f_i(d) : \mathbb{R}^{n_d} \to \mathbb{R}$ are symmetric[2] convex functions. A convenient property of problems of this type is that the optimal value can be obtained by solving n_c convex minimization problems. Given that algorithms that return the optimal value of convex minimization problems are well known, we are able to use these algorithms for finding the *exact* value of the attack impact. In the following lemma, we prove the aforementioned claim. We remark that a less general result was introduced in [20].

Lemma 1: Let I be the optimal value of Problem 1, and I' be the optimal value of the following set of n_c convex minimization problems

$$\begin{aligned} \underset{l \in \{1,\ldots,n_c\}}{\text{minimize}}\ \ &\underset{d}{\text{minimize}}\ \ -T(l,:)d \\ &\text{subject to}\ \ f_i(d) \leq \delta_i \ \ i = 1, \ldots, n_i \end{aligned}$$

[1] $\mathbf{0}_{n \times m}$ represents the matrix filled with zeros with n rows and m columns, while $\mathbf{I}_n$ represents the identity matrix with n rows and columns.

[2] A function $f(d)$ is symmetric if $f(d) = f(-d)$.

where $T(l,:)$ represents l-th row of the matrix T. Then the equality $I = |I'|$ holds.

Proof: Let d^* be an optimal solution of Problem 1, and let the optimal value of this problem be defined with

$$I = ||Td^*||_\infty = |T(l^*,:)d^*|$$

where l^* is the row of T for which the optimal value is achieved. Thus $-|T(l^*,:)d^*| \leq -T(l,:)d$ for every $l \in \{1, 2, \ldots, n_c\}$, and for every d that satisfies the constraints. Given that the constraints on d are equivalent for both of the problems, it follows that $-I \leq I'$. Assume that $-I < I'$. By the symmetry of the constraints, we have that both d^* and $-d^*$ are feasible points for both problems. However, that implies that either $T(l^*,:)d^*$ or $T(l^*,:)(-d^*)$ is less than 0. If we define $I'' := \min\{-T(l^*,:)d^*, T(l^*,:)d^*\}$ then it follows that $I'' = -I < I'$. This contradicts the assumption that I' is the optimal value of the problem stated in the lemma. Therefore, the only possibility is $-I = I'$, which concludes the proof. ■

In order to reduce the attack strategies to the form of Problem 1, we use that the detector constraints are convex and symmetric under a certain condition.

Lemma 2: Assume that $S(0) = 0$. If the residuals can be expressed as $r(k) = T_r(k)d$, where $T_r(k) \in \mathbb{R}^{n_r \times n_d}$ and $d \in \mathbb{R}^{n_d}$, then the constraints $S(k+1) \leq \delta_r$, $k = 0, \ldots, N$, represent convex and symmetric constraints in d for the stateless, CUSUM, and MEWMA detectors.

Proof: We first show that the stateless detector is convex and symmetric in d. By using the definition of the stateless detector, we have

$$S(k+1) = ||Q_r r(k)||_p^2 = ||Q_r T_r(k)d||_p^2 \leq \delta_r.$$

Since every norm is symmetric and convex, and the square of a convex function is convex, the stateless detector represents convex and symmetric constraint in d.

Using $r(k) = T_r(k)d$ in the definition of the CUSUM detector leads to

$$S(k+1) = \max\{S(k) + ||Q_r T_r(k)d||_p^2 - \delta, 0\}.$$

Since $||Q_r T_r(k)d||_p^2$ is symmetric in d, then $S(k+1)$ is also symmetric in d for every k. The proof that the CUSUM detector represents convex constraints follows the same lines as the proof of Proposition 2 in [20], for $p = 2$.

For the MEWMA detector we rewrite $\tilde{S}(k)$ in terms of d

$$\begin{aligned}\tilde{S}(k) &= \beta \sum_{i=0}^{k-1} (1-\beta)^{k-1-i} Q_r r(i)\\ &= \beta \sum_{i=0}^{k-1} (1-\beta)^{k-1-i} Q_r T_r(i) d.\end{aligned}$$

If d is replaced by $-d$, then $\tilde{S}(k)$ equals $-\tilde{S}(k)$. Thus, $S(k)$ represents a symmetric constraint in d due to the symmetry of the squared Euclidean norm. Given that $\tilde{S}(k)$ represents a linear transformation of d, it is a convex function. Hence, $S(k) = \frac{2-\beta}{\beta}||\tilde{S}(k)||_2^2$ is also convex in d for all k. ■

We now introduce the attack strategies, and prove that the problem of finding the attack impact can be reduced to the form of Problem 1 in all the cases.

1) Denial of Service Attack: In this attack strategy, the attacker starts blocking some of the signals of the sensors and actuators from reaching their destination. This can be achieved by making physical damage to devices, overflowing communication network with large amount of traffic, or jamming the network [13]. One possible way of modeling this type of attacks was suggested in [13], [14], where the control signals and measurements during the attack were modeled as

$$\tilde{u}(k) = \Lambda_u u(k) \qquad \tilde{y}(k) = \Lambda_y y(k) \tag{7}$$

where $\Lambda_u \in \mathbb{R}^{n_u \times n_u}$ and $\Lambda_y \in \mathbb{R}^{n_y \times n_y}$ are diagonal matrices defined as follows

$$\Lambda_u(i,i) = \begin{cases} 0 & i \in \mathcal{V}_u \\ 1 & i \notin \mathcal{V}_u \end{cases} \qquad \Lambda_y(i,i) = \begin{cases} 0 & i \in \mathcal{V}_y \\ 1 & i \notin \mathcal{V}_y. \end{cases} \tag{8}$$

By combining (1), (2), (4), and (7), the dynamics of the extended system under the denial of service attack can be expressed as

$$\begin{aligned} x_e(k+1) &= \tilde{A}_e x_e(k) + \tilde{B}_e y_r \\ r(k) &= \tilde{C}_e x_e(k) + \tilde{D}_e y_r \end{aligned} \tag{9}$$

where

$$\begin{aligned}
\tilde{A}_e &= \begin{bmatrix} A_p + B_p \Lambda_u D_f \Lambda_y C_p & B_p \Lambda_u C_f & \mathbf{0}_{n_x \times n_s} \\ B_f \Lambda_y C_p & A_f & \mathbf{0}_{n_z \times n_s} \\ (B_d D_f + K_d) \Lambda_y C_p & B_d C_f & A_d \end{bmatrix} \\
\tilde{C}_e &= \begin{bmatrix} (D_d D_f + E_d) \Lambda_y C_p & D_d C_f & C_d \end{bmatrix} \\
\tilde{B}_e &= \begin{bmatrix} B_p \Lambda_u E_f \\ K_f \\ B_d E_f \end{bmatrix} \qquad \tilde{D}_e = D_d E_f.
\end{aligned}$$

From (9), and by using the fact that $x_e(0) = Q_{ss} y_r$, the critical states after N steps and the residual signal after k steps can be expressed as

$$x_c(N) = T_x y_r \qquad r(k) = T_r(k) y_r \tag{10}$$

where

$$\begin{aligned}
T_x &= Q_c \left(\tilde{A}_e^N Q_{ss} + \sum_{i=0}^{N-1} \tilde{A}_e^i \tilde{B}_e \right) \\
T_r(k) &= \tilde{C}_e \left(\tilde{A}_e^k Q_{ss} + \sum_{i=0}^{k-1} \tilde{A}_e^i \tilde{B}_e \right) + \tilde{D}_e.
\end{aligned}$$

Note that the evolution of the system under the denial of service attack is only dependent on the value of the reference signal y_r. Thus, what we need to investigate is if there exists an y_r inside of the operating region defined by (3), such that the denial of service attack strategy drives some of the critical states far from the steady state while remaining undetected at the same time. Therefore, the problem of finding the worst case impact $I(\mathcal{V}_y, \mathcal{V}_u)$ in the case of this attack strategy can be formulated as the following optimization problem.

Problem 2:

$$\begin{aligned} \underset{y_r}{\text{maximize}} \quad & I\big(\mathcal{V}_y, \mathcal{V}_u\big) = ||T_x y_r||_\infty \\ \text{subject to} \quad & -\delta_{y_r} \le y_r \le \delta_{y_r} \\ & S(k+1) \le \delta_r \quad k = 0, \dots, N. \end{aligned}$$

In what follows, we prove that Problem 2 can be reduced to the form of Problem 1.

Proposition 1: Problem 2 is an instance of Problem 1 for the stateless, CUSUM, and MEWMA detectors.

Proof: The decision variable in Problem 2 is y_r, so $d = y_r$. The objective functions are of the same form, thus we only need to prove that all the constraints of the problem are convex and symmetric. Let $f_k(d) = S(k+1) \le \delta_r$, $k = 0, \dots, N$. We know from (10) that $r(k) = T_r(k)d$ for every k, so it follows from Lemma 2 that $f_0(d), ..., f_N(d)$ represent convex and symmetric constraints in d. It remains to prove that the reference constraint is symmetric and convex in d. Let Q_{y_r} be the diagonal matrix from $\mathbb{R}^{n_{y_r} \times n_{y_r}}$ whose elements are defined by $Q_{y_r}(i,i) = 1/\delta_{y_r}(i)$. We can represent the constraint (3) as $f_{y_r}(d) = ||Q_{y_r} d||_\infty \le 1$. This constraint is a convex and symmetric constraint in d due to infinity norm, which concludes the proof. ■

2) Rerouting Attacks: In this attack strategy, the attacker permutes the values of some of the measurements or control signals under its control. As stated in [16], the attacker can conduct this attack by physically re-wiring the sensor cables, or by modifying the sender's address. Thus, the control inputs and measurements during the rerouting attack are given by $\tilde{u}(k) = \Lambda_u u(k)$ and $\tilde{y}(k) = \Lambda_y y(k)$, where $\Lambda_u \in \mathbb{R}^{n_u \times n_u}$ and $\Lambda_y \in \mathbb{R}^{n_y \times n_y}$ are any permutation matrices that satisfy the following constraints

$$\Lambda_u(i,i) = 1 \;\; i \notin \mathcal{V}_u \qquad \Lambda_y(i,i) = 1 \;\; i \notin \mathcal{V}_y.$$

Note that the way we define $\tilde{u}(k)$ and $\tilde{y}(k)$ in this attack strategy is identical to the way we defined them for the denial of service attack strategy. The only difference is that Λ_u and Λ_y represent permutation matrices. Therefore, for fixed permutation matrices Λ_u and Λ_y, the problem of finding the worst case impact of the rerouting attack strategy can be reduced to Problem 2.

3) Sign Alternation Attack: In this attack strategy, the attacker simply flips the sign of the measurement and control signals under its control. Although simple, this attack can for instance turn negative feedback into positive, and in that way destabilize the system. Moreover, it was shown that in certain configurations with a Kalman filter, this attack strategy leads to strictly stealthy attacks [17], [18]. The control signal and measurement signal during the attack are given by $\tilde{u}(k) = \Lambda_u u(k)$ and $\tilde{y}(k) = \Lambda_y y(k)$, where $\Lambda_u \in \mathbb{R}^{n_u \times n_u}$ and $\Lambda_y \in \mathbb{R}^{n_y \times n_y}$ are in this case defined as

$$\Lambda_u(i,i) = \begin{cases} -1 & i \in \mathcal{V}_u \\ \;\;\,1 & i \notin \mathcal{V}_u \end{cases} \qquad \Lambda_y(i,i) = \begin{cases} -1 & i \in \mathcal{V}_y \\ \;\;\,1 & i \notin \mathcal{V}_y. \end{cases}$$

Therefore, the impact in case of this attack strategy can also be reduced to Problem 2, as it was the case with the denial of service and rerouting attack strategies.

4) Replay Attacks: This attack strategy is inspired by the well known Stuxnet malware [2]. The attacker keeps sending the steady state-sensor measurements from the sensors under its control, while at the same time applies malicious control signals to the actuators it controls. We assume the signals sent to the actuators are constant, but the analysis can be extended to scenarios where the attacker sends other forms of signals, or simply blocks the corresponding control of reaching the plant. The control signals and measurements during the attack can then be modeled as

$$\tilde{u}(k) = u(k) + D_u a_u \qquad \tilde{y}(k) = \tilde{\Lambda}_y y(k) + \Lambda_y y(0) \tag{11}$$

where $a_u \in \mathbb{R}^{n_{a_u}}$ is the malicious control signal sent to the actuators, and $\Lambda_y \in \mathbb{R}^{n_y \times n_y}$ and $\tilde{\Lambda}_y \in \mathbb{R}^{n_y \times n_y}$ are diagonal matrices defined as

$$\Lambda_y(i,i) = \begin{cases} 1 & i \in \mathcal{V}_y \\ 0 & i \notin \mathcal{V}_y \end{cases} \qquad \tilde{\Lambda}_y = \mathbf{I}_{n_y} - \Lambda_y.$$

From (1), (2), (4), (11), and

$$y(0) = [C_p \;\; \mathbf{0}_{n_y \times n_z} \;\; \mathbf{0}_{n_y \times n_s}] Q_{ss} y_r =: Q_{ss}^y y_r$$

it follows that the system under the replay attack propagates according to the equations

$$\begin{aligned} x_e(k+1) &= \tilde{A}_e x_e(k) + \tilde{B}_e a_u + \tilde{K}_e y_r \\ r(k) &= \tilde{C}_e x_e(k) + \tilde{D}_e y_r \end{aligned}$$

where

$$\begin{aligned} \tilde{A}_e &= \begin{bmatrix} A_p + B_p D_f \tilde{\Lambda}_y C_p & B_p C_f & \mathbf{0}_{n_x \times n_s} \\ B_f \tilde{\Lambda}_y C_p & A_f & \mathbf{0}_{n_z \times n_s} \\ (K_d + B_d D_f) \tilde{\Lambda}_y C_p & B_d C_f & A_d \end{bmatrix} \\ \tilde{B}_e &= \begin{bmatrix} B_p D_u \\ \mathbf{0}_{n_z \times n_{a_u}} \\ \mathbf{0}_{n_s \times n_{a_u}} \end{bmatrix} \tilde{K}_e = \begin{bmatrix} B_p(E_f + D_f \Lambda_y Q_{ss}^y) \\ K_f + B_f \Lambda_y Q_{ss}^y \\ B_d E_f + (B_d D_f + K_d) \Lambda_y Q_{ss}^y \end{bmatrix} \\ \tilde{C}_e &= \begin{bmatrix} (D_d D_f + E_d) \tilde{\Lambda}_y C_p & D_d C_f & C_d \end{bmatrix} \\ \tilde{D}_e &= D_d E_f + (D_d D_f + E_d) \Lambda_y Q_{ss}^y. \end{aligned}$$

The critical states after N steps and the residual signal after k steps are then given by

$$x_c(N) = T_x \begin{bmatrix} y_r \\ a_u \end{bmatrix} \quad r(k) = T_r(k) \begin{bmatrix} y_r \\ a_u \end{bmatrix} \tag{12}$$

where

$$\begin{aligned} T_x &= Q_c \left[\tilde{A}_e^N Q_{ss} + \sum_{i=0}^{N-1} \tilde{A}_e^i \tilde{K}_e \quad \sum_{i=0}^{N-1} \tilde{A}_e^i \tilde{B}_e \right] \\ T_r(k) &= \left[\tilde{C}_e (\tilde{A}_e^k Q_{ss} + \sum_{i=0}^{k-1} \tilde{A}_e^i \tilde{K}_e) + \tilde{D}_e \quad \tilde{C}_e \sum_{i=0}^{k-1} \tilde{A}_e^i \tilde{B}_e \right]. \end{aligned}$$

In what follows, we formulate the problem of finding the worst-case impact of replay attacks.

Problem 3:

$$\begin{aligned} \underset{y_r, a_u}{\text{maximize}} \quad & I(\mathcal{V}_y, \mathcal{V}_u) = \left|\left| T_x \begin{bmatrix} y_r \\ a_u \end{bmatrix} \right|\right|_\infty \\ \text{subject to} \quad & -\delta_{y_r} \le y_r \le \delta_{y_r} \\ & S(k+1) \le \delta_r \quad k = 0, \dots, N. \end{aligned}$$

This problem represents an instance of Problem 1.

Proposition 2: Problem 3 is an instance of Problem 1 for the stateless, CUSUM, and MEWMA detectors.

Proofs of Propositions 2–4 follow the same lines as the proof of Proposition 1, and are omitted due to the page limit.

5) False Data Injection Attacks: False data injection attacks represent very sophisticated attack strategy. The attack signal $a(0), \dots, a(N)$ is calculated based on the full model knowledge and then fed into the system through the corrupted sensors and actuators. Although very powerful, this attack is more unlikely than for example denial of service attack due to the need of full model knowledge.

This attack is additive in nature, thus the attack trajectory of the extended system (6) can be divided into the trajectory $x_e^0(k), r^0(k)$ driven by the initial value and the reference, and the trajectory $x_e^a(k), r^a(k)$ driven by the attack signal

$$x_e(k) = x_e^0(k) + x_e^a(k) \qquad r(k) = r^0(k) + r^a(k).$$

Under Assumption 1, the system has reached steady state before the attack starts. Since the attack does not change the system structure, it follows $x_e^0(k) = x_e(0)$, $r^0(k) = 0$. Based on the previous discussion, and using the extended system equations (6), the critical states after N steps and the residual signal after k steps can be expressed as

$$x_c(N) = T_x \begin{bmatrix} y_r \\ a_{0:N} \end{bmatrix} \qquad r(k) = T_r(k) a_{0:k}$$

where $a_{0:k} = [a(0)^T \ \dots \ a(k)^T]^T$, and

$$\begin{aligned} T_x &= Q_c[Q_{ss} \ A_e^{N-1}B_e \ \dots \ B_e \ \mathbf{0}_{n_e \times n_a}] \\ T_r(k) &= [C_e A_e^{k-1} B_e \ \dots \ C_e B_e \ D_e]. \end{aligned} \tag{13}$$

The worst case impact of the false data injection attacks can then be obtained by solving the following problem.

Problem 4:

$$\begin{aligned} \underset{y_r, a_{0:N}}{\text{maximize}} \quad & I(\mathcal{V}_y, \mathcal{V}_u) = \left\| T_x \begin{bmatrix} y_r \\ a_{0:N} \end{bmatrix} \right\|_\infty \\ \text{subject to} \quad & -\delta_{y_r} \le y_r \le \delta_{y_r} \\ & S(k+1) \le \delta_r \quad k = 0, \dots, N. \end{aligned}$$

This problem is also reducible to the form of Problem 1.

Proposition 3: Problem 4 is an instance of Problem 1 for the stateless, CUSUM, and MEWMA detectors.

6) Bias Injection Attack: Compared to the false data injection attack, the bias injection attack is less sophisticated since the attacker injects a constant bias in the corrupted signals instead of a time-varying signal [11]. The control inputs and measurements during the bias injection attack can be expressed as

$$\tilde{u}(k) = u(k) + D_u a_u \qquad \tilde{y}(k) = y(k) + D_y a_y.$$

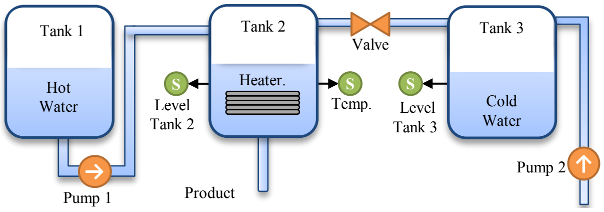

Fig. 1. Chemical process with four actuators (two pumps, heater, and valve), and three sensors (two level sensors and one temperature sensor).

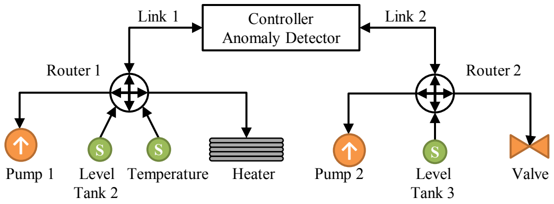

Fig. 2. Cyber infrastructure of the process.

Let $a = [a_u^T \ a_y^T]^T$. By inserting $a(0) = \dots = a(N) = a$ in (13), the critical states after N steps and the residual signal after k steps can be expressed as

$$\begin{aligned} x_c(N) &= Q_c \left(Q_{ss} y_r + \sum_{i=0}^{N-1} A_e^i B_e a \right) =: T_x \begin{bmatrix} y_r \\ a \end{bmatrix} \\ r(k) &= \left(C_e \sum_{i=0}^{k-1} A_e^i B_e + D_e \right) a =: T_r(k) a. \end{aligned}$$

The problem of finding the worst case bias injection attack can then be formulated as follows.

Problem 5:

$$\begin{aligned} \underset{y_r, a}{\text{maximize}} \quad & I(\mathcal{V}_y, \mathcal{V}_u) = \left\| T_x \begin{bmatrix} y_r \\ a \end{bmatrix} \right\|_\infty \\ \text{subject to} \quad & -\delta_{y_r} \le y_r \le \delta_{y_r} \\ & S(k+1) \le \delta_r \quad k = 0, \dots, N. \end{aligned}$$

Proposition 4: Problem 5 is an instance of Problem 1 for the stateless, CUSUM, and MEWMA detectors.

IV. Simulations

In this section, we illustrate how the attack models we proposed can be used to conduct risk assessment. We observe a part of a chemical process [23] shown in Figure 1. The control objective is to keep a constant liquid level and a constant temperature in Tank 2. This objective is achieved by injecting hot water from Tank 1, and cold water from Tank 3. The cyber infrastructure of the system is assumed to be as shown in Figure 2. The communication links that connect the routers with the controller are unprotected, and our task is to decide which one is more important to protect.

The states of the system are the volume in Tank 3 (x_1), the volume in Tank 2 (x_2), and the temperature in Tank 2 (x_3). All three states are measured. The control signals are the flow rate of Pump 2 (u_1), the openness of the valve (u_2), the flow rate of Pump 1 (u_3), and the power of the heater (u_4). We choose $\delta_{y_r} = 1$, $N = 20$ for the attack length, $Q_r = \mathbf{I}_3$, $Q_{ss} = [\mathbf{I}_3 \ \mathbf{I}_3 \ \mathbf{I}_3]^T$, and $Q_c = [\mathbf{0}_{2\times 1} \ \mathbf{I}_2 \ \mathbf{0}_{2\times 6}]$ such that the critical states correspond to x_2 and x_3. The MEWMA

TABLE I

IMPACT $I(\mathcal{V}_y, \mathcal{V}_u)$ OF DIFFERENT ATTACK STRATEGIES.

Attack strategy	Link 1	Link 2
Denial of Service	1.2107	1.3773
Rerouting	1.7271	1.0506
Sign Alternation	1.4488	1.6064
Replay	∞	1.7843
False Data Injection	∞	3.6314
Bias Injection	3.4919	1.6071

detector with parameters $\delta_r = 1$ and $\beta = 0.2$ is used to detect anomalies. For the attack on Link 1, the attacker can manipulate components $\mathcal{V}_u^1 = \{3, 4\}$ and $\mathcal{V}_y^1 = \{2, 3\}$, while for the attack on Link 2 we have $\mathcal{V}_u^2 = \{1, 2\}$ and $\mathcal{V}_y^2 = \{1\}$.

For the given configuration, we derive the impact $I(\mathcal{V}_y, \mathcal{V}_u)$ of the presented cyber-attacks. The results are shown in Table 1. We see from the table that the false data injection and replay attack can have devastating impacts on the system if the attacker has access to Link 1. This is according to expectation, since an attack on Link 1 can directly manipulate the measurements of the critical states x_2 and x_3, and these states are not measurable from sensors measurements transmitted over Link 2. In that case, $\ker(T_r) \nsubseteq \ker(T_x(l, :))$ for $l \in \{1, 2\}$, so the attacker can make an arbitrary large impact. This also shows that in certain cases simpler attack such as a replay attack, might be equally as dangerous as false data injection attacks with full model knowledge. We can also see that the attack impact on Link 1 is not always larger than the attack impact on Link 2. In particular, the impact of denial of service and sign alternation attacks on Link 2 is larger than the impact on Link 1. Nevertheless, given that the attack impact on Link 1 is higher for most of the attack strategies, the resources should be allocated to protect this link.

V. CONCLUSION AND FUTURE WORK

In this paper we proposed a framework that can be used for conducting risk assessment in industrial control systems. The framework can be used to estimate the impact of both simple and more complex attack strategies, and it is applicable for several types of anomaly detectors. Possible extensions of this work will be to include process and measurement noises into the framework, and to evaluate the attack impact under novel types of estimators and detectors, such as those proposed in [8], [24].

REFERENCES

[1] J. Slay and M. Miller, *Lessons Learned from the Maroochy Water Breach*. Boston, MA: Springer US, 2008, pp. 73–82.

[2] D. Kushner, "The real story of STUXNET," *IEEE Spectrum*, vol. 50, no. 3, pp. 48–53, March 2013.

[3] D. U. Case, "Analysis of the cyber attack on the Ukrainian power grid," 2016.

[4] S. Amin, X. Litrico, S. Sastry, and A. M. Bayen, "Cyber security of water SCADA systems - Part I: Analysis and experimentation of stealthy deception attacks," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 5, pp. 1963–1970, Sept 2013.

[5] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 1, pp. 13:1–13:33, Jun. 2011.

[6] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli, "False data injection attacks against state estimation in wireless sensor networks," in *49th IEEE Conference on Decision and Control (CDC)*, Dec 2010, pp. 5967–5972.

[7] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, Nov 2013.

[8] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1454–1467, June 2014.

[9] K. Stouffer, J. Falco, and K. Scarfone, "Guide to industrial control systems (ICS) security," *NIST special publication*, vol. 800, no. 82, 2011.

[10] A. A. Cárdenas, S. Amin, Z.-S. Lin, Y.-L. Huang, C.-Y. Huang, and S. Sastry, "Attacks against process control systems: Risk assessment, detection, and response," in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, ser. ASIACCS '11. New York, NY, USA: ACM, 2011, pp. 355–366.

[11] C. M. Ahmed, C. Murguia, and J. Ruths, "Model-based attack detection scheme for smart water distribution networks," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM, 2017, pp. 101–113.

[12] D. I. Urbina, J. A. Giraldo, A. A. Cárdenas, N. O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, and H. Sandberg, "Limiting the impact of stealthy attacks on industrial control systems," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: ACM, 2016, pp. 1092–1105.

[13] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems." in *HotSec*, 2008.

[14] S. Amin, A. A. Cárdenas, and S. Sastry, "Safe and secure networked control systems under denial-of-service attacks." in *HSCC*, vol. 5469. Springer, 2009, pp. 31–45.

[15] A. Teixeira, K. Paridari, H. Sandberg, and K. H. Johansson, "Voltage control for interconnected microgrids under adversarial actions," in *Emerging Technologies & Factory Automation (ETFA), 2015 IEEE 20th Conference on*. IEEE, 2015, pp. 1–8.

[16] R. M. Ferrari and A. M. Teixeira, "Detection and isolation of routing attacks through sensor watermarking," in *American Control Conference (ACC), 2017*. IEEE, 2017, pp. 5436–5442.

[17] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Optimal linear cyber-attack on remote state estimation," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 4–13, March 2017.

[18] C.-Z. Bai and V. Gupta, "On Kalman filtering in the presence of a compromised sensor: Fundamental performance bounds," in *2014 American Control Conference*. IEEE, 2014, pp. 3029–3034.

[19] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on scada systems," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, July 2014.

[20] D. Umsonst, H. Sandberg, and A. A. Cárdenas, "Security analysis of control system anomaly detectors," in *American Control Conference (ACC), 2017*. IEEE, 2017, pp. 5500–5506.

[21] C. A. Lowry, W. H. Woodall, C. W. Champ, and S. E. Rigdon, "A multivariate exponentially weighted moving average control chart," *Technometrics*, vol. 34, no. 1, pp. 46–53, 1992.

[22] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135–148, 2015.

[23] M. Blanke, M. Kinnaert, J. Lunze, M. Staroswiecki, and J. Schröder, *Diagnosis and fault-tolerant control*. Springer, 2006, vol. 691.

[24] M. Pajic, I. Lee, and G. J. Pappas, "Attack-resilient state estimation for noisy dynamical systems," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 82–92, March 2017.