This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCNS.2019.2940253, IEEE Transactions on Control of Network Systems

Estimating the Impact of Cyber-Attack Strategies for Stochastic Networked Control Systems

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

Abstract—**Risk assessment is an inevitable step in implementation of a cyber-defense strategy. An important part of this assessment is to reason about the impact of possible attacks. In this work, we study the problem of estimating the impact of cyber-attacks in stochastic linear networked control systems. For the stealthiness constraint, we adopt the Kullback-Leibler divergence between attacked and non-attacked residual sequences. Two impact metrics are considered: The probability that some of the critical states leave a safety region and the expected value of the infinity norm of the critical states. For the first metric, we prove that the optimal value of the impact estimation problem can be calculated by solving a set of convex problems. For the second, we derive efficient to calculate lower and upper bounds. Finally, we show compatibility of our framework with a number of attack strategies proposed in the literature, and demonstrate how it can be used for risk assessment on an example.**

I. Introduction

Networked control systems operate physical processes of great societal significance, such as electricity production, transportation, and water distribution. Unfortunately, it is known that numerous security vulnerabilities can be found within these systems [1], which if exploited, can lead to extremely dangerous attacks [2]–[4]. Hence, it is essential to prevent security vulnerabilities before an attacker exploits them.

However, preventing security vulnerabilities in a networked control system can be complicated and costly [1]. Thus, one should conduct a risk assessment to prioritize among the vulnerabilities. Prioritization is done based on the likelihood that vulnerabilities are exploited, and the impact that can happen if the exploitation occurs [5]. The resources can then be focused on preventing the most critical vulnerabilities.

Motivated by the risk assessment application, we study an impact estimation problem. By solving the impact estimation problem, we check if an attacker can inflict a large damage to the system while remaining stealthy. Hence, the objective function of the problem is an impact metric that is maximized, while the constraints include a stealthiness constraint. This problem is generally difficult to solve, since it usually reduces to a non-convex constrained maximization problem.

Related work: Significant effort have been dedicated towards estimating the impact of attacks that remain undetected by the *chi-square* anomaly detector [6]–[10]. In these studies, reachable sets were predominantly used to characterize the impact, and algorithms for calculating upper and lower bounds of these sets were proposed in [6]–[8]. The focus of these studies was on false data injection (FDI) [6]–[9] and bias injection [10] attack strategies.

The impact estimation problem for other types of detectors have also been considered [11]–[15]. The focus of [11]–[15] was also on powerful injection attacks. In this set of literature, the work especially relevant for our study is [15]. There, the authors used the infinity norm of critical states to quantify the impact under the cumulative sum detector, and showed that the exact value of the impact can be obtained by solving a set of convex problems. This useful property of the infinity norm based metric was also recognized in [16], where the impact was obtained by solving a set of linear programs. However, the works [15], [16] neglect the influence of noise, and do not propose a substitute for the infinity norm based metric that can be used in stochastic systems.

Our work differs from the existing literature in the following aspects. Different from the works on the infinity norm based metric [15], [16], we focus on more general stochastic systems. Particularly, we propose two metrics that can substitute the infinity norm based metric, and study the impact estimation problem based on these metrics. Compared to the studies on the impact estimation problem [6]–[16] that focus on powerful injection attacks, our analysis is more general. Particularly, our analysis covers both FDI and bias injection attack strategies, as well as Denial of Service (DoS) [17], [18], replay [19], rerouting [20], sign alternation [21], and combined DoS and FDI [22], [23] attack strategies. Additionally, the studies [6]–[15] focus their analysis on particular types of anomaly detectors, so the impact analysis is carried out for every detector separately. In our work, we use the idea from [24]–[27], and model the stealthiness constraints based on the Kullback-Leibler (KL) divergence. In this way, we make our analysis independent of the choice of anomaly detector.

Contributions: Firstly, we propose and study a novel type of impact estimation problem. We consider two impact metrics: (i) The probability that some of the critical states leave a safety region (I_P); (ii) The expected value of the infinity norm of the critical states (I_E). For the stealthiness constraint, we adopt the KL–divergence between attacked and non-attacked residual sequences. Furthermore, we introduce additional constraints on attack signals. Through these constraints, we impose different types of attack strategies.

Secondly, we introduce an auxiliary problem $\mathcal{P}$ which we use to analyze the impact estimation problem, and establish its convexity (Propositions 1). Using $\mathcal{P}$, we characterize conditions under which the impact estimation problem is infeasible or its optimal value equals to the maximum impact (Proposition 2). If these conditions are not satisfied, we prove that the metric I_P has the same desirable properties as the infinity norm based metric from [15]. That is, the exact value

*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 Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. Emails:{jezdimir, hsan, kallej}@kth.se.

of the impact in terms of I_P can be obtained by solving a set of convex problems $\mathcal{P}$ (Theorem 1). Unfortunately, the metric I_E does not have a closed form expression and is not trivial to evaluate. However, we derive efficient to calculate lower and upper bounds for the metric I_E using $\mathcal{P}$ (Theorem 2). We then discuss the tightness of these bounds, and explain how the bounds can be used if the tightness cannot be established.

Thirdly, we show that our framework allows us to analyze the impact of FDI, bias injection, DoS, replay, rerouting, sign alternation, and combined DoS and FDI attack strategies (Propositions 3–5). Finally, using a numerical example of a chemical process, we illustrate how our framework can be used for risk assessment and discuss how the tuning parameters influence the impact of different attack strategies.

The preliminary version of this work appeared in [28]. In [28], our focus was on deterministic systems and three classes of anomaly detectors. In the present work, we consider a more general stochastic system model, different impact metrics, stealthiness constraints, and an attack model.

Organization: Section II introduces the model setup and Section III the impact estimation problem. Section IV presents the main technical results. Section V introduces attacks compatible with our framework. Section VI illustrates the applicability of our framework on an example of a networked control system. Section VII concludes the paper. Appendix contains proofs of some technical lemmas and propositions.

Notation: We denote by: $0_{m\times n}$ the zero-matrix with m rows and n columns; I_n the identity matrix of size n; 1_n the vector of size n with all the elements equal to 1; $T(i,:)$ the i-th row of matrix T; $T(i,j)$ the element of matrix T positioned in the i-th row and j-th column; $x^{(i)}$ the i-th element of vector x; $\otimes$ the Kronecker product; $\mathcal{N}(\mu,\Sigma)$ the Gaussian distribution with the mean value μ and the covariance matrix Σ. If x is a discrete-time signal, then $x_{N:M} = [x(N)^T \ \dots \ x(M)^T]^T$. Let $A{\in}\mathbb{R}^{n\times n}$, $B{\in}\mathbb{R}^{n\times m}$, $C{\in}\mathbb{R}^{p\times n}$, and $D{\in}\mathbb{R}^{p\times m}$. Then

$$\mathcal{O}_N(A,C) = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^N \end{bmatrix}, \ \mathcal{C}_N(A,B) = \begin{bmatrix} A^{N-1}B & \dots & B \end{bmatrix},$$

$$\mathcal{T}_N(A,B,C,D) = \begin{bmatrix} D & 0_{p\times m} & \dots & 0_{p\times m} \\ CB & D & \dots & 0_{p\times m} \\ \vdots & \vdots & \ddots & \vdots \\ CA^{N-1}B & CA^{N-2}B & \dots & D \end{bmatrix}.$$

II. Model Setup

The system consists of the physical plant, the estimator, the controller, and the residual filter. The plant is modeled by

$$\begin{aligned} x(k+1) &= Ax(k) + B\tilde{u}(k) + v_x(k), \\ y(k) &= Cx(k) + v_y(k), \end{aligned} \tag{1}$$

where $x(k){\in}\mathbb{R}^{n_x}$ is the plant state, $y(k){\in}\mathbb{R}^{n_y}$ are the measurements, $\tilde{u}(k){\in}\mathbb{R}^{n_u}$ are the control actions applied to the plant, and $v_x(k){\in}\mathbb{R}^{n_x}$ (resp. $v_y(k){\in}\mathbb{R}^{n_y}$) is the process (resp. measurement) noise. The noises v_x and v_y are independent Gaussian white processes with zero mean and positive definite covariance matrices Σ_{v_x} and Σ_{v_y}, respectively. The pair (C,A) (resp. (B,A)) is observable (resp. controllable).

The estimator is a steady state Kalman filter defined by

$$\hat{x}(k+1) = (A-KC)\hat{x}(k) + Bu(k) + K\tilde{y}(k), \tag{2}$$

where $\hat{x}(k){\in}\mathbb{R}^{n_x}$ is the one step ahead prediction of $x(k)$, $u(k){\in}\mathbb{R}^{n_u}$ are the control actions calculated by the controller, and $\tilde{y}(k){\in}\mathbb{R}^{n_y}$ are the measurements received by the estimator. The steady state Kalman gain is given by $K{=}A\Sigma_e C^T(C\Sigma_e C^T + \Sigma_{v_y})^{-1}$, where Σ_e is the error covariance matrix obtained by solving the Riccati equation $\Sigma_e{=}A\Sigma_e A^T{+}\Sigma_{v_x}{-}A\Sigma_e C^T(C\Sigma_e C^T{+}\Sigma_{v_y})^{-1}C\Sigma_e A^T$. The gain K exists under the introduced assumptions, and it is known that $A-KC$ is asymptotically stable [29].

The controller is of the form

$$u(k) = L\hat{x}(k) + L_{y_r}y_r, \tag{3}$$

where $y_r{\in}\mathbb{R}^{n_{y_r}}$ is a constant reference. We assume that the controller ensures asymptotic stability and satisfactory performances in absence of attacks, and that the system has reached a stationary regime before an attack starts.

The residual signal is defined by

$$\tilde{r}(k) = \Sigma_r^{-\frac{1}{2}}\big(\tilde{y}(k) - C\hat{x}(k)\big), \tag{4}$$

where $\Sigma_r{=}C\Sigma_e C^T{+}\Sigma_w$. This signal is used to measure attack stealthiness. In absence of attacks, the residual sequence is a white Gaussian process with zero mean value and identity covariance matrix. We denote by r the non-attacked residual.

We assume that an attack starts at $k=0$. The attacked measurements $\tilde{y}$ and control actions $\tilde{u}$ are modeled by

$$\begin{aligned} \tilde{y}(k) &= \Lambda_y y(k) + \Gamma_y a_y(k) + \Gamma_y a_s(k), \\ \tilde{u}(k) &= \Lambda_u u(k) + \Gamma_u a_u(k). \end{aligned} \tag{5}$$

For example, the signals can be corrupted when they are communicated over a network. Here, $a_u(k){\in}\mathbb{R}^{n_u}$ (resp. $a_y(k){\in}\mathbb{R}^{n_y}$) is the deterministic part of the attack against the actuators (resp. sensors). The signal $a_s(k){\in}\mathbb{R}^{n_y}$ is stochastic in nature, and will be required to model the replay attack strategy (see Section V). Finally, the matrices Γ_y, Γ_u, Λ_y, and Λ_u depend on the attack strategy and the attacker's resources.

By combining the equations (1)–(5), the system dynamics under attack can be written as

$$\begin{aligned} x_e(k+1) &= \tilde{A}x_e(k){+}\tilde{B}v(k){+}\tilde{E}y_r{+}\tilde{G}a(k){+}\tilde{J}a_s(k), \\ \tilde{r}(k) &= \tilde{C}x_e(k){+}\tilde{D}v(k){+}\tilde{F}y_r{+}\tilde{H}a(k){+}\tilde{K}a_s(k), \end{aligned} \tag{C1}$$

where $x_e(k){=}[x(k)^T\,\hat{x}(k)^T]^T$, $v(k){=}[v_x(k)^T\,v_y(k)^T]^T$, and $a(k){=}[a_u(k)^T\,a_y(k)^T]^T$ are the extended state, noise, and attack vectors, respectively. We denote the dimension of $a(k)$ by n_a, and of $v(k)$ by n_v. The matrices $\tilde{A}$–$\tilde{K}$ are given by

$$\tilde{A} = \begin{bmatrix} A & -B\Lambda_u L \\ K\Lambda_y C & A-KC-BL \end{bmatrix}, \tilde{B} = \begin{bmatrix} I_{n_x} & 0_{n_x\times n_y} \\ 0_{n_x\times n_x} & K\Lambda_y \end{bmatrix},$$

$$\tilde{E} = \begin{bmatrix} B\Lambda_u L_{y_r} \\ BL_{y_r} \end{bmatrix}, \tilde{G} = \begin{bmatrix} B\Gamma_u & 0_{n_x\times n_y} \\ 0_{n_x\times n_u} & K\Gamma_y \end{bmatrix}, \tilde{J} = \begin{bmatrix} 0_{n_x\times n_y} \\ K\Gamma_y \end{bmatrix},$$

$\tilde{C}{=}\Sigma_r^{-\frac{1}{2}}[\Lambda_y C \ -C]$, $\tilde{D}{=}[0_{n_y\times n_x} \ \Sigma_r^{-\frac{1}{2}}\Lambda_y]$, $\tilde{F} = 0_{n_y\times n_{y_r}}$, $\tilde{H} = [0_{n_y\times n_u} \ \Sigma_r^{-\frac{1}{2}}\Gamma_y]$, and $\tilde{K}{=}\Sigma_r^{-\frac{1}{2}}\Gamma_y$.

III. Problem Formulation

This section defines two impact estimation problems $\mathcal{P}_1$ and $\mathcal{P}_2$ the rest of the paper is concerned about. We first introduce the decision variables, the impact metrics, and the constraints.

The decision variables are $d{=}[a_{0:N}^T\ y_r^T]^T$, where $N{\in}\mathbb{Z}^+$ is the length of the horizon over which we estimate the impact. Although the system trajectory is influenced by other signals as well, we show that the impact metrics and the constraints are only affected by the reference y_r and the attack sequence $a_{0:N}$. Since we perform off-line analysis, the exact value of y_r at the beginning of the attack is unknown to us. The same holds for $a_{0:N}$, since it depends on the attacker's choice. Hence, by optimizing over d, we identify the worst case impact.

The impact metrics are based on the concept of critical states. These states may model the flow of energy through the power line that should be maintained within predefined bounds, or a temperature that should not exceed some safety limit. We define the critical states as

$$z(k) = Q_z x(k), \tag{6}$$

where $Q_z{\in}\mathbb{R}^{n_z\times n_x}$ is a full row rank scaling matrix, and $n_z{\le}n_x$ is the number of the critical states. The matrix Q_z is chosen such that having magnitude of any of the critical states larger than one indicates a dangerous system state.

Example 1: Let $x{=}[x^{(1)}\ x^{(2)}]^T$ be the plant state. Assume $x^{(2)}$ is the critical state that should be kept within the interval $[-\bar{x},\bar{x}]$, where $\bar{x}{\ge}0$. The matrix Q_z is then defined by $Q_z{=}[0\ \ 1/\bar{x}]$. Therefore, if $|x^{(2)}(k)|\ge\bar{x}$, then $|z(k)|\ge 1$.

In the related work on deterministic systems [15], the impact metric was defined as $||z_{1:N}||_\infty$. If $||z_{1:N}||_\infty{\ge}1$, then the attacker can drive some of the critical states outside the safety region during N time steps. Yet, in our case, the state is influenced by the noise in addition to attacks. Hence, some of the critical states can leave the safety region with non-zero probability even in absence of attacks. To make the impact metric suitable for stochastic systems, we define a new metric

$$I_P(d) = \max\nolimits_{i\in\mathcal{I}}\ \mathbb{P}(|z_{1:N}^{(i)}|\ge 1),$$

where $\mathcal{I}=\{1,\dots,n_zN\}$. If I_P is close to one (resp. close to zero), the critical states leave (resp. stay within) the safety region with high probability, and the attack is dangerous (resp. harmless). Another possible impact metric is the expected value of the infinity norm of $z_{1:N}$, that is,

$$I_E(d) = \mathbb{E}\{||z_{1:N}||_\infty\}.$$

Unfortunately, I_E does not have a closed form expression, and is hard to evaluate in general.

The problem constrains are denoted by (C1)–(C5). Constraint (C1) was introduced in the previous section, and imposes that x_e and $\tilde{r}$ have to satisfy the system dynamics. Constraint (C2) is the reference constraint defined by

$$||Q_{y_r}y_r||_\infty\le 1, \tag{C2}$$

where $Q_{y_r}{\in}\mathbb{R}^{n_{y_r}\times n_{y_r}}$ is a full rank scaling matrix. Constraint (C3) is the stealthiness constraint

$$\tfrac{1}{N+1}\mathcal{D}(\tilde{r}_{0:N}||r_{0:N})\le\epsilon, \tag{C3}$$

where $\mathcal{D}(\tilde{r}_{0:N}||r_{0:N})$ is the KL–divergence between the distributions of attacked $\tilde{r}_{0:N}$ and non-attacked $r_{0:N}$ residual sequences, and $\epsilon\ge 0$ is a stealthiness level. The KL–divergence gives a distance between two distributions p and q over a sample space X, and is defined by $\mathcal{D}(p||q){=}\int_X \log\big(\frac{p(x)}{q(x)}\big)p(x)dx$. It is known that $\mathcal{D}(p||q){\ge}0$ with equality if and only if p equals q almost everywhere. Hence, if $\mathcal{D}(\tilde{r}_{0:N}||r_{0:N})$ is small, then the distributions of $\tilde{r}_{0:N}$ and $r_{0:N}$ are similar, and the attack stays stealthy. The constraints (C4) and (C5) are given by

$$F_a a_{0:N} = 0_{n_{F_a}\times 1}, \tag{C4}$$

$$a_{s0:N} = T_1 x_e(N_s) + T_2 v_{N_s:-1} + T_3 y_r, \tag{C5}$$

where $N_s{<}0$, the matrices T_1, T_2, T_3, and F_a have appropriate dimensions, and n_{F_a} is the number of rows of F_a. These constraints are used to impose a particular attack strategy.

We are now ready to introduce the impact estimation problem based on the metric I_P.

$$\mathcal{P}_1: \qquad \underset{d}{\text{maximize}}\ \ I_P(d)\quad \text{subject to (C1)–(C5).}$$

Although our main focus is on $\mathcal{P}_1$, we also investigate the impact estimation problem based on the metric I_E.

$$\mathcal{P}_2: \qquad \underset{d}{\text{maximize}}\ \ I_E(d)\quad \text{subject to (C1)–(C5).}$$

Both $\mathcal{P}_1$ and $\mathcal{P}_2$ are non-convex constrained maximization problems, and efficient algorithms for solving these type of problems are unknown in general. Nevertheless, we propose an efficient way to calculate the optimal value of $\mathcal{P}_1$. Additionally, we derive lower and upper bounds for $\mathcal{P}_2$. Prior to that, we outline some properties of these problems.

Remark 1: The tuning parameters in $\mathcal{P}_1$ and $\mathcal{P}_2$ are N and ϵ. Naturally, we first want to discover stealthy attacks that result in a high impact in a short amount of time. Thus, choosing small values of N and ϵ is a good starting point for the analysis. One can then start gradually increasing N and ϵ to discover less dangerous attacks.

Remark 2: One can also consider maximizing impact in N_z steps and imposing stealthiness in $N_r{\neq}N_z$ steps. The case $N_z{<}N_r$ captures attacks that maximize damage in N_z steps, and prevent the operator noticing this in additional $N_r{-}N_z$ steps. The case $N_z{>}N_r$ models ambush attacks [30], where the attacker stealthily prepares N_r steps, and then launches a not necessarily stealthy attack in the remaining time. Although we focus on the case $N_r{=}N_z{=}N$, the analysis that follows can be extended to cover the aforementioned cases as well.

Remark 3: Some of the advantages of using the stealthiness constraint (C3) are as follows: (i) As shown later, (C3) is a convex constraint in d for the class of attacks we observe; (ii) The impact analysis is made independent of the choice of anomaly detector; (iii) Generating attack signals that satisfy (C3) can be a reasonable choice by the attacker that does not know which anomaly detector is deployed; (iv) In some cases, other types of stealthiness constraints can be replaced by a KL–divergence based constraint [6].

Remark 4: As shown later in Proposition 2, $\mathcal{P}_1$ and $\mathcal{P}_2$ can be infeasible due to (C3). If that is the case, we define the impact to be 0.

IV. Main Results

In this section, we prove that the optimal value of $\mathcal{P}_1$ can be calculated by solving a set of convex problems. Additionally, we show that in the process of solving $\mathcal{P}_1$, we also obtain lower and upper bounds of $\mathcal{P}_2$. Prior to this, we introduce some auxiliary lemmas, present the problem $\mathcal{P}$ crucial for solving $\mathcal{P}_1$ and bounding $\mathcal{P}_2$, and highlight some special cases of the problems. The omitted proofs are provided in Appendix.

A. Preliminaries

We first introduce Lemma 1, which establishes that the vector of critical states $z_{1:N}$ and the vector of residuals $\tilde{r}_{0:N}$ are Gaussian random vectors with fixed covariance matrices and mean values linear in d.

Lemma 1: The critical states vector $z_{1:N}$ is distributed according to $\mathcal{N}(\mu_Z, \Sigma_Z)$ and the residual vector $\tilde{r}_{0:N}$ according to $\mathcal{N}(\mu_R, \Sigma_R)$, where $\mu_Z = T_Z d$ and $\mu_R = T_R d$. The matrices T_Z, T_R, Σ_Z, Σ_R are independent of d and given by the equations (28), (29), (32), and (33), respectively. The covariance matrix Σ_Z is a positive definite matrix.

While Σ_Z is provably positive definite, the same claim does not hold for Σ_R in general. Namely, due to attacks, Σ_R may be positive semi definite. In what follows, we assume Σ_R is positive definite, and we later justify this assumption.

Assumption 1: Σ_R is a positive definite matrix.

We now use Lemma 1 to show that the stealthiness constraint (C3) is a convex and symmetric[1] constraint in d.

Lemma 2: Under Assumption 1, (C3) can be written as $||T_R d||_2^2{\le}\epsilon'$, where $\epsilon'{=}(N{+}1)(2\epsilon{+}n_y){-}\text{tr}(\Sigma_R){+}\ln\det(\Sigma_R)$.

Remark 5: If $\epsilon'{<}0$, (C3) is impossible to satisfy, and $\mathcal{P}_1$ and $\mathcal{P}_2$ are infeasible. Particularly, ϵ' approaches $-\infty$ when an eigenvalue of Σ_R approaches 0, which justifies excluding the cases where Σ_R is positive semidefinite from the analysis.

Remark 6: Other types of stealthiness constraints based on the KL–divergence are also reducible to convex and symmetric constraints. For example, the claim can be proven for: (i) The window type constraints $\frac{1}{N_w+1}\mathcal{D}(\tilde{r}_{i:i+N_w}||r_{i:i+N_w}){\le}\epsilon$, where $i{=}0,\ldots,N{-}N_w$, $N_w{\in}\mathbb{Z}^+$, and $N_w{\le}N$; (ii) The constraints from [24] $\mathcal{D}(\tilde{r}_i||r_i) \le \epsilon$, where $i = 0, \ldots N$.

We now introduce an optimization problem $\mathcal{P}$ crucial for solving $\mathcal{P}_1$ and deriving bounds for $\mathcal{P}_2$.

$$\mathcal{P}: \qquad \underset{d}{\text{maximize}} \;\; \mathbb{E}\{z^{(i)}_{1:N}\} \quad \text{subject to (C1)–(C5)},$$

where i belongs to $\mathcal{I}$. In what follows, we use Lemma 1 and Lemma 2 to show that $\mathcal{P}$ is reducible to a convex problem with symmetric constraints. Thus, $\mathcal{P}$ can be solved efficiently using well known algorithms.

Proposition 1: Under Assumption 1, $\mathcal{P}$ is reducible to the following convex optimization problem

$$\begin{aligned} &\underset{d}{\text{maximize}} && T_Z(i,:)d \\ &\text{subject to} && ||Qd||_\infty \le 1,\; ||T_R d||_2^2 \le \epsilon',\; Fd = 0_{n_{F_a}\times 1}, \end{aligned} \tag{7}$$

where $Q{=}[0_{n_{y_r}\times(N+1)n_a}\; Q_{y_r}]$ and $F = [F_a\; 0_{n_{F_a}\times n_{y_r}}]$.

Proof: The constraints (C1) and (C5) impose that $z_{1:N}$ (resp. $\tilde{r}_{0:N}$) is distributed according to $\mathcal{N}(T_Z d, \Sigma_Z)$ (resp. $\mathcal{N}(T_R d, \Sigma_R)$) (Lemma 1). Hence, the objective function of $\mathcal{P}$ can be written as $\mathbb{E}\{z^{(i)}_{1:N}\}{=}T_Z(i,:)d$, which is the objective function of (7). Constraint (C2) can be rewritten as

$$||Q_{y_r} y_r||_\infty = ||[0_{n_{y_r}\times(N+1)n_a}\; Q_{y_r}][a^T_{0:N}\; y^T_r]^T||_\infty \le 1.$$

Hence, (C2) reduces to $||Qd||_\infty{\le}1$, which is the first constraint in (7). From Lemma 2, Constraint (C3) can be exchanged with the second constraint in (7). Finally, Constraint (C4) can be rewritten as $F_a a_{0:N}{=}[F_a\; 0_{n_{F_a}\times n_{y_r}}][a^T_{0:N}\; y^T_r]^T{=}Fd$, which is the third constraint in (7). ■

Next, we investigate when $\mathcal{P}$ is infeasible (there are no points that satisfy the constraints) and unbounded (the optimal value is infinite), and then explain the importance of this result.

Proposition 2: The following statements hold: (I) $\mathcal{P}$ is infeasible for any i from $\mathcal{I}$ if and only if $\epsilon'{<}0$; (II) $\mathcal{P}$ is unbounded for at least one i from $\mathcal{I}$ if and only if $\epsilon'{\ge}0$ and $\text{null}([Q^T\; T_R^T\; F^T]^T){\not\subseteq}\text{null}(\,T_Z)$.

Proof: Statement I. ($\Rightarrow$) If $\mathcal{P}$ is infeasible, $\epsilon' \ge 0$ cannot hold, since $d = 0$ would be a feasible point for any i from $\mathcal{I}$.

($\Leftarrow$) If $\epsilon'{<}0$, then the constraint $||T_R d||_2^2{\le}\epsilon'$ cannot be satisfied for any d and any i from $\mathcal{I}$, so $\mathcal{P}$ is infeasible.

Statement II. ($\Rightarrow$) The proof is by contradiction. If $\epsilon'{<}0$, then $\mathcal{P}$ is infeasible, so it cannot be unbounded. If $\epsilon'{\ge}0$ and $\text{null}([Q^T\; T_R^T\; F^T]^T){\subseteq}\text{null}(\,T_Z)$, then for every d for which $T_Z d{\neq}0$ we have $[Q^T\; T_R^T\; F^T]^T d{\neq}0$. Hence, $T_Z(i,:)d$ cannot be made arbitrary large for any i from $\mathcal{I}$, because at least one of the constraints would be violated.

($\Leftarrow$) If $\text{null}([Q^T\, T_R^T\; F^T]^T){\not\subseteq}\text{null}(\,T_Z)$ and $\epsilon'{\ge}0$, then there exists d that satisfies $T_Z d{\neq}0$ and $[Q^T\; T_R^T\; F^T]^T d{=}0$. By increasing the magnitude of d while keeping it's direction same, we can make $T_Z(i,:)d$ arbitrary large for at least one i, while the constraints remain satisfied. ■

Remark 7: Note that $\mathcal{P}$ is infeasible if and only if $\mathcal{P}_1$ and $\mathcal{P}_2$ are infeasible, since these problems have the same constraints. Hence, the only situation when $\mathcal{P}_1$ and $\mathcal{P}_2$ are infeasible is when (C3) cannot be satisfied, that is, when the attacker cannot achieve the predefined stealthiness level.

Remark 8: If $\mathcal{P}$ is unbounded, the system is seriously vulnerable. Namely, if the easy to check conditions $\epsilon'{\ge}0$ and $\text{null}([Q^T\; T_R^T\; F^T]^T){\not\subseteq}\text{null}(\,T_Z)$ are satisfied, the attacker can make the deterministic part of a critical state arbitrary large while remaining stealthy. In that case, the influence of the stochastic component becomes negligible, so the optimal value of $\mathcal{P}_1$ (resp. $\mathcal{P}_2$) goes to 1 (resp. $+\infty$).

In the remainder, we focus on the case when $\mathcal{P}$ is feasible and bounded ($\text{null}([Q^T\; T_R^T\; F^T]^T){\subseteq}\text{null}(\,T_Z)$ and $\epsilon' \ge 0$). Lemma 3 (resp. Lemma 4) is later used to establish a link between $\mathcal{P}_1$ (resp. $\mathcal{P}_2$) with the convex problem $\mathcal{P}$ in this case.

Lemma 3: Let $\mathcal{C}_d$ be a symmetric constraint and consider the following optimization problems

$$\underset{d}{\text{maximize}} \;\; \mathbb{E}\{z^{(i)}_{1:N}\} \qquad \text{subject to} \;\; \mathcal{C}_d, \tag{8}$$

$$\underset{d}{\text{maximize}} \;\; \mathbb{P}(|z^{(i)}_{1:N}| > 1) \qquad \text{subject to} \;\; \mathcal{C}_d. \tag{9}$$

If the optimal value of (8) is bounded and if d^* is a solution of (8), then d^* is also a solution of (9).

[1] We say that a constraint $\mathcal{C}$ is symmetric, if for every x that satisfies $\mathcal{C}$, then $-x$ also satisfies $\mathcal{C}$.

Algorithm 1 Calculating the optimal value of $\mathcal{P}_1$

```
1: Input: $T_R$, $T_Z$, $\Sigma_Z$, $\Sigma_R$, $Q$, $F$, $\epsilon$
2: Output: $\hat{I}_P^*$
3: for every $i$ from $\mathcal{I}$ do
4:     For a given $i$, find a solution $\hat{d}_i^*$ of $\mathcal{P}$
5:     Calculate $\hat{P}_i^*{=}\mathbb{P}(|z_{1:N}^{(i)}| > 1)$ assuming $d{=}\hat{d}_i^*$
6: end for
7: $\hat{I}_P^* = \max_{i\in\mathcal{I}} \hat{P}_i^*$
```

Lemma 4: (Jensen's inequality [31]) Let ϕ be a convex function defined on a convex subset $\mathcal{C}_\phi$ of $\mathbb{R}^n$, and let X be an n-dimensional integrable random vector that satisfies $\mathbb{P}(X{\in}\mathcal{C}_\phi){=}1$. Then $\phi(\mathbb{E}\{X\}){\leq}\mathbb{E}\{\phi(X)\}$.

B. Solving $\mathcal{P}_1$

We now introduce Algorithm 1 and prove that it solves $\mathcal{P}_1$. For each i from $\mathcal{I}$, Algorithm 1 calculates a solution $\hat{d}_i^*$ of $\mathcal{P}$, and then $\hat{P}_i^*{=}\mathbb{P}(|z_{1:N}^{(i)}|{>}1)$ based on $\hat{d}_i^*$. Since $z_{1:N}$ is a Gaussian random vector (Lemma 1), $z_{1:N}^{(i)}$ is a Gaussian random variable. Hence, the probability $\hat{P}_i^*$ can be calculated with sufficiently large accuracy and in the computational time that is negligible compared to the computational time of solving $\mathcal{P}$. Finally, Algorithm 1 returns the largest $\hat{P}_i^*$ as the attack impact $\hat{I}_P^*$. The following theorem establishes that $\hat{I}_P^*$ is the optimal value of $\mathcal{P}_1$.

Theorem 1: Assume that null($[Q^T\ T_R^T\ F^T]^T$)$\subseteq$null(T_Z) and $\epsilon'{\geq}0$. Let I_P^* be the optimal value of $\mathcal{P}_1$ and let $\hat{I}_P^*$ be the value returned by Algorithm 1. Then $I_P^* = \hat{I}_P^*$.

Proof: Since the constraints of $\mathcal{P}_1$ are independent of i, $\mathcal{P}_1$ can be solved in the two steps. In the first step, one calculates

$$P_i^* = \underset{d}{\text{maximize}}\ \mathbb{P}(|z_{1:N}^{(i)}| > 1) \quad \text{subject to (C1)–(C5)}, \quad (10)$$

for every i from $\mathcal{I}$. In the second, one calculates the optimal value of $\mathcal{P}_1$ as $I_P^* = \max_{i\in\mathcal{I}} P_i^*$.

Algorithm 1 performs these two steps. Firstly, Algorithm 1 computes a solution $\hat{d}_i^*$ of $\mathcal{P}$, and based on it, calculates $\hat{P}_i^*{=}\mathbb{P}(|z_{1:N}^{(i)}|{>}1)$ (Lines 3–6). Under the assumption $\epsilon'{\geq}0$ and null($[Q^T\ T_R^T\ F^T]^T$)$\subseteq$null(T_Z), we know that a solution $\hat{d}_i^*$ of $\mathcal{P}$ exists, and that $T_Z(i,:)\hat{d}_i^*$ is bounded for every i (Proposition 2). From Proposition 1, the constraints of $\mathcal{P}$ are convex and symmetric constraints in d. Since $\mathcal{P}$ and (10) have the same constraints, it follows from Lemma 3 that $\hat{d}_i^*$ is also a solution of (10). This implies that $\hat{P}_i^*{=}P_i^*$ for each i from $\mathcal{I}$, so Algorithm 1 performs the first step of the procedure. The algorithm then performs the second step of the procedure (Line 7). Hence, it follows that $I_P^*{=}\hat{I}_P^*$. ■

Remark 9: Theorem 1 represents an interesting extension of the works [15], [16] that used the infinity norm metric $||z_{1:N}||_\infty$. Particularly, Theorem 1 shows that the optimal value of $\mathcal{P}_1$ can be obtained by solving a set of convex problems. This is the same favorable property that the impact estimation problem had in the deterministic systems case.

Remark 10: Algorithm 1 needs to solve the convex problem $\mathcal{P}$ n_zN times, which may appear to be time consuming. However, since we are performing off-line analysis, the execution time is not of critical importance. Moreover, we can considerably reduce the execution time by solving the problem $\mathcal{P}$ in parallel for every i from $\mathcal{I}$.

Remark 11: Constraint (C5) on the signal a_s simplifies the derivation of Theorem 1. Thanks to (C5), $z_{1:N}$ and $\tilde{r}_{0:N}$ are Gaussian random vectors, the decision variables can be represented by the vector d, (C3) is a convex and symmetric constraint in d, and the connection between $\mathcal{P}_1$ and $\mathcal{P}$ can be established. These convenient properties do not hold if a_s is a random process with non-Gaussian distribution. For example, (C3) may become hard to evaluate, since it has closed form expression only in some special cases. Additionally, the connection between $\mathcal{P}_1$ and $\mathcal{P}$ would be lost in general.

C. Lower and Upper Bounds for $\mathcal{P}_2$

We now use $\mathcal{P}$ to derive lower and upper bounds for $\mathcal{P}_2$. Particularly, let $\hat{I}_E^*$ be defined as

$$\hat{I}_E^* = \max_{i\in\mathcal{I}} \mu_i^*, \quad (11)$$

where μ_i^* is the optimal value of $\mathcal{P}$ corresponding to i. Theorem 2 provides lower and upper bounds based on $\hat{I}_E^*$.

Theorem 2: Assume that null($[Q^T\ T_R^T\ F^T]^T$)$\subseteq$null(T_Z) and $\epsilon' \geq 0$. Let I_E^* be the optimal value of $\mathcal{P}_2$, $\hat{I}_E^*$ be defined as in (11), and $\sigma_Z^* = \max_{i\in\mathcal{I}} \sqrt{\Sigma_Z(i,i)}$. Then

$$\hat{I}_E^* \leq I_E^* \leq \hat{I}_E^* + Nn_z\sigma_Z^*. \quad (12)$$

Proof: Since $\mathbb{E}\{z_{1:N}\}{=}T_Zd$ (Lemma 1), then $\hat{I}_E^*$ is the optimal value of the following optimization problem:

$$\underset{i\in\mathcal{I}}{\text{maximize}}\ \underset{d}{\text{maximize}}\ T_Z(i,:)d \quad \text{subject to (C1)–(C5)}. \quad (13)$$

Let I_E' be the optimal value of the problem

$$\underset{d}{\text{maximize}}\ ||T_Zd||_\infty \quad \text{subject to (C1)–(C5)}. \quad (14)$$

Note that both (13) and (14) are feasible, since $\epsilon'{\geq}0$. We first show $I_E' = \hat{I}_E^*$. The proof is similar to the proof of [28, Lemma 1], but we include it here for the reader's convenience and for the sake of completeness.

Let d' be an optimal solution for which I_E' is obtained, and notice that $I_E' = ||T_Zd'||_\infty = |T_Z(i',:)d'|$, for some i' from $\mathcal{I}$. Thus, $|T_Z(i',:)d'| \geq T_Z(i,:)d$ for every i from $\mathcal{I}$, and every d that satisfies (C1)–(C5). We then have $\hat{I}_E^*{\leq}I_E'$, since (13) and (14) have the same constraints. We now show that $\hat{I}_E^*{<}I_E'$ cannot hold. Since (C1)–(C5) are symmetric (Proposition 1), then d' and $-d'$ are feasible points for (13). Then it follows $|T_Z(i',:)d'|{=}I_E'{>}\hat{I}_E^*$, which contradicts the assumption that $\hat{I}_E^*$ is the optimal value of (13). Hence, $I_E'{=}\hat{I}_E^*$ holds.

We now establish the lower bound. Let $Z'{\sim}\mathcal{N}(T_Zd', \Sigma_Z)$, and note that Z' is with the finite mean value (integrable) once null($[Q^T\ T_R^T\ F^T]^T$)$\subseteq$null(T_Z). We then have

$$\hat{I}_E^* = I_E' = ||\mathbb{E}\{Z'\}||_\infty \overset{\text{(i)}}{\leq} \mathbb{E}\{||Z'||_\infty\} = I_E(d') \overset{\text{(ii)}}{\leq} I_E^*,$$

where: (i) follows from Lemma 4, since every norm is convex; (ii) follows from the fact that d' is a feasible point of $\mathcal{P}_2$, so $I_E(d')$ has to be lower than the optimal value I_E^* of $\mathcal{P}_2$.

We now establish the upper bound. Let d^* be a solution of $\mathcal{P}_2$ and Z^* be distributed according to $\mathcal{N}(T_Zd^*, \Sigma_Z)$. Note that: (1) Z^* can be written as $Z^*{=}T_Zd^*{+}Z$, where

$Z \sim \mathcal{N}(0, \Sigma_Z)$; (2) $Z^{(i)}$ is a Gaussian random variable with the mean value 0 and the variance $\Sigma_Z(i,i)$, so $|Z^{(i)}|$ is a random variable distributed according to the folded normal distribution [32]. We then have

$$\begin{aligned}\mathbb{E}\{||Z^*||_\infty\} &\overset{\text{(i)}}{\leq} \mathbb{E}\{||Z||_\infty\} + ||T_Z d^*||_\infty \\ &\overset{\text{(ii)}}{\leq} \mathbb{E}\{||Z||_\infty\} + \hat{I}_E^* \overset{\text{(iii)}}{\leq} \sum_{i=1}^{Nn_z} \mathbb{E}\{|Z^{(i)}|\} + \hat{I}_E^* \\ &\overset{\text{(iv)}}{\leq} \sum_{i=1}^{Nn_z} \sqrt{\frac{2\Sigma_Z(i,i)}{\pi}} + \hat{I}_E^* \overset{\text{(v)}}{\leq} Nn_z\sigma_Z^* + \hat{I}_E^*,\end{aligned}$$

where: (i) follows from the triangle inequality and linearity of the expectation operator; (ii) follows from the fact that $||T_Z d^*||_\infty {\leq} ||T_Z d'||_\infty {=} \hat{I}'_E {=} \hat{I}_E^*$; (iii) follows from $||Z||_\infty {\leq} \sum_{i=1}^{Nnz} |Z^{(i)}|$ and linearity of the expectation operator; (iv) follows from the fact that $|Z^{(i)}|$ has the mean value $(\frac{2}{\pi}\Sigma_Z(i,i))^{\frac{1}{2}}$ [32]; (v) follows from the definition of σ_Z^*. ■

Remark 12: Since Σ_Z is independent of d and can be obtained from the system matrices, we only need to calculate $\hat{I}_E^*$ to calculate the bounds. Hence, the bounds can be obtained by solving $\mathcal{P}$ $n_z N$ times, same as the optimal value of $\mathcal{P}_1$.

Remark 13: From (12), we can see that the bounds are tight in at least two cases:(i) $\hat{I}_E^*$ is much larger than $Nn_z\sigma_Z^*$; (ii) σ_Z^* has small value (noise is negligible). Additionally, even if the tightness cannot be established, the bounds can still be useful. If the lower bound (resp. upper bound) is large (resp. small), then the optimal value I_E^* is for sure large (resp. small).

V. Applicability

This section introduces attack strategies whose impact can be analyzed using our framework. The omitted proofs are available in Appendix.

A. DoS, Rerouting, and Sign Alternation Attacks

We first consider three strategies that can be modeled by

$$\tilde{y}(k) = \Lambda_y y(k), \qquad \tilde{u}(k) = \Lambda_u u(k). \tag{15}$$

The first strategy is a DoS attack strategy [17], [18], where the attacker prevents the measurements $\mathcal{Y}_a$ and control actions $\mathcal{U}_a$ from reaching their destination. For example, the attacker can physically damage the corresponding sensors and actuators, or jam the network over which the signals are transmitted [17]. Here, Λ_y and Λ_u are diagonal matrices defined by

$$\Lambda_y(i,i) = \begin{cases} 1, & i \notin \mathcal{Y}_a, \\ 0, & i \in \mathcal{Y}_a, \end{cases} \qquad \Lambda_u(i,i) = \begin{cases} 1, & i \notin \mathcal{U}_a, \\ 0, & i \in \mathcal{U}_a. \end{cases} \tag{16}$$

Remark 14: There are alternative DoS attack models in the literature. For example, instead of setting missing measurement or control signals to zero, one can use the last received values [33], or their estimates [34].

In rerouting attacks [20], the attacker permutes the values of the measurements $\mathcal{Y}_a$ and control actions $\mathcal{U}_a$. The attack can be performed by physically re-wiring the sensor cables, or by modifying the sender's address [20]. In this strategy, Λ_y and Λ_u are permutation matrices that satisfy $\Lambda_y(i,i){=}1$ for $i{\notin}\mathcal{Y}_a$ and $\Lambda_u(i,i){=}1$ for $i{\notin}\mathcal{U}_a$.

Finally, in a sign alternation attack [21], the attacker flips the sign of the measurement $\mathcal{Y}_a$ and the control actions $\mathcal{U}_a$. This attack can turn negative feedback into positive, and potentially destabilize the system. Moreover, in certain configurations, this attack strategy leads to strictly stealthy attacks [21]. In this case, Λ_u and Λ_y are diagonal matrices given by

$$\Lambda_y(i,i) = \begin{cases} -1, & i \in \mathcal{Y}_a, \\ 1, & i \notin \mathcal{Y}_a, \end{cases} \qquad \Lambda_u(i,i) = \begin{cases} -1, & i \in \mathcal{U}_a, \\ 1, & i \notin \mathcal{U}_a. \end{cases}$$

The following proposition establishes compatibility of the above mentioned strategies with our framework.

Proposition 3: The impact estimation problems on DoS, rerouting, and sign alternation attack strategies can be formulated as optimization problems $\mathcal{P}_1$ or $\mathcal{P}_2$.

B. FDI, Bias Injection, and Combined FDI and DoS Attacks

In FDI attacks [6], [15], the attacker is able to manipulate the measurements $\mathcal{Y}_a$ and the control actions $\mathcal{U}_a$, and knows the whole system model. Using these resources, the attacker constructs an optimal attack sequence $a_{0:N}$ that maximizes some impact metric. Signals $\tilde{y}$ and $\tilde{u}$ are given by

$$\tilde{y}(k) = y(k) + \Gamma_y a_y(k), \qquad \tilde{u}(k) = u(k) + \Gamma_u a_u(k), \tag{17}$$

where Γ_y and Γ_u are diagonal matrices defined by

$$\Gamma_y(i,i) = \begin{cases} 1, & i \in \mathcal{Y}_a, \\ 0, & i \notin \mathcal{Y}_a, \end{cases} \qquad \Gamma_u(i,i) = \begin{cases} 1, & i \in \mathcal{U}_a, \\ 0, & i \notin \mathcal{U}_a. \end{cases} \tag{18}$$

In bias injection attacks, the attacker injects a constant bias to the measurements $\mathcal{Y}_a$ and control actions $\mathcal{U}_a$ [10], [13]. Hence, this strategy can be modeled by

$$\tilde{y}(k) = y(k) {+} \Gamma_y a_y(0), \qquad \tilde{u}(k) = u(k) {+} \Gamma_u a_u(0), \tag{19}$$

where Γ_y and Γ_u are defined same as in (18). In fact, one can notice that the only difference in comparison to (17) is that a_u and a_y are now constant.

Finally, one can imagine a situation where the attacker injects corrupted data to the measurements $\mathcal{Y}_I$ and the control actions $\mathcal{U}_I$, but can only deny access to $\mathcal{Y}_D$ and $\mathcal{U}_D$ [22], [23]. This combination of FDI and DoS attacks can be modeled by

$$\tilde{y}(k) = \Lambda_y y(k) {+} \Gamma_y a_y(k), \quad \tilde{u}(k) = \Lambda_u u(k) {+} \Gamma_u a_u(k), \tag{20}$$

where Λ_y and Λ_u are defined based on $\mathcal{Y}_D$ and $\mathcal{U}_D$ as in (16), and Γ_y and Γ_u are defined based on $\mathcal{Y}_I$ and $\mathcal{U}_I$ as in (18).

The injection strategies introduced in this subsection are also compatible with our framework.

Proposition 4: The impact estimation problems on FDI, bias injection, and combined FDI and DoS attack strategies can be formulated as optimization problems $\mathcal{P}_1$ or $\mathcal{P}_2$.

C. Replay Attacks

The replay attack strategy is inspired by the Stuxnet malware [3]. The replay attack on the sensors $\mathcal{Y}_a$ is modeled by

$$\tilde{y}(k) = \Lambda_y y(k) + \Gamma_y a_s(k), \tag{21}$$

where Λ_y is defined as in (16), Γ_y as in (18), and

$$a_s(k) = y(k - N - 1). \tag{22}$$

In other words, the attacker replaces the attacked measurements with the measurements of the normal operation previously recorded at the time steps $-N-1, \dots, -1$. The purpose of attacking the sensors $\mathcal{Y}_a$ is to cover attacks against the actuators $\mathcal{U}_a$, which can be modeled in different ways. For instance, in [28], we modeled actuator attacks as

$$\tilde{u}(k) = u(k) + \Gamma_u a_u(0), \tag{23}$$

where Γ_u is defined as in (18). This captures the case where the attacker sends some large signal to the attacked actuators. Another scenario is a DoS attack against the actuators

$$\tilde{u}(k) = \Lambda_u u(k), \tag{24}$$

where Λ_u is defined as in (16). Both of the previously introduced replay attack strategies are compatible with our framework, as stated in the following proposition.

Proposition 5: The impact estimation problems on replay attack strategies can be formulated as problems $\mathcal{P}_1$ or $\mathcal{P}_2$.

VI. Numerical Example

We now illustrate how the impact estimation framework we proposed can be used for risk assessment, and discuss how the tuning parameters influence the impact of different strategies.

A. System Model

We consider a chemical process from [35] shown in Fig. 1 a). The states are the volume in Tank 3 (x_1), the volume in Tank 2 (x_2), and the temperature in Tank 2 (x_3). 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 assume that the control objective is to keep a constant temperature in Tank 2. The objective is achieved by injecting hot water from Tank 1, and cold water from Tank 3.

The plant is described by

$$A = \begin{bmatrix} 0.96 & 0 & 0 \\ 0.04 & 0.97 & 0 \\ -0.04 & 0 & 0.90 \end{bmatrix}, B = \begin{bmatrix} 8.8 & -2.3 & 0 & 0 \\ 0.20 & 2.2 & 4.9 & 0 \\ -0.21 & -2.2 & 1.9 & 21 \end{bmatrix},$$

$C = I_3$, $\Sigma_{v_x} = 0.05\, I_3$, and $\Sigma_{v_y} = 0.01\, I_3$. The matrices of the controller are given by

$$L = -0.01 \begin{bmatrix} 10 & 1.8 & -0.1 \\ -2.0 & 7.1 & -0.5 \\ 1.4 & 16 & 0.2 \\ -0.4 & -0.7 & 4.2 \end{bmatrix}, L_{y_r} = 0.01 \begin{bmatrix} 11 & 11 & 0 \\ -1 & 44 & 0 \\ 0 & 0 & 0 \\ 0 & 4.7 & 4.7 \end{bmatrix}.$$

We adopted $Q_{y_r} = 0.4\, I_3$, and we used the steady state Kalman filter as an estimator.

A cyber-infrastructure is shown in Fig. 1 b). It was identified that the communication link between Router 1 and the controller is unprotected (vulnerability V_1). The same holds for the link between Router 2 and the controller (vulnerability V_2). If the attacker exploits V_1 (resp. V_2), he/she gains control over sensors y_2, y_3 (resp. y_1), and actuators u_3, u_4 (resp. u_1, u_2).

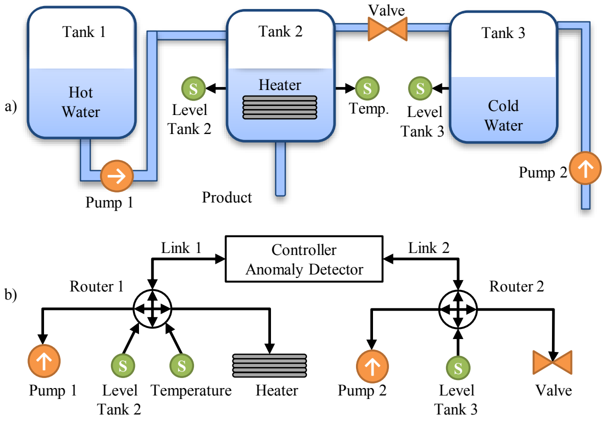

Fig. 1. a) The physical part of a chemical process with four actuators (two pumps, one heater, and one valve), and three sensors (two level sensors and one temperature sensor); b) The cyber part of the process.

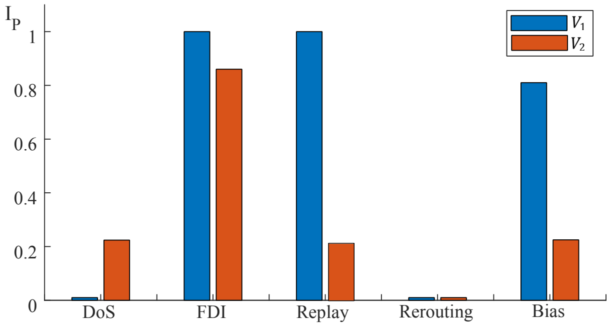

Fig. 2. The impact of different attack strategies when V_1 is exploited (blue) and V_2 is exploited (red).

B. Risk Assessment

We now use our framework to determine which of the vulnerabilities is more threatening. We set $N = 10$, $\epsilon = 0.3$, and $Q_z{=}[0_{1\times 2}\;\; 1/3]$. The metric I_P was used and we considered DoS [17], [18], rerouting [20], replay [19], FDI [6], [15], and bias injection [10], [13] attack strategies. Since the attacker can conduct DoS and rerouting attacks in multiple ways, we calculated the worst case impact for these strategies. For the replay strategy, the attack against the actuators was modeled according to (24).

The results of the analysis are illustrated in Fig. 2. Firstly, note that the impact of different strategies may result in different conclusions concerning the importance of vulnerabilities. Based on the impact of DoS attacks, it follows that V_2 is more important to be prevented than V_1. Yet, based on the impact of replay, FDI, and bias attacks, V_1 is more critical. The impact of rerouting attacks was not informative, since it was equal to zero in both of the cases. This illustrates that in some cases, several attack strategies need to be considered to decide on importance of vulnerabilities. In this case, we can give higher priority to V_1, since the impact of majority of attack strategies is larger for this vulnerability.

Secondly, we point out that sometimes less complex attacks can be just as dangerous as full model knowledge FDI attacks. For example, if V_1 is exploited, replay attacks lead to the

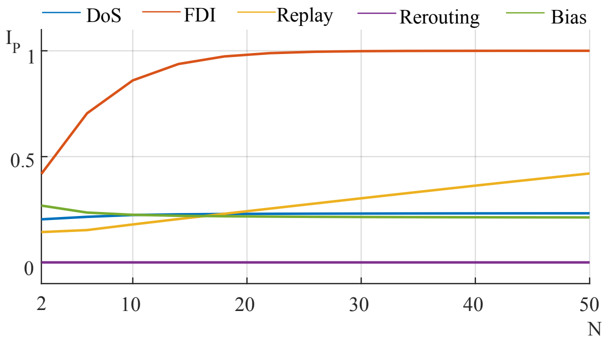

Fig. 3. The impact of different attack strategies with respect to N.

same impact as FDI attacks. Thirdly, we observe that the stochastic model of the system can considerably influence the impact of some attacks. Particularly, rerouting attacks proved to be harmless in this framework, because they were detectable in both of the scenarios. Yet, in our previous study on deterministic systems [28], these attacks had impact comparable with DoS and bias injection attacks.

Finally, once the attacker exploits V_1 and uses FDI attack strategy, he/she can make the deterministic part of the critical state x_3 arbitrarily large (Proposition 2). Namely, by manipulating the compromised actuators, the attacker affects the volume x_2 and the temperature x_3 of Tank 2. Additionally, these changes cannot be seen neither from y_2 and y_3 (controlled by the attacker), nor from y_1 (x_2 and x_3 do not influence x_1).

C. Tuning Parameters

Recall that ϵ and N are the tuning parameters in $\mathcal{P}_1$. By increasing ϵ, the stealthiness constraint becomes easier to satisfy, so the impact is clearly non-decreasing with respect to ϵ. However, the connection of the impact to the horizon length N is not obvious. To illustrate some interesting facts, we investigate how the impact changes when we vary N in the range 2 to 50. We fixed other modeling parameters to be the same as in the previous two sections, assumed V_2 to be exploited, and considered the same attack strategies.

A plot of the impact of different attack strategies with respect to N is shown in Fig. 3. The first observation is that the impact of almost all the strategies seems to converge to a steady state relatively quickly. In fact, only the impact of the replay strategy keeps increasing over time. The second observation is that the impact can also be decreasing with N, as in the case of bias injection attacks. We find the reason to be that the stealthiness constraint becomes harder to satisfy, while the number of decision variables in the problem effectively remains the same. Both of these observations point out that in certain cases, we do not have to consider long horizon lengths to find the worst case attack impact.

VII. Conclusion and Future Work

We proposed a framework for estimating impact of a range of cyber-attack strategies that is independent of the choice of anomaly detector. Furthermore, we suggested two alternatives for the impact metric based on the infinity norm that can be used in stochastic systems. For the first metric, we proved that the optimal value of the impact estimation problem can be obtained by solving a set of convex problems. For the second metric, lower and upper bounds were derived. Additionally, we demonstrated how our framework can be used for risk assessment on an illustrative example.

The future work may go in the following directions. Firstly, a possible extension would be to explore if the impact of feedback attacks can be analyzed using our framework. Secondly, it might be interesting to derive conditions under which the impact is decreasing or increasing with the horizon length N. This may help us to perform risk assessment faster. Finally, we plan to investigate how can one apply our framework to allocate security measures, tune anomaly detectors, or develop a game theoretic based defense strategy.

Appendix

Proof of Lemma 1: Let $N_s \in \mathbb{Z}$ and $N_s < 0$. We first prove that $x_e(N_s)$ is distributed according to $\mathcal{N}(T_0 y_r, \Sigma_0)$. In absence of attacks, the extended state x_e is given by

$$x_e(k+1) = A_e x_e(k) + B_e v(k) + E_e y_r, \tag{25}$$

where A_e,B_e, and C_e are respectively obtained from $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ by setting $\Lambda_y = I_{n_y}$, $\Lambda_u = I_{n_u}$, $\Gamma_y = 0_{n_y \times n_y}$, and $\Gamma_u = 0_{n_u \times n_u}$. We denote the covariance matrix of v by Σ_v. Let y_r=0, and recall that we assumed that the system has reached the stationary regime and that A_e is asymptotically stable. Under these assumptions, x_e is zero mean Gaussian stationary process with the covariance matrix obtained as the solution of the Lyapunov equation $\Sigma_0{=}A_e \Sigma_0 A_e^T{+}B_e \Sigma_v B_e^T$ (see [29, Chapter 4]). Once $y_r \neq 0$, only the mean value of the process changes. Since the system has reached the stationary regime, it follows that $\mathbb{E}\{x_e(N_s{+}1)\}{=}\mathbb{E}\{x_e(N_s)\}$. Hence, from (25), we have $\mathbb{E}\{x_e(N_s)\}{=}T_0 y_r$, where $T_0{=}(I_{2n_x}{-}A_e)^{-1}E_e$.

We now prove that $z_{1:N}$ is distributed according to $\mathcal{N}(T_Z d, \Sigma_Z)$. From (C1), (6), and (25), we have

$$z_{0:N} = P_1 x_e(N_s) + P_2 v_{N_s:N} + P_3 y_r + P_4 a_{0:N} + P_5 a_{s0:N}, \tag{26}$$

where $P_1{=}\mathcal{O}_N(\tilde{A}, Q_z')A_e^{|N_s|}$,

$$P_2 = [\mathcal{O}_N(\tilde{A}, Q_z')\mathcal{C}_{N_s}(A_e, B_e) \;\; \mathcal{T}_N(\tilde{A}, \tilde{B}, Q_z', 0_{n_z \times n_v})],$$

$$P_3 = \mathcal{O}_N(\tilde{A}, Q_z')\sum_{i=0}^{|N_s|-1} A_e^i E_e {+} \mathcal{T}_N(\tilde{A}, \tilde{E}, Q_z', 0_{n_z \times n_{y_r}})(1_{N+1} {\otimes} I_{n_{y_r}}),$$

$P_4{=}\mathcal{T}_N(\tilde{A}, \tilde{G}, Q_z', 0_{n_z \times n_a})$, $P_5{=}\mathcal{T}_N(\tilde{A}, \tilde{J}, Q_z', 0_{n_z \times n_y})$, and $Q_z'{=}[Q_z \; 0_{n_z \times n_x}]$. Next, from (26) and (C5), it follows that

$$z_{1:N} = P_1' x_e(N_s) + P_2' v_{N_s:N} + P_3' y_r + P_4' a_{0:N}, \tag{27}$$

where $P_1'{=}P_l(P_1{+}P_5 T_1)$, $P_2'{=}P_l(P_2{+}[P_5 T_2 \; 0_{N' n_z \times N' n_v}])$, $P_3'{=}P_l(P_3 + P_5 T_3)$, $P_4'{=}P_l P_4$, $P_l = [0_{N n_z \times n_z} \;\; I_{N n_z}]$, and $N'{=}N{+}1$. Since $x_e(N_s)$ and $v_{N_s:N}$ are independent Gaussian vectors, and $a_{0:N}$ and y_r are deterministic, $z_{1:N}$ is a Gaussian vector. Using the linearity property of the expected value and the fact that $x_e(N_s) \sim \mathcal{N}(T_0 y_r, \Sigma_0)$, we get

$$\mathbb{E}\{z_{1:N}\} = P_1' T_0 y_r + P_3' y_r + P_4' a_{0:N} = T_Z d, \tag{28}$$

where $T_Z{=}[P_4'\ P_1'T_0{+}P_3']$. Let the covariance matrix of $v_{N_s:N}$ be denoted with Σ_V. We then have

$$\Sigma_Z = P_1'\Sigma_0(P_1')^T + P_2'\Sigma_V(P_2')^T, \tag{29}$$

where we used that $x_e(N_s)$ and $v_{N_s:N}$ are independent. Notice that T_Z and Σ_Z are independent of d.

We similarly prove that $\tilde{r}_{0:N}$ is distributed according to $\mathcal{N}(T_Rd, \Sigma_R)$. From (C1) and (25), $\tilde{r}_{0:N}$ can be written as

$$\tilde{r}_{0:N} = M_1x_e(N_s) + M_2v_{N_s:N} + M_3y_r + M_4a_{0:N} + M_5a_{s0:N}, \tag{30}$$

where $M_1 = \mathcal{O}_N(\tilde{A},\tilde{C})A_e^{|N_s|}$,

$$M_2 = [\mathcal{O}_N(\tilde{A},\tilde{C})\mathcal{C}_{N_s}(A_e,B_e)\ \ \mathcal{T}_N(\tilde{A},\tilde{B},\tilde{C},\tilde{D})],$$

$$M_3 = \mathcal{O}_N(\tilde{A},\tilde{C})\sum_{i=0}^{|N_s|-1} A_e^iE_e + \mathcal{T}_N(\tilde{A},\tilde{E},\tilde{C},\tilde{F})(1_{N+1}\otimes I_{n_{y_r}}),$$

$M_4{=}\mathcal{T}_N(\tilde{A},\tilde{G},\tilde{C},\tilde{H})$, and $M_5{=}\mathcal{T}_N(\tilde{A},\tilde{J},\tilde{C},\tilde{K})$. From (30) and (C5), it follows that

$$\tilde{r}_{0:N} = M_1'x_e(N_s) + M_2'v_{N_s:N} + M_3'y_r + M_4a_{0:N}, \tag{31}$$

where $M_1'{=}M_1{+}M_5T_1$, $M_2'{=}M_2{+}[M_5T_2\ 0_{N'n_y\times N'n_v}]$, and $M_3'{=}M_3 + M_5T_3$. Thus, $\tilde{r}_{0:N}$ is a Gaussian vector, since $x_e(N_s)$ and $v_{N_s:N}$ are independent Gaussian vectors, and $a_{0:N}$ and y_r are deterministic. The mean value of $\tilde{r}_{0:N}$ is

$$\mathbb{E}\{\tilde{r}_{0:N}\} = M_1'T_0y_r + M_3'y_r + M_4a_{0:N} = T_Rd, \tag{32}$$

where $T_R{=}[M_4\ M_1'T_0{+}M_3']$. The covariance matrix of $\tilde{r}_{0:N}$ is

$$\Sigma_R = M_1'\Sigma_0(M_1')^T + M_2'\Sigma_V(M_2')^T. \tag{33}$$

Finally, we can see that T_R and Σ_R are independent of d.

It remains to prove that Σ_Z is positive definite. Equation (27) can be rewritten as

$$\begin{aligned} z_{1:N} = P_1'x_e(N_s) &+ P_{vp}v_{N_s:-1} + P_{v_x}v_{x0:N-1} + P_{v_y}v_{y0:N-1} \\ &+ P_3'y_r + P_4a_{0:N} + 0_{Nn_z\times n_x}v_x(N) + 0_{Nn_z\times n_y}v_y(N). \end{aligned}$$

Since $P_1'x_e(N_s)$, $P_{vp}v_{N_s:-1}$, $P_{v_x}v_{x0:N-1}$, $P_{v_y}v_{y0:N-1}$ are independent Gaussian vectors, Σ_Z is the sum of the covariance matrices of these vectors. Thus, it suffices to prove that one of these vectors has a positive definite covariance matrix. From (1) and (6), P_{vx} is of the form

$$P_{vx} = \begin{bmatrix} Q_z & \dots & 0_{n_z\times n_x} \\ \vdots & \ddots & \vdots \\ \times & \dots & Q_z \end{bmatrix}.$$

It then follows that $\text{null}(P_{vx}^T) = \emptyset$, since Q_z has full row rank. Let Σ_{V_x} be the covariance matrix of $v_{x0:N-1}$. Since Σ_{v_x} is positive definite, so it is Σ_{V_x}. Thus, $P_{vx}\Sigma_{V_x}(P_{vx})^T$ is positive definite, which implies that Σ_Z is positive definite. ■

Proof of Lemma 2: Let Y_1 and Y_2 be random vectors with the distributions $\mathcal{N}(\mu_1,\Sigma_1)$ and $\mathcal{N}(\mu_2,\Sigma_2)$, respectively. Let Σ_1 and Σ_2 be positive definite. Then

$$\mathcal{D}(Y_1||Y_2) = 0.5\left(\text{tr}(\Sigma_2^{-1}\Sigma_1) + ||\mu_2-\mu_1||_{\Sigma_2^{-1}} + \ln\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - n\right),$$

where n is the dimension of Y_1 and Y_2 [36]. In our case, the distributions of $\tilde{r}_{0:N}$ and $r_{0:N}$ are $\mathcal{N}(T_Rd,\Sigma_R)$ and $\mathcal{N}(0_{(N+1)n_y\times 1}, I_{(N+1)n_y})$, respectively. Thus, it follows

$$\begin{aligned} \mathcal{D}(\tilde{r}_{0:N}||r_{0:N}) = 0.5\big(&\text{tr}(\Sigma_R) + ||T_Rd||_2^2 - \ln\ \det(\Sigma_R) \\ &- (N+1)n_y\big) = c_{KL} + 0.5||T_Rd||_2^2, \end{aligned}$$

where $c_{KL}{=}0.5(\text{tr}(\Sigma_R){-}\ln\ \det(\Sigma_R){-}(N{+}1)n_y)$.

Hence, (C3) becomes $||T_Rd||_2^2 \le \epsilon'$, where $\epsilon' = 2((N{+}1)\epsilon{-}c_{KL}) = (N{+}1)(2\epsilon{+}n_y){-}\text{tr}(\Sigma_R){+}\ln\ \det(\Sigma_R)$. ■

Proof of Lemma 3: From Lemma 1, $z_{1:N}$ is a Gaussian random vector with a non-degenerate distribution $\mathcal{N}(T_Zd,\Sigma_Z)$. Thus, $z_{1:N}^{(i)}$ is a Gaussian random variable with the distribution $\mathcal{N}(\mu,\sigma^2)$, where $\mu = T_Z(i,:)d$, and $\sigma^2 = \Sigma_Z(i,i)$. Hence, d influences the distribution of $z_{1:N}^{(i)}$ only through μ.

Next, we outline two properties of $\mathbb{P}(|z_{1:N}^{(i)}|{\ge}1)$ with respect to μ. Let $f(\mu){=}\mathbb{P}(|z_{1:N}^{(i)}|{\ge}1)$, $c_1{=}(\sqrt{2}\sigma)^{-1}$, and note that $|z_{1:N}^{(i)}|$ is distributed according to the folded normal distribution [32]. Therefore, it follows

$$f(\mu) = 1 - 0.5\text{erf}(c_1 - c_1\mu) - 0.5\text{erf}(c_1 + c_1\mu), \tag{34}$$

where $\text{erf}(x){=}\frac{1}{\sqrt{\pi}}\int_{-x}^{x}e^{-t^2}dt$ is the error function [32]. From (34), we can observe that $f(-\mu){=}f(\mu)$. Hence, $f(\mu)$ is symmetric in μ (Property 1). Next, we have

$$\begin{aligned} \frac{df(\mu)}{d\mu} &= -0.5\frac{2}{\sqrt{\pi}}e^{-c_1^2(1-\mu)^2}(-c_1) - 0.5\frac{2}{\sqrt{\pi}}e^{-c_1^2(1+\mu)^2}c_1 \\ &= \frac{c_1}{\sqrt{\pi}}\big(e^{-c_1^2(1-\mu)^2} - e^{-c_1^2(1+\mu)^2}\big), \end{aligned}$$

where we used the property $\frac{d}{dz}\ \text{erf}(z) = \frac{2}{\sqrt{\pi}}e^{-z^2}$. Since $e^{-c_1^2(1+\mu)^2}{<}e^{-c_1^2(1-\mu)^2}$ (resp. $e^{-c_1^2(1+\mu)^2}{>}e^{-c_1^2(1-\mu)^2}$) once $\mu{>}0$ (resp. $\mu{<}0$) , it follows that f is monotonically increasing (resp. decreasing) with respect to μ on the interval $(0,+\infty)$ (resp. $(-\infty,0)$). Due to this fact and Property 1, we have that $f(\mu)$ is increasing with respect to $|\mu|$ (Property 2).

We are now ready to prove the claim of the lemma. Since $\mathcal{C}_d$ is a symmetric constraint and $\mathbb{E}\{z_{1:N}^{(i)}\} = T_Z(i,:)d$, we have that a solution of (8) also represents a solution of

$$\underset{d}{\text{maximize}}\ \ |\mathbb{E}\{z_{1:N}^{(i)}\}| \qquad \text{subject to } \mathcal{C}_d.$$

From Property 2, we have $\mathbb{P}(|z_{1:N}^{(i)}| \ge 1)$ to be increasing with respect to $|\mathbb{E}\{z_{1:N}^{(i)}\}|$. Hence, d^* is also a solution of (9). ■

Proof of Proposition 3: To show compatibility, we need to verify if the strategies can be imposed through (C4) and (C5). From (15), it follows that a_y, a_u, and a_s are zero. These constraints on a_y and a_u can be modeled by (C4), by setting $F_a{=}I_{(N+1)n_a}$. The constraint on a_s can be modeled by (C5), by setting T_1–T_3 to zero. ■

Proof of Proposition 4: Same as in the previous proposition, we verify compatibility by showing that the strategies can be imposed through (C4) and (C5). Consider first FDI attack strategy. No constraints are imposed on a_u and a_y, which can be modeled by (C4) by setting F_a to zero. The constraints $a_s(k){=}0$ for $k{=}0,\dots,N$ can be modeled by (C5) by setting T_1–T_3 to zero. The proof for the bias injection strategy is similar as for FDI attacks. The only difference are the additional constraints on a_y and a_u, that can be written as $a_y(k){=}a_y(0)$, $a_u(k){=}a_u(0)$, where $k{=}1,\dots,N$. These constraints are linear

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCNS.2019.2940253, IEEE Transactions on Control of Network Systems

equality constraints that can be modeled by (C4). In the case of combined FDI and DoS attacks (20), a_u and a_y are free to select, and $a_s(k) = 0$ for $k{=}0,\dots,N$. As discussed above, having a_u and a_y free to choose can be modeled by (C4), and having $a_s(k){=}0$ for $k{=}0,\dots,N$ can be modeled by (C5). ■

Proof of Proposition 5: From (21), we have $a_y(k) = 0$ for $k{=}0,\dots,N$, which can be modeled by (C4). The same holds for a_u if (24) is used to model the actuator attack. If (23) is used, a_u needs to satisfy $a_u(k) = a_u(0)$ for $k{=}1,\dots,N$, which can also be modeled by (C4). We now show that a_s can be imposed through (C5). Let $N_s{=}{-}N{-}1$, $C_e{=}[C\ 0_{n_y \times n_x}]$, and $D_e{=}[0_{n_y \times n_x}\ I_{n_y}]$. In absence of attacks, we have

$$y(k) = C_e x_e(k) + D_e v(k). \tag{35}$$

From (22), we have $a_{s0:N}{=}y_{N_s:-1}$. From the later, (25), and (35), it follows that

$$a_{s0:N} = T_1 x_e(N_s) + T_2 v_{N_s:-1} + T_3 y_r,$$

where $T_1{=}\mathcal{O}_N(A_e, C_e)$, $T_2{=}\mathcal{T}_N(A_e, B_e, C_e, D_e)$, and $T_3{=}\mathcal{T}_N(A_e, E_e, C_e, 0_{n_y \times n_{y_r}})(1_{N+1} \otimes I_{n_{y_r}})$. Hence, the constraint (22) on a_s can be modeled by (C5). ■

References

[1] K. Stouffer, J. Falco, and K. Scarfone, “Guide to industrial control systems security,” *National Institute of Standards & Technology*, 2011.

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

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

[4] R. M. Lee, M. J. Assante, and T. Conway, “Analysis of the cyber attack on the Ukrainian power grid,” *Electricity Information Sharing and Analysis Center & SANS*, 2016.

[5] G. Stoneburner, A. Y. Goguen, and A. Feringa, “Risk management guide for information technology systems,” *National Institute of Standards & Technology*, 2002.

[6] Y. Mo and B. Sinopoli, “On the performance degradation of cyber-physical systems under stealthy integrity attacks,” *IEEE Transactions on Automatic Control*, vol. 61, no. 9, pp. 2618–2624, 2016.

[7] C. Murguia, N. van de Wouw, and J. Ruths, “Reachable sets of hidden CPS sensor attacks: Analysis and synthesis tools,” *IFAC Proceedings Volumes*, vol. 50, no. 1, pp. 2088 – 2094, 2017.

[8] I. Jovanov and M. Pajic, “Sporadic data integrity for secure state estimation,” in *Proceedings of the 55th IEEE Conference on Decision and Control (CDC)*, 2017.

[9] Y. Chen, S. Kar, and J. M. F. Moura, “Optimal attack strategies subject to detection constraints against cyber-physical systems,” *IEEE Transactions on Control of Network Systems*, vol. PP, no. 99, pp. 1–1, 2017.

[10] J. Milošević, T. Tanaka, H. Sandberg, and K. H. Johansson, “Analysis and mitigation of bias injection attacks against a Kalman filter,” *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 6484–6489, 2017.

[11] 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*, 2011.

[12] T. R, C. Murguia, and J. Ruths, “Tuning windowed chi-squared detectors for sensor attacks,” in *Proceedings of the American Control Conference (ACC)*, 2018.

[13] C. M. Ahmed, C. Murguia, and J. Ruths, “Model-based attack detection scheme for smart water distribution networks,” in *Proceedings of the Asia Conference on Computer and Communications Security*, 2017.

[14] C. Murguia, I. Shames, J. Ruths, and D. Nesic, “Security metrics of networked control systems under sensor attacks (extended preprint),” *arXiv preprint arXiv:1809.01808*, 2018.

[15] D. Umsonst, H. Sandberg, and A. A. Cárdenas, “Security analysis of control system anomaly detectors,” in *Proceedings of the American Control Conference (ACC)*, 2017.

[16] N. H. Hirzallah and P. G. Voulgaris, “On the computation of worst attacks: a LP framework,” in *Proceedings of the American Control Conference (ACC)*, 2018.

[17] A. A. Cárdenas, S. Amin, and S. Sastry, “Research challenges for the security of control systems.” in *Proceedings of the 3rd USENIX Workshop on Hot Topics in Security (HotSec)*, 2008.

[18] S. Amin, A. A. Cárdenas, and S. Sastry, “Safe and secure networked control systems under denial-of-service attacks.” in *Proceedings of the 12th International Conference on Hybrid Systems: Computation and Control*, 2009.

[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, 2014.

[20] R. M. Ferrari and A. M. Teixeira, “Detection and isolation of routing attacks through sensor watermarking,” in *Proceedings of the American Control Conference (ACC)*, 2017.

[21] C.-Z. Bai and V. Gupta, “On kalman filtering in the presence of a compromised sensor: Fundamental performance bounds,” in *Proceedings of the American control conference (ACC)*. IEEE, 2014, pp. 3029–3034.

[22] R. Anguluri, V. Gupta, and F. Pasqualetti, “Periodic coordinated attacks against cyber-physical systems: Detectability and performance bounds,” in *Proceedings of the 55th Conference on Decision and Control (CDC)*, 2016.

[23] K. Pan, A. M. H. Teixeira, M. Cvetkovic, and P. Palensky, “Combined data integrity and availability attacks on state estimation in cyber-physical power grids,” in *Proceedings of the IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2016.

[24] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, “Worst-case stealthy innovation-based linear attack on remote state estimation,” *Automatica*, vol. 89, pp. 117 – 124, 2018.

[25] C. Z. Bai, V. Gupta, and F. Pasqualetti, “On Kalman filtering with compromised sensors: Attack stealthiness and performance bounds,” *IEEE Transactions on Automatic Control*, vol. 62, no. 12, pp. 6641–6648, 2017.

[26] S. Weerakkody, B. Sinopoli, S. Kar, and A. Datta, “Information flow for security in control systems,” in *Proceedings of the 55th Conference on Decision and Control (CDC)*, 2016.

[27] E. Kung, S. Dey, and L. Shi, “The performance and limitations of ϵ-stealthy attacks on higher order systems,” *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 941–947, 2017.

[28] J. Milošević, D. Umsonst, H. Sandberg, and K. H. Johansson, “Quantifying the impact of cyber-attack strategies for control systems equipped with an anomaly detector,” in *Proceedings of the European Control Conference*, 2018.

[29] B. D. O. Anderson and J. Moore, *Optimal filtering*. Courier Corporation, 2012.

[30] T. Lipp and S. Boyd, “Antagonistic control,” *Systems & Control Letters*, vol. 98, pp. 44 – 48, 2016.

[31] M. D. Perlman, “Jensen’s inequality for a convex vector-valued function on an infinite-dimensional space,” *Journal of Multivariate Analysis*, vol. 4, no. 1, pp. 52 – 65, 1974.

[32] M. Tsagris, C. Beneki, and H. Hassani, “On the folded normal distribution,” *Mathematics*, vol. 2, no. 1, pp. 12–28, 2014.

[33] L. Schenato, “To zero or to hold control inputs with lossy links?” *IEEE Transactions on Automatic Control*, vol. 54, no. 5, pp. 1093–1099, May 2009.

[34] E. Henriksson, H. Sandberg, and K. H. Johansson, “Reduced-order predictive outage compensators for networked systems,” in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. IEEE, 2009, pp. 3775–3780.

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

[36] J. Duchi, “Derivations for linear algebra and optimization,” *Berkeley, California*, vol. 3, 2007.

Jezdimir Milošević received his M.Sc. degree in Electrical Engineering and Computer Science in 2015 from the School of Electrical Engineering, University of Belgrade, Serbia. He is currently a Ph.D. student at the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden. He was a visiting student researcher at the University of Hawaii at Manoa in 2014 and Massachusetts Institute of Technology in 2018 and 2019. His research interests are within cyber-security of industrial control systems.

Henrik Sandberg is Professor at the Department of Automatic Control, KTH Royal Institute of Technology, Stockholm, Sweden. He received the M.Sc. degree in engineering physics and the Ph.D. degree in automatic control from Lund University, Lund, Sweden, in 1999 and 2004, respectively. From 2005 to 2007, he was a Post-Doctoral Scholar at the California Institute of Technology, Pasadena, USA. In 2013, he was a visiting scholar at the Laboratory for Information and Decision Systems (LIDS) at MIT, Cambridge, USA. He has also held visiting appointments at the Australian National University and the University of Melbourne, Australia. His current research interests include security of cyber-physical systems, power systems, model reduction, and fundamental limitations in control. Dr. Sandberg was a recipient of the Best Student Paper Award from the IEEE Conference on Decision and Control in 2004, an Ingvar Carlsson Award from the Swedish Foundation for Strategic Research in 2007, and Consolidator Grant from the Swedish Research Council in 2016. He has served on the editorial board of IEEE Transactions on Automatic Control and is currently Associate Editor of the IFAC Journal Automatica.

Karl Henrik Johansson is Director of the Stockholm Strategic Research Area ICT The Next Generation and Professor at the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology. He received MSc and PhD degrees from Lund University. He has held visiting positions at UC Berkeley, Caltech, NTU, HKUST Institute of Advanced Studies, and NTNU. His research interests are in networked control systems, cyber-physical systems, and applications in transportation, energy, and automation. He is a member of the IEEE Control Systems Society Board of Governors, the IFAC Executive Board, and the European Control Association Council. He has received several best paper awards and other distinctions. He has been awarded Distinguished Professor with the Swedish Research Council and Wallenberg Scholar. He has received the Future Research Leader Award from the Swedish Foundation for Strategic Research and the triennial Young Author Prize from IFAC. He is Fellow of the IEEE and the Royal Swedish Academy of Engineering Sciences, and he is IEEE Distinguished Lecturer.