Quantifying Cyber-Security for Networked Control Systems

[1]André Teixeira, [2]Kin C. Sou, [1]Henrik Sandberg, and [1]Karl H. Johansson

[1]ACCESS Linnaeus Centre and Automatic Control Lab,
KTH Royal Institute of Technology,
Stockholm, Sweden
[2]Mathematical Sciences,
Chalmers University of Technology and University of Gothenburg,
Gothenburg, Sweden

in: Control of Cyber-Physical Systems. See also BibTeX entry below.

BibTeX:

```
@incollection{
year={2013},
isbn={978-3-319-01158-5},
booktitle={Control of Cyber-Physical Systems},
volume={449},
series={Lecture Notes in Control and Information Sciences},
editor={Tarraf, Danielle C.},
doi={10.1007/978-3-319-01159-2_7},
title={Quantifying Cyber-Security for Networked Control Systems},
url={http://dx.doi.org/10.1007/978-3-319-01159-2_7},
publisher={Springer International Publishing},
keywords={Security; Networked Control Systems; Impact Analysis},
author={Teixeira, Andr\'{e} and Sou, Kin~C. and Sandberg, Henrik and Johansson, Karl~H.},
pages={123-142}
}
```

Quantifying Cyber-Security for Networked Control Systems

André Teixeira[1], Kin Cheong Sou[2], Henrik Sandberg[1], and Karl H. Johansson[1]

[1] ACCESS Linnaeus Centre and Automatic Control Lab,
KTH Royal Institute of Technology, Stockholm, Sweden,
{andretei,hsan,kallej}@kth.se.
[2] Mathematical Sciences,
Chalmers University of Technology and University of Gothenburg,
Gothenburg, Sweden,
kincheong.sou@chalmers.se

Abstract. In this paper we consider a typical architecture for a networked control system under false-data injection attacks. Under a previously proposed adversary modeling framework, various formulations for quantifying cyber-security of control systems are proposed and formulated as constrained optimization problems. These formulations capture trade-offs in terms of attack impact on the control performance, attack detectability, and adversarial resources. The formulations are then discussed and related to system theoretic concepts, followed by numerical examples illustrating the various trade-offs for a quadruple-tank process.

Keywords: Security, Networked Control Systems, Impact Analysis

1 Introduction

Critical infrastructure security is of utmost importance in modern society and has been a major concern in recent years. The increasing complexity of these systems and the desire to improve their efficiency and flexibility has led to the use of heterogeneous IT systems, which support the timely exchange of data among and across different system layers, from the corporate level to the local control level. Furthermore, IT infrastructures are composed of heterogeneous components from several vendors and often use non-proprietary communication networks. Therefore the amount of cyber threats to these IT infrastructures has greatly increased over the past years, given their large number of possible attack points across the system layers. There are several examples of cyber threats being exploited by attackers to disrupt the behavior of physical processes, including a staged attack on a power generator [10] and the recent Stuxnet virus attack on centrifuges' control system [16, 12]. Hence monitoring and mitigating cyber attacks to these systems is crucial, since they may bring disastrous consequences to society. This is well illustrated by recalling the consequences of the US-Canada 2003 blackout [19], partially due to lack of awareness in the control center.

A particular type of complex cyber attack is that of false-data injection, where the attacker introduces corrupted data in the communication network. Several instances of this scenario have been considered in the context of control systems, see [2, 4, 15] and references therein. In this paper we address stealthy false-data injection attacks that are constructed so that they are not detected based on the control input and measurement data available to anomaly detectors. A sub-class of these attacks have been recently addressed from a system theoretic perspective. In [14] the author characterizes the set of attack policies for covert (stealthy) false-data injection attacks with detailed model knowledge and full access to all sensor and actuator channels, while [11] described the set of stealthy false-data injection attacks for omniscient attackers with full-state information, but possibly compromising only a subset of the existing sensors and actuators. Similarly, the work in [5] considers a finite time-interval and characterizes the number of corrupted channels that cannot be detected during that time-interval. In the previous approaches the control input and measurement data available to the anomaly detector with and without the attack were the same, thus rendering the attack undetectable. Instead in this paper we allow more freedom to the adversary and consider attacks that may be theoretically detectable, but are still stealthy since they do not trigger any alarm by the anomaly detector.

Contributions and outline

In this paper we consider the typical architecture for a networked control system under false-data injection attacks and adversary models presented in [17]. Under this framework, various formulations for quantifying cyber-security of control systems are proposed and formulated as constrained optimization problems. These formulations capture trade-offs in terms of impact on the control system, attack detectability, and adversarial resources. In particular, one of the formulations considers the minimum number of data channels that need to be corrupted so that the adversary remains stealthy, similarly to the security index for static systems proposed in [13]. The formulations are related to system theoretic concepts.

The outline of the paper is as follows. The control system architecture and model are described in Section 2, followed by the adversary model in Section 3. Different formulations quantifying cyber-security of control systems are introduced in Section 4 for a given time-horizon and in Section 5 for steady-state. A particular formulation is posed as a mixed integer linear program and illustrated through numerical examples in Section 6, followed by conclusions in Section 7.

Notation and Preliminaries

Let $\mathbf{x}_{[k_0,\, k_f]} = \{x_{k_0},\ x_{k_0+1},\ \dots,\ x_{k_f}\}$ be a discrete-time signal in the time-interval $[k_0,\ k_f] = \{k_0,\ \dots,\ k_f\}$ with $x_k \in \mathbb{R}^n$ for $k \in [k_0,\ k_f]$. For simplicity, we also denote the time-domain signal of x_k in vector form as $\mathbf{x}_{[k_0,\, k_f]} \in \mathbb{R}^{n(k_f - k_0 + 1)}$, with $\mathbf{x}_{[k_0,\, k_f]} = [x_{k_0}^\top,\ \dots,\ x_{k_f}^\top]^\top$. When the time-interval at consideration is clear, the short-form notation $\mathbf{x}$ will be used in place of $\mathbf{x}_{[k_0,\, k_f]}$.

For $y \in \mathbb{C}^n$, denote the p-norm of y as $\|y\|_p \triangleq \left(\sum_{i=1}^{n} |y_{(i)}|^p\right)^{1/p}$ for $1 \leq p < \infty$, where $y_{(i)}$ is the i-th entry of the vector y, and let $\|y\|_\infty \triangleq \max_i \ |y_{(i)}|$. Additionally, we denote $\|y\|_0$ as the number of non-zero elements of y and define $\mathbb{S} = \{z \in \mathbb{C} : \ |z| = 1\}$ as the unit circle in the complex plane.

As for the discrete-time signal $\mathbf{x}$, denote its ℓ_p-norm in the time-interval $[k_0, \ k_f]$ as $\|\mathbf{x}\|_{\ell_p[k_0,\, k_f]} \triangleq \|\mathbf{x}_{[k_0,\, k_f]}\|_p = \left(\sum_{k=k_0}^{k_f} \|x_k\|_p^p\right)^{1/p}$ for $1 \leq p < \infty$, and let $\|\mathbf{x}\|_{\ell_\infty[k_0,\, k_f]} \triangleq \sup_{k \in [k_0,\, k_f]} \|x_k\|_\infty$.

For a given matrix $G \in \mathbb{C}^{n \times m}$, denote its Hermitian conjugate as G^H and, supposing G is full-column rank, let $G^\dagger = (G^H G)^{-1} G^H$ be its pseudo-inverse.

2 Networked Control System

In this section we describe the networked control system structure, where we consider three main components as illustrated in Fig. 1: the physical plant and communication network, the feedback controller, and the anomaly detector.

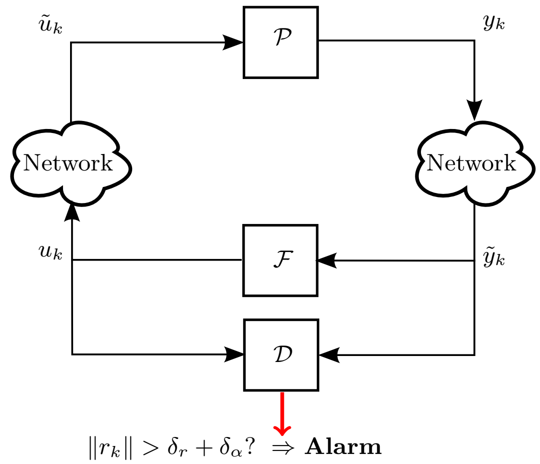


Fig. 1. Schematic of networked control system.

2.1 Physical Plant and Communication Network

The physical plant is modeled in a discrete-time state-space form

$$\mathcal{P} : \begin{cases} x_{k+1} = Ax_k + B\tilde{u}_k + Gw_k + Ff_k \\ \quad y_k = Cx_k + v_k \end{cases}, \tag{1}$$

where $x_k \in \mathbb{R}^n$ is the state variable, $\tilde{u}_k \in \mathbb{R}^{n_u}$ the control actions applied to the process, $y_k \in \mathbb{R}^{n_y}$ the measurements from the sensors at the sampling instant

$k \in \mathbb{Z}$, and $f_k \in \mathbb{R}^d$ is the unknown signal representing the effects of anomalies, usually denoted as fault signal in the fault diagnosis literature [3]. The process and measurement Gaussian noise, $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^{n_y}$, represent the discrepancies between the model and the real process, due to unmodeled dynamics or disturbances, for instance, and we assume their means are respectively bounded by δ_w and δ_v, i.e. $\bar{w} = \|\mathbb{E}\{w_k\}\| \leq \delta_w$ and $\bar{v} = \|\mathbb{E}\{v_k\}\| \leq \delta_v$.

The physical plant operation is supported by a communication network through which the sensor measurements and actuator data are transmitted, which at the plant side correspond to y_k and $\tilde{u}_k$, respectively. At the controller side we denote the sensor and actuator data by $\tilde{y}_k \in \mathbb{R}^{n_y}$ and $u_k \in \mathbb{R}^{n_u}$, respectively. Since the communication network may be unreliable, the data exchanged between the plant and the controller may be altered, resulting in discrepancies in the data at the plant and controller ends. In this paper we do not consider the usual communication network effects such as packet losses and delays. Instead we focus on data corruption due to malicious cyber attacks, as described in Section 3. Therefore the communication network *per se* is supposed to be reliable, not affecting the data flowing through it.

Given the physical plant model (1) and assuming an ideal communication network, the networked control system is said to have a *nominal behavior* if $f_k = 0$, $\tilde{u}_k = u_k$, and $\tilde{y}_k = y_k$. The absence of either one of these condition results in an abnormal behavior of the system.

2.2 Feedback Controller

In order to comply with performance requirements in the presence of the unknown process and measurement noises, we consider that the physical plant is controlled by an appropriate linear time-invariant feedback controller [20]. The output feedback controller can be written in a state-space form as

$$\mathcal{F} : \begin{cases} z_{k+1} = A_c z_k + B_c \tilde{y}_k \\ u_k = C_c z_k + D_c \tilde{y}_k \end{cases} \tag{2}$$

where the states of the controller, $z_k \in \mathbb{R}^{n_z}$, may include the process state and tracking error estimates. Given the plant and communication network models, the controller is supposed to be designed so that acceptable performance is achieved under nominal behavior.

2.3 Anomaly Detector

In this section we consider the anomaly detector that monitors the system to detect possible anomalies, i.e. deviations from the nominal behavior. The anomaly detector is supposed to be collocated with the controller, therefore it only has access to $\tilde{y}_k$ and u_k to evaluate the behavior of the plant.

Several approaches to detecting malfunctions in control systems are available in the fault diagnosis literature [3, 6]. Here we consider the following observerbased fault detection filter

$$\mathcal{D}:\begin{cases}\hat{x}_{k+1|k} = A\hat{x}_{k|k} + Bu_k\\ \hat{x}_{k|k} = \hat{x}_{k|k-1} + K(\tilde{y}_k - C\hat{x}_{k|k-1})\ ,\\ r_k = V(\tilde{y}_k - \hat{y}_{k|k})\end{cases} \tag{3}$$

where $\hat{x}_{k|k} \in \mathbb{R}^n$ and $\hat{y}_{k|k} = C\hat{x}_{k|k} \in \mathbb{R}^{n_y}$ are the state and output estimates given measurements up until time k, respectively, and $r_k \in \mathbb{R}^{n_r}$ the residue evaluated to detect and locate existing anomalies. The previous filter dynamics can be rewritten as

$$\mathcal{D}:\begin{cases}\hat{x}_{k+1|k} = A(I-KC)\hat{x}_{k|k-1} + Bu_k + AK\tilde{y}_k\\ r_k = V[(I-CK)\tilde{y}_k - (I-CK)C\hat{x}_{k|k-1}].\end{cases} \tag{4}$$

The anomaly detector is designed by choosing K and V such that

1. under nominal behavior of the system (i.e., $f_k = 0$, $u_k = \tilde{u}_k$, $y_k = \tilde{y}_k$), the expected value of r_k converges asymptotically to a neighborhood of zero, i.e., $\lim_{k\to\infty}\mathbb{E}\{r_k\} \in \mathcal{B}_{\delta_r}$, with $\delta_r \in \mathbb{R}^+$ and $\mathcal{B}_{\delta_r} \triangleq \{r \in \mathbb{R}^{n_r} :\ \|r\|_p \leq \delta_r\}$;
2. the residue is sensitive to the anomalies ($f_k \not\equiv 0$).

The characterization of $\mathcal{B}_{\delta_r}$ depends on the noise terms and can be found in [3] for particular values of p. Given the residue signal over the time-interval $[d_0,\ d_f]$, $\mathbf{r}_{[d_0,\ d_f]}$, an alarm is triggered if

$$\mathbf{r}_{[d_0,\ d_f]} \notin \mathcal{U}_{[d_0,\ d_f]}, \tag{5}$$

where the set $\mathcal{U}_{[d_0,\ d_f]}$ is chosen so that the false-alarm rate does not exceed a given threshold $\alpha \in [0,\,1]$. This necessarily requires no alarm to be triggered in the noiseless nominal behavior i.e., $\mathbf{r}_{[d_0,\ d_f]} \in \mathcal{U}_{[d_0,\ d_f]}$ if for all $k \in [d_0,\ d_f]$ it holds that $r_k \in \mathcal{B}_{\delta_r}$. For instance, one can take $\mathcal{U}_{[d_0,\ d_f]}$ to be a bound on the energy of the residue signal over the time-interval $[d_0,\ d_f]$, resulting in $\mathcal{U}_{[d_0,\ d_f]} = \{\mathbf{r} : \|\mathbf{r}\|_{\ell_2[d_0,\ d_f]} \leq \delta\}$.

3 Adversary Model

In this section we discuss the adversary models composed of adversarial goals and the system dynamics under attack. In particular, we consider attack scenarios where the adversary's goal is to drive the system to an unsafe state while remaining stealthy. Below we describe the networked control system under attack with respect to the attack vector $a_k \in \mathbb{R}^{q_a}$.

3.1 Networked Control System under Attack

The system components under attack are now characterized for the attack vector a_k. Considering the plant and controller states to be augmented as $\eta_k =$

$[x_k^\top \quad z_k^\top]^\top$, the dynamics of the closed-loop system composed by $\mathcal{P}$ and $\mathcal{F}$ under the effect of a_k can be written as

$$\begin{aligned}\eta_{k+1} &= \mathbf{A}\eta_k + \mathbf{B}a_k + \mathbf{G}\begin{bmatrix} w_k \\ v_k \end{bmatrix} \\ \tilde{y}_k &= \mathbf{C}\eta_k + \mathbf{D}a_k + \mathbf{H}\begin{bmatrix} w_k \\ v_k \end{bmatrix},\end{aligned} \tag{6}$$

where the system matrices are

$$\begin{aligned}\mathbf{A} &= \begin{bmatrix} A + BD_cC & BC_c \\ B_cC & A_c \end{bmatrix}, & \mathbf{G} &= \begin{bmatrix} G & BD_c \\ 0 & B_c \end{bmatrix}, \\ \mathbf{C} &= \begin{bmatrix} C & 0 \end{bmatrix}, & \mathbf{H} &= \begin{bmatrix} 0 & I \end{bmatrix},\end{aligned}$$

and $\mathbf{B}$ and $\mathbf{D}$ capture how the attack vector a_k affects the plant and controller.

Similarly, using $\mathcal{P}$ and $\mathcal{D}$ as in (1) and (4), respectively, the anomaly detector error dynamics under attack are described by

$$\begin{aligned}\xi_{k+1|k} &= \mathbf{A}_e\xi_{k|k-1} + \mathbf{B}_e a_k + \mathbf{G}_e\begin{bmatrix} w_k \\ v_k \end{bmatrix} \\ r_k &= \mathbf{C}_e\xi_{k|k-1} + \mathbf{D}_e a_k + \mathbf{H}_e\begin{bmatrix} w_k \\ v_k \end{bmatrix},\end{aligned} \tag{7}$$

where $\xi_{k|k-1} \in \mathbb{R}^n$ is the estimation error and

$$\begin{aligned}\mathbf{A}_e &= A(I - KC), & \mathbf{G}_e &= \begin{bmatrix} G & -AK \end{bmatrix}, \\ \mathbf{C}_e &= VC(I - KC), & \mathbf{H}_e &= \begin{bmatrix} 0 & V(I - CK) \end{bmatrix}.\end{aligned}$$

The matrices $\mathbf{B}_e$ and $\mathbf{D}_e$ are specific to the available disruptive resources and are characterized below for data deception attacks.

Data Deception Resources. The deception attacks modify the control actions u_k and sensor measurements y_k from their calculated or real values to the corrupted signals $\tilde{u}_k$ and $\tilde{y}_k$, respectively. Denoting $\mathcal{R}^u_I \subseteq \{1, \ldots, n_u\}$ and $\mathcal{R}^y_I \subseteq \{1, \ldots, n_y\}$ as the deception resources, i.e. set of actuator and sensor channels that can be affected, and $|\mathcal{R}^u_I|$ and $|\mathcal{R}^u_I|$ as the respective cardinality of the sets, the deception attacks are modeled as

$$\tilde{u}_k \triangleq u_k + \Gamma^u b^u_k, \quad \tilde{y}_k \triangleq y_k + \Gamma^y b^y_k, \tag{8}$$

where the signals $b^u_k \in \mathbb{R}^{|\mathcal{R}^u_I|}$ and $b^y_k \in \mathbb{R}^{|\mathcal{R}^y_I|}$ represent the data corruption and $\Gamma^u \in \mathbb{B}^{n_u \times |\mathcal{R}^u_I|}$ and $\Gamma^y \in \mathbb{B}^{n_y \times |\mathcal{R}^y_I|}$ ($\mathbb{B} \triangleq \{0, 1\}$) are the binary incidence matrices mapping the data corruption to the respective data channels. The matrices Γ^u and Γ^y indicate which data channels can be accessed by the adversary and are therefore directly related to the adversary resources in deception attacks. Recalling that $a \in \mathbb{R}^{q_a}$, the number of data channels that may be compromised by the adversary are given by $q_a = |\mathcal{R}_I^u| + |\mathcal{R}_I^y|$. Defining $a_k = [b_k^{u\top} \quad b_k^{y\top}]^\top$, the system dynamics are given by (6) and (7) with

$$\mathbf{B} = \begin{bmatrix} B\Gamma^u & BD_c\Gamma^y \\ 0 & B_c\Gamma^y \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 0 \; \Gamma^y \end{bmatrix},$$

$$\mathbf{B}_e = \begin{bmatrix} B\Gamma^u & -AK\Gamma^y \end{bmatrix}, \quad \mathbf{D}_e = \begin{bmatrix} 0 & V(I-CK)\Gamma^y \end{bmatrix}.$$

3.2 Attack Goals and Constraints

In addition to the attack resources, the attack scenarios need to include the adversary's intent, namely the attack goals and constraints shaping the attack policy. The attack goals can be stated in terms of the attack impact on the system operation, while the constraints may be related to the attack detectability.

Several physical systems have tight operating constraints which if not satisfied might result in physical damage to the system. In this work we use the concept of safe sets to characterize the safety constraints.

Definition 1. *For a given time-interval $[k_0,\ k_f]$, the system is said to be safe if $\mathbf{x}_{[k_0,\ k_f]} \in \mathcal{S}_{[k_0,\ k_f]}$, where $\mathcal{S}_{[k_0,\ k_f]}$ is a compact set with non-empty interior.*

The above definition of safe set $\mathcal{S}_{[k_0,\ k_f]}$ allows one to consider both time-interval and time-instant characterizations of safe regions, for instance signal energy and safe regions of the state space, respectively.

Assumption 1. *The system is in a safe state at the beginning of the attack, i.e. $\mathbf{x}_{(-\infty,\ k_0-1]} \in \mathcal{S}_{(-\infty,\ k_0-1]}$.*

The physical impact of an attack can be evaluated by assessing whether or not the state of the system remained in the safe set during and after the attack. The attack is considered successful if the state is driven out of the safe set. For simplicity of notation, the safe set $\mathcal{S}_{[k_0,\ k_f]}$ will be simply denoted as $\mathcal{S}$ whenever the time-interval is not ambiguous. Moreover, the safe sets considered in the remaining of this paper are of the form $\mathcal{S}^p_{[k_0,\ k_f]} = \{\mathbf{x} :\ \|\mathbf{x}\|_{\ell_p[k_0,\ k_f]} \leq 1\}$.

Regarding the attack constraints, we consider that attacks are constrained to remain stealthy. Furthermore, we consider the disruptive attack component consists of only physical and data deception attacks, and thus we have the attack vector $a_k = [b_k^{u\top} \quad b_k^{y\top}]^\top$. Given the anomaly detector described in Section 2, denoting $\mathbf{a}_{[k_0,\ k_f]} = \{a_{k_0}, \ldots, a_{k_f}\}$ as the attack signal, and recalling that the residue signal $\mathbf{r}_{[k_0,\ +\infty)}$ is a function of the attack signal, the set of stealthy attacks are defined as follows.

Definition 2. *The attack signal $\mathbf{a}_{[k_0,\ k_f]}$ is stealthy over the time-interval $[k_0,\ d_f]$ with $d_f \geq k_f$ if $\mathbf{r}_{[k_0,\ d_f]} \in \mathcal{U}_{[k_0,\ d_f]}$.*

Note that the above definition is dependent on the initial state of the system at k_0, as well as the noise terms w_k and v_k. Furthermore, it also requires the attack to be stealthy even after it has been performed, as $d_f \geq k_f$.

Since the closed-loop system (6) and the anomaly detector (7) under linear attack policies are linear systems, each of these systems can be separated into two components, the nominal component with $a_k = 0 \ \forall k$ and the following systems with zero initial conditions $\eta_0^a = \xi_{0|0}^a = 0$

$$\begin{aligned} \eta_{k+1}^a &= \mathbf{A}\eta_k^a + \mathbf{B}a_k \\ \tilde{y}_k^a &= \mathbf{C}\eta_k^a + \mathbf{D}a_k, \end{aligned} \tag{9}$$

$$\begin{aligned} \xi_{k|k}^a &= \mathbf{A}_e\xi_{k-1|k-1}^a + \mathbf{B}_e a_{k-1} \\ r_k^a &= \mathbf{C}_e\xi_{k-1|k-1}^a + \mathbf{D}_e a_{k-1}. \end{aligned} \tag{10}$$

Assuming the system is behaving nominally before the attack and that, given the linearity of (7), there exists a set $\mathcal{U}_{[k_0,\, d_f]}^a \triangleq \{\mathbf{r} : \ \|\mathbf{r}\|_{\ell_p[k_0,\, d_f]} \leq \delta_\alpha\}$ such that $\mathbf{r}_{[k_0,\, d_f]}^a \in \mathcal{U}_{[k_0,\, d_f]}^a \Rightarrow \mathbf{r}_{[k_0,\, d_f]} \in \mathcal{U}_{[k_0,\, d_f]}$, we have the following definition:

Definition 3. *The attack signal* $\mathbf{a}_{[k_0,\, k_f]}$ *is stealthy over the time-interval* $[k_0,\ d_f]$ *if* $\mathbf{r}_{[k_0,\, d_f]}^a \in \mathcal{U}_{[k_0,\, d_f]}^a$.

Albeit more conservative than Definition 2, this definition only depends on the attack signals $\mathbf{a}_{[k_0,\, k_f]}$. Similarly, the impact of attacks on the closed-loop system can also be analyzed by looking at the linear system (9).

4 Quantifying Cyber-Security: Transient Analysis

As mentioned in Section 3.2, the adversary aims at driving the system to an unsafe state while remaining stealthy. Additionally we consider that the adversary also has resource constraints, in the sense that only a small number of attack points to the system are available. In the following, several formulations for quantifying cyber-security of networked control systems are discussed.

Consider the dynamical system in (9) and the time-interval $[0,\ N]$ with $d_0 = k_0 = 0$ and $k_f = d_f = N$. Defining $\mathbf{n} = [\eta_0^\top \ \dots \ \eta_N^\top]^\top$, $\mathbf{a} = [a_0^\top \ \dots \ a_N^\top]^\top$, and $\mathbf{y} = [y_0^\top \ \dots \ y_N^\top]^\top$, the state and output trajectories can be described by the following mappings

$$\begin{aligned} \mathbf{n} &= \mathcal{O}_\eta\eta_0 + \mathcal{T}_\eta\mathbf{a} \\ \mathbf{y} &= \mathcal{C}_\eta\mathbf{n} + \mathcal{D}_\eta\mathbf{a}, \end{aligned} \tag{11}$$

where

$$\mathcal{O}_\eta = \begin{bmatrix} I \\ \mathbf{A} \\ \mathbf{A}^2 \\ \vdots \\ \mathbf{A}^N \end{bmatrix}, \quad \mathcal{T}_\eta = \begin{bmatrix} \mathbf{D} & 0 & \dots & 0 \\ \mathbf{B} & 0 & \dots & 0 \\ \mathbf{AB} & \mathbf{B} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \mathbf{A}^{N-1}\mathbf{B} & \mathbf{A}^{N-2}\mathbf{B} & \dots & \mathbf{B} \end{bmatrix}, \tag{12}$$
$$\mathcal{C}_\eta = I_{N+1} \otimes \mathbf{C}, \quad \mathcal{D}_\eta = I_{N+1} \otimes \mathbf{D}$$

Similarly for (10), defining $\mathbf{e} = [\xi_{-1|-1}^\top \ \cdots \ \xi_{N-1|N-1}^\top]^\top$, $\mathbf{r} = [r_0^\top \ \dots \ r_N^\top]^\top$ yields

$$\begin{aligned} \mathbf{e} &= \mathcal{O}_\xi\xi_{-1|-1} + \mathcal{T}_\xi\mathbf{a} \\ \mathbf{r} &= \mathcal{C}_\xi\mathbf{e} + \mathcal{D}_\xi\mathbf{a}. \end{aligned} \tag{13}$$

Recall that the system is operating safely during the time-interval $[k_0,\ k_f]$ if $\mathbf{x} \in \mathcal{S}_{[k_0,\ k_f]}$. Supposing $\mathcal{S}^p_{[k_0,\ k_f]} = \{\mathbf{x} :\ \|\mathbf{x}\|_{\ell_p[k_0,\ k_f]} \leq 1\}$ for $p \geq 1$, the system is safe during the time-interval $\{0, 1, \dots, N\}$ if

$$\mathbf{x} \triangleq \mathcal{C}_x \mathbf{n} \in \mathcal{S}^p_{[0,\ N]}, \tag{14}$$

where $\mathcal{C}_x = I_{N+1} \otimes [I_n\ 0]$. In particular, for $p = \infty$ we have that the system is safe if $\|\mathbf{x}\|_\infty = \|\mathcal{C}_x \mathbf{n}\|_\infty \leq 1$.

4.1 Maximum-Impact Attacks

One possible way to quantify cyber-security is by analyzing the impact of attacks on the control system, given some pre-defined resources available to the adversary. Recalling the safe set introduced earlier, $\mathcal{S}^p_{[0,\ N]} = \{\mathbf{x} :\ \|\mathbf{x}\|_{\ell_p[0,\ N]} \leq 1\}$, the attack impact during the time-interval $[0,\ N]$ is characterized by

$$g_p(\mathbf{n}) = \begin{cases} \|\mathcal{C}_x \mathbf{n}\|_p & \text{, if } \ \mathcal{C}_x \mathbf{n} \in \mathcal{S}^p_{[0,\ N]} \\ \quad +\infty & \text{, otherwise,} \end{cases} \tag{15}$$

since the adversary aims at driving the system to an unsafe state. Similarly, recall the set of stealthy attacks $\mathbf{a}$ such that $\mathbf{r} \in \mathcal{U}^a_{[k_0,\ d_f]} \triangleq \{\mathbf{r} :\ \|\mathbf{r}\|_{\ell_p[k_0,\ d_f]} \leq \delta_\alpha\}$.

The attack yielding the maximum impact can be computed by solving

$$\begin{aligned} \max_{\mathbf{a}} \quad & g_p(\mathbf{n}) \\ \text{s.t.} \quad & \|\mathcal{C}_\xi \mathbf{e} + \mathcal{D}_\xi \mathbf{a}\|_q \leq \delta_\alpha, \\ & \mathbf{e} = \mathcal{O}_\xi \xi_{-1|-1} + \mathcal{T}_\xi \mathbf{a}, \\ & \mathbf{n} = \mathcal{O}_\eta \eta_0 + \mathcal{T}_\eta \mathbf{a}, \end{aligned} \tag{16}$$

with p and q possibly different. Given the objective function $g_p(\mathbf{n})$, the adversary's optimal policy is to drive the system to an unsafe state while keeping the residue below the threshold. When the unsafe state is not reachable while remaining stealthy, the optimal attack drives the system as close to the unsafe set as possible by maximizing $\|\mathbf{x}\|_{\ell_p[0,\ N]} = \|\mathcal{C}_x \mathbf{n}\|_p$.

Letting $\xi_{-1|-1} = 0$ and $\eta_0 = 0$, the optimal values of (16) can be characterized by analyzing the following modified problem

$$\begin{aligned} \max_{\mathbf{a}} \quad & \|\mathcal{T}_x \mathbf{a}\|_p \\ \text{s.t.} \quad & \|\mathcal{T}_r \mathbf{a}\|_q \leq \delta_\alpha, \end{aligned} \tag{17}$$

where $\mathcal{T}_x = \mathcal{C}_x \mathcal{T}_\eta$ and $\mathcal{T}_r = \mathcal{C}_\xi \mathcal{T}_\xi + \mathcal{D}_\xi$. The conditions under which (17) admits bounded optimal values are characterized in the following result.

Lemma 1. *The problem* (17) *is bounded if and only if* $\ker(\mathcal{T}_r) \subseteq \ker(\mathcal{T}_x)$.

Proof. Suppose that $\ker(\mathcal{T}_r) \neq \emptyset$ and consider the subset of solutions where $\mathbf{a} \in \ker(\mathcal{T}_r)$. For this subset of solutions, the optimization problem then becomes $\max_{\mathbf{a}\in\ker(\mathcal{T}_r)} \|\mathcal{T}_x\mathbf{a}\|_p$. Since the latter corresponds to a maximization of a convex function, its solution is unbounded unless $\mathcal{T}_x\mathbf{a} = 0$ for all $\mathbf{a} \in \ker(\mathcal{T}_r)$ i.e., $\ker(\mathcal{T}_r) \subseteq \ker(\mathcal{T}_x)$. For $\mathbf{a} \notin \ker(\mathcal{T}_r)$ the feasible set is compact and thus the objective function over the feasible set is bounded, which concludes the proof.

Supposing that the optimization problem (17) is bounded and $p = q = 2$, (17) can be rewritten as a generalized eigenvalue problem and solved analytically.

Theorem 1. *Let $p = q = 2$ and suppose that $\ker(\mathcal{T}_r) \subseteq \ker(\mathcal{T}_x)$. The optimal attack policy for (17) is given by*

$$\mathbf{a}^\star = \frac{\delta_\alpha}{\|\mathcal{T}_r\mathbf{v}^\star\|_2}\mathbf{v}^\star, \tag{18}$$

where $\mathbf{v}^\star$ is the eigenvector associated with λ^*, the largest generalized eigenvalue of the matrix pencil $\left(\mathcal{T}_x^\top\mathcal{T}_x,\ \mathcal{T}_r^\top\mathcal{T}_r\right)$. Moreover, the corresponding optimal value is given by $\|\mathcal{T}_x\mathbf{a}^\star\|_2 = \sqrt{\lambda^*}\delta_\alpha$.

Proof. The proof is similar to that of [17, Thm. 12].

Given the solution to (17) characterized by the previous result, the maximum impact with respect to (16) is given by

$$g_p(\mathcal{T}_x\mathbf{a}^\star) = \begin{cases} \sqrt{\lambda^*}\delta_\alpha & \text{, if } \sqrt{\lambda^*}\delta_\alpha \leq 1 \\ +\infty & \text{, otherwise.} \end{cases}$$

4.2 Minimum-Resource Attacks

Cyber-security of control systems can also be quantified by assessing the number of resources needed by the adversary to perform a given set of attacks, without necessarily taking into account the attack impact, as formulated below.

Consider the set of attacks $\mathcal{G}$ such that $\mathbf{a} \in \mathcal{G}$ satisfies the goals of a given attack scenario. Recall that $a_k \in \mathbb{R}^{q_a}$ for all $k \in [k_0,\ k_f]$ and denote $\mathbf{a}_{(i),\,[k_0,\,k_f]} = \{a_{(i),k_0},\ \ldots,\ a_{(i),k_f}\}$ as the signal corresponding to the $i-$th attack resource. Consider the function

$$h_p(\mathbf{a}) = [\|\mathbf{a}_{(1)}\|_{\ell_p}\ \ldots\ \|\mathbf{a}_{(q_a)}\|_{\ell_p}]^\top \tag{19}$$

with $1 \leq p \leq +\infty$. The number of resources employed in a given attack are $\|h_p(\mathbf{a})\|_0$. For the set of attacks $\mathcal{G}$, the minimum-resource attacks are computed by solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{a}} \quad & \|h_p(\mathbf{a})\|_0 \\ \text{s.t.} \quad & \|\mathcal{C}_\xi\mathbf{e} + \mathcal{D}_\xi\mathbf{a}\|_q \leq \delta_\alpha, \\ & \mathbf{e} = \mathcal{O}_\xi\xi_{-1|-1} + \mathcal{T}_\xi\mathbf{a}, \\ & \mathbf{a} \in \mathcal{G}. \end{aligned} \tag{20}$$

Although the set $\mathcal{G}$ may be chosen depending on the attack impact $g_p(\mathbf{n})$, i.e., $\mathcal{G} = \{\mathbf{a} : \ g_p(\mathbf{n}) = \|\mathcal{T}_x\mathbf{a}\|_{\ell_p} > \gamma\}$, this generally results in non-convex constraints that increase the computational complexity of the problem. As an example, the set $\mathcal{G} = \{\mathbf{a} : \ \|\mathcal{T}_x\mathbf{a}\|_{\ell_\infty} > \gamma\}$ is formulated as a set of linear constraints with binary variables in (35). However, $\mathcal{G}$ might not be directly related to the impact of the attack in terms of $g_p(\mathbf{n})$. For instance, the formulation (20) captures the security-index proposed for static systems in [13], where the adversary aims at corrupting a given measurement i without being detected. The security-index formulation is retrieved by having $\xi_{-1|-1} = 0$, $N = 0$, $\delta_\alpha = 0$, and $\mathcal{G} = \{\mathbf{a} \in \mathbb{R}^{q_a} : \ \mathbf{a}_{(i)} = 1\}$. However, for dynamic systems when $N > 0$, the specification of the attack scenario and corresponding set of attacks $\mathcal{G}$ is more involved. The same scenario where the adversary aims at corrupting a given channel i can be formulated by having $\delta_\alpha = 0$ and $\mathcal{G} = \{\mathbf{a} : \ \|\mathbf{a}_{(i)}\|_{\ell_p} = \epsilon\}$. However, for positive values of δ_α the feasibility of the problem depends on both δ_α and ϵ, which need to be carefully chosen.

The optimization problem (20) is also related to the estimation of sparse unknown-input signals. For instance, the ℓ_1/ℓ_q decoder proposed in [5] relaxes (20) using a ℓ_1/ℓ_q-norm regularization [8] and can be obtained by having $p \geq 1$ and solving the following modified problem for given $\mathbf{r}$

$$\begin{aligned} \min_{\mathbf{a},\ \xi_{-1|-1}} \quad & \|h_p(\mathbf{a})\|_1 \\ \text{s.t.} \quad & \mathbf{r} = \mathcal{C}_\xi\mathbf{e} + \mathcal{D}_\xi\mathbf{a}, \\ & \mathbf{e} = \mathcal{O}_\xi\xi_{-1|-1} + \mathcal{T}_\xi\mathbf{a} \end{aligned}$$

The optimal solution $\mathbf{a}^\star$ and $\xi^\star_{-1|-1}$ can then be used reconstruct the state trajectory according to (9). However, note that using $\|h_p(\mathbf{a})\|_1$ as the objective function instead of $\|h_p(\mathbf{a})\|_0$ may lead to substantially different solutions, since $\|h_p(\mathbf{a})\|_1$ mixes the time and physical dimensions of the attack signal. In fact, letting $p = 1$ so that $\|h_p(\mathbf{a})\|_1 = \|\mathbf{a}\|_1$ and supposing the number of available channels is given by $q_a = 2$ and $N = 1$, having $\mathbf{a} = [\mathbf{a}_{(1)}^\top \ \mathbf{a}_{(2)}^\top]^\top = [1\ 0\ 0\ 1]$ leads to $\|h_p(\mathbf{a})\|_1 = \|h_p(\mathbf{a})\|_0 = 2$ and corrupts two channels, while $\mathbf{a} = [1\ 1\ 0\ 0]^\top$ yields $\|h_p(\mathbf{a})\|_1 = 2$ and $\|h_p(\mathbf{a})\|_0 = 1$, thus corrupting only one channel. These attacks are significantly different in terms of adversarial resources, as corrupting two channels requires much larger effort than corrupting only one.

4.3 Maximum-Impact Minimum-Resource Attacks

The previous formulations considered impact and resources independently when quantifying cyber-security. Here the impact and resources and addressed simultaneously by considering the multi-objective optimization problem

$$\begin{aligned} \max_{\mathbf{a}} \quad & [g_p(\mathbf{n}),\ -\|h_p(\mathbf{a})\|_0]^\top \\ \text{s.t.} \quad & \|\mathcal{C}_\xi\mathbf{e} + \mathcal{D}_\xi\mathbf{a}\|_q \leq \delta_\alpha, \\ & \mathbf{e} = \mathcal{O}_\xi\xi_{-1|-1} + \mathcal{T}_\xi\mathbf{a}, \\ & \mathbf{n} = \mathcal{O}_\eta\eta_0 + \mathcal{T}_\eta\mathbf{a}. \end{aligned} \tag{21}$$

The vector-valued objective function indicates that the adversary desires to simultaneously maximize and minimize $g_p(\mathbf{n})$ and $\|h_p(\mathbf{a})\|_0$, respectively. Solutions to multi-objective problems are related to the concept of Pareto optimality [9] and correspond to the optimal trade-off manifold between the several objectives. These solutions can be obtained through several techniques, for instance the bounded objective function method in which all but one of the objectives are posed as constraints, thus obtaining a scalar-valued objective function. Applying this method to (21) and constraining $\|h_p(\mathbf{a})\|_0$ yields

$$\begin{aligned} \max_{\mathbf{a}} \quad & g_p(\mathbf{n}) \\ \text{s.t.} \quad & \left\|\mathcal{C}_\xi \mathbf{e} + \mathcal{D}_\xi \mathbf{a}\right\|_q \leq \delta_\alpha, \\ & \mathbf{e} = \mathcal{O}_\xi \xi_{-1|-1} + \mathcal{T}_\xi \mathbf{a}, \\ & \mathbf{n} = \mathcal{O}_\eta \eta_0 + \mathcal{T}_\eta \mathbf{a}, \\ & \|h_p(\mathbf{a})\|_0 < \epsilon, \end{aligned} \tag{22}$$

which can be interpreted as a maximum-impact resource-constrained attack policy. The Pareto frontier that characterizes the optimal trade-off manifold can be obtained by iteratively solving (22) for $\epsilon \in \{1, \ldots, q_a\}$. This approach is illustrated in Section 6 for the quadruple-tank process.

5 Quantifying Cyber-Security: Steady-State Analysis

Here we consider the steady-state of the system under attack. Let $z \in \mathbb{C}$ and define

$$\begin{aligned} G_{xa}(z) &= [I_n \; 0](zI - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}, \\ G_{ra}(z) &= C_e(zI - A_e)^{-1}B_e + D_e, \end{aligned} \tag{23}$$

which correspond to the transfer functions from a_k to x_k and r_k respectively. Considering exponential attack signals of the form $a_k = gz^k$ for fixed z, denote $a(z) = g \in \mathbb{C}^{q_a}$, $x(z) = G_{xa}(z)a(z)$, and $r(z) = G_{ra}(z)a(z)$ as the phasor notation of a_k, x_k, and r_k, respectively. Since the analysis in this section is restricted to steady-state, we consider z to be on the unit circle, $z \in \mathbb{S}$, and thus $a(z)$ corresponds to sinusoidal signals of constant magnitude. Defining the frequency-domain safe set as $\mathcal{S}^p_\infty = \{x \in \mathbb{C}^n : \;\|x\|_p \leq 1\}$, the system under attack is said to be safe at steady-state if $x(z) = G_{xa}(z)a(z) \in \mathcal{S}^p_\infty$.

5.1 Maximum-Impact Attacks

For a given $z \in \mathbb{S}$, the steady-state attack impact is characterized by

$$g_p(x(z)) = \begin{cases} \|x(z)\|_p \text{ , if } \; x(z) \in \mathcal{S}^p_\infty \\ \quad +\infty \qquad \text{, otherwise.} \end{cases} \tag{24}$$

Similarly, recall the set of steady-state stealthy attacks $a(z)$ such that $r(z) \in \mathcal{U}^a \triangleq \{r \in \mathbb{C}^{p_d} : \;\|r\|_p \leq \delta_\alpha\}$, where $r(z) = G_{ra}(z)a(z)$.

The attack yielding the maximum impact can be computed by solving

$$\begin{aligned} \sup_{z\in\mathbb{S}} \; \max_{a(z)} \quad & g_p(G_{xa}(z)a(z)) \\ \text{s.t.} \quad & \left\|G_{ra}(z)a(z)\right\|_p \leq \delta_\alpha. \end{aligned} \tag{25}$$

The maximum impact over all stealthy attacks can be computed by replacing the objective function $g_p(G_{xa}(z)A(z))$ with $\|G_{xa}(z)a(z)\|_p$, solving

$$\begin{aligned} \sup_{z\in\mathbb{S}} \; \max_{a(z)} \quad & \|G_{xa}(z)a(z)\|_p \\ \text{s.t.} \quad & \left\|G_{ra}(z)a(z)\right\|_q \leq \delta_\alpha, \end{aligned} \tag{26}$$

and evaluating $g_p(G_{xa}(z)a(z))$ for the obtained solution. The conditions under which (26) admits bounded optimal values are characterized as follows.

Lemma 2. *The optimization problem* (26) *is bounded if and only if* $\ker(G_{ra}(z)) \subseteq \ker(G_{xa}(z))$ *for all* $z \in \mathbb{S}$*.*

Proof. The proof follows the same reasoning as that of Lemma 1.

The previous statement is related to the concept of invariant-zeros of dynamical systems [18] as discussed below.

Definition 4. *Consider a linear time-invariant system in discrete-time with the state-space realization* (A, B, C, D) *and the equation*

$$\begin{bmatrix} zI - A & -B \\ C & D \end{bmatrix} \begin{bmatrix} x_0 \\ u_z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{27}$$

with $z \in \mathbb{C}$ *and* $x_0 \neq 0$*. For a given solution to the previous equation* (z, u_z, x_0)*, denote* z *as the invariant-zero,* u_z *as the input-zero direction, and* x_0 *as the state-zero direction.*

Lemma 3. *The optimization problem* (26) *is bounded if and only if either of the following hold:*

1. *the transfer function* $G_{ra}(z)$ *does not contain invariant-zeros on the unit circle;*
2. *all the invariant-zeros of the transfer function* $G_{ra}(z)$ *on the unit circle are also invariant-zeros of* $G_{xa}(z)$*, with the same input-zero direction.*

Proof. For the first statement, note that if $G_{ra}(z)$ does not contain invariant-zeros on the unit circle, then $\ker(G_{ra}(z)) = \emptyset$ for $z \in \mathbb{S}$ and thus (26) is bounded. As for the second statement, suppose that $G_{ra}(z)$ contains an invariant-zero $\bar{z} \in \mathbb{S}$ and recall that (A_e, B_e, C_e, D_e) is the state-space realization of $G_{ra}(z)$. Applying the Schur complement to (27) we see that, for a non-zero state-zero direction x_0, (27) can be rewritten as

$$\begin{aligned} &(\bar{z}I - A_e)x_0 - B_e u_z = 0, \\ &C_e x_0 + D_e u_z = 0. \end{aligned} \tag{28}$$

Since A_e is stable and $|\bar{z}| = 1$ we have that $\bar{z}I - A_e$ is invertible and thus (28) can be rewritten as $\left(C_e(\bar{z}I - A_e)^{-1}B_e + D_e\right) u_z \; = \; G_{ra}(\bar{z})u_z \; = \; 0$. Hence we conclude that the input-zero direction u_z lies in the null-space of $G_{ra}(\bar{z})$. In this case, applying Lemma 2 shows that the problem is bounded if and only if u_z also lies in the null-space of $G_{xa}(\bar{z})$, which concludes the proof.

Supposing that the optimization problem (26) is bounded and $p = 2$, (26) can be rewritten as a generalized eigenvalue problem and solved analytically.

Theorem 2. *Let $p = q = 2$ and suppose that $\ker(G_{ra}(z)) \subseteq \ker(G_{xa}(z))$ for all $z \in \mathbb{S}$. The optimal maximum-impact attack policy is given by*

$$a^\star(z^\star) = \frac{\delta_\alpha}{\|G_{ra}(z^\star)\mathbf{v}^\star\|_2}\mathbf{v}^\star, \tag{29}$$

where $\mathbf{v}^\star$ is the eigenvector associated with λ^*, the largest generalized eigenvalue of the matrix pencil $\left(G_{xa}^H(z)G_{xa}(z),\; G_{ra}^H(z)G_{ra}(z)\right)$ maximized over $z \in \mathbb{S}$. Moreover, the corresponding impact is given by $\|G_{xa}(z^\star)a^\star(z^\star)\|_2 = \sqrt{\lambda^*}\delta_\alpha$.

Proof. The proof is similar to that of [17, Thm. 12].

Given the solution to (26) characterized by the previous result, the maximum impact with respect to (25) is given by

$$g_p(G_{xa}(z^\star)a^\star(z^\star)) = \begin{cases} \sqrt{\lambda^*}\delta_\alpha & \text{, if } \sqrt{\lambda^*}\delta_\alpha \leq 1 \\ +\infty & \text{, otherwise.} \end{cases}$$

Theorem 3. *Supposing $G_{ra}(z)$ is left-invertible for all $z \in \mathbb{S}$, the largest generalized eigenvalue of the matrix pencil $\left(G_{xa}^H(z)G_{xa}(z),\; G_{ra}^H(z)G_{ra}(z)\right)$, $\lambda^\star(z^\star)$, maximized over $z^\star \in \mathbb{S}$ corresponds to the $\mathcal{H}_\infty$-norm of $G_{xa}(z)G_{ra}^\dagger(z)$ with $G_{ra}^\dagger(z) = \left(G_{ra}^H(z)G_{ra}(z)\right)^{-1} G_{ra}^H(z)$.*

Proof. First observe that $\ker\left(G_{ra}(z)\right) = \emptyset$, since $G_{ra}(z)$ is left-invertible for all $z \in \mathbb{S}$. Letting $\delta_\alpha = 1$, from Theorem 2 we then have that

$$\lambda^\star(z^\star) = \sup_{z \in \mathbb{S}} \max_{a(z)\colon \|G_{ra}(z)a(z)\|_2 = 1} \|G_{xa}(z)a(z)\|_2.$$

The proof concludes by noting that, since $G_{ra}(z)$ is left-invertible and $G_{xa}(z)$ and $G_{ra}(z)$ are stable, we have $a(z) = G_{ra}^\dagger(z)b(z)$ for some $b(z) \in \mathbb{C}^{n_r}$ and so $\lambda^\star(z^\star)$ can be rewritten as

$$\lambda^\star(z^\star) = \sup_{z \in \mathbb{S}} \max_{b(z)\colon \|b(z)\|_2 = 1} \|G_{xa}(z)G_{ra}^\dagger(z)b(z)\|_2^2 \triangleq \|G_{xa}(z)G_{ra}^\dagger(z)\|_\infty.$$

5.2 Minimum-Resource Attacks

Consider the set of attacks $\mathcal{G}$ such that $a(z) \in \mathcal{G}$ satisfies the goals of a given attack scenario. For the set of attacks $\mathcal{G}$, the minimum-resource steady-state attacks are computed by solving the following optimization problem

$$\begin{aligned}\inf_{z\in\mathbb{S}}\ \min_{a(z)} \quad & \|a(z)\|_0\\ \text{s.t.} \quad & \|G_{ra}(z)a(z)\|_q \leq \delta_\alpha,\\ & a(z) \in \mathcal{G}.\end{aligned} \tag{30}$$

As in the security-index formulation for a given channel i [13], one can define $\mathcal{G} \triangleq \{a(z) \in \mathbb{C}^{q_a} :\ a_{(i)}(z) = 1\}$.

5.3 Maximum-Impact Minimum-Resource Attacks

Similarly as for the transient analysis, the impact and adversarial resources can be treated simultaneously in the multi-objective optimization problem

$$\begin{aligned}\sup_{z\in\mathbb{S}}\ \max_{a(z)} \quad & [g_p(G_{xa}(z)a(z)),\ -\|a(z)\|_0]^\top\\ \text{s.t.} \quad & \|G_{ra}(z)a(z)\|_q \leq \delta_\alpha.\end{aligned} \tag{31}$$

Using the bounded objective function method [9], the Pareto frontier can be obtained by iteratively solving the following problem for $\epsilon \in \{1,\ \ldots,\ q_a\}$

$$\begin{aligned}\sup_{z\in\mathbb{S}}\ \max_{a(z)} \quad & g_p(G_{xa}(z)a(z))\\ \text{s.t.} \quad & \|G_{ra}(z)a(z)\|_q \leq \delta_\alpha,\\ & \|a(z)\|_0 < \epsilon.\end{aligned} \tag{32}$$

6 Computational Algorithms and Examples

In this section the maximum-impact resource-constrained formulation proposed in the transient analysis with $p = \infty$ is formulated as a mixed integer linear programming problem. Numerical examples are also presented to illustrate some of the proposed formulations for quantifying cyber-security of control systems.

6.1 Mixed Integer Linear Programming

Consider the maximum-impact resource-constrained formulation from the transient analysis (22) reproduced below

$$\begin{aligned}\max_{\mathbf{a}} \quad & g_p(\mathbf{n})\\ \text{s.t.} \quad & \|\mathcal{C}_\xi\mathbf{e} + \mathcal{D}_\xi\mathbf{a}\|_p \leq \delta_\alpha,\\ & \|h_p(\mathbf{a})\|_0 \leq \epsilon,\\ & \mathbf{e} = \mathcal{O}_\xi\xi_{-1|-1} + \mathcal{T}_\xi\mathbf{a},\\ & \mathbf{n} = \mathcal{O}_\eta\eta_0 + \mathcal{T}_\eta\mathbf{a}.\end{aligned}$$

For $1 \leq p \leq \infty$, the constraint $\|h_p(\mathbf{a})\|_0 \leq \epsilon$ models the fact that the number of channels the adversary can compromise is upper bounded by epsilon. By introducing the binary decision variables $\mathbf{z}_i$, one for each channel, the constraint can be modeled as follows:

$$\begin{aligned} \mathbf{a}_{(i)} &\leq M_h \mathbf{z}_i \mathbf{1} && \forall\, i = 1,\, \ldots,\, q_a \\ -\mathbf{a}_{(i)} &\leq M_h \mathbf{z}_i \mathbf{1} && \forall\, i = 1,\, \ldots,\, q_a \\ \sum_{i=1}^{q_a} \mathbf{z}_i &\leq \epsilon \\ \mathbf{z}_i &\in \{0, 1\} && \forall\, i = 1,\, \ldots,\, q_a. \end{aligned} \tag{33}$$

In (33), $\mathbf{1}$ is a vector of ones of appropriate dimension. M_h is a given large number used to model "infinity". Its value is typically chosen according to the physical limitation of the system. The binary decision variables $\mathbf{z}_i$ serve to count the number of channels the adversary can compromise. That is, $\mathbf{z}_i = 1$ if and only if channel i can be compromised. Once a channel is compromised, the adversary is expected to be able to modify the time signal in that channel in any way he desires. This is modeled by the first two sets of constraints in (33).

In the constraint $\|C_\xi \mathbf{e} + D_\xi \mathbf{a}\|_p \leq \delta_\alpha$, the ℓ_p norm is chosen to be the ℓ_∞ norm modeling a constraint on the worst case output violation. This constraint can be modeled as

$$\begin{aligned} C_\xi \mathbf{e} + D_\xi \mathbf{a} &\leq \delta_\alpha \mathbf{1} \\ -C_\xi \mathbf{e} - D_\xi \mathbf{a} &\leq \delta_\alpha \mathbf{1}. \end{aligned} \tag{34}$$

In the objective function $g_p(n)$, the safety set $\mathcal{S}^p$ is chosen to be a ℓ_∞ norm ball. That is, $\mathcal{C}_x \mathbf{n} \in \mathcal{S}^p$ if and only if $\|\mathcal{C}_x \mathbf{n}\|_\infty \leq M_\mathcal{S}$ for some given safety tolerance $M_\mathcal{S}$. This is to model the fact that if any component of $\mathcal{C}_x \mathbf{n}$ is too large, then the system is considered to be unsafe. Consequently, the adversary's goal is to maximize $g_p(\mathbf{n})$ so that at least one component of $\mathcal{C}_x \mathbf{n}$ is larger than the safety tolerance $M_\mathcal{S}$. In hypograph form [1], maximizing $g_p(\mathbf{n})$ amounts to maximizing a slack variable γ with the additional constraint that $g_p(\mathbf{n}) \geq \gamma$. The later constraint can be modeled as

$$\begin{aligned} \mathcal{C}_x \mathbf{n} &\geq +\gamma \mathbf{1} - M_{\mathcal{C}_x} \left(\mathbf{1} - \mathbf{z}^+\right) \\ \mathcal{C}_x \mathbf{n} &\leq -\gamma \mathbf{1} + M_{\mathcal{C}_x} (\mathbf{1} - \mathbf{z}^-) \\ \mathbf{z}_i^+ + \mathbf{z}_i^- &\leq 1 && \forall\, i \\ \sum_i \left(\mathbf{z}_i^+ + \mathbf{z}_i^-\right) &\geq 1 \\ \mathbf{z}_i^+ &\in \{0, 1\} && \forall\, i \\ \mathbf{z}_i^- &\in \{0, 1\} && \forall\, i. \end{aligned} \tag{35}$$

In (35), $M_{\mathcal{C}_x}$ is another given large number used to represent "infinity". For each i, when the binary decision variable $\mathbf{z}_i^+ = 1$, the ith constraint of $\mathcal{C}_x \mathbf{n} \geq \gamma \mathbf{1} - M_{\mathcal{C}_x} \left(\mathbf{1} - \mathbf{z}^+\right)$ implies that the ith component of $\mathcal{C}_x \mathbf{n}$ is greater than or equal to γ. On the other hand, if $\mathbf{z}_i^+ = 0$ then this constraint component can be ignored. A similar interpretation holds for the combination of $\mathbf{z}^-$ and $\mathcal{C}_x \mathbf{n} \leq -\gamma \mathbf{1} + M_{\mathcal{C}_x} (\mathbf{1} - \mathbf{z}^-)$. Furthermore, the constraint $\mathbf{z}_i^+ + \mathbf{z}_i^- \leq 1$ models the fact that the ith component of $\mathcal{C}_x \mathbf{n}$ cannot be both greater than γ and less than $-\gamma$

when $\gamma > 0$. Together with the above discussion, the constraint $\sum_i \left(\mathbf{z}_i^+ + \mathbf{z}_i^-\right) \geq 1$ indicates that at least one component of $\mathcal{C}_x\mathbf{n}$ must be greater than or equal to γ in absolute value. Since the objective is to maximize γ, it holds that $\gamma = \|\mathcal{C}_x\mathbf{n}\|_\infty$ at optimality. Finally, to model the fact that once the goal $\|\mathcal{C}\mathbf{n}\|_\infty > M_\mathcal{S}$ is achieved the adversary no longer needs to maximize γ. An additional constraint

$$\gamma \leq M_\mathcal{S} \tag{36}$$

can be imposed.

In conclusion, the maximum-impact resource-constrained attack can be modeled by the following mixed integer linear program:

$$\begin{aligned} \max_{\mathbf{a},\gamma,\mathbf{z},\mathbf{z}^+,\mathbf{z}^-} \quad & \gamma \\ \text{s.t.} \quad & \mathbf{e} = \mathcal{O}_\xi \xi_{-1|-1} + \mathcal{T}_\xi \mathbf{a}, \\ & \mathbf{n} = \mathcal{O}_{\eta_0}\eta_0 + \mathcal{T}_\eta \mathbf{a}, \\ & (33), (34), (35), (36). \end{aligned} \tag{37}$$

6.2 Numerical Example

Next we illustrate some of the proposed formulations for the Quadruple-Tank Process (QTP) illustrated in Fig. 2. The plant model can be found in [7]

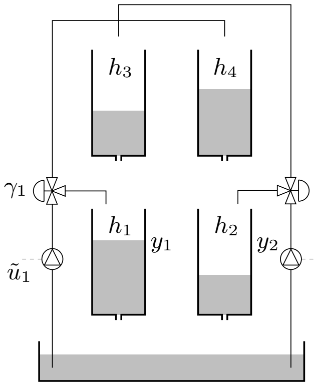


Fig. 2. Schematic of the Quadruple-Tank Process.

$$
\begin{aligned}
\dot{h}_1 &= -\frac{a_1}{A_1}\sqrt{2gh_1} + \frac{a_3}{A_1}\sqrt{2gh_3} + \frac{\gamma_1 k_1}{A_1}u_1, \\
\dot{h}_2 &= -\frac{a_2}{A_2}\sqrt{2gh_2} + \frac{a_4}{A_2}\sqrt{2gh_4} + \frac{\gamma_2 k_2}{A_2}u_2, \\
\dot{h}_3 &= -\frac{a_3}{A_3}\sqrt{2gh_3} + \frac{(1-\gamma_2)k_2}{A_3}u_2, \\
\dot{h}_4 &= -\frac{a_4}{A_4}\sqrt{2gh_4} + \frac{(1-\gamma_1)k_1}{A_4}u_1,
\end{aligned}
\tag{38}
$$

where h_i are the heights of water in each tank, A_i the cross-section area of the tanks, a_i the cross-section area of the outlet hole, k_i the pump constants, γ_i the flow ratios and g the gravity acceleration. The nonlinear plant model is linearized for a given operating point and sampled with a sampling period $T_s = 2\,s$. The QTP is controlled using a centralized LQG controller with integral action and a Kalman-filter-based anomaly detector is used so that alarms are triggered according to (5), for which we chose $\delta_\alpha = 0.25$ for illustration purposes.

For the time-interval $[0,\ 50]$, the maximum-impact minimum-resource attacks were computed for the process in minimum and non-minimum phase settings by iteratively solving (22) with $p = q = 2$. The respective impacts are presented in Table 1. As expected, the non-minimum phase system is less resilient than the minimum-phase one. In both settings the attack impact can be made arbitrarily large by corrupting 3 or more channels and thus the adversary can drive the state out of the safe set while remaining stealthy.

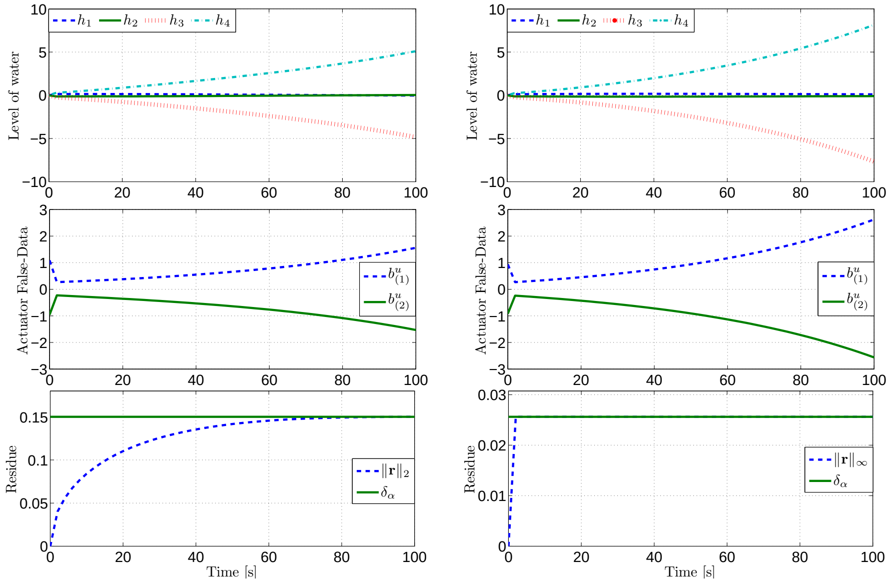
Table 1. Values of $\|\mathbf{x}\|_p$ for the maximum-impact formulation with $p = q = 2$ and $\delta_\alpha = 0.15$.

	$\|h_p(\mathbf{a})\|_0$			
	1	**2**	**3**	**4**
Minimum phase	1.15	140.39	∞	∞
Non-minimum phase	2.80	689.43	∞	∞

The maximum-impact attack signal for the non-minimum phase system with $\epsilon = 2$, $\delta_\alpha = 0.15$, and $p = q = 2$ is presented in Fig. 3(a). For the parameters $\epsilon = 2$, $\delta_\alpha = 0.025$, and $p = q = \infty$, the maximum-impact attack signal was computed using the mixed-integer linear programming problem (37) and is shown in Fig. 3(b). In both cases the optimal attack corrupts both actuator channels and ensures $\|\mathbf{r}\|_{\ell_p} \leq \delta_\alpha$.

7 Conclusions

Several formulations for quantifying cyber-security of networked control systems were proposed and formulated as constrained optimization problems, capturing trade-offs among adversary goals and constraints such as attack impact on the



(a) Parameters: $p = q = 2$, $\delta_\alpha = 0.15$.

(b) Parameters: $p = q = \infty$, $\delta_\alpha = 0.025$.

Fig. 3. Simulation results of the multi-objective problem (22) with $\epsilon = 2$ for the non-minimum phase system.

control system, attack detectability, and adversarial resources. Although the formulations are non-convex, some can be related to system theoretic concepts such as invariant-zeros and weighted $\mathcal{H}_\infty$ norm of the closed-loop system and thus may be solved efficiently. The maximum-impact resource-constrained attack policy was also formulated as a mixed-integer linear program for a particular choice of parameters. The results were illustrated for the quadruple-tank process.

Acknowledgments

This work was supported in part by the European Commission through the HYCON2 project, the Swedish Research Council under Grants 2007-6350 and 2009-4565, the Swedish Foundation for Strategic Research, and the Knut and Alice Wallenberg Foundation.

References

1. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
2. Cárdenas, A., Amin, S., Lin, Z., Huang, Y., Huang, C., Sastry, S.: Attacks against process control systems: risk assessment, detection, and response. In: Proceedings

of the 6th ACM Symposium on Information, Computer and Communications Security. pp. 355–366. ASIACCS '11, ACM, New York, NY, USA (2011)

3. Ding, S.X.: Model-based Fault Diagnosis Techniques: Design Schemes. Springer Verlag (2008)
4. Esfahani, P., Vrakopoulou, M., Margellos, K., Lygeros, J., Andersson, G.: Cyber attack in a two-area power system: Impact identification using reachability. In: American Control Conference, 2010. pp. 962–967 (Jul 2010)
5. Fawzi, H., Tabuada, P., Diggavi, S.: Security for control systems under sensor and actuator attacks. In: Proceedings of the 51st IEEE Conference on Decision and Control. Maui, Hawaii, USA (Dec 2012)
6. Hwang, I., Kim, S., Kim, Y., Seah, C.E.: A survey of fault detection, isolation, and reconfiguration methods. IEEE Transactions on Control Systems Technology 18(3), 636–653 (May 2010)
7. Johansson, K.: The quadruple-tank process: a multivariable laboratory process with an adjustable zero. IEEE Transactions on Control Systems Technology 8(3), 456–465 (May 2000)
8. Liu, J., Ye, J.: Efficient L1/Lq Norm Regularization. ArXiv e-prints (Sep 2010)
9. Marler, R.T., Arora, J.S.: Survey of multi-objective optimization methods for engineering. Structural and Multidisciplinary Optimization 26(6), 369–395 (Apr 2004)
10. Meserve, J.: Sources: Staged cyber attack reveals vulnerability in power grid. CNN (2007), available at `http://edition.cnn.com/2007/US/09/26/power.at.risk/index.html`.
11. Pasqualetti, F., Dorfler, F., Bullo, F.: Cyber-physical attacks in power networks: Models, fundamental limitations and monitor design. In: Proc. of the 50th IEEE Conf. on Decision and Control and European Control Conference. Orlando, FL, USA (Dec 2011)
12. Rid, T.: Cyber war will not take place. Journal of Strategic Studies 35(1), 5–32 (2011)
13. Sandberg, H., Teixeira, A., Johansson, K.H.: On security indices for state estimators in power networks. In: Preprints of the First Workshop on Secure Control Systems, CPSWEEK 2010. Stockholm, Sweden (Apr 2010)
14. Smith, R.: A decoupled feedback structure for covertly appropriating networked control systems. In: Proc. of the 18th IFAC World Congress. Milano, Italy (Aug–Sep 2011)
15. Sundaram, S., Hadjicostis, C.: Distributed function calculation via linear iterative strategies in the presence of malicious agents. Automatic Control, IEEE Transactions on 56(7), 1495–1508 (july 2011)
16. Symantec: Stuxnet introduces the first known rootkit for industrial control systems. Symantec (August 6th 2010), available at: `http://www.symantec.com/connect/blogs/stuxnet-introduces-first-known-rootkit-scada-devices`
17. Teixeira, A., Shames, I., Sandberg, H., Johansson, K.H.: A Secure Control Framework for Resource-Limited Adversaries. ArXiv e-prints (Dec 2012), submitted to Automatica.
18. Tokarzewski, J.: Finite zeros in discrete time control systems. Lecture notes in control and information sciences, Springer (2006)
19. U.S.-Canada PSOTF: Final report on the August 14th blackout in the United States and Canada. Tech. rep., U.S.-Canada Power System Outage Task Force (Apr 2004)
20. Zhou, K., Doyle, J.C., Glover, K.: Robust and Optimal Control. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1996)