Automatica 51 (2015) 135–148

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

A secure control framework for resource-limited adversaries☆

André Teixeira [a,1], Iman Shames [b], Henrik Sandberg [a], Karl Henrik Johansson [a]

[a] ACCESS Linnaeus Centre, KTH Royal Institute of Technology, School of Electrical Engineering, Stockholm, Sweden

[b] Department of Electrical and Electronic Engineering, University of Melbourne, Australia

ARTICLE INFO

Article history:
Received 14 November 2012
Received in revised form
18 August 2014
Accepted 27 August 2014
Available online 19 November 2014

Keywords:
Cyber–physical systems
System security
Attack space
Secure control systems

ABSTRACT

Cyber-secure networked control is modeled, analyzed, and experimentally illustrated in this paper. An attack space defined by the adversary's model knowledge, disclosure, and disruption resources is introduced. Adversaries constrained by these resources are modeled for a networked control system architecture. It is shown that attack scenarios corresponding to denial-of-service, replay, zero-dynamics, and bias injection attacks on linear time-invariant systems can be analyzed using this framework. Furthermore, the attack policy for each scenario is described and the attack's impact is characterized using the concept of safe sets. An experimental setup based on a quadruple-tank process controlled over a wireless network is used to illustrate the attack scenarios, their consequences, and potential countermeasures.

1. Introduction

Safe and reliable operation of infrastructures is of major societal importance. These systems need to be engineered in such a way so that they can be continuously monitored, coordinated, and controlled despite a variety of potential system disturbances. Given the strict operating requirements and system complexity, such systems are operated through IT infrastructures enabling the timely data flow between digital controllers, sensors, and actuators. However, the use of communication networks and heterogeneous IT components has made these cyber–physical systems vulnerable to cyber threats. One such example is the industrial systems and critical infrastructures operated through Supervisory Control and Data Acquisition (SCADA) systems. The measurement and control data in these systems are commonly transmitted through unprotected communication channels, leaving the system vulnerable to several threats (Giani, Sastry, Johansson, & Sandberg, 2009). As illustrative examples, we mention the cyber attacks on power transmission networks operated by SCADA systems reported in the public media (Gorman, 2009), and the Stuxnet malware that supposedly infected an industrial control system and disrupted its operation (Falliere, Murchu, & Chien, 2011; Rid, 2011).

There exists a vast literature on computer security focusing on three main properties of data and IT services, namely confidentiality, integrity, and availability (Bishop, 2002). Confidentiality relates to the non-disclosure of data by unauthorized parties. Integrity on the other hand concerns the trustworthiness of data, meaning there is no unauthorized change of the data contents or properties, while availability means that timely access to the data or system functionalities is ensured. Unlike other IT systems where cyber-security mainly involves the protection of data, cyber attacks on networked control systems may influence physical processes through feedback actuation. Therefore networked control system security needs to consider threats at both the cyber and physical layers. Furthermore, it is of the utmost importance in the study of cyber attacks on control systems to capture the adversary's resources and knowledge. Cyber threats can be captured and qualitatively categorized in the attack space illustrated in Fig. 1, which depicts several attack scenarios as points. Note that each example corresponds to a given instance of an attack scenario. As examples, eavesdropping and denial-of-service (DoS) attacks that do not have any model knowledge are indicated in the figure. However, some instances of DoS attacks may use additional resources and model knowledge, see Gupta, Langbort, and Başar (2010).

☆ The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 608224, the EIT-ICT Labs through the project SESSec-EU, the Swedish Research Council under Grants 2009-4565 and 2013-5523, and the Knut and Alice Wallenberg Foundation. The material in this paper was partially presented at the 1st International Conference on High Confidence Networked Systems (HiCoNS), April 17–18, 2012, Beijing, China. This paper was recommended for publication in revised form by Associate Editor Giancarlo Ferrari-Trecate under the direction of Editor Ian R. Petersen.

E-mail addresses: andretei@kth.se (A. Teixeira), iman.shames@unimelb.edu.au (I. Shames), hsan@kth.se (H. Sandberg), kallej@kth.se (K.H. Johansson).

[1] Tel.: +46 73 429 78 31; fax: +46 8 790 73 29.

http://dx.doi.org/10.1016/j.automatica.2014.10.067
0005-1098/

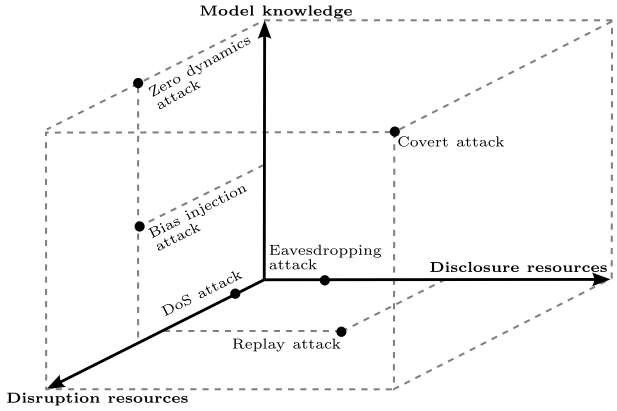


Fig. 1. The cyber–physical attack space. Each point depicts the qualitative classification of a given attack scenario.

We propose three dimensions for the attack space: the adversary's *a priori* system model knowledge, disclosure, and disruption resources. Although adversaries possess several other features, the proposed three dimensions are quite relevant from a control system's perspective and allow a straightforward categorization of many attack scenarios studied in the literature. The *a priori* model knowledge can be used by the adversary to construct more complex attacks, possibly harder to detect and with more severe consequences. Similarly, the disclosure resources enable the adversary to obtain sensitive information about the system during the attack by violating data confidentiality. Note that disclosure resources alone cannot disrupt the system operation. An example of an attack using only disclosure resources is the eavesdropping attack illustrated in Fig. 1. On the other hand, disruption resources can be used to affect the system operation, which happens for instance when data integrity or availability properties are violated. In particular this characterization fits the Stuxnet malware, which had resources to record and manipulate data in the SCADA network (Falliere et al., 2011). Moreover, the complexity and operation of Stuxnet also indicate that its developers had access to a reasonable amount of knowledge of both physical and cyber components of the target control system.

1.1. Security versus fault-tolerance

Control theory has contributed with frameworks to handle model uncertainties and disturbances as well as fault diagnosis and mitigation, see, for example, Chen and Patton (1999) and Hwang, Kim, Kim, and Seah (2010), Zhou, Doyle, and Glover (1996), respectively. In particular, on-line fault diagnosis uses real-time data of the system to monitor its behavior and detect faults in physical components, e.g. actuators and sensors. Once faults are detected, fault-tolerant techniques can be used to ensure graceful degradation of the system until the fault is repaired. Since cyber attacks on networked control systems also affect the physical behavior of the system, these tools can be used to detect and attenuate the consequences of cyber attacks, as has recently been done in the literature. However, there are substantial conceptual and technical differences between the fault-tolerant and secure control frameworks that motivate the need for specific theories and methodologies to address security issues in control systems.

Cyber attacks and faults have inherently distinct characteristics, which pose different challenges on the secure control and fault-tolerant approaches. Faults are considered as physical events that affect the system behavior, where simultaneous events are assumed to be non-colluding, i.e., the events do not act in a coordinated way. On the other hand, cyber attacks may be performed over a significant number of attack points in a coordinated fashion (Smith, 2011; Teixeira, Dán, Sandberg, & Johansson, 2011). Moreover, faults do not have an intent or objective to fulfill, as opposed to cyber attacks that do have a malicious intent. Later one attack scenario exploiting several of the previous issues is discussed in detail, namely the zero-dynamics attack.

The distinct characteristics of faults and attacks lead to quite different approaches for increasing the system resiliency. Increased resiliency may be achieved through mainly three actions: prevention, detection, and mitigation (Bishop, 2002; Isermann, 2006). These actions need to be tailored to the specific properties of faults and attacks to efficiently and effectively ensure resiliency. For instance, prevention, detection, and mitigation of faults may be achieved by maintenance, on-line monitoring, and timely repair of the physical components of the system, respectively. On the other hand, preventing, detecting, and mitigating cyber attacks on control systems must use mechanisms that consider both the cyber and physical realms, such as encryption and improved algorithms (Pang & Liu, 2012). Furthermore, ensuring security may involve addressing a large number of threats, thus requiring attack impact analysis and the use of risk assessment methods (Sridhar, Hahn, & Govindarasu, 2012). Several of these issues are present in this paper and have also been addressed in recent work on secure control systems.

1.2. Related work

Cyber attacks on control systems compromising measurement and actuator data integrity and availability have been considered in Cárdenas, Amin, and Sastry (2008), where the authors modeled the attack effects on the physical dynamics. Several attack scenarios have been simulated and evaluated on the Tennessee-Eastman process control system (Cárdenas et al., 2011) to study the attack impact and detectability. The attack scenarios in Cárdenas et al. (2011) are related to the ones considered in this paper, but we quantify the attack resources and policies in a systematic way.

Availability attacks have been analyzed in Amin, Cárdenas, and Sastry (2009) and Gupta et al. (2010) for resource-constrained adversaries with full-state information. Particularly, the authors considered DoS attacks in which the adversary could tamper with the communication channels and prevent measurement and actuator data from reaching their destination, rendering the data unavailable. A particular instance of the DoS attack in which the adversary does not have any *a priori* model knowledge, as the attack in Amin et al. (2009), is represented in the attack space in Fig. 1.

Deception attacks compromising integrity have recently received attention. In Pang and Liu (2012) the authors proposed an encryption and predictive control scheme to prevent and mitigate deception attacks on control systems. Replay attacks on the sensor measurements, which is a particular kind of deception attack, have been analyzed in Mo and Sinopoli (2009). The authors considered the case where all the existing sensors were attacked and suitable counter-measures to detect the attack were proposed. In this attack scenario the adversary does not have any model knowledge but is able to access and corrupt the sensor data, thus having disclosure and disruptive resources, as depicted in Fig. 1.

Another class of deception attacks, false-data injection attacks, has been studied in recent work. For instance, in the case of power networks, an adversary with perfect model knowledge has been considered in Liu, Reiter, and Ning (2009). The work in Kosut, Jia, Thomas, and Tong (2010) considered stealthy attacks with limited resources and proposed improved detection methods, while Sandberg, Teixeira, and Johansson (2010) analyzed the minimum number of sensors required for stealthy attacks. A corresponding measurement security metric for studying sets of vulnerable sensors was proposed in Sandberg et al. (2010). The

consequences of these attacks have also been analyzed in Teixeira et al. (2011), Teixeira, Sandberg, Dán, and Johansson (2012) and Xie, Mo, and Sinopoli (2010). In particular, in Teixeira et al. (2011) the authors analyzed attack policies with limited model knowledge and performed experiments on a power system control software, showing that such attacks are stealthy and can induce the erroneous belief that the system is at an unsafe state. The models used in the previous work are static, hence these attack scenarios are closest to the bias injection attack shown in Fig. 1.

Data injection attacks on dynamic control systems were also considered. In Smith (2011) the author characterizes the set of attack policies for covert (undetectable) false-data injection attacks with detailed model knowledge and full access to all sensor and actuator channels, while Pasqualetti, Dorfler, and Bullo (2011) described the set of undetectable false-data injection attacks for omniscient adversaries with full-state information, but possibly compromising only a subset of the existing sensors and actuators. In the context of multi-agent systems, optimal adversary policies for data injection using full model knowledge and state information were derived in Khanafer, Touri, and Başar (2012). The work in Sundaram, Revzen, and Pappas (2012) considers the detection and mitigation of false-data injection attacks on linear information dissemination algorithms over communication networks. In these attack scenarios confidentiality was violated, as the adversary had access to either measurement and actuator data or full-state information. These attacks are therefore placed close to the covert attack in Fig. 1.

Impact of false-data injection attacks has also been considered in the literature. Using dynamic nonlinear models of power networks, in Esfahani, Vrakopoulou, Margellos, Lygeros, and Andersson (2010) the authors use reachability methods to analyze the impact of a false-data injection attack on a two-area power network. For linear networked control systems under false-data injection attacks, Mo and Sinopoli (2012) propose methods to approximate the reachable set of states for stealthy adversaries.

Most of the recent work on cyber-security of control systems has considered scenarios where the adversary has access to a large set of resources and knowledge, thus being placed far from the origin of the attack space in Fig. 1. A large part of the attack space has not been explored yet. In particular, the class of detectable attacks that do not trigger conventional alarms has yet to be covered in depth.

1.3. Contributions and outline

In this paper we consider a typical networked control architecture under both cyber and physical attacks. A generic adversary model applicable to several attack scenarios is discussed and the attack resources are mapped to the corresponding dimensions of the attack space depicted in Fig. 1. Although the framework is presented for linear time-invariant (LTI) systems, the conceptual components and methodology may be applied to other classes of systems.

To illustrate the proposed framework, we consider a LTI system under several attack scenarios where the adversary's goal is to drive the system to an unsafe state while remaining stealthy. Exploiting the properties of LTI systems, for each scenario we formulate the corresponding stealthy attack policy and comment on the attack's performance. Furthermore, we describe the adversary's capabilities along each dimension of the attack space in Fig. 1, namely the disclosure resources, disruption resources, and model knowledge. Some of the attack scenarios analyzed in the paper have been staged on a wireless quadruple-tank testbed for security of control systems. The testbed architecture and results from the staged attacks are presented and discussed.

One of the attack scenarios analyzed corresponds to a novel type of detectable attack, the bias injection attack. Although this attack may be detected, it can drive the system to an unsafe region and it only requires limited model knowledge and no information about the system state. Stealthiness conditions for this attack are provided, as well as a methodology to assess the attack impact on the physical state of the system.

The material in this paper is an extension of the authors' preliminary work, see Teixeira, Pérez, Sandberg, and Johansson (2012). Particularly, in the current work the attack goals are formalized using the notion of safe regions of the state space and two additional attack scenarios are described and analyzed. Furthermore, the attack performance of each scenario is analyzed in more detail and additional results for the zero-dynamics and bias injection attacks are presented.

The outline of the paper is as follows. The system architecture and model are described in Section 2, while Section 3 contains the adversary model and a detailed description of the attack resources on each dimension of the attack space. The framework introduced in the previous sections is then illustrated for five particular attack scenarios in Section 4, supposing that the adversary aims at driving the system to an unsafe state while remaining stealthy. The attack policy, attack performance, and required model knowledge, disclosure, and disruption resources are described in detail for each attack scenario. The results of the experiments for four of the attack scenarios in a secure control systems testbed are presented and discussed in Section 5, followed by conclusions and future work directions in Section 6.

2. Networked control system

In this section we describe the networked control system structure, where we consider three main components: the physical plant and communication network, the feedback controller, and the anomaly detector. Although the networked control system architecture is presented for LTI systems, the same components can be used when considering other classes of systems.

2.1. Physical plant and communication network

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

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

where $x_k \in \mathbb{R}^{n_x}$ 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 $f_k \in \mathbb{R}^{n_f}$ is the unknown signal representing the effects of anomalies, usually denoted as fault signal in the fault diagnosis literature (Ding, 2008). The process and measurement noise, $w_k \in \mathbb{R}^{n_x}$ 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, it is assumed that, first, any possible mismatches between the transmitted and received data are due to malicious adversaries alone. Second, the communication network is assumed to be reliable and 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 a linear time-invariant feedback controller (Zhou et al., 1996). The output-feedback controller can be written 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 (Ding, 2008; Hwang et al., 2010). Here we consider a general form of an observer-based Fault Detection Filter

$$\mathcal{D}: \begin{cases} s_k = A_e s_k + B_e u_k + K_e \tilde{y}_k \\ r_k = C_e s_k + D_e u_k + E_e \tilde{y}_k, \end{cases} \tag{3}$$

where $s_k \in \mathbb{R}^{n_s}$ is the state of the anomaly detector and $r_k \in \mathbb{R}^{n_r}$ is the residue evaluated to detect and locate existing anomalies.

The anomaly detector is designed by choosing A_e, B_e, K_e, C_e, D_e, and E_e 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 $\mathcal{B}_{\delta_r} \triangleq \{r \in \mathbb{R}^{n_r} : \|r\|_q \leq \delta_r\}$ for $q \geq 1$;
(2) the residue is sensitive to the anomalies (i.e., $f_k \not\equiv 0$ and $f_k \equiv 0$ for all k result in different residues).

The characterization of $\mathcal{B}_{\delta_r}$ depends on the noise terms and can be found in Ding (2008) for particular values of q. Given the residue signal over the time-interval $[d_0,\ d_f]$, $\mathbf{r}_{[d_0,\ d_f]} = [r_{d_0}^\top \quad \cdots \quad r_{d_f}^\top]^\top$, an alarm is triggered if

$$\mathbf{r}_{[d_0,\ d_f]} \notin \mathcal{U}_{[d_0,\ d_f]}, \tag{4}$$

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 $r_k \in \mathcal{B}_{\delta_r}$. Such set-based detection fits several residual evaluation techniques presented in Frank and Ding (1997). 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 $\mathcal{U}_{[d_0,\ d_f]} = \{\mathbf{r}_{[d_0,\ d_f]} : \|\mathbf{r}_{[d_0,\ d_f]}\|_2 \leq \delta\}$ for some $\delta \in (0,\ \infty)$.

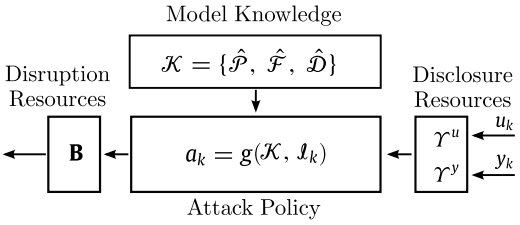


Fig. 2. Adversary model for a point in the attack space in Fig. 1.

3. Adversary models

The adversary model considered in this paper is illustrated in Fig. 2 and is composed of an attack policy and the adversary resources i.e., the system model knowledge, the disclosure resources, and the disruption resources. Each of the adversary resources can be mapped to a specific axis of the attack space in Fig. 1: $\mathcal{K} = \{\hat{\mathcal{P}},\ \hat{\mathcal{F}},\ \hat{\mathcal{D}}\}$ is the *a priori* model knowledge possessed by the adversary; $\mathcal{I}_k$ corresponds to the set of sensor and actuator data available to the adversary at time k as illustrated in (8), thus being mapped to the disclosure resources; a_k is the attack vector at time k that may affect the system behavior using the disruption resources captured by $\mathbf{B}$, as defined in the current section. The attack policy mapping $\mathcal{K}$ and $\mathcal{I}_k$ to a_k at time k is denoted as

$$a_k = g(\mathcal{K},\ \mathcal{I}_k). \tag{5}$$

Examples of attacks policies for different attack scenarios are given in Section 4.

In this section we describe the networked control system under attack with respect to the attack vector a_k. Then we detail the adversary's model knowledge, the disclosure resources, and the disruption resources. Models of the attack vector a_k for particular disruption resources are also given.

3.1. Networked control system under attack

The system components under attack are now characterized for the attack vector a_k, which also includes the fault vector f_k. Stacking the states of the plant and controller 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

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

where the system matrices are

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

and $\mathbf{B}$ and $\mathbf{D}$ capture the way in which the attack vector a_k affects the plant and controller. These matrices are characterized for some attack scenarios in Section 3.4. Similarly, using $\mathcal{P}$, $\mathcal{F}$, and $\mathcal{D}$ as in (1), (2), and (3), respectively, and stacking the states of the plant, controller, and anomaly detector as $\xi_k = [\eta_k^\top \quad s_k^\top]^\top$ the residue dynamics under attack are described by

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

where

$$\mathbf{A}_e = \begin{bmatrix} \mathbf{A} & 0 \\ B_e\mathbf{C}_u + K_e\mathbf{C} & A_e \end{bmatrix}, \qquad \mathbf{B}_e = \begin{bmatrix} \mathbf{B} \\ (B_eD_c + K_e)\mathbf{D} \end{bmatrix},$$

$$\mathbf{C}_e = \begin{bmatrix} D_e\mathbf{C}_u + E_e\mathbf{C} & C_e \end{bmatrix}, \qquad \mathbf{G}_e = \begin{bmatrix} \mathbf{G} \\ (B_eD_c + K_e)\mathbf{H} \end{bmatrix},$$

$$\mathbf{D}_e = (D_eD_c + E_e)\mathbf{D}, \qquad \mathbf{H}_e = (D_eD_c + E_e)\mathbf{H}.$$

3.2. Model knowledge

The amount of *a priori* knowledge regarding the control system is a core component of the adversary model, as it may be used, for instance, to render the attack undetectable. In general, we may consider that the adversary approximately knows the model of the plant ($\hat{\mathcal{P}}$) and the algorithms used in the feedback controller ($\hat{\mathcal{F}}$) and the anomaly detector ($\hat{\mathcal{D}}$), thus denoting the adversary knowledge by $\mathcal{K} = \{\hat{\mathcal{P}}, \hat{\mathcal{F}}, \hat{\mathcal{D}}\}$. Fig. 1 illustrates several types of attack scenarios with different levels of model knowledge. In particular, note that the replay attacks do not need any knowledge of the system components, therefore having $\mathcal{K} = \emptyset$, while the covert attack requires full knowledge about the system, hence $\mathcal{K} = \{\mathcal{P}, \mathcal{F}, \mathcal{D}\}$.

3.3. Disclosure resources

The disclosure resources enable the adversary to gather sequences of data from the calculated control actions u_k and the real measurements y_k through disclosure attacks. Denote $\mathcal{R}^u \subseteq \{1, \ldots, n_u\}$ and $\mathcal{R}^y \subseteq \{1, \ldots, n_y\}$ as the disclosure resources, i.e. the set of actuator and sensor channels that can be accessed during disclosure attacks, and let $\mathcal{I}_k$ be the control and measurement data sequence gathered by the adversary from time k_0 to k. The disclosure attacks can then be modeled as

$$\mathcal{I}_k := \mathcal{I}_{k-1} \cup \left\{ \begin{bmatrix} \Upsilon^u & 0 \\ 0 & \Upsilon^y \end{bmatrix} \begin{bmatrix} u_k \\ y_k \end{bmatrix} \right\}, \tag{8}$$

where $\mathcal{I}_{k_0} = \emptyset$ and $\Upsilon^u \in \mathbb{B}^{|\mathcal{R}^u| \times n_u}$ and $\Upsilon^y \in \mathbb{B}^{|\mathcal{R}^y| \times n_y}$ are the binary incidence matrices mapping the data channels to the corresponding data gathered by the adversary. As seen in the above description of disclosure attacks, the physical dynamics of the system are not affected by these type of attacks. Instead, these attacks gather intelligence that may enable more complex attacks, such as the replay attacks depicted in Fig. 1.

3.4. Disruption resources

In the system dynamics under attack, (6) and (7), disruption resources are related to the attack vector a_k and may be used to affect the several components of the system. The way a particular attack disturbs the system operation depends not only on the respective resources, but also on the nature of the attack. For instance, a physical attack directly perturbs the system dynamics, whereas a cyber attack disturbs the system through the cyber–physical couplings. To better illustrate this discussion we now consider physical and data deception attacks.

3.4.1. Physical resources

Physical attacks may occur in control systems, often in conjunction with cyber attacks. For instance, in the experiments reported in Amin, Litrico, Sastry, and Bayen (2010), water was pumped out of an irrigation system while the water level measurements were corrupted so that the attack remained stealthy. Since physical attacks are similar to the fault signals in (1), in the following sections we consider f_k to be the physical attack modifying the plant dynamics as

$$x_{k+1} = Ax_k + B\tilde{u}_k + Gw_k + Ff_k$$
$$y_k = Cx_k.$$

Considering $a_k = f_k$, the resulting system dynamics are described by (6) and (7) with

$$\mathbf{B} = \begin{bmatrix} F \\ 0 \end{bmatrix}, \qquad \mathbf{D} = 0.$$

Note that the disruption resources in this attack are captured in the matrix F.

3.4.2. 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 $\mathcal{R}_l^u \subseteq \{1, \ldots, n_u\}$ and $\mathcal{R}_l^y \subseteq \{1, \ldots, n_y\}$ as the deception resources, i.e. set of actuator and sensor channels that can be affected, the deception attacks are modeled as

$$\tilde{u}_k := u_k + \Gamma^u b_k^u, \qquad \tilde{y}_k := y_k + \Gamma^y b_k^y, \tag{9}$$

where the signals $b_k^u \in \mathbb{R}^{|\mathcal{R}_l^u|}$ and $b_k^y \in \mathbb{R}^{|\mathcal{R}_l^y|}$ represent the data corruption and $\Gamma^u \in \mathbb{B}^{n_u \times |\mathcal{R}_l^u|}$ and $\Gamma^y \in \mathbb{B}^{n_y \times |\mathcal{R}_l^y|}$ ($\mathbb{B} := \{0, 1\}$) 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.

Defining $a_k = [b_k^{u\top} \quad b_k^{y\top}]^\top$, the system dynamics are given by (6) and (7) with

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

Note that deception attacks do not possess any disclosure capabilities, as depicted in Fig. 1 for examples of deception attacks such as the bias injection attack.

4. Attack scenarios

In this section we discuss the general goal of an adversary and likely choices of the attack policy $g(\cdot, \cdot)$. In particular, using the framework introduced in the previous sections, we consider several attack scenarios where the adversary's goal is to drive the system to an unsafe state while remaining stealthy. For each scenario we formulate the corresponding stealthy attack policy, comment on the attack's performance, and describe the adversary's capabilities along each dimension of the attack space in Fig. 1, namely the disclosure resources, disruption resources, and model knowledge. A subset of these scenarios is illustrated by experiments on a process control testbed in Section 5.

4.1. Attack goals and constraints

In addition to the attack resources, the attack scenarios need to also include the intent of the adversary, 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 regions to characterize safety constraints.

Definition 1. At a given time instant k, the system is said to be safe if $x_k \in \mathcal{X}_x$, where $\mathcal{X}_x$ is a closed and compact set with non-empty interior.

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. Assuming that the system is in a safe state at the beginning of the attack, i.e. $x_{k_0} \in \mathcal{X}_x$, the attack is considered successful if the state is driven out of the safe set.

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, therefore, we have the attack vector $a_k = [f_k^\top \quad b_k^{u\top} \quad b_k^{y\top}]^\top$. Given the anomaly detector described in Section 2 and denoting $\mathbf{a}_{[k_0,\,k_f]} = [a_{k_0}^\top \quad \cdots \quad a_{k_f}^\top]^\top$ as 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.

Since the closed-loop system (6) and the anomaly detector (7) under linear attack policies are LTI systems, each of these systems can be separated into two components: the nominal component with $a_k = 0\ \forall k$ and the following systems

$$\begin{aligned} \eta^a_{k+1} &= \mathbf{A}\eta^a_k + \mathbf{B}a_k \\ \tilde{y}^a_k &= \mathbf{C}\eta^a_k + \mathbf{D}a_k \end{aligned} \tag{10}$$

and

$$\begin{aligned} \xi^a_{k+1} &= \mathbf{A}_e\xi^a_k + \mathbf{B}_e a_k \\ r^a_k &= \mathbf{C}_e\xi^a_k + \mathbf{D}_e a_k, \end{aligned} \tag{11}$$

with $\eta^a_0 = \xi^a_0 = 0$.

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

Albeit more conservative than Definition 2, this definition only depends on the attack signals $\mathbf{a}_{[k_0,\,k_f]}$. Therefore, the stealthiness of linear attacks on LTI systems may be analyzed independently of the noise inputs. Similarly, the impact of attacks on the closed-loop system can be analyzed through the linear system (10), as illustrated in Section 4.6 for the bias injection attack. For other classes of systems, e.g., nonlinear or switched systems, the analysis and characterization of attacks may have to consider the noise terms directly.

4.2. Denial-of-service attack

The DoS attacks prevent the actuator and sensor data from reaching their respective destinations and results in the absence of data. To model absent data, we consider one of the typical mechanisms used by digital controllers to deal with unavailable data (Schenato, 2009), in which the absent data is replaced with the last received data, u_{τ_u} and y_{τ_y}.

Attack policy: Denote $\mathcal{R}^u_A \subseteq \{1, \ldots, n_u\}$ and $\mathcal{R}^y_A \subseteq \{1, \ldots, n_y\}$ as the set of actuator and sensor channels that can be made unavailable and define $S^u_k \in \mathbb{B}^{|\mathcal{R}^u_A|\times|\mathcal{R}^u_A|}$ and $S^y_k \in \mathbb{B}^{|\mathcal{R}^y_A|\times|\mathcal{R}^y_A|}$ as Boolean diagonal matrices where the ith diagonal entry indicates whether a DoS attack is performed ($[S^{(\cdot)}_k]_{ii} = 1$) or not ($[S^{(\cdot)}_k]_{ii} = 0$) on the corresponding channel. Using the latter variables, DoS attacks can be modeled as deception attacks in (9) with

$$\begin{aligned} b^u_k &:= -S^u_k \Gamma^{u\top}(u_k - u_{\tau_u}) \\ b^y_k &:= -S^y_k \Gamma^{y\top}(y_k - y_{\tau_y}) \end{aligned} \tag{12}$$

and $a_k = [b^{u\top}_k \ b^{y\top}_k]^\top$. Therefore DoS attacks on the data are a type of disruptive attacks, as depicted in Fig. 1.

The attack scenario analyzed in this paper considers a Bernoulli adversary (Amin et al., 2009) on the sensor channels following the random policy

$$\begin{aligned} &\mathbb{P}([S^y_k]_{ii} = 1) = 0, \quad \forall i = 1, \ldots, |\mathcal{R}^u_A|,\ k < k_0 \\ &\mathbb{P}([S^y_k]_{ii} = 1) = p, \quad \forall i = 1, \ldots, |\mathcal{R}^u_A|,\ k \geq k_0 \end{aligned}$$

where p is the probability of blocking the data packet at any given time.

Attack performance: Although the absence of data packets is not stealthy, since it is trivially detectable, DoS attacks may be misdiagnosed as a poor network condition. As for the impact on the closed-loop system, the results available for Bernoulli packet losses readily apply to the current attack scenario (Schenato, 2009; Schenato, Sinopoli, Franceschetti, Poolla, & Sastry, 2007; Zhang, Branicky, & Phillips, 2001). In particular, we recall the following result applied to (12).

Proposition 4 (*Theorem 8 in Zhang et al., 2001*)**.** *Assume that the closed-loop system with no DoS attack is stable. Then the closed-loop system with Bernoulli DoS attacks is exponentially stable for* $p \in [0,\ 1)$ *if the open-loop system*

$$\eta_{k+1} = \begin{bmatrix} A & BC_c \\ 0 & A_c \end{bmatrix} \eta_k$$

is marginally stable.

Disclosure resources: Although the proposed model of DoS attacks in (12) contains the control and output signals, note that no disclosure resources are needed in the actual implementation of the attack. Hence we have $\mathcal{R}^u = \mathcal{R}^y = \emptyset$.

Disruption resources: The disruption capabilities correspond to the data channels that the adversary is able to make unavailable, $\mathcal{R}^u_A$ and $\mathcal{R}^y_A$.

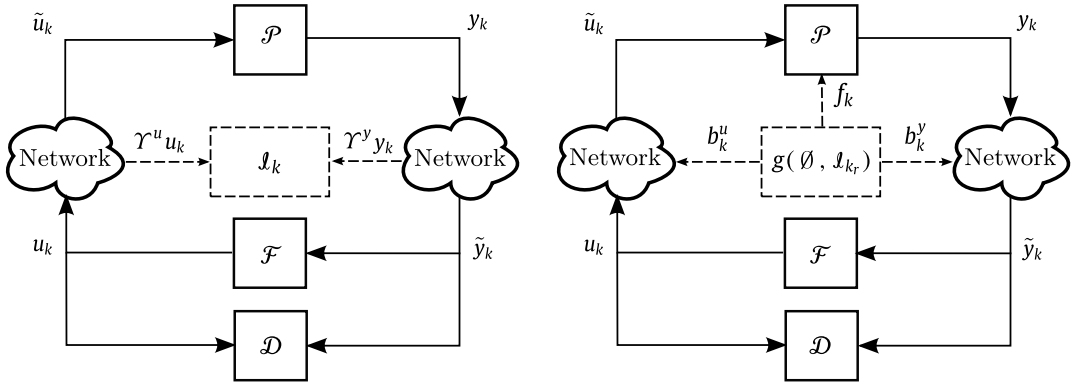
Model knowledge: For the Bernoulli attack policy, no *a priori* knowledge of the system model is needed.

4.3. Replay attack

In replay attacks the adversary first performs a disclosure attack from $k = k_0$ until k_r, gathering sequences of data $\mathcal{I}_{k_r}$, and then begins replaying the recorded data at time $k = k_r + 1$ until the end of the attack at $k = k_f > k_r$, as illustrated in Fig. 3. In the scenario considered here the adversary is also able to perform a physical attack while replaying the recorded data, which covers the experiment on a water management SCADA system reported in Amin et al. (2010) and one of Stuxnet's operation mode (Falliere et al., 2011).

Attack policy: Similar to the work in Mo and Sinopoli (2009), assuming $\mathcal{R}^{(\cdot)} = \mathcal{R}^{(\cdot)}_l$ i.e., the adversary can corrupt the digital channels from which the data sequences are gathered, the replay attack policy can be described as

$$\text{Phase I}: \begin{cases} a_k = 0 \\ \mathcal{I}_k = \mathcal{I}_{k-1} \cup \left\{ \begin{bmatrix} \Upsilon^u & 0 \\ 0 & \Upsilon^y \end{bmatrix} \begin{bmatrix} u_k \\ y_k \end{bmatrix} \right\}, \end{cases} \tag{13}$$



(a) Phase I of the replay attack (13).

(b) Phase II of the replay attack (14).

Fig. 3. Schematic of the replay attack.

with $k_0 \leq k \leq k_r$ and $\mathcal{I}_{k_0} = \emptyset$ and

$$\text{Phase II}: \begin{cases} a_k = \begin{bmatrix} g_f(\mathcal{K}, \mathcal{I}_{k_r}) \\ \Upsilon^u(u_{k-T} - u_k) \\ \Upsilon^y(y_{k-T} - y_k) \end{bmatrix} \\ \mathcal{I}_k = \mathcal{I}_{k-1}, \end{cases} \tag{14}$$

where $T = k_r - 1 + k_0$ and $k_r + 1 \leq k \leq k_f$. An interesting instance of this attack scenario consists of applying a pre-defined physical attack to the plant, while using replay attacks to render the attack stealthy. In this case the physical attack signal f_k corresponds to an open-loop signal, $f_k = g_f(k)$. Note that, while (14) resembles a time-delay of length T, replay attacks differ from delayed data in a subtle but important way: all measurement data during the attack interval $[k_r + 1,\ k_f]$ is never available to the anomaly detector. As in Amin et al. (2010), this allows the adversary to design the attack so that no alarm is triggered by the anomaly detector.

Attack performance: The work in Mo and Sinopoli (2009) provided conditions under which replay attacks with access to all measurement data channels are stealthy. However, these attacks are not guaranteed to be stealthy when only a subset of the data channels is attacked. In this case, the stealthiness constraint may require additional knowledge of the system model. For instance, the experiment presented in Section 5 requires knowledge of the physical system structure, so that f_k only excites the attacked measurements. Hence f_k can be seen as a zero-dynamics attack with respect to the uncompromised measurements, which is characterized in the section below. Since the impact of the replay attack is dependent only on f_k, we refer the reader to Section 4.4 for a characterization of the replay attack's impact.

Disclosure resources: The required disclosure capabilities correspond to the data channels that can be eavesdropped by the adversary, namely $\mathcal{R}^u$ and $\mathcal{R}^y$.

Disruption resources: In this case the deception capabilities correspond to the data channels that the adversary can tamper with, $\mathcal{R}_I^u$ and $\mathcal{R}_I^y$. In particular, for replay attacks the adversary can only tamper with the data channels from which data has been previously recorded, i.e. $\mathcal{R}_I^u \subseteq \mathcal{R}^u$ and $\mathcal{R}_I^y \subseteq \mathcal{R}^y$.

Direct disruption of the physical system through the signal f_k depends on having direct access to the physical system, modeled by the matrix F in (1).

Model knowledge: Note that no *a priori* knowledge $\mathcal{K}$ on the system model is needed for the cyber component of the attack, namely the data disclosure and deception attack, as seen in the attack policy (13) and (14). As for the physical attack, f_k, the required knowledge is scenario dependent. In the scenario considered in the experiments described in Section 5, this component was modeled as an open-loop signal, $f_k = g_f(k)$.

4.4. Zero-dynamics attack

Recalling that for linear attack policies the plant and the anomaly detector are LTI systems, (10) and (11) respectively, Definition 3 states that this type of attacks are 0-stealthy if $r_k^a = 0,\ k = k_0, \ldots, d_f$. The idea of 0-stealthy attacks consists of designing an attack policy and attack signal $\mathbf{a}_{[k_0,\ k_f]}$ so that the residue r_k does not change due to the attack. In other words, these attacks are decoupled from the output of the closed-loop linear system (7), namely r_k, and their design in general depends on the plant, controller, and anomaly detector dynamics. A particular subset of 0-stealthy attacks that only depend on the plant dynamics are characterized in the following lemma.

Lemma 5. *The attack signal $\mathbf{a}_{[k_0,\ k_f]}$ is 0-stealthy with respect to an arbitrary anomaly detector $\mathcal{D}$ if $\tilde{y}_k^a = 0,\ \forall k \geq k_0$.*

Proof. Consider arbitrary controller and anomaly detector and their corresponding attacked components in (10) and (11) with $s_0^a = 0$. From the controller dynamics it directly follows that $\tilde{y}_k^a = 0,\ \forall k \geq k_0$ results in $u_k^a = 0,\ \forall k \geq k_0$, as the input to the controller $(\tilde{y}_k^a)$ is zero. Since $s_0^a = 0$ and $\tilde{y}_k^a = u_k^a = 0,\ \forall k \geq k_0$, meaning that the detector's inputs are zero, we then conclude $r_k^a = 0,\ \forall k \geq k_0$.

Lemma 5 indicates that these attacks are decoupled from the plant output y_k, thus being stealthy with respect to arbitrary anomaly detectors. Hence finding 0-stealthy attack signals relates to the output-zeroing problem or zero-dynamics studied in the control theory literature (Zhou et al., 1996). Note that such an attack requires the perfect knowledge of the plant dynamics $\mathcal{P}$ and the attack signal is based on the open-loop prediction of the output changes due to the attack. This is illustrated in Fig. 4 where $\mathcal{K}_z$ denote the zero-dynamics and there is no disclosure of sensor or actuator data.

Attack policy: The attack policy corresponds to the input sequence (a_k) that makes the outputs of the process $(\tilde{y}_k^a)$ identically zero for all k and is illustrated in Fig. 4. It can be shown (Zhou et al., 1996) that the solution to this problem is given by the sequence

$$a_k = \nu^k g, \tag{15}$$

parameterized by the system zero ν and the corresponding input-zero direction g.

For sake of simplicity we consider a particular instance of this attack, where only the actuator data is corrupted. In this case the zero attack policy corresponds to the transmission zero-dynamics of the plant. The plant dynamics due to an attack on the actuator data are described by

$$\begin{aligned} x_{k+1}^a &= Ax_k^a + Ba_k \\ \tilde{y}_k^a &= Cx_k^a \end{aligned} \tag{16}$$

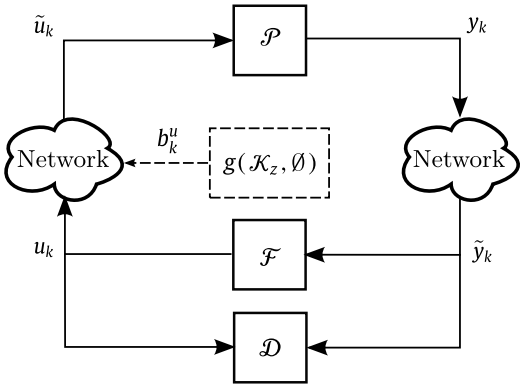


Fig. 4. Schematic of the zero-dynamics attack.

with $a_k = b_k^u$. Given the discrete-time system (16) with B having full column rank, the transmission zeros can be calculated as the values $\nu \in \mathbb{C}$ that cause the matrix $P(\nu)$ to lose rank, where

$$P(\nu) = \begin{bmatrix} \nu I - A & -B \\ C & 0 \end{bmatrix}.$$

Those values are called minimum phase or non-minimum phase zeros depending on whether they are stable or unstable zeros, respectively. In discrete-time systems a zero is stable if $|\nu| < 1$ and unstable otherwise.

The input-zero direction can be obtained by solving the following equation:

$$\begin{bmatrix} \nu I - A & -B \\ C & 0 \end{bmatrix} \begin{bmatrix} x_0 \\ g \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{17}$$

where x_0 is the initial state of the system for which the input sequence (15) results in an identically zero output.

Lemma 6. *Let x_0 be the initial state of the system, where x_0 satisfies (17). The state trajectories generated by the zero-dynamics attack are contained in span(x_0) i.e., $x_k^a \in$ span(x_0) $\forall k \geq 0$.*

Proof. The proof follows an induction argument. Consider the zero-dynamics attack parameterized by x_0 and g and the state evolution under attack, $x_{k+1}^a = Ax_k^a + \nu^k Bg$ with $x_0^a = x_0$. For $k = 0$ it follows from (17) that $x_1^a = Ax_0 + Bg = \nu x_0$. Supposing that $x_k^a = \nu^k x_0$ yields $x_{k+1}^a = Ax_k^a + \nu^k Bg = \nu^k(Ax_0^a + Bg) = \nu^{k+1}x_0$ for all $k \geq 0$, which concludes the proof.

Attack performance: Note that the zero-dynamics attack is 0-stealthy only if $x_0^a = x_0$. However the initial state of the system under attack x_0^a is defined to be zero at the beginning of the attack. Therefore stealthiness of the attack may be violated for large differences between $x_0^a = 0$ and x_0. We refer the reader to Teixeira, Shames, Sandberg, and Johansson (2012) for a detailed analysis of the effects of zero initial conditions on zero-dynamics attacks.

If the zero is stable, that is $|\nu| < 1$, the attack will asymptotically decay to zero, thus having little effect on the plant. However, in the case of unstable zeros the attack grows geometrically, which could cause a great damage to the process. This statement is captured in the following result.

Theorem 7. *A zero-dynamics attack with $|\nu| > 1$ leads the system to an unsafe state if and only if span(x_0) is not contained in $\mathcal{S}_x$.*

Proof. Follows directly from Lemma 6 and from the fact that the zero-attack with $|\nu| > 1$ generates an unstable state trajectory moving away from the origin along span(x_0).

Disclosure resources: This attack scenario considers an open-loop attack policy and so no disclosure capabilities are required, resulting in $\mathcal{R}^u = \mathcal{R}^y = \emptyset$ and $\mathcal{I}_k^u = \mathcal{I}_k^y = \emptyset$.

Disruption resources: The disruption capabilities in this attack scenario correspond to the ability of performing deception attacks on the actuator data channels. Therefore the required resources are $\mathcal{R}_l^u = \{1, \ldots, n_u\}$, $\mathcal{R}_l^y = \emptyset$, and $F = 0$.

Model knowledge: The ability to compute the open-loop attack policy requires perfect knowledge of the zero-dynamics, which we denote as $\mathcal{K}_z$. Moreover, the zero-dynamics can be computed from the plant dynamics, namely A, B, and C. No knowledge of the feedback controller or anomaly detector is assumed in this scenario.

Although the former analysis considers LTI systems, the concept of zero-dynamics has been extended to other classes of system, e.g., nonlinear systems (Isidori, 1995). Hence zero-dynamics attacks could be directly extended to other classes of system in the noiseless case. In the presence of noise however, the interplay between the zero-dynamics and the noise inputs is not trivial and requires further analysis.

4.5. Local zero-dynamics attack

In the previous scenario the zero-dynamics attack was characterized in terms of the entire system. Here we further restrict the adversary resources by considering that the adversary has disruption resources and knows the model of only a subset of the system. In particular, we rewrite the plant dynamics (16) as

$$\begin{aligned} \begin{bmatrix} x_{k+1}^1 \\ x_{k+1}^2 \end{bmatrix} &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_k^1 \\ x_k^2 \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} a_k \\ \tilde{y}_k^a &= \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} x_k^1 \\ x_k^2 \end{bmatrix} \end{aligned} \tag{18}$$

and assume the adversary has access to only A_{11}, A_{21}, B_1, and C_1. From the adversary's view, this local system is characterized by

$$\begin{aligned} x_{k+1}^1 &= A_{11}x_k^1 + B_1 a_k + A_{12}x_k^2 \\ y_k^l &= \begin{bmatrix} C_1 \\ A_{21} \end{bmatrix} x_k^1, \end{aligned}$$

where y_k^l encodes the measurements depending on the local state, $C_1 x_k^1$, and the interaction between the local subsystem and the remaining subsystems, $A_{21}x_k^1$.

Attack policy: Similar to the zero-dynamics attack, the attack policy is given by the sequence $a_k = \nu^k g^1$, where g^1 is the input zero direction for the chosen zero ν. The input zero direction can be obtained by solving

$$\begin{bmatrix} \nu I - A_{11} & -B_1 \\ C_1 & 0 \\ A_{21} & 0 \end{bmatrix} \begin{bmatrix} x_0^1 \\ g^1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Note that the zero-dynamics parameterized by g^1 and ν correspond to local zero-dynamics of the global system.

Attack performance: A similar discussion as for the global zero-dynamics attack applies to this scenario. In particular, the stealthiness of the local zero-dynamics attack may be violated for large differences between x_0^1 and 0. Additionally, as stated in Theorem 7, attacks associated with unstable zeros yielding $|\nu| > 1$ are more dangerous and may lead the system to an unsafe state.

Disclosure resources: This attack scenario considers an open-loop attack policy and so no disclosure capabilities are required, resulting in $\mathcal{R}^u = \mathcal{R}^y = \emptyset$ and $\mathcal{I}_k^u = \mathcal{I}_k^y = \emptyset$ $\forall k$.

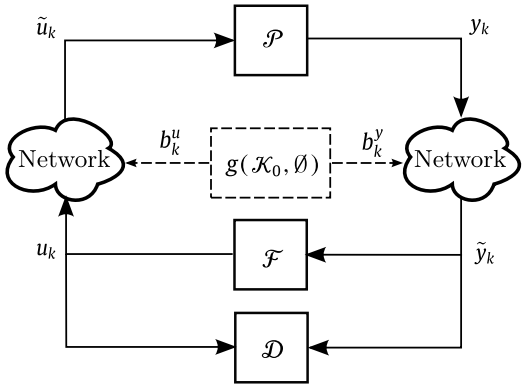


Fig. 5. Schematic of the bias injection attack.

Disruption resources: The disruption capabilities in this attack scenario correspond to the ability of performing deception attacks on the actuator data channels of the local subsystem. Therefore the required resources are $\mathcal{R}_l^u = \{1, \ldots, n_u^1\}$, $\mathcal{R}_l^y = \emptyset$, and $F = 0$.

Model knowledge: The open-loop attack policy requires the perfect knowledge of the local zero-dynamics, denoted as $\tilde{\mathcal{K}}_z$ and obtained from A_{11}, B_1, C_1, and A_{21}.

4.6. Bias injection attack

Consider a false-data injection scenario where the adversary's goal is to inject a constant bias in the system without being detected. Furthermore, the bias is computed so that the impact at steady-state is maximized.

Attack policy: The bias injection attack is illustrated in Fig. 5. The attack policy is composed of a steady-state component, the desired bias denoted as a_∞, and a transient component. For the transient, we consider that the adversary uses a low-pass filter so that the data corruptions are slowly converging to the steady-state values. As an example, for a set of identical first-order filters the open-loop attack sequence is described by

$$a_{k+1} = \beta a_k + (1-\beta) a_\infty^*, \tag{19}$$

where $a_0 = 0$ and $0 < \beta < 1$ is chosen to ensure that the attack is α-stealthy during the transient regime. The steady-state attack policy yielding the maximum impact on the physical system is described below, where the computation of a_∞ is summarized in Theorems 11 and 13.

Attack performance: First the steady-state policy is considered. Denote a_∞ as the bias to be injected and recall the anomaly detector dynamics under attack (7). The steady-state detectability of the attack is then dependent on the steady-state value of the residual

$$r_\infty^a = \left(\mathbf{C}_e (I - \mathbf{A}_e)^{-1} \mathbf{B}_e + \mathbf{D}_e\right) a_\infty =: G_{ra} a_\infty.$$

Consider the set $\mathcal{U}_{[0,\infty]}^a = \left\{\mathbf{r}_{[0,\infty]}^a : \|r_k^a\|_2 \le \delta_\alpha,\ \forall k \ge 0\right\}$ and recall Definition 3 for α-stealthy attacks. A necessary condition for the bias injection attack to be α-stealthy is

$$\|G_{ra} a_\infty\|_2 \le \delta_\alpha. \tag{20}$$

Although attacks satisfying (20) could be detected during the transient, incipient attack signals slowly converging to a_∞ may go undetected. In fact, sufficient conditions for the bias attack to be α-stealthy are given in Theorem 14 and the results are illustrated through experiments in Section 5.

The impact of such attacks can be evaluated using the closed-loop dynamics under attack given by (6). Recalling that $\eta_k^a = [x_k^{a\top}\ z_k^{a\top}]^\top$, the steady-state impact on the state is given by

$$x_\infty^a = [I\ 0]\,(I - \mathbf{A})^{-1}\,\mathbf{B} a_\infty =: G_{xa} a_\infty.$$

Consider the following safe set defined in terms of x_k^a.

Definition 8. The safe set $\mathcal{S}_{x^a}^2$ is defined as

$$\mathcal{S}_{x^a}^2 = \left\{x \in \mathbb{R}^{n_x} : \|x\|_2^2 \le 1\right\},$$

and the system is said to be in a safe state if $x_k^a \in \mathcal{S}_{x^a}^2$.

For the 2-norm safe set $\mathcal{S}_{x^a}^2$, the most dangerous bias injection attack corresponds to the α-stealthy attack yielding the largest bias in the 2-norm sense, which can be computed by solving

$$\begin{aligned} \max_{a_\infty} \quad & \|G_{xa} a_\infty\|_2^2 \\ \text{s.t.} \quad & \|G_{ra} a_\infty\|_2^2 \le \delta_\alpha^2. \end{aligned} \tag{21}$$

Lemma 9. *The optimization problem (21) is bounded if and only if* $\ker(G_{ra}) \subseteq \ker(G_{xa})$.

Proof. Suppose that $\ker(G_{ra}) \neq \emptyset$ and consider the subset of solutions where $a_\infty \in \ker(G_{ra})$. For this subset of solutions, the optimization problem then becomes $\max_{a_\infty \in \ker(G_{ra})} \|G_{xa} a_\infty\|_2^2$. Since the objective function does not have an upper-bound and the feasible set is unbounded, the optimal value is unbounded unless $G_{xa} a_\infty = 0$ for all $a_\infty \in \ker(G_{ra})$ i.e., $\ker(G_{ra}) \subseteq \ker(G_{xa})$. Noting that the feasible set and the objective function are bounded for all solutions $a_\infty \notin \ker(G_{ra})$ concludes the proof.

Given Lemma 9, below we consider the non-trivial case for which it holds that $\ker(G_{ra}) \subseteq \ker(G_{xa})$. The above optimization problem can be transformed into a generalized eigenvalue problem and the corresponding optimal solution characterized in terms of generalized eigenvalues and eigenvectors. Before formalizing this statement, we introduce the following result.

Lemma 10. *Let $Q \in \mathbb{R}^{n\times n}$ and $P \in \mathbb{R}^{n\times n}$ be positive semi-definite matrices satisfying* $\ker(Q) \subseteq \ker(P)$. *Denote λ^* as the largest generalized eigenvalue of the matrix pencil (P, Q) and v^* as the corresponding eigenvector. The matrix $P - \lambda Q$ is negative semi-definite for a generalized eigenvalue λ if and only if $\lambda = \lambda^* \ge 0$. Moreover, we have $x^\top (P - \lambda^* Q) x = 0$ with $Qx \neq 0$ if and only if $x \in span(v^*)$.*

Proof. Recall that $\lambda \in \mathbb{C}$ is a finite generalized eigenvalue of (P, Q) associated with the generalized eigenvector v if $(P - \lambda Q)v = 0$ and $v \notin \ker(Q)$. Using the assumption $\ker(Q) \subseteq \ker(P)$ and the matrix congruence property, we have that $P - \lambda Q \preceq 0$ is equivalent to $\tilde{P} - \lambda I \preceq 0$, where the matrix $\tilde{P} \in \mathbb{R}^{r\times r}$, with $r = \operatorname{rank}(Q)$, is positive semi-definite. Moreover, the eigenvalues of $\tilde{P}$ correspond to the generalized eigenvalues of (P, Q), which implies that $\lambda^\star$ is non-negative. Therefore, we conclude $P - \lambda Q \preceq 0$ holds if and only if $\tilde{P} \preceq \lambda I$, which can be rewritten as $\lambda \ge \lambda^\star$. Constraining λ to be a generalized eigenvalue of (P, Q), we have $\lambda = \lambda^\star \ge 0$. The remaining of the proof follows directly from the definition of generalized eigenvectors.

The optimal bias injection attack in the sense of (21) is characterized by the following result.

Theorem 11. *Consider the 2-norm safe set $\mathcal{S}_{x^a}^2$ and the corresponding optimal α-stealthy bias injection attack parameterized by the optimization problem (21), which is assumed to be bounded. Denote λ^* and v^* as the largest generalized eigenvalue and corresponding unit-norm eigenvector of the matrix pencil $(G_{xa}^\top G_{xa},\ G_{ra}^\top G_{ra})$. The optimal bias injection attack is given by*

$$a_\infty^* = \pm \frac{\delta_\alpha}{\|G_{ra} v^*\|_2} v^*, \tag{22}$$

and the corresponding optimal value is $\|G_{xa} a_\infty\|_2^2 = \lambda^* \delta_\alpha^2$. Moreover, at steady-state the system is in a safe state if and only if $\lambda^* \delta_\alpha^2 \le 1$.

Proof. The necessary and sufficient conditions for the optimization problem (21) are given by Hiriart-Urruty (2001)

$$0 = (G_{xa}^\top G_{xa} - \lambda^* G_{ra}^\top G_{ra}) a_\infty^*,$$
$$0 = a_\infty^{*\top} G_{ra}^\top G_{ra} a_\infty^* - \delta_\alpha^2,$$
$$0 \geq y^\top (G_{xa}^\top G_{xa} - \lambda^* G_{ra}^\top G_{ra}) y, \quad \text{for } y \neq 0.$$

Suppose λ^* is the largest generalized eigenvalue of $(G_{xa}^\top G_{xa},\ G_{ra}^\top G_{ra})$ and let v^* be the corresponding eigenvector. Scaling v^* by κ so that $a_\infty^* = \kappa v^*$ satisfies $\|G_{ra} a_\infty^*\|_2^2 = \delta_\alpha^2$ leads to $\kappa = \pm \frac{\delta_\alpha}{\|G_{ra} v^*\|_2}$, and the first and second conditions are satisfied. As for the third condition, note that $G_{xa}^\top G_{xa} - \lambda^* G_{ra}^\top G_{ra}$ is negative semi-definite by Lemma 10, given that λ^* is the largest generalized eigenvalue, $G_{xa}^\top G_{xa}$ and $G_{ra}^\top G_{ra}$ are positive semi-definite, and the assumption that $\ker(G_{ra}) \subseteq \ker(G_{xa})$. To conclude the proof, observe that the optimal value is given by $a_\infty^{*\top} G_{xa}^\top G_{xa} a_\infty^* = \lambda^* a_\infty^{*\top} G_{ra}^\top G_{ra} a_\infty^* = \lambda^* \delta_\alpha^2 = \|x_\infty^a\|_2^2$ and thus, by definition, $x_\infty^a \in \mathcal{X}_{x^a}^2$ if and only if $\lambda^* \delta_\alpha^2 \leq 1$.

More generally, the optimal bias injection attacks for ellipsoidal safe sets $\mathcal{X}_{x^a} = \left\{x^a \in \mathbb{R}^{n_x} : {x^a}^\top P x^a \leq 1\right\}$, with P positive definite, can be found by replacing the objective function in (21) by $\|P^{1/2} G_{xa} a_\infty\|_2^2$. Similarly, the steady-state attack policy is derived below for the following safe set.

Definition 12. The safe set $\mathcal{X}_{x^a}^\infty$ is defined as

$$\mathcal{X}_{x^a}^\infty = \left\{x \in \mathbb{R}^{n_x} : \|x\|_\infty \leq 1\right\},$$

and the system is said to be in a safe state if $x_k^a \in \mathcal{X}_{x^a}^\infty$.

Given the infinity-norm safe set $\mathcal{X}_{x^a}^\infty$, the bias injection attack with the largest impact corresponds to the α-stealthy attack yielding the largest bias in the infinity-norm sense. This attack can be obtained by solving the following optimization problem

$$\begin{aligned} \max_{a_\infty} \quad & \|G_{xa} a_\infty\|_\infty \\ \text{s.t.} \quad & \|G_{ra} a_\infty\|_2 \leq \delta_\alpha. \end{aligned} \tag{23}$$

To solve this problem, observe that

$$\|G_{xa} a_\infty\|_\infty = \max_i \|e_i^\top G_{xa} a_\infty\|_2,$$

where the vector e_i is ith column of the identity matrix. Therefore, one can transform the optimization problem (23) into a set of problems with the same structure as (21), obtaining

$$\begin{aligned} \max_i \max_{a_\infty^i} \quad & \left\|e_i^\top G_{xa} a_\infty^i\right\|_2 \\ \text{s.t.} \quad & \left\|G_{ra} a_\infty^i\right\|_2 \leq \delta_\alpha. \end{aligned} \tag{24}$$

Theorem 13. *Consider the infinity-norm safe set $\mathcal{X}_{x^a}^\infty$ and the corresponding optimal α-stealthy bias injection attack parameterized by the optimization problem (23), which is assumed to be bounded. Let e_i be the ith column of the identity matrix and denote λ_i^* and v_i^* as the largest generalized eigenvalue and corresponding unit-norm eigenvector of the matrix pencil $G_{xa}^\top e_i e_i^\top G_{xa} - \lambda G_{ra}^\top G_{ra}$. Letting $\lambda^* = \max_i \lambda_i^*$, with v^* as the corresponding generalized eigenvector, the optimal bias attack is given by*

$$a_\infty^* = \pm \frac{\delta_\alpha}{\|G_{ra} v^*\|_2} v^*, \tag{25}$$

and the corresponding optimal value is $\|G_{xa} a_\infty\|_\infty = \sqrt{\lambda^*}\delta_\alpha$. Moreover, at steady-state the system is in a safe state if and only if $\lambda^* \delta_\alpha^2 \leq 1$.

Proof. The proof follows from considering the optimization problems (24) and applying Theorem 11.

Recall that the steady-state value of the data corruption a_∞^* is not sufficient for the attack to be α-stealthy, since the transients are disregarded. In practice, however, it has been observed in the fault diagnosis literature that faults with slow dynamics, also known as incipient faults, are difficult to distinguish from model uncertainty and noise (Chen & Patton, 1999; Zhang, Polycarpou, & Parisini, 2002). Therefore the low-pass filter dynamics in the attack policy (19) could be designed sufficiently slow as to make detection more difficult. Below we provide sufficient conditions under which a given filter parameter β renders the bias attack α-stealthy with respect to $\mathcal{U}_{[0,\,\infty]}^a = \left\{\mathbf{r}_{[0,\,\infty]}^a : \|r_k^a\|_2 \leq \delta_\alpha,\ \forall k \geq 0\right\}$.

Theorem 14. *Consider the attack policy $a_{k+1} = \beta a_k + (1-\beta) a_\infty^*$ with $\beta \in (0,\ 1)$. The residual r_k^a is characterized as the output of the autonomous system*

$$\begin{aligned} \psi_{k+1}^a &= \bar{A} \psi_k^a \\ r_k^a &= \bar{C} \psi_k^a \end{aligned} \tag{26}$$

with $\bar{C} = [\mathbf{C}_e\ \mathbf{D}_e\ 0]$ and

$$\bar{A} = \begin{bmatrix} \mathbf{A}_e & \mathbf{B}_e & 0 \\ 0 & \beta I & (1-\beta) I \\ 0 & 0 & I \end{bmatrix}, \qquad \psi_0^a = \begin{bmatrix} 0 \\ 0 \\ a_\infty^* \end{bmatrix}.$$

Moreover, the attack policy is α-stealthy for a given β if the following optimization problem admits a solution

$$\begin{aligned} \min_{\gamma, P} \quad & \gamma \\ \text{s.t.} \quad & \gamma \leq \delta_\alpha^2,\ P \succ 0,\ {\psi_0^a}^\top P \psi_0^a \leq 1, \\ & 0 \preceq \begin{bmatrix} P & \bar{C}^\top \\ \bar{C} & \gamma I \end{bmatrix}, \\ & 0 \succ \bar{A}^\top P \bar{A} - P. \end{aligned} \tag{27}$$

Proof. The autonomous system is directly obtained by considering the augmented state $\psi^a = [\xi_{k|k}^{a\top}\ l_k^\top\ \upsilon_k^\top]^\top$, where l_k is the state of the low-pass filter bank and υ_k the integral state initialized at $\upsilon_0 = a_\infty$. Given this autonomous system, one observes that the attack is α-stealthy if and only if the corresponding output-peak $\|r_k^a\|_2^2$ is bounded by δ_α^2 for all $k \geq 0$, given the initial condition parameterized by α_∞^*. The remainder of the proof follows directly from the results in Boyd, El Ghaoui, Feron, and Balakrishnan (1994) regarding output-peak bounds for autonomous systems.

Disclosure resources: Similarly to the zero attack, no disclosure capabilities are required for this attack, since the attack policy is open-loop. Therefore we have $\mathcal{R}^u = \mathcal{R}^y = \emptyset$ and $\mathcal{I}_k^u = \mathcal{I}_k^u = \emptyset$ for all k.

Disruption resources: The biases may be added to both the actuator and sensor data, hence the required resources are $\mathcal{R}_l^u \subseteq \{1, \ldots, n_u\}$, $\mathcal{R}_l^y \subseteq \{1, \ldots, n_y\}$. Since no physical attack is performed, we have $F = 0$.

Model knowledge: As seen in (21), the open-loop attack policy (19) requires the knowledge of the closed-loop system and anomaly detector steady-state gains G_{ra} and G_{xa}, which we denoted as $\mathcal{K}_0$ as shown in Fig. 5.

5. Experiments

In this section we present our testbed and report experiments on staged cyber attacks following the different scenarios described in the previous section.

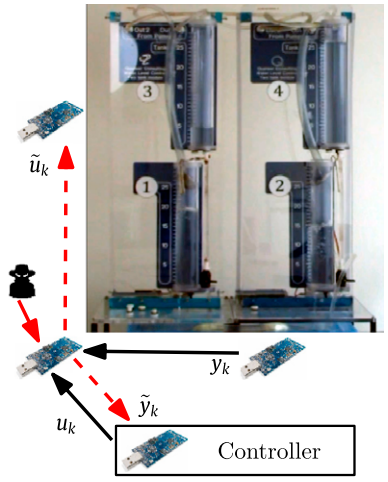


Fig. 6. Schematic diagram of the testbed with the Quadruple-Tank Process and a multi-hop communication network.

5.1. Quadruple-Tank Process

Our testbed consists of a Quadruple-Tank Process (QTP) (Johansson, 2000) controlled through a wireless communication network, as shown in Fig. 6.

The plant model can be found in Johansson (2000)

$$\begin{aligned}
\dot{h}_1 &= -\frac{a_1}{A_1}\sqrt{2gh_1} + \frac{a_3}{A_1}\sqrt{2gh_3} + \frac{\gamma_1 k_1}{A_1}u_1,\\
\dot{h}_2 &= -\frac{a_2}{A_2}\sqrt{2gh_2} + \frac{a_4}{A_2}\sqrt{2gh_4} + \frac{\gamma_2 k_2}{A_2}u_2,\\
\dot{h}_3 &= -\frac{a_3}{A_3}\sqrt{2gh_3} + \frac{(1-\gamma_2)k_2}{A_3}u_2,\\
\dot{h}_4 &= -\frac{a_4}{A_4}\sqrt{2gh_4} + \frac{(1-\gamma_1)k_1}{A_4}u_1,
\end{aligned} \tag{28}$$

where $h_i \in [0,\ 30]$ 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. Moreover, given the range of the water levels, the following safe set is considered $\mathcal{S}_x = \{x \in \mathbb{R}^{n_x} : \ \|x - \sigma\mathbf{1}\|_\infty \leq 15,\ \sigma = 15\}$, where $\mathbf{1} \in \mathbb{R}^{n_x}$ is a vector with all entries set to 1.

The QTP is controlled using a centralized LQG controller with integral action running in a remote computer and a wireless network is used for the communications. A Kalman-filter-based anomaly detector is also running in the remote computer and alarms are triggered according to (4), for which we have considered $\mathcal{U}_{[k_0,\,\infty]} = \left\{\mathbf{r}_{[k_0,\,\infty]} : \|r_k\|_2 \leq \delta_\alpha + \delta_r,\ \forall k \geq k_0\right\}$ and $\mathcal{U}^a_{[k_0,\,\infty]} = \left\{\mathbf{r}^a_{[k_0,\,\infty]} : \|r^a_k\|_2 \leq \delta_\alpha,\ \forall k \geq k_0\right\}$ with $\delta_r = 0.15$ and $\delta_\alpha = 0.25$ for illustration purposes. The communication network is multi-hop, having one additional wireless device relaying the data.

5.2. Denial-of-service attack

Here we consider the case where the QTP suffers a DoS attack on both sensors, while operating at a constant set-point. The state and residual trajectories from this experiment are presented in Fig. 7. The DoS attack follows a Bernoulli model (Amin et al., 2009) with $p = 0.9$ as the probability of packet loss and the last received data is used in the absence of data. From Proposition 4, we have that the closed-loop system under such DoS attack is exponentially stable. The DoS attack initiates at $t \approx 100$ s, leading to an increase in the residual due to packet losses. However the residual remained below the threshold during the attack and there were no significant changes in the system's state.

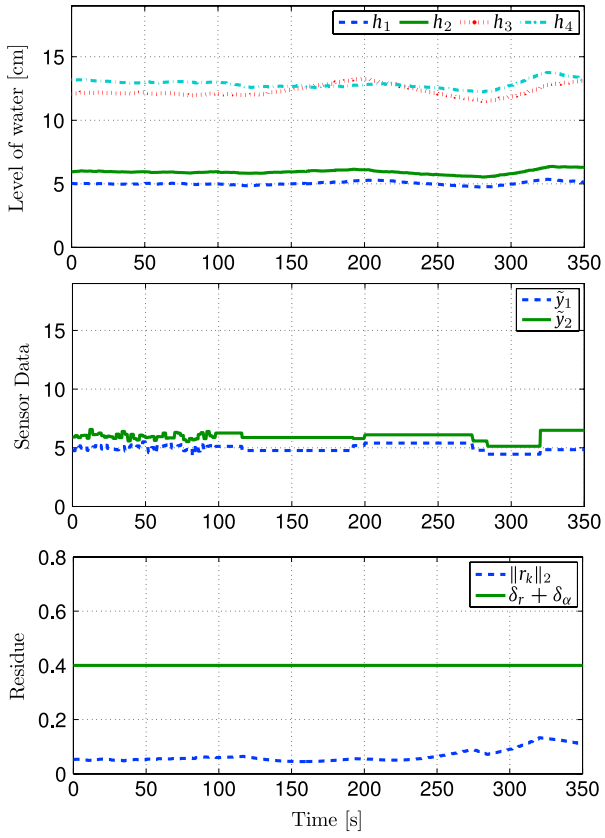


Fig. 7. Results for the DoS attack performed against both sensors since $t \approx 100$ s.

5.3. Replay attack

In this scenario, the QTP is operating at a constant set-point while a hacker desires to steal water from tank 4, the upper tank on the right. An example of this attack is presented in Fig. 8, where the replay attack policy is the one described in Section 4.3. The adversary starts by replaying past data from y_2 at $t \approx 90$ s and then begins stealing water from tank 4 at $t \approx 100$ s. Tank 4 is successfully emptied and the adversary stops removing water at $t \approx 180$ s. To ensure stealthiness, the replay attack continues until the system recovered its original setpoint at $t \approx 280$ s. Note that the attack is not detected, since the residue stays below the alarm threshold.

5.4. Zero-dynamics attack

The QTP has a non-minimum phase configuration in which the plant possesses an unstable zero. In this case, as discussed in Section 4.4, an adversary able to corrupt all the actuator channels may launch a false-data injection attack where the false-data follows the zero-dynamics. Moreover, since the safe region is described by the set $\mathcal{S}_x = \{x \in \mathbb{R}^{n_x} : \ \|x - \sigma\mathbf{1}\|_\infty \leq 15,\ \sigma = 15\}$, from Theorem 7 we expect that the zero-dynamics attack associated with the unstable zero can drive the system to an unsafe region. This scenario is illustrated in Fig. 9.

The adversary's goal is to either empty or overflow at least one of the tanks, considered as an unsafe state. The attack on both actuators begins at $t \approx 30$ s, causing a slight increase in the residual. Tank 3 becomes empty at $t \approx 55$ s and shortly after actuator 2 saturates, producing a steep increase in the residual which then crosses the threshold. However, note that the residual was below the threshold when the unsafe state was reached.

Note that the system dynamics change after saturation of the water levels and actuators and, consequently, the attack signal

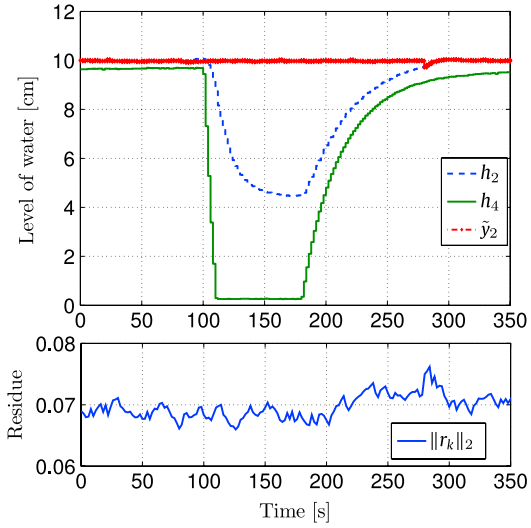


Fig. 8. Results for the replay attack performed against sensor 2 from $t \approx 90$ s to $t \approx 280$ s. The adversary opens the tap of tank 4 at $t \approx 100$ s and closes it at $t \approx 180$ s.

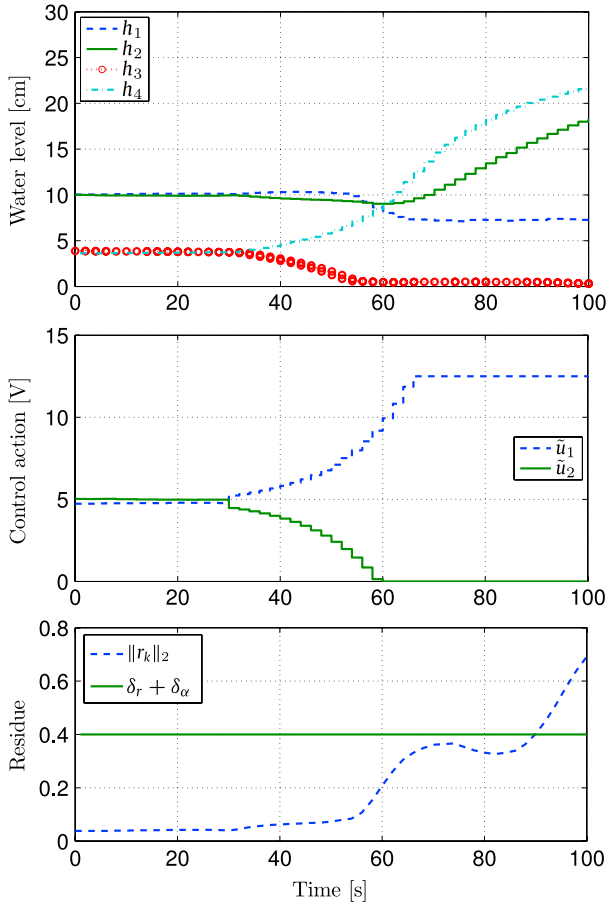


Fig. 9. Results for the zero-dynamics attack starting at $t \approx 30$ s. Tank 3 is emptied at $t \approx 55$ s, resulting in a steep increase in the residual since the linearized model is no longer valid.

no longer corresponds to the zero-dynamics and is detected. However, the attack has already damaged the system before being detected. Therefore, these attacks are particularly dangerous in processes that have unstable zero-dynamics and in which the actuators are over-dimensioned, allowing the adversary to perform longer attacks before saturating.

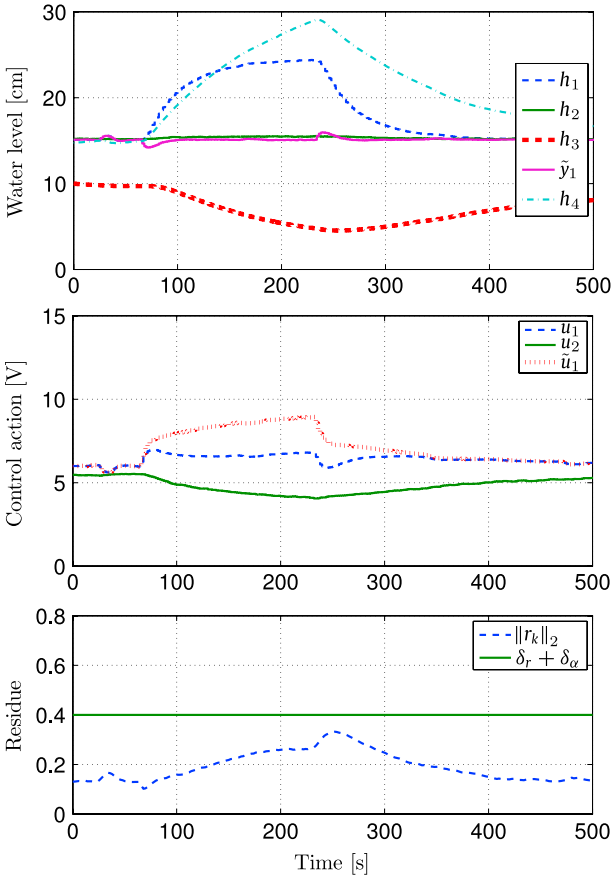


Fig. 10. Results for the bias attack against the actuator 1 and sensor 1 in the minimum phase QTP. The attack is launched using a low-pass filter in the instant $t \approx 70$ s and stopped at $t \approx 230$ s.

5.5. Bias injection attack

The results for the case where u_1 and y_1 are respectively corrupted with b^u_∞ and b^y_∞ are presented in Fig. 10. In this scenario, the adversary aimed at driving the system out of the safe set $\mathcal{X}$ while remaining stealthy for $\delta_\alpha = 0.25$. The bias was slowly injected using a first-order low-pass filter with $\beta = 0.95$ and the following steady-state value, computed using Theorem 13, $a_\infty = [b^u_\infty \quad b^y_\infty]^\top = [2.15 \quad -9.42]^\top$.

The bias injection began at $t \approx 70$ s and led to an overflow in tank 4 at $t \approx 225$ s. At that point, the adversary started removing the bias and the system recovered the original setpoint at $t \approx 350$ s. The residual remained within the allowable bounds throughout the attack, thus the attack was not detected.

6. Conclusions and future work

In this paper we have analyzed the security of networked control systems. A novel attack space based on the adversary's model knowledge, disclosure, and disruption resources was proposed and the corresponding adversary model described. Attack scenarios corresponding to DoS, replay, zero-dynamics, and bias injection attacks were analyzed using this framework. In particular the maximum impact of stealthy bias injection attacks was derived and it was shown that the corresponding policy does not require perfect model knowledge. These attack scenarios were illustrated using an experimental setup based on a quadruple-tank process controlled over a wireless network.

Future research directions include the extension of the framework to other classes of systems, e.g., nonlinear systems. Analyzing

attack scenarios under non-ideal communication network models and considering closed-loop attack policies are also relevant research directions.

References

Amin, S., Cárdenas, A. A., & Sastry, S. S. (2009). Safe and secure networked control systems under denial-of-service attacks. In *Lecture notes in computer science, Hybrid systems: computation and control* (pp. 31–45). Berlin, Heidelberg: Springer.

Amin, S., Litrico, X., Sastry, S. S., & Bayen, A. M. (2010). Stealthy deception attacks on water SCADA systems. In *Proc. of the 13th ACM int. conf. on hybrid systems: computation and control*, HSCC'10. New York, NY, USA: ACM.

Bishop, M. (2002). *Computer security: art and science*. Addison-Wesley Professional.

Boyd, S., El Ghaoui, L., Feron, E., & Balakrishnan, V. (1994). *Studies in applied mathematics*: *Vol. 15. Linear matrix inequalities in system and control theory*. Philadelphia, PA: SIAM.

Cárdenas, A., Amin, S., Lin, Z., Huang, Y., Huang, C., & Sastry, S. (2011). Attacks against process control systems: risk assessment, detection, and response. In *Proceedings of the 6th ACM symposium on information, computer and communications security*, ASIACCS'11. (pp. 355–366). New York, NY, USA: ACM.

Cárdenas, A.A., Amin, S., & Sastry, S.S. (2008). Research challenges for the security of control systems. In *Proc. 3rd USENIX workshop on hot topics in security, San Jose, CA, USA, July*.

Chen, J., & Patton, R. J. (1999). *Robust model-based fault diagnosis for dynamic systems*. Kluwer Academic Publishers.

Ding, S. X. (2008). *Model-based fault diagnosis techniques: design schemes*. Springer Verlag.

Esfahani, P., Vrakopoulou, M., Margellos, K., Lygeros, J., & Andersson, G. (2010). Cyber attack in a two-area power system: Impact identification using reachability. In *American control conference, 2010, July* (pp. 962–967).

Falliere, N., Murchu, L., & Chien, E. (2011). W32.Stuxnet dossier, February.

Frank, P. M., & Ding, X. (1997). Survey of robust residual generation and evaluation methods in observer-based fault detection systems. *Journal of Process Control*, *7*(6), 403–424.

Giani, A., Sastry, S., Johansson, K.H., & Sandberg, H. (2009). The VIKING project: an initiative on resilient control of power networks. In *Proc. 2nd int. symp. on resilient control systems, Idaho Falls, ID, USA, August*.

Gorman, S. (2009). Electricity grid in US penetrated by spies. *The Wall Street Journal*, A1.

Gupta, A., Langbort, C., & Başar, T. (2010). Optimal control in the presence of an intelligent jammer with limited actions. In *Proc. of the 49th IEEE conf. on decision and control, Atlanta, GA, USA, December*.

Hiriart-Urruty, J. B. (2001). Global optimality conditions in maximizing a convex quadratic function under convex quadratic constraints. *Journal of Global Optimization*, *21*(4), 443–453.

Hwang, I., Kim, S., Kim, Y., & Seah, C. E. (2010). A survey of fault detection, isolation, and reconfiguration methods. *IEEE Transactions on Control Systems Technology*, *18*(3), 636–653.

Isermann, R. (2006). *Fault-diagnosis systems: an introduction from fault detection to fault tolerance*. Springer.

Isidori, A. (1995). *Nonlinear control systems* (3rd ed.). Springer-Verlag.

Johansson, K. H. (2000). The quadruple-tank process: a multivariable laboratory process with an adjustable zero. *IEEE Transactions on Control Systems Technology*, *8*(3), 456–465.

Khanafer, A., Touri, B., & Başar, T. (2012). Consensus in the presence of an adversary. In *Proc. 3rd IFAC workshop on estimation and control of networked systems, NecSys'12, Santa Barbara, CA, USA, September*.

Kosut, O., Jia, L., Thomas, R., & Tong, L. (2010). Malicious data attacks on smart grid state estimation: attack strategies and countermeasures. In *Proceedings of the first IEEE international conference on smart grid communications, Gaithersburg, MD, USA, October*.

Liu, Y., Reiter, M. K., & Ning, P. (2009). False data injection attacks against state estimation in electric power grids. In *Proc. 16th ACM conf. on computer and communications security, Chicago, IL, USA, November*.

Mo, Y., & Sinopoli, B. (2009). Secure control against replay attack. In *Proceedings of the 47th annual Allerton conference on communication, control, and computing, Allerton, IL, USA, October*.

Mo, Y., & Sinopoli, B. (2012). Integrity attacks on cyber-physical systems. In *Proc. 1st international conference on high confidence networked systems, CPSWeek 2012, Beijing, China* (pp. 47–54).

Pang, Z.-H., & Liu, G.-P. (2012). Design and implementation of secure networked predictive control systems under deception attacks. *IEEE Transactions on Control Systems Technology*, *20*(5), 1334–1342.

Pasqualetti, F., Dorfler, F., & Bullo, F. (2011). 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, December*.

Rid, T. (2011). Cyber war will not take place. *Journal of Strategic Studies*, *35*(1), 5–32.

Sandberg, H., Teixeira, A., & Johansson, K.H. (2010). On security indices for state estimators in power networks. In *Preprints of the first workshop on secure control systems, CPSWEEK 2010, Stockholm, Sweden, April*.

Schenato, L. (2009). To zero or to hold control inputs with lossy links? *IEEE Transactions on Automatic Control*, *54*(5), 1093–1099.

Schenato, L., Sinopoli, B., Franceschetti, M., Poolla, K., & Sastry, S. (2007). Foundations of control and estimation over lossy networks. *Proceedings of the IEEE*, *95*(1), 163–187.

Smith, R. (2011). A decoupled feedback structure for covertly appropriating networked control systems. In *Proc. of the 18th IFAC world congress, Milano, Italy, August–September*.

Sridhar, S., Hahn, A., & Govindarasu, M. (2012). Cyber-physical system security for the electric power grid. *Proceedings of the IEEE*, *100*(1), 210–224.

Sundaram, S., Revzen, S., & Pappas, G. (2012). A control-theoretic approach to disseminating values and overcoming malicious links in wireless networks. *Automatica*, *48*(11), 2894–2901.

Teixeira, A., Dán, G., Sandberg, H., & Johansson, K.H. (2011). Cyber security study of a SCADA energy management system: stealthy deception attacks on the state estimator. In *Proc. of the 18th IFAC world congress, Milano, Italy, August–September*.

Teixeira, A., Pérez, D., Sandberg, H., & Johansson, K. H. (2012). Attack models and scenarios for networked control systems. In *Proc. 1st international conference on high confidence networked systems, CPSWeek 2012, Beijing, China*.

Teixeira, A., Sandberg, H., Dán, G., & Johansson, K. H. (2012). Optimal power flow: closing the loop over corrupted data. In *Proc. American control conference, Montreal, Canada, June*.

Teixeira, A., Shames, I., Sandberg, H., & Johansson, K. H. (2012). Revealing stealthy attacks in control systems. In *Proceedings of the 50th annual Allerton conference on communication, control, and computing, Allerton, IL, USA, October*.

Xie, L., Mo, Y., & Sinopoli, B. (2010). False data injection attacks in electricity markets. In *Proceedings of the first IEEE international conference on smart grid communications, Gaithersburg, MD, USA, October*.

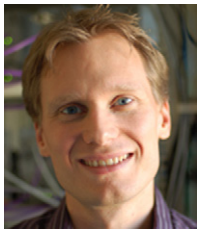
Zhang, W., Branicky, M. S., & Phillips, S. M. (2001). Stability of networked control systems. *IEEE Control Systems Magazine*, *21*, 84–99.

Zhang, X., Polycarpou, M., & Parisini, T. (2002). A robust detection and isolation scheme for abrupt and incipient faults in nonlinear systems. *IEEE Transactions on Automatic Control*, *47*(4), 576–593.

Zhou, K., Doyle, J. C., & Glover, K. (1996). *Robust and optimal control*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc..

André Teixeira is a Ph.D. candidate in Automatic Control at KTH Royal Institute of Technology, Stockholm, Sweden. He received the M.Sc. degree in Electrical and Computers Engineering in 2009 from the Faculdade de Engenharia da Universidade do Porto, Porto, Portugal. He was a finalist for the NecSys 2012 Best Student Paper Award and one of his publications is listed in ACM Computing Review's Notable Computing Books and Articles of 2012. His main research interests include secure control, distributed fault detection and isolation, distributed optimization, power systems, and multi-agent systems.

Iman Shames is a McKenzie fellow at the department of Electrical and Electronic Engineering, the University of Melbourne. Previously, he was an ACCESS Postdoctoral Researcher at the ACCESS Linnaeus Centre, the KTH Royal Institute of Technology, Stockholm, Sweden. He received his B.S. degree in Electrical Engineering from Shiraz University, Iran in 2006, and the Ph.D. degree in Engineering and Computer Science from the Australian National University, Canberra, Australia in 2011. His current research interests include optimization, sensor networks, distributed fault detection and isolation, and security in networked systems.



Henrik Sandberg 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. He is an Associate Professor with the Department of Automatic Control, KTH Royal Institute of Technology, Stockholm, Sweden. From 2005 to 2007, he was a Post-Doctoral Scholar with 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 with the Australian National University and the University of Melbourne, Australia. His current research interests include secure networked control, power systems, model reduction, and fundamental limitations in control. He was a recipient of the Best Student Paper Award from the IEEE Conference on Decision and Control in 2004 and an Ingvar Carlsson Award from the Swedish Foundation for Strategic Research in 2007. He is currently an Associate Editor of the IFAC Journal Automatica.

Karl Henrik Johansson is Director of the KTH ACCESS Linnaeus Centre and Professor at the School of Electrical Engineering, Royal Institute of Technology, Sweden. He is a Wallenberg Scholar and has held a six-year Senior Researcher Position with the Swedish Research Council. He is Director of the Stockholm Strategic Research Area ICT The Next Generation. He received M.Sc. and Ph.D. degrees in Electrical Engineering from Lund University. He has held visiting positions at UC Berkeley (1998–2000) and California Institute of Technology (2006–2007). His research interests are in networked control systems, hybrid and embedded system, and applications in transportation, energy, and automation systems. He has been a member of the IEEE Control Systems Society Board of Governors and the Chair of the IFAC Technical Committee on Networked Systems. He has been on the Editorial Boards of several journals, including Automatica, IEEE Transactions on Automatic Control, and IET Control Theory and Applications. He is currently on the Editorial Board of IEEE Transactions on Control of Network Systems and the European Journal of Control. He has been Guest Editor for special issues, including the one on "Wireless Sensor and Actuator Networks" of IEEE Transactions on Automatic Control 2011. He was the General Chair of the ACM/IEEE Cyber–Physical Systems Week 2010 in Stockholm and IPC Chair of many conferences. He has served on the Executive Committees of several European research projects in the area of networked embedded systems. In 2009, he received the Best Paper Award of the IEEE International Conference on Mobile Ad-hoc and Sensor Systems. In 2009, he was also awarded Wallenberg Scholar, as one of the first ten scholars from all sciences, by the Knut and Alice Wallenberg Foundation. He was awarded an Individual Grant for the Advancement of Research Leaders from the Swedish Foundation for Strategic Research in 2005. He received the triennial Young Author Prize from IFAC in 1996 and the Peccei Award from the International Institute of System Analysis, Austria, in 1993. He received Young Researcher Awards from Scania in 1996 and from Ericsson in 1998 and 1999. He is a Fellow of the IEEE.