# Cyber-Attack Detection and Mitigation in Networked Control Systems

Mohsen Ghaderi

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of The Requirements

for The Degree of

Master of Applied Science (Electrical and Computer Engineering)

Concordia University

Montréal, Québec, Canada

November 2019

**CONCORDIA UNIVERSITY**
**SCHOOL OF GRADUATE STUDIES**

This is to certify that the thesis prepared

By: Mohesen Ghaderi

Entitled: Cyber-Attack Detection and Mitigation in Networked Control Systems

and submitted in partial fulfillment of the requirements for the degree of

**Master of Applied Science (Electrical and Computer Engineering)**

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. A.G. Aghdam

_____ External Examiner
Dr. Y. Zhang (MIAE)

_____ Internal Examiner
Dr. A.G. Aghdam

_____ Supervisor
Dr. W. Lucia

Approved by: _____
Dr. Y.R. Shayan, Chair
Department of Electrical and Computer Engineering

_____ 20___ _____
Dr. Amir Asif, Dean,
Faculty of Engineering and Computer Science

# Abstract

Cyber-Attack Detection and Mitigation in Networked Control Systems

Mohsen Ghaderi

Cyber-Physical System (CPS) is the term used to describe the physical systems equipped with computation and communication capabilities. CPSs can be used in different applications e.g. autonomous vehicles, water distribution systems, smart grids, industry 4.0 and Internet of Things (IoT). CPSs have expectation of improving the capability of traditional engineering system but on the other hand, they arise several concerns about their security against cyber-attacks. In the last decade, several cyber-attacks targeting SCADA systems have been reported, see e.g. Maroochy water breach and the Stuxnet worm aimed Iran's nuclear facility. From a control point of view, a CPS can be interpreted as a Networked Control System (NCS) where the risk of cyber-attacks can be modeled as the possibility that malicious agents could compromise the communication channels. In order to benefit from CPSs, specially in safety critical systems, their vulnerabilities to cyber-attacks must be properly faced. In this thesis two control architectures for CPS are developed. In the first, starting from the analysis of active detection mechanisms available in the literature, we propose a novel architecture capable of detecting a broad class of False Data Injection (FDI) attacks. Such strategy has been contrasted with the well-known watermarking detection mechanism and it is shown that our solution is capable of detecting replay attacks without degrading the closed-loop performance of the system. Moreover, it is shown that compared to detection schemes resorting to auxiliary systems, the proposed strategy is less involved and of easier implementation. In particular, it can be installed on the existing NCS infrastructure without changing communications, controller or state estimator. In the second architecture, we propose another novel architecture capable of detecting and mitigating a broad class of FDI attacks. First, we propose a detection mechanism based on a coding scheme to limit the attacker's disclosure and disruptive resources and prevent

the existence of stealthy attacks. Second, we propose an emergency local controller that is activated when an attack is detected or the plant's safety is in danger. It is proved that the proposed architecture always guarantees the safety of the system, regardless of the attack actions and detector performance. Moreover, plant's normal operation recovery is ensured once the attack is terminated.

# To my Parents

# Acknowledgments

I would like to express my gratitude to my supervisor Dr. Walter Lucia for his immense help and support throughout my master's degree. This thesis would not be possible without his continuous guidance and insightful comments. I am grateful to have such a knowledgeable, considerate and outstanding supervisor.

I am very grateful to my supervisor and Concordia University for the financial support that I received, which was very crucial to complete this research work. I would like to acknowledge the financial support from the School of Graduate Studies, GSA, ECSGA at Concordia University and EUCA Student Travel Support to support the costs related to presentation of this work in European Control Conference (ECC 2019). I acknowledge the support and help offered to me by the Concordia Institute for Information Systems Engineering (CIISE).

I wish to thank my friends and colleagues at Concordia University: Hamid Nabati, Dr. Omid Saatlou, Maryam Bagherzadeh, Kian Gheitasi, Shima Savehshemshaki, Amirreza Mousavi, Ehsan Agah, Sepehr Radmannia, Flavia Grandinetti, Antonello Venturino, Rezvan Nozari and Farnaz Yarali who walked by my side during the last two years, shared my moments of distress and joy, and made this period the most memorable one in my life.

Last but not the least, I would like to thank my beloved parents, my brother and his wife for their unconditional love and inspiration throughout my life. None of this would have been possible without their encouragement.

# Contents

# List of Figures

# List of Tables

# List of Symbols

| | |
|---|---|
| $\mathbf{O^T}$: | Transpose of $\mathbf{O}$ |
| $\mathbf{O}^{-1}$: | Inverse of matrix $\mathbf{O}$ |
| $\mathbf{O_k}$: | $\mathbf{O}$ at time step $\mathbf{k}$ |
| $\mathbf{x}$: | State vector |
| $\mathbf{u}$: | Input vector |
| $\mathbf{y}$: | Sensor measurement vector |
| $\mathbf{r}$: | Reference signal vector |
| $\hat{\mathbf{x}}_\mathbf{k}$: | Estimated value of $\mathbf{x}$ at time $\mathbf{k}$ |
| $\mathbf{A}$: | State matrix |
| $\mathbf{B}$: | Input matrix |
| $\mathbf{C}$: | Output matrix |
| $\mathbf{D}$: | Direct transition matrix |
| $\mathcal{N}(.,.)$: | Normal distribution |
| $\mathcal{X}$: | State constraints set |
| $\mathcal{U}$: | Input constraints set |
| $\mathbb{Z}_+$: | Non-negative integers $\{0, 1, ...\}$ |
| $\mathbf{e}$: | Estimation error |
| $\mathbf{u^a}$: | Attack injected input vector |
| $\mathbf{y^a}$: | Attack injected measurement vector |

# List of Abbreviations

| | |
|---|---|
| CPS: | Cyber-Physical Systems |
| IoT: | Internet of Things |
| CIA: | Confidentiality, Integrity or Availability |
| SCADA: | Supervisory Control and Data Acquisition |
| FDI: | False Data Injection |
| DoS: | Denial of Service |
| LTI: | Linear Time-Invariant |
| IID: | Independently and Identically Distributed |
| NCS: | Networked Control Systems |
| DoA: | Domain of Attraction |
| MPC: | Model Predictive Controller |
| RPI: | Robust Positive Invariant |
| RCI: | Robust Control Invariant |
| LQ : | Linear Quadratic |
| E_STC: | Emergency Set-Theoretic Controller |
| SG : | Safety Guard |
| UUB: | Uniformly Ultimately Bounded |

# Chapter 1

# Introduction

Cyber-Physical System is the term used to describe the physical systems equipped with computation and communication capabilities. In Fig. 1 some of the applications of CPS e.g. autonomous vehicles, water distribution systems, smart grids, industry 4.0 and Internet of Things (IoT) are shown. CPSs have expectation of improving the capability of traditional engineering system but on the other hand, they arise several concerns about their security against cyber-attacks. In the last decade, several cyber-attacks targeting SCADA systems have been reported [1], see e.g. Maroochy water breach [2] and the Stuxnet worm aimed Iran's nuclear facility [3]. Therefore, such a systems must be properly controlled and protected [4–6].

From a control point of view, a CPS can be interpreted as a networked control system where the risk of cyber-attacks can be modeled as the possibility that malicious agents could compromise the communication channels, see Fig. 2. In order to benefit from CPS, specially in safety critical systems, its vulnerability to cyber-attacks must be properly faced.

In order to have secure control systems, attack-detection and attack-mitigation techniques are needed. For attack detection, several solutions have been proposed in the literature, see e.g. the survey paper [7]. Nevertheless, most of them target specific attack-scenarios and no solutions are capable of detecting a broad class of attacks. On the other

hand, very few attack mitigation and recovery strategy have been proposed. Generally speaking, attack mitigation is considered a very challenging task.

The above considerations, motivate the work in this thesis where novel detection and mitigation strategies are proposed.

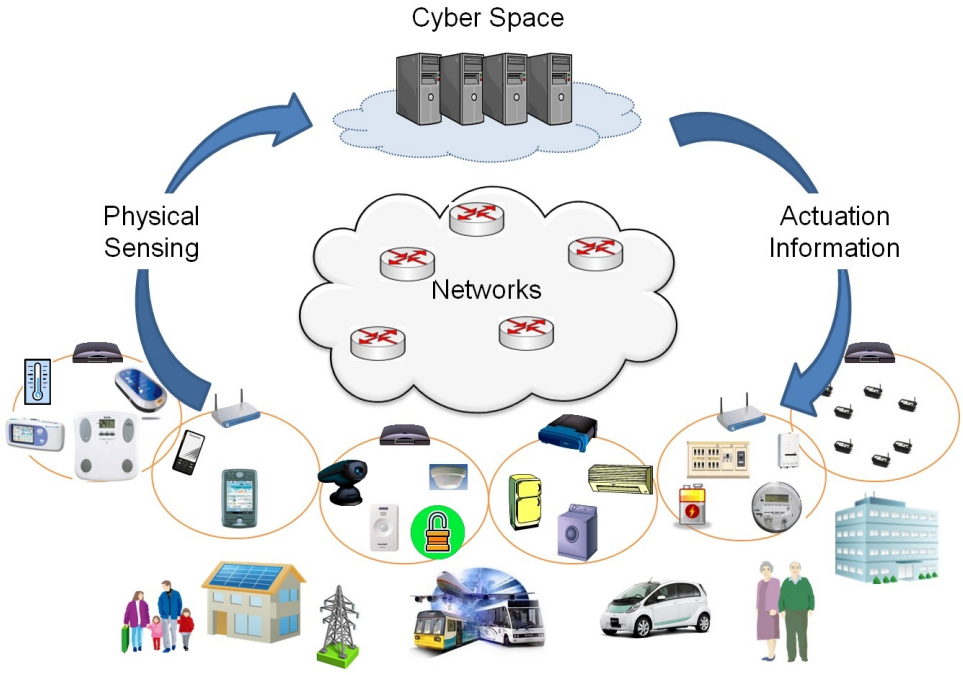



Figure 1: CPS Application [1]

In networked control systems, the controller is located remotely with respect to plant so communication links are needed to connect them. In addition, the reference signal might be locally available to the controller or might be sent through a network from another control center. These networks might be prone to cyber-attacks and an adversary might launch different attack scenarios by violating the Confidentiality, Integrity or Availability (CIA) properties of the communication channels.

[1]The figure is taken from https://devicesmart.wordpress.com/tag/cyber-physical-systems/

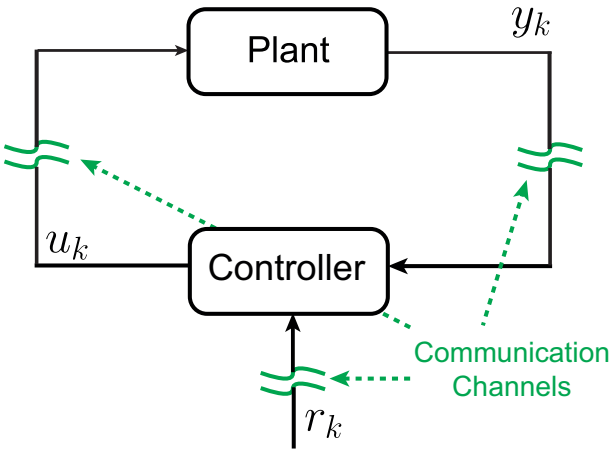


Figure 2: Networked Control Architectures

## 1.1 Literature Review

In this thesis, we focus on the class of FDI attacks which alter the data transmitted through communication channels. Different FDI attacks can be performed according to the attacker's available resources. A complete taxonomy of existing FDI attacks can be found in [8] where a 3D classification is reported. Notable example are replay, zero-dynamics and covert attacks.

Most of the existing strategies on detection of FDIs have focused on attack on the sensor or actuation channels, see e.g. [9–18] and references therein. Detection mechanism are usually classified in passive and active strategies. In [10], passive static and dynamic detectors are proposed to deal with FDIs affecting smart grid systems. In [17], a Bayesian approach based on random sets is deployed against switching signals and fake measurements attacks to ensure resilient state estimation. In [9], the class of attacks undetectable by any passive detector has been defined.

Due to the existence of stealthy attacks against passive detectors, active detection methods have been proposed in the literature. In [11], a watermarked input signal has been proposed to actively detect steady-state replay attacks. [10] has studied the effects of imposing the watermarking signal on the closed-loop performance. In [12], a stochastic

game approach to compensate the effect of watermarking signal and to reach a trade-off between detection rate and performance degradation has been proposed. In [13], a coding scheme for sensor measurements has been introduced to detect stealthy sensor false data injection attacks.

Although active detection methods presented above outperform passive detectors [10], there is still the possibility of performing stealthy FDI attacks, see e.g. covert [19] and zero-dynamics [14] attacks. In [14], detecting zero-dynamics attack is solved by properly modifying the system's structural matrices. In [19], the author shows that it is possible to design an intelligent attack, namely covert attack, that is undetectable regardless of the used remote controller and detector. In [18], the covert attack detection problem is solved by adding a modulation matrix in the input channel. Such solution is proved to be also effective against zero-dynamics attacks. In [15], an auxiliary time-varying system, coupled with the plant dynamics, is designed to avoid the existence of covert attacks. A similar idea is pursued in [16] where the auxiliary dynamics are designed to resemble the dynamics of the plant. In both [15, 16], the added dynamics are changed randomly to prevent attackers from estimating their behavior.

As long as the resilience of CPS to cyber attack is concerned, several methods have been proposed to perform resilient state estimation. In the literature, several methods are proposed to perform resilient state estimation in presence of malicious agents, see in this regards [10, 20–23] and references therein. In [10], by assuming an upper bound on cardinality of attack vector, robust estimators are designed to detect the presence of faults/attacks in systems. In [20], the authors present a novel algorithm that uses a satisfiability modulo theory approach to harness estimation complexity. In [21], the authors proposed a resilient state estimator capable of reconstructing the state vector if at least half of the measurements are not under attack. In [22], the attack resilient state estimation is addressed in presence of the bounded-size noise and a $l_0$-based state estimator is designed. In [23], a security-oriented cyber-physical state estimation system for smart

grid systems is presented.

Finally, of particular interest for this thesis are the control solutions in [24, 25] where both attack detection and countermeasures have been jointly proposed. In [24], a set-theoretic control strategy is introduced to guarantee the safety of system until the communication channels are reestablished after attack detection. In [25], an adaptive control strategy is proposed to detect and mitigate FDI attacks.

## 1.2 Thesis Motivations and Contributions

From the state-of-art in cyber-attack detection and compensation problem, it is possible to appreciate that there does not exist a single solution capable of detecting multiple advanced FDI attacks such as replay, zero-dynamics and covert attacks. In particular, each solution in [11,13,14] only deal with a single category of FDI attacks. Moreover, some of the existing solutions [10–12] achieve detection by degrading the closed-loop system performance or with detection mechanisms which are too involved, see e.g. [15, 16]. Also, there are few solutions for recovering normal behavior of the control system after the detection task is completed.

In this thesis, starting from the existing solutions [11, 13, 15, 16], we propose novel control architectures capable of detecting the above mentioned classes of undetectable attacks. Different from the existing literature, the proposed solutions do not affect the system performance and they can be easier deployed on the existing networked control architecture. Moreover, we propose novel attack countermeasures capable of ensuring satefy of the system during the attack and perform recovery when the attack is terminated. In this regard, this thesis introduces the concept of one-step attack safe region, that is the state space region where the plant safety is assured for at least one step regardless of any admissible attack action.

## 1.3 Thesis Layout

In chapter 2, the main concepts and definitions used along this thesis are presented. In chapter 3, an architecture to detect FDI attacks in NCS is proposed. In chapter 4, another control architecture equipped with compensation action is proposed. Finally, chapter 5 concludes the thesis and highlights future research directions.

## 1.4 Publications

- [26] M. Ghaderi, K. Gheitasi, and W. Lucia. "A Novel Control Architecture for the Detection of False Data Injection Attacks in Networked Control Systems." In American Control Conference (ACC), pp. 139-144, 2019.

- [27] K. Gheitasi, M. Ghaderi, and W. Lucia. "A Novel Networked Control Scheme with Safety Guarantees for Detection and Mitigation of Cyber-Attacks." In European Control Conference (ECC), pp. 1449-1454, 2019.

- [28] W. Lucia, K. Gheitasi, and M. Ghaderi. "A Command Governor Based Approach for Detection of Setpoint Attacks in Constrained Cyber-Physical Systems." In IEEE Conference on Decision and Control (CDC), pp. 4529-4534, 2018.

# Chapter 2

# Preliminaries and Definitions

## 2.1 Standard Networked Control Architecture

In a networked control architecture, the controller and the plant are spatially distributed and they are connected through a communication network as shown in Fig. 3. This architecture consists of four main ingredients: (I) Plant (II) Controller (III) State Estimator (IV) Anomaly/Attack Detector which are described as follows:

### 2.1.1 Plant

In this manuscript, we assume that the physical system is modeled as a discrete-time Linear Time Invariant (LTI) system and is represented in the state-space as follows:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + \omega_k \\ y_k &= Cx_k + \eta_k \end{aligned} \tag{1}$$

where $k \in \mathbb{Z}_+ := \{0, 1, ...\}$ is the sampling time instant, $x_k \in \mathbb{R}^n$ is the vector of the states of the plant, $u_k \in \mathbb{R}^m$ is the vector of control commands and $y_k \in \mathbb{R}^p$ is the vector of sensor measurements. A, B and C are assumed to be time independent matrices with compatible dimensions. $\omega_k$ is an Independently and Identically Distributed(IID) process noise distributed normally with zero mean, namely $\omega_k \sim \mathcal{N}(0, \mathcal{Q})$ and $\mathcal{Q}$ is positive

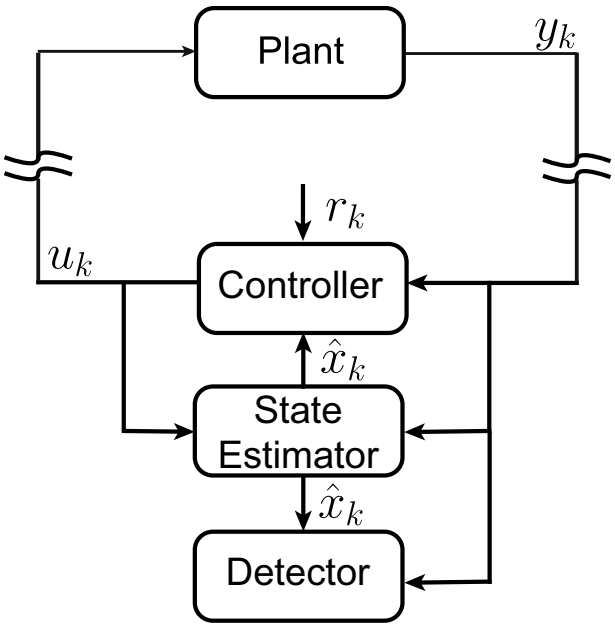


Figure 3: Networked Control System

definite matrix. $\eta_k$ is an Independently and Identically Distributed(IID) measurement noise distributed normally with zero mean, namely $\eta_k \sim \mathcal{N}(0, \mathcal{R})$ and $\mathcal{R}$ is positive definite matrix.

**Assumption 1.** *Given the plant model (1), it is assumed the pairs (A,C) and (A,B) are detectable and stabilizable, respectively.*

The system (1) can be subject to state or input constraints.

$$x_k \in \mathcal{X}, u_k \in \mathcal{U} \tag{2}$$

where $\mathcal{X}$ and $\mathcal{U}$ are compact subset of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, with $0_n \in \mathcal{X}$ and $0_m \in \mathcal{U}$.

### 2.1.2 Controller

A state-feedback controller is responsible to satisfy the plant constraints (2) and ensures tracking of the reference signal $r_k$ in absence of attacks. The controller working region will be hereafter denoted as the Domain of Attraction (DoA), namely $\mathcal{X}_\eta \subset \mathcal{X}$.

### 2.1.3 State Estimator

In order to design a state-feedback controller, the state of the system is required. When this is not entirely measurable, a state estimator is used to reconstruct the state from the available sensor measurements and input signals [29]. In order to estimate the states of the system, the following Kalman Filter is used:

$$\hat{x}_k = A\hat{x}_{k-1} + Bu_{k-1} + L_K(y_{k-1} - C\hat{x}_{k-1}) \tag{3}$$

where $y_k \in \mathbb{R}^p$ is the measurement vector received on the controller side, $\hat{x}_k$ is the estimated state. By defining the estimation error as $e_k := x_k - \hat{x}_k$, in the Kalman filter, the gain matrix $L_K$ is designed to minimize the covariance matrix $P_k := E[e_k e_k^T]$ in the absence of attack. If the pair (A,C) is detectable (see assumption 1), then the covariance matrix $P_k$ converge to a steady-state solution $P$ and the steady-state kalman gain is:

$$L_K := APC^T(CPC^T + \mathcal{R})^{-1} \tag{4}$$

where P is the only positive semi-definite solution of the following Riccati equation

$$P = APA^T + \mathcal{Q} - APC^T(CPC^T + \mathcal{R})^{-1}CPA^T \tag{5}$$

### 2.1.4 Anomaly Detector

Given the Kalman Filter introduced in (3), the residual signal is defined as follows:

$$r_k = y_k - C\hat{x}_k = Ce_k + \eta_k \tag{6}$$

which evolves according to the following equation

$$\begin{cases} e_{k+1} = (A - L_K C)e_k + \omega_k - L_K \eta_k \\ r_k = Ce_k + \eta_k \end{cases} \tag{7}$$

In attack-free condition, the mean of the residual signal is

$$E[r_k] = CE[e_k] + E[\eta_k] = 0_{p\times 1} \tag{8}$$

and the covariance is [30]:

$$\Sigma = E[r_k r_k^T] = CPC^T + \mathcal{R} \tag{9}$$

Such signal, can be exploited to detect the presence of cyber attacks in the communication channels. In particular, the following binary hypotesis test can be defined:

$$\mathcal{H}_0 : \begin{cases} E[r_k] = 0_{p\times 1}, \\ E[r_k r_k^T] = \Sigma, \end{cases} \qquad \mathcal{H}_1 : \begin{cases} E[r_k] \neq 0_{p\times 1}, \\ E[r_k r_k^T] \neq \Sigma, \end{cases} \tag{10}$$

where hypothesis $\mathcal{H}_0$ denotes the normal mode and hypothesis $\mathcal{H}_1$ indicates anomaly/attack mode. Such test in the literature [31,32] is often approximated by means of a $\chi^2$ test as the following distance measure:

$$z_k = \sum_{i=k-\mathcal{J}+1}^{k} r_i^T \Sigma^{-1} r_i \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \beta \tag{11}$$

where $\mathcal{J}$ is the length of the detection window, and $z_k$ is the scalar value resulted by $\chi^2$ test. $\beta > 0$ is the threshold value which is chosen according to the desired false alarm rate [33].

## 2.2 Attack Classification

In the networked control system shown in Fig. 4, we assume that the communication channels between the controller and the plant are insecure. Therefore a malicious agent can alter the closed-loop evolution of (1).

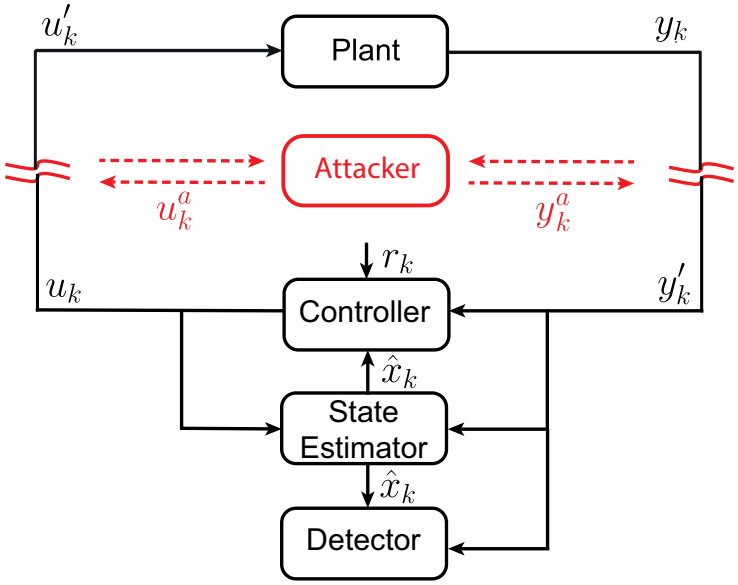



Figure 4: Networked Control System under attack

**Definition 1.** *(**Secure and insecure channel**) A communication channel is considered secure if the Confidentiality, Integrity, and Availability properties (CIA triad) are satisfied [34]. A channel is insecure if at least one of the CIA properties is not met.* □

This ability of an attacker to disrupt the plant operations depends on the available set of disruptive and disclosure resources. Moreover, the capability of the attacker to perform sophisticated attacks also depends on the available information on the closed-loop system operations, namely plant model, controller and detector.

**Definition 2.** *[8] (**Attacker's resources**) Let us consider an insecure communication channel, namely channel $-\ i$, where the data packet $h_k \in \mathbb{R}^{n_h}$ is transmitted at each sampling time $k \in \mathbb{Z}_+$.*

- *Disclosure Resources: An attacker has disclosure resource on the channel $-$ $i$ if he/she can violate the confidentiality property, i.e. intercept/read the vector $h_k$.*

- *Disruptive Resources: An attacker has disruptive resources on the channel $-$ $i$ if he/she can violate the authentication or integrity properties, i.e. the attacker can arbitrary change the transmitted vector $h_k$ into a new compatible vector $h'_k \in \mathbb{R}^{n_h}$ .*

- *Model Knowledge: An attacker has model knowledge when the attacker has a subset $\mathcal{I}_{attacker}$ of the information characterizing the closed loop evolution of the system, i.e.*

$$\mathcal{I}_{attacker} \subseteq \{A, B, C, D, \mathcal{X}, \mathcal{U}, f(\cdot, \cdot)\} \tag{12}$$

□

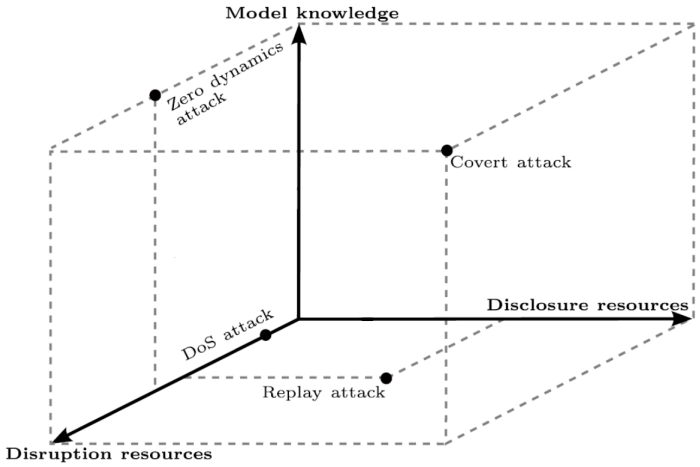



Figure 5: The attack-space in Cyber-Physical Systems [8]

Model knowledge, disclosure and disruptive resources are the basis to shape the attack space. This is well-shown in Fig. 5 where the different attacks are shown with respect to their required resources. Here, a more formal definition of FDI attacks is provided.

**Definition 3.** *(**False data injection attacks**) Let us consider an insecure communication channel where data packet $h_k \in \mathbb{R}^{n_h}$ is transmitted and the attacker has disruptive*

*resources. An FDI attack is a deception attack [35], [10] where the attacker modifies the vector $h_k$ by either injecting an arbitrary vector $h_k^a$ (Integrity Attack) or by substituting $h_k$ with a fake unstructured vector $h_k^a$ (Substitution Attack) [34], and in general both attacks can be modeled as additive data injection i.e.*

$$h'_k = h_k + h_k^a \tag{13}$$

*with $h'_k \in \mathbb{R}^{n_h}$ denoting the resulting corrupted vector.* □

### 2.2.1 Attack Model

In this subsection, FDI attacks on the sensor and actuator channels are described. By resorting to Definition 3 and the standard networked architecture in Fig. 4, we model networked FDI attacks as follows:

$$\begin{aligned} u'_k &:= u_k + u_k^a \\ y'_k &:= y_k + y_k^a \end{aligned} \tag{14}$$

where $u_k^a \in \mathbb{R}^m$ and $y_k^a \in \mathbb{R}^p$ are vectors injected by the attackers on the actuation and measurement channels, respectively.

**Definition 4.** *[8] (**Stealthy FDI attack**) An FDI attack is considered stealthy if it is capable of injecting false data on communication channels for an arbitrary time interval while remaining undetected.* □

**Definition 5.** *[19] (**Covert attack**) A covert attack requires perfect system knowledge (1) and (2), $\mathcal{I}_{attacker} = \{A, B, C, \mathcal{X}, \mathcal{U}\}$, disclosure and disruptive resources on both actuator and measurement channels. This coordinated attack injects the vector $u_k^a$ into the system to arbitrary deteriorate the control system performance while the vector $y_k^a$ is injected to completely remove the attack's effect in the measurement vector. Due to the*

*linearity, $y_k^a$ can be simply computed as follows:*

$$y_k^a := -C \sum_{j=0}^{k-1} (A^j B u_{k-1-j}^a) \tag{15}$$

□

**Definition 6.** *[11] (**Replay attack**) A replay attack requires disclosure and disruption resources on a given channel. This attack is executed in two steps: first a disclosure attack is launched to record the transmitted data for an arbitrary number of step $\tau > 0$. Then in the second phase of the attack, the recorded data is replayed in the same channel instead of the legitimate one.* □

**Definition 7.** *[14] (**Zero-dynamics attack**) A zero-dynamics attack requires perfect model knowledge and disruption resources on actuation channel. 0-stealthy attacks exploits the transmission zeroes of a system to inject an input vector $u_k^a$ which produce a zero response, $y_k^a \equiv 0$, on the output vector.* □

**Remark 1.** *In zero-dynamics attack, $u_k^a := \theta^k g$ where $\theta$ is the zero of the system and $g$ is the corresponding input-zero direction. By assuming matrix $B$ to be full column rank in (1), the transmission zeros can be found as the values of $\theta$ than make the following $P(\theta)$ to lose rank:*

$$P(\theta) = \begin{bmatrix} \theta I - A & -B \\ C & 0 \end{bmatrix} \tag{16}$$

*If $|\theta| < 1$, the zero is called minimum phase or stable zero and if $|\theta| \geq 1$, it is called non-minimum phase or unstable zero. The input-zero direction is found by solving the following equation:*

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

*where $x_0$ is the initial state of the system.*

## 2.3 Set Theoretic Control

In this section, the set-theoretic control approach is described. Model Predictive Control (MPC) is a well-known control strategy capable of dealing with plant constraints and disturbances. Traditionally, the MPC problem is formulated as a constrained optimization problem over a prediction control horizon. Such optimization is executed at each sampling time and according to the receding horizon paradigm, only the first computed action is applied to the system [36]. Such paradigm histrionically suffered for the required computationally high demand. Therefore, in the literature, different approaches have been proposed to mitigate such a burden [37]. In particular, of interest here is the set-theoretic control paradigm developed in [38–40], for its capability to move most of the calculations offline and solve a simpler optimization problem on-line.

### 2.3.1 Set Theoretic Control Design

Let us consider the following linear plant model subject to an exogenous bounded disturbances

$$x_{k+1} = Ax_k + Bu_k + B_d d_k \tag{18}$$

where $d_k \in \mathcal{D} \subset \mathbb{R}^d$ with $0_d \in \mathcal{D}$.

**Definition 8.** *[40] (**Minkowski/Pontryagin set sum and difference**) Given two sets $\mathcal{A} \subset \mathbb{R}^n$ and $\mathcal{B} \subset \mathbb{R}^n$, the Minkowski/Pontryagin set sum and difference are defined as follows:*

$$\begin{aligned} \mathcal{A} \oplus \mathcal{B} &:= \{a + b | a \in \mathcal{A}, b \in \mathcal{B}\} \\ \mathcal{A} \sim \mathcal{B} &:= \{a \in \mathcal{A} | a + b \in \mathcal{A}, \forall b \in \mathcal{B}\} \end{aligned} \tag{19}$$

□

**Definition 9.** *[41] (**Robust Positive Invariant(RPI) Set**) A set $\mathcal{O} \subseteq \mathcal{X}$ is robust*

*positive invariant set for the autonomous system* ($x_{k+1} = Ax_k + B_d d_k$)*, if*

$$x_0 \in \mathcal{O} \Rightarrow x_k \in \mathcal{O}, \forall d_k \in \mathcal{D}, k \in \mathbb{Z}_+ \tag{20}$$

□

**Definition 10.** *[41] (**Robust Control Invariant(RCI) region**) A set* $\mathcal{C} \subseteq \mathcal{X}$ *is robust control-invariant for the system (18) subject to constraints (2) if*

$$x_k \in \mathcal{C} \Rightarrow \exists u_k \in \mathcal{U}, \textit{such that } Ax_k + Bu_k + B_d d_k \in \mathcal{C}, \forall d_k \in \mathcal{D}, k \in \mathbb{Z}_+ \tag{21}$$

□

**Definition 11.** *[41] (**One-Step Controllable Set**) Given a set* $\mathcal{T}$*, the set of states* $\mathcal{T}_1$*, controllable in one-step towards* $\mathcal{T}$ *regardless of the disturbances in the system is defined as follows:*

$$\begin{aligned} \mathcal{T}_1 :=& \{x \in \mathcal{X} : \exists u \in \mathcal{U}\ s.t.\ \forall d \in \mathcal{D}, Ax + Bu + B_d d \in \mathcal{T}\} \\ =& \{x \in \mathcal{X} : \exists u \in \mathcal{U}\ s.t.\ Ax + Bu \in \tilde{\mathcal{T}}\} \end{aligned} \tag{22}$$

*where* $\tilde{\mathcal{T}} := \mathcal{T} \sim B_d \mathcal{D}$. □

Let us consider the regulation problem for (18). According to the set-theoretic paradigm, the control law is built in two phases: offline and online. In the offline phase, the following steps are taken:

- Step 1- By considering the unconstrained disturbance-free model of (18), a stabilizing state-feedback controller $u_0(x_k)$ is designed to regulate the states of the system towards the equilibrium point as shown in Fig. 6.
- Step 2- The smallest RCI set $\mathcal{T}_0$ associated to the state-feedback controller designed in Step 1 is computed [42] such that constraints (2) are satisfied, see Fig. 7.

- Step 3- The terminal controller designed in the steps 1-2 might have a small Domain of Attraction(DoA). In order to cover all the possible system's initial conditions i.e. $\forall x_0 \in \mathcal{X}_\eta \subset \mathcal{X}$, the recursion (23) can be employed to enlarge the DoA. These sets are enlarged until the family of one-step controllable sets cover the set of initial conditions namely, $\bigcup_{i=0}^{N} \mathcal{T}_i \supseteq \mathcal{X}_\eta$ where N is the number of sets. Moreover, N represents the maximum number of control moves required to reach the terminal region $\mathcal{T}_0$ starting from any initial condition in $\mathcal{X}_\eta$ as shown in Fig. 8, see [43].

$$
\begin{aligned}
\mathcal{T}_0 :=& \mathcal{T} \\
\mathcal{T}_i :=& \{x \in \mathcal{X} : \exists u \in \mathcal{U}\ s.t.\ \forall d \in \mathcal{D}, Ax + Bu + B_d d \in \mathcal{T}_{i-1}\} \\
=& \{x \in \mathcal{X} : \exists u \in \mathcal{U}\ s.t.\ Ax + Bu \in \tilde{\mathcal{T}}_{i-1}\}
\end{aligned}
\tag{23}
$$

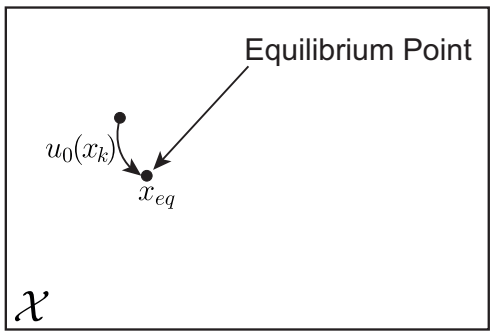



Figure 6: Regulator design

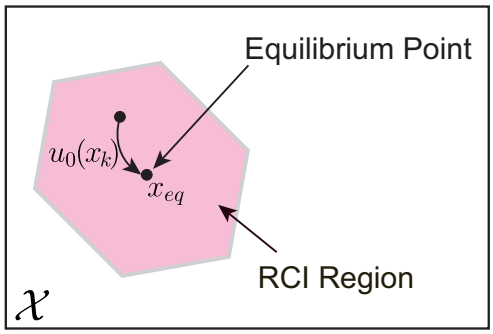



Figure 7: Terminal controller design

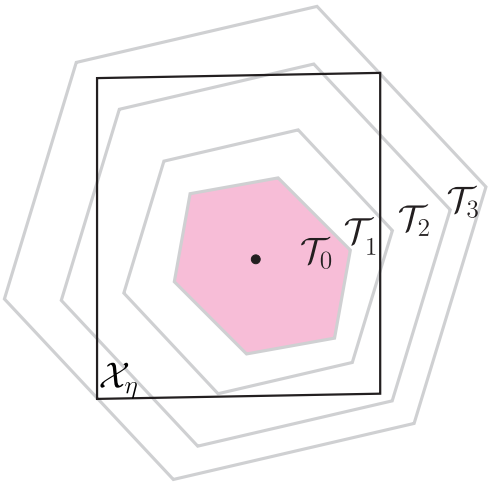



Figure 8: Enlarging the DoA of the controller

**Remark 2.** *In Step 1, a simple way to design the state-feedback controller $u_0(x_k)$ is by resorting to the well-known Linear Quadratic(LQ) controller*

$$u_0(x_k) = K_0(x_k - x_{eq}) + u_{eq} \tag{24}$$

*where $K_0 \in \mathbb{R}^{m\times n}$ is the LQ controller gain, $x_{eq} \in \mathbb{R}^n$ and $u_{eq} \in \mathbb{R}^m$ are the states and the input vector associated to the equilibrium point, respectively.*

**Remark 3.** *As the number of sets i increases, the complexity of the recursive computation (23) increases and becomes intractable. Therefore, approximation methods are proposed in the literature e.g. the ellipsoidal inner approximation in [40, 44] or Zonotopes based method in [45].*

In the online phase, the offline computed family is used to compute the control input. To this end, the following algorithm is utilized:

---

Set-Theoretic Control Paradigm

---

**Off-line computations:** $\{\mathcal{T}_i\}_{i=0}^{N}$, $\mathcal{X} \subseteq \bigcup_{i=0}^{N} \mathcal{T}_i$

**On-line computations:** $u_k$

1: Find the smallest set index $i_k$ containing $x_k$,

$$i_k := min\{i : x_k \in \mathcal{T}_i\}$$

2: **if** $i_k == 0$ **then** $u_k = K_0(x - x_{eq}) + u_{eq}$ (see (24))

3: **else**

$$\begin{aligned} u_k &= \arg\min_u J(x_k, u) \quad s.t. \\ Ax_k + Bu &\in \tilde{\mathcal{T}}_{i_k-1},\ u \in \mathcal{U} \end{aligned} \tag{25}$$

4: **end if**

5: $k \leftarrow k + 1$ goto Step 1

---

where $J(x_k, u)$ is any convex cost function of interest.

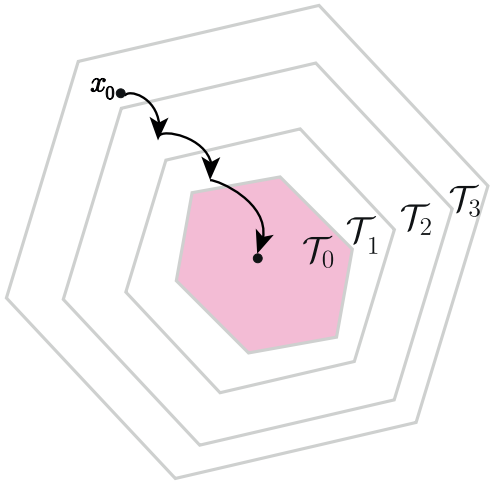



Figure 9: The one-step controllable sets

**Remark 4.** *It is possible to prove that the described set-theoretic control paradigm enjoys the following properties:*

- *The plant's state vector evolution converges to the terminal region in a finite number of steps.*
- *The trajectory is Uniformly Ultimately Bounded (UUB) in $\mathcal{T}_0$ regardless of any disturbance realization*
- *In the absence of disturbance, then the stability is asymptotic*

# Chapter 3

# A Control Architecture to Detect FDI Attacks

The control architecture proposed in this chapter is published as a conference paper in ACC 2019, see [26].

In recent years, different solutions have been proposed to detect advanced stealthy cyber-attacks against networked control systems. In this manuscript, we propose a blended detection scheme that properly leverages and combines two existing detection ideas, namely *watermarking* and *moving target*. In particular, a watermarked signal and a nonlinear static auxiliary function are combined to both limit the attacker's disclosure resources and obtain an unidentifiable moving target. The proposed scheme is capable of detecting a broad class of intelligent attacks, including zero-dynamics, replay, and covert attacks. Moreover, it is shown that the proposed approach mitigates the drawbacks of standard moving target and watermarking defense strategies. Finally, an extensive simulation study is reported to contrast the proposed detector with recent competitor schemes and provide tangible evidence of the effectiveness of the proposed solution.

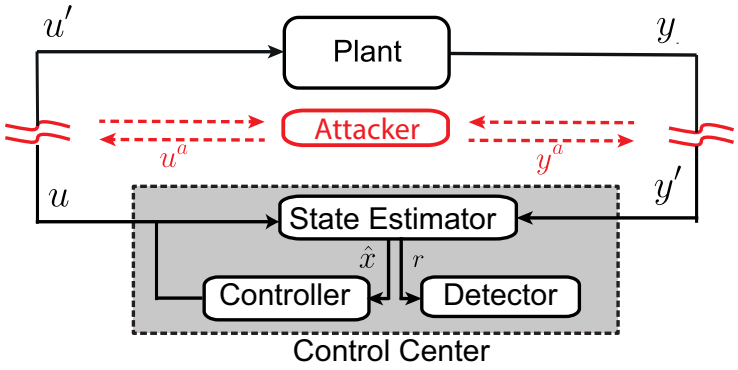
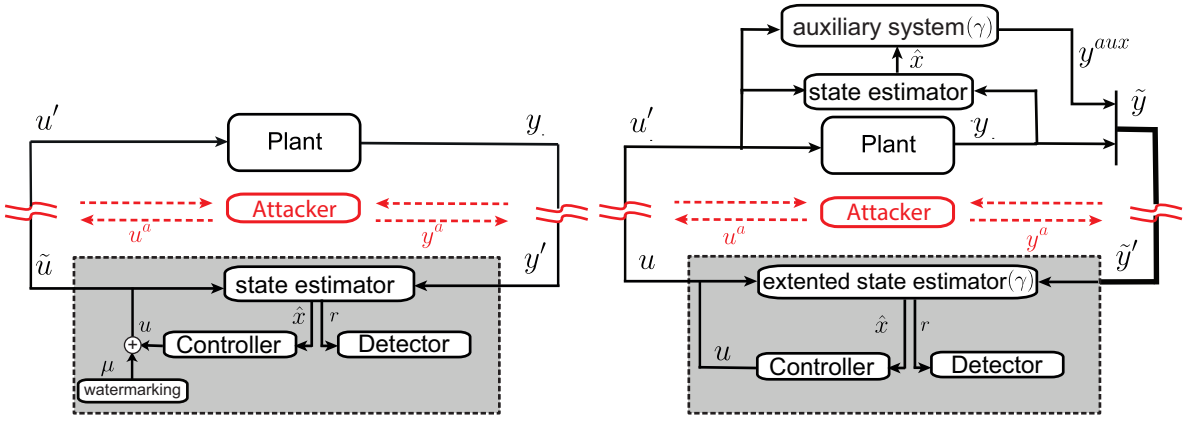



Figure 10: Networked control system under cyber-attacks

## 3.1 Problem Formulation

In this section, first, general limitations and drawbacks of existing watermarking [11, 12] and moving target [15, 16, 46] detection solutions (see Fig. 11) are summarized, then, the objective of this chapter is stated.




(a) Watermarked Control Inputs. (b) Moving Target (Auxiliary System)

Figure 11: Networked control system equipped with watermarked inputs or moving target (auxiliary system)

- *Watermarking detection scheme*: In order to authenticate the system dynamics and detect steady-state replay attacks, a watermarking signal $\mu_k$ with zero-mean and covariance $\mathcal{S}$ ($\mu_k \sim \mathcal{N}(0, \mathcal{S})$) is added to the optimal control input computed by the controller, i.e. $\tilde{u}_k := u_k + \mu_k$ (see Fig. 11.a). It has been shown in [47] that the

probability of detecting replay attack is directly proportional to the covariance $\mathcal{S}$ and that the control performance are inverse proportional to $\mathcal{S}$. As a consequence, a trade-off between contrasting objectives (attack detection and control performance) must be reached [12]. Moreover, the watermarking detection scheme in Fig. 11.a is unable to detect zero-dynamics or covert attacks.

- *Moving target (auxiliary system) detection scheme*: In order to reveal covert-attacks, a randomly switching dynamical auxiliary system (moving target) is deployed on plant side and its dynamics are coupled with the plant dynamics, see Fig. 11.b. While this detection scheme has been proved to be effective against covert-attacks and zero-dynamics [15, 16, 46], different drawbacks can be highlighted, especially for its practical implementation: a state estimator module needs to be deployed on the plant side for coupling the auxiliary dynamics; switching random dynamics must be generated/emulated in the plant side; auxiliary sensor measurements must be transmitted; the state-estimator in the control center must be changed into a switching state estimator on an extended state-space vector.

***Objective (O1)**: Given the networked control system in Fig. 10, the plant model (1) and the control center's detector (11), design a novel active detection scheme capable of*

- *Assuring the absence of stealthy FDI attacks (e.g. replay (Definition 6), zero-dynamics (Definition 7), covert (Definition 5));*
- *Overcoming the drawbacks of watermarking [11] and moving-target [15, 16, 46] detection schemes.*

## 3.2 Proposed Networked Control Architecture

In this section, a detection strategy meeting the objective (**O1**) is designed. The section is organized as follows: first, the proposed control architecture is presented and the detection mechanism is illustrated; then, it is formally shown that the proposed solution does

not affect the closed-loop system performance and stealthy attacks cannot be launched; Finally, the advantages of the proposed solution are highlighted and some remarks and guidelines for the design of the auxiliary system are given.

### 3.2.1 Control Architecture Operation and Detection Strategy

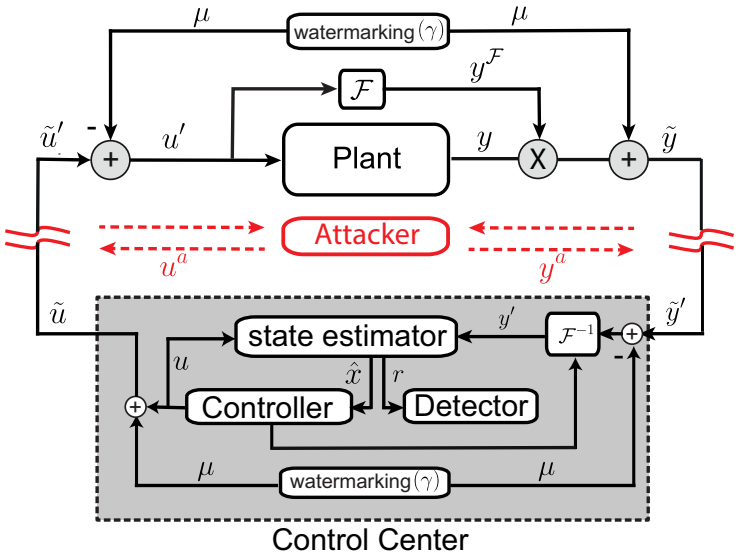



Figure 12: Proposed networked control architecture

The proposed control architecture, see Fig. 12, consists of the following main ingredients:

- A standard networked control architecture, see Fig. 10;
- A pseudo-random number generator producing a watermarking signal $\mu_k$. The pseudo-random sequence is generated starting from seed number $\gamma$ which is shared between the plant and control center;
- An auxiliary injective non-zero single-valued nonlinear function $\mathcal{F}: u'_k \rightarrow \mathcal{F}(u'_k)$ where $u'_k \in \mathbb{R}^m$ and $y^{\mathcal{F}}_{k+1} := \mathcal{F}(u'_k) \in \mathbb{R}_+$ .

where we assume the following:

**Assumption 2.** *In the proposed networked architecture Fig. 12, only the seed number $\gamma$ is assumed secretly shared between the plant and controller and unknown to the attacker [48]. The plant model, the control center operations, the auxiliary function $\mathcal{F}$ and the reference architecture are known to both the defender and attacker.* □

We can summarize the networked control operations by means of the following pseudo-procedure:

---

Networked Control System - Operations (***NCS-O***)

---

——— *control center* ———

**Receive:** $\tilde{y}'_k$, **Send:** $\tilde{u}_k$

1: The watermarking signal $\mu_k$ and the auxiliary output $\mathcal{F}(u_{k-1})$ are removed from $\tilde{y}'_k$, i.e.

$$y'_k = \mathcal{F}^{-1}(u_{k-1})(\tilde{y}'_k - \mu_k) \tag{26}$$

2: The state-estimator computes the best estimation $\hat{x}_k$ and the residual signal $r_k$ according to (3) and (6), respectively

3: The $\chi^2$ detection rule (11) checks for FDIs attacks

4: The controller computes the optimal control action $u_k$;

5: The watermarking signal $\mu_k$ is superimposed on $u_k$, i.e. $\tilde{u}_k = u_k + \mu_k$

6: The watermarked command $\tilde{u}_k$ is transmitted.

——— *plant side* ———

**Receive:** $\tilde{u}'_k$, **Send:** $\tilde{y}_{k+1}$

1: The watermarking signal $\mu_k$ is removed from the received command $\tilde{u}'_k$, i.e. $u'_k = \tilde{u}'_k - \mu_k$

2: The auxiliary output is computed, i.e. $y^{\mathcal{F}}_{k+1} = \mathcal{F}(u'_k)$

3: The plant output vector $y_{k+1}$ is multiplied by the scalar $y^{\mathcal{F}}_{k+1}$ and the watermarking signal $\mu_k$ is added

$$\tilde{y}_{k+1} = (y_{k+1} y^{\mathcal{F}}_{k+1}) + \mu_{k+1}; \tag{27}$$

4: The watermarked measurement vector $\tilde{y}_{k+1}$ is transmitted

---

The detection strategy can be summarized as follows. The same watermarking random signal $\mu_k$ is generated in the control center and locally to the plant. Such a randomly changing signal is added on top of the transmitted actuation and sensor measurements to limit the attacker's disclosure resources. Moreover, such a signal is removed at the receivers' sides to avoid any associated performance loss as in [47]. The auxiliary function $\mathcal{F}$ is used to generate a moving target. Nevertheless, contrary to existing solutions [15, 16, 46], the moving target is not achieved by means of a switching auxiliary system, but jointly combining the action of the watermarking signals and the nonlinear multiplicative coupling between the system's outputs $y_k$ and the auxiliary's output $y^{\mathcal{F}}_k$. In particular, the transmitted watermarked signal, prevents the attacker to understand the exact value of $u_k$ or $y_k$, while the non-linearity in $\mathcal{F}$ and in the measurement coupling $(y^{\mathcal{F}}_k y_k)$ does not allow the attacker to generate a perfect replay or covert action (see e.g (15) for the covert-attacks). Moreover, by designing $\mathcal{F}$ to be a non-zero function, zero-dynamics attacks detection is also enabled. Indeed, any zero-dynamic input vector $u^a_k = \theta^k g$ will never produce a non-zero output in the auxiliary system.

### 3.2.2 Correctness of the control operations in absence of attacks

The correctness of the proposed control architecture under an attack-free scenario is proved in the following proposition.

**Proposition 1.** *Let us consider the control architecture in Fig. 12. Under an attack-free scenario,* $u^a_k \equiv 0$ *and* $y^a_k \equiv 0$, *the proposed architecture does not interfere with the closed-loop control system operations.*

*Proof* - The proposition's proof is obtained by showing that, in absence of attacks, the proposed control architecture (Fig. 12) and the standard networked control system (Fig. 10) operations are equivalent. In particular:

- $\forall\, k \rightarrow u'_k \equiv u_k$ : In the Step 1 of the ***NCS-O*** algorithm (plant-side) the signal $u'_k$ is recovered as $u'_k = \tilde{u}'_k - \mu_k$. Under an attack-free scenario, $u^a_k \equiv 0$, we can write

$$u'_k = (u_k + \mu_k) - \mu_k \equiv u_k, \;\; \forall\, k;$$

  which concludes the first part of the proof;

- $\forall\, k \rightarrow y'_k \equiv y_k$ : In the Step 1 of the ***NCS-O*** algorithm (controller-side) the signal $y'_k$ is recovered as

$$y'_k = \mathcal{F}^{-1}(u_{k-1})(\tilde{y}'_k - \mu_k)$$

  First, it is important to remark that such operation is well-posed because $\mathcal{F}$ is a non-zero single-valued function. Then, under an attack-free scenario, $y^a_k \equiv 0$, we can re-write $y'_k$ as

$$y'_k = \frac{y_k \mathcal{F}(u_{k-1}) + \mu_k - \mu_k}{\mathcal{F}(u_{k-1})} \equiv y_k, \;\; \forall\, k \tag{28}$$

  which concludes the second part of the proof. ■

### 3.2.3 Absence of stealthy attacks and auxiliary system design

In the presence of FDI attacks (14), the output vector $y'_k$ becomes

$$\begin{aligned} y'_k &= \frac{\tilde{y}'_k - \mu_k}{\mathcal{F}(u_{k-1})} = \frac{(\tilde{y}_k + y^a_k) - \mu_k}{\mathcal{F}(u_{k-1})} \\ &= \frac{(y_k \mathcal{F}(u'_{k-1}) + \mu_k + y^a_k) - \mu_k}{\mathcal{F}(u_{k-1})} = \frac{y_k \mathcal{F}(u_{k-1} + u^a_{k-1}) + y^a_k}{\mathcal{F}(u_{k-1})} \end{aligned} \tag{29}$$

and the following proposition can be stated:

**Proposition 2.** *Let us consider the control architecture in Fig. 12 and the FDI attack*

*model (14). The proposed control architecture ensures that intelligent FDI attacks (replay, zero-dynamics, covert-attacks) cannot remain stealthy.*

*Proof -*

- *replay attacks:* In [11], it has been shown that stealthy replay attacks can be launched only when the closed-loop system is in steady-state conditions. Nevertheless, from (29), it is possible to notice that if the system is in stationary conditions, i.e. $E[y_k] \equiv \bar{y}$ and $E[u_k] \equiv \bar{u}$, and an input attack is performed, then the transmitted and received outputs signal, $\tilde{y}_k$ and $y'_k$, respectively, are not stationary, i.e.

$$\begin{aligned} \tilde{y}_k &= \bar{y}\mathcal{F}(\bar{u} + u^a_{k-1}) + y^a_k \\ y'_k &= \frac{\bar{y}\mathcal{F}(\bar{u}+u^a_{k-1})+y^a_k}{\mathcal{F}(u_{k-1})} = \frac{\tilde{y}_k+y^a_k}{\mathcal{F}(u_{k-1})} \end{aligned}$$

  As a consequence, any substituting of $\tilde{y}_k$ with a previously recorded vector, i.e. $y^a_k = \tilde{y}_{k-T} - \tilde{y}_k$, $T > 0$, will be detected by (11).

- *zero-dynamics attacks:* Let us denote with $y^u_k$ and $y^{u^a}_k$ the outputs of the plant due the quadruple ($u_k$, $\omega_k$, $\eta_k$, $x_0$) and to attack vector $u^a_k$, respectively. For linearity, we can write that

$$y_k = y^u_k + y^{u^a}_k$$

  and

$$\tilde{y}_k = (y^u_k + y^{u^a}_k)\mathcal{F}(u_{k-1} + u^a_{k-1}) + \mu_k \tag{30}$$

  By definition of zero-dynamic attack we have that $y^{u^a}_k \equiv 0$, $\forall\, k$. Nevertheless, given the non-zero nature of the auxiliary function $\mathcal{F}$ and according to (30), the effect of the input attacks will never be zero in the transmitted vector $\tilde{y}_k$. As a consequence, stealthy zero-dynamics attacks are not possible in the proposed architecture.

- *Covert attacks:* A covert attack will be successful if the attacker is capable of removing from the output vector $y_k$, the effect of input attack ($u^a_k$), namely $y^{u^a}_k$. From

(29), the latter translates into the following problem:

$$\text{Find } y_k^a: \ \ y_k\mathcal{F}(u_{k-1}+u_{k-1}^a)+y_k^a=y_k\mathcal{F}(u_k) \tag{31}$$

which the solution is:

$$y_k^a=y_k\left(\mathcal{F}(u_{k-1})-\mathcal{F}(u_{k-1}+u_{k-1}^a)\right) \tag{32}$$

However, in the proposed architecture, both $u_k$ and $y_k$ are unknown to the attacker (the attacker has disclosure information only related to the transmitted watermarked signals $\tilde{u}_k$ and $\tilde{y}_k$). Moreover, given the non-linear nature of the function $\mathcal{F}$ and coupling between the system outputs and the auxiliary output ($y_k\mathcal{F}(u_{k-1}+u_{k-1}^a)$), the attacker cannot simply exploits (15) to perfectly cancel the effect of input attacks. As a consequence, a perfect covert attack cannot be launched. ■

**Remark 5.** *Given the impossibility for the attacker to perform the perfect cancellation (32) and the injective non-zero nature of $\mathcal{F}(\cdot)$, it is straightforward to show that the residual signal under attack, namely $r'(k)$, is different from the residual signal in absence of attack, i.e.*

$$E[r'(k)] = E\left[\frac{y_k\mathcal{F}(u_{k-1}+u_{k-1}^a)+y_k^a}{\mathcal{F}(u_{k-1})}-C\hat{x}_{k-1}\right]\neq E[r(k)]$$

*Moreover, $r'(k)$ is a function of $\mathcal{F}(\cdot)$ which can be designed to increase the sensitivity in response to any attacker's input vector $u_{k-1}^a\in\mathbb{R}^m$, i.e.*

$$\mathcal{F}(u_{k-1}+u_{k-1}^a)>>\mathcal{F}(u_{k-1})\ or\ \mathcal{F}(u_{k-1}+u_{k-1}^a)<<\mathcal{F}(u_{k-1})$$

*In the simulation section, by considering as auxiliary function the exponential law $\mathcal{F}(u_k')=e^{||u_k'||_2}$, experimental results are conducted to characterize the sensitivity of the $\chi^2$ detection rule for different input attack vectors.* □

### 3.2.4 Advantages of the proposed solution

The main capabilities and advantages of the proposed control architecture, in terms of attack detection capability, control performance and architecture, can be summarized as follows:

- *Attack detection:* The proposed detection mechanism ensures the absence of stealthy replay, zero-dynamics and covert attacks. Moreover, the auxiliary system can be designed to achieve any desired level of sensitivity to FDI input attacks.

- *Control performance:* Contrary to existing watermarking solutions, the proposed detection scheme does not introduce any performance loss.

- *Architecture advantages:* Contrary to existing moving-target solutions, the moving target is here obtained by using a nonlinear static function $\mathcal{F}$ instead of a randomly changing dynamical system. Moreover, the auxiliary system does not need to be coupled with the plant dynamics. Furthermore, the size of the transmitted measurement vector is not increased and control center operations are unchanged (e.g. extra sensor data and extended switching state-estimators are not needed).

## 3.3 Simulation example

In this section, the effectiveness of the proposed strategy is testified against three different intelligent FDI attacks: covert, replay and zero-dynamics attacks.

The quadruple-tank water system introduced in [49] and shown in Fig. 13 is used as the testbed. The system consists of four tanks where the water levels $h_i$, $i = 1, \ldots, 4$ are the state components of the system, i.e. $x = [h_1, ..., h_4]^T$ while the two valves $v_1$ and $v_2$ are the control inputs, i.e. $u = [v_1, v_2]^T$. Two sensors are available to measure the water levels in $h_1$, $h_2$.

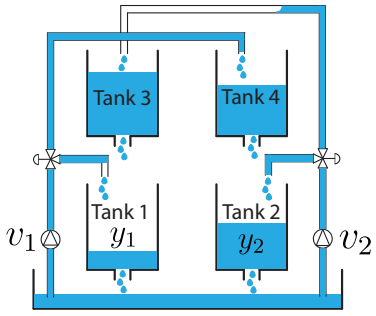


Figure 13: Quadruple-tank process

The nonlinear system dynamics have been discretized with a sampling time $T_s = 1\,\text{sec}$ and linearized around the operating equilibrium point

$$x_{eq} = [5,\ 5,\ 2.044,\ 1.399]^T, \quad u_{eq} = [0.724,\ 1.165]^T.$$

obtaining the linearized system matrices:

$$A = \begin{bmatrix} 0.975 & 0 & 0.042 & 0 \\ 0 & 0.977 & 0 & 0.044 \\ 0 & 0 & 0.958 & 0 \\ 0 & 0 & 0 & 0.956 \end{bmatrix}$$
$$B = \begin{bmatrix} 0.0515 & 0.0016 \\ 0.0019 & 0.0447 \\ 0 & 0.0737 \\ 0.0850 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0.2 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \end{bmatrix} \tag{33}$$

The following subsystems have been used:

- An LQ controller to regulate the level of water in each tank around the equilibrium. The controller gain is:

$$K = \begin{bmatrix} 3.0993 & 4.0721 & -2.0528 & 2.8417 \\ 3.9353 & 3.3330 & 2.8461 & -1.9997 \end{bmatrix}$$

- A Kalman filter is used to estimate states of the plant. The steady-state gain $L$ is:

$$L = 10^{-4} \times \begin{bmatrix} 0.8349 & 0 & 0.2325 & 0 \\ 0 & 0.8688 & 0 & 0.2292 \\ 0.2325 & 0 & 0.5808 & 0 \\ 0 & 0.2292 & 0 & 0.5557 \end{bmatrix}$$

- A $\chi^2$ detector is used as the anomaly detector. The detector threshold is $\beta = 7.013$ which has been tuned for a 3% false alarm rate.

- As the auxiliary system, the following exponential function is used: $\mathcal{F}(u'_k) := e^{||u'_k||_2}$

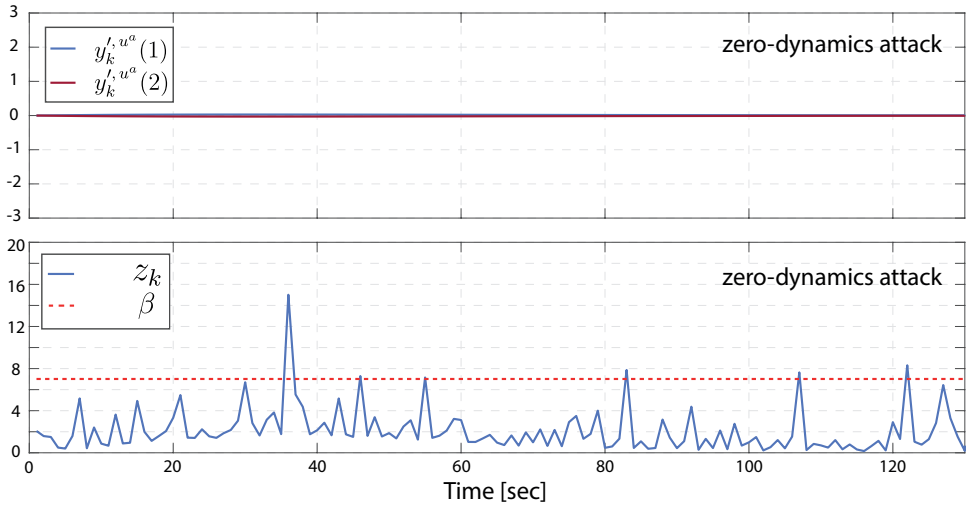
### 3.3.1 Zero-dynamics Attack

In this section, the ability of the proposed architecture to detect zero-dynamics attacks is evaluated. Since the plant presents two zeros, $\theta_1 = 0.89$ and $\theta_2 = 1.03$, the zero-dynamic attack is designed to excite the unstable zero $\theta_2$ as in (16). The designed input attack is the following

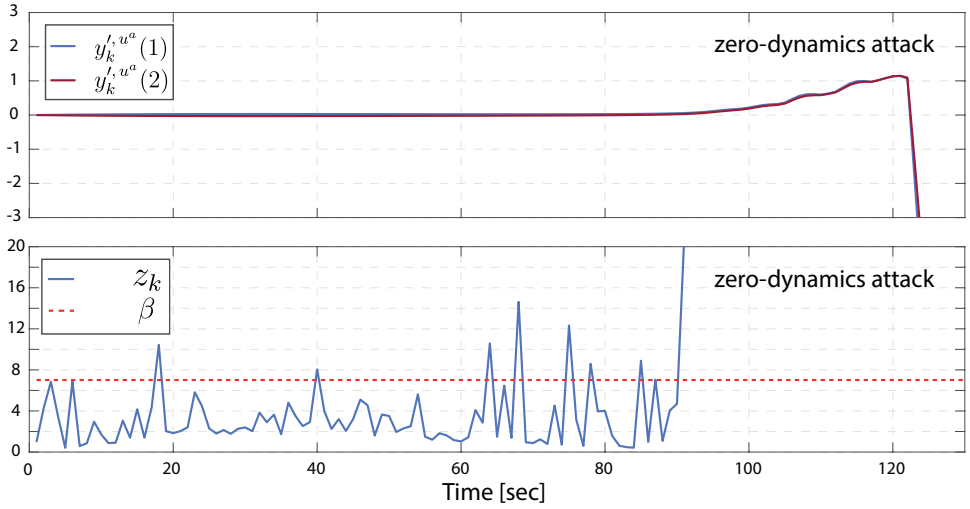
$$u_k^a = \begin{cases} 1.03^k \begin{bmatrix} -0.26 \\ 0.3 \end{bmatrix} & \text{if } \ 0 \leq k \leq 130 \\ 0 & \text{if } \quad k > 130 \end{cases}, \ x_0 = \begin{bmatrix} 5 \\ 5 \\ 2.35 \\ 1.10 \end{bmatrix}$$

The obtained simulation results are shown in Fig. 14. In the upper subplots, we show the measurement vector received by the state estimator and due to only the effect of zero-dynamics attack vector $u^a$, namely $y_k'^{,u^a}$. In the lower subplots we show the $\chi^2$ performance. As expected, when the auxiliary module is deactivated (subplots a), the zero-dynamics attack does not appear in the measurement vector and, as a consequence, the detection probability stays below the designed false alarm rate; on the other hand,

when the auxiliary module is active (subplots b), the used injective non-zero exponential function $\mathcal{F}$, shows the presence of the attack in the received measurements and, as a consequence, the $\chi^2$ test is able to detect the presence of the attack.




(a) Without the auxiliary system




(b) With the auxiliary system

Figure 14: Zero-dynamic attack against the proposed control architecture: without auxiliary system (a) vs with auxiliary system (b).

### 3.3.2 Replay Attack

In this section, the capability of the proposed architecture to detect replay attacks is investigated. Moreover, the proposed strategy is contrasted with the competitor scheme

[11] both in terms of attack's detection rate and control performance degradation.

We assume that the system starts from an initial condition $x_0 = [5.1,\ 5.2,\ 2.344,\ 1.799]^T$ and that a replay attack affects the system for $800 \leq t < 1200\,\text{sec}$. The replay attack is evaluated for four different

watermarking signals $\mu_k \sim \mathcal{N}(0, M)$. Moreover, to quantify detection rate and performance loss, the following performance indices are used and averaged over 1000 trials.

$$J_a\% = \frac{\sum_{k=800}^{1199}(z_k > \beta))}{400}\%,\ \ J_e = \sum_{k=200}^{799} \frac{||x(k) - x_{eq}||^2}{600}$$

where $J_a$ defines the attack detection rates while $J_e$ is the covariance of the tracking error signal in the absence of attacks.

The obtained results are summarized in Table 3 and Fig. 15. In Table 3, it is possible to appreciated that the proposed detection strategy does not affect the controller performance and that the detection rate remains above 99% for small watermarking signals. On the other hand, in [11], it is possible to notice that the tracking error covariance is proportional to the watermarking signal covariance while the detection rate is inverse proportional. Moreover, in Fig. 15 it is possible to qualitative notice how the tracking error signal degrades in the presence of a watermarking signal when $M = 10$. As a consequence, in [11], the watermarking signal must be properly designed to achieve the best compromise between detection rate and performance loss, while in the proposed solution such drawback is not present.

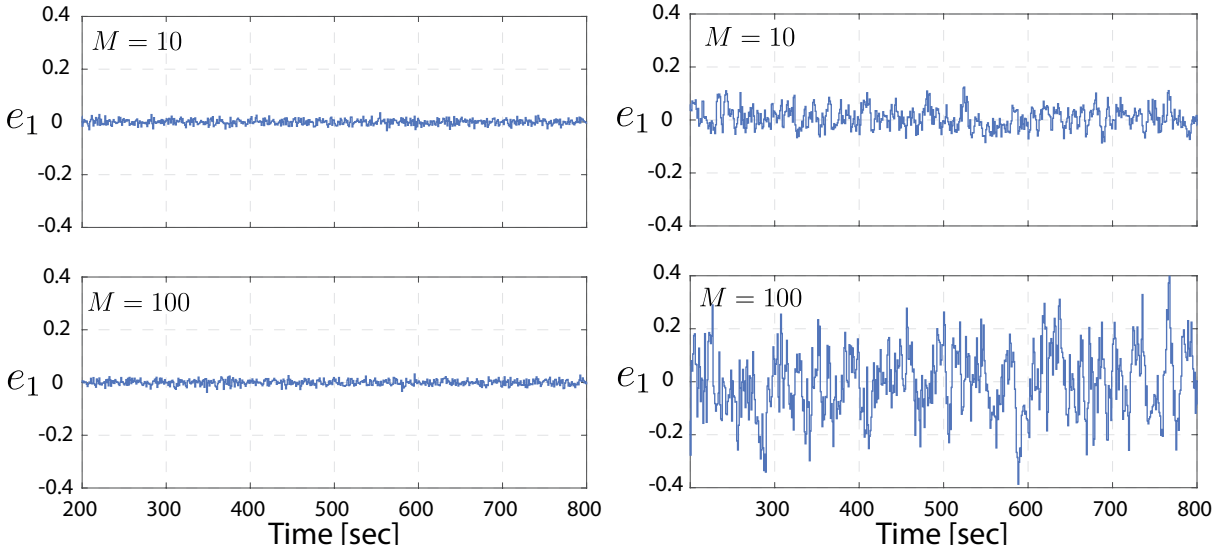
### 3.3.3 Covert Attack

In this section, the proposed detection scheme is validated by showing its effectiveness to detect covert attacks.

The detection performance of the proposed architecture is contrasted with solution proposed in [46]. In particular, we have assumed a plant's initial condition $x_0 = [5.1,\ 5.2,\ 2.344,\ 1.799]^T$ and a covert attack for $300 \leq t < 350$. We have investigated the detector's performance

Table 3: Detection rate $J_a$ and tracking error covariance $J_e$ for different watermarking signal covariance $M$: proposed control architecture vs [11]

| | Proposed Architecture | | [11] | |
|---|---|---|---|---|
| M | $J_a\%$ | $J_e$ | $J_a\%$ | $J_e$ |
| 100 | 99.72% | $3.40 \times 10^{-5}$ | 92.10% | 0.217 |
| 10 | 99.79% | $3.40 \times 10^{-5}$ | 72.91% | 0.022 |
| 1 | 99.80% | $3.40 \times 10^{-5}$ | 33.12% | 0.002 |
| 0.1 | 99.77% | $3.40 \times 10^{-5}$ | 8.68% | $2.21 \times 10^{-4}$ |




(a) Proposed control architecture

(b) proposed control architecture in [11]

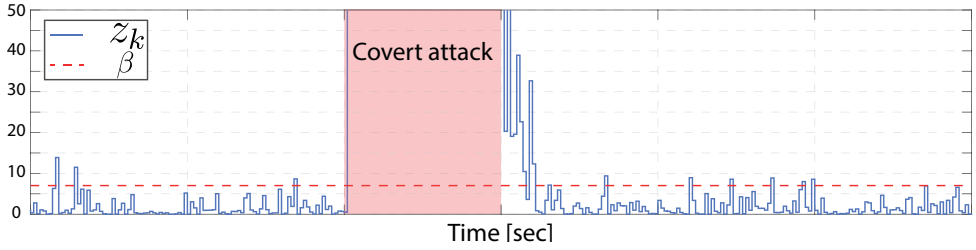
Figure 15: Tracking error on the first state component $e_1 := x_1 - x_{eq}(1)$ for $M = 10$ (top subplots) and $M = 100$ (bottom subplots): proposed control architecture vs [11].

$J_a = \frac{\sum_{k=300}^{349}(z_k > \beta))}{50}$ for different input attack vectors $u_k^a$. The experiment results, averaged over 1000 trials, are summarized in Table 4 while Fig. 16, shows the detection results on a single run. In Table 4, it is possible to appreciate that the detection rate of the proposed method is always bigger than the competitor scheme. Moreover, the detection rate of the proposed scheme drops less significantly of [46] when the magnitude of the input attack vector is decreased. This finds justification in the used exponential auxiliary function which results to be very sensitive to even small variations of its inputs. As a consequence, the proposed detector is more sensitive than the competitor scheme. Finally, in Fig. 16 it

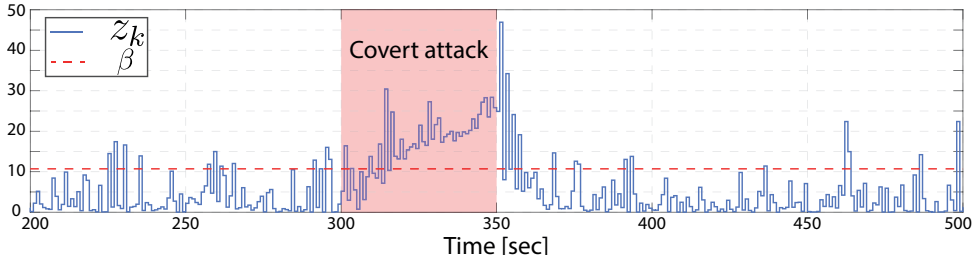
is possible to appreciate the $\chi^2$ signal for the proposed scheme (subplot (a)) and for [46] (subplot (b)) when $u^a(k) = [-0.1,\ -0.1]^T$. In the proposed solution, the $\chi^2$ signal increase abruptly and detection is achieved instantaneously while in [46], detection is achieved with some delay. The latter can be mainly explained for the different nature of the auxiliary system which in our approach is a static function while in [46] is a dynamical system with its own non-instantaneously dynamics.

Table 4: Detection rate $J_a$ for different input attack vectors $u^a$ : proposed detection strategies vs [46]

| $u^a$ | Proposed Detector $J_a\%$ | [46] $J_a\%$ |
|---|---|---|
| $[-0.5, -0.5]^T$ | 100% | 99.27% |
| $[-0.1, -0.1]^T$ | 100% | 77.59% |
| $[-0.03, -0.03]^T$ | 78.33% | 14.20% |




(a) Proposed detection strategy




(b) Proposed detection strategy in [46]

Figure 16: Covert attack: proposed detection strategy vs [46]

# Chapter 4

# A Control Architecture to Detect and Mitigate FDI Attacks

The control architecture proposed in this chapter is published as a conference paper in ECC 2019, see [27].

In this chapter, a novel networked control architecture capable of ensuring plant safety in presence of cyber-attacks on the communication channels is proposed. First, by combining a coding mechanism and a safety risk detection rule, an attack detection mechanism local to the plant is designed. Then, a set-theoretic controller is proposed as an emergency controller whenever an attack is detected and communication channels cannot be trusted. It is formally proved that the proposed control scheme enjoys plant safety regardless of any admissible attack scenario. A numerical simulation involving a two-tank water system is performed with the aim of clarifying the capabilities of the proposed solution.

## 4.1 Preliminaries and Definitions

In this chapter, the system model (1) is considered by assuming that entire state vector is available to the controller directly:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + B_d d_k \\ y_k &= I_n x_k \end{aligned} \tag{34}$$

where $k \in \mathbb{Z}_+ := \{0, 1, \ldots\}$, $x_k \in \mathbb{R}^n$ is the state vector, $u_k \in \mathbb{R}^m$ is the input vector and $d_k$ is a bounded disturbance, i.e.,

$$d_k \in \mathcal{D} \subset \mathbb{R}^d, \quad 0_d \in \mathcal{D} \tag{35}$$

Moreover, set-membership state and input constraints are prescribed as (2) namely, $u_k \in \mathcal{U}, \quad x_k \in \mathcal{X} \quad \forall\, k \in \mathbb{Z}_+$ Where $\mathcal{U} \subseteq \mathbb{R}^m$ and $\mathcal{X} \subseteq \mathbb{R}^n$ are compact subsets with $0_m \in \mathcal{U}$ and $0_n \in \mathcal{X}$, respectively.

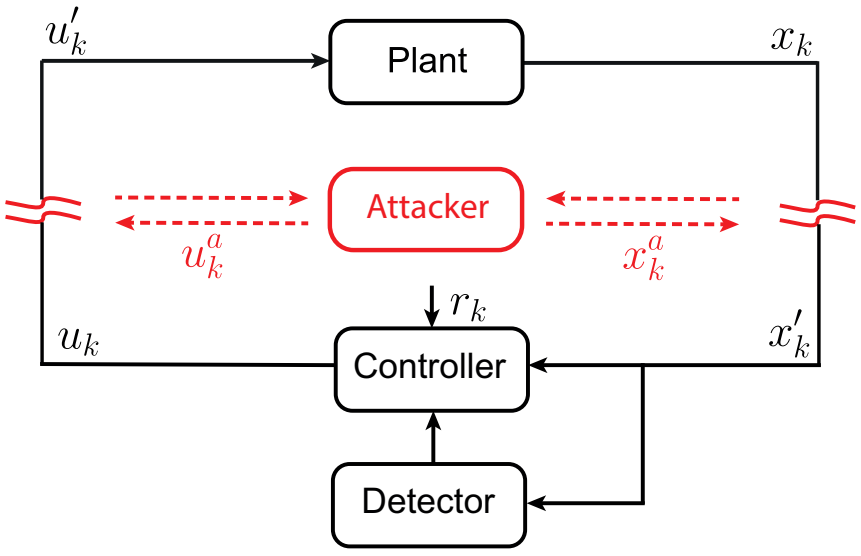



Figure 17: Networked control system vulnerable to cyber-attacks

## 4.2 Problem Formulation

In this section, first the considered attack scenario is presented and then the problem of interest is formally stated. By referring to the control architecture in Fig. 17, we model the plant behavior and controller actions under FDI attacks (14) as follows:

$$x_{k+1} = Ax_k + Bu'_k + B_d d_k \tag{36}$$

$$u_k = \eta(x'_k, r_k) \tag{37}$$

where, $u'_k := u_k + u^a_k$, $x'_k := x_k + x^a_k$, $u^a_k$ and $x^a_k$ are the attacker input and state signals, respectively, $r_k \in \mathbb{R}^r$ is the reference signal and $\eta : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^m$ is the function describing the remote controller logic.

By referring to Fig. 17, it is assumed that an FDI attacks (see Definition 3 and equation (14)) could affect both the actuation and measurement channels. The attacker aims to sabotage the tracking controller operations while remaining stealthy. To this end, the attacker is assumed to be aware of the networked control system operations (plant model (36) and controller logic (37)). Moreover, disclosure and disruptive (Definition 2) capabilities are assumed on both channels.

**Assumption 3.** *(**Controller Side**) We assume that a tracking controller 37 is available. Such a controller, in absence of attacks, satisfies the plant constraints (2) and ensures tracking of $r_k$. The controller working region, also known as controller DoA is $\mathcal{X}_\eta \subseteq \mathcal{X}$. Moreover, an observer-based anomaly detector [50] is present in the control side to detect faults or cyber-attacks in the closed-loop system, see e.g. [8].* □

**Assumption 4.** *(**Emergency Working Configuration**) It is assumed that for the plant (34) there exists, an a-priori defined equilibrium pair $(x^{em}_{eq}, u^{em}_{eq})$, compatible with the constraints (2), i.e. $x^{em}_{eq} \in \mathcal{X}$, $u^{em}_{eq} \in \mathcal{U}$, and acceptable, in terms of plant performance, under emergency attack scenarios.*

**Remark 6.** *The a-priori defined equilibrium pair is essential in this architecture because the reference signal is not available to the emergency controller. As a consequence, Assumption 4 is instrumental to assure that we can confine the plant's state trajectory in a safe region under attack.*

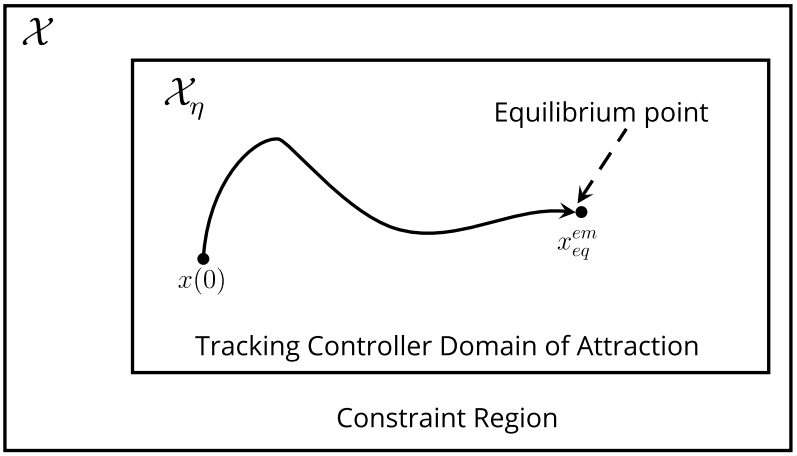



Figure 18: Equilibrium point

Two objectives are considered in this chapter. The first is detecting attacks on the plant side in order to prevent malicious inputs to be applied to the plant and avoid the existence of undetectable attacks. The second is designing a local safe controller which can be activated whenever an attack scenario is detected. The aims of this chapter can be formally stated as follows:

*Given the networked control system (36)-(37) (Fig. 17), the state and input constraints (2) and the attack model (14), the goal is to design a novel control architecture capable of*

- ***(O1)** - Detecting cyber-attacks occurrences (14) with the insurance that detection is accomplished before a harmful input sequence could violate the safety of the plant.*
- ***(O2)** - Activating an emergency controller, local to the plant, in response to an attack scenario detection. Such a controller has only objective of maintaining the plant safe operations until an attack-free scenario is recovered.*

Next sections provide a solution to **(O1)** and **(O2)**.

## 4.3 Proposed Distributed Control Architecture

In what follows, first the proposed control architecture(Fig. 19) and the role of each subsystem is introduced. Then, the emergency controller and the detector are designed and their effectiveness is formally proved.

In order to deal with the objectives **(O1)** and **(O2)**, the standard networked control system in Fig. 17 is extended as shown in Fig. 19. The proposed control architecture introduces the following subsystems:

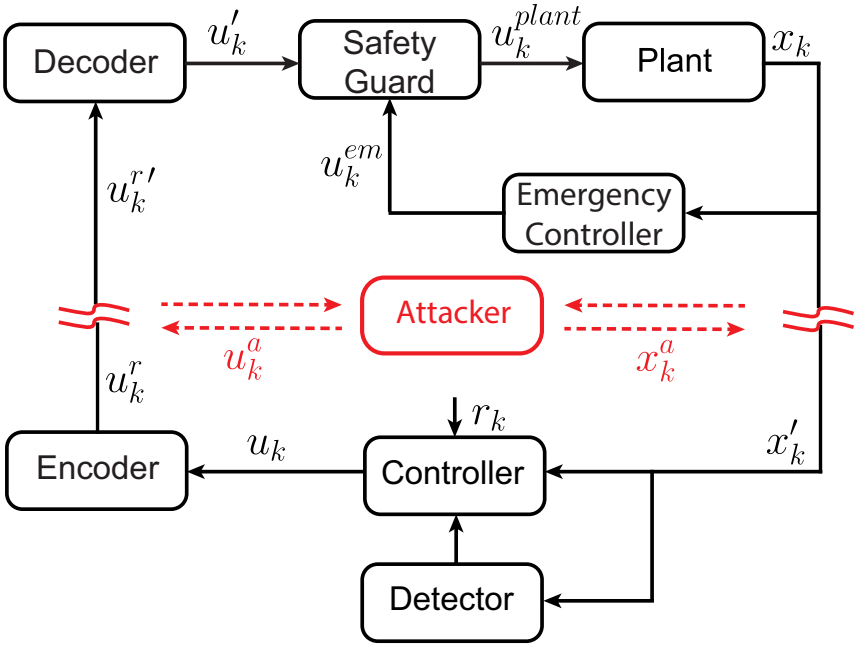



Figure 19: Proposed Control Architecture

- ***Encoder and Decoder:*** These blocks encode and decode the control input vector $u_k$, respectively, in such a way that the attacker's disclosure and disruptive resources on the actuation channel are deceived (detailed discussion in Section 4.3.1);

- ***Safety Guard:*** This subsystem detects attack occurrences before compromised input sequences could affect the plant safety (detailed discussion in Section 4.3.4);

- ***Emergency Controller:*** Such a local controller replaces, under attack, the spatially distributed tracking controller in order to preserve the plant safety (detailed discussion in Section 4.3.2)

**Remark 7.** *The rationale behind the emergency controller is to keep the plant states in a safe region. This controller cannot guarantee reference tracking since it does not have access to reference signal. Therefore, the performance of closed-loop system is degraded when the emergency controller is applied (under attack), but the safety is guaranteed.*

The rationale behind the proposed architecture can be briefly summarized as follows: it is assumed that attackers have disclosure and disruptive resources on both channels. Consequently, the considered networked control system is prone to advanced stealthy attacks [19] in which the detection task is impossible to be achieved regardless of any anomaly detector employed in the Control Center [9]. Following this reasoning, it is important to add an active component on the plant side of the network to limit the attacker's disruptive capabilities on at least one of the communication channels. In this respect, the inspiration is taken from the sensor coding ideas in [13] to propose a novel coding/decoding scheme (**Encoder** and **Decoder**) that is applied on the actuation channel. Moreover, under attack scenarios of arbitrary length, the communication channels cannot be trusted and the only way to ensure plant safety is to have a local **Emergency Controller** to be activated whenever attacks are detected. Such a controller cannot be aware of the reference signal $r_k$; therefore, its objective is only to maintain the plant safety until an attack-free scenario is re-established. Finally, to ensure safety regardless of any attack scenario, an attack detector module is added, namely **Safety Guard**, on the plant side. This module will trigger the safety controller any time an attack scenario is detected.

Next sections are devoted to design each component of the proposed control architecture:

### 4.3.1 Encoder and Decoder

The encoder block is placed between the tracking controller and the communication network. It performs two main tasks:

- The control signal vector $u_k$ is extended with an auxiliary random vector, namely $u_k^e := [u_k^{1,e}, \ldots, u_k^{p,e}]^T \in \mathbb{R}^p$ with $p \in \mathbb{Z}_+$;
- A random invertible matrix, namely $\Omega_k \in \mathbb{R}^{m+p}$, is applied on the augmented input vector $u_k^{aug} := [u_k^T, u_k^{eT}]^T$

Such actions produce in outcome the randomized input vector $u_k^r$

$$u_k^r = \Omega_k \underbrace{\begin{bmatrix} u_k \\ u_k^e \end{bmatrix}}_{u_k^{aug}} \tag{38}$$

which is transmitted through the network.

On the other hand, the decoder subsystem, placed between the communication channel and the plant, reconstructing the input signal on the plant side, namely $u'_k$, starting from the received signal $u_k^{r\prime} := u_k^r + u_k^a$. In particular, $u'_k$ is recovered as follows: first the received augmented signal, namely $u_k^{aug\prime} := [u_k'^T, u_k^{r\prime T}]^T$ is determined

$$u_k^{aug\prime} \;=\; \Omega_k^{-1} u_k^{r\prime} \;=\; \begin{bmatrix} u'_k \\ u_k^{e\prime} \end{bmatrix} \tag{39}$$

then, the first $m$ components of $u_k^{aug\prime}$, namely $u'_k$, are the input commands applied to the safety guard.

**Remark 8.** *It is important to justify why the proposed coding scheme is effective to compromise the attacker's capabilities on the actuation channel. Since the attacker is aware of the plant model (36) and tracking controller logic (37), the performed coding operations*

*(38) generates a random input vector $u_k^r$ which contains more components than the actual control signals $u_k$ and where none of the components resembles the real input signal. As a consequence, although, the attacker knows $u_k$ and can read the transmitted $u_k^r$, he/she cannot infer which component is meaningful. Moreover, even-tough, the attacker can inject an arbitrary input vector $u_k^a$, it cannot a-priori understand which component of $u_k^r$ will be affected. As a consequence, stealthy attacks on both input and output channels are prevented [9], [10]. In the proposed architecture, the detector and the safety guard are spatially distributed, see Fig. 19. In such a scheme, even if the detector detects the presence of attacks, it cannot securely inform the plant. In this regard, the introduced coding scheme can be also useful to embed such information in the auxiliary inputs. Please refer to Section 4.3.4 for further details.* □

**Remark 9.** *It is worth mentioning that the above coding/decoding scheme works under the assumption that the matrix $\Omega_k$ is the same for both the encoder and decoder. A possible way to do so is to assume that the matrix is generated from a pseudo-random algorithm which is initialized by a seed number that is secretly off-line shared between encoder and decoder [48].* □

### 4.3.2 Emergency Controller

In this section, a possible implementation of an emergency controller is proposed and it can be activated whenever the received control signals $u_k'$ cannot be trusted. The controller aims to keep the plant within a safe region which is shaped by the plant state and input constraints (2) and to drive the state trajectory towards the emergency equilibrium point $(x_{eq}^{em}, u_{eq}^{em})$ (see *Assumption* 4).

Hereafter, the proposed solution is based on an MPC idea exploiting set-theoretic arguments [40], [51]. Such a regulator is chosen for its capability of dealing with both state and input constraints, and for its modest computational demand during the on-line operations [40], which guarantees that the resulting control scheme can be executed

within a small sampling time interval. Nevertheless, it is important to remark that other constrained/robust controllers [52] could be successfully employed instead of the one here proposed.

The proposed controller [40] exploits two main ingredients: a RCI region and a family of robust one-step controllable sets. A detailed explanation on how to design this MPC controller can be found in sections 2.3 and 2.3.1.

The stopping condition for enlarging DoA of emergency controller is $\bigcup_{i=0}^{N}\{\mathcal{T}_i\} \supseteq \mathcal{X}_\eta$ which guarantees that the emergency controller has a DoA that is bigger or equal to the primary tracking controller (37) and as a consequence, the emergency controller can be safely activated starting from any plant condition $x_k \in \mathcal{X}_\eta$. Further details are provided in Section 4.3.4.

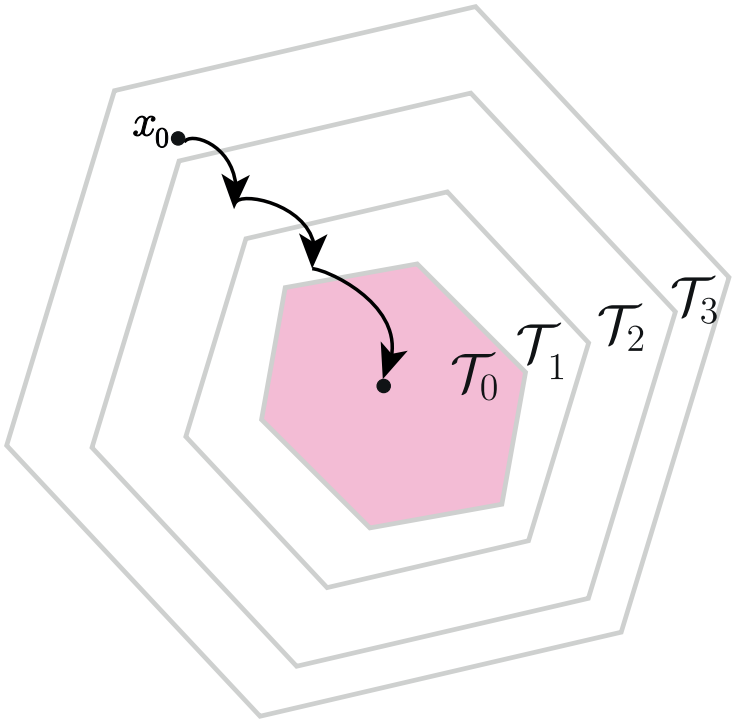



Figure 20: The one-step controllable sets

**Remark 10.** *According to the recursion (23), if the current state $x_k \in \mathcal{T}_i$, $1 \leq i \leq N$, then there exists (by construction) an admissible control input capable of steering the one step evolution within the successor of the current set, i.e. $x_{k+1} \in \mathcal{T}_{i-1}$ as shown in Fig.*

*20. Therefore, starting from any state* $x_k \in \bigcup_{i=0}^{N}\{\mathcal{T}_i\}$, *the state trajectory of the system can be steered in at most* $N$ *steps in the terminal region and uniformly ultimately bounded within it. As a consequence,* $\bigcup_{i=0}^{N}\{\mathcal{T}_i\} \subseteq \mathcal{X}$ *represents the total DoA of the emergency controller.* □

As long as the actual online computation of the emergency controller action, namely $u_k^{em}$, is concerned, this can be obtained by resorting to the following receding-horizon computation algorithm, where $J(x_k, u)$ denotes a generic convex cost function of interest:

---

Emergency Set-Theoretic Controller **(E-STC)**

---

**Off-line computations:** $\{\mathcal{T}_i\}_{i=0}^{N}$, $\mathcal{X}_\eta \subseteq \bigcup_{i=0}^{N} \mathcal{T}_i \subseteq \mathcal{X}$

**On-line computations:** $u_k^{em}$

1: Find the smallest set index $i_k$ containing $x_k$, i.e.

$$i_k := min\{i : x_k \in \mathcal{T}_i\}$$

2: **if** $i_k == 0$ **then** $u_k^{em} = g_0(x_k, x_{eq}^{em})$

3: **else**

$$u_k^{em} = \arg\min_{u} J(x_k, u) \quad s.t. \tag{40}$$

$$Ax_k + Bu \in \tilde{\mathcal{T}}_{i_k-1},\ u \in \mathcal{U} \tag{41}$$

4: **end if**

5: $k \leftarrow k+1$ goto Step 1

---

where $J(x_k, u)$ is any convex cost function of interest. It is important to notice that for the **E-STC** controller the main computational demand is due to the optimization (40)-(41) which turns out to be a convex optimization solvable in polynomial time.

### 4.3.3 One-step attack-safe region

In this section, the concept of one-step attack-safe region for (36) is introduced and characterized.

**Definition 12.** *(**One-Step Attack Safe**) The system (36) is said one-step attack-safe under the action of the controller and regardless of any FDI attack scenario and disturbance realization, iff its one-step evolution, namely $x^+$, will remain confined withing the controller domain of attraction $\mathcal{X}_\eta$, i.e.*

$$x^+ := Ax_k + Bu_k + B_d d_k \in \mathcal{X}_\eta, \forall\, d_k \in \mathcal{D} \tag{42}$$

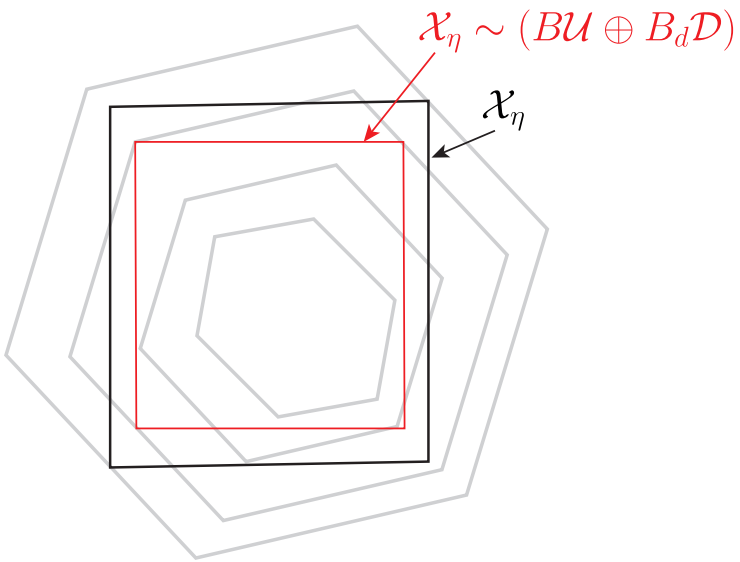



Figure 21: One-step attack safe region

**Proposition 3.** *Let us consider the networked control system (36)-(37), the state and input constraints (2), the tracking controller domain of attraction $\mathcal{X}_\eta$, and the current state vector $x_k$. The plant's state space evolution is guaranteed to be one-step attack-safe, regardless of any FDI attack and disturbance occurrences, i.e. $x^+ \in \mathcal{X}$, if the current state*

*$x_k$ belongs to $\mathcal{S}$ defined as follows:*

$$x_k \in \mathcal{S} := \{x \in \mathcal{X} : Ax \in (\mathcal{X}_\eta \sim (B\mathcal{U} \oplus B_d\mathcal{D}))\} \tag{43}$$

*Proof.* In an attack-free scenario ($u_k \equiv u'_k$ and $x_k \equiv x'_k$), the controller logic ensures that $u_k \in \mathcal{U}$ produces a one-step ahead evolution satisfying (42). Nevertheless, under a generic FDI attack scenario on the channels, e.g.

$$\begin{aligned} u_k^{r\prime} &= u_k^r + u_k^a \\ x'_k &= x_k + x_k^a \end{aligned} \tag{44}$$

Either with an attack on the state or on the actuation channel, the control input $u'_k$ received by the plant might be compromised and produces a plant evolution $x^+$ which is not expected (42) and a-priori unknown. It is only reasonable to assume that, for physical limitations on the power deliverable by the actuators, any input applied to the plant will be bounded within $\mathcal{U}$.

Then, starting from the current state $x_k$, it is possible to characterize the set of all admissible one-step ahead state evolutions as follows:

$$\mathcal{X}^+(x_k) = \{x^+ \in \mathbb{R}^n : \exists u \in \mathcal{U}, \exists d \in \mathcal{D} \;\; s.t. \;\; x^+ = Ax_k + Bu + B_d d \,\} \tag{45}$$

which can be rewritten in terms of minkowsky/pontryagin set sum as:

$$\mathcal{X}^+(x_k) = Ax_k + (B\mathcal{U} \oplus B_d\mathcal{D}) \tag{46}$$

Starting from (46), we can write the set of one-step ahead safe states $\mathcal{S}$ as

$$\mathcal{S} := \{x \in \mathcal{X} : Ax + (B\mathcal{U} \oplus B_d\mathcal{D}) \subseteq \mathcal{X}_\eta\} \tag{47}$$

which can be rewritten by using minkowsky/pontryagin set difference as

$$\mathcal{S} := \{x \in \mathcal{X} : Ax \in (\mathcal{X}_\eta \sim (B\mathcal{U} \oplus B_d\mathcal{D}))\} \tag{48}$$

concluding the proof. □

The set $\mathcal{S}$ will be hereafter used by the safety guard module to ensure that attack detection can be achieved at least one step before the attack could violate the safety of the plant.

### 4.3.4 Safety Guard

In this subsection, all the previous developed ingredients i.e.

- *Control Center* (*Assumption* 3);
- *Encoder* and *Decoder* scheme (4.3.1);
- *Emergency Controller* (4.3.2).
- *One-step attack-safe region* $\mathcal{S}$ (4.3.3);

are collected to develop a *Safety Guard* which aims at preventing the plant from reaching unsafe configurations, i.e. $x_k \notin \mathcal{X}$ regardless of any attack scenario. To this end, first the safety guard operations are described, then its effectiveness is proved.

The Safety Guard must activate the emergency controller as soon as a safety risk is detected, and must restore the normal plant operation when an attack-free scenario is recovered. Although the emergency controller can be activated instantaneously, the recovery phase should be done carefully to prevent that switching attacks [53] could produce instability. In the switching system related literature, this problem is well-known and different solutions, based on the concept of dwell-time, have been proposed e.g. [54]. A guaranteed, although not optimal, dwell-time $\tau \in \mathbb{Z}_+$ can be straightforwardly obtained by considering

a waiting time(dwell-time) equal to the number of one-step controllable set regions, i.e. $\tau \geq N$.

The Safety Guard pseudo-algorithm can be summarized as follows:

---

Safety Guard **(SG) -** $\forall k$

---

**Configuration:** dwell-time:=$\tau \geq N$

**Initialization:** attack-flag=0, counter=0.

**Output:** Control input applied to the plant $u_k^{plant}$

1: **if** $(u_k^{e\prime} \neq u_k^e \,||\, x_k \notin \mathcal{S} \,||\, u_k \notin \mathcal{U})$ **then**

2: attack-flag=1, counter=0;

3: **else**

4: **if** (attack-flag==1) **then** counter=counter+1;

5: **else** counter=0;

6: **end if**

7: **end if**

8: **if** (attack-flag==1 & counter$< \tau$) **then**

9: ▷ (Emergency Controller)

$$u_k^{plant} = u_k^{em}$$

10: **else** ▷ (Tracking Controller)

$$u_k^{plant} = u_k'$$

11: **end if**

---

**Remark 11.** *It is important to underline that the attack detection rules* $u_k^{e\prime} \neq u_k^e$ *in Step*

*1 of **SG** is sufficient to trigger an alarm for attacks on both the input and state vectors. While the firsts are straightforward in virtue of the adopted coding and decoding scheme (see 4.3.1), the seconds need to be further explained. In this chapter, the nature of the anomaly detector module in the Control Center is not specified (see Assumption 3) but a novel control architecture is proposed where secret information can be shared from the controller to the safety guard (see Remark 8). Therefore, if an attack is detected on the controller-side, a simple way to trigger a flag is to create a "fake" attack on the command signal. This will have the straightforward consequence of triggering the attack detection rule $u_k^{e\prime} \neq u_k^e$.*

**Proposition 4.** *By considering the networked control system (36), the tracking controller, the emergency controller algorithm (**E-STC**), the one-step attack-safe region (48), and the encoder and decoder (38)-(39) functions. The Safety Guard (**SG**) algorithm provides a solution for the objectives **(O1)** and **(O2)**.*

*Proof.* By collecting all the above developments, it is straightforward to prove that no admissible FDI attacks can put in risk the safety of the plant without being detected **(O1)**. Indeed, in virtue of the nature of the one-step attack safe region $\mathcal{S}$ (Section 4.3.3), in the worst-case, any attack is detected one-step before it could harm the plant (see the attack detection rule $x_k \notin \mathcal{S}$ in the Step 1 of **SG** algorithm). Moreover, since the emergency controller **E-STC** contains, by construction, the domain of the tracking controller $\mathcal{X}_\eta$ and the safe region $\mathcal{S}$, it can be safety activated regardless of the current state of the plant $x_k \in \mathcal{S}$. Moreover, the dwell-time condition $\tau \geq N$ ensures that the recovery of normal plant operation can be attempted only when the state of the system is surely contained withing the **E-STC** terminal region (see *Remark* 10). The latter limits the maximum admissible rate of switching attacks and, as a consequence, no attacks can bring the state of the system outside of the safety region (**O2**). □

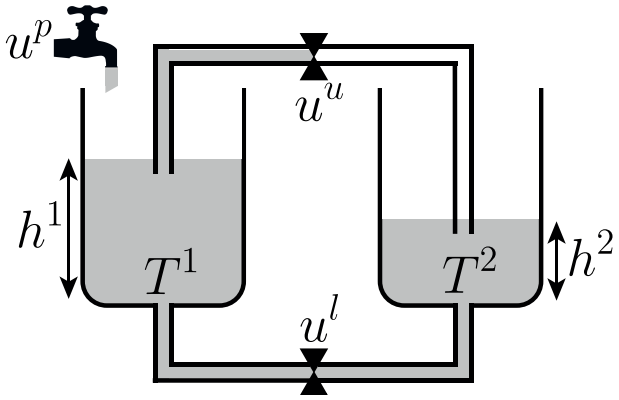


Figure 22: Two-Tank water system

## 4.4 Simulation Example

The two-tank water system [55] depicted in Fig.22 is here used to show the effectiveness of the proposed control architecture against cyber-attacks. The plant consists of two tanks, denoted with $T^1$ and $T^2$, which water's levels are $h^1$ and $h^2$, respectively. The input vector is $u = \left[u^p, u^l, u^u\right]^T$, where $u^p$ is the command input that regulate the valve injecting water within $T^1$, while $u^l$ and $u^u$ are the lower and upper valves between $T^1$ and $T^2$. The state-space vector of the system contains the water's levels of both tanks, namely $x = \left[h^1, h^2\right]^T$. The continuous-time nonlinear model of two-tank water system is linearized and discretized [56] using a sampling time $T_s = 1$ and the equilibrium pair $x_{eq} = \left[0.4, 0.06\right]^T$, $u_{eq} = \left[0.48, 0.75, 0.2\right]^T$. The resulting discrete-time linear model (34) is governed by the matrices:

$$A = \begin{bmatrix} 0.9931 & 0.0035 \\ 0.0068 & 0.9823 \end{bmatrix}, \quad B_d = - \begin{bmatrix} 0.9966 \\ 0.0034 \end{bmatrix} \times 10^{-3}$$

$$B = \begin{bmatrix} 0.0081 & -0.0032 & -0.0034 \\ 0 & 0.0032 & 0.0034 \end{bmatrix} \tag{49}$$

where the bounded disturbance $d \in \mathcal{D} = \{d : -10^{-3} \leq d \leq 10^{-3}\}$ models possible model mismatches and/or disturbance outflows. Notice that, in what follows, for sake of clarity, constraints, vectors, regions and figures are w.r.t. the linearized model.

The following state and input constraints are assumed:

$$\begin{gathered} \mathcal{U} : -0.5 \leq u^p \leq 1.5, \;\; -0.25 \leq u^l \leq 1.75, \;\; -0.8 \leq u^u \leq 1.2 \\ \mathcal{X} : 0.02 \leq h^1 \leq 0.60, \;\; 0.02 \leq h^2 \leq 0.60 \end{gathered} \tag{50}$$

Since the objective of the conducted simulations is to investigate the behavior of the proposed architecture in the "worst-case" scenario, we assume that the anomaly detector module is not available in the Control Center. The Command Governor in [56] is used to ensure constraints satisfaction. In the following simulations, the reference water levels are $r = [0.4,\, 0.3]^T$ and the emergency working condition (see Assumption 4) is

$$x_{eq}^{em} = [0.2525, 0.2834]^T, \;\; u_{eq}^{em} = [0.5, 0.5, 0.5]^T \tag{51}$$

According to the proposed architecture, to design the Emergency Controller, first the terminal state feedback controller is computed:

$$u_0^{em} = K_0(x - x_{eq}^{em}) + u_{eq}^{em},\; K_0 = \begin{bmatrix} -26.040 & -13.073 \\ 4.903 & -23.680 \\ 5.209 & -25.160 \end{bmatrix}$$

and the associated RPI region (see the green region in Fig. 25). Then a family or robust one-step controllable sets has been determined to cover the tracking controller domain $\mathcal{X}_\eta := \mathcal{X}$. In particular a family of 71 sets, $\{\mathcal{T}_i\}_{i=0}^{71}$, has been computed (see Fig. 25). Finally, an encoder by using 3 auxiliary inputs is designed, e.g. $p = 3$.

In the sequel, two different attack scenarios are investigated: "*Attack on the Actuation Channel*" and "*Stealthy Attack on the Measurement Channel*."

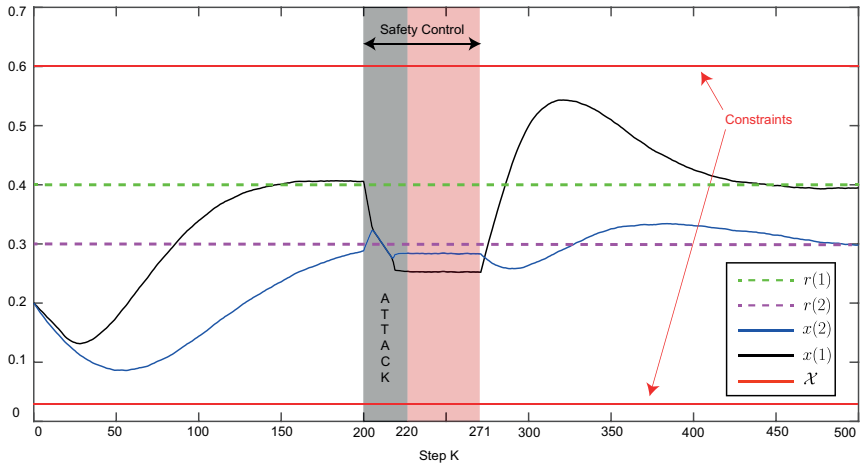



Figure 23: Case A: States Evolution

### 4.4.1 Attack on the Actuation Channel

In the first scenario, the attacker performs an FDI attack on the actuation channel. The attack scenario can be summarized as follows:

$$\text{Attack (Case A)}: \begin{cases} \text{Start} & k = 200 \\ \text{End} & k = 225 \\ \text{Action} & u_k^{r\prime} = u_k^r + u_k^a \end{cases}$$

Where $u_k^a = \Big[1, 1, 1, 1, 1, 1\Big]^T$ is the attack control signal. The simulation results are shown in Fig. 23. It is shown that for $k < 200$ the tracking controller, starting from $x_0 = [0.2,\ 0.2]^T$, is capable to steer the states of the system towards the desired reference. At $k = 200$, an FDI attack on the input is attempted. Nevertheless, since the attacker is not aware of randomness of the proposed encoding/decoding scheme, its presence is trivially instantaneously detected by the safety guard because ${u_{200}^e}' \neq u_{200}^e$, see Step 1 of the **SG** algorithm. The latter, has the consequences of activating the **E-STC** emergency controller (see Step 9 of the **SG** algorithm) and the tracking control action until safety of the channel and of the plant are re-ensured (i.e counter$\leq$ 71 and attack-flag=0). The controller **E-STC**, activated at $k = 200$, is capable of steering the plant trajectory $x_k$,

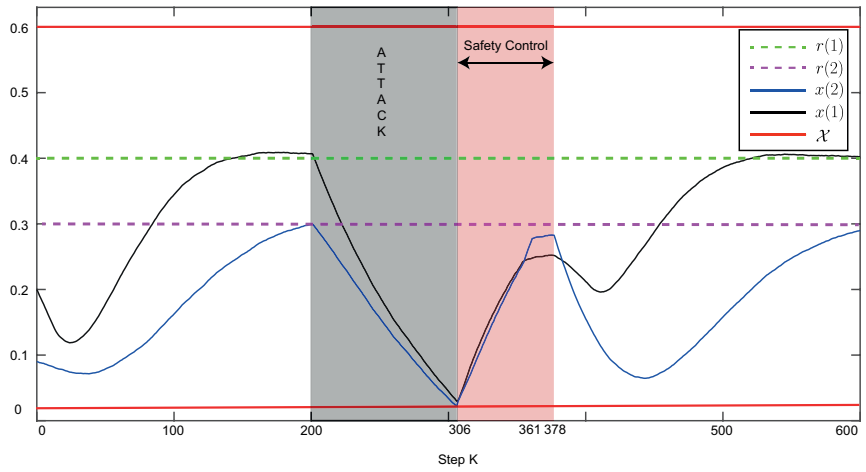



Figure 24: Case B: States Evolution

without violating the plant constraints, within the safety RPI region $\mathcal{T}_0$ at $k = 220$. Finally, at $k = 271$, since the condition in Step 10 of **SG** is satisfied and the input attack is terminated, the networked tracking controller is resumed and the plant starts tracking again the reference signal.

### 4.4.2 Stealthy Attack on the Measurement Channel

In the second scenario, the attacker performs an FDI attack on the state measurements, and it is assumed that there is no detector in the Control Center. The attack scenario is the following:

$$\text{Attack (Case B)}: \begin{cases} \text{Start} & k = 200 \\ \text{End} & k = 350 \\ \text{Action} & x'_k = x_k + x^a_k \end{cases}$$

where $x^a_k = \begin{bmatrix} 0.2, 0.2 \end{bmatrix}^T$ is the bias injected. The simulations results are collected in Figs. 24-25. Since it is assumed that there is no detector in the Control Center, the attack cannot be revealed neither by the control center nor by the proposed input encoding/decoding scheme. Nevertheless, it is proved (see *Proposition* 4) that no attacks can harm the plant while remaining undetected. The latter still holds true in the considered worst-case

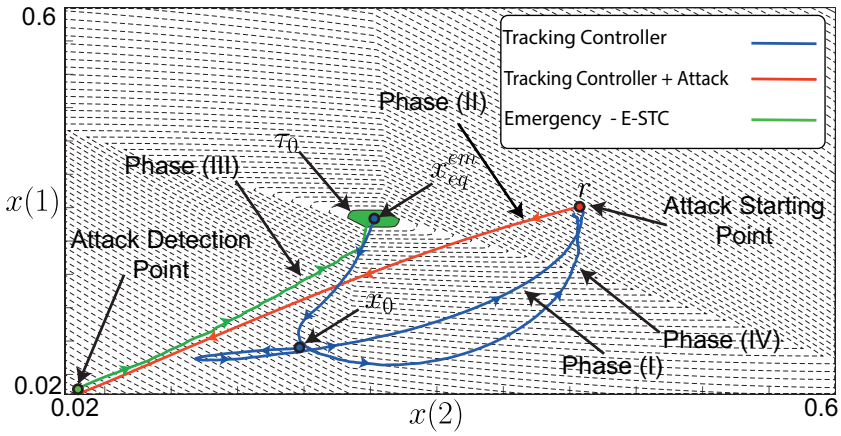



Figure 25: Case B: State Trajectory. The state trajectory can be divided in 4 phases: phase (I) - the networked tracking controller is active (blue line), phase (II) - an FDI attack is started but not yet detected (red line), phase (III) - the attack is detected and E-STC is activated (green line), phase (IV) - the attack is over and the tracking controller is reactivated (blue line).

scenario and this is testified by the state trajectory shown in Fig. 25 which is confined withing the plant state constraints. In particular, the attack starts at $k = 200$ and its consequence is that the tracking controller (misleaded by the received corrupted state measurements) starts bringing the state of the system outside of the admissible state space region (red trajectory in Fig. 25). This is not revealed until $k = 306$ when the safety condition $x_{306} \notin \mathcal{S}$ (Step 1 of the **SG** algorithm) is violated. Therefore, At $k = 306$ the attack detection is accomplished and the safety controller is activated. Finally, similarly to what commented for the first scenario, the state trajectory, at $k = 361$, first safely reaches the emergency RPI region (see green region in Fig. 24) and then, at $k = 378$, the tracking controller is re-activated.

# Chapter 5

# Conclusion and Future Work

## 5.1 Conclusion

In this thesis, the problem of detecting and mitigating of FDI attacks in networked control systems was considered. In the introduction, existing solutions to the attack detection problem were explained and their advantages and disadvantages were highlighted. Due to the importance of the research topic and the drawbacks of the available methods in the literature, two novel control architectures were proposed in order to detect FDI attacks affecting networked control systems.

In Chapter 3, the watermarking and moving target detection ideas are jointly exploited to design a novel architecture capable of detecting FDI attacks. Contrary to watermarking idea in [11], where the watermarked input affects the system performances, our solution does not suffer from the same drawback. As a consequence, while in [11], the watermarking signal must be chosen to obtain the best trade-off between detection and performance, in our architecture the amplitude is a free design parameter that can be tuned to achieve the desired detection rate. With respect to existing covert detection solutions, namely moving target/auxiliary system [15, 16], our approach has the advantages of using a static auxiliary system that its dynamics are not coupled with the physical plant dynamics or with the detection mechanism. As a consequence, it can be installed on the existing

NCS without having to affect the existing communication infrastructure, estimation or detection schemes. Finally, simulation results for a four tanks system was shown to testify the effectiveness of the proposed architecture and its advantages against the competitor schemes.

In Chapter 4, first, we propose a detection scheme based on auxiliary inputs in order to deteriorate the disclosure and disruption resources of the attacker. Then, a safety guard is proposed to switch the plant's controller to a local emergency controller whenever an attack is detected or safety conditions of the plant are at risk. Although the local controller cannot guarantee reference tracking, its aim is to guarantee the plant constraints satisfaction until an attack-free scenario is recovered. Such a controller has been designed by resorting to a set-theoretic MPC scheme capable of steering the state of the system within an RCI region centered in a-priori defined emergency equilibrium point. Finally, numerical results on a two tanks water system are provided to show the effectiveness of the proposed architecture.

## 5.2 Future Work

Some suggestions for future research in this area are outlined below:

- In chapter 3, an auxiliary system is proposed for attack detection. In particular, the required conditions for an effective auxiliary system were defined. Nevertheless, it did not propose how to optimize the design of the auxiliary function to maximize the detection sensitivity.

- The architecture proposed in chapter 3 suffers from lack of the mitigation actions. Designing a controller for attack compensation can be another extension to this work.

- In chapter 4, once the safety guard decides to switch to the emergency controller, we have to wait until the plant's states converge to the terminal region. In other words, we have to wait for N steps; however, the attack might be removed before

the waiting time. The recovery procedure can be optimized by designing a real-time monitoring to recover the normal behavior as soon as possible.

# References


[1] M. N. Al-Mhiqani, R. Ahmad, W. Yassin, A. Hassan, Z. Z. Abidin, N. S. Ali, and K. H. Abdulkareem, "Cyber-security incidents: a review cases in cyber-physical systems," *International Journal of Advanced Computer Science and Applications*, vol. 9, pp. 499–508, 2018.

[2] J. Slay and M. Miller, "Lessons learned from the maroochy water breach," *International Conference on Critical Infrastructure Protection*, pp. 73–82, 2007.

[3] T. Chen, "Stuxnet, the real start of cyber warfare?[editor's note]," *IEEE Network*, vol. 24, no. 6, pp. 2–3, 2010.

[4] N. Jazdi, "Cyber physical systems in the context of industry 4.0," *IEEE International Conference on Automation, Quality and Testing, Robotics*, pp. 1–4, 2014.

[5] G. Xiong, F. Zhu, X. Liu, X. Dong, W. Huang, S. Chen, and K. Zhao, "Cyber-physical-social system in intelligent transportation," *IEEE/CAA Journal of Automatica Sinica*, vol. 2, no. 3, pp. 320–333, 2015.

[6] S. Sridhar, A. Hahn, M. Govindarasu *et al.*, "Cyber-physical system security for the electric power grid." *Proceedings of the IEEE*, vol. 100, no. 1, pp. 210–224, 2012.

[7] S. M. Dibaji, M. Pirani, D. B. Flamholz, A. M. Annaswamy, K. H. Johansson, and A. Chakrabortty, "A systems and control perspective of cps security," *Annual Reviews in Control*, 2019.

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

[9] Y. Chen, S. Kar, and J. Moura, "Dynamic attack detection in cyber-physical systems with side initial state information," *IEEE Transactions on Automatic Control*, vol. 62, no. 9, pp. 4618–4624, 2017.

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

[11] Y. Mo and B. Sinopoli, "Secure control against replay attacks," *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 911–918, 2009.

[12] F. Miao, M. Pajic, and G. J. Pappas, "Stochastic game approach for replay attack detection," *IEEE Conference on Decision and Control (CDC)*, pp. 1854–1859, 2013.

[13] F. Miao, Q. Zhu, M. Pajic, and G. J. Pappas, "Coding schemes for securing cyber-physical systems against stealthy data injection attacks," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 106–117, 2017.

[14] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1806–1813, 2012.

[15] S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," *IEEE Conference on Decision and Control (CDC)*, pp. 5820–5826, 2015.

[16] C. Schellenberger and P. Zhang, "Detection of covert attacks on cyber-physical systems by extending the system dynamics with an auxiliary system," *IEEE Conference on Decision and Control (CDC)*, pp. 1374–1379, 2017.

[17] N. Forti, G. Battistelli, L. Chisci, and B. Sinopoli, "A bayesian approach to joint attack detection and resilient state estimation," *IEEE Conference on Decision and Control (CDC)*, pp. 1192–1198, 2016.

[18] A. Hoehn and P. Zhang, "Detection of covert attacks and zero dynamics attacks in cyber-physical systems," *American Control Conference (ACC)*, pp. 302–307, 2016.

[19] R. S. Smith, "Covert misappropriation of networked control systems: Presenting a feedback structure," *IEEE Control Systems*, vol. 35, no. 1, pp. 82–92, 2015.

[20] Y. Shoukry, P. Nuzzo, A. Puggelli, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and P. Tabuada, "Secure state estimation for cyber-physical systems under sensor attacks: A satisfiability modulo theory approach," *IEEE Transactions on Automatic Control*, vol. 62, no. 10, pp. 4917–4932, 2017.

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

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

[23] S. Zonouz, K. M. Rogers, R. Berthier, R. B. Bobba, W. H. Sanders, and T. J. Overbye, "Scpse: Security-oriented cyber-physical state estimation for power grid critical infrastructures," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 1790–1799, 2012.

[24] W. Lucia, B. Sinopoli, and G. Franze, "A set-theoretic approach for secure and resilient control of cyber-physical systems subject to false data injection attacks," *SOSCYPS*, pp. 1–5, 2016.

[25] X. Jin, W. M. Haddad, and T. Yucelen, "An adaptive control architecture for mitigating sensor and actuator attacks in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 6058–6064, 2017.

[26] M. Ghaderi, K. Gheitasi, and W. Lucia, "A novel control architecture for the detection of false data injection attacks in networked control systems," *American Control Conference (ACC)*, pp. 139–144, 2019.

[27] K. Gheitasi, M. Ghaderi, and W. Lucia, "A novel networked control scheme with safety guarantees for detection and mitigation of cyber-attacks," *European Control Conference (ECC)*, pp. 1449–1454, 2019.

[28] W. Lucia, K. Gheitasi, and M. Ghaderi, "A command governor based approach for detection of setpoint attacks in constrained cyber-physical systems," *IEEE Conference on Decision and Control (CDC)*, pp. 4529–4534, 2018.

[29] K. Ogata, "Modern control engineering," *Prentice-Hall Electrical Engineering Series, Englewood Cliffs: Prentice-Hall,— c1970*, 1970.

[30] C. Murguia and J. Ruths, "Cusum and chi-squared attack detection of compromised sensors," *IEEE Conference on Control Applications (CCA)*, pp. 474–480, 2016.

[31] C. Kwon, W. Liu, and I. Hwang, "Security analysis for cyber-physical systems against stealthy deception attacks," *2013 American control conference*, pp. 3344–3349, 2013.

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

[33] R. Tunga, C. Murguia, and J. Ruths, "Tuning windowed chi-squared detectors for sensor attacks," *American Control Conference (ACC)*, pp. 1752–1757, 2018.

[34] A. Shostack, *Threat modeling: Designing for security.* John Wiley & Sons, 2014.

[35] N. Forti, G. Battistelli, L. Chisci, and B. Sinopoli, "A bayesian approach to joint attack detection and resilient state estimation," *IEEE Conference on Decision and Control (CDC)*, pp. 1192–1198, 2016.

[36] W. H. Kwon and S. H. Han, *Receding horizon control: model predictive control for state models*. Springer Science & Business Media, 2006.

[37] S. V. Raković and W. S. Levine, *Handbook of model predictive control*. Springer, 2018.

[38] F. Blanchini, "Set invariance in control," *Automatica*, vol. 35, no. 11, pp. 1747–1767, 1999.

[39] D. Bertsekas and I. Rhodes, "Recursive state estimation for a set-membership description of uncertainty," *IEEE Transactions on Automatic Control*, vol. 16, no. 2, pp. 117–128, 1971.

[40] D. Angeli, A. Casavola, G. Franzè, and E. Mosca, "An ellipsoidal off-line mpc scheme for uncertain polytopic discrete-time systems," *Automatica*, vol. 44, no. 12, pp. 3113–3119, 2008.

[41] F. Borrelli, A. Bemporad, and M. Morari, *Predictive control for linear and hybrid systems*. Cambridge University Press, 2017.

[42] S. V. Rakovic, E. C. Kerrigan, K. I. Kouramas, and D. Q. Mayne, "Invariant approximations of the minimal robust positively invariant set," *IEEE Transactions on Automatic Control*, vol. 50, no. 3, pp. 406–410, 2005.

[43] F. Blanchini, "Ultimate boundedness control for uncertain discrete-time systems via set-induced lyapunov functions," *IEEE Transactions on Automatic Control*, vol. 39, no. 2, pp. 428–433, 1994.

[44] A. A. Kurzhanskiy and P. Varaiya, "Ellipsoidal techniques for reachability analysis of discrete-time linear systems," *IEEE Transactions on Automatic Control*, vol. 52, no. 1, pp. 26–38, 2007.

[45] M. Althoff, O. Stursberg, and M. Buss, "Verification of uncertain embedded systems by computing reachable sets based on zonotopes," *IFAC Proceedings Volumes*, vol. 41, no. 2, pp. 5125–5130, 2008.

[46] P. Griffioen, S. Weerakkody, and B. Sinopoli, "An optimal design of a moving target defense for attack detection in control systems," *American Control Conference (ACC)*, pp. 4527–4534, 2019.

[47] Y. Mo and B. Sinopoli, "On the performance degradation of cyber-physical systems under stealthy integrity attacks," *IEEE Trans. on Automatic Control*, vol. 61, no. 9, pp. 2618–2624, 2015.

[48] B. Ripley, "Thoughts on pseudorandom number generators," *Journal of Computational and Applied Mathematics*, vol. 31, no. 1, pp. 153–163, 1990.

[49] K. H. Johansson, "The quadruple-tank process: A multivariable laboratory process with an adjustable zero," *IEEE Transactions on Control Systems Technology*, vol. 8, no. 3, pp. 456–465, 2000.

[50] I. Hwang, S. Kim, Y. Kim, and C. E. Seah, "A survey of fault detection, isolation, and reconfiguration methods," *IEEE Transactions on Control Systems Technology*, vol. 18, no. 3, pp. 636–653, 2010.

[51] F. Blanchini and S. Miani, *Set-theoretic methods in control*. Springer, 2008.

[52] G. Goodwin, M. M. Seron, and J. A. De Doná, *Constrained control and estimation: an optimisation approach*. Springer Science & Business Media, 2006.

[53] S. Z. Yong, M. Zhu, and E. Frazzoli, "Resilient state estimation against switching attacks on stochastic cyber-physical systems," *IEEE Conference on Decision and Control (CDC)*, pp. 5162–5169, 2015.

[54] D. Liberzon, *Switching in systems and control*. Springer Science & Business Media, 2003.

[55] T. Heckenthaler and S. Engell, "Approximately time-optimal fuzzy control of a two-tank system," *IEEE Control Systems Magazine*, vol. 14, no. 3, pp. 24–30, 1994.

[56] G. Franzè and W. Lucia, "A set-theoretic control architecture for constrained switching systems," *American Control Conference (ACC)*, pp. 685–690, 2016.