Cyber-Attack Detection Methodologies for Cyber-Physical Systems: A System Theoretic Approach

Mahdi Taheri

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Electrical and Computer Engineering) at

Concordia University

Montréal, Québec, Canada

May 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mahdi Taheri**

Entitled: **Cyber-Attack Detection Methodologies for Cyber-Physical Systems: A System Theoretic Approach**

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

Doctor of Philosophy (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. Javad Dargahi

_____ External Examiner
Dr. Bahram Shafai

_____ Examiner
Dr. Rastko Selmic

_____ Examiner
Dr. Youmin Zhang

_____ Examiner
Dr. Shahin Hashtrudi Zad

_____ Supervisor
Dr. Khashayar Khorasani

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

_____ 2024

_____
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Cyber-Attack Detection Methodologies for Cyber-Physical Systems: A System Theoretic Approach

Mahdi Taheri, Ph.D.

Concordia University, 2024

Cyber-physical systems (CPS) are integral to critical infrastructures such as power networks, transportation systems, and water treatment networks. Despite the advancements in developing more secure CPS and monitoring systems, the number of successfully executed cyber-attacks in CPS has increased over the past decade. The mentioned cyber-attacks, which can make CPS unstable, are performed by intelligent adversaries who try to maintain their malicious attacks undetected. This thesis addresses several crucial challenges related to cyber-attacks in CPS and multi-agent systems (MAS).

The first part of the thesis focuses on simultaneous cyber-attacks and fault detection and isolation (CAFDI) in centralized and large-scale interconnected CPS. Proposed methodologies include centralized and distributed CAFDI approaches, incorporating two filters and an unknown input observer (UIO)-based detector to identify various deception attacks such as covert, zero dynamics, and replay attacks. The effectiveness of the distributed CAFDI approach is demonstrated through a hardware-in-the-loop (HIL) simulation of a four-area power network system.

The second part studies stealthy cyber-attacks in CPS, particularly zero dynamics, covert, and controllable attacks. Conditions for executing these attacks are derived from CPS Markov parameters and the system observability matrix. Dynamic coding schemes are proposed as countermeasures, increasing the number of actuators needed to execute cyber-attacks.

In the third part, zero dynamics and undetectable cyber-attacks in linear and nonlinear CPS are explored. A new security metric, security effort (SE), is introduced to determine the minimum number of secured actuators and sensors required to prevent such attacks in linear CPS. For nonlinear CPS, the study uses Koopman operator theory and the extended dynamic mode decomposition (EDMD) algorithm to create a

finite-dimensional linear representation of the system to identify critical sensor measurements that need securing to prevent zero dynamics and covert attacks.

The fourth part addresses privacy-preserving consensus control, controllability cyber-attacks, undetectable cyber-attacks, and detection methodologies in MAS. A distributed transformation-based consensus control method is developed to protect agent privacy from eavesdroppers. Conditions for adversaries to control the MAS network by attacking a few agents are explored, defining these as controllability cyber-attacks. Undetectable cyber-attacks in MAS are defined and an event-triggered detection module to detect such attacks is proposed.

Acknowledgments

First and foremost, I extend my sincere appreciation to my dissertation advisor, Prof. K. Khorasani, for his invaluable guidance, mentorship, and unwavering support. Your expertise and encouragement have been instrumental in shaping the direction of my research and academic growth.

I am also indebted to Prof. N. Meskin and Prof. I. Shames for their significant contributions in writing papers, providing insightful ideas, and offering invaluable technical comments. Your guidance and collaboration have enhanced the quality and depth of my work.

I extend my heartfelt thanks to the members of my dissertation committee, Prof. R. Selmic, Prof. Y. Zhang, Dr. S. Hashtrudi Zad, Prof. J. Dargahi, and Prof. B. Shafai, for their expert evaluation, constructive feedback, and invaluable suggestions.

Special appreciation goes to Dr. Amir Baniamerian for his friendship, support, and encouragement throughout my Ph.D. journey. Your presence and camaraderie have made the challenges more manageable and the successes more meaningful.

I am also grateful to my friends in Montreal, Hassan, Mohammadreza, Kaoutar, and other lab members whose companionship have made the journey of my Ph.D. enjoyable and memorable.

To my parents, Mehri and Ahmad, I owe a profound debt of gratitude for their unconditional love, encouragement, and sacrifices. Without your support and guidance, I would not have been able to pursue and complete my Ph.D. journey.

I would also like to acknowledge the constant support and encouragement of my family members, Kharaman, Sepideh, Mohammadreza, Fereshteh, and Vagar. Your belief in me has been a source of strength and motivation throughout this endeavor.

Table of Contents

List of Figures

List of Tables

List of Acronyms

C&C	Command and Control
CAFDI	Cyber Attacks and Faults Detection and Isolation
CPS	Cyber-Physical Systems
DoS	Denial of Service
EDMD	Extended Dynamic Mode Decomposition
FDI	Fault Detection and Isolation
IoT	Internet of things
KCT	Koopman Canonical Transform
LTI	Linear Time-Invariant
MAS	Multi-Agent Systems
PLCs	Programmable Logic Controllers
SCADA	Supervisory Control and Data Acquisition

SISO	Single-Input Single-Output
UIO	Unknown Input Observer

List of Symbols

$0_{n\times m}$	Matrix with n rows, m columns, and all enteries equal to zero
$A^\dagger$	Moore-Penrose pseudoinverse of matrix A
$A^\top$	Transpose of matrix A
$L_{f_i}h_i(x)$	Lie Derivative of $h_i(x)$ along $f_i(x)$
$\|\cdot\|_2$	$\|x\|_2$, the Euclidean norm of x
$\|\cdot\|_\infty$	$\|x\|_\infty$, the infinity norm of the vector x
$\|\cdot\|_{\mathrm{F}}$	$\|x\|_{\mathrm{F}}$, the Forbenius norm of the vector x
$\mathbb{C}$	Set of complex numbers
$\mathbb{N}$	Set of natural numbers
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^{n\times m}$	Set of matrices with entries in $\mathbb{R}$, n rows, and m columns
Im$(\cdot)$	Im(A), range space of matrix A
Ker$(\cdot)$	Ker(A), null space of matrix A
k	Discrete temporal variable
t	Continues time

Chapter 1

Introduction

1.1 Motivation

Cyber-physical systems (CPS) are monitored and controlled by sensors, actuators, and computational capabilities of embedded computers, and are linked via communication networks [1]. Our today's life massively depends on CPS due to their wide range of applications in different areas, such as power systems and smart grid, next generation aerospace and transportation systems, Internet of things(IoT), unmanned aerial vehicles (UAV), process control and water treatment networks [2, 3]. These systems provide us with unique capabilities and high level performance and reliability in performing complex tasks [4].

Supervisory Control and Data Acquisition(SCADA) systems are considered as CPS [5]. These systems have been utilized in controlling and monitoring critical infrastructures such as electrical power distribution networks, oil and gas pipelines, and water distribution and water treatment plants. The disruption of SCADA controlling units in any of the mentioned systems can lead to significant economic losses at a nationwide scale.

In recent years, due to the increased use of wireless communication networks in control systems, such as SCADA, security concerns related to these systems have been raised [5–8]. For instance, in 2000, the SCADA system at Maroochy Water Services in Queensland, Australia, was subject to an attack through its communication network [8]. A former employee, by using a laptop and a radio transmitter, was able to take control over 150 pumping stations for three months [8]. In June 2010, it was discovered that the Iranian

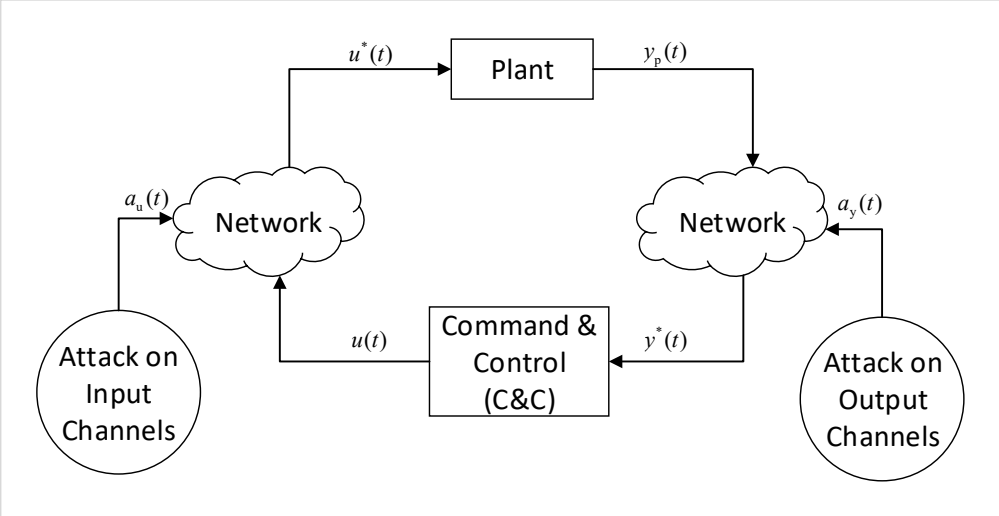

Figure 1.1: Cyber-physical system under deception attack on both input and output channels, where $u(t)$ denotes the control command, $a_{\mathrm{u}}(t)$ represents the cyber-attack signal on the input channel, $u^*(t)$ represents the control input of the plant, $y_{\mathrm{p}}(t)$ denotes the output on the plant side, $a_{\mathrm{y}}(t)$ denotes the attack signal on the output channel, and $y^*(t)$ denotes the output on the C&C side.

nuclear facilities were struck by the Stuxnet computer worm [9]. The Stuxnet targeted SCADA systems in Iranian facilities and compromised Siemens programmable logic controllers (PLCs) to manipulate the electrical power fed to the motor of gas centrifuges for short intervals to increase their speed beyond their nominal limits, which eventually resulted in the breakdown of a number of operational centrifuges [9]. These incidents indicate the existence of a potential cyber threat against communication networks of CPS, which cannot be tackled by conventional fault detection and isolation (FDI) methodologies.

Anomalies in physical components of CPS such as actuators and sensors are either cyber-attacks or faults. Generally speaking, there are three types of cyber-attacks, namely integrity attacks in which trustworthiness of communicated data are compromised by adversaries and result in *deception*, availability attacks which are considered as the lack of system's availability due to cyber-attacks lead to *denial of service* (DoS), and confidentiality attacks in which unauthorized adversaries can read system's information which causes *disclosure* of information [10]. In deception (or integrity) type of cyber-attacks an adversary changes the transmitted information of the system's input or output by compromising the CPS network communication channels as depicted in Figure 1.1. In this work, only deception cyber-attacks are considered.

In order to protect the CPS against malicious cyber-attacks, one needs to develop methodologies and

monitoring systems that to preserve the privacy of communicated data in the communication networks and to detect violations of data integrity. Hence, in this thesis, various methodologies and monitoring systems for addressing the problem of cyber-attack detection in the CPS are developed and studied.

1.2 Literature Review

1.2.1 Centralized Cyber-Physical Systems (CPS)

Anomalies and machine induced faults as well as malicious cyber-attacks in physical components of the CPS do occur and are observed in actuators and sensors. In recent years, cybersecurity challenges in CPS, that include cyber-attacks on communication networks have attracted significant interest [2, 11–13]. A special type of cyber-attacks is defined as the deception attack in which an adversary changes the transmitted information of the system's input or output by compromising the CPS network communication channels. Covert attacks, controllable attacks, and zero dynamics attacks are defined as undetectable attacks [14–17], since they have no impact on the received output measurements on the command and control (C&C) side of the CPS.

In [1] and [18], it has been shown that if the CPS have non-minimum phase zeros, adversaries can perform zero dynamics attacks and make the system internally unstable, while the sensor measurements are not affected by the attack signals. It has been demonstrated in [14] and [15] that both zero dynamics attacks and controllable attacks belong to the weakly unobservable subspaces of the CPS. Furthermore, controllable attacks are in the controllability subspace within the weakly unobservable subspace of the system [15]. Also, in [19], data-driven approaches have been utilized to derive sufficient conditions under which adversaries can carry out zero dynamics attacks. However, necessary and sufficient conditions in terms of the required disruption resources and system knowledge for performing the zero dynamics and controllable attacks have not been fully studied in the above work.

The minimum number of required actuators and sensors, i.e., the disruption resources, to execute an undetectable or a perfectly undetectable cyber-attack such as the zero dynamics attacks, controllable attacks, and covert attacks has been defined as the security index for the CPS [16, 20–22]. In the case of undetectable cyber-attacks, if the initial conditions of the CPS are known, the impact of the cyber-attack can still be

detected in the sensor measurements [2,20], but perfectly undetectable attacks leave no impact on the outputs of the CPS [22, 23].

Various monitoring systems and active cyber-attack detection methodologies have been proposed to either detect stealthy cyber-attacks or prevent adversaries from executing them [17, 24–32]. Coding schemes [29], modulation matrices [27], moving target approaches [33], and watermarking schemes [17, 28] which are considered as active cyber-attack detection methodologies have been developed and employed that distort the system knowledge from adversaries point of view and prevent them from executing stealthy cyber-attacks.

To detect stealthy cyber-attacks on output measurements in the CPS such as replay attacks, a bank of multiplicative sensor watermarking filters was proposed in [17]. In [24], geometric theory was used to define zero dynamics attacks and show their impact on the system. A method was also proposed to add perturbations to the system matrices of the system (A, B, C) to change the zero dynamics of the system so that the adversary can no longer excite these new zero dynamics modes. In [29], a sensor coding method was proposed that reveals stealthy false data injection attacks by changing the direction of cyber-attacks where an algorithm to compute the coding matrices was designed, and finally, a time-varying coding approach was developed for the case when the adversary is capable of estimating a static coding matrix. A two-way coding scheme is developed in [30], which in addition to distorting the adversary's system knowledge, under certain conditions for single-input single-output (SISO) systems, it can change the non-minimum phase zeros to minimum phase ones.

On the other hand, for cyber-attack detection and monitoring systems, fault detection methods have been utilized. However, due to the inherent differences between cyber-attacks and machine induced faults, in some cases, such as covert attacks, zero dynamics attacks, and replay attacks fault detection methods fail to detect cyber-attacks. This is due to the fact that faults represent structural physical anomalies in the system, whereas cyber-attacks are injected intentionally by an intelligent adversary with the purpose of damaging the nominal behavior of the system without being detected. Consequently, conventional fault diagnosis algorithms should be fundamentally generalized to accommodate the malicious intelligent adversary cyber-attacks threats.

As a brief overview, the geometric-based fault detection methodologies were proposed in [34, 35] to

obtain necessary and sufficient conditions for existence of observers that can be used to generate a residual signal for the purpose of fault detection and isolation (FDI). In addition to geometric approaches, many algebraic model-based FDI methods have been introduced in the literature, such as unknown input observer (UIO) [36, 37], interacting multiple model [38], multiple model [39], distributed detection algorithms [40, 41], and parity equation based approaches [42, 43].

In [44], a distributed covert attack detection methodology is proposed that utilizes two local observers for each subsystem in a large-scale interconnected CPS. The proposed work in [44] assumes the existence of certain secure communication channels among subsystems and that the adversaries are capable of performing covert attack in one subsystem at any instance of time. Hence, each subsystem can detect covert cyber-attacks in the neighboring subsystems. In [25], monitoring systems that utilize auxiliary filters have been developed to detect stealthy cyber-attacks such as the zero dynamics attacks. In [45], the state-space model of a linear system was augmented by adding switching auxiliary dynamics that are unknown to the adversary and a switched Luenberger observer was designed to detect covert and zero dynamics attacks, however, for implementation purposes the extended system and the switched observer need to be synchronized.

In all the above work, the problem of detection and isolation of cyber-attacks and faults has not been addressed. Hence, by using the above methods, a fault in the system can misleadingly be detected as a cyber-attack and vice versa. Consequently, if a fault is misleadingly detected as a cyber-attack, the CPS operators may consider a certain course of action, and therefore countermeasure strategies that are designed to cope with the cyber-attack threats will not resolve the machine induced fault problem and will not recover the CPS. On the other hand, if a cyber-attack is misleadingly detected as a fault, the CPS operators cannot resolve the problem by utilizing fault-tolerant control methodologies or by repairing components of the CPS that are misleadingly diagnosed to be defective. This issue has been taken into consideration in [46] where a methodology to detect and isolate faults and cyber-attacks has been suggested. An event triggered adaptive estimator is designed and proposed in [46] which can be used to isolate sensor replay attacks and sensor faults. In [47], Bayesian Network models have been utilized and constructed to distinguish between cyber-attacks and faults in sensor measurements for floodgates in water management infrastructures.

Due to stealthiness of covert and zero dynamics attacks, it is of paramount importance to develop

methodologies that can be used to detect and isolate them. However, in [46] and [47], undetectable cyber-attacks such as covert attacks and zero dynamics attacks have not been considered. In addition, due to existence of physical component faults in the CPS, one needs to also clearly detect and isolate actuator and sensor faults and cyber-attacks on actuators and sensors. Moreover, in [46] and [47], only detection and isolation of sensor faults and cyber-attacks on sensor measurements have been investigated.

The security index (SI) for perfectly undetectable cyber-attacks in the CPS is defined as the minimum number of actuators and sensors that should be compromised by adversaries to execute a perfectly undetectable cyber-attack [15, 16, 21, 22]. Computing the security index is an NP-hard problem [21, 48]. Therefore, in [21] and [22] structural system framework has been utilized to describe the CPS by using graph theory to compute the security index in a generic sense by using computationally efficient algorithms. In [15] and [16], an upper bound of the security index is defined and geometric control theory is utilized to compute the security index over the weakly unobservable and controllable weakly unobservable subspaces of the CPS. However, the notion of security index considers the CPS from the adversary's point of view. Hence, in certain cases, it may not provide CPS operators with adequate information to prevent zero dynamics attacks, covert attacks, and controllable attacks. Consequently, one needs to study a security measure that studies the CPS from the operator's perspective.

As for the case of nonlinear CPS, the stealthiness of zero dynamics cyber-attacks for the quadruple-tank process is investigated in [49] and it is shown that for a finite amount of time the executed zero dynamics cyber-attack remains stealthy. Furthermore, in [50], a method to implement a stealthy type of cyber-attacks for a class of nonlinear CPS is introduced. However, all the above works have assumed that adversaries have a complete knowledge of the dynamics of the CPS, which may not be the case always.

The Koopman operator provides a linear infinite-dimensional representation of a given nonlinear system which can be used in spectral analysis of nonlinear flows and dynamics [51–54]. In [55], by employing Koopman eigenfunctions, Koopman eigenvalues, and Koopman modes, Koopman canonical form (KCF) for nonlinear control affine systems has been introduced, which is then used to develop an observer for the system. Considering that the Koopman operator is infinite-dimensional, it will be challenging to employ and apply tools and methods that are available in linear systems to infinite-dimensional representation of a given nonlinear system that the Koopman operator yields. Hence, data-driven algorithms such as the dynamic

mode decomposition (DMD) and extended dynamic mode decomposition (EDMD) have been utilized to come up with a linear finite-dimensional representation of a given nonlinear system for developing data-driven model predictive controllers and data-driven fault diagnosis methodologies [56–60].

1.2.2 Multi-Agent Systems (MAS)

Multi-agent systems (MAS), due to their wide range of applications, such as in unnamed aerial vehicles (UAV), next generation aerospace and transportation systems, autonomous and drive-less cars, have been a major topic of research during the past decade [40, 41, 61–63]. One of the challenges in MAS is to reach a consensus among the agents in a distributed manner. This problem has been addressed for systems having various types of linear and nonlinear dynamics [61, 64–66]. To achieve consensus among agents, each agent needs to transmit its information to its nearest neighboring agents. This communication is carried out through network channels that exist among the agents.

The existence of communication networks makes the multi-agent systems vulnerable to cyber-attacks. Suppose a group of agents are on an intelligence, surveillance, and reconnaissance (ISR) mission and an intelligent adversary performs an attack on the incoming communication links for a subset of these agents. The adversary, using the incoming communication signals can directly modify the received data associated with the compromised agents.

The problem of data privacy protection calls for developing privacy preserving control approaches for Internet of Things (IoT) devices over the edges of the network which leverage the distributed edge computing capabilities of local servers as opposed to centralized cloud computing platforms. The practical limitation of using cloud computing services can be pointed out as transmitting all the data to a central server which could cause data latency, putting computational stress on a central cloud server, and requiring a high bandwidth communication network [67]. Moreover, in the context of Internet of Battlefield Things (IoBT) one of the main requirements is to protect the privacy of data, which contains sensitive information [68, 69].

There may exist honest-but-curious agents in the network that attempt to violate the privacy of other agents by using the received information and learning about them (e.g. discovering sensitive information such as location of an agent) [70]. Moreover, the transmitted information can be intercepted by eavesdropper adversaries to perform confidentiality cyber-attacks [71]. Hence, protecting the data privacy of agents

against adversaries while achieving their control objectives is of paramount importance.

Several privacy preserving control methodologies have been proposed in the literature [72–78]. In [75] and [76], implementation of linear time-invariant (LTI) dynamic controllers for IoT networks by utilizing Paillier semihomomorphic encryption have been studied. By using the proposed semihomomorphic encryption, one can outsource the processing of encrypted plant data to a cloud platform.

Encryption-based methodologies have also been proposed to address the problem of privacy preserving control in MAS [73, 79–82]. In [73], homomorphic encryption was utilized to address the average consensus problem with finite-time convergence over cloud-based platforms. A decentralized multi-party computational method for MAS was suggested in [80] which is developed based on a combination of private sum aggregation and homomorphic encryption. In [81], the problem of average consensus for distributed systems under undirected communication graphs was studied and a decentralized architecture by utilizing homomorphic encryption was proposed.

Although efficiency of encryption-based methodologies in preserving the data privacy has been proven, these methods have a number of disadvantages. To name a few, the process of encrypting data is computationally excessive for agents and IoT devices with limited computational resources, and compared to the raw plant data, the encrypted data require higher bandwidth to be transmitted.

Differential privacy-based methodologies have been developed to address the problem of average consensus in MAS by deliberately adding noise to each agent [74,83,84]. In [74], a privacy preserving algorithm was proposed which guarantees that the initial values of agents will not be discovered by adversaries while they communicate information reach an average consensus. A noise with Laplace distribution characteristics was added to agents equipped with event-triggered controllers proposed in [84] to ensure that the privacy of agents is preserved and they reach a consensus. However, in above works, the considered agents are governed by single integrator dynamics.

In addition to encryption-based and differential privacy-based methodologies, isomorphisms and transformation based methodologies have recently been used to address privacy issues in controlling systems over cloud-based platforms [68,78,85]. In [68], isomorphisms and a communication protocol for control over the cloud were utilized and proposed to transform dynamics of agents to a new basis and make the transformed dynamics indistinguishable in the cloud. However, in this work a centralized cloud architecture is used to

control the agents, which is not desirable due to bandwidth requirements and data latency problems that can occur.

The work in [78], suggests transformation-based schemes to protect the privacy of an LTI system against various types of adversaries, while being controlled over the cloud. Moreover, a method to measure the privacy of the system is proposed in [78]. As an advantage, the computational overhead added due to using transformation-based methodologies is light since the size of transmitted data remains the same and one only requires matrix multiplication operations [78].

In addition to confidentially cyber-attacks, MAS are prone to integrity cyber-attacks. Secure consensus tracking control strategies considering two types of attacks were proposed for MAS in [86]. A distributed impulsive control for achieving synchronization in MAS subject to false data injection attacks has also been proposed in [87]. The work in [88] has suggested a control scheme for multi-agent systems with nonlinearities to reach a consensus while the agents are under deception attacks. In [89], cyber-physical attacks on MAS using a system theoretic approach has been studied. It was shown that the attack on one agent can spread into other agents that are reachable from the attacked agent. However, there are limitations and shortcomings in the above work as all cyber-attacks on MAS are treated as similar to attacks on standard LTI systems. On the other hand, cyber-attacks on communication channels among the agents and their significance and impacts have not been addressed and studied in the literature.

The impact of a certain type of cyber-attacks on MAS in which the adversary uses the model of the system to generate its attack signals has been studied in [90]. It is shown if the root of the directed spanning tree contained in the network graph is under "cyber-physical" attacks, the entire MAS can become unstable. In [91], a distributed methodology to detect cyber-attacks on communication networks among interconnected systems and MAS that are equipped with consensus-based controllers has been proposed. However, conditions under which the adversary is capable of performing undetectable cyber-attacks in MAS has not been investigated in above references.

To increase security and reduce consumption of energy, one can employ an event-triggered protocol so agents in MAS can communicate with one another. In [92–94], various types of event-triggered observer-based methodologies have been proposed for linear MAS and LTI systems. The work in [95] has studied an event-triggered unit that can be used to simultaneously reach a consensus and detect faults in MAS.

1.3 General Problem Statement and Thesis Objectives

Given that CPS consist of both physical components and communication networks, they are prone to both machine induced faults and cyber-attacks. One of the main challenges in theses systems is to address the problem of cyber-attack and fault detection and isolation (CAFDI). In other words, to detect and isolate machine induced faults and malicious deception cyber-attacks, such as covert attacks, zero dynamics attacks, replay attacks, and false data injection attacks. Furthermore, given that the CPS have applications in both centralized and geographically dispersed systems, both centralized and large-scale interconnected CPS should be considered to develop centralized and distributed CAFDI methodologies. In certain applications, such as a single unmanned aerial vehicle (UAV), due to the centralized architecture of the system, having a centralized CAFDI monitoring system is desirable, whereas in large-scale and geographically dispersed CPS, such as power networks and smart grids, using a distributed CAFDI methodology is more suitable and practical. Hence, our first objective in this thesis is to develop and investigate centralized and distributed CAFDI monitoring methodologies for CPS.

In addition to cyber-attack monitoring methodologies in CPS, there exists active cyber-attack detection approaches. Coding schemes can be utilized as active countermeasures against stealthy cyber-attacks in CPS. A coding scheme can be used to disrupt adversary's system knowledge and to target its required disruption resources for performing stealthy cyber-attacks. The latter can be achieved by increasing the required disruption resources for executing stealthy cyber-attacks, such as zero dynamics attacks, controllable attacks, and covert attacks. In particular, such a coding scheme is designed and developed such that it increases the CPS security index. The security index is a measure that indicates the minimum number of compromised sensors and actuators necessary for performing certain stealthy cyber-attacks. Hence, in presence of the above mentioned coding schemes, having only a certain number of secured actuators and sensors will prevent the adversaries from executing undetectable cyber-attacks. Our second objective in this thesis is to develop and study dynamic coding schemes that increase the security index of CPS.

Given that one of the main objectives of CPS operators is to secure their systems against undetectable cyber-attacks, it is imperative that they are made aware of the baseline security requirements in terms of disruption resources to accomplish this goal. As it was mentioned, adversaries require access to certain disruption resources to perform stealthy cyber-attacks. Hence, one is interested in developing a security

measure that denotes the minimum number of actuators and sensors that should be secured by CPS operators to prevent adversaries from executing stealthy cyber-attacks. Moreover, in the case of nonlinear CPS, the execution of zero dynamics and covert cyber-attacks and finding sensor measurements that should be secured to prevent adversaries from performing the mentioned cyber-attacks are challenging problems that have not been addressed in the literature. As for our third objective in this thesis, we aim to study and investigate actuators and sensors that should be secured to prevent stealthy cyber-attacks in both linear and nonlinear CPS.

Multi-agent systems (MAS) can be considered as distributed CPS. In MAS, agents share their information with their neighboring agents to achieve goals such as consensus and formation control. Given the existence of communication networks in MAS, they are susceptible to confidentially and integrity cyber-attacks. Hence, as one of our objectives in this thesis, we aim to develop a control protocol for MAS which ensures reaching a consensus in a distributed manner while agents' data privacy is protected. Furthermore, given that adversaries can inject their attack signals in the communication channels of MAS, under certain conditions they may take control over the entire MAS. Also, it would be possible for adversaries to attack certain nodes of the communication graph in a manner that all the agents follow that attack signal and reach a new consensus point. This implies that one is injecting and introducing an undetectable cyber-attack in the sense that residual signals, which become unbounded if the consensus is not achieved, in presence of cyber-attacks approach to zero as time approaches to infinity. Consequently, one of the main challenges that is addressed in this thesis is to develop a detector and generate residuals that are sensitive to undetectable cyber-attacks.

1.4 Contributions of Thesis

The main contributions of this thesis are as follows.

- **Cyber-attack and machine induced fault detection and isolation methodologies for cyber-physical systems**
    - Based on both the plant side and the C&C side centralized estimation and observation methodology, design conditions are developed and provided that can be used to detect and isolate actuator

cyber-attacks, sensor cyber-attacks, actuator faults, and sensor faults in a centralized architecture.

- A distributed filter design methodology based on observing the system from both the plant side and the C&C side is introduced and developed that can be utilized to detect and isolate both cyber-attacks and machine induced faults in large-scale interconnected systems.
- By utilizing our proposed methodologies, cyber-attacks such as covert attacks, zero dynamics attacks, replay attacks, and false data injection attacks can be detected and isolated.

- **Dynamic coding schemes as active countermeasures for cyber-attacks in cyber-physical systems**
  - Under certain assumptions, conditions under which one can carry out the zero dynamics and controllable attacks are obtained. These conditions are derived in terms of the Markov parameters of the CPS, elements of the observability matrix, and characteristic matrices of the system. Therefore, these conditions outline both the required disruption resources, i.e., the required actuators to be attacked, and the level of system knowledge that adversaries need to execute the zero dynamics and controllable cyber-attacks.
  - By utilizing the proposed conditions for existence of zero dynamics and controllable attacks, their implementation methodologies are then provided. As for the case of zero dynamics attacks, the implementation solely relies on the Markov parameters of the CPS and elements of the observability matrix.
  - A dynamic coding scheme is then developed and proposed that under certain conditions can increase the number of actuators that are needed to execute the zero dynamics and controllable cyber-attacks to its maximum possible value. Therefore, the proposed dynamic coding scheme can increase the actuators security index for the CPS.
  - Necessary and sufficient conditions under which covert cyber-attacks can be performed in the CPS are derived. The developed conditions can be used to determine which disruption resources in terms of input and output communication channels of the CPS should be compromised to carry out covert attacks.

- An upper bound on the SI for covert attacks is defined which relies on the developed necessary and sufficient conditions on the existence of covert attacks. Moreover, we provide and algorithm that can be used to compute the upper bound on SI for covert attacks.
- As an active countermeasure against covert attacks, we develop and propose a dynamic coding scheme. The proposed coding scheme includes an encoder on the C&C side and a decoder on the plant side of the CPS. Under certain conditions, if there exists one secure input and two secure output communication channels, adversaries will not be capable of performing covert cyber-attacks in the CPS.

- **The security requirement to prevent zero dynamics attacks and perfectly undetectable cyber-attacks in linear and nonlinear cyber-physical systems**
  - The notion of SE is formally defined as a measure that denotes the minimum number of actuators and sensors that should be secured to prevent adversaries from executing zero dynamics attacks, covert attacks, and controllable attacks.
  - Conditions under which the weakly unobservable subspace of CPS becomes zero are developed and investigated. If these conditions are satisfied, no zero dynamics attacks, covert attacks, and controllable attacks can be performed by the adversaries on the CPS.
  - In order to study perfectly undetectable cyber-attacks, conditions under which the controllable weakly unobservable subspace of CPS becomes zero are investigated. Therefore, under these conditions, adversaries cannot execute perfectly undetectable cyber-attacks, i.e., covert attacks and controllable attacks.
  - The ϵ-stealthy cyber-attacks in terms of Koopman operator are defined which can be used to categorize various types of cyber-attacks.
  - A relative degree of the CPS by means of Koopman eigenfunction, Koopman eigenvalue, and Koopman modes is defined. The proposed definition of the relative degree only requires matrix multiplications and is easy to check and verify. Moreover, we use the relative degree to discover internal dynamics of the CPS.

- A method to identify sensor measurements that are needed by adversaries to execute zero dynamics and covert cyber-attacks in nonlinear CPS is developed. Hence, by securing certain sensor measurements, one can prevent the execution of zero dynamics and covert cyber-attacks. Moreover, data-driven strategies for executing and implementing the zero dynamics and covert cyber-attacks by using the KCF of the CPS and the EDMD algorithm are proposed.

- **Addressing confidentially and integrity cyber-attacks in multi-agent systems by utilizing privacy preserving consensus control and event-triggered cyber-attack detection methodologies**

  - A unique isometric isomorphisms is developed and designed and used for each agent so that adversaries require discovering all the utilized isometric isomorphisms to disclose information of the entire network.

  - To preserve the privacy of agents when they are communicating with agents in their nearest neighborhood, a distributed consensus control is proposed that requires the transformed output measurements and dynamic controller states of the nearest neighboring agents to ensure reaching consensus.

  - We introduce the notion of controllability attacks on communication channels of the MAS systems. The importance of these attacks by studying and developing conditions that would provide the adversary full control over the entire MAS system is developed and formalized.

  - It is shown that the adversary is not capable of exciting zero dynamics of the directly attacked and healthy agents simultaneously.

  - A definition is introduced and proposed that specifies characteristics of undetectable cyber-attacks on MAS. Then conditions on the graph topology and its Laplacian matrix along with detectors of MAS are developed so that an adversary is capable of performing undetectable cyber-attacks. Moreover, if the above does not hold, we investigate under what conditions cyber-attacks are detectable on a certain team of agents.

  - Quasi-covert cyber-attacks are introduced where malicious hackers can inject in order to maintain their attacks undetected provided only non-root agents are compromised.

- An event-triggered detector is proposed for quasi-covert cyber-attacks that given its event-based communication strategy is more secure in comparison with conventional communication protocols.

1.5 Thesis Outline

In Chapter 2, cyber-attacks in CPS are briefly reviewed and the implementation of stealthy cyber-attacks such as zero dynamics attacks, covert attacks, and replay attacks are discussed.

In Chapter3, the problem of CAFDI is studied. Section 3.1 presents mathematical models considering faults and cyber-attacks. The proposed centralized CAFDI methodology with UIO-based detector and residual signals is explored in Section 3.2. Design conditions for the distributed CAFDI methodology are outlined in Section 3.3. In Section 3.4, a hardware-in-the-loop simulation and case studies demonstrate the effectiveness of the analytical results. Concluding remarks are provided in Section 3.5.

Chapter 4 is devoted to the development of active cyber-attack detection methodologies, namely dynamic coding schemes. Section 4.1 provides the state-space representation of the CPS system and defines specific cyber-attacks. In Section 4.2, the input/output (I/O) representation of CPS and ϵ-stealthy cyber-attacks are examined. Section 4.2 also explores zero dynamics, controllable cyber-attacks, and their existence conditions. Conditions for covert attacks and the investigation of Security Index (SI) for covert attacks are discussed in Sections 4.2.3 and 4.3. Dynamic coding schemes against zero dynamics attacks, controllable attacks, and covert cyber-attacks are proposed in Sections 4.4 and 4.5. Finally, Section 4.6 presents three numerical case studies to showcase the effectiveness of the proposed methodologies.

In Chapter 5, methods for finding the minimum number of actuators and sensors that should be secured to prevent certain stealthy cyber-attacks in linear and nonlinear CPS are investigated. Section 5.1 provides the state-space representation of both linear and nonlinear CPS systems, including objectives, definitions for specific cyber-attacks, and the Koopman operator theory. Investigation into conditions leading to zero weakly unobservable and controllable weakly unobservable subspaces of CPS is covered in Section 5.2. The formalization of security effort (SE) for linear CPS is discussed in Section 5.3. Section 5.4 introduces ϵ-stealthy cyber-attacks and methodologies for executing zero dynamics and covert attacks in nonlinear CPS using the Koopman operator theory. Data-driven implementation of zero dynamics and covert cyber-attacks

in nonlinear CPS is explored in Section 5.5. To showcase the effectiveness of the proposed methodologies, Section 5.6 presents numerical case studies.

Confidentially and integrity cyber-attacks are explored in Chapter 6. Section 6.1 presents essential graph theory concepts and establishes a model for MAS systems, including assumptions and lemmas. In Section 6.2, we introduce a model for MAS systems with attacked communication channels and outline the chapter objectives. The investigation of our privacy-preserving consensus control methodology for MAS is covered in Section 6.3. Section 6.4 formulates necessary and sufficient conditions for an adversary to gain full control over the MAS systems network. The limitations on zero dynamics attacks injected through compromised communication channels are explored in Section 6.4.3. Section 6.5 formally defines undetectable cyber-attacks in MAS, while Section 6.6 develops an event-triggered cyber-attack detection methodology. Finally, Section 6.7 provides illustrative numerical case studies demonstrating the capabilities of our proposed methodologies.

Concluding remarks and future research directions are presented in Chapter 7.

Chapter 2

Background: Cyber-Attacks in Cyber-Physical Systems (CPS)

In this chapter, the state-space representation of cyber-physical systems (CPS) under cyber-attacks and machine induced faults is provided. Moreover, the implementation of certain stealthy attacks such as covert attacks, replay attacks and zero dynamics attack are investigated and discussed. Results in this chapter are based on [1, 2, 18, 96]

2.1 Cyber-Physical Systems in Presence of Cyber-Attacks and Faults

Consider a strictly proper linear time-invariant (LTI) CPS represented by the following:

$$\begin{aligned} \dot{x}(t) =& Ax(t) + Bu^*(t) + L_1 f_1(t) + N\omega(t), \\ y_{\mathrm{p}}(t) =& Cx(t) + L_2 f_2(t) + \nu(t), \end{aligned} \tag{1}$$

where $x(t) \in \mathbb{R}^n$ denotes the state, $y_{\mathrm{p}}(t) \in \mathbb{R}^p$ is the measured output on the plant side, $u^*(t) \in \mathbb{R}^m$ is the control input, and $f_1(t) \in \mathbb{R}^{m_{\mathrm{f}}}$ and $f_2(t) \in \mathbb{R}^{p_{\mathrm{f}}}$ represent actuator and sensor faults, respectively. Additionally, $\omega(t) \in \mathbb{R}^m$ and $\nu(t) \in \mathbb{R}^p$ are zero mean wide-sense stationary (WSS) random Gaussian processes representing process and measurement noise with covariance matrices Q and R, respectively. The matrices $(A,\ C,\ B,\ N)$ have appropriate dimensions and describe the CPS characteristics, while the known

pair $(L_1,\ L_2)$ captures the fault signatures.

In the event of a cyber-attack on actuators, the control input is modified as follows:

$$u^*(t) = u(t) + S_{\mathrm{a}} a_{\mathrm{u}}(t), \tag{2}$$

where $u(t) \in \mathbb{R}^m$ is the control command from the Command and Control (C&C), $a_{\mathrm{u}}(t) \in \mathbb{R}^{m_{\mathrm{a}}}$ describes the effects of unknown cyber-attacks on actuators, and S_{a} is a matrix indicating the control input channels under attack.

The output of the CPS on the C&C side during sensor cyber-attacks is expressed as:

$$y^*(t) = Cx(t) + L_2 f_2(t) + D_{\mathrm{a}} a_{\mathrm{y}}(t) + \nu(t), \tag{3}$$

where $y^*(t) \in \mathbb{R}^p$ denotes the outputs, $a_{\mathrm{y}}(t) \in \mathbb{R}^{p_{\mathrm{a}}}$ denotes the attack signal, and the known matrix D_a describes the sensor attack signature. A CPS in the presence of both actuator and sensor cyber-attacks is illustrated in Figure 1.1.

Equations (1) and (2) provide a state-space realization of the CPS from the C&C side in the form:

$$\dot{x}(t) = Ax(t) + Bu(t) + B_{\mathrm{a}} a_{\mathrm{u}}(t) + L_1 f_1(t) + N\omega(t), \tag{4}$$

where $B_{\mathrm{a}} = BS_{\mathrm{a}}$ is interpreted as the actuator cyber-attack signature.

In (2) and (3), $a_{\mathrm{u}}(t)$ and $a_{\mathrm{y}}(t)$ represent the impacts of the adversary's attack on the control input and output of the CPS, respectively. These signals can be arbitrarily manipulated by the malicious adversary, intending to inflict maximum possible damage on the system components while remaining undetected.

2.1.1 Invariant zeros and output zeroing

Due to the linear representation of the system (4) and according to the superposition principle to study the invariant zeros of the system, one can consider $a_{\mathrm{u}}(t) = 0$, $a_{\mathrm{y}}(t) = 0$, $f_1(t) = 0$, $f_2(t) = 0$, $\omega(t) = 0$, and $\nu(t) = 0$. Given $s = z_0 \in \mathbb{R}$, invariant zeros of the system (4) are those z_0 in which the Rosenbrock system matrix

$$P_{\Sigma}(s) = \begin{bmatrix} sI - A & -B \\ C & 0 \end{bmatrix}$$

is rank deficient, i.e., the rank of $P_{\Sigma}(s)$ falls below its normal rank. Suppose $s = z_0$ is an invariant zero of system with associated zero state direction $x_0 \neq 0$ and zero input direction $u_0 \neq 0$ such that

$$\begin{bmatrix} z_0 I - A & -B \\ C & 0 \end{bmatrix} \begin{bmatrix} x_0 \\ u_0 \end{bmatrix} = 0. \tag{5}$$

Hence, given the nonzero initial condition $x(0) = x_0$ and control input $u(t) = u_0 e^{z_0 t}$, the state response of (4) is $X(t) = x_0 e^{z_0 t} \neq 0$, whereas the received output measurement on the C&C side is $y^*(t) = 0$. For a negative z_0 one has a minimum phase invariant zero, which can not exert a major damage to the system since $u(t)$ and $x(t)$ converge to zero as time approaches infinity. In contrary, excitation of a positive z_0, which is defined as a non-minimum phase invariant zero results in the increase of $u(t)$ and $X(t)$ as $t \to \infty$.

Below definitions of various types of invariant zeros for LTI systems are given [97]:

- Transmission zeros of the system (4) are defined as the invariant zeros of its observable and controllable (minimal) subsystem.

- Output decoupling zeros are the unobservable modes of the system (4). A given z_0 is an output decoupling zero of the presented LTI system if and only if the matrix pencil $P_{\Sigma}(z_0)$ loses its normal column rank. It follows that there exists a nonzero x_0 such that $(z_0 I - A)x_0 = 0$, and $Cx_0 = 0$.

- Input decoupling zeros are the unreachable modes of the system. A given z_0 is an input decoupling zero of the system (4) if and only if the matrix

$$\begin{bmatrix} z_0 I - A & -B \end{bmatrix}$$

  loses its normal row rank.

- The unreachable and unobservable modes of the LTI system are defined as its input-output zeros.

It is worth noting that all the output zeroing internal dynamics are called zero dynamics of the system.

2.2 Modeling Deception Attacks

In (2) and (3), $a_{\mathrm{u}}(t)$ and $a_{\mathrm{y}}(t)$ represent actuator and sensor cyber-attacks, respectively. Attack signals $a_{\mathrm{u}}(t)$ and $a_{\mathrm{y}}(t)$ can be arbitrarily designed by the malicious attacker. Consequently, the adversary's objective is to properly design $a_{\mathrm{u}}(t)$ and $a_{\mathrm{y}}(t)$ to exert the maximum possible damage to the components of the system, while reducing its detection risk.

Moreover, since the adversary compromises the system communication network to inject its malicious attack signals, $a_{\mathrm{u}}(t)$ and $a_{\mathrm{y}}(t)$ are limited by the network bandwidth, time delay in the network, and packet dropouts. In this thesis, it is assumed that the network is ideal and the mentioned limitations are not considered.

2.2.1 Replay Attack

In replay attacks, adversaries attempt to change the control input of the system, while they are manipulating the output in a manner that the received output on the C&C side (3) shows a nominal operational condition of the system [1]. Hence, adversaries record the system output $y_{\mathrm{p}}(t)$ from $t = t_1$ which is $y_{\mathrm{p}}(t_1)$ to $t = t_2$ that is $y_{\mathrm{p}}(t_2)$. In the next step, by compromising all the output communication channels, i.e., $D_{\mathrm{a}} = I_p$, and considering $a_{\mathrm{y}}(t) = -y_{\mathrm{p}}(t) + \tilde{y}(t)$, where $\tilde{y}(t)$ is the recorded nominal output of the system and belongs to $\{y_{\mathrm{p}}(t_1), \ldots, y_{\mathrm{p}}(t_2)\}$, adversaries manipulate $y^*(t)$ such that it shows the previously recorded output measurement $\tilde{y}(t)$. Consequently, adversaries inject their actuator cyber-attack signals through $a_{\mathrm{u}}(t)$ to either control the system or exert a damage to it. It is worth noting that to perform replay attack, adversaries do not need to know the parameters of the system.

2.2.2 Covert Attack

One of the main objectives of malicious attackers is to hack into a system without being detected. To this end, adversaries need to perform undetectable cyber-attacks. A more sophisticated version of replay attack is covert attack. In covert attacks, the sensor cyber-attack signal $a_{\mathrm{y}}(t)$ is intelligently designed to eliminate the impact of actuator cyber-attack on $y^*(t)$ [2, 96], which makes this type of cyber-attacks undetectable.

Assume that the adversary has access to all the communication channels of the system and knows all

system parameters. Consider the LTI system (4) under covert attacks such that for $a_{\mathrm{u}}(t) \neq 0$ one has

$$\dot{a}_{\mathrm{y}}(t) = A a_{\mathrm{y}}(t) + B_{\mathrm{a}} a_u(t),$$

where $a_{\mathrm{y}}(0) = x(0)$. Due to the adversary's full access to the communication channels, they can set $D_{\mathrm{a}} = -C$ and one can conclude that the impact of actuator cyber-attack $a_{\mathrm{u}}(t)$ is removed from the output measurements on the C&C side that is given in (3).

2.2.3 Zero Dynamics Attack

Given the initial condition $x(0) = x_0$, to excite the zero dynamics of the CPS (4), the adversary needs to generate the actuator attack signal $a_u(t) = a_{\mathrm{u}0} e^{z_0 t} \neq 0$ that satisfies the following [1]:

$$\begin{bmatrix} z_0 I - A & -B_{\mathrm{a}} \\ C & 0 \end{bmatrix} \begin{bmatrix} x_0 \\ a_{\mathrm{u}0} \end{bmatrix} = 0. \tag{6}$$

In zero dynamics attacks, the adversary is capable of exciting non-minimum phase ($z_0 > 0$) and minimum phase ($z_0 < 0$) zeros of the system. Due to their significant and dangerous impact on the system, only non-minimum phase type of zero dynamics attacks are considered in this thesis. In theory, for a given $z_0 > 0$, actuator cyber-attack signal $a_{\mathrm{u}}(t)$ can be unbounded as $t \to \infty$ that results in an unbounded $x(t)$, while the resulting output is identically zero. This implies that the non-minimum phase zero dynamics attacks result in a zero output and can lead the system to dangerous trajectories. To perform a zero dynamics attack, the adversary needs to know all the system parameters and has access to the input communication channels.

Chapter 3

Cyber-Attack and Machine Induced Fault Detection and Isolation Methodologies for Cyber-Physical Systems

In this chapter, the problem of simultaneous cyber-attack and fault detection and isolation (CAFDI) for both centralized and large-scale interconnected cyber-physical systems (CPS) is studied. The proposed methodologies include centralized and distributed CAFDI approaches, which involve the use of two filters on the plant and command and control (C&C) sides of the CPS, as well as an unknown input observer (UIO)-based detector on the plant side. The chapter characterizes the conditions under which the proposed methodologies can detect various types of deception attacks, such as covert attacks, zero dynamics attacks, and replay attacks. In the proposed centralized CAFDI methodology, the transmission of estimates from the C&C side filter to the plant side is required, with the assumption that a certain number of communication channels are secured. Consequently, a bank UIO-based detectors are utilized on the plant side to detect and isolate anomalies. It is also assumed that adversaries have knowledge of system parameters, filters, and the UIO-based detector. To address the limitations of secure communication channels, modifications to the two side filters and the UIO-based detector have been developed and implemented that eliminates the need for any secured communication channel in the modified CAFDI module. However, information must now be sent to and received from the plant side filter. Consequently, we develop a distributed CAFDI methodology

for the interconnected large-scale CPS which consists of several subsystems. Finally, a hardware-in-the-loop (HIL) simulation of a four area power network system under presence of both cyber-attacks and faults by using an OPAL-RT real-time simulator and Raspberry Pi is provided to illustrate the effectiveness of our proposed distributed CAFDI methodology. The work presented in this chapter has appeared in [26].

To summarize, the main contributions of this chapter are stated as follows:

(1) Based on both the plant side and the C&C side centralized estimation and observation methodology, design conditions are developed and provided that can be used to detect and isolate actuator cyber-attacks, sensor cyber-attacks, actuator faults, and sensor faults in a centralized architecture.

(2) A distributed filter design methodology based on observing the system from both the plant side and the C&C side is introduced and developed that can be utilized to detect and isolate both cyber-attacks and machine induced faults in large-scale interconnected systems.

(3) By utilizing our proposed methodologies, cyber-attacks such as covert attacks, zero dynamics attacks, replay attacks, and false data injection attacks can be detected and isolated.

The remainder of the chapter is organized as follows. Mathematical models of the systems that take into account faults and cyber-attacks and the definition of undetectable attacks are provided in Section 3.1. In Section 3.2, our proposed centralized CAFDI methodology that consists of two side filters, the UIO-based detector and residual signals are developed and investigated. Design conditions for the distributed CAFDI methodology are proposed and developed in Section 3.3. To illustrate and demonstrate the effectiveness and capabilities of our analytical results, a hardware-in-the-loop (HIL) simulation environment and case studies are presented and extensively investigated in Section 3.4. Conclusions are provided in Section 3.5.

3.1 Problem Statement and Formulation

3.1.1 Cyber-Physical Systems (CPS) Model

In this chapter, a strictly proper linear time-invariant (LTI) CPS of the form given below is studied:

$$\begin{aligned} \dot{x}^{\mathrm{s}}(t) =& A^{\mathrm{s}}x^{\mathrm{s}}(t) + B^{\mathrm{s}}u^{*}(t) + L_1 f_1(t) + N^{\mathrm{s}}\omega^{\mathrm{s}}(t), \\ y_{\mathrm{p}}(t) =& C^{\mathrm{s}}x^{\mathrm{s}}(t) + L_2 f_2^{\mathrm{s}}(t) + \nu^{\mathrm{s}}(t), \end{aligned} \tag{7}$$

where $x^{\mathrm{s}}(t) \in \mathbb{R}^n$ represents the state, $y_{\mathrm{p}}(t) \in \mathbb{R}^p$ denotes the measured output on the plant side, $u^*(t) \in \mathbb{R}^m$ denotes the control input, $f_1(t) \in \mathbb{R}^{m_{\mathrm{f}}}$ and $f_2^{\mathrm{s}}(t) \in \mathbb{R}^{p_{\mathrm{f}}}$ correspond to actuator and sensor faults, respectively. Moreover, $\omega^{\mathrm{s}}(t) \in \mathbb{R}^n$ and $\nu^{\mathrm{s}}(t) \in \mathbb{R}^p$ denote zero mean wide-sense stationary (WSS) random Gaussian processes that represent process and measurement noise with the covariance matrices Q and R, respectively. The quadruple $(A^{\mathrm{s}},\, C^{\mathrm{s}},\, B^{\mathrm{s}},\, N^{\mathrm{s}})$ has appropriate dimensions and describe the CPS characteristics, and the known pair $(L_1,\, L_2)$ capture the fault signatures.

In case of injection of a cyber-attack on actuators, the control input is expressed and changed to

$$u^*(t) = u(t) + S_{\mathrm{a}}a_{\mathrm{u}}(t), \tag{8}$$

where $u(t) \in \mathbb{R}^m$ represents the control command which is the output of the C&C, $a_{\mathrm{u}}(t) \in \mathbb{R}^{m_{\mathrm{a}}}$ denotes a vector describing the effects of unknown cyber-attacks on actuators, and S_{a} is a matrix of appropriate dimension that indicates the control input channels that are under attack.

The output of the CPS on the C&C side when sensors are under cyber-attack can be expressed as

$$y^*(t) = y_{\mathrm{p}}(t) + D_{\mathrm{a}}a_{\mathrm{y}}(t), \tag{9}$$

where $y^*(t) \in \mathbb{R}^p$ denotes the output, $a_y(t) \in \mathbb{R}^{p_{\mathrm{a}}}$ denotes the attack signal, and the matrix D_a describes the sensor attack signature.

Equations (7) and (8) provide a state space realization of the CPS from the C&C side in the following

form:

$$\dot{x}^{\mathrm{s}}(t) = A^{\mathrm{s}}x^{\mathrm{s}}(t) + B^{\mathrm{s}}u(t) + B^{\mathrm{s}}_{\mathrm{a}}a_{\mathrm{u}}(t) + L_1 f_1(t) + N^{\mathrm{s}}\omega^{\mathrm{s}}(t), \tag{10}$$

where $B^{\mathrm{s}}_{\mathrm{a}} = B^{\mathrm{s}}S_{\mathrm{a}}$ is to be interpreted as the actuator cyber-attack signature.

Definition 3.1 (Weakly Unobservable Subspace [98]). *Let us denote the CPS by* $\Sigma = (A^s, B^s, B^s_a, L_1, N^s, C^s, L_2, D_a)$. *Under the fault free scenario* $f_1(t) = 0$ *and* $f^s_2(t) = 0$, *the noise free scenario* $\omega^s(t) = 0$ *and* $\nu^s(t) = 0$, *and the cyber-attack free scenario* $a_u(t) = 0$ *and* $a_y(t) = 0$, *a point* $x^s(0) = x^s_0 \in \mathbb{R}^n$ *is called weakly unobservable if there exists an input function* $u(t)$ *such that the output satisfies* $y^*(t) = 0$, $\forall\, t \geq 0$. *The set of all weakly unobservable points is called weakly unobservable subspace and is denoted by* $\mathscr{V}(\Sigma)$. *Moreover, the largest weakly unobservable subspace is denoted by* $\mathscr{V}^*(\Sigma)$.

Let us denote $X^{\mathrm{s}}(x^{\mathrm{s}}(0), u(t), a_{\mathrm{u}}(t), a_{\mathrm{y}}(t))$ as the solution to (10) under the fault free condition, and $Y(x^{\mathrm{s}}(0), u(t), a_{\mathrm{u}}(t), a_{\mathrm{y}}(t)) = C^{\mathrm{s}}X^{\mathrm{s}}(x^{\mathrm{s}}(0), u(t), a_{\mathrm{u}}(t), a_{\mathrm{y}}(t))$ as the corresponding output of the CPS, $\forall\, t \geq 0$.

Definition 3.2 (Undetectable Cyber-Attacks [15]). *Given initial conditions* $x^s_0 \in \mathbb{R}^n$ *and* $\bar{x}^s_0 \in \mathbb{R}^n$, *in the CPS* (10) *under the fault free scenario, the cyber-attack on actuators and sensors using* $a(t) = [a_u(t)^\top\ a_y(t)^\top]^\top \neq 0$, *is designated as undetectable if* $Y(x^s_0, u(t), a_u(t), a_y(t)) = Y(\bar{x}^s_0, u(t), 0, 0)$, $\forall t \geq 0$, *otherwise, the cyber-attack is defined as detectable.*

Definition 3.3 (Input Observable Systems [99]). *The system* $(C,\ A,\ B)$ *is input observable if* B *is monic and* $Im(B)$ *does not intersect with the unobservable subspace of* $(C,\ A)$, *where* $Im(B)$ *denotes the image of* B.

In the same manner as described in [35] and [99], the sensor fault and sensor noise can be represented by pseudo actuator fault and pseudo process noise, respectively. It is worth noting that in this representation, as described below, sensor faults are mapped into and represented by pseudo actuator faults.

Towards the above end, the following auxiliary invertible LTI system that is driven by the appropriate $f_2(t)$, which represents the pseudo actuator fault, and $\omega^{\mathrm{a}}(t)$, which captures the pseudo process noise, is

expressed as:

$$\begin{aligned} \dot{x}^{\mathrm{a}}(t) &= A^{\mathrm{a}}x^{\mathrm{a}}(t) + L_2^{\mathrm{a}}f_2(t) + N^{\mathrm{a}}\omega^{\mathrm{a}}(t), \\ C^{\mathrm{a}}x^{\mathrm{a}}(t) &= L_2 f_2^{\mathrm{s}}(t) + \nu^{\mathrm{s}}(t), \end{aligned} \tag{11}$$

where $x^{\mathrm{a}}(t) \in \mathbb{R}^{p_{\mathrm{f}}+p}$, $f_2(t) \in \mathbb{R}^{p_{\mathrm{f}}}$, and $\omega^{\mathrm{a}}(t) \in \mathbb{R}^{p}$. By incorporating the dynamics of (10) and (11), one can obtain the augmented and extended CPS in the following form:

$$\begin{aligned} \dot{x}(t) =& Ax(t) + Bu(t) + B_{\mathrm{a}}a_{\mathrm{u}}(t) + F_1 f_1(t) + F_2 f_2(t) \\ &+ N\omega(t), \\ y^*(t) =& Cx(t) + D_{\mathrm{a}}a_{\mathrm{y}}(t), \end{aligned} \tag{12}$$

where $x(t) = [x^{\mathrm{s}}(t)^\top,\, x^{\mathrm{a}}(t)^\top]^\top$, $A = \mathrm{diag}(A^{\mathrm{s}}, A^{\mathrm{a}})$, $B = [B^{\mathrm{s}\top},\, 0_{m\times(p_{\mathrm{f}}+p)}]^\top$, $B_{\mathrm{a}} = [B_{\mathrm{a}}^{\mathrm{s}\top},\, 0_{m_{\mathrm{a}}\times(p_{\mathrm{f}}+p)}]^\top$, $F_1 = [{L_1}^\top,\, 0_{m_{\mathrm{f}}\times(p_{\mathrm{f}}+p)}]^\top$, $F_2 = [0_{p_{\mathrm{f}}\times n},\, {L_2^{\mathrm{a}}}^\top]^\top$, $N = \mathrm{diag}(N^{\mathrm{s}}, N^{\mathrm{a}})$, $\omega(t) = [\omega^{\mathrm{s}}(t)^\top,\, \omega^{\mathrm{a}}(t)^\top]^\top$, and $C = [C^{\mathrm{s}},\, C^{\mathrm{a}}]$. It should be noted that the defined output $y^*(t)$ in (9) is equal to the one that is given by (12), however, the representations are different.

3.1.2 Model of the Interconnected CPS

In this section, our objective is to provide a representation of a class of interconnected CPS that are distributed in nature and consist of several subsystems as depicted in Figure 3.1. Consequently, considering the large-scale of interconnected CPS, one needs to develop a scalable distributed CAFDI methodology that is more desirable for this type of CPS.

We consider the interconnected CPS as consisting of N subsystems. The dynamics of the i-th subsystem is expressed as:

$$\mathcal{S}_i : \begin{cases} \dot{x}_i(t) = & A_i x_i(t) + \sum\limits_{j\in\mathcal{N}_i} A_{ij}x_j(t) + B_i u_i^*(t) \\ & +F_1^i f_1^i(t) + F_2^i f_2^i(t) + N_i\omega_i(t), \\ y_i^*(t) = & C_i x_i(t) + D_{\mathrm{a}}^i a_{\mathrm{y}}^i(t), \quad i = 1, \ldots, N, \end{cases} \tag{13}$$

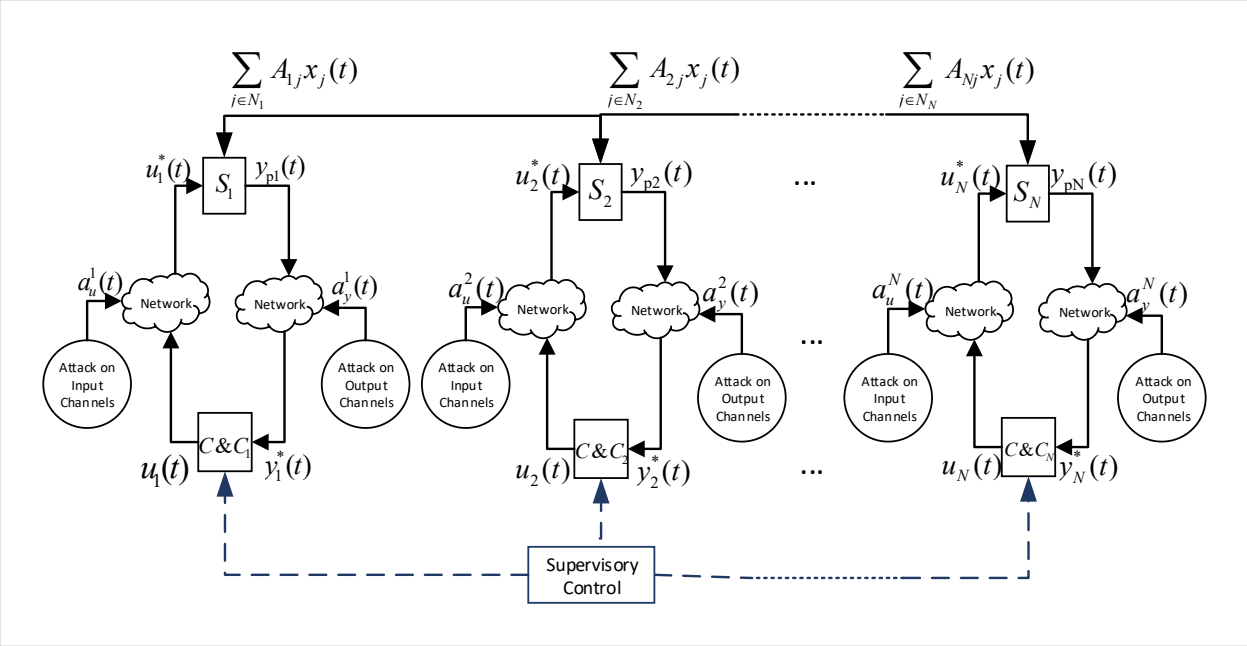

Figure 3.1: Distributed interconnected cyber-physical system consisting of N subsystems under actuator and sensor attacks. Dashed lines indicate the possible interconnections among C&C centers of different subsystems.

where $x_i(t) \in \mathbb{R}^{n_i+p_{\mathrm{f}i}+p_i}$ denotes the state of the i-th subsystem, $u_i^*(t) \in \mathbb{R}^{m_i}$ denotes the control input of the subsystem i, $y_i^*(t) \in \mathbb{R}^{p_i}$ denotes the measured output on the C&C side of $\mathcal{S}_i$, $a_{\mathrm{y}}^i(t) \in \mathbb{R}^{p_{\mathrm{a}i}}$ denotes the sensor attack signal in the i-th subsystem, $\omega_i(t) \in \mathbb{R}^{n_i+p_i}$ denotes the zero mean WSS random Gaussian noise of $\mathcal{S}_i$ with the covariance matrix Q_i, $f_1^i(t) \in \mathbb{R}^{m_{\mathrm{f}i}}$ and $f_2^i(t) \in \mathbb{R}^{p_{\mathrm{f}i}}$ correspond to actuator and pseudo actuator faults in the subsystem i, respectively, that are derived using a similar method as shown in Section 3.1.1.

Moreover, the matrix A_{ij} represents the physical coupling between subsystems i and $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ is the set of neighboring subsystems that are coupled with the i-th subsystem. Furthermore, one has

$$B_i u_i^*(t) = B_i u_i(t) + B_{\mathrm{a}}^i a_{\mathrm{u}}^i(t),$$

where $u_i(t) \in \mathbb{R}^{m_i}$ denotes the control input generated on the C&C side of $\mathcal{S}_i$, $a_{\mathrm{u}}^i(t) \in \mathbb{R}^{m_{\mathrm{a}i}}$ denotes the actuator attack signal in the subsystem i, $B_{\mathrm{a}}^i = B_i S_{\mathrm{a}}^i$ describes the actuator attack signature of the i-th subsystem, and the matrix S_{a}^i of appropriate dimension indicates the control input channels of $\mathcal{S}_i$ that are

compromised by adversaries. The quadruple (A_i, A_{ij}, B_i, C_i) has appropriate dimensions and describes the characteristics of the i-th subsystem, N_i is the noise signature of $\mathcal{S}_i$, and the known pairs (F_1^i, F_2^i) and $(B_{\mathrm{a}}^i, D_{\mathrm{a}}^i)$ capture fault and attack signatures of $\mathcal{S}_i$, respectively.

We consider the following assumptions throughout this chapter.

Assumption 3.1. *In the CPS* (7)*, the number of states is greater than the number of actuators and sensors, i.e.,* $n > m$ *and* $n > p$*.*

Assumption 3.2. *The adversary has full knowledge on the parameters of* (7) *and* (13)*, and has access to all the input and output communication channels, i.e.,* $m_a = m$*,* $p_a = p$*,* $m_{ai} = m_i$*, and* $p_{ai} = p_i$*.*

Remark 3.1. *It should be emphasized that there may exist local controllers on the plant side of the CPS* (12) *and the interconnected CPS* (13)*. Hence, the C&C as shown in Figures 1.1 and 3.1 act as outer control loops of the CPS.*

Remark 3.2. *It should be noted that in the interconnected CPS* (13)*, communication channels among the C&C centers of subsystems can be compromised by adversaries. However, detection of this type of cyber-attack is not within the scope of this chapter and is not addressed here. Methodologies for detecting cyber-attacks on the communication channels among the C&C centers of subsystems can be found in [100] and [101].*

3.1.3 Objectives

Our main objective in this chapter is to address the simultaneous cyber-attack and fault detection and isolation (CAFDI) problem for the CPS both corresponding to centralized and distributed architectures. Towards this end, we design a bank of observers such that each set of residual signals corresponding to observers is sensitive and specified to detect one specific type of anomaly, namely either an actuator cyber-attack $a_{\mathrm{u}}(t)$, a sensor cyber-attack $a_{\mathrm{y}}(t)$, an actuator fault $f_1(t)$, and/or a pseudo actuator fault $f_2(t)$, while each residual is decoupled from all the other anomalies.

Decoupling the residuals from one another implies that the occurrence of anomalies only affects those residual signals that are designated to them. We also do not limit our focus on detecting only detectable

cyber-attacks (see Definition 3.2). Our goal and objective is to further detect the so-called undetectable cyber-attacks in the sense of Definition 3.2, e.g., covert and zero dynamics attacks.

First, we assume that the adversary cannot compromise all the communication channels among the proposed C&C side filter and the UIO-based detector, although they have a complete knowledge of parameters of the filters and detectors. Next, we investigate conditions under which adversaries have access to all the communication channels among the C&C side filter and UIO-based detector. We modify our proposed centralized CAFDI module to address the latter problem. Moreover, we extend our results to develop and propose a distributed CAFDI methodology for the interconnected large-scale CPS.

3.2 Centralized Cyber-Attack and Fault Detection and Isolation Methodology

The presence of network layer in the CPS has enabled malicious adversaries to perform cyber-attacks on the entire system. On the other hand, due to the existence of this network layer, it is possible to observe the CPS from both the plant side and its C&C side. The idea of observing the CPS from both the plant side and the C&C side is illustrated in Figure 3.2. Our goal in this framework is to utilize information from the designed filters on both sides via a communication channel and generate residuals that are specifically sensitive to faults and cyber-attacks. Using these residuals, the isolation between faults or cyber-attacks can also be achieved.

Two filters having the same characteristics on both sides are designed in Subsections 3.2.1 and 3.2.2. By using communication channels, states of the C&C side filter are transmitted to the plant side to generate a residual signal that is sensitive to only cyber-attacks while this communication channel may still be compromised by an adversary. However, we assume that there exists a certain number of secure communication channels to transmit states of the C&C side filter to the plant side.

A detector on the plant side that utilizes an unknown input observer (UIO) is designed in Subsection 3.2.3. The detector utilizes the previously generated residuals as additional input so that the UIO-based detector is sensitive to both cyber-attacks and faults. The reason for selecting UIO as the main detector is that it enables one to utilize a general design structure to simultaneously address the considered CAFDI

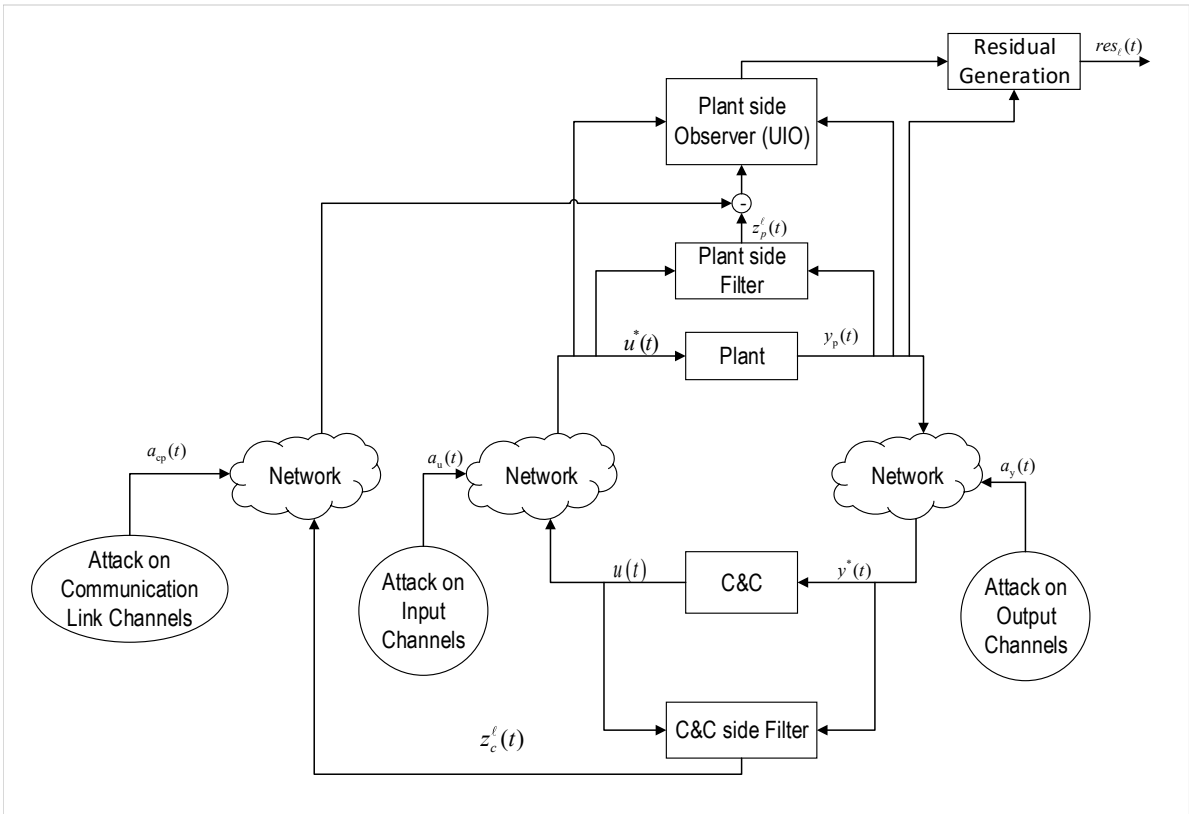

Figure 3.2: Observers/filters on both the plant side and the C&C side of the CPS, where $z_{\rm c}(t)$ represents the states of the C&C side filter, $z_{\rm p}(t)$ denotes the states of the plant side filter, $a_{\rm c}(t)$ denotes the cyber-attack on the communication channels, and $res(t)$ denotes the residual signals that are generated on the plant side.

problems.

Our proposed centralized CAFDI methodology is presented in Subsection 3.2.4. It is worth noting that by utilizing the proposed methodology, one is still capable of detecting several types of stealthy cyber-attacks on the system, such as covert attacks and zero dynamics attacks. Moreover, in Subsection 3.2.5, the case where all the communication channels between the C&C side filter and the plant side module are non-secure is studied. Consequently, the C&C side and the plant filters as well as the UIO-based detector are modified to address the CAFDI problem for the CPS in which all the communication channels among the two side filters and detectors are compromised by adversaries.

3.2.1 Command & Control (C&C) Side Filter

From the C&C side and according to (12), the output of the CPS is governed by

$$y^*(t) = Cx(t) + D_{\mathrm{a}} a_{\mathrm{y}}(t). \tag{14}$$

We have the following standing assumption to be considered throughout this chapter.

Assumption 3.3. *Only the communication channels can be compromised and attacked. Consequently, on the C&C side one has access to the control signal, $u(t)$, before its manipulation by the adversary. Moreover, on the plant side one has access to $y_p(t)$ before its manipulation by the malicious attacker.*

The proposed filter on the C&C side can be expressed as

$$\dot{z}_{\mathrm{c}}^{\ell}(t) = F_{\mathrm{p}}^{\ell} z_{\mathrm{c}}^{\ell}(t) + T_{\mathrm{p}}^{\ell} B u(t) + K_{\mathrm{p}}^{\ell} y^*(t), \tag{15}$$

where $z_{\mathrm{c}}^{\ell}(t) \in \mathbb{R}^n$ represents the filter state that estimates $x^{\mathrm{s}}(t)$ from the C&C side, and the matrices F_{p}^{ℓ}, T_{p}^{ℓ}, and K_{p}^{ℓ} are of appropriate dimensions that are designed and selected subsequently. The index $\ell \in \{\mathrm{SA}, \mathrm{AA}, \mathrm{SF}, \mathrm{AF}\}$ designates if the filter is designed for detecting sensor attacks, actuator attacks, sensor faults, and actuator faults, respectively.

3.2.2 Plant Side Filter

On the plant side, sensor measurements are carried out before sensor cyber-attacks, and the output of the CPS can be expressed as follows:

$$y_{\mathrm{p}}(t) = Cx(t).$$

Moreover, on this side one has access to the potentially manipulated control signal $u^*(t) = u(t) + S_{\mathrm{a}} a_{\mathrm{u}}(t)$.

The proposed filter on the plant side is expressed in the following form:

$$\dot{z}_{\mathrm{p}}^{\ell}(t) = F_{\mathrm{p}}^{\ell} z_{\mathrm{p}}^{\ell}(t) + T_{\mathrm{p}}^{\ell} B u^*(t) + K_{\mathrm{p}}^{\ell} y_{\mathrm{p}}(t), \tag{16}$$

where $z_{\mathrm{p}}^{\ell}(t) \in \mathbb{R}^n$ denotes the filter state estimating $x^{\mathrm{s}}(t)$ from the plant side. Similar to the C&C side

filters, the index $\ell \in \{\text{SA}, \text{AA}, \text{SF}, \text{AF}\}$, indicates if the filter is designed for detecting sensor attacks, actuator attacks, sensor faults, and actuator faults, respectively.

The error signals between estimated states for both sides can be defined as $e_{\text{p}}^{\ell}(t) = z_{\text{p}}^{\ell}(t) - z_{\text{c}}^{\ell}(t)$. The representation of the error dynamics between the two filter states can be derived as follows:

$$\dot{e}_{\text{p}}^{\ell}(t) = F_{\text{p}}^{\ell} e_{\text{p}}^{\ell}(t) + T_{\text{p}}^{\ell} B_{\text{a}} a_{\text{u}}(t) - K_{\text{p}}^{\ell} D_{\text{a}} a_{\text{y}}(t). \tag{17}$$

It follows from (17) that the error dynamics is only sensitive to cyber-attacks.

3.2.3 UIO-Based Detector and Residual Signal Generation

We have adopted the UIO design from [36] to develop a UIO-based detector on the plant side with the following representation:

$$\begin{aligned} \dot{z}^{\ell}(t) =& F^{\ell} z^{\ell}(t) + T^{\ell} B u^{*}(t) + K^{\ell} y_{\text{p}}(t) + L^{\ell}(z_{\text{p}}^{\ell}(t) - (z_{\text{c}}^{\ell}(t) + D_{\text{cp}} a_{\text{cp}}(t))), \\ \hat{x}^{\ell}(t) =& z(t)^{\ell} + H^{\ell} y_{\text{p}}(t), \end{aligned} \tag{18}$$

where $z^{\ell}(t) \in \mathbb{R}^{n+p_{\text{f}}+p}$, and $\hat{x}^{\ell}(t) \in \mathbb{R}^{n+p_{\text{f}}+p}$ denotes the estimated states by the detector, and $a_{\text{cp}}(t) \in \mathbb{R}^{n_c}$ denotes the cyber-attack on the communication channel between the two filters with the signature D_{cp}. The matrices F^{ℓ}, T^{ℓ}, K^{ℓ}, L^{ℓ}, and H^{ℓ} are of appropriate dimensions and will be specified subsequently, with $\ell \in \{\text{SA}, \text{AA}, \text{SF}, \text{AF}\}$, denoting the categories defined previously.

The error between the states of the detector and the CPS is defined as $e^{\ell}(t) = x(t) - \hat{x}^{\ell}(t)$. Let

$$res_{\ell}(t) = y_{\text{p}}(t) - C\hat{x}^{\ell}(t) = Ce^{\ell}(t), \tag{19}$$

denote a residual signal. By selecting $K^{\ell} = K_1^{\ell} + K_2^{\ell}$, $F^{\ell} = A - H^{\ell}CA - K_1^{\ell}C$, with K_1^{ℓ} of appropriate dimension, and $K_2^{\ell} = FH^{\ell}$, the dynamics associated with $e^{\ell}(t)$ can now be expressed in the following form:

$$\begin{aligned}\dot{e}^{\ell}(t) =&(A - H^{\ell}CA - K_1^{\ell}C)e^{\ell}(t) + (I - T^{\ell} - H^{\ell}C)(Bu(t) \\ &+ B_{\mathrm{a}}a_{\mathrm{u}}(t)) + (I - H^{\ell}C)F_1 f_1(t) + (I - H^{\ell}C)F_2 f_2(t) \\ &+ (I - H^{\ell}C)N\omega(t) - L^{\ell}e_{\mathrm{p}}^{\ell}(t) - L^{\ell}D_{\mathrm{cp}}a_{\mathrm{cp}}(t).\end{aligned} \tag{20}$$

Definition 3.4. *A cyber-attack/fault is detected if the residual signal $res_{\ell}(t)$ given by (19) exceeds a pre-specified threshold $\eta > 0$ as follows:*

$$\|res_{\ell}(t)\|_2 > \eta,$$

where $\|.\|_2$ indicates the Euclidean norm.

Remark 3.3. *To select the threshold η, one may need to perform Monte Carlo simulation runs for the healthy system, i.e., for the fault free and cyber-attack free system in presence of external disturbances and noise and choose the maximum value of $\|res(t)_{\ell}\|_2$ as η.*

Definition 3.5 (Decoupled Residual)**.** *The residual signal $res_{\ell}(t)$ given by (19) is decoupled from an anomalous signal in the set $\{a_u(t), a_y(t), a_{cp}(t), f_1(t), f_2(t)\}$ if the dynamics and trajectory of $res_{\ell}(t)$ are not affected by that anomalous signal.*

The following assumption stands throughout this section.

Assumption 3.4. *The malicious adversary knows the parameters of the C&C side filter in (15), the plant side filter in (16), and the UIO-based detector in (18).*

3.2.4 Filters and Detectors Design for Cyber-Attack and Fault Detection and Isolation Objectives

The error dynamics in (17) and (20) can now be augmented as follows:

$$\dot{\check{e}}^{\ell}(t) = \check{F}^{\ell}\check{e}^{\ell}(t) + \check{B}^{\ell}u(t) + \check{B}_{\mathrm{a}}^{\ell}a_{\mathrm{u}}(t) + \check{F}_1^{\ell}f_1(t) + \check{F}_2^{\ell}f_2(t) - \check{K}_{\mathrm{p}}^{\ell}a_{\mathrm{y}}(t) - \check{L}^{\ell}a_{\mathrm{cp}}(t) + \check{N}^{\ell}\omega(t), \tag{21}$$

where $\check{e}^{\ell}(t) = [e^{\ell}(t)^{\top}\ e_{\mathrm{p}}^{\ell}(t)^{\top}]^{\top}$, and

$$
\begin{aligned}
\check{F}^{\ell} &= \begin{bmatrix} F^{\ell} & -L^{\ell} \\ 0 & F_{\mathrm{p}}^{\ell} \end{bmatrix},\ \check{B} = \begin{bmatrix} (I - T^{\ell} - H^{\ell}C)B \\ 0 \end{bmatrix}, \\
\check{B}_{\mathrm{a}}^{\ell} &= \begin{bmatrix} (I - T^{\ell} - H^{\ell}C)B_{\mathrm{a}} \\ T_{\mathrm{p}}^{\ell}B_{\mathrm{a}} \end{bmatrix},\ \check{F}_{1}^{\ell} = \begin{bmatrix} (I - H^{\ell}C)F_1 \\ 0 \end{bmatrix}, \\
\check{F}_{2}^{\ell} &= \begin{bmatrix} (I - H^{\ell}C)F_2 \\ 0 \end{bmatrix},\ \check{K}_{\mathrm{p}}^{\ell} = \begin{bmatrix} 0 \\ K_{\mathrm{p}}^{\ell}D_{\mathrm{a}} \end{bmatrix},\ \check{L}^{\ell} = \begin{bmatrix} L^{\ell}D_{\mathrm{cp}} \\ 0 \end{bmatrix}, \\
\check{N}^{\ell} &= \begin{bmatrix} (I - H^{\ell}C)N \\ 0 \end{bmatrix},
\end{aligned}
\tag{22}
$$

where $\ell \in \{\mathrm{SA}, \mathrm{AA}, \mathrm{SF}, \mathrm{AF}\}$.

Assumption 3.5. *There exist $q = max\{m_a, p_a\}$ secure communication channels among the C&C side filter in (15) and the UIO-based detector in (18), i.e., rank$(D_{cp}) = n - q$. Moreover, $C_q = \{c_1, \ldots, c_q\}$ denotes the set of secured communication channels, where $c_\zeta \in \{1, \ldots, n\}$, for $\zeta \in \{1, \ldots, q\}$.*

In the following, it is shown how one can generate four residual signals $res_{\mathrm{AA}}(t)$, $res_{\mathrm{SA}}(t)$, $res_{\mathrm{AF}}(t)$, and $res_{\mathrm{SA}}(t)$ to detect the actuator cyber-attack, the sensor cyber-attack, the actuator fault, and the sensor fault, respectively, by using a bank of filters and four UIO-based detectors.

Proposition 3.1. *Under Assumption 3.5, the residual signal $res_{AA}(t) = y_p(t) - C\hat{x}^{AA}(t)$ is affected by the actuator cyber-attacks $a_u(t)$ and is decoupled from $a_y(t)$, $a_{cp}(t)$, $f_1(t)$, and $f_2(t)$ in the sense of Definition 3.5, if the following conditions for the augmented dynamics (21) hold for $\ell = AA$, namely:*

(1) $T^{\ell} = I - H^{\ell}C$;

(2) $(I - H^{\ell}C)F_1 = 0$;

(3) $(I - H^{\ell}C)F_2 = 0$;

(4) $L^{\ell}D_{cp} = 0$;

(5) $K_p^{AA} D_a = 0$*;*

(6) the triplet $(C,\ F^\ell,\ \bar{L}^\ell)$ *is left-invertible, where* $\bar{L}^\ell = [l_{c_1}^\ell \cdots l_{c_q}^\ell]$*, and* $l_{c_\zeta}^\ell$ *is the* c_ζ*-th column of* L^ℓ*, for* $\zeta = 1, \ldots, q$*;*

(7) the Rosenbrock system matrix

$$P_{\Sigma_u}(s) = \begin{bmatrix} sI - F_p^{AA} & -T_p^{AA} B_a \\ L^{AA} & 0_{(n+p_f+p)\times m_a} \end{bmatrix},$$

does not have any non-minimum phase zero dynamics;

(8) $rank\,(L^{AA} T_p^{AA} B_a) = rank\,(T_p^{AA} B_a)$*;*

(9) $\check{F}^\ell$ *is Hurwitz.*

Proof. The augmented error dynamics associated with $e^{\mathrm{AA}}(t)$ and $e_p^{\mathrm{AA}}(t)$ are governed by (21) where $\ell = \mathrm{AA}$. Under Conditions 1) to 5), the dynamics (21) become

$$\dot{\check{e}}^{\mathrm{AA}}(t) = \check{F}^{\mathrm{AA}} \check{e}^{\mathrm{AA}}(t) + \check{B}_{\mathrm{a}}^{\mathrm{AA}} a_{\mathrm{u}}(t) + \check{N}^{\mathrm{AA}} \omega(t). \tag{23}$$

Consequently, the error signal $\check{e}^{\mathrm{AA}}(t)$ is not affected by the control command $u(t)$, the actuator fault $f_1(t)$, the sensor fault $f_2(t)$, the sensor attack $a_{\mathrm{y}}(t)$, and the communication channel attack signal $a_{\mathrm{c}}(t)$. It should be noted that having $L^{\mathrm{AA}} D_{\mathrm{cp}} = 0$ implies that

$$\mathrm{Im}(D_{\mathrm{cp}}) \subseteq \mathrm{Ker}(L^{\mathrm{AA}}).$$

Hence, since as per Assumption 3.5 we consider $\mathrm{rank}(D_{\mathrm{ac}}) = n - q$, there exists a nonzero L^{AA} with $\mathrm{rank}(L^{\mathrm{AA}}) = q$ which satisfies $L^{\mathrm{AA}} D_{\mathrm{cp}} = 0$.

Furthermore, (23) can be partitioned into the following two subsystems:

$$\dot{e}_{\mathrm{p}}^{\mathrm{AA}}(t) = F_{\mathrm{p}}^{\mathrm{AA}} e_{\mathrm{p}}^{\mathrm{AA}}(t) + T_{\mathrm{p}}^{\mathrm{AA}} B_{\mathrm{a}} a_{\mathrm{u}}(t), \tag{24}$$

and

$$\begin{aligned} \dot{e}^{\mathrm{AA}}(t) =& F^{\mathrm{AA}} e^{\mathrm{AA}}(t) - L^{\mathrm{AA}} e_{\mathrm{p}}^{\mathrm{AA}}(t) + (I - H^{\mathrm{AA}} C) N \omega(t), \\ res_{\mathrm{AA}}(t) =& C e^{\mathrm{AA}}(t). \end{aligned} \tag{25}$$

Based on Condition 6) and according to (25), the impact of $e_{\mathrm{p}}^{\mathrm{AA}}(t)$ will appear in $res_{\mathrm{AA}}(t)$ for any $a_{\mathrm{u}}(t) \neq 0$.

Consider $e_{\mathrm{p}}^{\mathrm{AA}}(t)$ in (24) with the output $L^{\mathrm{AA}} e_{\mathrm{p}}^{\mathrm{AA}}(t)$ in order to construct the Rosenbrock system matrix $P_{\Sigma_{\mathrm{u}}}(s)$. To prevent stealthy attacks on the plant side filter, one needs to design this filter and L^{AA} such that the Rosenbrock system matrix $P_{\Sigma_{\mathrm{u}}}(s)$ has no non-minimum phase zero dynamics and is left-invertible [14]. The Rosenbrock system matrix $P_{\Sigma_{\mathrm{u}}}(s)$ being left-invertible is equivalent to the largest controllability subspace of the system $(L^{\mathrm{AA}}, F_{\mathrm{p}}^{\mathrm{AA}}, T_{\mathrm{p}}^{\mathrm{AA}} B_{\mathrm{a}})$ contained in $\ker(L^{\mathrm{AA}})$, and designated as $\mathscr{R}^*(\Sigma_{\mathrm{u}})$ being zero [98]. One has (refer to Theorem 8.22 in [98] and Theorem 5.6 in [102])

$$\mathscr{R}^*(\Sigma_{\mathrm{u}}) = \mathscr{V}^*(\Sigma_{\mathrm{u}}) \cap \mathscr{W}^*(\Sigma_{\mathrm{u}}), \tag{26}$$

where $\mathscr{V}^*(\Sigma_{\mathrm{u}})$ is the largest weakly unobservable subspace that is equivalent to the largest output-nulling subspace of the triplet $(L^{\mathrm{AA}}, F_{\mathrm{p}}^{\mathrm{AA}}, T_{\mathrm{p}}^{\mathrm{AA}} B_{\mathrm{a}})$, and $\mathscr{W}^*(\Sigma_{\mathrm{u}})$ is the smallest conditioned invariant subspace containing $\mathrm{Im}(T_{\mathrm{p}}^{\mathrm{AA}} B_{\mathrm{a}})$ [15].

As described in [98] and [102], these subspaces can be computed by using the following algorithm

$$\begin{aligned} \mathscr{V}_0 &= \mathrm{Ker}(L^{\mathrm{AA}}), \\ \mathscr{V}_k &= \mathscr{V}_0 \cap F_{\mathrm{p}}^{\mathrm{AA}\,-1}(\mathscr{V}_{k-1} + \mathrm{Im}(T_{\mathrm{p}}^{\mathrm{AA}} B_{\mathrm{a}})), \end{aligned} \tag{27}$$

and

$$\begin{aligned} \mathscr{W}_0 &= \mathrm{Im}(T_{\mathrm{p}}^{\mathrm{AA}} B_{\mathrm{a}}), \\ \mathscr{W}_k &= \mathscr{W}_0 + F_{\mathrm{p}}^{\mathrm{AA}}(\mathscr{W}_{k-1} \cap \mathrm{Ker}(L^{\mathrm{AA}})), \end{aligned} \tag{28}$$

where $\mathscr{V}_k$ and $\mathscr{W}_k$ converge to $\mathscr{V}^*(\Sigma_u)$ and $\mathscr{W}^*(\Sigma_u)$, respectively, in at most $k = n$ steps.

Given (26), (27), and (28), $\mathscr{R}^*(\Sigma_{\mathrm{u}}) = 0$, if $\mathscr{V}_0 \cap \mathscr{W}_0 = 0$, or equivalently,

$$\mathrm{Ker}(L^{\mathrm{AA}}) \cap \mathrm{Im}(T_{\mathrm{p}}^{\mathrm{AA}} B_{\mathrm{a}}) = 0. \tag{29}$$

The equation (29) implies that $\mathrm{Im}(T_{\mathrm{p}}^{\mathrm{AA}} B_{\mathrm{a}})$ should not be in the null space of L^{AA}, which is equivalent to

$$\mathrm{rank}\,(L^{\mathrm{AA}} T_{\mathrm{p}}^{\mathrm{AA}} B_{\mathrm{a}}) = \mathrm{rank}\,(T_{\mathrm{p}}^{\mathrm{AA}} B_{\mathrm{a}}). \tag{30}$$

Given that as per Assumption 3.5 one has $\mathrm{rank}(L^{\mathrm{AA}}) \geq \mathrm{rank}\,(T_{\mathrm{p}}^{\mathrm{AA}} B_{\mathrm{a}})$, therefore the matrix L^{AA} can be obtained such that (30) holds.

The Rosenbrock system matrix $P_{\Sigma_{\mathrm{u}}}(s)$ being left-invertible implies that for any $a_{\mathrm{u}}(t) \neq 0$, $L^{\mathrm{AA}} e_{\mathrm{p}}^{\mathrm{AA}}(t) \neq 0$.

Finally, in order to detect actuator cyber-attacks, the governing dynamics in (23) should be stable. This completes the proof of the Proposition 1. □

Proposition 3.2. *Under Assumption 3.5, the residual signal $res_{SA}(t) = y_p(t) - C\hat{x}^{SA}(t)$ is affected by the sensor cyber-attacks $a_y(t)$ and is decoupled from $a_u(t)$, $a_{cp}(t)$, $f_1(t)$, and $f_2(t)$ in the sense of Definition 3.5, if Conditions 1)-4), 6), and 9) of the Proposition 3.1 for $\ell = SA$, and the following conditions for the augmented error dynamics (21) hold, namely:*

(1) $T_p^{SA} B_a = 0$;

(2) the Rosenbrock system matrix

$$P_{\Sigma_y}(s) = \begin{bmatrix} sI - F_p^{SA} & K_p^{SA} D_a \\ L^{SA} & 0_{(n+p_f+p)\times p_a} \end{bmatrix},$$

does not have any non-minimum phase zero dynamics; and

(3) $\mathrm{rank}\,(L^{SA} K_p^{SA} D_a) = \mathrm{rank}\,(K_p^{SA} D_a)$.

Proof. The proof follows along similar lines to that of Proposition 3.1 and is omitted for sake of brevity. □

Remark 3.4. *Suppose Condition 8) of the Proposition 3.1 is not satisfied and $P_{\Sigma_u}(s)$ is not left-invertible. In this case, it has been shown in [14] that one can find an actuator cyber-attack $a_u(t) \neq 0$ such that $L^{AA}e_p^{AA}(t) = 0$. This type of cyber-attack has been represented in [14] and has been defined as "undetectable controllable attacks" in [15]. According to (24) and (25) the actuator cyber-attack signal $a_u(t)$ can affect the error $e^{AA}(t)$ only through $L^{AA}e_p^{AA}(t)$. Hence, the adversary has the capability of injecting a stealthy cyber-attack by using $a_u(t)$ that does not affect the residual signal $res_{AA}(t) = Ce^{AA}(t)$. Similarly, it can be shown that if Condition 3) of Proposition 3.2 is not satisfied and $P_{\Sigma_y}(s)$ is not left-invertible, the adversary can inject stealthy attack using $a_y(t)$ which does not affect the residual $res_{SA}(t)$.*

Remark 3.5. *In Propositions 3.1 and 3.2, there is no assumption on the nature, characteristics, and type of sensor and actuator cyber-attacks. This implies that by using the proposed methodology, one is capable of detecting and isolating detectable attacks, such as false data injection attacks, as well as undetectable attacks (refer to Definition 3.2), such as covert attacks and zero dynamics attacks. Furthermore, since as per Definition 3.4, we are using a threshold checking mechanism to make a decision on the anomalous status of the CPS, it would be still possible for adversaries to design their attack signals such that the residual remains below the threshold. Hence, in such a scenario, adversaries will try to reduce the amplitude of their attack signals to remain undetected which implies that the attack signals may not necessarily lead the CPS to dangerous conditions.*

Proposition 3.3. *The residual signal $res_{AF}(t) = y_p(t) - C\hat{x}^{AF}(t)$ is affected by the actuator fault $f_1(t)$ and is decoupled from $a_u(t)$, $a_y(t)$, $a_{cp}(t)$, and $f_2(t)$ in the sense of Definition 3.5, provided that $L^{AF} = 0$ and Conditions 1), 3), and 9) of the Proposition 3.1 hold for $\ell = AF$.*

Proof. In light of the Conditions 1) and 3) of the Proposition 3.1, and setting $\ell = \text{AF}$, (21) yields

$$\dot{\check{e}}^{\text{AF}}(t) = \check{F}^{\text{AF}}\check{e}^{\text{AF}}(t) + \check{B}_{\text{a}}^{\text{AF}}a_{\text{u}}(t) + \check{F}_1^{\text{AF}}f_1(t) - \check{K}_{\text{p}}^{\text{AF}}a_{\text{y}}(t) - \check{L}^{\text{AF}}a_{\text{cp}}(t) + \check{N}^{\text{AF}}\omega(t).$$

Moreover, by setting $L^{\text{AF}} = 0$, the dynamics of $e^{\text{AF}}(t)$ is governed by:

$$\dot{e}^{\text{AF}}(t) = F^{\text{AF}}e^{\text{AF}}(t) + (I - H^{\text{AF}}C)F_1f_1(t) + N\omega(t).$$

and consequently, the residual signal $res_{\text{AF}}(t) = Ce^{\text{AF}}(t)$ is only sensitive to the actuator fault $f_1(t)$. In addition, $\check{F}^{\text{AF}}$ should be Hurwitz in order to have a stable error dynamics $e^{\text{AF}}(t)$. This completes the proof of the Proposition 3. □

Proposition 3.4. *The residual signal $res_{SF}(t) = y_p(t) - C\hat{x}^{SF}(t)$ is affected by the pseudo actuator fault $f_2(t)$ and is decoupled from $a_u(t)$, $a_y(t)$, $a_{cp}(t)$, and $f_1(t)$ in the sense of Definition 3.5, provided that $L^{SF} = 0$ and Conditions 1), 2), and 9) of the Proposition 3.1 hold for $\ell = SF$.*

Proof. Setting $\ell = \text{SF}$, the proof follows along similar lines to that of Proposition 3.3 and is omitted for sake of brevity. □

As stated in [36], the Conditions 2) and 3) in the Proposition 3.1 are solvable if and only if $\text{rank}(CF_1) = \text{rank}(F_1)$; and $\text{rank}(CF_2) = \text{rank}(F_2)$. The next theorem provides sufficient conditions for isolability of sensor and actuator faults.

Theorem 3.1. *The residuals $res_{AF}(t)$ and $res_{SF}(t)$ can be simultaneously generated to detect and isolate $f_1(t)$ and $f_2(t)$ if $F_1^\top F_2 = 0$.*

Proof. In order to generate the residual signal $res_{\text{AF}}(t)$, the Condition 2) in Proposition 3.3 should hold, which can be interpreted as requiring

$$\text{Im}(I - H^{\text{AF}}C) \subset \text{Ker}(F_2^\top). \tag{31}$$

and at the same time, the impact of $f_1(t)$ should show up in the dynamics of $e(t)$, that implies $(I - H^{\text{AF}}C)F_1 \neq 0$. The latter condition is equivalent to having

$$\text{Im}(F_1^\top) \subset \text{Im}(I - H^{\text{AF}}C). \tag{32}$$

From (31) and (32), it can be inferred that $\text{Im}(F_1^\top) \subset \text{Ker}(F_2^\top)$, which implies that $F_1^\top F_2 = 0$. Note that the case of generating the residual signal $res_{\text{SF}}(t)$ provides one with the same result. This completes the proof of the Theorem 3.1. □

It follows from the definitions of F_1 and F_2 that the condition $F_1^\top F_2 = 0$ is always satisfied. Therefore, as long as Conditions 2) and 3) in Proposition 3.1 are solvable, the actuator faults and pseudo actuator faults can be detected and isolated.

Remark 3.6. *Given that L^{AF} and L^{SF} are equal to zero in the Propositions 3.3 and 3.4, in order to generate the residual signals $res_{AA}(t)$, $res_{SA}(t)$, $res_{AF}(t)$, and $res_{SF}(t)$ one needs to construct a bank of four filters (two on each side) with the states $z_p^{AA}(t)$, $z_c^{AA}(t)$, $z_p^{SA}(t)$, and $z_c^{SA}(t)$, and four UIO-based detectors with the states $\hat{x}^{AA}(t)$, $\hat{x}^{SA}(t)$, $\hat{x}^{AF}(t)$, and $\hat{x}^{SF}(t)$ according to the Propositions 3.1-3.4. In the Propositions 3.1 and 3.2, the matrices K_p^{AA} and T_p^{SA} have been utilized to decouple sensor cyber-attacks and actuator cyber-attacks in the sense of Definition 3.5 from the generated residual signals, respectively. Hence, one can conclude that there is no contradiction among the conditions to generate $res_{AA}(t)$ and $res_{SA}(t)$. Subsequently, from Theorem 3.1 it can be seen that no contradiction exists among the design conditions in the Propositions 3.3 and 3.4 to generate $res_{AF}(t)$ and $res_{SF}(t)$. Moreover, in the Propositions 3.3 and 3.4, the matrix L^ℓ has been employed to decouple the cyber-attack signals from $res_{AF}(t)$ and $res_{SF}(t)$, which indicates that there are no contradictions in the design conditions for the Propositions 3.1-3.4.*

Remark 3.7. *One application of the proposed centralized CAFDI methodology could be the detection and isolation of anomalies in a single UAV. Consider a UAV that is remotely controlled and receives its way points from the C&C center. An adversary is capable of performing man-in-the-middle cyber-attack to either hijack or destroy the UAV. Considering the availability of relatively cheap and powerful microcontrollers, it is reasonable to assume that a UAV can have adequate computational resources to deploy and run the proposed UIO-based detector and filters in its plant side. Moreover, given that in our proposed centralized CAFDI methodology, detection of anomalies occurs in the plant side, this information can be relayed back to the C&C side as a flag for corrective actions to be considered.*

3.2.5 The Case of Fully Non-Secure Communication Channels

In the previous subsection, under Assumption 3.5, we considered the existence of secure communication channels between the two side filters. Consequently, the generated residuals in the Propositions 3.1-3.4 were decoupled from the communication channel attack signal $a_{\text{cp}}(t)$. However, it is possible that adversaries compromise all the communication channels among the two side filters. Hence, in this subsection we remove

the Assumption 3.5 and consider the case where there exists no secure communication channel among the filters. Furthermore, the proposed filters in (15) and (16) are modified to address the CAFDI problem.

In order to develop the modified filters, one requires two communication channels, one from the C&C side filter to the plant side filter and the other from the plant side filter to the C&C side filter to transmit states of the two filters to one another. Moreover, the specified communication channels are assumed to be fully compromised by the adversaries.

The proposed filter on the plant side is modified in the following form:

$$\dot{z}_{\mathrm{p}}^{\ell}(t) = F_{\mathrm{p}}^{\ell} z_{\mathrm{p}}^{\ell}(t) + T_{\mathrm{p}}^{\ell} B u^{*}(t) + K_{\mathrm{p}}^{\ell} y_{\mathrm{p}}(t) + \delta_{\mathrm{p}}(t) L_{\mathrm{p}}^{\ell}(z_{\mathrm{p}}^{\ell}(t) - (z_{\mathrm{c}}^{\ell}(t) + D_{\mathrm{cp}} a_{\mathrm{cp}}(t))), \tag{33}$$

where $\delta_{\mathrm{p}}(t)$ denotes a scalar random variable in the interval $[\delta_1, \delta_2]$, δ_1 and δ_2 are positive real numbers, $z_{\mathrm{c}}^{\ell}(t)$ denotes the state of the C&C side filter transmitted to the plant side, and $a_{\mathrm{cp}}(t) \in \mathbb{R}^{n_{\mathrm{c}}}$ denotes the cyber-attack on the communication channel from the C&C side filter to the plant side filter with the signature D_{cp}.

Consequently, the modified filter on the C&C side can be expressed as

$$\dot{z}_{\mathrm{c}}^{\ell}(t) = F_{\mathrm{p}}^{\ell} z_{\mathrm{c}}^{\ell}(t) + T_{\mathrm{p}}^{\ell} B u(t) + K_{\mathrm{p}}^{\ell} y^{*}(t) + \delta_{\mathrm{c}}(t) L_{\mathrm{p}}^{\ell}(z_{\mathrm{c}}^{\ell}(t) - (z_{\mathrm{p}}^{\ell}(t) + D_{\mathrm{pc}} a_{\mathrm{pc}}(t))), \tag{34}$$

where $\delta_{\mathrm{c}}(t) \in [\delta_1, \delta_2]$ denotes a scalar random variable and $a_{\mathrm{pc}}(t) \in \mathbb{R}^{n_{\mathrm{c}}}$ denotes the cyber-attack on the communication channel from the plant side filter to the C&C side filter associated with the signature D_{pc}.

Assumption 3.6. *The adversary has access to all the communication channels among the two side filters, i.e.,* $D_{cp} = D_{pc} = I_n$.

Considering Assumption 3.6, and the dynamics (33) and (34), the governing dynamics of the error $e_{\mathrm{p}}^{\ell}(t) = z_{\mathrm{p}}^{\ell}(t) - z_{\mathrm{c}}^{\ell}(t)$ can be derived as follows:

$$\dot{e}_{\mathrm{p}}^{\ell}(t) = F_{\mathrm{p}}^{\ell}(t) e_{\mathrm{p}}^{\ell}(t) + T_{\mathrm{p}}^{\ell} B_{\mathrm{a}} a_{\mathrm{u}}(t) - K_{\mathrm{p}}^{\ell} D_{\mathrm{a}} a_{\mathrm{y}}(t) - \delta_{\mathrm{p}}(t) L_{\mathrm{p}}^{\ell} a_{\mathrm{cp}}(t) + \delta_{\mathrm{c}}(t) L_{\mathrm{p}}^{\ell} a_{\mathrm{pc}}(t), \tag{35}$$

where $F_{\mathrm{p}}^{\ell}(t) = F_{\mathrm{p}}^{\ell} + \delta(t) L_{\mathrm{p}}^{\ell}$, and $\delta(t) = \delta_{\mathrm{p}}(t) + \delta_{\mathrm{c}}(t)$.

Moreover, the dynamics of the UIO-based detector on the plant side, as given in (18), can now be rewritten in the following form:

$$\begin{aligned}\dot{z}^{\ell}(t) =& F^{\ell}z^{\ell}(t) + T^{\ell}Bu^{*}(t) + K^{\ell}y_{\mathrm{p}}(t) + L^{\ell}(z_{\mathrm{p}}^{\ell}(t) - (z_{\mathrm{c}}^{\ell}(t) + a_{\mathrm{cp}}(t))),\\ \hat{x}^{\ell}(t) =& z(t)^{\ell} + H^{\ell}y_{\mathrm{p}}(t).\end{aligned} \tag{36}$$

Consequently, the error dynamics of $e^{\ell}(t) = x(t) - \hat{x}^{\ell}(t)$ in (20) can be reformulated as follows:

$$\begin{aligned}\dot{e}^{\ell}(t) =& (A - H^{\ell}CA - K_1^{\ell}C)e^{\ell}(t) + (I - T^{\ell} - H^{\ell}C)(Bu(t)\\ &+ B_{\mathrm{a}}a_{\mathrm{u}}(t)) + (I - H^{\ell}C)F_1f_1(t) + (I - H^{\ell}C)F_2f_2(t)\\ &+ (I - H^{\ell}C)N\omega(t) - L^{\ell}(e_{\mathrm{p}}^{\ell}(t) - a_{\mathrm{cp}}(t)).\end{aligned} \tag{37}$$

Assumption 3.7. *The malicious adversary does not have knowledge on the parameters $\delta_p(t)L_p^{\ell}$ in (33) and $\delta_c(t)L_p^{\ell}$ in (34), however, the remaining parameters in (33), (34), and (36) are known to the adversary.*

Remark 3.8. *Given the randomness of the variables $\delta_p(t)$ and $\delta_c(t)$, it is reasonable to assume that adversaries do not know the values of the parameters stated in Assumption 3.7. Moreover, $\delta_p(t)$ and $\delta_c(t)$ can have any arbitrary probability distributions associated with them.*

Lemma 3.1. *Let Assumptions 3.6 and 3.7 hold and consider the CPS (12) under cyber-attacks and faults. Given the modified plant side filter (33) and the modified C&C side filter (34), adversaries cannot design communication channel attack signals $a_{cp}(t)$ and $a_{pc}(t)$ to ensure $L^{\ell}(e_p^{\ell}(t) - a_{cp}(t)) = 0$, $\forall t > 0$ in the error dynamics (37).*

Proof. We consider three possible scenarios, namely $a_{\mathrm{cp}}(t) = 0$ and $a_{\mathrm{pc}}(t) \neq 0$, $a_{\mathrm{pc}}(t) = 0$ and $a_{\mathrm{cp}}(t) \neq 0$, and finally $a_{\mathrm{cp}}(t) \neq 0$ and $a_{\mathrm{pc}}(t) \neq 0$. Let $a_{\mathrm{cp}}(t) = 0$ and consider the error dynamics $e_{\mathrm{p}}^{\ell}(t)$ in (35) with the output $L^{\ell}e_{\mathrm{p}}^{\ell}(t)$. Given the unknown random parameter $\delta_{\mathrm{c}}(t)L_{\mathrm{p}}^{\ell}$ in the dynamics of $e_{\mathrm{p}}^{\ell}(t)$, adversaries cannot design $a_{\mathrm{pc}}(t) \neq 0$ such that $-K_{\mathrm{p}}^{\ell}D_{\mathrm{a}}a_{\mathrm{y}}(t) + \delta_{\mathrm{c}}(t)L_{\mathrm{p}}^{\ell}a_{\mathrm{pc}}(t) = 0$ or $T_{\mathrm{p}}^{\ell}B_{\mathrm{a}}a_{\mathrm{u}}(t) + \delta_{\mathrm{c}}(t)L_{\mathrm{p}}^{\ell}a_{\mathrm{pc}}(t) = 0$, $\forall t > 0$. Moreover, since adversaries do not know $F_{\mathrm{p}}^{\ell}(t)$ and $\delta_{\mathrm{c}}(t)L_{\mathrm{p}}^{\ell}$, they cannot execute zero dynamics attacks or perform "undetectable controllable attacks" (refer to Remark 3.4) on the triplet $(L^{\ell}, F_{\mathrm{p}}^{\ell}(t), \delta_{\mathrm{c}}(t)L_{\mathrm{p}}^{\ell})$. A similar argument can be used to show that adversaries cannot make $L^{\ell}(e_{\mathrm{p}}^{\ell}(t) - a_{\mathrm{cp}}(t)) = 0$, $\forall t > 0$, in the case of $a_{\mathrm{pc}}(t) = 0$ and $a_{\mathrm{cp}}(t) \neq 0$.

Let $a_{\rm cp}(t) \neq 0$ and $a_{\rm pc}(t) \neq 0$. Since the parameter $F_{\rm p}^{\ell}(t)$ is unknown to adversaries, they cannot design the communication attack signal $a_{\rm cp}(t)$ to eliminate the impact of cyber-attacks in (35) from the signal $L^{\ell}(e_{\rm p}^{\ell}(t) - a_{\rm cp}(t))$. This completes the proof of the lemma. □

Proposition 3.5. *Let Assumptions 3.6 and 3.7 hold and consider the modified plant side filter (33), the modified C&C side filter (34), and the UIO-based detector (36), where $\ell = AA$. The residual signal $res_{AA}(t) = y_p(t) - C\hat{x}^{AA}(t)$ is affected by $a_u(t)$, $a_{cp}(t)$, and $a_{pc}(t)$ and is decoupled from $a_{\rm y}(t)$, $f_1(t)$, and $f_2(t)$ in the sense of Definition 3.5 provided that the Conditions 1)-3), and 5) of the Proposition 3.1 and the following conditions hold for the error dynamics (35) and (37), namely:*

(1) the triplet $(C,\ F^{\ell},\ L^{\ell})$ is input observable in the sense of Definition 3.3;

(2) F^{ℓ} is Hurwitz;

(3) $F_p^{\ell}(t)$ is designed such that there exists a symmetric positive definite matrix $Q_p^{\ell}(t)$ that satisfies

$$F_p^{\ell}(t)^{\top} + F_p^{\ell}(t) = -Q_p^{\ell}(t), \tag{38}$$

where $\beta_{ep} I_n \leq Q_p^{\ell}(t)$, and β_{ep} is a positive scalar.

Proof. Let $\ell = \text{AA}$. Under the Conditions 1)-3), and 5) in the Proposition 3.1, the error dynamics (35) and (37) become

$$\dot{e}_{\rm p}^{\rm AA}(t) = F_{\rm p}^{\rm AA}(t) e_{\rm p}^{\rm AA}(t) + T_{\rm p}^{\rm AA} B_{\rm a} a_{\rm u}(t) - \delta_{\rm p}(t) L_{\rm p}^{\rm AA} a_{\rm cp}(t) + \delta_{\rm c}(t) L_{\rm p}^{\rm AA} a_{\rm pc}(t), \tag{39}$$

and

$$\dot{e}^{\rm AA}(t) = F^{\rm AA} e^{\rm AA}(t) - L^{\rm AA}(e_{\rm p}^{\rm AA}(t) - a_{\rm cp}(t)) + (I - H^{\rm AA} C) N \omega(t), \tag{40}$$

respectively.

Consider the error dynamics in (39) with the output $L^{\rm AA}(e_{\rm p}^{\rm AA}(t) - a_{\rm cp}(t))$. Given that adversaries do not know $F_{\rm p}^{\rm AA}(t)$, by utilizing the actuator attack signal $a_{\rm u}(t)$, they cannot excite the zero dynamics of the triplet

$(L^{\text{AA}}, F_{\text{p}}(t), T_{\text{p}}^{\text{AA}} B_{\text{a}})$ or carry out "undetectable controllable attacks" (refer to Remark 3.4). Moreover, according to Lemma 3.1, and given that $(C,\ F^{\ell},\ L^{\ell})$ is input observable, the impact of cyber-attacks $a_{\text{u}}(t)$, $a_{\text{cp}}(t)$, and $a_{\text{pc}}(t)$ in (39) will be manifested in the residual signal $res_{\text{AA}}(t) = Ce^{\text{AA}}(t)$ through the error dynamics (40).

Consequently, in order to detect cyber-attacks, one needs to show that the error dynamics (40) and (39) are stable. The dynamics (40) is stable if F^{AA} is Hurwitz. Furthermore, consider the Lyapunov function candidate $V_{\text{ep}}(e_{\text{p}}^{\text{AA}}(t)) = e_{\text{p}}^{\text{AA}}(t)^{\top} e_{\text{p}}^{\text{AA}}(t)$.

The derivative of $V_{\text{ep}}(e_{\text{p}}^{\text{AA}}(t))$ along the trajectories of (39) can be obtained as

$$\begin{aligned} \dot{V}_{\text{ep}}(e_{\text{p}}^{\text{AA}}(t)) =& \dot{e}_{\text{p}}^{\text{AA}}(t)^{\top} e_{\text{p}}^{\text{AA}}(t) + e_{\text{p}}^{\text{AA}}(t)^{\top} \dot{e}_{\text{p}}^{\text{AA}}(t) \\ =& - e_{\text{p}}^{\text{AA}}(t)^{\top} Q_{\text{p}}^{\text{AA}}(t) e_{\text{p}}^{\text{AA}}(t). \end{aligned} \tag{41}$$

It follows from (41) that $\dot{V}_{\text{ep}}(e_{\text{p}}^{\text{AA}}(t)) \leq -\beta_{\text{ep}} \|e_{\text{p}}^{\text{AA}}(t)\|^2$, which implies that under Condition 2), the error dynamics (39) is stable [103]. This completes the proof of the proposition. □

Remark 3.9. *According to the Condition 3) in the Proposition 3.5, one needs to design $F_p^{\ell}(t)$ such that (38) holds. Given that $F_p^{\ell}(t) = F_p^{\ell} + \delta(t) L_p^{\ell}$, one has $F_p^{\ell}(t) + F_p^{\ell}(t)^{\top} = \tilde{F}_p^{\ell} + \delta(t)\tilde{L}_p^{\ell}$, where $\tilde{F}_p^{\ell} = F_p^{\ell} + F_p^{\ell^{\top}}$ and $\tilde{L}_p^{\ell} = L_p^{\ell} + L_p^{\ell^{\top}}$ are symmetric matrices. Thus, one can use F_p^{ℓ} to design $\tilde{F}_p^{\ell}$ such that $\tilde{F}_p^{\ell} + 2\delta_1 \tilde{L}_p^{\ell}$ and $\tilde{F}_p^{\ell} + 2\delta_2 \tilde{L}_p^{\ell}$ are negative definite, which is the sufficient condition for (38) to hold.*

Proposition 3.6. *Under the Assumptions 3.6 and 3.7, the residual signal $res_{SA} = y_p(t) - C\hat{x}^{SA}(t)$ generated by utilizing the modified plant side filter (33), the modified C&C side filter (34), and the UIO-based detector (36) is affected by $a_y(t)$, $a_{cp}(t)$, and $a_{pc}(t)$ and is decoupled from $a_u(t)$, $f_1(t)$, and $f_2(t)$ in the sense of the Definition 3.5 provided that the Conditions 1)-3) of the Proposition 3.1, the Condition 1) of the Proposition 3.2, and the Conditions 1)-3) of the Proposition 3.5 for $\ell = SA$ hold.*

Proof. The proof follows along similar lines to that of the Proposition 3.5 and is omitted for sake of brevity. □

Remark 3.10. *In the Proposition 3.1, the residual signal is only sensitive to the actuator attack signal $a_u(t)$, however, due to not having any secure communication channel in the Proposition 3.5, the residual signal is*

affected by the set of signals $\{a_u(t), a_{cp}(t), a_{pc}(t)\}$. *Similarly, the residuals in the Propositions 3.2 and 3.6 are affected by the sensor attack signal* $a_y(t)$ *and the set* $\{a_y(t), a_{cp}(t), a_{pc}(t)\}$. *Therefore, when there is no secure communication channel, the generated residuals in the Propositions 3.5 and 3.6 cannot be decoupled from cyber-attacks on the communication channels between the two side filters. Moreover, the given conditions in the Propositions 3.3 and 3.4 can be used to design and implement actuator and sensor fault detection and isolation modules, respectively. In other words, having non-secure communication channels among the two side filters does not affect the performance of our proposed fault detection and isolation methodologies and modules.*

3.3 Distributed Cyber-Attack and Fault Detection and Isolation Methodology for Interconnected CPS

In this section, our objective is to extend our results in the Subsection 3.2.5 and address the CAFDI problem for large scale interconnected CPS given by (13) through a distributed architecture. The proposed CAFDI methodology is distributed in the sense that CAFDI modules on each subsystem communicate information with their neighboring subsystems. Hence, each subsystem can detect and isolate its cyber-attacks and faults as well as anomalies in its neighboring subsystems. Furthermore, we consider the detection of both detectable and undetectable cyber-attacks (refer to Definition 3.2) in our proposed methodology.

3.3.1 UIO-Based Detectors and Filters Design for the i-th Subsystem

In this subsection, we use a similar approach as in the Subsection 3.2.5 to design both side filters and UIO-based detectors. Each subsystem is equipped with a bank of filters both on its plant side and its C&C side, where both side filters transmit their states to one another over compromised communication channels. In this methodology, we consider the existence of one secure communication channel among the two side filters. Moreover, a detector by using the UIO is designed and utilized on the plant side of each subsystem. Each UIO-based detector receives estimated states of the UIO-based detectors in its neighborhood through compromised communication channels. It should be noted that in our proposed distributed CAFDI methodology for interconnected CPS, two side filters of each subsystem do not transmit information to the

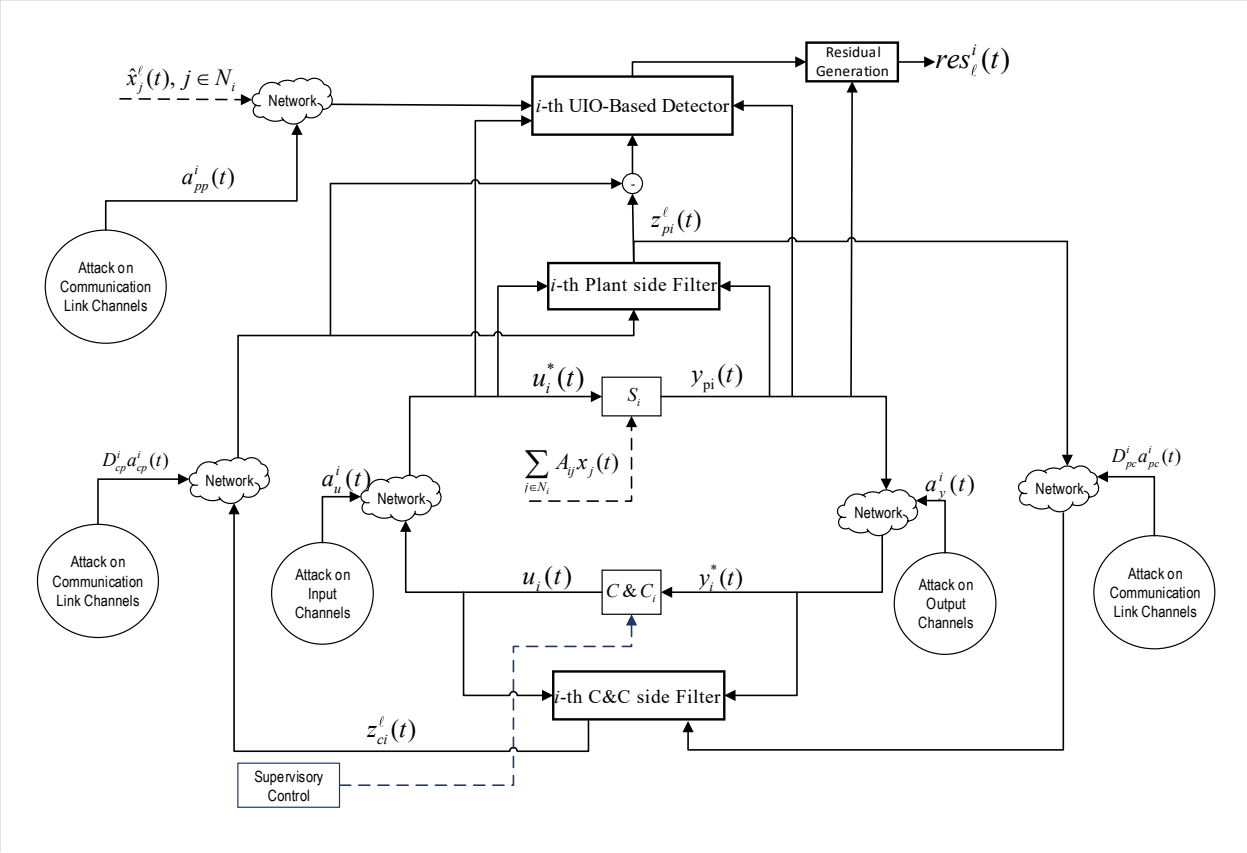

Figure 3.3: The distributed CAFDI methodology for the i-th subsystem, where D_{cp}^i and D_{pc}^i are rank deficient matrices that denote the signatures of the cyber-attack signals on the communication channels between the C&C and the plant side filters.

nearby filters that are in the other subsystems. The proposed distributed CAFDI methodology is depicted in Figure 3.3.

Consequently, the proposed filter on the plant side of $\mathcal{S}_i$ is given by

$$\begin{aligned}\dot{z}_{\mathrm{p}i}^\ell(t) =& F_{\mathrm{p}i}^\ell z_{\mathrm{p}i}^\ell(t) + T_{\mathrm{p}i}^\ell B_i u_i^*(t) + K_{\mathrm{p}i}^\ell y_{\mathrm{p}i}(t) + \delta_{\mathrm{p}i}(t) L_{\mathrm{p}i}^\ell (z_{\mathrm{p}i}^\ell(t) \\ & - (z_{\mathrm{c}i}^\ell(t) + D_{\mathrm{cp}}^i a_{\mathrm{cp}}^i(t))),\end{aligned} \tag{42}$$

where $z_{\mathrm{p}i}^\ell(t) \in \mathbb{R}^{n_i}$ denotes the state of the plant side filter on the i-th subsystem, $y_{\mathrm{p}i}(t) = C_i x_i(t)$ denotes the measured output of $\mathcal{S}_i$ on the plant side, $z_{\mathrm{c}i}(t)^\ell \in \mathbb{R}^{n_i}$ denotes the state of the C&C side filter of $\mathcal{S}_i$ which is transmitted to the plant side, $a_{\mathrm{cp}}^i(t) \in \mathbb{R}^{n_{\mathrm{c}i}}$ denotes the cyber-attack on the communication channel from the C&C side filter of $\mathcal{S}_i$ to its plant side filter with the signature D_{cp}^i, and $\delta_{\mathrm{p}i}(t) \in [\delta_1, \delta_2]$ denotes a

scalar random variable.

Moreover, the proposed filter on the C&C side of the i-th subsystem is governed by

$$\begin{aligned}\dot{z}^{\ell}_{\mathrm{c}i}(t) =& F^{\ell}_{\mathrm{p}i} z^{\ell}_{\mathrm{c}i}(t) + T^{\ell}_{\mathrm{p}i} B_i u_i(t) + K^{\ell}_{\mathrm{p}i} y^{*}_i(t) \\ &+ \delta_{\mathrm{c}i}(t) L^{\ell}_{\mathrm{c}i}(z^{\ell}_{\mathrm{c}i}(t) - (z_{\mathrm{p}i}(t) + D^{i}_{\mathrm{pc}} a^{i}_{\mathrm{pc}}(t))),\end{aligned} \tag{43}$$

where $\delta_{\mathrm{c}i}(t) \in [\delta_1, \delta_2]$ is a scalar random variable, and $a^{i}_{\mathrm{pc}}(t) \in \mathbb{R}^{n_{\mathrm{c}i}}$ denotes the cyber-attack on the communication channel from the plant side filter of $\mathcal{S}_i$ to its C&C side filter with the signature D^{i}_{pc}.

Let us define the error $e^{\ell}_{\mathrm{p}i}(t) = z^{\ell}_{\mathrm{p}i}(t) - z^{\ell}_{\mathrm{c}i}(t)$ for the i-th subsystem. The dynamics of the error $e^{\ell}_{\mathrm{p}i}(t)$ can be derived in the following form:

$$\begin{aligned}\dot{e}^{\ell}_{\mathrm{p}i}(t) =& F^{\ell}_{\mathrm{p}i}(t) e^{\ell}_{\mathrm{p}i}(t) + T^{\ell}_{\mathrm{p}i} B^{i}_{\mathrm{a}} a^{i}_{\mathrm{u}}(t) - K^{\ell}_{\mathrm{p}i} D^{i}_{\mathrm{a}} a^{i}_{\mathrm{y}}(t) \\ &- \delta_{\mathrm{p}i}(t) L^{\ell}_{\mathrm{p}i} D^{i}_{\mathrm{cp}} a^{i}_{\mathrm{cp}}(t) + \delta_{\mathrm{c}i}(t) L^{\ell}_{\mathrm{c}i} D^{i}_{\mathrm{pc}} a^{i}_{\mathrm{pc}}(t),\end{aligned} \tag{44}$$

where $F^{\ell}_{\mathrm{p}i}(t) = F^{\ell}_{\mathrm{p}i} + \delta_{\mathrm{c}i}(t) L^{\ell}_{\mathrm{c}i} + \delta_{\mathrm{p}i}(t) L^{\ell}_{\mathrm{p}i}$.

The UIO-based detector of $\mathcal{S}_i$ can be expressed in the following form:

$$\begin{aligned}\dot{z}^{\ell}_{i}(t) =& F^{\ell}_{i} z^{\ell}_{i}(t) + T^{\ell}_{i} B_i u^{*}_i(t) + K^{\ell}_{i} y_{\mathrm{p}i}(t) + \sum_{j \in \mathcal{N}_i} A_{ij}(\hat{x}^{\ell}_{j}(t) \\ &+ D^{ij}_{\mathrm{pp}} a^{ij}_{\mathrm{pp}}(t)) + \delta_{\mathrm{z}i}(t) L^{\ell}_{i}(z^{\ell}_{\mathrm{p}i}(t) - (z^{\ell}_{\mathrm{c}i}(t) + D^{i}_{\mathrm{cp}} a^{i}_{\mathrm{cp}}(t))),\end{aligned} \tag{45}$$

where $\delta_{\mathrm{z}i}(t) \in [\delta_1, \delta_2]$ denotes a random variable, $\hat{x}^{\ell}_{j}(t) = z^{\ell}_{i}(t) + H^{\ell}_{i} y_{\mathrm{p}i}(t)$ denotes the estimation of the state in the j-th subsystem which is transmitted through a communication channel to the UIO-based detector in $\mathcal{S}_i$, and $a^{ij}_{\mathrm{pp}}(t) \in \mathbb{R}^{n_j}$ denotes the malicious attack signal on the communication channel among $\mathcal{S}_i$ and $\mathcal{S}_j$ with the signature D^{ij}_{pp}, for $j \in \mathcal{N}_i$. Moreover, the matrices F^{ℓ}_{i}, T^{ℓ}_{i}, K^{ℓ}_{i}, L^{ℓ}_{i}, and H^{ℓ}_{i} are of appropriate dimensions and satisfy $K^{\ell}_{i} = K^{\ell}_{1i} + K^{\ell}_{2i}$, $F^{\ell}_{i} = A_i - H^{\ell}_{i} C_i A_i - K^{\ell}_{1i} C_i$, where K^{ℓ}_{1i} is a matrix of appropriate dimension, and $K^{\ell}_{2i} = F^{\ell}_{i} H^{\ell}_{i}$.

Consequently, the error dynamics of $e_i^\ell(t) = x_i(t) - \hat{x}_i^\ell(t)$ in the subsystem i can be expressed by

$$\begin{aligned}\dot{e}_i^\ell(t) =& F_i^\ell e_i^\ell(t) + (I - T_i^\ell - H_i^\ell C_i)(B_i u_i(t) + B_{\mathrm{a}}^i a_{\mathrm{u}}^i(t)) \\ &+ (I - H_i^\ell C_i) \sum_{j \in \mathcal{N}_i} (A_{ij}(e_j^\ell(t) - D_{\mathrm{pp}}^{ij} a_{\mathrm{pp}}^{ij}(t))) \\ &+ (I - H_i^\ell C_i)(F_1^i f_1^i(t) + F_2^i f_2^i(t)) \\ &- \delta_{\mathrm{z}i}(t) L_i^\ell (e_{\mathrm{p}i}^\ell(t) - D_{\mathrm{cp}}^i a_{\mathrm{cp}}^i(t)). \end{aligned} \tag{46}$$

where $z_i^\ell(t) \in \mathbb{R}^{n_i + p_{\mathrm{f}i} + p_i}$, and $\hat{x}_i^\ell(t) \in \mathbb{R}^{n_i + p_{\mathrm{f}i} + p_i}$ is the estimated states by the detector of $\mathcal{S}_i$.

Moreover, by utilizing (45), one can generate a residual signal on the plant side of $\mathcal{S}_i$ in the following form:

$$res_\ell^i(t) = y_{\mathrm{p}i}(t) - C_i \hat{x}_i^\ell(t) = C_i e_i^\ell(t). \tag{47}$$

Definition 3.6. *A cyber-attack/fault on the i-th subsystem is detected if the following inequality is satisfied for the residual signal (47):*

$$\|res_\ell^i(t)\|_2 > \eta_i^\ell,$$

where $\eta_i^\ell > 0$ is a pre-specified threshold.

Similar to the value of η in Definition 3.4 (see also Remark 3.3), the prescribed threshold η_i^ℓ can be computed by means of Monte Carlo simulation runs for the healthy system.

Definition 3.7. *The generated residual signal in (47) which belongs to the i-th subsystem is decoupled from an anomalous signal in the set $\{a_u^i(t), a_y^i(t), a_{cp}^i(t), a_{pc}^i(t), a_{pp}^i(t), f_1^i(t),\ f_2^i(t)\}$ if that anomalous signal does not affect the dynamics and trajectories of $res_\ell^i(t)$.*

We consider the following assumptions throughout this subsection.

Assumption 3.8. *The q_{cp}^i-th communication channel from the C&C side filter to the plant side filter and the q_{pc}^i-th communication channel from the plant side filter to the C&C side filter are secured, where $q_{cp}^i, q_{pc}^i \in \{1, \ldots, n_i\}$, i.e., rank$(D_{cp}^i) = n_i - 1$ and rank$(D_{pc}^i) = n_i - 1$, for $i, j = 1, \ldots, N$. Moreover, all the communication channels among the nearby UIO-based detectors can be compromised, i.e., rank$(D_{pp}^{ij}) = n_j + p_{fj} + p_j$.*

Assumption 3.9. *In the plant side filter (42), the C&C side filter (43), and the UIO-based detector (45) of $\mathcal{S}_i$, only the random parameters $\delta_{pi}(t)L^{\ell}_{pi}$, $\delta_{ci}(t)L^{\ell}_{pi}$, and $\delta_{zi}(t)L^{\ell}_{i}$ are unknown to the adversary, respectively, and the adversary knows the other parameters.*

Remark 3.11. *The random variables $\delta_{pi}(t)$, $\delta_{ci}(t)$, and $\delta_{zi}(t)$ can have any arbitrary probability distributions.*

Proposition 3.7. *Consider the Assumptions 3.8 and 3.9, the plant side filter (42), the C&C side filter (43), and the UIO-based detector (45). The residual signal $res^{AA}_{i}(t) = y_{pi}(t) - C_i\hat{x}^{AA}_{i}(t)$ in the i-th subsystem is affected by actuator cyber-attack signals $a^{i}_{u}(t)$ in $\mathcal{S}_i$ and $a^{j}_{u}(t)$ in $\mathcal{S}_j$ and malicious attack signals $a^{ij}_{pp}(t)$ and $a^{jr}_{pp}(t)$, for $j \in \mathcal{N}_i$ and $r \in \mathcal{N}_j$. Moreover, the generated $res^{AA}_{i}(t)$ is decoupled from the sets of anomalous signals $\{a^{i}_{y}(t), f^{i}_{1}(t), f^{i}_{2}(t), a^{j}_{y}(t), f^{j}_{1}(t), f^{j}_{2}(t)\}$ and $\{a^{i}_{cp}(t), a^{i}_{pc}(t), a^{j}_{cp}(t), a^{j}_{pc}(t)\}$ in the sense of Definition 3.7 if for the error dynamics (44) and (46) and every $i = 1, \ldots, N$, the following conditions hold:*

(1) $T^{\ell}_{i} = I - H^{\ell}_{i}C_i$;

(2) $(I - H^{\ell}_{i}C_i)F^{i}_{1} = 0$;

(3) $(I - H^{\ell}_{i}C_i)F^{i}_{2} = 0$;

(4) $L^{\ell}_{ci}D^{i}_{pc} = 0$, $L^{\ell}_{pi}D^{i}_{cp} = 0$, $L^{\ell}_{i}D^{i}_{cp} = 0$;

(5) the triplet $(C_i, F^{\ell}_{i}, \bar{l}^{\ell}_{i})$ is input observable, where $\bar{l}^{\ell}_{i}$ is the q^{i}_{cp}-th column of L^{ℓ}_{i};

(6) F^{ℓ}_{i} is Hurwitz;

(7) $F^{\ell}_{pi}(t)$ is designed such that

$$F^{\ell}_{pi}(t)^{\top} + F^{\ell}_{pi}(t) = -Q^{\ell}_{pi}(t), \tag{48}$$

where $Q^{\ell}_{pi}(t)$ is a symmetric positive definite matrix that satisfies $\beta^{i}_{ep}I_n \leq Q^{\ell}_{pi}(t)$, and β^{i}_{ep} is a positive scalar;

(8) and $K^{AA}_{pi}D^{i}_{a} = 0$.

Proof. Let $\ell = \text{AA}$. Considering the provided conditions in this proposition, the dynamics of errors $e_i^{\text{AA}}(t)$ and $e_j^{\text{AA}}(t)$ can be derived in the following form, namely:

$$\dot{e}_i^{\text{AA}}(t) = F_i e_i^{\text{AA}}(t) + (I - H_i^{\text{AA}} C_i) \sum_{j \in \mathcal{N}_i} A_{ij}(e_j^{\text{AA}}(t) - D_{\text{pp}}^{ij} a_{\text{pp}}^{ij}(t)) - \delta_{\text{z}i}(t) L_i^{\text{AA}} e_{\text{p}i}^{\text{AA}}(t), \tag{49}$$

and

$$\dot{e}_j^{\text{AA}}(t) = F_j e_j^{\text{AA}}(t) + (I - H_j^{\text{AA}} C_j) \sum_{r \in \mathcal{N}_j} A_{jr}(e_r^{\text{AA}}(t) - D_{\text{pp}}^{jr} a_{\text{pp}}^{jr}(t)) - \delta_{\text{z}j}(t) L_j^{\text{AA}} e_{\text{p}j}^{\text{AA}}(t), \tag{50}$$

where $r \in \mathcal{N}_j$. From (49) and (50) it can be inferred that $res_i^{\text{AA}}(t) = C_i e_i^{\text{AA}}$ is affected by the actuator attack signals $a_{\text{u}}^i(t)$ and $a_{\text{u}}^j(t)$.

The remainder of the proof follows along similar lines to that of Propositions 3.1 and 3.5. □

Proposition 3.8. *Let the Assumptions 3.8 and 3.9 hold and set $\ell = SA$. The residual signal $res_i^{SA}(t) = y_{pi}(t) - C_i \hat{x}_i^{SA}(t)$ generated by using the plant side filter (42), the C&C side filter (43), and the UIO-based detector (45) in $\mathcal{S}_i$ is affected by sensor cyber-attack signals $a_y^i(t)$ in the subsystem i and $a_y^j(t)$ in the j-th subsystem and communication attack signals $a_{pp}^{ij}(t)$ and $a_{pp}^{jr}(t)$, for $j \in \mathcal{N}_i$ and $r \in \mathcal{N}_j$. Moreover, $res_i^{SA}(t)$ is decoupled from the sets of anomalous signals $\{a_u^i(t), f_1^i(t), f_2^i(t), a_u^j(t), f_1^j(t), f_2^j(t)\}$ and $\{a_{cp}^i(t), a_{pc}^i(t), a_{cp}^j(t), a_{pc}^j(t)\}$ in the sense of Definition 3.7 if for every $i = 1, \ldots, N$, $T_{pi}^{SA} B_a^i = 0$ and the Conditions 1)-7) of the Proposition 3.7 hold for the error dynamics (44) and (46).*

Proof. The proof follows along similar lines to that of Propositions 3.5 and 3.7 and is omitted for sake of brevity. □

Proposition 3.9. *Let the Assumption 3.8 holds and set $\ell = AF$. The residual signal $res_{AF}^i(t) = y_{pi}(t) - C_i \hat{x}_i^{AF}(t)$ that is generated in the i-th subsystem by using the UIO-based detector (45) is affected by actuator faults $f_1^i(t)$ and $f_1^j(t)$ and malicious attack signals $a_{pp}^{ij}(t)$ and $a_{pp}^{jr}(t)$, for $j \in \mathcal{N}_i$ and $r \in \mathcal{N}_j$. Moreover, $res_{AF}^i(t)$ is decoupled from the sets of anomalous signals $\{a_u^i(t), a_y^i(t), f_2^i(t), a_u^j(t), a_y^j(t), f_2^j(t)\}$ and $\{a_{cp}^i(t), a_{pc}^i(t), a_{cp}^j(t), a_{pc}^j(t)\}$ in the sense of the Definition 3.7, if the Conditions 1), 3), 4) and 6) of the Proposition 3.7 hold and $L_i^\ell = 0$.*

Proof. The proof follows along similar lines to that of the Propositions 3.3 and 3.7. □

Proposition 3.10. *Consider the Assumption 3.8 and let $\ell = SF$ for the modified UIO-based detector in (45). The generated residual signal $res^i_{SF}(t) = y_{pi}(t) - C_i\hat{x}^{SF}_i(t)$ in $\mathcal{S}_i$ is affected by pseudo actuator faults $f^i_2(t)$ and $f^j_2(t)$ and communication channel attack signals $a^{ij}_{pp}(t)$ and $a^{jr}_{pp}(t)$, for $j \in \mathcal{N}_i$ and $r \in \mathcal{N}_j$. Furthermore, $res^i_{SF}(t)$ is decoupled from the sets of anomalous signals $\{a^i_u(t), a^i_y(t), f^i_1(t), a^j_u(t), a^j_y(t), f^j_1(t)\}$ and $\{a^i_{cp}(t), a^i_{pc}(t), a^j_{cp}(t), a^j_{pc}(t)\}$ in the sense of the Definition 3.7, if the Conditions 1), 2), 4), and 6) of the Proposition 3.7 hold and $L^\ell_i = 0$.*

Proof. The proof follows along similar lines to that of the Propositions 3.3 and 3.7 and is omitted for sake of brevity. □

Theorem 3.2. *The residual signals $res^i_{AF}(t)$ and $res^i_{SF}(t)$ in the Propositions 3.9 and 3.10, respectively, can be simultaneously generated to detect and isolate $f^i_1(t)$ and $f^i_2(t)$ if ${F^i_1}^\top F^i_2 = 0$, for $i = 1, \ldots, N$.*

Proof. The proof follows along similar lines to that of Theorem 3.1. □

Remark 3.12. *It should be pointed out that in the Propositions 3.7-3.10, for detecting anomalies, i.e., faults and cyber-attacks, in the neighboring subsystems, the matrix H^ℓ_i should be designed such that $(I - H^\ell_i C_i)A_{ij} \neq 0$, for $i, j = 1, \ldots, N$ and $\ell \in \{SA, AA, SF, AF\}$.*

3.3.2 Non-Secure Communication Channels Among Two Side Filters and Nearby UIO-Based Detectors

In this subsection, we consider the case where there is no secure communication channel among the plant side filter (42) and the C&C side filter (43), and vice versa. Hence, in the following, we investigate the performance of our proposed CAFDI methodology in the previous Subsection under the following assumption.

Assumption 3.10. *The adversary has access to all the communication channels among the two side filters and nearby UIO-based detectors, i.e., rank(D^i_{cp}) = rank$(D^i_{pc}) = n_i$ and rank$(D^{ij}_{pp}) = n_j + p_{fj} + p_j$, for $i, j = 1, \ldots, N$.*

Under the Assumption 3.10, the Condition 4) in the Proposition 3.7 cannot be satisfied. Hence, the impact of cyber-attacks on the communication channels among the two side filters (43) and (42) and the nearby UIO-based detectors in (45) cannot be eliminated from the generated residuals in the Propositions 3.7-3.10.

Corollary 3.1 (Proposition 3.7)**.** *Consider the Assumptions 3.9 and 3.10. Since the Condition 4) in the Proposition 3.7 cannot be satisfied, the generated residual* $res_i^{AA}(t) = y_{pi}(t) - C_i\hat{x}_i^{AA}(t)$ *cannot be decoupled from the sets of cyber-attack signals* $U_a^i = \{a_{cp}^i(t), a_{pc}^i(t), a_{pp}^{ij}(t)\}$ *in* $\mathcal{S}_i$ *and* $U_a^j = \{a_{cp}^j(t), a_{pc}^j(t), a_{pp}^{jr}(t)\}$ *in the subsystem* $\mathcal{S}_j$ *in the sense of the Definition 3.7, for* $j \in \mathcal{N}_i$ *and* $r \in \mathcal{N}_j$.

Proof. The proof follows along similar lines to that of the Lemma 3.1 and Proposition 3.7. □

Corollary 3.2 (Proposition 3.8)**.** *Let the Assumptions 3.9 and 3.10 hold. The residual signal* $res_i^{SA}(t) = y_{pi}(t) - C_i\hat{x}_i^{SA}(t)$ *generated in the Proposition 3.8 cannot be decoupled from the sets of attack signals* $Y_a^i = \{a_{cp}^i(t), a_{pc}^i(t), a_{pp}^{ij}(t)\}$ *in the subsystem* i *and* $Y_a^j = \{a_{cp}^j(t), a_{pc}^j(t), a_{pp}^{jr}(t)\}$ *in the* j*-th subsystem in the sense of the Definition 3.7, where* $j \in \mathcal{N}_i$ *and* $r \in \mathcal{N}_j$.

Proof. The proof follows along similar lines to that of the Lemma 3.1 and Proposition 3.8 and is omitted for sake of brevity. □

Remark 3.13. *The Corollaries 3.1 and 3.2 indicate the importance of having one secure communication channel from the C&C side filter (43) to the plant side filter (42), and vice versa. Hence, under the Assumption 3.10, although the residual signals* $res_i^{AA}(t)$ *and* $res_i^{SA}(t)$ *can be used to detect and isolate actuator and sensor cyber-attacks, respectively, they are also affected by the sets of anomalous signals* $\{a_{cp}^i(t), a_{pc}^i(t), a_{pp}^{ij}(t)\}$ *and* $\{a_{cp}^j(t), a_{pc}^j(t), a_{pp}^{jr}(t)\}$.

Remark 3.14. *In this chapter, two main centralized and distributed CAFDI methodologies have been developed and proposed. In the Propositions 3.1-3.4, a centralized CAFDI methodology is proposed while it is assumed that there exists a set of secure communication channels to transmit information from the C&C side filter of the CPS (15) to its plant side UIO-based detector (18). Moreover, the designed CAFDI module in the Propositions 3.1-3.4 only requires transmission of information from the C&C side to the plant side of the CPS. In order to eliminate our assumption on the number of secure communication channels, a modified version of our centralized CAFDI module is developed in the Propositions 3.5 and 3.6 where one does not*

need to have any secure communication channel on the two sides. However, the proposed methodology in the Propositions 3.5 and 3.6 requires one to transmit information both from the C&C side of the CPS to its plant side and from the plant side to its C&C side. Finally, in the Propositions 3.7-3.10, our proposed CAFDI methodologies in previous sections have been extended to detect and isolate cyber-attacks and faults in large-scale interconnected CPS through a distributed architecture. In the Propositions 3.7-3.10, we assume that there exist one secure communication channel from the C&C side filter (43) to the plant side filter (42) and one secure communication channel from the plant side filter (42) to the C&C side filter (43). The proposed CAFDI methodology in the Propositions 3.7-3.10 requires one to transmit information from the C&C side of the CPS to its plant side and from the plant side to its C&C side as well as the transmission of information among the nearby UIO-based detectors.

3.4 Case Studies

In this section, two case studies are provided to demonstrate and verify the capabilities and advantages of our proposed methodologies as compared to the available results in the literature. In the first case study, the effectiveness of the proposed CAFDI methodology in Propositions 3.1-3.4 for simultaneous detection and isolation of cyber-attacks and faults in a Quadruple-Tank Process (QTP) are illustrated. The continues-time linear QTP model given in [104] is used. Moreover, to simulate the covert and zero dynamics attacks the models in [2] and [1] are used, respectively. It is worth noting that the considered QTP is a positive system (see [105, 106] for more details on positive systems).

In the second case study, a hardware-in-the-loop (HIL) simulation is provided to demonstrate and verify the capabilities of our proposed methodologies. A four area power network system is simulated by utilizing the OPAL-RT real-time simulator and 4 Raspberry Pis.

3.4.1 Quadruple-Tank Process

In our first case study, two types of cyber-attacks are considered, namely covert attacks and zero dynamics attacks. Moreover, detection and isolation of simultaneous loss of effectiveness fault in the actuator and sensor bias fault with cyber-attacks are also demonstrated and validated. The linearized state-space

representation of the QTP (system (7)) with a non-minimum phase zero is given by [104],

$$A^{\mathrm{s}} = \begin{bmatrix} -0.015821 & 0 & 0.025633 & 0 \\ 0 & -0.010941 & 0 & 0.017822 \\ 0 & 0 & -0.025633 & 0 \\ 0 & 0 & 0 & -0.017822 \end{bmatrix},$$

$$B^{\mathrm{s}} = \begin{bmatrix} 0.048221 & 0 \\ 0 & 0.034956 \\ 0 & 0.07755 \\ 0.055931 & 0 \end{bmatrix}, C^{\mathrm{s}} = \begin{bmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \end{bmatrix},$$

$$D_{\mathrm{cp}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, A^{\mathrm{a}} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix}, \tag{51}$$

$C^{\mathrm{a}} = [0\,1\,1; 1\,1\,1]$, $B^{\mathrm{s}}_{\mathrm{a}} = B^{\mathrm{s}}$, $L_1 = [0.048221\,0\,0\,0.055931]^\top$, $L^{\mathrm{a}}_2 = [1\,0\,0]^\top$, $N^{\mathrm{s}} = [1\,1\,1\,1]^\top$, $D_{\mathrm{a}} = 0.5 \times I_2$, $N^{\mathrm{a}} = [0\,1\,1]^\top$, where all the input and output channels are compromised by adversaries as they have access to two out of the four communication channels. Hence, the third and the fourth communication channels among the C&C side filter (15) and the UIO-based detector (18) are secured, i.e., $q = 2$. The covariance matrices of $\omega^{\mathrm{s}}(t)$ and $\omega^{\mathrm{a}}(t)$ are specified as $Q = \mathrm{diag}(0.1,\ 0.1,\ 0.1,\ 0.1)$ and $R^{\mathrm{a}} = \mathrm{diag}(0.2,\ 0.2)$, respectively.

A bank of plant side filters as given by (16), C&C side filters as presented by (15), and detectors as provided in (18) are designed such that the conditions of Propositions 3.1-3.4 are satisfied. Moreover, the residual signals $res_{\mathrm{AA}}(t)$, $res_{\mathrm{SA}}(t)$, $res_{\mathrm{AF}}(t)$, and $res_{\mathrm{SF}}(t)$ are generated according to Propositions 3.1-3.4, respectively.

To determine the threshold for the residual signals $res_{\mathrm{AA}}(t)$ and $res_{\mathrm{SA}}(t)$ of the actuator and sensor cyber-attacks 100 Monte Carlo simulation runs are conducted according to Remark 3.3, and the threshold is determined as $\eta = 0.1$. Moreover, the threshold for residuals that are used to detect actuator and sensor faults is computed according to the method provided in Remark 3.3 and is set to $\eta = 0.4$.

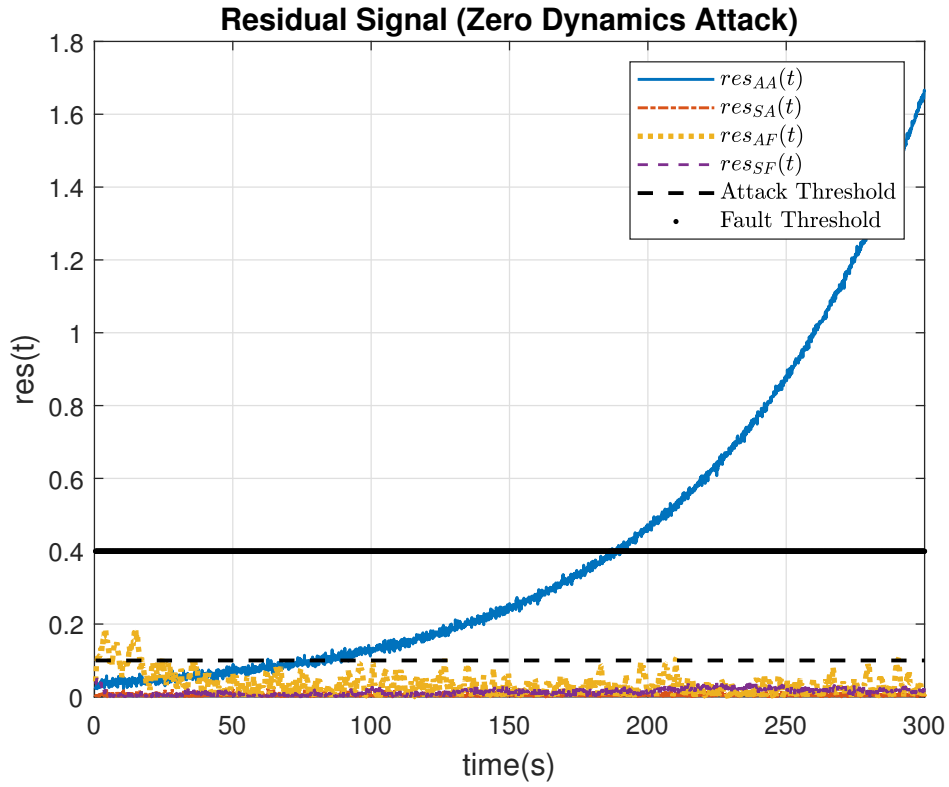

Figure 3.4: Detection of a zero dynamics attack that is injected at $t = 0$ (s).

Scenario 1 (Zero Dynamics Attacks): The system presented in (51) has a non-minimum phase zero at $s = 0.00127798$, that is associated with the zero state direction $x_0^{\mathrm{s}} = [0,\ 0,\ -0.63604564,\ 0.61796063]^\top$ and the zero input direction $u_0 = [-0.33810175,\ 0.31505206]^\top$.

As can be seen in Fig. 3.4, the residual signal $res_{\mathrm{AA}}(t) = y_{\mathrm{p}}(t) - C\hat{x}^{\mathrm{AA}}(t)$ that is designed to detect actuator cyber-attacks has increased (due to a zero dynamics attack) while the other residuals are successfully below the threshold.

Scenario 2 (Covert Attacks): In this scenario, a covert attack scenario is considered. The adversary is capable of completely removing the impact of actuator cyber-attack $a_{\mathrm{u}}(t) = [-2,\ -1]^\top$ which starts at $t = 10$ (s) from the sensor measurements by using the sensor cyber-attack $D_{\mathrm{a}}a_{\mathrm{y}}(t) = -Cx_{\mathrm{cov}}(t)$, where $\dot{x}_{\mathrm{cov}}(t) = Ax_{\mathrm{cov}}(t) + B_{\mathrm{a}}a_{\mathrm{u}}(t)$. As shown in Fig. 3.5, the increase in actuator and sensor cyber-attacks residuals, $res_{\mathrm{AA}}(t)$ and $res_{\mathrm{SA}}(t)$, respectively, that exceed the threshold indicate the occurrence of these cyber-attacks.

Scenario 3 (Faults): In this scenario, 50 percent loss of effectiveness in the first actuator, has occurred at

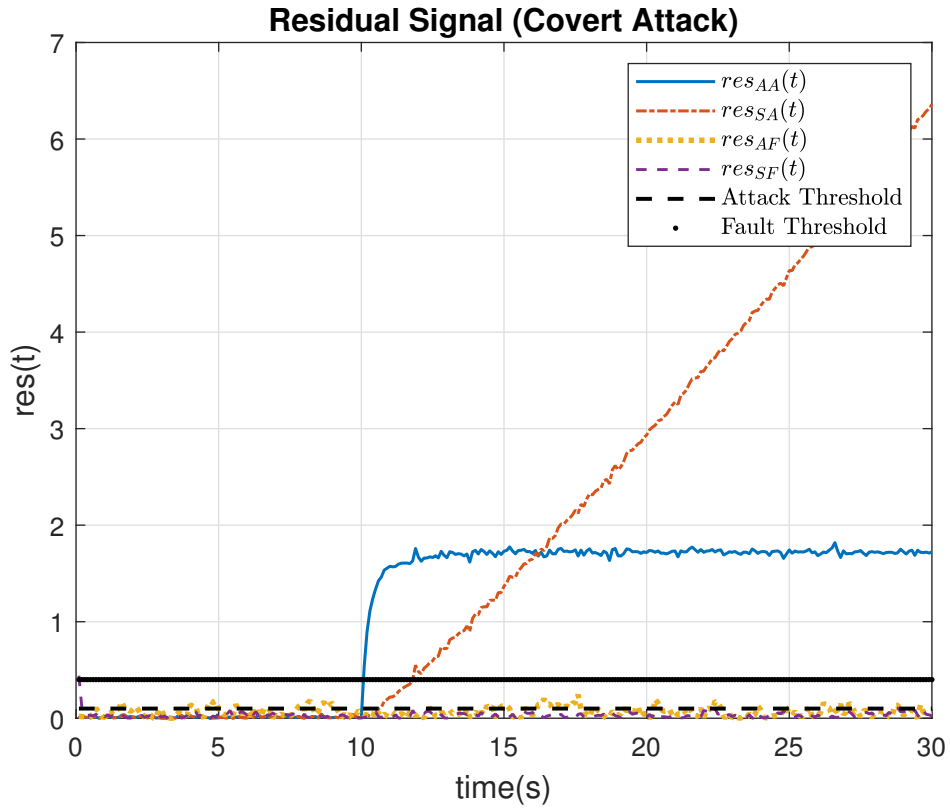

Figure 3.5: Detection of actuator and sensor cyber-attacks in case of covert attacks.

$t = 5$ (s). Moreover, bias fault in the second sensor which is modeled as pseudo actuator fault, $f_2(t) = 1$, also exists in the system from $t = 15$ (s) onward. It can be observed from Fig. 3.6 that due to occurrence of faults the corresponding residuals have been increased.

Scenario 4 (Simultaneous Injection of Cyber-Attacks and Faults): In this scenario, the detection and isolation of simultaneous cyber-attacks and faults is demonstrated. In this scenario, the system is under a covert attack at $t = 0$ (s) and actuator and sensor faults occur at $t = 5$ (s) and $t = 15$ (s), respectively. As depicted in Fig. 3.7, these anomalies can both be detected and isolated successfully.

3.4.2 Four Area Power Network

As depicted in Figure 3.8, the power system and all the plant side dynamics are simulated on the OPAL-RT simulator and the C&C side controllers and filters are deployed and simulated on the Raspberry Pis. Our HIL simulation setup is shown in Figure 3.9. Similar to the distributed wide area monitoring systems

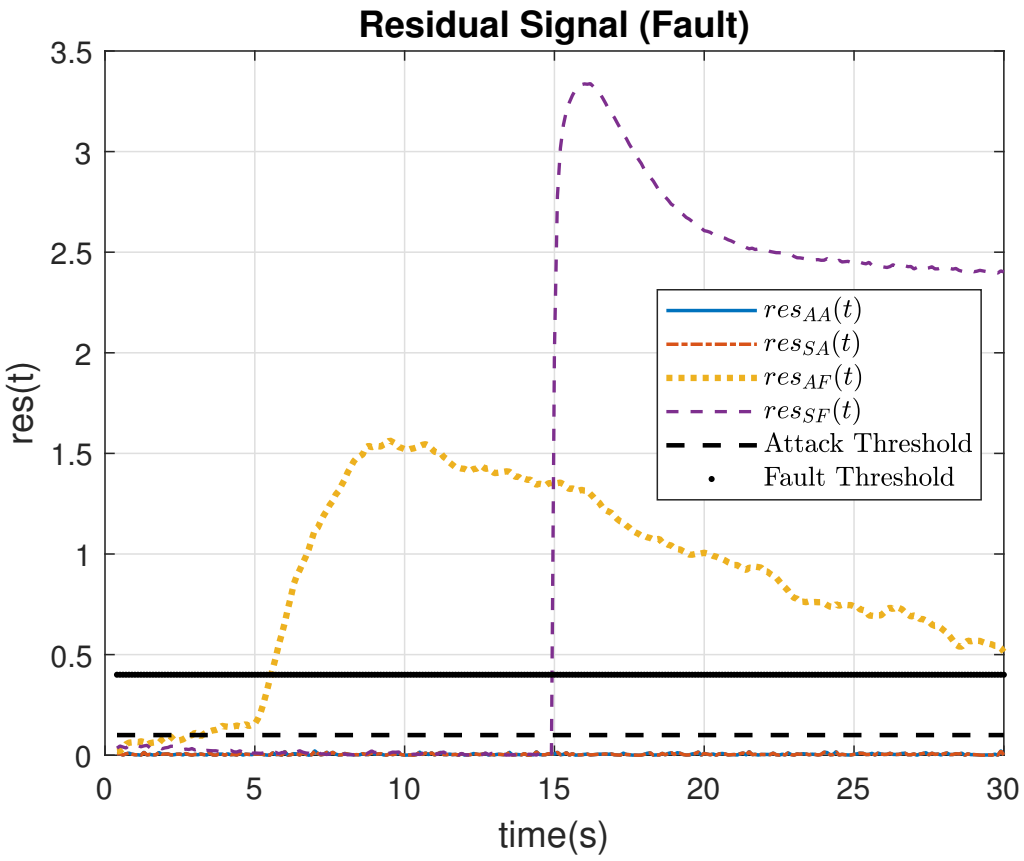

Figure 3.6: Detection of actuator and sensor faults.

(WAMS), in Figure 3.8, we have utilized Phasor Data Concentrator (PDC) units which acquire, archive, exchange, and process data within each area [107, 108]. Furthermore, in [109] a methodology has been proposed which can be used to represent the IEEE New England 39-bus power system as a 4 area network.

The governing dynamics of the power system in the i-th area is given by [110]:

$$\begin{aligned}
\dot{\delta}_i^{\mathrm{v}}(t) =& f_i^{\mathrm{b}}(t), \\
T_{\mathrm{p}i}\dot{f}_i^{\mathrm{b}}(t) =& -(f_i^{\mathrm{b}}(t) - f^{\mathrm{nom}}) + K_{\mathrm{p}i}(P_{\mathrm{t}i}(t) - P_{\mathrm{d}i} \\
&+ \sum_{j\in\mathcal{N}_i} V_i(t)V_j(t)B_{ij}\sin(\delta_i^{\mathrm{v}}(t) - \delta_j^{\mathrm{v}}(t))), \\
T_{\mathrm{V}i}\dot{V}_i(t) =& \bar{E}_{\mathrm{f}i} - (1 - (X_{\mathrm{d}i} - X'_{\mathrm{d}i})B_{ii})V_i(t) - (X_{\mathrm{d}i} - X'_{\mathrm{d}i}) \\
&\times \sum_{j\in\mathcal{N}_i} V_j(t)B_{ij}\cos(\delta_i^{\mathrm{v}}(t) - \delta_j^{\mathrm{v}}(t))),
\end{aligned} \tag{52}$$

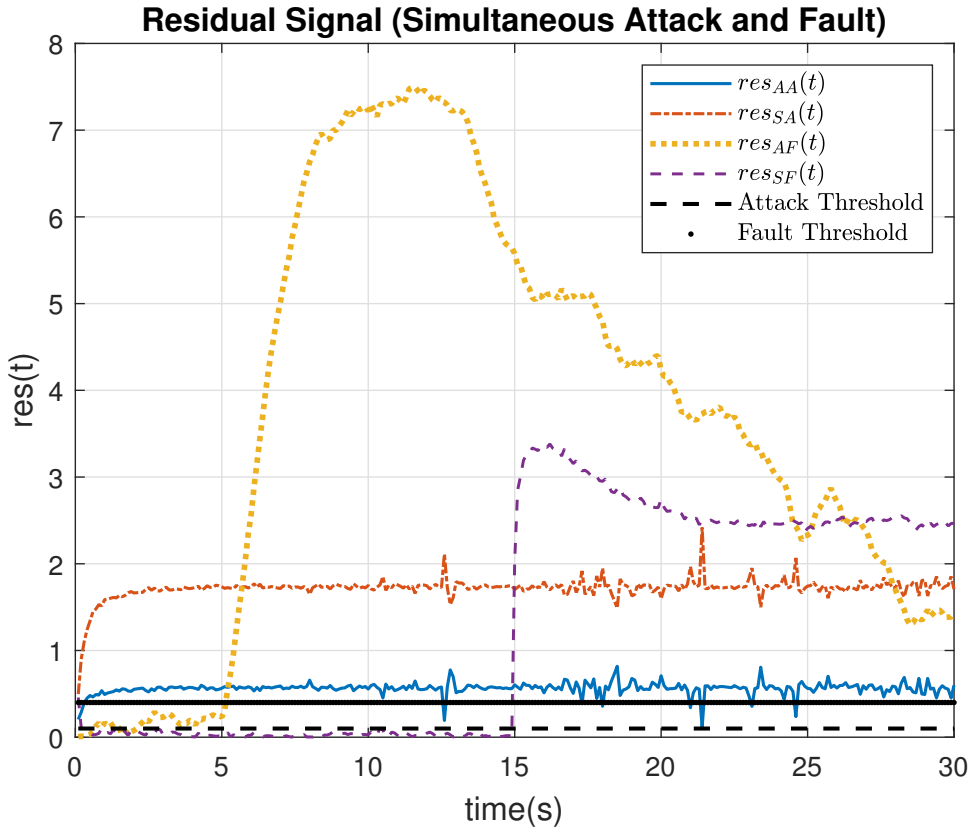

Figure 3.7: Detection and isolation of different simultaneous cyber-attacks and faults.

for $i = 1, \ldots, 4$. Moreover, the dynamics of the turbine and the governor can be expressed by [110]:

$$\begin{aligned} T_{ti}\dot{P}_{ti}(t) &= -P_{ti}(t) + P_{gi}(t), \\ T_{gi}\dot{P}_{gi}(t) &= -\frac{1}{R_i}(f_i^{\mathrm{b}}(t) - f^{\mathrm{nom}}) - P_{gi}(t) + u_i, \end{aligned} \tag{53}$$

where the definition of parameters and their values in (52) and (53) are provided in Tables 3.1 and 3.2, respectively. Also, $\mathcal{N}_i$ is the set of neighborhoods of the subsystem $\mathcal{S}_i$, for $i = 1, \ldots, 4$. In this case study, we have $\mathcal{N}_1 = \{2, 4\}$, $\mathcal{N}_2 = \{1, 3\}$, $\mathcal{N}_3 = \{2, 4\}$, and $\mathcal{N}_4 = \{1, 3\}$.

It should be noted that the nonlinear dynamics in (52) is used in the HIL simulation. However, in order to design our proposed CAFDI methodology for the interconnected large-scale CPS in the Propositions 3.7-3.10, we have linearized the power system dynamics using Simulink "Model Linearizer" app. Consequently, by utilizing the linearized model, we design and implement a bank of two side filters and UIO-based detectors to detect and isolate cyber-attacks and faults for the four area power network system in the HIL

Table 3.1: List of symbols used in (52) and (53).

Symbol	Description
$\delta_i^{\mathrm{v}}(t)$	Voltage Angle
$f_i^{\mathrm{b}}(t)$	Frequency
$V_i(t)$	Voltage
$P_{\mathrm{t}i}(t)$	Turbine Output Power
$P_{\mathrm{g}i}(t)$	Governor Output Power
f^{nom}	Nominal Frequency
$T_{\mathrm{p}i}$	Time Constant of the Generator
$T_{\mathrm{t}i}$	Time Constant of the Turbine
$T_{\mathrm{g}i}$	Time Constant of the Governor
$T_{\mathrm{V}i}$	Direct Axis Transient Open-Circuit Constant
$K_{\mathrm{p}i}$	Governor Gain
R_i	Speed Regulation Coefficient
$X_{\mathrm{d}i}$	Direct Synchronous Reactance
$X'_{\mathrm{d}i}$	Direct Synchronous Transient Reactance
B_{ij}	Transmission Line Susceptance
u_i	Control Input to the Governor
$\bar{E}_{\mathrm{f}i}$	Constant Exciter Voltage
$P_{\mathrm{d}i}$	Unknown Power Demand

Table 3.2: Values of parameters used in (52) and (53).

Symbol	Area 1	Area 2	Area 3	Area 4
$T_{\mathrm{p}i}$ (s)	21	25	23	22
$T_{\mathrm{t}i}$ (s)	0.30	0.33	0.35	0.28
$T_{\mathrm{g}i}$ (s)	0.080	0.072	0.070	0.081
$T_{\mathrm{V}i}$ (s)	5.54	7.41	6.11	6.22
$K_{\mathrm{p}i}$ $(s^{-1}\,\mathrm{p.u.}^{-1})$	120	112.5	115	118.5
R_i $(s^{-1}\,\mathrm{p.u.}^{-1})$	2.5	2.7	2.6	2.8
$X_{\mathrm{d}i}$ (p.u.)	1.85	1.84	1.86	1.83
$X'_{\mathrm{d}i}$ (p.u.)	0.25	0.24	0.26	0.23
B_{ii} (p.u.)	-13.6	-12.9	-12.3	-12.3
$\bar{E}_{\mathrm{f}i}$ (p.u.)	1	1	1	1
$P_{\mathrm{d}i}$ (p.u.)	0.01	0.015	0.012	0.014

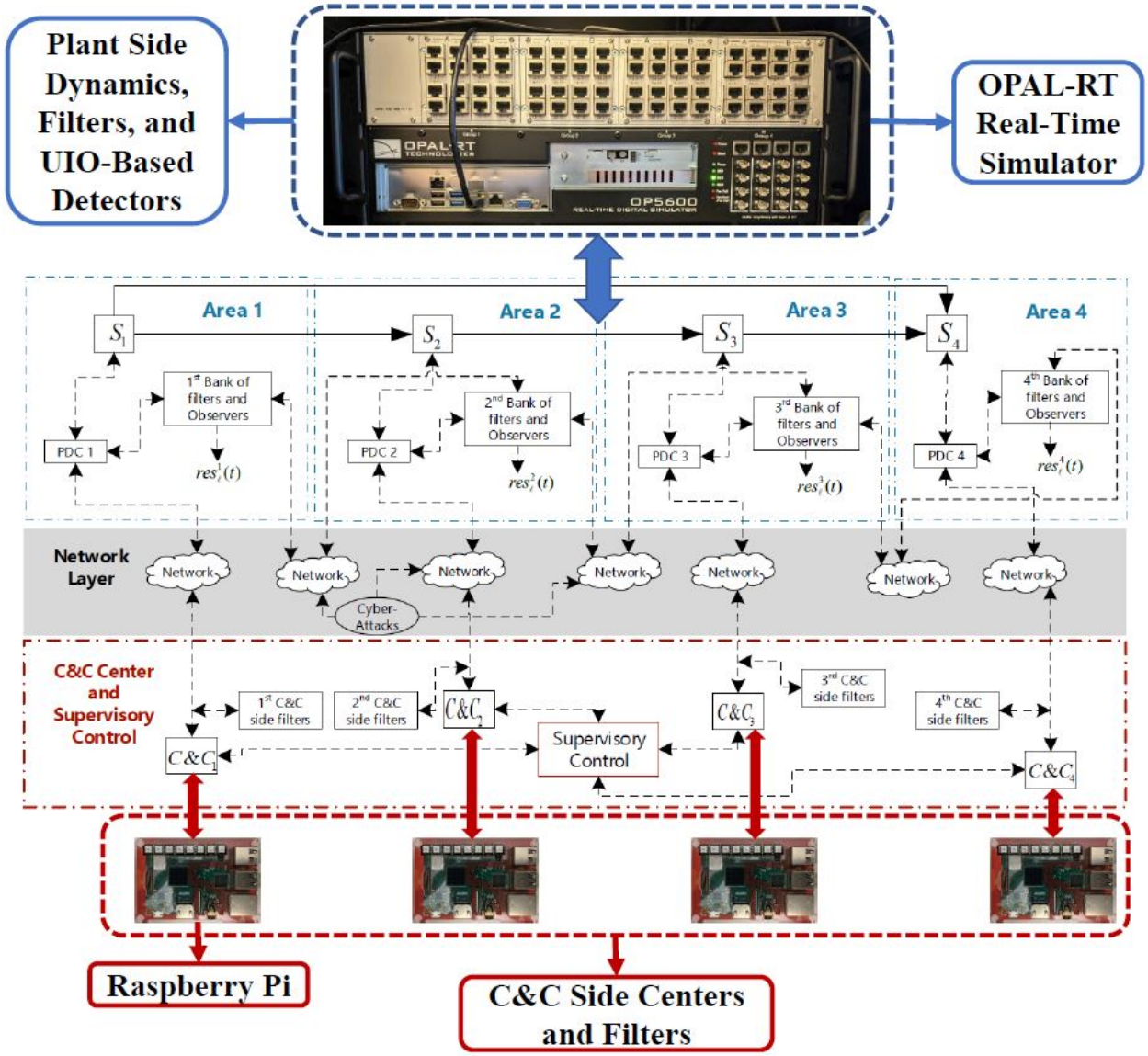

Figure 3.8: The architecture of the HIL simulation for our proposed distributed CAFDI in the four area power network system. Black dashed and solid lines denote communication of data and physical couplings among subsystems, respectively.

simulation platform.

HIL Simulation Results for the Four Area Power Network

In this case study, we develop a CAFDI methodology for a four area power network system under cyber-attacks and faults. Each subsystem is connected to its neighboring subsystems through tie-lines with

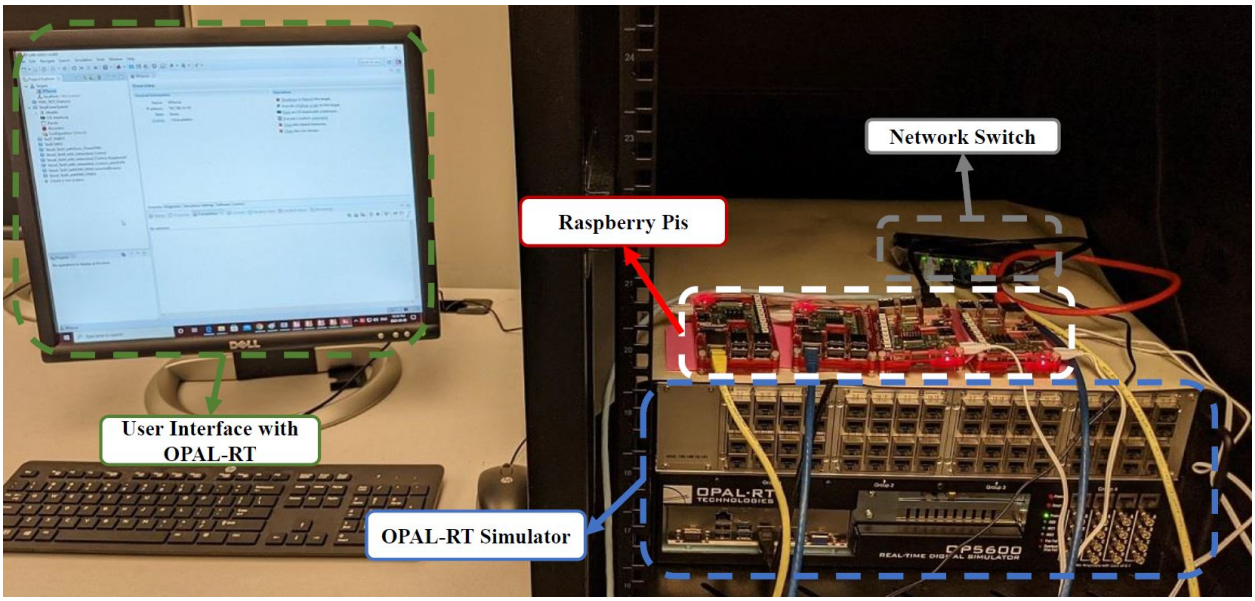

Figure 3.9: The implemented HIL simulation platform.

$B_{12} = -5.4$ p.u., $B_{23} = -5$ p.u., $B_{34} = -4.5$ p.u., and $B_{14} = -5.2$ p.u., with the base power of 1000 MW.

In this case study, we consider that the actuator of each subsystem as well as the first and the third sensors of all the subsystems are under faulty conditions. Furthermore, all the input and output channels of all the subsystems are compromised by adversaries. As depicted in Figure 3.8, each subsystem is equipped with a bank of plant side filters given by (42), C&C side filters in (43), and UIO-based detectors provided in (45) that are designed according to the Propositions 3.7-3.10. As per the Assumption 3.8, we assume that there exists one secured communication channel from (43) to (42) and one secured communication channel from (42) to (43).

Scenario 1 (Covert Attacks): In this scenario, a covert attack on the subsystem $\mathcal{S}_1$ is considered. The covert attack starts at at $t = 50$ (s) and ends at $t = 300$ (s). As shown in Figure 3.10 the impact of both the actuator and sensor attacks on $\mathcal{S}_1$ can be seen in the generated residual of $\mathcal{S}_1$ and nearby subsystems which are $\mathcal{S}_2$ and $\mathcal{S}_4$. Hence, the covert cyber-attack on the subsystem $\mathcal{S}_1$ is detected by this subsystem and its neighboring subsystems.

Scenario 2 (Faults): Actuator and sensor faults are injected to the subsystem $\mathcal{S}_1$. The subsystem $\mathcal{S}_1$ is under an actuator fault from $t = 60$ (s) to $t = 300$ (s), and a sensor fault starting from $t = 150$ (s) to $t = 350$ (s).

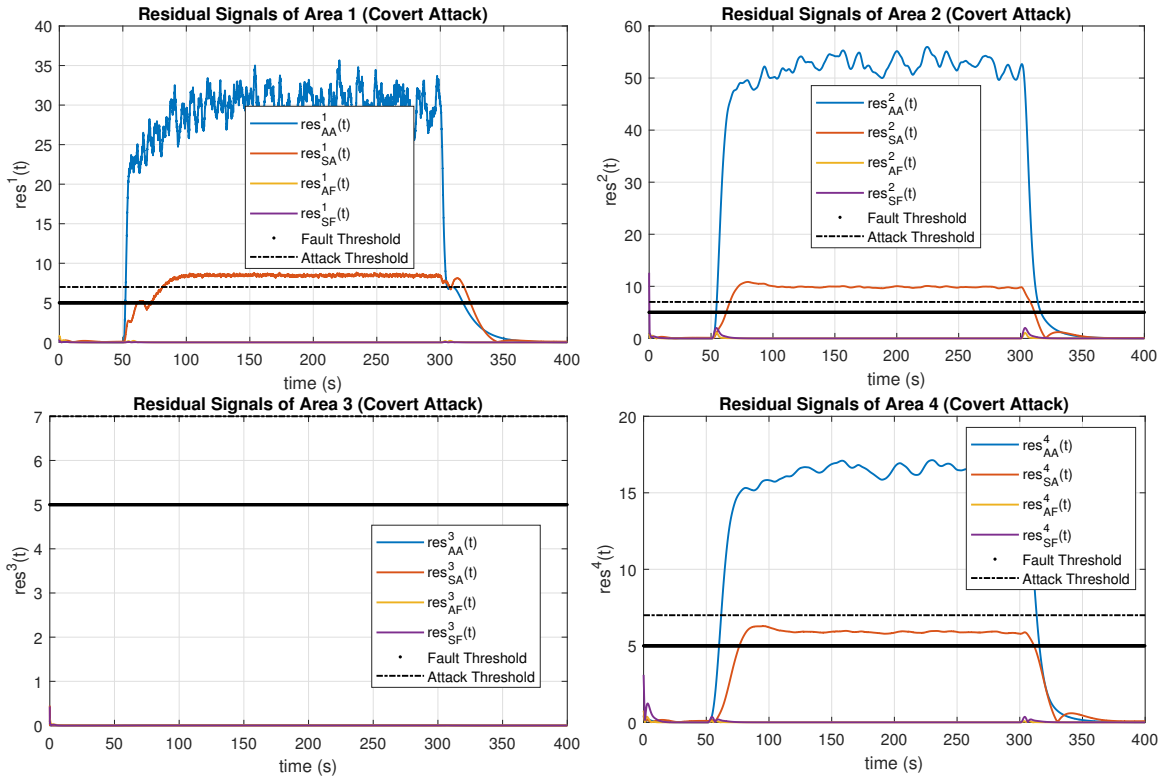

Figure 3.10: Detection of covert cyber-attack on the subsystem $\mathcal{S}_1$.

As depicted in Figure 3.11, $\mathcal{S}_1$ has been able to detect its local actuator fault, but the neighboring subsystems, i.e., $\mathcal{S}_2$ and $\mathcal{S}_4$, have not detected the actuator fault. This is due to the linearization that was made for (52) which has resulted in having $(I - H_1^{\mathrm{AF}} C_1) A_{1j} = 0$, for $j = 2$ and 4 (see Remark 3.12). However, the sensor fault is successfully detected and isolated in the subsystems $\mathcal{S}_1$, $\mathcal{S}_3$, and $\mathcal{S}_4$.

Scenario 3 (Simultaneous Injection of Cyber-Attacks and Faults): In this scenario, actuator and sensor cyber-attacks as well as actuator and sensor faults are simultaneously injected to the subsystem $\mathcal{S}_1$. The subsystem $\mathcal{S}_1$ is under an actuator attack starting from $t = 50$ (s) to $t = 220$ (s), a sensor attack starting from $t = 100$ (s) to $t = 250$ (s), an actuator fault from $t = 150$ (s) to $t = 300$ (s), and a sensor fault starting from $t = 200$ (s) to $t = 350$ (s). As depicted in Figure 3.12, all the anomalies are successfully detected and isolated.

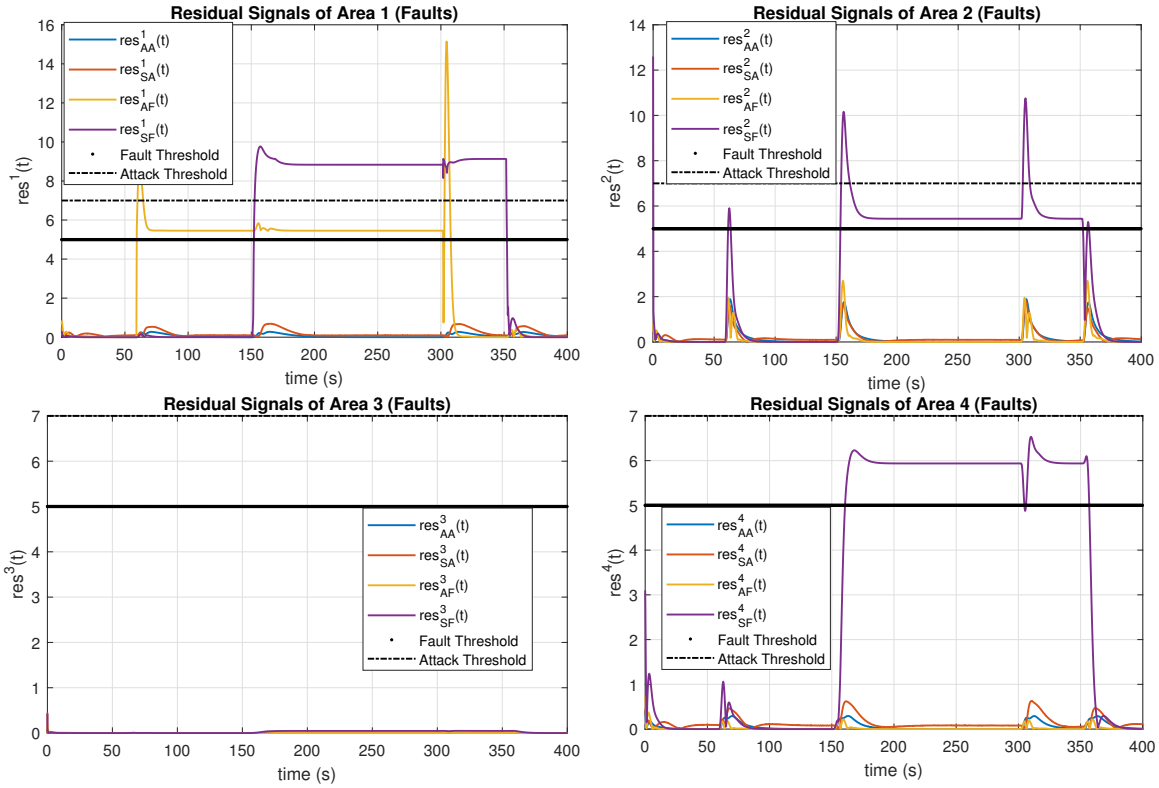

Figure 3.11: Detection and isolation of faults in the subsystem $\mathcal{S}_1$ and nearby subsystems.

Performance Evaluation

A confusion matrix analysis [111] is employed to evaluate the performance of our proposed CAFDI methodology for the 4 area power network. Given a classifier and its corresponding instances, four possible outcomes are obtained as (1) TP (True Positive), if the instance is positive and is truly classified as positive, (2) FN (False Negative), if the instance is positive and incorrectly classified as negative, (3) TN (True Negative), if the instance is negative and correctly classified as negative, and (4) FP (False Positive), if the instance is negative and incorrectly classified as positive [111]. Based on the possible outcomes, true positive rate (TPR) and false positive rate (FPR) can be used as performance metrics and measures, where $TPR = TP/(TP+FN)$ and $FPR = FP/(FP+TN)$. In particular, the Receiver Operating Characteristic (ROC) curve which shows the TPR versus FPR for various threshold levels is also utilized in this subsection.

The ROC curves for $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_4$ are depicted in Figure 3.13. Moreover, as illustrated in the previous subsection, cyber-attacks and faults on $\mathcal{S}_1$ cannot be detected in $\mathcal{S}_3$, therefore, the ROC curve for the 3-rd

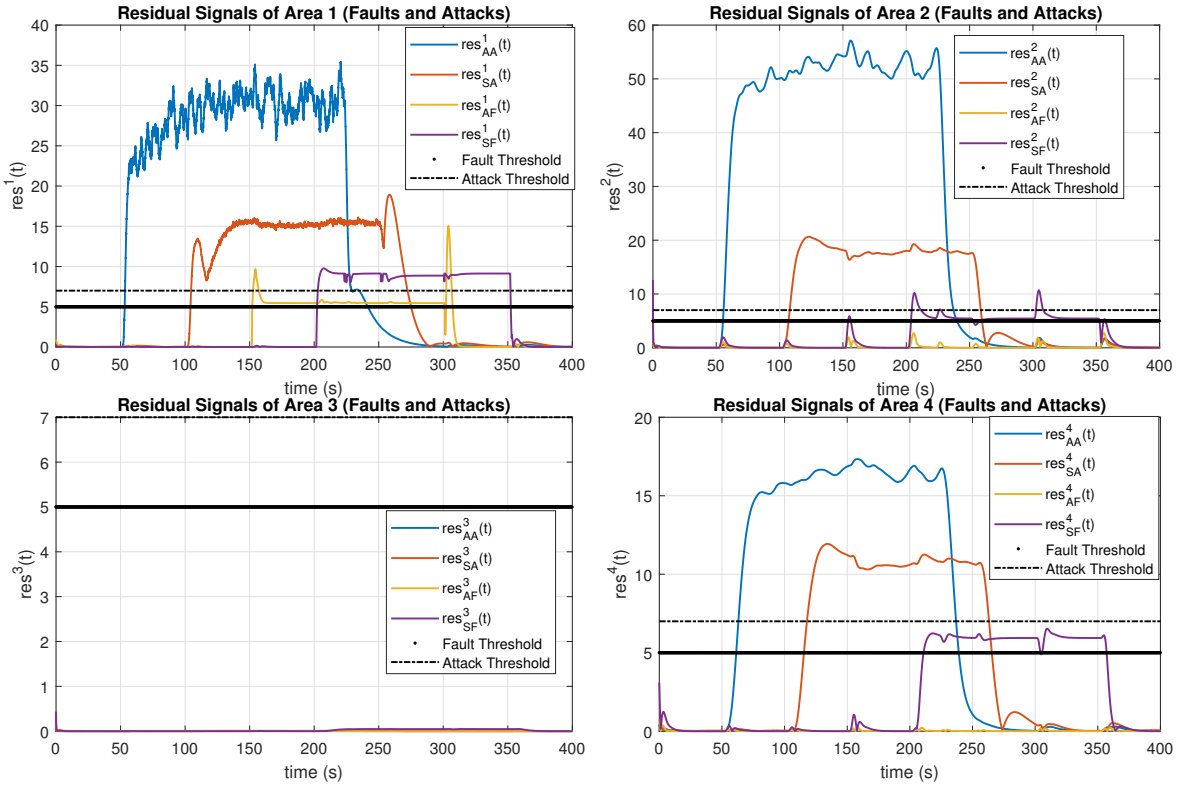

Figure 3.12: Detection and isolation of simultaneous cyber-attacks and faults in the subsystem $\mathcal{S}_1$ and nearby subsystems.

subsystem is not provided. It can be seen in Figure 3.13 that for the case of actuators cyber-attacks, sensors cyber-attacks, and sensor faults $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_4$ have high TPR. Moreover, the neighboring subsystems $\mathcal{S}_2$ and $\mathcal{S}_4$ have lower ROC for actuator faults, that was also observed in Figures 3.11 and 3.12.

3.5 Conclusion

In this chapter, the problem of simultaneous detection and isolation of machine induced faults and intelligent malicious adversarial cyber-attacks has been studied. Centralized and distributed methodologies based on the cyber-physical systems (CPS) two side filters and UIO-based detectors have been proposed and developed. In both methodologies, a bank of filters along with UIO-based detectors are designed on the plant side and a bank of filters was implemented on the C&C side of the CPS. In case of the proposed distributed CAFDI methodology, the UIO-based detector of each subsystem communicates information with

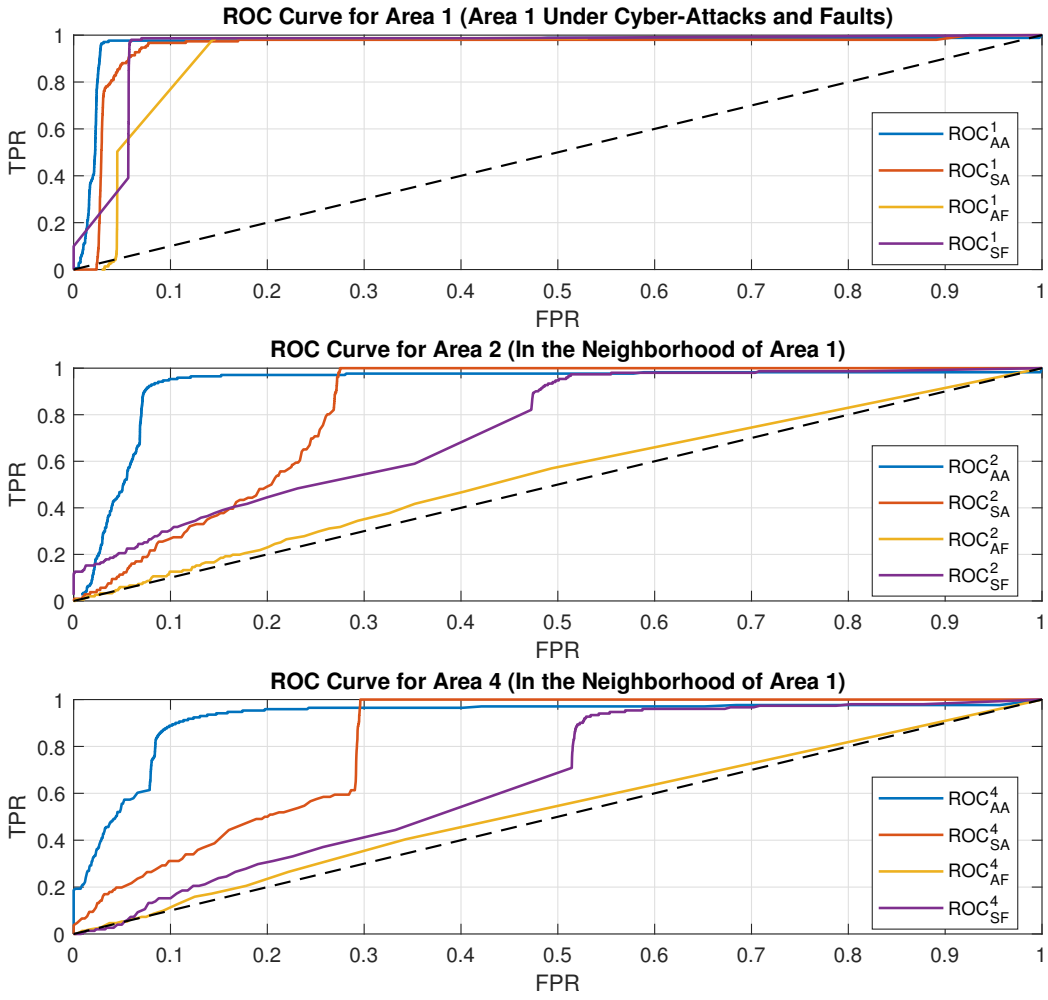

Figure 3.13: ROC curves where the subsystem $\mathcal{S}_1$ is under cyber-attacks and faults.

UIO-based detectors in the nearby subsystems. Hence, under certain conditions, each subsystem can detect and isolate its cyber-attacks and faults as well as anomalies in its nearby subsystems. Using the proposed centralized and distributed strategies, one is capable of *simultaneously* detecting machine induced actuator and sensor faults as well as undetectable cyber-attacks, such as covert and zero dynamics attacks, and detectable cyber-attacks, such as false data injection attacks.

Chapter 4

Dynamic Coding Schemes as Active Countermeasures for Cyber-Attacks in Cyber-Physical Systems

In this chapter, we study stealthy cyber-attacks in cyber-physical systems (CPS), namely zero dynamics attacks, covert attacks, and controllable attacks. In particular, under certain assumptions, we investigate and propose conditions under which one can execute zero dynamics and controllable attacks in the CPS. The above conditions are derived based on the Markov parameters of the CPS and elements of the system observability matrix. Consequently, in addition to outlining the number of required actuators to be attacked, these conditions provide one with the minimum system knowledge needed to perform zero dynamics and controllable cyber-attacks. As a countermeasure against the above stealthy cyber-attacks, we develop a dynamic coding scheme that increases the minimum number of the CPS required actuators to carry out zero dynamics and controllable cyber-attacks to its maximum possible value. It is shown that if at least one secure input channel exists, the proposed dynamic coding scheme can prevent adversaries from executing the zero dynamics and controllable attacks even if they have complete knowledge of the coding system. Moreover, we address three main problems in the context of covert cyber-attacks in the CPS. Firstly, we aim to investigate and develop necessary and sufficient conditions in terms of disruption resources of the CPS that enable adversaries to execute covert cyber-attacks. These conditions can be utilized to identify the input

and output communication channels that are needed by adversaries to execute covert attacks. Secondly, we utilize the derived conditions to define an upper bound on the security index (SI) for covert attacks in the CPS. The upper bound determines the minimum number of actuators and sensors that are needed to be attacked to execute a covert cyber-attack. Lastly, this chapter introduces and develops a dynamic coding scheme as a countermeasure against covert attacks. Under certain conditions and assuming the existence of one secure input and two secure output communication channels, the proposed dynamic coding scheme prevents adversaries from executing covert cyber-attacks. Finally, three illustrative numerical case studies are provided to demonstrate the effectiveness and capabilities of our derived conditions and proposed methodologies.

To summarize, the main contributions of this chapter are as follows:

(1) Under certain assumptions, conditions under which one can carry out the zero dynamics and controllable attacks are obtained. These conditions are derived in terms of the Markov parameters of the CPS, elements of the observability matrix, and characteristic matrices of the system. Therefore, these conditions outline both the required disruption resources, i.e., the required actuators to be attacked, and the level of system knowledge that adversaries need to execute the zero dynamics and controllable cyber-attacks.

(2) By utilizing the proposed conditions for existence of zero dynamics and controllable attacks, their implementation methodologies are then provided. As for the case of zero dynamics attacks, the implementation solely relies on the Markov parameters of the CPS and elements of the observability matrix.

(3) A dynamic coding scheme is then developed and proposed that under certain conditions can increase the number of actuators that are needed to execute the zero dynamics and controllable cyber-attacks to its maximum possible value. Therefore, the proposed dynamic coding scheme can increase the actuators security index for the CPS.

(4) Necessary and sufficient conditions under which covert cyber-attacks can be performed in the CPS are derived. The developed conditions can be used to determine which disruption resources in terms of input and output communication channels of the CPS should be compromised to carry out covert

attacks.

(5) An upper bound on the SI for covert attacks is defined which relies on the developed necessary and sufficient conditions on the existence of covert attacks. Moreover, we provide and algorithm that can be used to compute the upper bound on SI for covert attacks.

(6) As an active countermeasure against covert attacks, we develop and propose a dynamic coding scheme. The proposed coding scheme includes an encoder on the C&C side and a decoder on the plant side of the CPS. Under certain conditions, if there exists one secure input and two secure output communication channels, adversaries will not be capable of performing covert cyber-attacks in the CPS.

The remainder of the chapter is organized as follows. State-space representation of the CPS system along with the definitions for certain cyber-attacks are provided in Section 4.1. In Section 4.2, the I/O representation of the CPS and the definition of ϵ-stealthy cyber-attacks are studied. Moreover, definitions of the zero dynamics and controllable cyber-attacks along with the conditions for their existence are investigated in Section 4.2. In Sections 4.2.3, conditions for execution of covert attacks are studied. Moreover, the SI for covert attacks is investigated in Section 4.3. Dynamic coding schemes against the zero dynamics attacks, controllable attacks, and covert cyber-attacks are proposed in Sections 4.4 and 4.5. Finally, three numerical case studies are presented in Section 4.6 to illustrate and demonstrate the effectiveness and capabilities of our proposed methodologies.

4.1 Problem Statement and Formulation

Let us consider the following discrete-time linear time-invariant CPS:

$$\begin{aligned} x(k+1) =& Ax(k) + Bu(k) + \omega(k), \\ y(k) =& Cx(k) + \nu(k), \end{aligned} \tag{54}$$

where $x(k) \in \mathbb{R}^n$ denotes the state, $u(k) \in \mathbb{R}^m$ denotes the control input, and $y(k) \in \mathbb{R}^p$ denotes the sensor measurement. Moreover, $\omega(k) \in \mathbb{R}^n$ and $\nu(k) \in \mathbb{R}^p$ are process and measurement noise that are represented by zero mean Gaussian distributions, respectively. The system characteristic matrices A, B, and C are of

appropriate dimensions.

In the presence of actuator and sensor cyber-attacks, the CPS (54) can be expressed as:

$$\begin{aligned} x(k+1) =& Ax(k) + B(u(k) + L_{\mathrm{a}} a_{\mathrm{u}}(k)) + \omega(k), \\ y(k) =& Cx(k) + D_{\mathrm{a}} a_{\mathrm{y}}(k) + \nu(k), \end{aligned} \tag{55}$$

where $a_{\mathrm{u}}(k) \in \mathbb{R}^{m_{\mathrm{a}}}$ represents the actuator attack signal, and $a_{\mathrm{y}}(k) \in \mathbb{R}^{p_{\mathrm{a}}}$ denotes the sensor attack signal. Additionally, the signatures of actuator and sensor cyber-attack signals are captured by $B_{\mathrm{a}} = BL_{\mathrm{a}}$ and D_{a}, respectively. Furthermore, in our examination, we omit the consideration of noise effects on the CPS and assume $\omega(k) = 0$ and $\nu(k) = 0$ for all $k \geq 0$. Nevertheless, to demonstrate the robustness of our proposed methodologies against noise, the numerical case study in Section 4.6.3 is conducted in the presence of both process and sensor noise.

Consider $\mathcal{I}_{\mathrm{a}} = \{u1, \ldots, um\mathrm{a}\}$ and $\mathcal{S}_{\mathrm{a}} = \{s1, \ldots, sp\mathrm{a}\}$ as the sets of compromised input and output communication channels with $|\mathcal{I}\mathrm{a}| = m\mathrm{a}$ and $|\mathcal{S}\mathrm{a}| = p\mathrm{a}$, respectively, where $|\cdot|$ denotes the cardinality of a set. The matrices L_{a} and D_{a} are defined with entries corresponding to the compromised elements in the communication channels. We have $L_{\mathrm{a}} = [l_{\mathrm{u}1} \cdots l_{um_{\mathrm{a}}}]$, where $l_i \in \mathbb{R}^m$ is a vector with all its entries equal to zero except for the i-th element that is equal to one, for $i = u1, \ldots, um_{\mathrm{a}}$. Moreover, $D_{\mathrm{a}} = [d_{\mathrm{s}1} \cdots d_{sp_{\mathrm{a}}}]$, where all entries of $d_q \in \mathbb{R}^p$ are zero except for the q-th element, for $q = s1, \ldots, sp_{\mathrm{a}}$.

Let $y_{\mathrm{o}}(x(0), u(k), a_{\mathrm{u}}(k), a_{\mathrm{y}}(k))$ denote the output of (55) as a function of the initial state $x(0)$, the control input $u(k)$, and attack signals $[a_{\mathrm{u}}(k), a_{\mathrm{y}}(k)]$. In the following, by utilizing the notion of $y_{\mathrm{o}}(x(0), u(k), a_{\mathrm{u}}(k), a_{\mathrm{y}}(k))$, we define the "zero dynamics attacks" (studied in [24] and [18]), "covert attacks" [2,96], and "controllable attacks" (studied in [14–16,31]).

Definition 4.1. *By considering $y_o(x(0), u(k), a_u(k), a_{\mathrm{y}}(k))$ the following cyber-attacks can be defined:*

(1) Let $x(0) = x_0 \neq 0$ and $a_{\mathrm{y}}(k) = 0$, $\forall k \geq 0$. The actuator attack signal $a_u(k) \neq 0$ is defined as a zero dynamics cyber-attack if $y_o(x_0, 0, a_u(k), 0) = 0$, for every $k \geq 0$.

(2) The set of attack signals $a_u(k) \neq 0$ and $a_{\mathrm{y}}(k) \neq 0$ is a covert attack if $y_o(0, 0, a_u(k), a_{\mathrm{y}}(k)) = 0$, $\forall t \geq 0$. In this type of cyber-attacks, adversaries need to have access to both input and output communication channels.

(3) Let $a_y(k) = 0, \forall k \geq 0$. The cyber-attack signal $a_u(k) \neq 0$ is designated as a controllable cyber-attack if $y_o(0, 0, a_u(k), 0) = 0$, for every $k \geq 0$.

It should be emphasized that the cyber-attack 3 in Definition 4.1 is referred to as the "zero stealthy attack" in [14], "zero state induced attack" in [31], and "controllable attack" in [15] and [16]. However, we have adopted the convention from [15] and [16] since the above cyber-attack is related to a certain controllable subspace of the system (see [16] for more details).

Definition 4.2 ([98, 112]). *The CPS (55) is left invertible with respect to the cyber-attack signal $a_u(k)$ if for all $a_u^1(k)$, $a_u^2(k) \in \mathbb{R}^{m_a}$, having $y_o(0, 0, a_u^1(k), 0) = y_o(0, 0, a_u^2(k), 0)$ implies $a_u^1(k) = a_u^2(k)$, for every $k \geq 0$.*

Remark 4.1. *By considering the linearity of the CPS, it follows from Definition 4.2 that the CPS is left invertible if and only if $y_o(0, 0, a_u(k), 0) = 0$ implies that $a_u(k) = 0$. Hence, controllable cyber-attacks in Definition 4.1 can be executed if and only if the CPS (55) is not left invertible in the sense of Definition 4.2 (see [14–16] for more details).*

Definition 4.3 (Invariant Zeros). *$\lambda \in \mathbb{C}$ is an invariant zero of the triple (C, A, B_a) if and only if there exist a nonzero state-zero direction $x_0 \in \mathbb{R}^n$ and an input-zero direction $g \in \mathbb{R}^{m_a}$ such that the following holds [97, Definition 2.1]:*

$$\begin{bmatrix} \lambda I - A & -B_a \\ C & 0 \end{bmatrix} \begin{bmatrix} x_0 \\ g \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{56}$$

Consequently, given an invariant zero λ, a state-zero direction $x_0 \neq 0$, and a nonzero input-zero direction g, $a_u(k) = g\lambda^k \neq 0$, which is designated as the zero dynamics cyber-attack signal in Definition 4.1, satisfies $y_o(x_0, 0, a_u(k), 0) = 0$, for every $k \geq 0$ [97, Lemma 2.7].

Below, we have adopted and modified the definitions in [2, 15–18, 21, 24] in order to provide a formal and unified definition for the *stealthy* and *perfectly undetectable* cyber-attacks on actuators of the CPS.

Definition 4.4 (ϵ-Stealthy Cyber-Attacks). *Let $x(0) = x_0 \in \mathbb{R}^n$. A cyber-attack that is performed by utilizing $a_u(k)$ and $a_y(k)$ on the CPS (55) is ϵ-stealthy if $\|y_o(x_0, u(k), 0, 0) - y_o(x_0, u(k), a_u(k), a_y(k))\|_\infty \leq \epsilon$, $\forall\, k \in \mathbb{N}$, where ϵ is a positive scalar. Moreover, an actuator attack is designated as perfectly undetectable if it is 0-stealthy, i.e., $y_o(x_0, u(k), 0, 0) = y_o(x_0, u(k), a_u(k), a_y(k))$.*

It should be noted that in Definition 4.4, we have adopted the notion of "perfectly undetectable attacks" from [21]. The main reason for choosing this designation is that the impact of perfectly undetectable attacks cannot be seen in the output measurements.

Due to the linearity of the CPS (55) and according to Definition 4.4, a cyber-attack is perfectly undetectable if and only if $y_{\mathrm{o}}(0, 0, a_{\mathrm{u}}(k), a_{\mathrm{y}}(k)) = 0, \forall\, k \geq 0$ [21]. Consequently, covert attacks and controllable attacks are perfectly undetectable cyber-attacks. Moreover, zero dynamics attacks are considered as ϵ-stealthy cyber-attacks.

4.1.1 Objectives

We have six objectives in this chapter. Our *first* objective is to study and propose conditions under which adversaries are capable of performing zero dynamics and controllable attacks in the sense of Definition 4.1. These conditions can be utilized to investigate vulnerability of the CPS to zero dynamics and controllable attacks that are designated as ϵ-stealthy and perfectly undetectable cyber-attacks in Definition 4.4, respectively. The *second* objective is to investigate the problem of implementation of zero dynamics and controllable cyber-attacks. In particular, we aim to utilize the above conditions for the existence of zero dynamics and controllable attacks and propose methodologies for designing actuator attack signals that lead to zero dynamics and controllable attacks. As for the *third* objective, we consider a countermeasure and develop a dynamic coding scheme that can be utilized to increase the security index of the CPS (55) to its maximum possible value, i.e., m actuators. Consequently, if the proposed dynamic coding scheme is used, adversaries will need to compromise all the input communication channels of the CPS to perform the zero dynamics and controllable cyber-attacks. Our *fourth* objective is to investigate and study necessary and sufficient conditions in terms of disruption resources under which adversaries are capable of performing covert cyber-attacks. These conditions determine input and output communication channels that should be attacked by adversaries to execute covert attacks in the CPS. As for our *fifth* objective, we utilize the derived necessary and sufficient conditions for performing covert attacks to find an upper bound on the security index (SI) for covert attacks in the CPS. By utilizing the upper bound on the SI for covert attacks, one can find the number of actuators and sensors that should be attacked to execute a covert cyber-attack. Our *sixth* objective is to develop and propose a dynamic coding scheme as a countermeasure against covert attacks that could be

utilized by the CPS operators. Hence, in presence of the proposed dynamic coding scheme, if the CPS operators secure 1 input and 2 output communication channels, adversaries will not be capable of performing covert cyber-attacks in the CPS.

4.2 Input/Output Model of the CPS and Stealthy Cyber-Attacks

In this section, we present the Input/Output (I/O) model for the CPS (55) within a specified time window. Subsequently, we define ϵ-stealthy cyber-attacks in terms of the I/O model.

The I/O model of the CPS (55) over the time window $\{0, 1, \ldots, N-1\}$ can be represented as:

$$Y(N) = \mathcal{O}_{\mathrm{N}} x(0) + \mathcal{C}_{\mathrm{N}} U(N) + \mathcal{C}_{\mathrm{a}} U_{\mathrm{a}}(N) + \mathcal{D}_{\mathrm{a}} Y_{\mathrm{a}}(N), \tag{57}$$

where $Y(N) = [y(0)^\top, y(1)^\top, \ldots, y(N-1)^\top]^\top$ represents the output of the I/O model. Other vectors include $U(N)$ for inputs, $U_{\mathrm{a}}(N)$ for actuator attack signals, and $Y_{\mathrm{a}}(N)$ for sensor attack signals. Moreover, $\mathcal{D}_{\mathrm{a}} = I_N \otimes D_{\mathrm{a}}$.

The matrices $\mathcal{O}_{\mathrm{N}}$, $\mathcal{C}_{\mathrm{N}}$, and $\mathcal{C}_{\mathrm{a}}$ are structured as shown in (59). Specifically,

$$\mathcal{O}_{\mathrm{N}} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{N-1} \end{bmatrix}, \mathcal{C}_{\mathrm{N}} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ CB & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{N-2}B & CA^{N-3}B & \cdots & 0 \end{bmatrix}, \tag{58}$$

$$\mathcal{C}_{\mathrm{a}} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ CB_{\mathrm{a}} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{N-2}B_{\mathrm{a}} & CA^{N-3}B_{\mathrm{a}} & \cdots & 0 \end{bmatrix}. \tag{59}$$

Let $\mathcal{Y}(x(0), U(N), \check{U}_{\mathrm{a}}(N))$ denote the output of the I/O model in (57) over the time window $\{0, 1, \ldots, N-1\}$. This function depends on the initial state $x(0)$, the vector of control inputs $U(N)$, and the vector of attack signals denoted by $\check{U}_{\mathrm{a}}(N) = [U_{\mathrm{a}}(N)^\top, Y_{\mathrm{a}}(N)^\top]^\top$. In the following, the stealthy and perfectly undetectable actuator cyber-attacks in terms of the I/O model (57) are defined.

Definition 4.5 (I/O ϵ-Stealthy Cyber-Attacks)**.** *A cyber-attack $\check{U}_a(N) \neq 0$ in the I/O CPS model (57) is ϵ-stealthy if $\|\mathcal{Y}(x_0, U(N), 0) - \mathcal{Y}(x_0, U(N), \check{U}_a(N))\|_\infty \leq \epsilon$ holds, $\forall\, N \geq 1$, where $x_0 \in \mathbb{R}^n$. Moreover, $\check{U}_a(N)$ is perfectly undetectable if it is 0-stealthy, i.e., $\mathcal{Y}(x_0, U(N), 0) = \mathcal{Y}(x_0, U(N), \check{U}_a(N))$.*

It should be emphasized that since $Y(N)$ is defined as the concatenated vector of output measurements $y(k)$, any type of ϵ-stealthy actuator attack in the CPS (55) will result in the same level of stealthiness in the I/O model of the CPS (57), and vice versa.

Lemma 4.1. *A cyber-attack is perfectly undetectable in the sense of Definition 4.5 if and only if $\mathcal{C}_a U_a(N) + \mathcal{D}_a Y_a(N) = 0$ holds, $\forall\, N \geq 1$.*

Proof. It follows readily that

$$\mathcal{Y}(x_0, U(N), \check{U}_{\rm a}(N)) = \mathcal{Y}(x_0, U(N), 0) + \mathcal{Y}(0, 0, \check{U}_{\rm a}(N)).$$

Hence, a cyber-attack is perfectly undetectable in the sense of Definition 4.5 if and only if $\mathcal{Y}(0, 0, \check{U}_{\rm a}(N)) = 0$, $\forall\, N \geq 1$, which is equivalent to having $\mathcal{C}_{\rm a} U_{\rm a}(N) + \mathcal{D}_{\rm a} Y_{\rm a}(N) = 0$. This completes the proof of the lemma. □

In the following, we utilize the I/O model (57) to study the zero dynamics cyber-attacks, covert attacks, and controllable cyber-attacks. Therefore, we consider the following assumption throughout this chapter.

Assumption 4.1. *Adversaries do not know the characteristic matrices of the CPS (54), i.e., the triple (A, B, C). However, they know components of $\mathcal{O}_N$ and elements of $\mathcal{C}_N$ that are given by (58).*

4.2.1 Zero Dynamics Cyber-Attacks

Adversaries need to compromise input communication channels of the CPS in order to perform the zero dynamics attacks. Moreover, according to Definitions 4.1 and 4.3, the zero dynamics attack signals are designed such that for a certain $x(0) = x_0 \neq 0$ one has $y_o(x_0, 0, a_{\rm u}(k), 0) = 0$, where $a_{\rm u}(k) \neq 0$. Similar to the latter description of the zero dynamics attacks which is based on the state-space representation of the CPS in (55), we define the zero dynamics attacks by utilizing the I/O model of the CPS as given by (57).

Definition 4.6 (Zero Dynamics Cyber-Attacks in I/O Models)**.** *Let* $a_u(k) = g\lambda^k \neq 0$, $\forall k \geq 0$, *with a nonzero* $g \in \mathbb{R}^{m_a}$ *and* $\lambda \in \mathbb{C}$ *(cf. Definition 4.3). The actuator cyber-attack signal* $a_u(k)$ *in the I/O model of the CPS (57) is defined as a zero dynamics attack if for a certain* $x(0) = x_0 \neq 0$ *one has* $\mathcal{Y}(x_0, 0, [U_a(N)^\top, 0]^\top) = 0$, $\forall N \geq 1$, *i.e.,* $\mathcal{O}_N x_0 + \mathcal{C}_a U_a(N) = 0$.

In order to further investigate the zero dynamics attacks, we need to define controlled invariant subspaces and the relative degree of the CPS (55).

Definition 4.7. *Let* $\mathcal{I} = \{1, \ldots, m\}$ *denote the set of all input channels of the CPS (55). The relative degree of the CPS (55) with respect to the q-th input channel is given by* r_q *if* $CA^iB_q = 0$, *for all* $i < r_q - 1$ *and* $CA^{r_q-1}B_q \neq 0$, *for every* $q \in \mathcal{I}$, *where* B_q *is the q-th column of* B. *If for any positive integer* i *one has* $CA^iB_q = 0$, *the relative degree with respect the q-th input channel cannot be defined. Moreover, let* $\mathcal{I}_a = \{a_1, \ldots, a_{m_a}\}$ *denote the set of attacked input communication channels. Hence, one has* $r_a = min\{r_{a_1}, \ldots, r_{a_{m_a}}\}$.

It should be noted that in [19] and [31], it has been shown that $\mathcal{O}_{\mathrm{N}} x_0 + \mathcal{C}_{\mathrm{a}} U_{\mathrm{a}}(N) = 0$ is equivalent to having a zero dynamics cyber-attack. In the following theorem, the existence of zero dynamics cyber-attacks by utilizing the Markov parameters and elements of the observability matrix of the CPS is studied.

Theorem 4.1. *Let* $x(0) = x_0 \in \bigcap_{i=0}^{r_a-1} ker(CA^i)$ *and assume that* $ker(CA^{r_a-1}) \bigcap ker(CA^{r_a}) = 0$. *A zero dynamics cyber-attack in the sense of Definitions 4.3 and 4.6 that is associated with* x_0 *can be performed if and only if there exists* $g \in \mathbb{R}^{m_a}$ *such that* $CA^{r_a}x_0 + CA^{r_a-1}B_a g = 0$ *and* $Im(CA^{r_a+1}x_0 + CA^{r_a}B_a g) \subseteq Im(CA^{r_a-1}B_a g)$.

Proof. According to Definition 4.6, in case of zero dynamics attacks, $a_{\mathrm{u}}(k)$ is designed such that every row of $\mathcal{O}_{\mathrm{N}} x_0 + \mathcal{C}_{\mathrm{a}} U_{\mathrm{a}}(N)$ is equal to zero. The latter is equivalent to having

$$CA^{r_{\mathrm{a}}}x_0 + CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}}a_{\mathrm{u}}(0) = 0, \tag{60a}$$

$$CA^{r_{\mathrm{a}}+1}x_0 + CA^{r_{\mathrm{a}}}B_{\mathrm{a}}a_{\mathrm{u}}(0) + CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}}a_{\mathrm{u}}(1) = 0, \tag{60b}$$

$$\vdots$$

$$CA^{N-1}x_0 + CA^{N-2}B_{\mathrm{a}}a_{\mathrm{u}}(0) + CA^{N-3}B_{\mathrm{a}}a_{\mathrm{u}}(1)$$
$$+ CA^{N-4}B_{\mathrm{a}}a_{\mathrm{u}}(2) + \cdots + CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}}a_{\mathrm{u}}(N-2) = 0. \tag{60c}$$

Next, we show that having $CA^{r_a}x_0+CA^{r_a-1}B_ag=0$ and $\text{Im}(CA^{r_a+1}x_0+CA^{r_a}B_ag)\subseteq\text{Im}(CA^{r_a-1}B_ag)$ are necessary and sufficient conditions for (60) to hold for every $N\geq r_a$.

Necessary Condition: Suppose for any given $g\in\mathbb{R}^{m_a}$ one has $CA^{r_a}x_0+CA^{r_a-1}B_ag\neq 0$ and $\text{Im}(CA^{r_a+1}x_0+CA^{r_a}B_ag)\nsubseteq\text{Im}(CA^{r_a-1}B_ag)$.

Moreover, assume that $a_u(k)$ is a zero dynamics cyber-attack. Consequently, there does not exist any $a_u(0)=g$ such that $CA^{r_a}x_0+CA^{r_a-1}B_aa_u(0)=0$ holds. Furthermore, there does not exist any $a_u(1)=g\lambda$ that can satisfy (60b). Hence, (60) does not hold, which contradicts our assumption.

Sufficient Condition: Assume there exists g such that

$$CA^{r_a}x_0+CA^{r_a-1}B_ag=0. \tag{61}$$

Moreover, let $\text{Im}(CA^{r_a+1}x_0+CA^{r_a}B_ag)\subseteq\text{Im}(CA^{r_a-1}B_ag)$. Hence, there exists a nonzero $\lambda\in\mathbb{R}$ such that the following holds:

$$CA^{r_a+1}x_0+CA^{r_a}B_ag+CA^{r_a-1}B_ag\lambda=0. \tag{62}$$

Now, we show that under the above assumptions, there exists a zero dynamics attack signal, i.e., $a_u(k)=g\lambda^k$, that is a solution to (60), for $k\in\{0,1,\ldots,N-2\}$ and every $N\geq r_a$.

Let us rewrite (61) as $CA^{r_a-1}(Ax_0+B_ag)=0$. Therefore, g satisfies

$$Ax_0+B_ag=\hat{\alpha}_0x_0+\hat{\alpha}_1x_1, \tag{63}$$

where $x_0\neq x_1$ and $x_1\in\ker(CA^{r_a-1})$ such that x_0 is a basis of the null space of CA^{r_a-1} and x_1 denotes a linear combination of its other bases, that implies $\text{Im}(x_0)\neq\text{Im}(x_1)$. Also, $\hat{\alpha}_0\in\mathbb{R}$ and $\hat{\alpha}_1\in\mathbb{R}$ are scalars.

One can substitute $CA^{r_a-1}B_ag=-CA^{r_a}x_0$ into the left-hand side of (62) such that $CA^{r_a+1}x_0+CA^{r_a}B_ag-CA^{r_a}x_0\lambda=0$. Hence, considering (63), one has

$$CA^{r_a}(Ax_0+B_ag-\lambda x_0)=CA^{r_a}(\hat{\alpha}_0x_0+\hat{\alpha}_1x_1-\lambda x_0)=0. \tag{64}$$

Since $\text{Im}(x_0)\neq\text{Im}(x_1)$ and it is assumed that $\ker(CA^{r_a-1})\bigcap\ker(CA^{r_a})=0$, therefore (64) is satisfied if and only if $\hat{\alpha}_1=0$ and $\hat{\alpha}_0=\lambda$.

Hence, in (63), one has $\hat{\alpha}_0 = \lambda$ and $\hat{\alpha}_1 = 0$. Since $x_0 \in \ker(C)$ and $(\lambda I - A)x_0 - B_{\rm a} g = 0$, according to the Definition 4.3, λ is an invariant zero of the triple $(C, A, B_{\rm a})$. Moreover, one has $g = (B_{\rm a})^{\dagger}(\lambda I - A)x_0$, which is also consistent with the definition of the input-zero direction in [97, Chapter 2].

We have $Ax_0 + B_{\rm a} g = \lambda x_0$. Moreover, it can be shown that $A^2 x_0 + AB_{\rm a} g + B_{\rm a} g\lambda = A(Ax_0 + B_{\rm a} g) + B_{\rm a} g\lambda = \lambda^2 x_0$. Consequently, one can derive $A^i x_0 + \sum_{j=1}^{i} A^{i-j} B_{\rm a} g \lambda^{j-1} = \lambda^i x_0$, for $i \geq 1$. Moreover, it can be easily shown that the i-th equation in (60) is equal to $CA^{r_{\rm a}-1}(A^i x_0 + \sum_{j=1}^{i} A^{i-j} B_{\rm a} g \lambda^{j-1}) = CA^{r_{\rm a}-1}(\lambda^i x_0) = 0$, for every $i \geq 1$. This completes the proof of the theorem. □

Corollary 4.1. *Let hypotheses of Theorem 4.1 hold. The input-zero direction g and the invariant zero λ of the CPS (55) that are associated with $x_0 \in \bigcap_{i=0}^{r_a - 1} ker(CA^i)$ can be expressed in the following form:*

(1) $g = -(CA^{r_a - 1} B_a)^{\dagger} CA^{r_a} x_0$;

(2) $\lambda = -(CA^{r_a - 1} B_a g)^{\dagger}(CA^{r_a} B_a g + CA^{r_a + 1} x_0)$.

Proof. Given that the hypotheses of Theorem 4.1 hold, g and λ that satisfy (61) and (62) are the input-zero direction and the invariant zero of the triple $(C, A, B_{\rm a})$, respectively, as per Definition 4.3.

It follows from (60a) that the CPS has an input-zero direction $g = -(CA^{r_{\rm a}-1} B_{\rm a})^{\dagger} CA^{r_{\rm a}} x_0$. Moreover, since $CA^{r_{\rm a}-1} B_{\rm a} g$ is a nonzero vector, it has a unique left pseudoinverse. Consequently, from (60b) one obtains $\lambda = -(CA^{r_{\rm a}-1} B_{\rm a} g)^{\dagger}(CA^{r_{\rm a}} B_{\rm a} g + CA^{r_{\rm a}+1} x_0)$. This completes the proof of the corollary. □

It should be noted that in Corollary 4.1, if the initial condition of the CPS, i.e., $x(0) = x_0$, satisfies the condition in Theorem 4.1 and $\ker(CA^{r_{\rm a}-1}) \bigcap \ker(CA^{r_{\rm a}}) = 0$, an invariant zero $\lambda = -(CA^{r_{\rm a}-1} B_{\rm a} g)^{\dagger} \times (CA^{r_{\rm a}} B_{\rm a} g + CA^{r_{\rm a}+1} x_0)$ exists. Therefore, in general, the given λ does not depend on the initial condition of the system, but rather on those states that belong to the weakly unobservable or output-nulling subspace of the system (refer to [98] for more information). Also, it can be easily seen that since the vector $CA^{r_{\rm a}-1} B_{\rm a} g$ has a unique left pseudoinverse, both states x_0 and $\hat{x}_0 = \alpha_1 x_0$, where α_1 is a nonzero scalar, result in the same invariant zero λ.

As stated in Theorem 4.1, under certain assumptions, if and only if $CA^{r_{\rm a}} x_0 + CA^{r_{\rm a}-1} B_{\rm a} g = 0$ and $\mathrm{Im}(CA^{r_{\rm a}+1} x_0 + CA^{r_{\rm a}} B_{\rm a} g) \subseteq \mathrm{Im}(CA^{r_{\rm a}-1} B_{\rm a} g)$, adversaries are capable of performing the zero dynamics attacks. In particular, the above necessary and sufficient conditions indicate the existence of all invariant

zeros of the system that for a nonzero input result in a zero output (see Definition 4.3), and not just the existence of the transmission zeros (refer to [113] for a detailed discussion). In the case where the CPS is controllable and observable, the set of invariant zeros only contains transmission zeros of the system [113]. Hence, if the triple (C, A, B_{a}) denotes a controllable and observable CPS and conditions in the Theorem 4.1 are satisfied, adversaries can execute a zero dynamics cyber-attack that excites a transmission zero of the CPS.

Note that in [114, Lemma 4] and [97, Lemma 3.2], it has been shown that having $\mathrm{Im}(CA^{r_{\mathrm{a}}}x_0) \subseteq \mathrm{Im}(CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}})$ is a necessary condition for the existence of invariant zeros, which implies the existence of a g that satisfies $CA^{r_{\mathrm{a}}}x_0 + CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}}g = 0$. However, in Theorem 4.1, it has been shown that under the assumption that $\ker(CA^{r_{\mathrm{a}}-1}) \bigcap \ker(CA^{r_{\mathrm{a}}}) = 0$, having the above condition and $\mathrm{Im}(CA^{r_{\mathrm{a}}+1}x_0 + CA^{r_{\mathrm{a}}}B_{\mathrm{a}}g) \subseteq \mathrm{Im}(CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}}g)$ are necessary and sufficient for the existence of the zero dynamics attacks which are those invariant zeros of the CPS that given a nonzero input, result in a zero output.

Moreover, similar to the proposed g in Corollary 4.1, the input-zero direction has been studied in [114, Propositions 1 and 2] and [97, Chapter 3]. It should be noted that results in the Corollary 4.1 hold under the assumption that $\ker(CA^{r_{\mathrm{a}}-1}) \bigcap \ker(CA^{r_{\mathrm{a}}}) = 0$, but in [114] and [97] the above assumption has not been considered. Moreover, the given formula in Corollary 4.1 to compute the invariant zero λ has not been provided in [114] and [97].

Under certain assumptions, Theorem 4.1 provides the necessary and sufficient conditions under which adversaries are capable of performing the zero dynamics cyber-attacks. Therefore, adversaries may compromise certain input communication channels of the CPS (55) such that the corresponding L_{a} results in having $CA^{r_{\mathrm{a}}}x_0 + CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}}g = 0$ and $\mathrm{Im}(CA^{r_{\mathrm{a}}+1}x_0 + CA^{r_{\mathrm{a}}}B_{\mathrm{a}}g) \subseteq \mathrm{Im}(CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}}g)$. Moreover, in Theorem 4.1, one has $x_0 \in \ker(CA^{r_{\mathrm{a}}-1})$ and $\ker(CA^{r_{\mathrm{a}}-1}) \bigcap \ker(CA^{r_{\mathrm{a}}}) = 0$. Hence, $CA^{r_{\mathrm{a}}}x_0 \neq 0$, which implies that x_0 does not belong to the unobservable subspace of the system.

It should be noted that in order for adversaries to investigate conditions in the Theorem 4.1, they need to know the initial condition of the CPS $x(0)$. Given that as per Assumption 4.1 adversaries have access to $\mathcal{C}_{\mathrm{N}}$ and $\mathcal{O}_{\mathrm{N}}$, once $N \geq n$, if $\mathcal{O}_{\mathrm{N}}$ has a full column rank, i.e., the CPS is observable, adversaries can estimate the initial condition by solving the equation $Y(N) = \mathcal{O}_{\mathrm{N}}x(0) + \mathcal{C}_{\mathrm{N}}U(N)$ for $x(0)$.

Remark 4.2. *Let x_0, g, and λ be the state-zero direction, the input-zero direction, and the invariant zero*

of the CPS (57)*, as per Corollary 4.1, respectively. Moreover, consider* k_a *as the time instant at which the adversary initiates the zero dynamics attack such that* $x(k_a) \neq x_0$*. According to [97, Lemma 2.6], the zero dynamics attack signal* $a_u(k) = g\lambda^k$ *results in having* $y_o(x(k_a), 0, a_u(k-k_a), 0) = CA^{k-k_a}(x(k_a) - x_0)$*, for* $k \geq k_a$*. Consequently, in the case where* A *is a Schur stable matrix, the error due to initiating the zero dynamics attack at the time instant* k_a*, i.e.,* $CA^{k-k_a}(x(k_a) - x_0)$*, will converge to zero. The latter is demonstrated in Section 4.6.1.*

Moreover, by utilizing the Corollary 4.1, adversaries can compute g and λ in terms of the components of $\mathcal{O}_{\mathrm{N}}$, i.e., the observability matrix, and elements of $\mathcal{C}_{\mathrm{N}}$, i.e., the Markov parameters in (58) to design their cyber-attack signals. Also, based on results in the Corollary 4.1, adversaries are capable of discovering whether the zero dynamics of the system is minimum phase, i.e., $|\lambda| < 1$, or it is non-minimum phase, i.e., $|\lambda| > 1$.

4.2.2 Controllable Cyber-Attacks

As discussed in Definition 4.1, under controllable cyber-attacks, one has $y_{\mathrm{o}}(0, 0, a_{\mathrm{u}}(k)) = 0$, where $a_{\mathrm{u}}(k) \neq 0$. In the following, an equivalent definition for controllable cyber-attacks is provided.

Definition 4.8 (Controllable Cyber-Attacks in the I/O Model)**.** *Let* $a_u(k) \neq 0$*,* $\forall k \geq 0$*. The attack signal* $U_a(N)$ *in the I/O model of the CPS* (57) *is designated as a controllable cyber-attack if one has* $\mathcal{Y}(0, 0, [U_a(N)^\top, 0]^\top) = 0$*,* $\forall N \geq 1$*, i.e.,* $\mathcal{C}_a U_a(N) = 0$*.*

Consequently, as per Lemma 4.1, the controllable cyber-attack in Definition 4.8 is perfectly undetectable. In the following theorem, we investigate conditions under which controllable cyber-attacks in Definition 4.8 can be performed on the I/O model of the CPS (57).

Theorem 4.2. *A controllable cyber-attack in the sense of Definition 4.8 can be executed in the CPS if there exists a nonzero* $\hat{a}_0 \in$ *ker*$(CA^{r_a-1}B_a)$*, such that Im*$(AB_a\hat{a}_0) \subseteq$ *Im*(B_a)*.*

Proof. According to Definition 4.8, in the case of controllable cyber-attacks, the actuator cyber-attack signal

$a_{\mathrm{u}}(k)$ should be designed such that $\mathcal{C}_{\mathrm{a}}U_{\mathrm{a}}(N)=0$, which is equivalent to

$$CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}}a_{\mathrm{u}}(0)=0, \tag{65a}$$

$$CA^{r_{\mathrm{a}}}B_{\mathrm{a}}a_{\mathrm{u}}(0)+CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}}a_{\mathrm{u}}(1)=0, \tag{65b}$$

$$CA^{r_{\mathrm{a}}+1}B_{\mathrm{a}}a_{\mathrm{u}}(0)+CA^{r_{\mathrm{a}}}B_{\mathrm{a}}a_{\mathrm{u}}(1)+CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}}a_{\mathrm{u}}(2)=0, \tag{65c}$$

$$\vdots$$

$$CA^{N-2}B_{\mathrm{a}}a_{\mathrm{u}}(0)+CA^{N-3}B_{\mathrm{a}}a_{\mathrm{u}}(1)+CA^{N-4}B_{\mathrm{a}}a_{\mathrm{u}}(2)$$
$$+\cdots+CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}}a_{\mathrm{u}}(N-2)=0. \tag{65d}$$

Suppose that $a_{\mathrm{u}}(0)\in\ker(CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}})$, and there exists $\hat{a}_0\in\ker(CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}})$ such that $\mathrm{Im}(AB_{\mathrm{a}}\hat{a}_0)\subseteq\mathrm{Im}(B_{\mathrm{a}})$. Let $a_{\mathrm{u}}(0)=\hat{a}_0$. Let us rewrite the left-hand side of (65b) as $CA^{r_{\mathrm{a}}-1}(AB_{\mathrm{a}}a_{\mathrm{u}}(0)+B_{\mathrm{a}}a_{\mathrm{u}}(1))$. Consequently, since $\mathrm{Im}(AB_{\mathrm{a}}\hat{a}_0)\subseteq\mathrm{Im}(B_{\mathrm{a}})$, one can design $a_{\mathrm{u}}(1)$ such that

$$AB_{\mathrm{a}}\hat{a}_0+B_{\mathrm{a}}a_{\mathrm{u}}(1)=B_{\mathrm{a}}\hat{a}_0. \tag{66}$$

Hence, by substituting (66) in (65b), one can conclude that there exists the cyber-attack signal $a_{\mathrm{u}}(1)$ that satisfies (65b).

The left-hand side of (65c) can be rewritten as $CA^{r_{\mathrm{a}}-1}(A^2B_{\mathrm{a}}\hat{a}_0+AB_{\mathrm{a}}a_{\mathrm{u}}(1)+B_{\mathrm{a}}a_{\mathrm{u}}(2))$. Considering that $\mathrm{Im}(AB_{\mathrm{a}}\hat{a}_0)\subseteq\mathrm{Im}(B_{\mathrm{a}})$ and $AB_{\mathrm{a}}\hat{a}_0+B_{\mathrm{a}}a_{\mathrm{u}}(1)=B_{\mathrm{a}}\hat{a}_0$, $a_{\mathrm{u}}(2)$ can be designed in the following form

$$A(AB_{\mathrm{a}}\hat{a}_0+B_{\mathrm{a}}a_{\mathrm{u}}(1))+B_{\mathrm{a}}a_{\mathrm{u}}(2)=B_{\mathrm{a}}\hat{a}_0, \tag{67}$$

which satisfies (65c).

Consequently, it can be shown that since $\mathrm{Im}(AB_{\mathrm{a}}\hat{a}_0)\subseteq\mathrm{Im}(B_{\mathrm{a}})$, there exists $a_{\mathrm{u}}(j)$ in the j-th equation of (65) that satisfies

$$AB_{\mathrm{a}}\hat{a}_0+B_{\mathrm{a}}a_{\mathrm{u}}(j)=B_{\mathrm{a}}\hat{a}_0, \tag{68}$$

for $j\geq 2$. Consequently, there exists an $a_{\mathrm{u}}(j)$ that is the solution to the j-th equation in (65). This completes the proof of the theorem. □

In [31], cyber-attacks that satisfy the condition in Definition 4.8, i.e., $\mathcal{C}_{\mathrm{a}}U_{\mathrm{a}}(N) = 0$, are defined as "zero state inducing" attacks. Moreover, necessary and sufficient conditions for the existence of this type of cyber-attack based on weakly unobservable and output-nulling reachable subspaces of the system have been provided in [31, Theorem 3]. However, as opposed to [31], the studied conditions in Theorem 4.2 rely only on A, B_{a}, and the first Markov parameter of the CPS and are easier to be verified and validated. In the following corollary, the implementation of controllable cyber-attacks is studied.

Corollary 4.2. *Assume that the conditions in Theorem 4.2 hold and let $a_u(0) \in ker(CA^{r_a-1}B_a)$. The actuator cyber-attack signal to perform a controllable attack in the sense of Definition 4.8 can be expressed as*

$$a_u(k) = a_u(0)h(k) - B_a^{\dagger}AB_a a_u(0)h(k-1), \tag{69}$$

for $k \geq 1$, where $h(k) \in \mathbb{R}$ such that $h(0) = 1$ and $h(k)$ for $k \geq 1$ can be any arbitrary function.

Proof. Let $h(k) \in \mathbb{R}$ such that $h(0) = 1$. Moreover, as per Theorem 4.2, consider $a_{\mathrm{u}}(0) \in \ker(CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}})$ such that $\mathrm{Im}(AB_{\mathrm{a}}a_{\mathrm{u}}(0)) \subseteq \mathrm{Im}(B_{\mathrm{a}})$. Consequently, $a_{\mathrm{u}}(1)$ can be designed such that

$$AB_{\mathrm{a}}a_{\mathrm{u}}(0)h(0) + B_{\mathrm{a}}a_{\mathrm{u}}(1) = B_{\mathrm{a}}a_{\mathrm{u}}(0)h(1). \tag{70}$$

Given that the left-hand side of (65b) can be rewritten as $CA^{r_{\mathrm{a}}-1}(AB_{\mathrm{a}}a_{\mathrm{u}}(0)h(0) + B_{\mathrm{a}}a_{\mathrm{u}}(1))$, (70) satisfies (65b). Moreover, since B_{a} is an injective map, $a_{\mathrm{u}}(1)$ can be uniquely derived as $a_{\mathrm{u}}(1) = a_{\mathrm{u}}(0)h(1) - B_{\mathrm{a}}^{\dagger}AB_{\mathrm{a}}a_{\mathrm{u}}(0)h(0)$.

Similar to $a_{\mathrm{u}}(1)$, one can design $a_{\mathrm{u}}(2)$ to satisfy

$$AB_{\mathrm{a}}a_{\mathrm{u}}(0)h(1) + B_{\mathrm{a}}a_{\mathrm{u}}(2) = B_{\mathrm{a}}a_{\mathrm{u}}(0)h(2). \tag{71}$$

Also, considering (70), the left-hand side of (65c) can be rewritten as $CA^{r_{\mathrm{a}}-1}(AB_{\mathrm{a}}a_{\mathrm{u}}(0)h(1) + B_{\mathrm{a}}a_{\mathrm{u}}(2))$. Thus, the given $a_{\mathrm{u}}(2)$ in (71) satisfies (65c). Hence, one has $a_{\mathrm{u}}(2) = a_{\mathrm{u}}(0)h(2) - B_{\mathrm{a}}^{\dagger}AB_{\mathrm{a}}a_{\mathrm{u}}(0)h(1)$.

Consequently, $a_{\mathrm{u}}(k)$ can be designed to satisfy

$$AB_{\mathrm{a}}a_{\mathrm{u}}(0)h(k-1) + B_{\mathrm{a}}a_{\mathrm{u}}(k) = B_{\mathrm{a}}a_{\mathrm{u}}(0)h(k), \tag{72}$$

for $k \geq 1$. Moreover, considering (72), the left-hand side of the $(k+1)$-th equation in (65) can be derived as $CA^{r_a-1}(AB_a a_u(0)h(k-1) + B_a a_u(k))$. Given that $a_u(k)$ is designed according to (72), $a_u(k)$ satisfies the $(k+1)$-th equation in (65), for $k \geq 1$. Therefore, from (72) it follows that the controllable cyber-attack signal $a_u(k)$ can be designed according to (69). This completes the proof of the corollary. □

Theorem 4.2 can be used to study existence of controllable cyber-attacks in the CPS. Furthermore, one needs to know the first Markov parameter of the CPS, i.e., $CA^{r_a-1}B_a$, and matrices A and B_a to investigate the proposed conditions in Theorem 4.2. However, as per Assumption 4.1, adversaries do not know the matrices A and B. Hence, in the following corollary, under certain conditions, the existence and implementation of controllable cyber-attacks by utilizing only the Markov parameters of the CPS are studied.

Corollary 4.3. *Let us assume that* $ker(CA^{r_a-1}) \bigcap ker(CA^{r_a}) = 0$. *Adversaries can execute a controllable cyber-attack in the CPS according to the Definition 4.8 if there exist nonzero* $\hat{a}_0 \in ker(CA^{r_a-1}B_a)$ *and* $\hat{a}_1 \in \mathbb{R}^{m_a}$ *that satisfy* $CA^{r_a}B_a\hat{a}_0 + CA^{r_a-1}B_a\hat{a}_1 = 0$ *and* $CA^{r_a+1}B_a\hat{a}_0 + CA^{r_a}B_a\hat{a}_1 = 0$. *Moreover, by considering* $a_u(0) = \hat{a}_0$, *a controllable cyber-attack signal can be expressed as*

$$a_u(k) = \hat{a}_0 h(k) + \hat{a}_1 h(k-1), \tag{73}$$

for $k \geq 1$, *where* $h(k) \in \mathbb{R}$ *such that* $h(0) = 1$ *and* $h(k)$ *for* $k \geq 1$ *can be any arbitrary function.*

Proof. From $CA^{r_a}B_a\hat{a}_0 + CA^{r_a-1}B_a\hat{a}_1 = 0$, it follows that

$$CA^{r_a-1}(AB_a\hat{a}_0 + B_a\hat{a}_1) = 0. \tag{74}$$

Moreover, having $CA^{r_a+1}B_a\hat{a}_0 + CA^{r_a}B_a\hat{a}_1 = 0$ implies that

$$CA^{r_a}(AB_a\hat{a}_0 + B_a\hat{a}_1) = 0. \tag{75}$$

Consequently, since $\ker(CA^{r_a-1}) \bigcap \ker(CA^{r_a}) = 0$, it follows from (74) and (75) that $AB_a\hat{a}_0 + B_a\hat{a}_1 = 0$, which implies that $\mathrm{Im}(AB_a\hat{a}_0) \subseteq \mathrm{Im}(B_a)$. Hence, conditions in the Theorem 4.2 for existence of controllable cyber-attacks are satisfied.

From $AB_{\mathrm{a}}\hat{a}_0 + B_{\mathrm{a}}\hat{a}_1 = 0$, it follows that $\hat{a}_1 = -B_{\mathrm{a}}^{\dagger}AB_{\mathrm{a}}\hat{a}_0$. Hence, as per Corollary 4.2, one can set $a_{\mathrm{u}}(0) = \hat{a}_0$ and design a controllable cyber attack in the following form:

$$a_{\mathrm{u}}(k) = a_{\mathrm{u}}(0)h(k) - B_{\mathrm{a}}^{\dagger}AB_{\mathrm{a}}a_{\mathrm{u}}(0)h(k-1),$$

for $k \geq 1$, where $h(0) = 1$ and $h(k) \in \mathbb{R}$ can be any arbitrary function. This completes the proof of the corollary. □

Remark 4.3. *In order to find $\hat{a}_0$ and $\hat{a}_1$ in the Corollary 4.3, one needs to solve*

$$\begin{bmatrix} CA^{r_a-1}B_a & 0 \\ CA^{r_a}B_a & CA^{r_a-1}B_a \\ CA^{r_a+1}B_a & CA^{r_a}B_a \end{bmatrix} \begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{76}$$

for $\hat{a}_0$ and $\hat{a}_1$. It is worth notifying that if the hypothesis of Corollary 4.3 holds, (76) can be easily solved by using the `mldivide` *MATLAB function.*

The derived conditions in the Theorem 4.1 and Corollary 4.3 for the CPS (55), rely on components of $\mathcal{O}_{\mathrm{N}}$, i.e., the observability matrix, and elements of $\mathcal{C}_{\mathrm{N}}$, i.e., the Markov parameters, given by (58). Hence, by employing methodologies in [115–118], and utilizing the results in Theorem 4.1 and Corollary 4.3, one can study the vulnerability of the CPS to zero dynamics attacks and controllable attacks, respectively, in a data-driven manner.

4.2.3 Covert Attacks

According to [2], in the case of covert attacks, adversaries compromise both input and output communication channels of the CPS and design their attack signals $a_{\mathrm{u}}(k) \neq 0$ and $a_{\mathrm{y}}(k) \neq 0$ such that the impact of actuator attacks cannot be observed in the transmitted sensor measurements to the control side of the CPS. In the following, a definition for covert attacks is given in terms of the I/O model of the CPS (57).

Definition 4.9 (Covert Attacks for the I/O Model)**.** *Let each element of $a_u(k) \in \mathbb{R}^{m_a}$ and $a_y(k) \in \mathbb{R}^{p_a}$ be nonzero for some $k \geq 0$. The attack signal $\breve{U}_a(N) = [U_a(N)^{\top}\ Y_a(N)^{\top}]^{\top}$ in the I/O model of the CPS (57) is designated as a covert attack if one has $\mathcal{Y}(0, 0, \breve{U}_a(N)) = 0$, $\forall N \geq 1$, i.e., $\mathcal{C}_a U_a(N) + \mathcal{D}_a Y_a(N) = 0$.*

In order to further investigate covert cyber-attacks, we need to define a relative degree for the CPS (55).

Definition 4.10. *The relative degree of the q-th output of the CPS (55) with respect to the actuator attack signal $a_u(k)$ is r_a^q if $C_q A^{i_q} B_a = 0$ for all $i_q < r_a^q - 1$ and $C_q A^{r_a^q - 1} B_a \neq 0$, for $q = 1, \ldots, p$, where C_q is the q-th row of C. If for any positive integer i_q one has $C_q A^{i_q} B_a = 0$, the relative degree for the q-th output with respect to $a_u(k)$ cannot be defined. Moreover, one has $r_a = min\{r_a^1, \ldots, r_a^p\}$.*

We are now in a position to provide necessary and sufficient conditions under which covert attacks in the sense of Definition 4.9 can be carried out in the I/O model of the CPS (57).

Theorem 4.3. *Given any actuator attack signal $a_u(k) \in \mathbb{R}^{m_a}$, a covert attack in the sense of Definition 4.9 can be executed in the CPS (57) if and only if relative degrees of all the outputs for triples (C, A, B_a) and $(D_a^* C, A, B_a)$ are equal as per Definition 4.10, where $D_a^* = D_a D_a^\top$, i.e., $C_q A^{r_a^q - 1} B_a = (D_a^* C)_q A^{r_a^q - 1} B_a$, for $q = 1, \ldots, p$, where $(D_a^* C)_q$ denotes the q-th row of $D_a^* C$.*

Proof. It follows from Definition 4.9 that in the case of covert attacks, adversaries should design the actuator attack signal $a_{\rm u}(k)$ and the sensor attack signal $a_{\rm y}(k)$ such that the following holds true:

$$C A^{r_{\rm a}-1} B_{\rm a} a_{\rm u}(0) + D_{\rm a} a_{\rm y}(1) = 0, \tag{77a}$$

$$C A^{r_{\rm a}} B_{\rm a} a_{\rm u}(0) + C A^{r_{\rm a}-1} B_{\rm a} a_{\rm u}(1) + D_{\rm a} a_{\rm y}(2) = 0, \tag{77b}$$

$$\vdots$$

$$C A^{N-2} B_{\rm a} a_{\rm u}(0) + C A^{N-3} B_{\rm a} a_{\rm u}(1) + C A^{N-4} B_{\rm a} a_{\rm u}(2)$$
$$+ \cdots + C A^{r_{\rm a}-1} B_{\rm a} a_{\rm u}(N-2) + D_{\rm a} a_{\rm y}(N-1) = 0. \tag{77c}$$

Consequently, in order to cancel out the impact of actuator attacks in (77), adversaries need to design the sensor attack signal in the following form:

$$a_{\rm y}(j+1) = -\sum_{\gamma=0}^{j} D_{\rm a}^\top C A^{r_{\rm a}+\gamma-1} B_{\rm a} a_{\rm u}(j-\gamma), \tag{78}$$

for $j = 0, \ldots, N-2$.

Necessary Condition: Suppose (77) holds and relative degrees of triples $(C, A, B_{\rm a})$ and $(D_{\rm a}^* C, A, B_{\rm a})$

are not equal. Given the definition of D_{a}^{*}, the relative degree of the q-th row of $D_{\mathrm{a}}^{*}C$ with respect to the actuator attack signal $a_{\mathrm{u}}(k)$ is either equal to that of C_q and $a_{\mathrm{u}}(k)$ or does not exist. Consider $\hat{s} \in \mathcal{S}_{\mathrm{a}}$ as a sensor at which $C_{\hat{s}}A^{r_{\mathrm{a}}^{\hat{s}}-1}B_{\mathrm{a}} \neq 0$ and $(D_{\mathrm{a}}^{*}C)_{\hat{s}}A^{r_{\mathrm{a}}^{\hat{s}}-1}B_{\mathrm{a}} = 0$, where $(D_{\mathrm{a}}^{*}C)_{\hat{s}}$ denotes the $\hat{s}$-th row of $D_{\mathrm{a}}^{*}C$.

Considering (78), let us rewrite the left-hand side of the j-th equation of (77) in the following form:

$$[(I_p - D_{\mathrm{a}}^{*})CA^{r_{\mathrm{a}}-1}B_{\mathrm{a}} \cdots (I_p - D_{\mathrm{a}}^{*})CA^{r_{\mathrm{a}}+j-1}B_{\mathrm{a}}] \begin{bmatrix} a_{\mathrm{u}}(j-1) \\ \vdots \\ a_{\mathrm{u}}(0) \end{bmatrix}, \tag{79}$$

for $j = 1, \ldots, N-1$. Consequently, (77) holds if and only if all rows of (79) are equal to zero. However, since we are considering any actuator attack signal, and $C_{\hat{s}}A^{r_{\mathrm{a}}^{\hat{s}}-1}B_{\mathrm{a}} \neq 0$ and $(D_{\mathrm{a}}^{*}C)_{\hat{s}}A^{r_{\hat{s}}-1}B_{\mathrm{a}} = 0$, there exists at least one nonzero row in (79), i.e., the row that corresponds to the $\hat{s}$-th sensor, which contradicts the assumption.

Sufficient Condition: Assume that the relative degrees of triples (C, A, B_{a}) and $(D_{\mathrm{a}}^{*}C, A, B_{\mathrm{a}})$ for all the sensors are equal. Hence, $C_qA^iB_{\mathrm{a}} = (D_{\mathrm{a}}^{*}C)_qA^iB_{\mathrm{a}}$, for any $i \in \mathbb{N}$ and $\forall q \in \{1, \ldots, p\}$. The latter implies that (78) is the solution to (77), and all rows in (79) are equal to zero. This completes the proof of the theorem. □

Theorem 4.3 states that in order to perform a covert attack, only those sensor measurements that are affected by actuator attacks are required to be manipulated. In other words, those sensors that one cannot define a relative degree for with respect to the actuator attack signal $a_{\mathrm{u}}(k)$ are not needed to be compromised in the case of covert attacks since they will not be affected by actuator attacks.

Remark 4.4. *Consider a case where the hypothesis of Theorem 4.3 is not hold. Consequently, if the triple* $((I_P - D_a^{*})C, A, B_a)$ *is not left invertible (see [98] for more details), adversaries can design a certain actuator attack signal* $a_u(k)$ *that makes* (79) *equal to zero despite having* $(I_p - D_a^{*})CA^{r_a-1}B_a \neq 0$. *However, in this case, the actuator attack signal* $a_u(k)$ *should be specifically designed to make* (79) *equal to zero whereas in Theorem 4.3, if the conditions are satisfied, adversaries can perform covert attacks for any arbitrary actuator attack signal. Hence, it can be concluded that the given necessary and sufficient conditions in Theorem 4.3 provide one with an upper bound for the number of actuators and sensors that*

should be manipulated to execute covert attacks.

Corollary 4.4. *Assume that the hypothesis of Theorem 4.3 holds. Given any $a_u(k) \in \mathbb{R}^{m_a}$, the sensor attack signal in a covert attack can be expressed by*

$$a_y(k+1) = -\sum_{\gamma=0}^{k} D_a^\top C A^{r_a+\gamma-1} B_a a_u(k-\gamma),$$

for $k \geq 0$, where $a_y(0) = 0$.

Proof. The proof follows along similar lines to that of Theorem 4.3 and is omitted for the sake of brevity. □

The conditions derived in Theorem 4.3 and Corollary 4.4 for the CPS (55) rely on certain elements within $\mathcal{O}_{\mathrm{N}}$ and $\mathcal{C}_{\mathrm{N}}$, namely, the observability matrix and the Markov parameters, which are defined by (59). Therefore, through the application of methodologies in [115–118], and by utilizing results presented in Theorem 4.3 and Corollary 4.4, one can investigate the vulnerability of the CPS to covert attacks by employing a data-driven approach.

4.3 Computing the Security Index for Covert Cyber-Attacks

In this section, we adopt the notion of security index from [16, 20, 21] to quantify the minimum number required actuators and sensors that should be compromised to carry out a covert attack. The security index in the CPS (55) for covert attacks can be defined in the following form [16, 20, 21], namely:

$$\begin{aligned} SI_{\mathrm{c}} \coloneqq \min_{a_{\mathrm{u}}(\cdot), a_{\mathrm{y}}(\cdot)} \quad & \|a_{\mathrm{u}}(\cdot)\|_0 + \|a_{\mathrm{y}}(\cdot)\|_0 \\ \text{s.t.} \quad & a(k) = [a_{\mathrm{u}}(k)^\top \; a_{\mathrm{y}}(k)^\top]^\top \text{ is a covert attack,} \end{aligned} \tag{80}$$

where $\|\cdot\|_0$ denotes the L_0 norm.

The computation of the SI_{c} in (80) is an NP-hard problem. Hence, graph-based methods for the structural representation of the CPS as well as geometric approaches have been utilized to find an upper bound for SI_{c} [16, 20]. On the other hand, in this section, we define an upper bound for covert cyber-attacks security index based on necessary and sufficient algebraic conditions provided in Theorem 4.3 (see Remark 4.4). In

particular, given the necessary and sufficient conditions in Theorem 4.3, below, we formulate the problem of computing an upper bound for $SI_{\rm c}$ in (80) as a trace minimization problem.

Let us define $L_{\rm a}^* = L_{\rm a} L_{\rm a}^\top$. Considering that $L_{\rm a}^*$ is a diagonal matrix with 0 and 1 entries, the provided condition in Theorem 4.3 for the existence of covert attacks can be rewritten as $C_q A^{r_{\rm a}^q - 1} B L_{\rm a}^* = (D_{\rm a}^* C)_q A^{r_{\rm a}^q - 1} B L_{\rm a}^*$, for $q = 1, \ldots, p$.

Definition 4.11. *In the CPS* (55)*, an upper bound for covert cyber-attacks security index is defined as*

$$\begin{aligned} \bar{SI}_c \coloneqq \min_{L_a^*, D_a^*} \quad & trace \begin{bmatrix} L_a^* & 0 \\ 0 & D_a^* \end{bmatrix} \\ s.t. \quad & \begin{bmatrix} (I_p - D_a^*)_1 C A^{r_a^1 - 1} B L_a^* \\ \vdots \\ (I_p - D_a^*)_p C A^{r_a^p - 1} B L_a^* \end{bmatrix} = 0. \end{aligned} \tag{81}$$

It should be noted that given $B_{\rm a} = B L_{\rm a}$, at each instance of the minimization problem in (81), the relative degree $r_{\rm a}^q$ should be updated accordingly, for $q = 1, \ldots, p$. Algorithm 1 can be utilized to find the provided $\bar{SI}_{\rm c}$ in (81).

4.4 Dynamic Coding Scheme to Prevent Zero dynamics and Controllable Cyber-Attacks

The studied cyber-attacks in Section 4.2 can be executed by adversaries that cause damage to the CPS while remaining undetected. Hence, in this section, a dynamic coding scheme on the input communication channels is developed that, under certain conditions, can be used to prevent adversaries from performing stealthy cyber-attacks such as the zero dynamics and controllable attacks on actuators. The coding scheme is designed such that having only one secure input communication channel will result in preventing adversaries from executing the zero dynamics attacks and controllable cyber-attacks.

Algorithm 1 Pseudo code to find $\bar{SI}_{\mathrm{c}}$

Input: (A, B, C), and the set of all inputs and outputs $S = \{u_1, \ldots, u_{\mathrm{m}}, y_1, \ldots, y_{\mathrm{p}}\}$

Output: $\bar{SI}$, and $\bar{S}_{\mathrm{min}}$ which is the set of actuators and sensors that should be attacked

Initialize $\bar{SI} = m + p$, $\bar{S}_{\mathrm{min}} = S$, $L_{\mathrm{a}}^* = I_m$, $D_{\mathrm{a}}^* = I_p$
Set $l = |S|$, where $|\cdot|$ denotes the cardinality of a set
for $i = 1 : 2^l - 1$ **do**
  Create the empty set $\hat{S} = \{\}$
  for $j = 1 : l$ **do**
    if the j-th bit of the binary representation of i is equal to 1 **then**
      Add j-th member of S to $\hat{S}$
    end if
  end for
  Compromise actuators and sensors that belong to the set $\hat{S}$, and update L_{a}^*, D_{a}^*, and $r_{\mathrm{a}}^1, \ldots, r_{\mathrm{a}}^p$ accordingly
  if $L_{\mathrm{a}}^* \neq 0$ **and** $D_{\mathrm{a}}^* \neq 0$ **then**
    if $\begin{bmatrix} (I_p - D_{\mathrm{a}}^*)_1 C A^{r_{\mathrm{a}}^1 - 1} B L_{\mathrm{a}}^* \\ \vdots \\ (I_p - D_{\mathrm{a}}^*)_p C A^{r_{\mathrm{a}}^p - 1} B L_{\mathrm{a}}^* \end{bmatrix} = 0$ **and** $|\hat{S}| \leq \bar{SI}$ **then**
      $\bar{SI} = |\hat{S}|$
      $\bar{S}_{\mathrm{min}} = \hat{S}$
    end if
  end if
end for

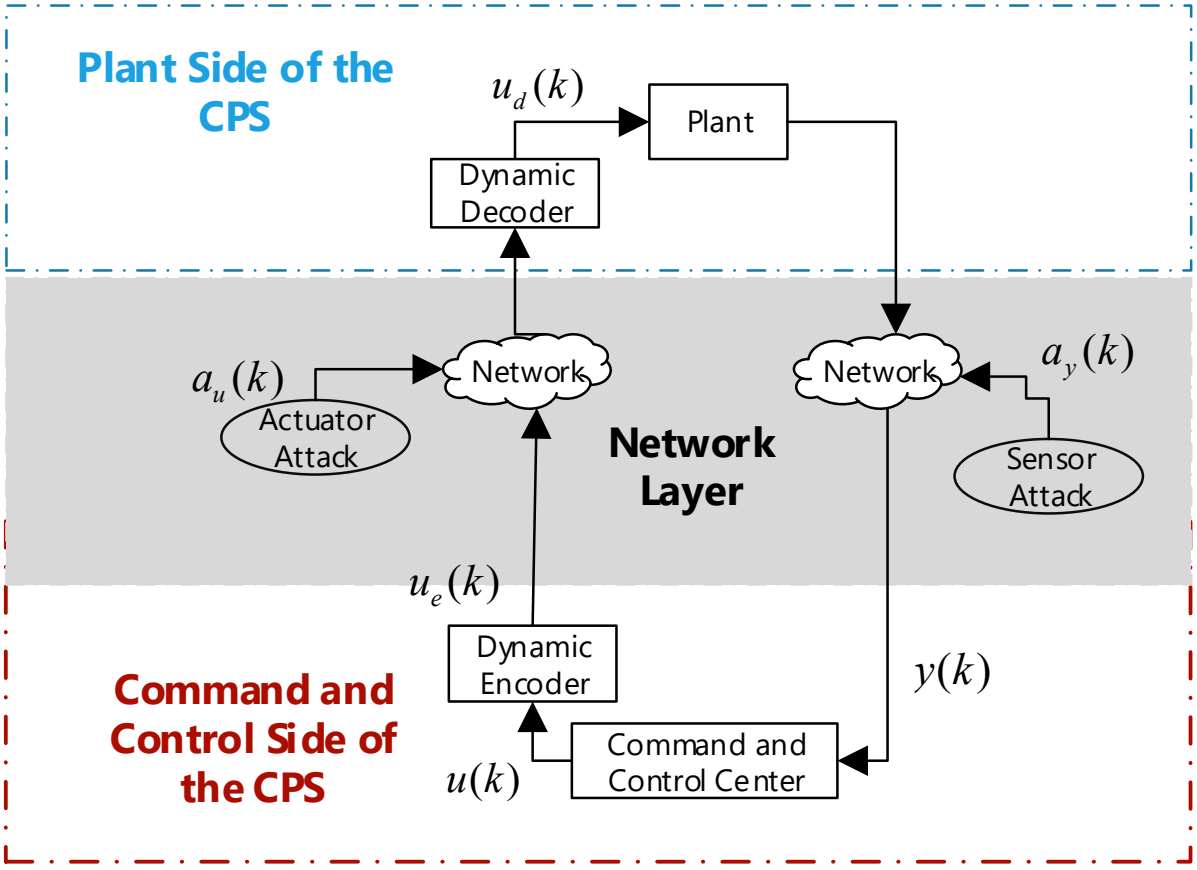

Figure 4.1: The architecture of the CPS and the dynamic coding scheme, where $u_{\mathrm{e}}(k)$ is the output of the encoder and $u_{\mathrm{d}}(k)$ is the output of the decoder.

4.4.1 CPS Model in Presence of the Dynamic Coding Scheme

An encoder, that is denoted by $\mathcal{E}$, on the command and control (C&C) side and a decoder, denoted by $\mathcal{D}$, on the plant side of the CPS are designed. The CPS along with the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$ are depicted in Figure 4.1.

The dynamics of the encoder and the decoder on the input communication channels of the CPS are

governed by

$$\mathcal{E}: \begin{cases} x_{\mathrm{e}}(k+1) & = A_{\mathrm{e}}x_{\mathrm{e}}(k) + B_{\mathrm{e}}u(k), \\ u_{\mathrm{e}}(k) & = C_{\mathrm{e}}x_{\mathrm{e}}(k) + D_{\mathrm{e}}u(k), \end{cases} \tag{82}$$

$$\mathcal{D}: \begin{cases} x_{\mathrm{d}}(k+1) & = A_{\mathrm{d}}x_{\mathrm{d}}(k) + B_{\mathrm{d}}(u_{\mathrm{e}}(k) + L_{\mathrm{a}}a_{\mathrm{u}}(k)), \\ u_{\mathrm{d}}(k) & = C_{\mathrm{d}}x_{\mathrm{d}}(k) + D_{\mathrm{d}}(u_{\mathrm{e}}(k) + L_{\mathrm{a}}a_{\mathrm{u}}(k)), \end{cases} \tag{83}$$

where $x_{\mathrm{e}}(k), x_{\mathrm{d}}(k) \in \mathbb{R}^{n_e}$ and $u_{\mathrm{e}}(k), u_{\mathrm{d}}(k) \in \mathbb{R}^m$ denote the states and outputs of the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$, respectively. Moreover, one has $x_{\mathrm{e}}(0) = x_{\mathrm{d}}(0) = 0$. The following lemma provides necessary and sufficient conditions under which the decoder $\mathcal{D}$ is the inverse of $\mathcal{E}$ such that once $a_{\mathrm{u}}(k) = 0$, one has $u_{\mathrm{d}}(k) = u(k), \forall k \geq 0$.

Lemma 4.2 ([17]). *Let $a_u(k) = 0$. One has $u_d(k) = u(k)$, $\forall k \geq 0$, if and only if there exists an invertible matrix T that satisfies the following:*

$$D_dC_e + C_dT = 0,\ T^{-1}B_dD_e = B_e,\ D_d = D_e^{-1},$$
$$T^{-1}A_dT + T^{-1}B_dC_e = T^{-1}A_dT - B_eC_dT = A_e.$$

In presence of $\mathcal{E}$ and $\mathcal{D}$, the dynamics of the CPS (55) under actuator attacks can be expressed as

$$\begin{aligned} x(k+1) =& Ax(k) + Bu_{\mathrm{d}}(k) + \omega(k), \\ y(k) =& Cx(k) + \nu(k). \end{aligned} \tag{84}$$

Consequently, the I/O model of the CPS (84) under noise free conditions is derived in the following form:

$$Y(N) = \mathcal{O}_{\mathrm{N}}x(0) + \mathcal{C}_{\mathrm{N}}(U(N) + \mathcal{C}_{\mathrm{d}}U_{\mathrm{a}}(N)), \tag{85}$$

where $\mathcal{C}_{\mathrm{d}} = \Gamma_{\mathrm{d}} \otimes L_{\mathrm{a}}$ and

$$\Gamma_{\mathrm{d}} = \begin{bmatrix} D_{\mathrm{d}} & 0 & \cdots & 0 \\ C_{\mathrm{d}}B_{\mathrm{d}} & D_{\mathrm{d}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C_{\mathrm{d}}A_{\mathrm{d}}^{N-2}B_{\mathrm{d}} & C_{\mathrm{d}}A_{\mathrm{d}}^{N-3}B_{\mathrm{d}} & \cdots & D_{\mathrm{d}} \end{bmatrix}. \tag{86}$$

Assumption 4.2. *The encoder* (82) *and the decoder* (83) *are designed according to Lemma 4.2. Moreover, adversaries have knowledge on the parameters of* Γ_d *in* (86)*.*

As can be seen from (85), due to existence of the coding scheme, the impact of actuator cyber-attack signals shows up as $\mathcal{C}_{\mathrm{N}}\mathcal{C}_{\mathrm{d}}U_{\mathrm{a}}(N)$ in the sensor measurements. It should be noted that since Γ_{d} is by definition an invertible matrix, $\ker(\mathcal{C}_{\mathrm{N}}) = \ker(\mathcal{C}_{\mathrm{N}}\Gamma_{\mathrm{d}})$. According to Definition 4.8, $\ker(\mathcal{C}_{\mathrm{N}})$ determines the existence of controllable cyber-attacks. Hence, having Γ_{d} in (85) does not result in introducing additional controllable cyber-attacks that can be executed in the CPS.

Furthermore, since Γ_{d} is invertible, one has $\mathrm{Im}(\mathcal{C}_{\mathrm{N}}) = \mathrm{Im}(\mathcal{C}_{\mathrm{N}}\Gamma_{\mathrm{d}})$. Moreover, as per Definition 4.6, in order for the zero dynamics cyber-attacks to exist, one needs to have $\mathrm{Im}(\mathcal{O}_{\mathrm{N}}x(0)) \subseteq \mathrm{Im}(\mathcal{C}_{\mathrm{N}})$. Consequently, since $\mathrm{Im}(\mathcal{C}_{\mathrm{N}}) = \mathrm{Im}(\mathcal{C}_{\mathrm{N}}\Gamma_{\mathrm{d}})$, Γ_{d} does not introduce any new zero dynamics in the CPS (85). However, in order to take into account the impact of the decoder $\mathcal{D}$, the zero dynamics and controllable attacks in presence of the coding scheme need to be redefined. Also, the definition of the relative degree should be modified for the CPS in presence of the dynamic coding scheme.

Definition 4.12 (Cyber-Attacks and the Coding Scheme)**.** *In the I/O model of the CPS* (85)*, let* $U_a(N) = \tilde{\mathcal{C}}_d\tilde{U}_a(N)$*, where* $\tilde{\mathcal{C}}_d$ *is designed such that* $\mathcal{C}_N\mathcal{C}_d\tilde{\mathcal{C}}_d = \mathcal{C}_a$*. Consequently, the following can be stated:*

(1) There exists a zero dynamics cyber-attack if $\tilde{U}_a(N)$ *is designed according to Definition 4.6 such that* $\mathcal{O}_N x(0) + \mathcal{C}_N\mathcal{C}_d\tilde{\mathcal{C}}_d\tilde{U}_a(N) = 0$*.*

(2) The actuator cyber-attack is a controllable attack if $\tilde{U}_a(N)$ *is designed according to Definition 4.8 such that* $\mathcal{C}_N\mathcal{C}_d\tilde{\mathcal{C}}_d\tilde{U}_a(N) = 0$*.*

Moreover, if one cannot execute the zero dynamics and controllable cyber-attacks in the CPS (84)*, the CPS is considered to be secure against these cyber-attacks.*

Definition 4.13. *Let $\mathcal{I}_a = \{a_1, \ldots, a_{m_a}\}$ denote the set of compromised input communication channels. The relative degree of the CPS (84) with respect to the q-th attacked input channel is r_q^d if $CA^iB(D_dL_a)_q = 0$ for all $i < r_q^d - 1$ and $CA^{r_q^d-1}B(D_dL_a)_q \neq 0$, for every $q \in \mathcal{I}_a$, where $(D_dL_a)_q$ denotes the q-th column of D_dL_a. Moreover, $r_a^d = min\{r_{a_1}^d, \ldots, r_{a_m}^d\}$.*

As stated in Definition 4.12, the adversary's objective is to cancel out the impact of the dynamic coding scheme by designing the actuator cyber-attack signals and maintaining the cyber-attack stealthy. Hence, the design conditions for $\mathcal{E}$ and $\mathcal{D}$ under which adversaries cannot evade the coding scheme to maintain their attacks undetected are discussed in the next subsection.

The following assumption holds throughout this section.

Assumption 4.3. *In the CPS (84), there exists at least one secure input communication channel, i.e., rank$(L_a) < m$.*

4.4.2 Designing the Dynamic Coding Scheme for Securing the CPS Against Zero Dynamics and Controllable Cyber-Attacks

As per Definition 4.12, in order to execute the zero dynamics and controllable cyber-attacks, adversaries need to first eliminate the impact of the coding scheme by designing $\tilde{\mathcal{C}}_{\mathrm{d}}$. Hence, our objective is to design and develop the coding scheme such that having only one secured input channel will prevent adversaries from having $\mathcal{C}_{\mathrm{N}}\mathcal{C}_{\mathrm{d}}\tilde{\mathcal{C}}_{\mathrm{d}} = \mathcal{C}_{\mathrm{a}}$. If the latter objective is achieved, the impact of the actuator cyber-attacks will always show up in the sensor measurements and cannot be eliminated by adversaries.

Theorem 4.4. *Under Assumption 4.3, for any set of attacked input channels $\mathcal{I}_a$, i.e., any L_a, adversaries cannot perform zero dynamics and controllable cyber-attacks in the sense of Definition 4.12 if C_dB_d is a full rank matrix and Im$((C_dB_d)_q) \nsubseteq$ Im$((D_d)_q)$, for $q = 1, \ldots, m$, where $(C_dB_d)_q$ and $(D_d)_q$ are the q-th columns of C_dB_d and D_d, respectively.*

Proof. According to the Definition 4.12, in zero dynamics cyber-attacks, one has $\mathcal{O}_{\mathrm{N}}x(0)+\mathcal{C}_{\mathrm{N}}\mathcal{C}_{\mathrm{d}}U_{\mathrm{a}}(N) = 0$,

which at instances $k = r_{\mathrm{a}}^{\mathrm{d}}$ and $k = r_{\mathrm{a}}^{\mathrm{d}} + 1$ results in

$$CA^{r_{\mathrm{a}}^{\mathrm{d}}}x(0) + CA^{r_{\mathrm{a}}^{\mathrm{d}}-1}BD_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}(0) = 0, \tag{87a}$$

$$CA^{r_{\mathrm{a}}^{\mathrm{d}}+1}x(0) + (CA^{r_{\mathrm{a}}^{\mathrm{d}}}BD_{\mathrm{d}} + CA^{r_{\mathrm{a}}^{\mathrm{d}}-1}BC_{\mathrm{d}}B_{\mathrm{d}})L_{\mathrm{a}}a_{\mathrm{u}}(0) + CA^{r_{\mathrm{a}}^{\mathrm{d}}-1}BD_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}(1) = 0. \tag{87b}$$

The actuator cyber-attacks signal can be written as $a_{\mathrm{u}}(1) = a_{\mathrm{u}}^{\mathrm{z}}(1) + a_{\mathrm{u}}^{\mathrm{d}}(1)$, where $a_{\mathrm{u}}^{\mathrm{z}}(1)$ is designed according to the Corollary 4.1 and is the zero dynamics cyber-attack signal for the triple $(C, A, BD_{\mathrm{d}}L_{\mathrm{a}})$. Moreover, $a_{\mathrm{u}}^{\mathrm{d}}(1)$ is designed to eliminate the impact of the dynamic coding from the sensor measurements such that

$$CA^{r_{\mathrm{a}}^{\mathrm{d}}-1}BC_{\mathrm{d}}B_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}(0) + CA^{r_{\mathrm{a}}^{\mathrm{d}}-1}BD_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}^{\mathrm{d}}(1) = 0. \tag{88}$$

In case of controllable cyber-attacks, at $k = r_{\mathrm{a}}^{\mathrm{d}}$ and $k = r_{\mathrm{a}}^{\mathrm{d}} + 1$ one has

$$CA^{r_{\mathrm{a}}^{\mathrm{d}}-1}BD_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}(0) = 0, \tag{89a}$$

$$CA^{r_{\mathrm{a}}^{\mathrm{d}}}BD_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}(0) + CA^{r_{\mathrm{a}}^{\mathrm{d}}-1}BC_{\mathrm{d}}B_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}(0) + CA^{r_{\mathrm{a}}^{\mathrm{d}}-1}BD_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}(1) = 0. \tag{89b}$$

Similar to the case of the zero dynamics cyber-attacks the actuator attack signal can be recast as $a_{\mathrm{u}}(1) = a_{\mathrm{u}}^{\mathrm{c}}(1) + a_{\mathrm{u}}^{\mathrm{d}}(1)$, where $a_{\mathrm{u}}^{\mathrm{c}}(1)$ is designed according to the Corollary 4.2 for the triple $(C, A, BD_{\mathrm{d}}L_{\mathrm{a}})$, and $a_{\mathrm{u}}^{\mathrm{d}}(1)$ is designed to cancel out the impact of the coding scheme such that it satisfies (88).

The condition (88) is satisfied if

$$C_{\mathrm{d}}B_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}(0) + D_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}^{d}(1) = \zeta\hat{a}_0, \tag{90}$$

where ζ is a scalar and $\hat{a}_0 \in \ker(CA^{r_{\mathrm{a}}^{\mathrm{d}}-1}B)$. If $\zeta \neq 0$, $C_{\mathrm{d}}B_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}(0) + D_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}^{d}(1)$ will show up in the next instances of the output, i.e., $k \geq r_{\mathrm{a}}^{\mathrm{d}} + 2$. Hence, adversaries may try to design $a_{\mathrm{u}}^{d}(1)$ to satisfy (90) for $\zeta = 0$.

There exists $a_{\mathrm{u}}^{\mathrm{d}}(1)$ that can satisfy (90) for $\zeta = 0$ if $\mathrm{Im}(C_{\mathrm{d}}B_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}(0)) \subseteq \mathrm{Im}(D_{\mathrm{d}}L_{\mathrm{a}})$. Since $C_{\mathrm{d}}B_{\mathrm{d}}$

is a square matrix, having a full rank $C_{\rm d}B_{\rm d}$ such that $\text{Im}((C_{\rm d}B_{\rm d})_q) \nsubseteq \text{Im}((D_{\rm d})_q)$, for $q = 1, \ldots, m$, implies that the q-th column of $C_{\rm d}B_{\rm d}$ is a basis of $\mathbb{R}^m$ which is different from all the columns of $D_{\rm d}$. The latter implies that all the columns of $D_{\rm d}$ should be accessible by the adversaries, i.e., $L_{\rm a} = I_m$, to have $\text{Im}(C_{\rm d}B_{\rm d}L_{\rm a}) \subseteq \text{Im}(D_{\rm d}L_{\rm a})$. Hence, under Assumption 4.3, if $\text{Im}((C_{\rm d}B_{\rm d})_q) \nsubseteq \text{Im}((D_{\rm d})_q)$, for $q = 1, \ldots, m$, having any rank deficient $L_{\rm a}$ results in $\text{Im}(C_{\rm d}B_{\rm d}L_{\rm a}a_{\rm u}(0)) \subseteq \text{Im}(C_{\rm d}B_{\rm d}L_{\rm a}) \nsubseteq \text{Im}(D_{\rm d}L_{\rm a})$. This completes the proof of the theorem. □

As the main implication of the proposed dynamic coding scheme in Theorem 4.4, the security index for the CPS (84) is now equal to m. Hence, even if adversaries know the dynamics of the CPS (54), the encoder $\mathcal{E}$ in (82), and the decoder $\mathcal{D}$ given by (83) (as considered in Assumption 4.2), the adversaries still need to compromise all the input channels of the CPS to execute the zero dynamics and controllable cyber-attacks.

Note that a static coding scheme where $A_{\rm d} = 0$, $B_{\rm d} = 0$, $C_{\rm d} = 0$, and $D_{\rm d} \neq 0$ may require more than one secure communication channel to prevent adversaries from performing the zero dynamics and controllable cyber-attacks. For instance, in case of a static coding scheme, for a certain $x(0)$ and $\text{rank}(L_{\rm a}) = m - 1$, i.e., one secure input channel, it is possible to have $CA^{r_{\rm a}^{\rm d}}x(0) + CA^{r_{\rm a}^{\rm d}-1}BD_{\rm d}L_{\rm a}g = 0$ such that $\text{Im}(CA^{r_{\rm a}+1}x_0 + CA^{r_{\rm a}}BD_{\rm d}L_{\rm a}g) \subseteq \text{Im}(CA^{r_{\rm a}-1}BD_{\rm d}L_{\rm a}g)$ (refer to Theorem 4.1), which are the necessary and sufficient conditions for existence of zero dynamics cyber-attacks, under certain assumptions. On the other hand, if Theorem 4.4 is utilized to design the dynamic coding scheme, having only one secure input channel will prevent adversaries from executing the zero dynamics and controllable cyber-attacks.

Since in the Theorem 4.4 there is no condition on $A_{\rm d}$, it can be selected as $A_{\rm d} = I_{\rm m}$. However, $B_{\rm d}$, $C_{\rm d}$, and $D_{\rm d}$ should be designed according to Theorem 4.4. Consequently, $A_{\rm e}$, $B_{\rm e}$, $C_{\rm e}$, and $D_{\rm e}$ should be designed to satisfy the conditions in the Lemma 4.2.

4.5 Dynamic Coding Scheme to Prevent Covert Cyber-Attacks

In this section, a dynamic coding scheme on the input communication channels is developed and proposed which can be used to prevent adversaries from performing covert attacks. The coding scheme is designed such that having only one secure input and two secure output communication channels will result in preventing adversaries from executing covert cyber-attacks.

We utilize the encoder (82) and the decoder (83). In presence of $\mathcal{E}$ and $\mathcal{D}$, the dynamics of the CPS (55) under both actuator and sensor attacks can be described as follows:

$$\begin{aligned} x(k+1) =& Ax(k) + Bu_{\mathrm{d}}(k) + \omega(k), \\ y(k) =& Cx(k) + D_{\mathrm{a}}a_{\mathrm{y}}(k) + \nu(k). \end{aligned} \tag{91}$$

Consequently, the noise free I/O model of the CPS (91) takes the following form:

$$Y(N) = \mathcal{O}_{\mathrm{N}}x(0) + \mathcal{C}_{\mathrm{N}}U(N) + \mathcal{C}_{\mathrm{N}}\mathcal{C}_{\mathrm{d}}U_{\mathrm{a}}(N) + \mathcal{D}_{\mathrm{a}}Y_{\mathrm{a}}(N), \tag{92}$$

where $\mathcal{C}_{\mathrm{d}} = \Gamma_{\mathrm{d}} \otimes L_{\mathrm{a}}$ and Γ_{d} is defined in (86).

The presence of the coding scheme leads to the appearance of actuator attack signal impact as $\mathcal{C}_{\mathrm{N}}\mathcal{C}_{\mathrm{d}}U_{\mathrm{a}}(N)$ in sensor measurements. Given the existence of the coding scheme in (91), we need to redefine the relative degree for our system.

Definition 4.14. *The relative degree of the q-th output of the CPS (91) with respect to the actuator attack signal $a_u(k)$ is r_d^q if $C_qA^{i_q}BD_dL_a = 0$ for all $i_q < r_d^q - 1$ and $C_qA^{r_d^q-1}BD_dL_a \neq 0$, for $q = 1, \ldots, p$, where C_q represents the q-th row of C. If for any positive integer i_q one has $C_qA^{i_q}BD_dL_a = 0$, the relative degree for the q-th output with respect to $a_u(k)$ cannot be defined. Also, we define $r_d = min\{r_d^1, \ldots, r_d^p\}$.*

The following assumption holds throughout this section.

Assumption 4.4. *In the CPS (91), there exist at least one secure input communication channel and two secure output communication channels, i.e., rank$(L_a) < m$ and rank$(D_a) < p - 1$.*

4.5.1 Design Specifications of the Dynamic Coding Scheme for Securing the CPS Against Covert Attacks

In presence of the coding scheme and in the case of covert attacks, adversaries may try to use sensor attack signals to eliminate the impact of actuator attack signals from output measurements. Hence, the proposed coding schemes should satisfy two requirements. First, when there exists a secure input communication channel, the impact of the coding scheme cannot be canceled out through actuator attack signals.

Second, to design the coding matrices such that all sensors are affected by actuator attacks. Under the latter condition, adversaries need to have access to all output communication channel to eliminate the impact of their actuator cyber-attacks from sensor measurements. Design specification of $\mathcal{E}$ and $\mathcal{D}$ are provided in the following theorem.

Theorem 4.5. *Under Assumption 4.4, consider q_{s1} and q_{s2} as two secure output communication channels. In the CPS (91), adversaries cannot perform covert cyber-attacks in the sense of Definition 4.9 if the following conditions are satisfied:*

(1) $C_{q_{s1}} A^{r_d-1} B D_d L_a = 0$;

(2) $ker(C_{q_{s2}} A^{r_d-1} B D_d L_a) \cap ker(C_{q_{s1}} A^{r_d} B D_d L_a) = 0$;

(3) and $C_{q_{s1}} A^{r_d-1} B C_d B_d L_a = 0$.

Proof. Adversaries need to design their attack signals such that $\mathcal{C}_{\mathrm{N}}\mathcal{C}_{\mathrm{d}}U_{\mathrm{a}}(N) + \mathcal{D}_{\mathrm{a}}Y_{\mathrm{a}}(N) = 0$. Hence, at the r_{d}-th and $r_{\mathrm{d}}+1$-th instances of the output, actuator and sensor attack signals should satisfy the following:

$$CA^{r_{\mathrm{d}}-1}BD_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}(0) + D_{\mathrm{a}}a_{\mathrm{y}}(1) = 0, \tag{93a}$$

$$CA^{r_{\mathrm{d}}}BD_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}(0) + CA^{r_{\mathrm{d}}-1}BC_{\mathrm{d}}B_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}(0) + CA^{r_{\mathrm{d}}-1}BD_{\mathrm{d}}L_{\mathrm{a}}a_{\mathrm{u}}(1) + D_{\mathrm{a}}a_{\mathrm{y}}(2) = 0. \tag{93b}$$

Under Assumption 4.4, measurements that are transmitted through the communication channels q_{s1} and q_{s2} cannot be manipulated by means of sensor attacks. Moreover since in condition 1) we have $C_{q_{\mathrm{s1}}} A^{r_{\mathrm{d}}-1} B D_{\mathrm{d}} L_{\mathrm{a}} = 0$, in order to satisfy (93a), the actuator attack signal should be designed such that $a_{\mathrm{u}}(0) \in \ker(C_{q_{\mathrm{s2}}} A^{r_{\mathrm{d}}-1} B D_{\mathrm{d}} L_{\mathrm{a}})$. Furthermore, it follows from condition 2) that $C_{q_{\mathrm{s1}}} A^{r_{\mathrm{d}}} B D_{\mathrm{d}} L_{\mathrm{a}} a_{\mathrm{u}}(0) \neq 0$. Since q_{s1} is a secure output communication channel and as per conditions 1) and 3) we have $C_{q_{\mathrm{s1}}} A^{r_{\mathrm{d}}-1} B D_{\mathrm{d}} L_{\mathrm{a}} \times a_{\mathrm{u}}(1) = 0$ and $C_{q_{\mathrm{s1}}} A^{r_{\mathrm{d}}-1} B C_{\mathrm{d}} B_{\mathrm{d}} L_{\mathrm{a}} a_{\mathrm{u}}(0) = 0$, respectively, the impact of $a_{\mathrm{u}}(0)$ cannot be removed from the q_{s1}-th communication channel and (93b) cannot be satisfied. This completes the proof of the theorem. □

The main objective in Theorem 4.5 is to design the coding scheme such that the impact of actuator attack signals show up in the sensor measurements that are secured. Moreover, the proposed design specifications

in Theorem 4.5 ensure that adversaries are not capable of removing the impact of their actuator attacks from sensor measurements that are transmitted through secure output communication channels q_{s1} and q_{s2}. Hence, as the main implication of Theorem 4.5, the CPS operators can prevent adversaries from executing covert attacks by securing 3 input and output communication channels and employing the proposed coding scheme in this subsection.

According to Theorem 4.5, in order to design the coding scheme, D_d should satisfy $C_{q_{s1}} A^{r_d-1} B D_d L_a = 0$ and $\ker(C_{q_{s2}} A^{r_d-1} B D_d L_a) \cap \ker(C_{q_{s1}} A^{r_d} B D_d L_a) = 0$, simultaneously. Consequently, C_d and B_d should satisfy $C_{q_{s1}} A^{r_d-1} B C_d B_d L_a = 0$, where one simple design choice could be $C_d B_d = D_d$. Also, Theorem 4.5 does not impose any design conditions on A_d. After designing the decoder $\mathcal{D}$, the encoder $\mathcal{E}$ should be developed as per Lemma 4.2.

4.6 Numerical Case Studies

In this section, three case studies are provided. In the first case study, the zero dynamics cyber-attacks are investigated. Motivated by the studied Quadruple-Tank Process (QTP) in [24, 104], a modified state-space representation of the QTP with an additional input is considered. We use the results in Theorem 4.1 to investigate the existence of the zero dynamics cyber-attacks for the modified QTP and implement the attack by utilizing Corollary 4.1. Moreover, following results in Section 4.4 and Theorem 4.4, a dynamic coding scheme is designed and implemented for the modified QTP.

The second case study is concerned with controllable cyber-attacks in the flight control system of a small single-engine fighter aircraft. We obtain the dynamics of the aircraft from [119, 120]. By utilizing Theorem 4.2 and Corollary 4.3, the existence and implementation of the controllable cyber-attacks in the considered flight control system are studied. Moreover, the proposed dynamic coding scheme in Section 4.4 and Theorem 4.4 is implemented and analyzed for the second case study. In the third case study, we study covert cyber-attacks, as per Definition 4.9, in the flight control system of the fighter aircraft.

4.6.1 Zero Dynamics Attacks in the Quadruple Tank Process

In the first case study, we consider the zero dynamics cyber-attacks in the sense of Definition 4.6. In [104] and [24], the QTP has two main pumps which their input voltages correspond to two control inputs of

the system. However, in this case study, we consider an additional pump which pumps water into the first tank. The linearized characteristic matrices of the QTP with the additional pump and the sampling period of $T_{\mathrm{s}} = 0.5$ (s) are expressed by [24]

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

We assume that the first two actuators of the QTP are under cyber-attacks and the third actuator is secured, i.e., we have

$$B_{\mathrm{a}} = \begin{bmatrix} 0.0515 & 0.0016 \\ 0.0019 & 0.0447 \\ 0 & 0.0737 \\ 0.0850 & 0 \end{bmatrix}.$$

Consequently, the relative degree r_{a} is equal to 1 (refer to Definition 4.7). Let us consider the initial condition $\bar{x}_0 = [0,\ 0,\ -1,\ 2]^\top$, that belongs to the null space of C and $\mathrm{Im}(CA\bar{x}_0) \subseteq \mathrm{Im}(CB_{\mathrm{a}})$. However, since there does not exist any $g \in \mathbb{R}^2$ such that $\mathrm{Im}(CA^2\bar{x}_0 + CAB_{\mathrm{a}}g) \subseteq \mathrm{Im}(CB_{\mathrm{a}}g)$, according to Theorem 4.1, the initial condition $\bar{x}_0$ is not associated with any invariant zero.

From Theorem 4.1, it follows that the given QTP with (A, B_{a}, C) has two zeros that correspond to initial conditions $x_{01} = [0,\ 0,\ 0.1,\ 0.1]^\top$ and $x_{02} = [0,\ 0,\ -0.72,\ 0.69]^\top$. In particular, for each initial condition one has $\mathrm{Im}(CAx_0) \subseteq \mathrm{Im}(CB_{\mathrm{a}})$, which result in having $CAx_0 + CB_{\mathrm{a}}g = 0$ such that $\mathrm{Im}(CA^2x_0 + CAB_{\mathrm{a}}g) \subseteq \mathrm{Im}(CB_{\mathrm{a}}g)$. Moreover, for the QTP system, one has $\ker(C) \bigcap \ker(CA) = 0$. Hence, one can utilize the Corollary 4.1 to compute the input-zero direction that corresponds to x_{01} as $g_1 = [-0.0786,\ -0.0951]^\top$ with a minimum phase zero $\lambda_1 = 0.8886$.

Moreover, the initial condition x_{02} corresponds to the input-zero direction $g_2 = [0.6091,\ -0.7051]$ and a non-minimum phase zero $\lambda_2 = 1.0306$. It should be emphasized that using the Rosenbrock system matrix of the triple (C, A, B_{a}) or the `tzero` MATLAB function, one can also find both invariant zeros of the QTP as $\lambda_1 = 0.8886$ and $\lambda_2 = 1.0306$ (refer to [24]). Consequently, for the initial condition x_{02}, we design

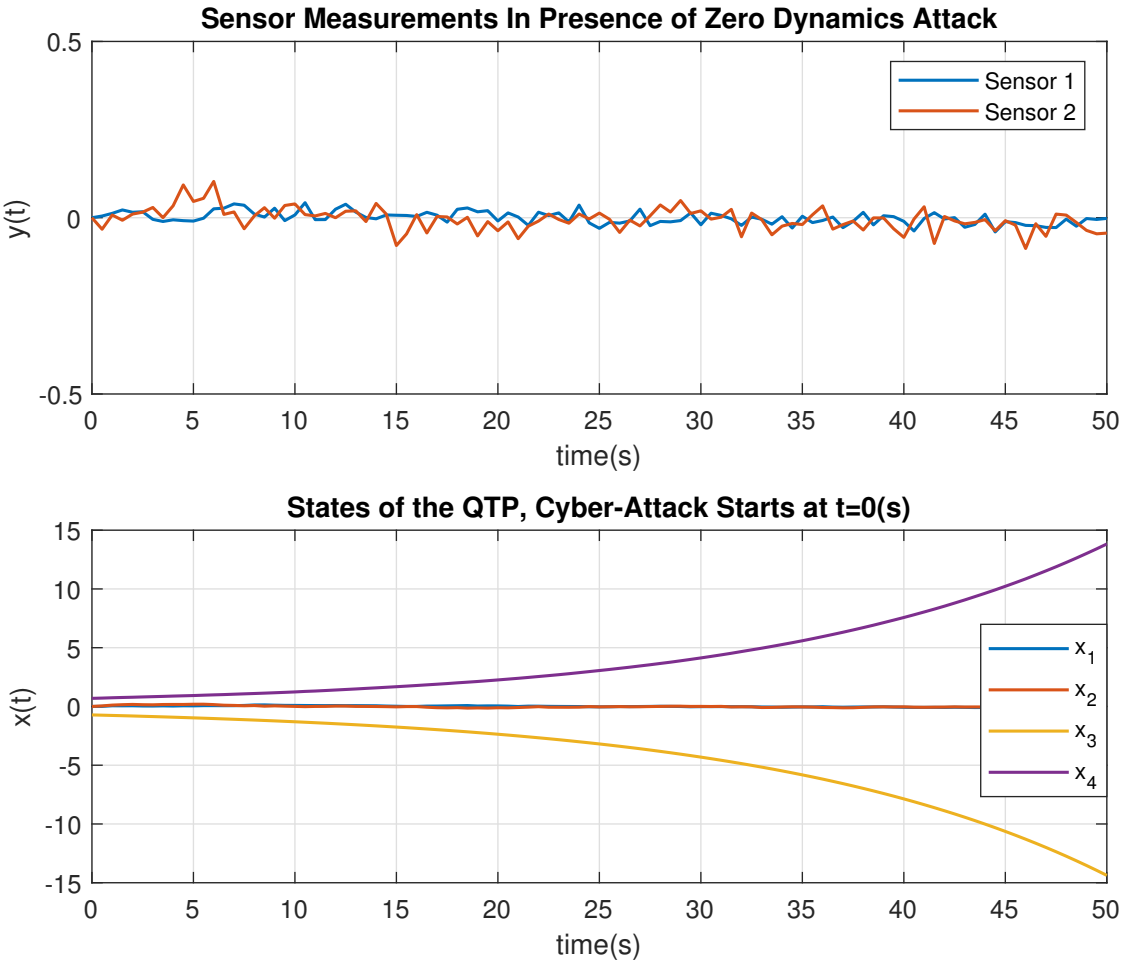

Figure 4.2: The QTP system under the zero dynamics cyber-attack injected at $t = 0$ (s).

the zero dynamics cyber-attack signal $a_{\mathrm{u}}(k) = g_2\lambda_2^k$. As shown in Figure 4.2, once the zero dynamics cyber-attack is injected at $k = 0$, i.e., $t = 0$ (s), in presence of the process and measurement noise, sensor measurements of the QTP remain close to zero, while the values of states grow unbounded. Moreover, since A is a Schur stable matrix and as discussed in Remark 4.2, considering Figure 4.3, the error due to starting the zero dynamics cyber-attack at $k = 20$, i.e., $t = 10$ (s) where $x(10) \neq x_{02}$ is minimal and not significant.

Given that the third actuator in the modified QTP is secured, Assumption 4.3 is satisfied and the proposed dynamic coding scheme in Section 4.4 can be employed. Hence, we design an encoder $\mathcal{E}$ given by (82) and a decoder $\mathcal{D}$ (83) that satisfy the conditions in Lemma 4.2. The parameters of $\mathcal{E}$ and $\mathcal{D}$ are $A_{\mathrm{d}} = I_3$,

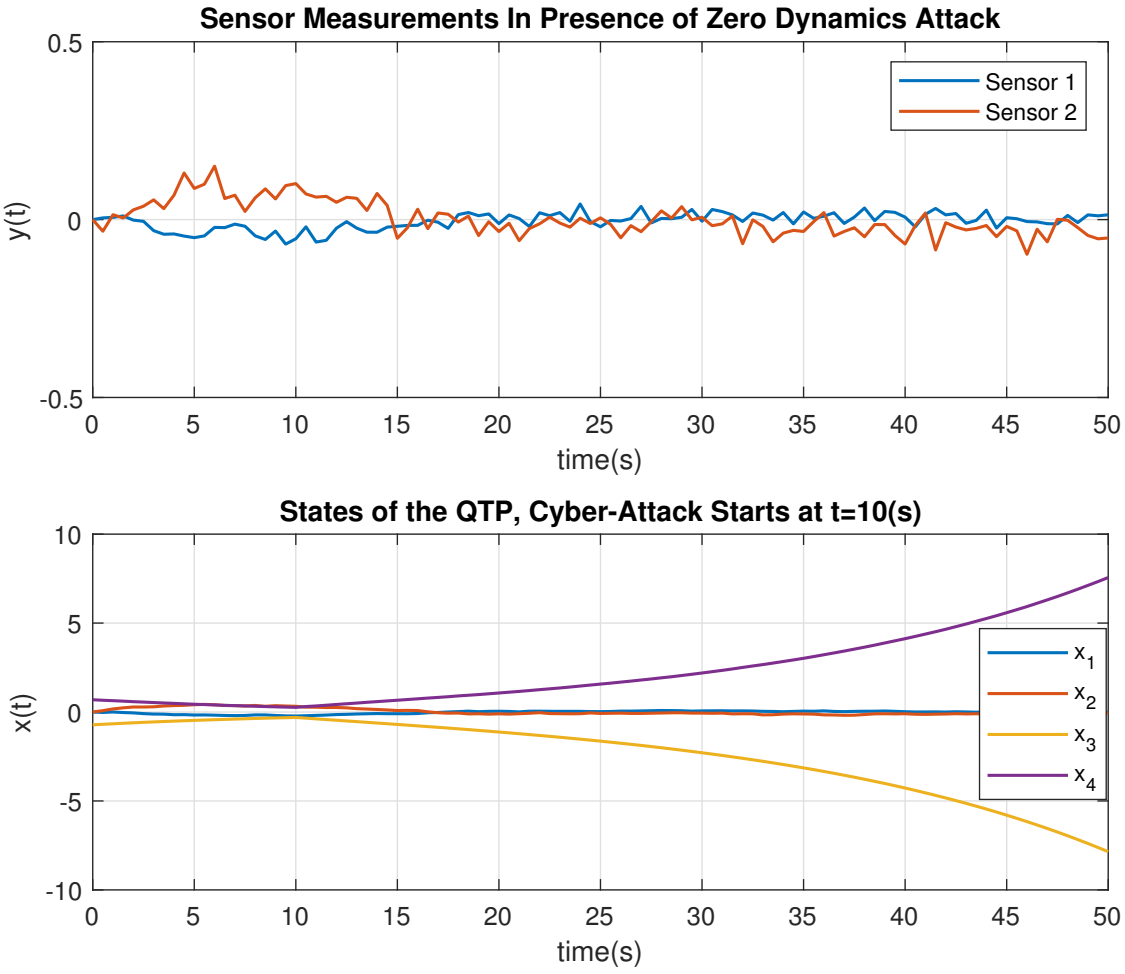

Figure 4.3: The zero dynamics cyber-attack injected at $t = 10$ (s) in the QTP system.

$C_{\rm d} = -C_{\rm e} = I_3$, $D_{\rm d} = D_{\rm e}^{-1} = I_3$, $A_{\rm e} = A_{\rm d} - B_{\rm e}C_{\rm d}$, and

$$B_{\rm e} = B_{\rm d} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

It can be seen that the q-th column of $C_{\rm d}B_{\rm d}$ is different from the q-th column of $D_{\rm d}$, for $q = 1, \ldots, 3$. Hence, conditions in Theorem 4.4 are satisfied and in presence of the dynamic coding scheme, adversaries will not be able to execute zero dynamics cyber-attacks. As shown in Figure 4.4, by utilizing the designed

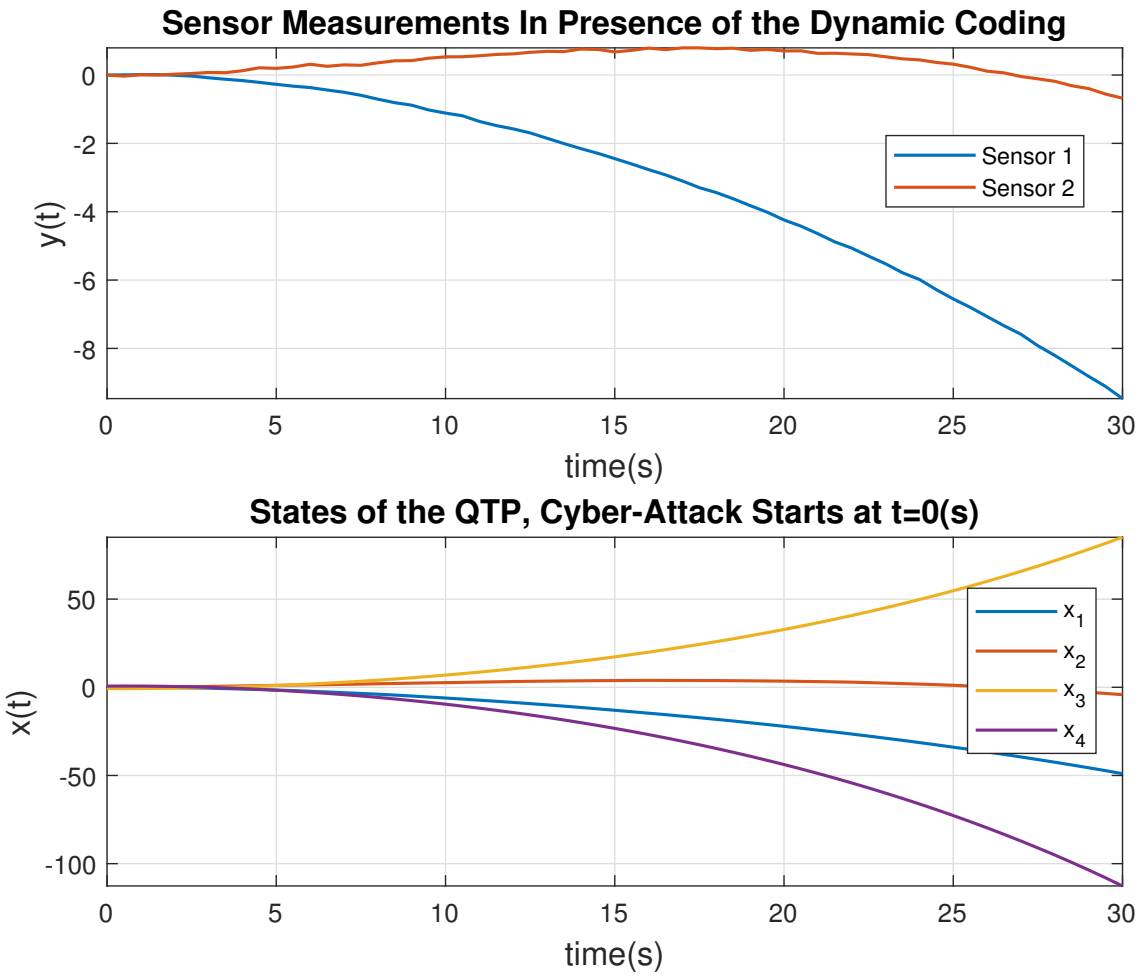

Figure 4.4: Impact of the dynamic coding scheme in securing the modified QTP against the zero dynamics cyber-attack injected at $t = 0$ (s).

dynamic coding scheme, the impact of the zero dynamics cyber-attack can be seen and detected in the sensor measurements.

4.6.2 Controllable Attacks in the Flight Control System of a Fighter Aircraft

In our second case study, we consider the controllable cyber-attacks that are described in Definition 4.8. The characteristic matrices of the linearized aircraft system with the sampling period of $T_s = 0.5$ (s) are

given by [119, 120]

$$A_{\mathrm{f}} = \begin{bmatrix} 1.0214 & 0.0054 & 0.0003 & 0.4176 & -0.0013 \\ 0 & 0.6307 & 0.0821 & 0 & -0.3792 \\ 0 & -3.4485 & 0.3779 & 0 & 1.1569 \\ 1.1199 & 0.0024 & 0.0001 & 1.0374 & -0.0003 \\ 0 & 0.3802 & -0.0156 & 0 & 0.8062 \end{bmatrix},$$

$$B_{\mathrm{f}} = \begin{bmatrix} 0.1823 & -0.1798 & -0.1795 & 0.0008 \\ 0 & -0.0639 & 0.0639 & 0.1397 \\ 0 & -1.5840 & 1.5840 & 0.2936 \\ 0.8075 & -0.6456 & -0.6456 & 0.0013 \\ 0 & -0.1005 & 0.1005 & -0.4114 \end{bmatrix},$$

$$C_{\mathrm{f}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

The matrix B_{f} has a full column rank and is an injective map. In this case study, the first 3 actuators of the system are compromised by adversaries, i.e., $\mathcal{I}_{\mathrm{a}} = \{1, 2, 3\}$, and the input channel 4 is attack free. Hence, one has $\mathrm{rank}(L_{\mathrm{a}}) = 3$, and the actuator cyber-attack signature is $B_{\mathrm{a}} = [(B_{\mathrm{f}})_1, (B_{\mathrm{f}})_2, (B_{\mathrm{f}})_3]$, where $(B_{\mathrm{f}})_q$ denotes the q-th column of B_{f}. Since $C_{\mathrm{f}}(B_{\mathrm{f}})_q \neq 0$, for every $q = 1, \ldots, 4$, each actuator of the system $\Sigma_{\mathrm{f}} = (C_{\mathrm{f}}, A_{\mathrm{f}}, B_{\mathrm{f}})$ yields a relative degree equal to 1 which implies that $r_{\mathrm{a}} = 1$.

The basis of the null space of $C_{\mathrm{f}}B_{\mathrm{a}}$ is $\hat{a}_0 = [-0.8124, -0.4122, -0.4122]^\top$. Given that $\ker(C_{\mathrm{f}}) \bigcap \ker(C_{\mathrm{f}}A_{\mathrm{f}}) = 0$, by utilizing Remark 4.3, there exists $\hat{a}_1 = [-0.3764, -0.3349, -0.3349]^\top$ that satisfies $C_{\mathrm{f}}A_{\mathrm{f}}B_{\mathrm{a}}\hat{a}_0 + C_{\mathrm{f}}B_{\mathrm{a}}\hat{a}_1 = 0$ and $C_{\mathrm{f}}A_{\mathrm{f}}^2B_{\mathrm{a}}\hat{a}_0 + C_{\mathrm{f}}A_{\mathrm{f}}B_{\mathrm{a}}\hat{a}_1 = 0$. Consequently, according to Corollary 4.3 and Definition 4.2, adversaries are capable of performing controllable cyber-attacks in the sense of Definition 4.8 and the flight control system Σ_{f} is not left invertible. We set $a_{\mathrm{u}}(0) = \hat{a}_0$ and as per Corollary 4.3 , we design $h(k) = (k+1)^2$ and the actuator attack signal in the following form:

$$a_{\mathrm{u}}(k) = \hat{a}_0 h(k) + \hat{a}_1 h(k-1), \tag{94}$$

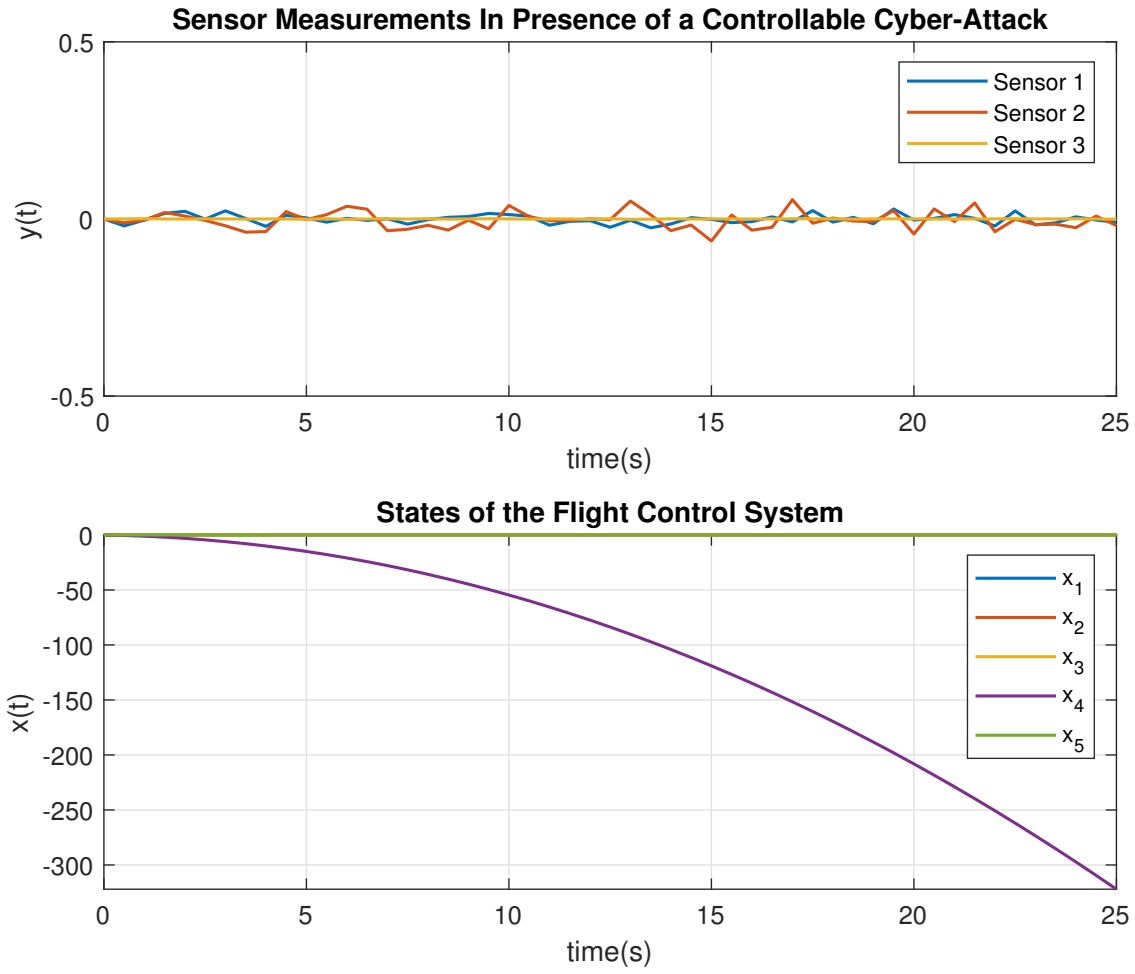

Figure 4.5: Controllable cyber-attacks in the flight control system.

for $k \geq 1$. As shown in Figure 4.5, the sensor measurements of the flight control system in presence of the controllable attacks and noise are close to zero, while the state of the system is growing unbound.

In order to make the flight control system Σ_{f} secure against controllable cyber-attacks in the sense of Definition 4.12, we design an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ with their dynamics given by (82) and (83), respectively. The decoder $\mathcal{D}$ and the encoder $\mathcal{E}$ are designed to satisfy the conditions in Lemma 4.2 such

that $A_{\rm d} = I_4$, $C_{\rm d} = -C_{\rm e} = I_4$, $D_{\rm d} = D_{\rm e}^{-1} = I_4$, $A_{\rm e} = A_{\rm d} - B_{\rm e}C_{\rm d}$, and

$$B_{\rm e} = B_{\rm d} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}.$$

Since the q-th column of $C_{\rm d}B_{\rm d}$ is different from the q-th column of $D_{\rm d} = I_4$, for $q = 1, \ldots, 4$, it follows from Theorem 4.4 that in presence of the dynamic coding scheme, adversaries will not be able to execute controllable cyber-attacks on the flight control system $\Sigma_{\rm f}$. Moreover, as depicted in Figure 4.6, in presence of the proposed dynamic coding scheme, the impact of the controllable cyber-attack given by (94) can now be observed and detected in the sensor measurements.

4.6.3 Covert Attacks in the Flight Control System of a Fighter Aircraft

In this subsection, we study covert cyber-attacks, as per Definition 4.9, in the flight control system of a fighter aircraft.

Executing Covert Attacks (Theorem 4.3): We consider scenarios where each actuator of the system is attacked separately. If the first input channel is compromised by adversaries, i.e., $L_{\rm a} = [1, 0, 0, 0]$, we have $r_{\rm a}^1 = 1$ and $C_{\rm f1}A_{\rm f}^{r_{\rm a}^1-1}B_{\rm a} = [0.1823, 0, 0]^\top$, where by definition $B_{\rm a} = B_{\rm f}L_{\rm a}$. Moreover, according to Definition 4.10, assuming that only the first input communication channel is compromised, a relative degree cannot be defined for the second and the third outputs since $C_qA^{i_q}B_{\rm a} = 0$, for any positive integer i_q and $q = 2$ and 3. Hence, adversaries need to only compromise the first output communication channel to satisfy the condition $C_{\rm f1}A_{\rm f}^{r_{\rm a}^1-1}B_{\rm a} = (D_{\rm a}^*C_{\rm f})_1A_{\rm f}^{r_{\rm a}^q-1}B_{\rm a}$ in Theorem 4.3 and perform a covert cyber-attack, i.e., $D_{\rm a} = [1, 0, 0]^\top$.

Following Corollary 4.4, we design a covert attack signal for the case where only the first input and the first output communication channels of the flight control system are under cyber-attacks. As shown in Figure 4.7, in presence of the system and the process noise, a covert cyber-attack is executed and the sensor measurements are close to zero while the values of states are increasing.

Having either the second or the third or the fourth input communication channels compromised yields a

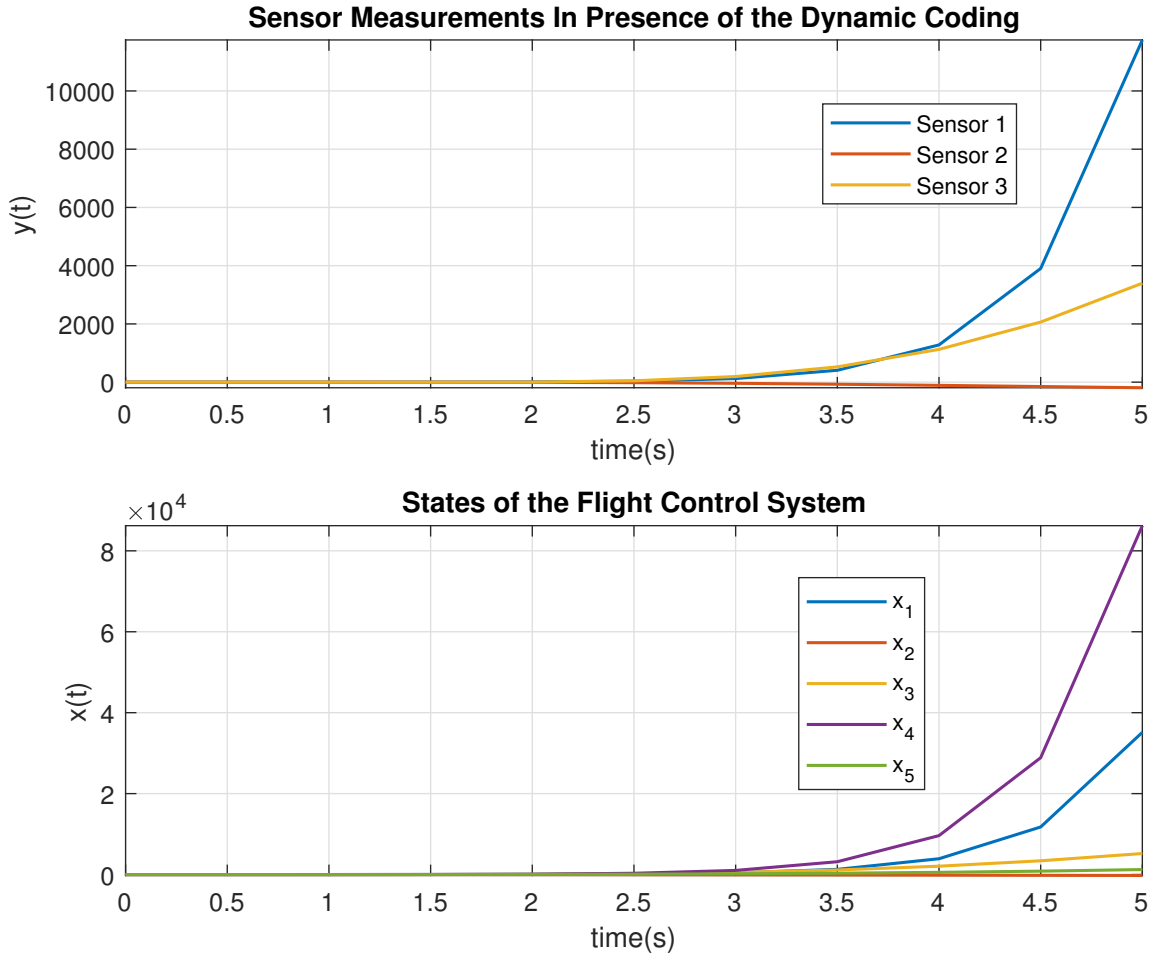

Figure 4.6: Impact of the dynamic coding scheme in securing the flight control system against controllable cyber-attacks.

relative degree equal to 1 for all the outputs, i.e., $r_{\mathrm{a}} = 1$ for $q = 1, 2, 3$ if $L_{\mathrm{a}} = [0, 1, 0, 0]$ or $L_{\mathrm{a}} = [0, 0, 1, 0]$ or $L_{\mathrm{a}} = [0, 0, 0, 1]$. Consequently, in order to execute covert attacks, when any actuator other than the first one is under cyber-attacks, adversaries need to compromise all 3 output communication channels of the system to satisfy the condition $C_{\mathrm{f}q} A_{\mathrm{f}}^{r_{\mathrm{a}}^q - 1} B_{\mathrm{a}} = (D_{\mathrm{a}}^* C_{\mathrm{f}})_q A_{\mathrm{f}}^{r_{\mathrm{a}}^q - 1} B_{\mathrm{a}}$ in Theorem 4.3, for $q = 1, \ldots, 3$.

Calculating the $\bar{SI}_{\mathbf{c}}$ (Definition 4.11, Algorithm 1): We utilize Algorithm 1 to compute an upper bound for the security index for covert attacks in the flight control system. In order to initialize the algorithm, we use the given characteristic matrices of the flight control system and create the set of inputs and outputs $S = \{u_1, \ldots, u_4, y_1, \ldots, y_3\}$. There exists 127 different combinations of members of the set S. Algorithm 1

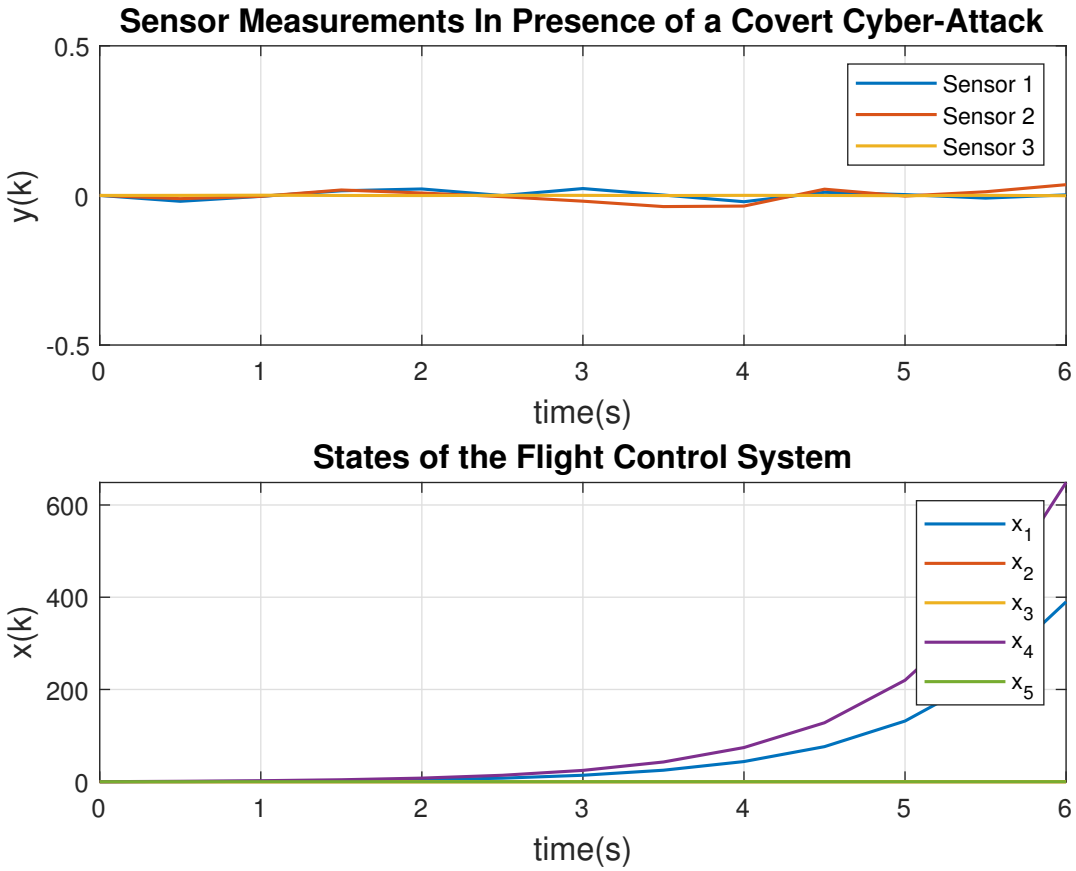

Figure 4.7: Covert attack while the first input and the first output communication channels are compromised.

yields the outputs $\bar{SI}_{\mathrm{c}} = 2$ and $\bar{S}_{\min} = \{u_1, y_1\}$. Hence, the minimum number of actuators and sensors that should be attacked by adversaries to perform a covert attack is 2 which includes the first actuator and the first sensor.

The Coding Scheme (Theorem 4.5): Our objective is to make the flight control system secure against covert cyber-attacks. Under Assumption 4.2, an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ with their dynamics given by (82) and (83) are designed, respectively. As per Assumption 4.4, we consider that the actuator 1 and sensors 1 and 2 are secured, i.e., $q_{s1} = 1$ and $q_{s2} = 2$. Moreover, the decoder $\mathcal{D}$ and the encoder $\mathcal{E}$ satisfy the conditions in Theorem 4.5. The characteristic matrices of the encoder and the decoder are $A_{\mathrm{d}} = I_4$,

$C_{\mathrm{d}} = -D_{\mathrm{d}}$, $B_{\mathrm{d}} = D_{\mathrm{d}}$, $D_{\mathrm{e}} = D_{\mathrm{d}}^{-1}$, $B_{\mathrm{e}} = I_4$, $C_{\mathrm{e}} = -B_{\mathrm{d}}^{-1}B_{\mathrm{e}}C_{\mathrm{d}}$, $A_{\mathrm{e}} = A_{\mathrm{d}} - B_{\mathrm{e}}C_{\mathrm{d}}$, and

$$D_{\mathrm{d}} = \begin{bmatrix} 1 & 0.575 & -0.0026 & 0.574 \\ 0 & 0.7911 & 0.0009 & -0.2085 \\ 0 & -0.2085 & 0.0009 & 0.7918 \\ 0 & 0.0009 & 1 & 0.0009 \end{bmatrix}.$$

Considering that the first and the second sensors are secured, we design the actuator attack signal such that its impact on these two sensors is zero. In order to achieve this goal, we have used the results in [14, 15] to design the actuator attack signal to be a controllable cyber-attack that its impact cannot be observed in sensors 1 and 2. Also, since sensor 3 is not secured and can be manipulated by adversaries, the sensor attack signal is designed to cancel out the impact of the actuator attack signal on this sensor. Consequently, in Figure 4.8, it can be seen that the impact of the actuator attack signal cannot be seen in the sensor readings. As depicted in Figure 4.9, in presence of the proposed dynamic coding scheme, adversaries cannot eliminate the impact of their actuator attack signals from sensor measurements despite knowing the parameters of encoder and the decoder and the attack signal can now be observed in the sensor measurements.

4.7 Conclusion

This chapter has studied the vulnerability of the CPS to zero dynamics attacks, covert attacks, and controllable cyber-attacks. Given that these cyber-attacks are considered to be stealthy, they can cause damage to the CPS without being detected. Under certain assumptions, we have studied and derived conditions for existence of these cyber-attacks in terms of nonzero Markov parameters of the CPS and entries of the observability matrix. Moreover, in addition to providing the number of required actuators to be attacked in case of zero dynamics and controllable attacks, the derived conditions represent the level of required system knowledge to carry out these stealthy cyber-attacks. Furthermore, a dynamic coding scheme was developed to increase the security index of the CPS to its maximum possible value. Hence, in presence of deploying the dynamic coding scheme, if one actuator is secured, adversaries will not be capable of performing the zero dynamics and controllable cyber-attacks. Moreover, three key challenges related to covert cyber-attacks in

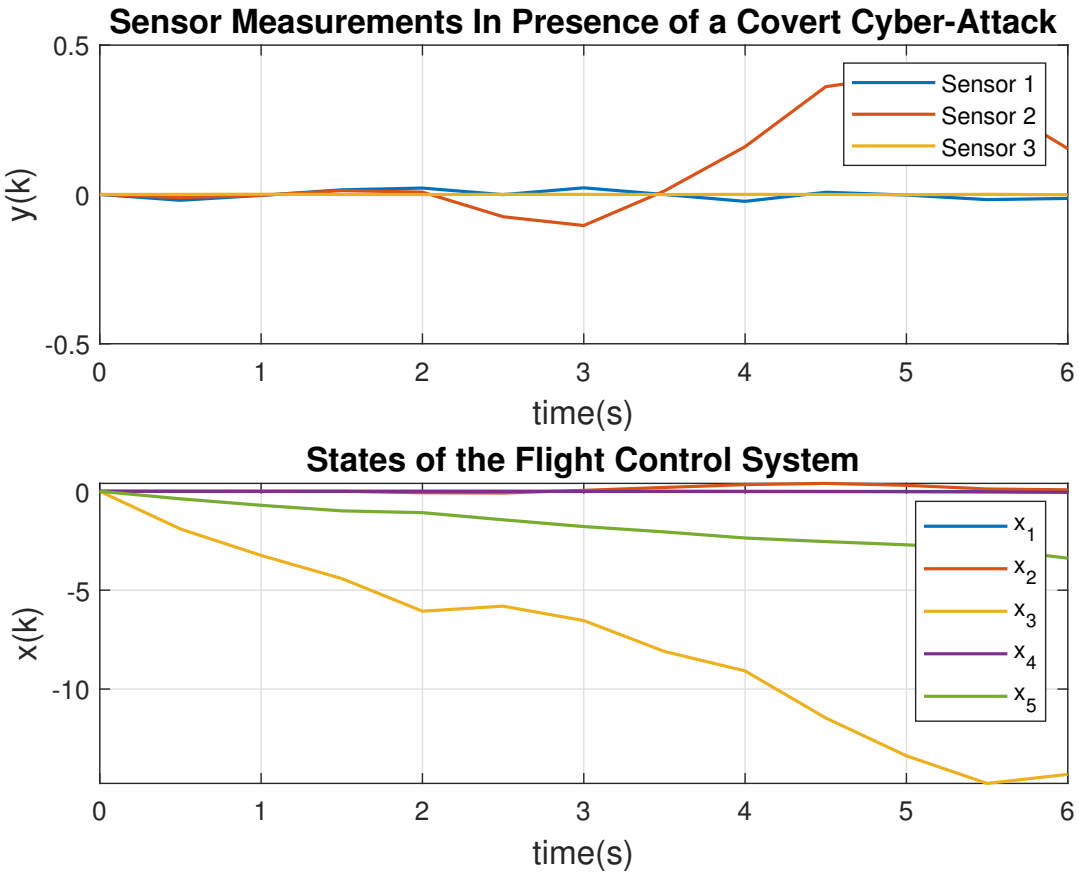

Figure 4.8: Covert attack while actuators 2, 3, and 4 along with sensor 1 are compromised.

the CPS were addressed. We have investigated and formulated necessary and sufficient conditions in the sense of disruption resources of the CPS that adversaries need in order to carry out covert cyber-attacks. These conditions can be employed to determine the input and output communication channels required for executing covert attacks. Furthermore, we have utilized the developed conditions to define an upper bound on the security index (SI) for covert attacks in CPS which determines the minimum number of actuators and sensors that should be attacked to execute a covert cyber-attack. As a countermeasure against covert cyber-attacks, a dynamic coding scheme has been introduced and developed. Under certain conditions and assuming the existence of one secure input and two secure output communication channels, the proposed coding scheme effectively prevents adversaries from executing covert cyber-attacks.

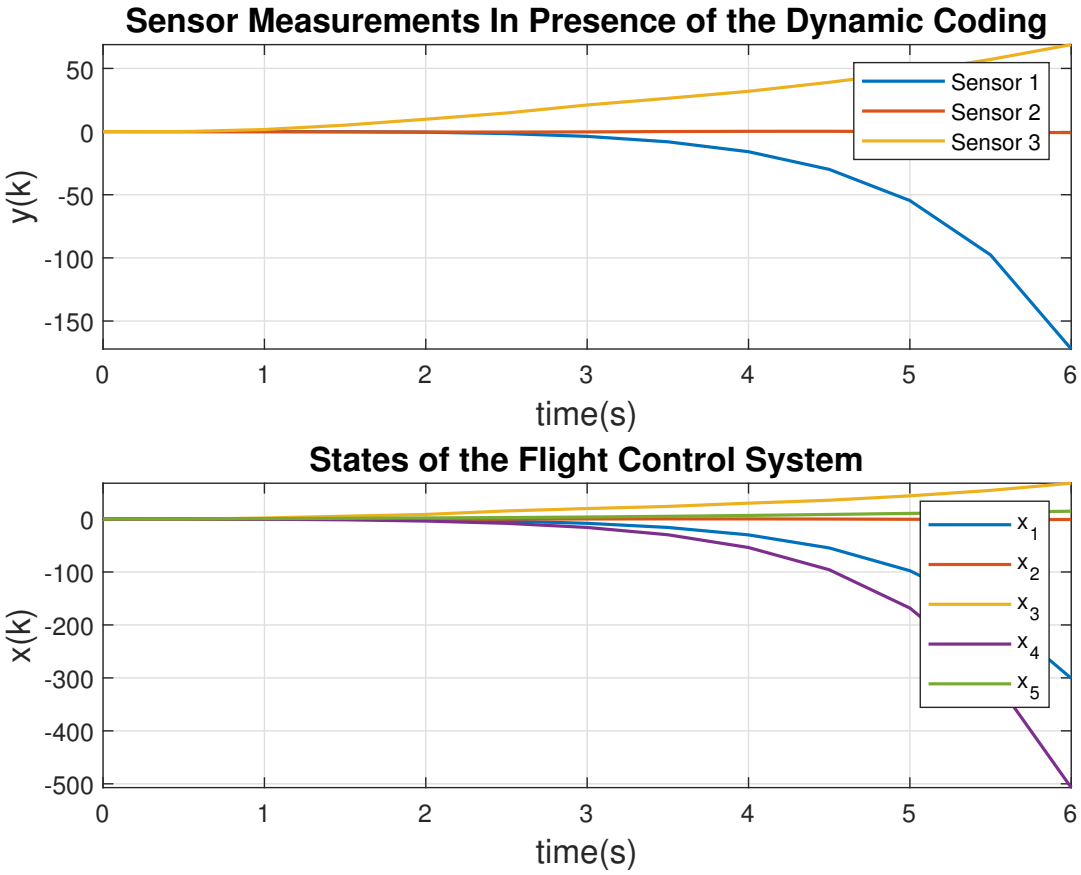

Figure 4.9: Covert attack in presence of the dynamic coding scheme.

Chapter 5

The Security Requirement to Prevent Zero Dynamics Attacks and Perfectly Undetectable Cyber-Attacks in Linear and Nonlinear Cyber-Physical Systems

In this chapter, zero dynamics and perfectly undetectable cyber-attacks in linear and nonlinear cyber-physical systems (CPS) are studied. The impacts of zero dynamics attacks and perfectly undetectable cyber-attacks cannot be observed in outputs of the CPS. Adversaries are capable of executing these cyber-attacks and leading the CPS to undesirable trajectories while remaining undetected. In this chapter, we introduce and formally define the notion of security effort (SE) as a novel security metric for linear CPS that determines the minimum number of actuators and sensors that should be secured and kept attack free in order to prevent adversaries from executing zero dynamics attacks, covert attacks, and controllable attacks. Moreover, since zero dynamics attacks, covert attacks, and controllable attacks belong to weakly unobservable and controllable weakly unobservable subspaces of the CPS, conditions under which these subspaces become zero are obtained and investigated. Subsequently, in this chapter, we study the data-driven implementation of stealthy cyber-attacks for a class of nonlinear CPS. In particular, we consider and study zero dynamics and covert cyber-attacks. By utilizing the Koopman operator theory, a given control affine CPS is transformed into the

Koopman canonical form, and its relative degree is defined in terms of the Koopman modes, Koopman eigenvalues, and Koopman eigenfunctions. Consequently, the relative degree of the CPS is utilized to determine zero dynamics cyber-attacks. In contrast to the linear case, adversaries need to compromise both input and output communication channels of the CPS to maintain their attacks undetected. Moreover, the Koopman canonical form of the CPS is used to define and implement covert cyber-attacks in nonlinear CPS. Hence, by utilizing the Koopman canonical form of the CPS, we find sensors that should be secured to prevent zero dynamics and covert cyber-attacks in nonlinear CPS. The extended dynamic mode decomposition (EDMD) provides a linear finite-dimensional approximation of the CPS. Consequently, approximated dynamics of the CPS are utilized to introduce data-driven zero dynamics and covert cyber-attacks. Finally, numerical case studies are provided to illustrate the effectiveness of our proposed methods. The work presented in this chapter has partly appeared in [121, 122].

To summarize, the main contributions of this chapter are stated as follows:

(1) The notion of SE is formally defined as a measure that denotes the minimum number of actuators and sensors that should be secured to prevent adversaries from executing zero dynamics attacks, covert attacks, and controllable attacks.

(2) Conditions under which the weakly unobservable subspace of CPS becomes zero are developed and investigated. If these conditions are satisfied, no zero dynamics attacks, covert attacks, and controllable attacks can be performed by the adversaries on the CPS.

(3) In order to study perfectly undetectable cyber-attacks, conditions under which the controllable weakly unobservable subspace of CPS becomes zero are investigated. Therefore, under these conditions, adversaries cannot execute perfectly undetectable cyber-attacks, i.e., covert attacks and controllable attacks.

(4) The ϵ-stealthy cyber-attacks in terms of Koopman operator are defined which can be used to categorize various types of cyber-attacks.

(5) A relative degree of the CPS by means of Koopman eigenfunction, Koopman eigenvalue, and Koopman modes is defined. The proposed definition of the relative degree only requires matrix multiplications and is easy to check and verify. Moreover, we use the relative degree to discover internal

dynamics of the CPS.

(6) A method to identify sensor measurements that are needed by adversaries to execute zero dynamics and covert cyber-attacks is developed. Hence, by securing certain sensor measurements, one can prevent the execution of zero dynamics and covert cyber-attacks. Moreover, data-driven strategies for executing and implementing the zero dynamics and covert cyber-attacks by using the KCF of the CPS and the EDMD algorithm are proposed.

The remainder of the chapter is organized as follows. State-space representation of the linear and nonlinear CPS systems along with objectives and the definitions for certain cyber-attacks and the Koopman operator theory are provided in Section 5.1. Conditions under which weakly unobservable and controllable weakly unobservable subspaces of the CPS become zero are investigated in Section 5.2. Moreover, the security effort (SE) for linear CPS is formally in Section 5.3. In Section 5.4, by utilizing the Koopman operator theory, ϵ-stealthy cyber-attacks and methodologies to execute zero dynamics and covert attacks in nonlinear CPS are presented. Furthermore, data-driven implementation of zero dynamics and covert cyber-attacks in nonlinear CPS are discussed in Section 5.5. Numerical case studies are presented in Section 5.6 to illustrate and demonstrate the effectiveness of our proposed methodologies.

5.1 Problem Statement and Formulation

5.1.1 Model of the Linear CPS

We consider a linear time-invariant (LTI) CPS in the following form:

$$\begin{aligned} \dot{x}(t) =& Ax(t) + Bu(t), \\ y(t) =& Cx(t), \end{aligned} \tag{95}$$

where $x(t) \in \mathbb{R}^n$ is the state, $y(t) \in \mathbb{R}^p$ is the output, and $u(t) \in \mathbb{R}^m$ denotes the control input. The characteristic matrices of the system, i.e., (A, B, C), are of appropriate dimensions. We assume that B is an injective map, i.e., B has full column rank, since otherwise, its linearly dependent columns can be removed.

5.1.2 Linear CPS Under Cyber-Attacks

Let $\mathcal{U} = \{u_1, \ldots, u_{\mathrm{m}}\}$ and $\mathcal{Y} = \{y_1, \ldots, y_{\mathrm{p}}\}$ denote the sets of input and output communication channels in the CPS (95) with $|\mathcal{U}| = m$ and $|\mathcal{Y}| = p$, respectively, where $|\cdot|$ denotes the cardinality of a set. Moreover, let $\mathcal{U}_{\mathrm{s}}$ and $\mathcal{Y}_{\mathrm{s}}$ denote the sets of secured input and output channels of the CPS, respectively. Consequently, $\mathcal{U}_{\mathrm{a}} = \mathcal{U}/\mathcal{U}_{\mathrm{s}} = \{u_1^{\mathrm{a}}, \ldots, u_{m_{\mathrm{a}}}^{\mathrm{a}}\}$, with $|\mathcal{U}_{\mathrm{a}}| = m_{\mathrm{a}}$ is the set of attacked inputs and $\mathcal{Y}_{\mathrm{a}} = \mathcal{Y}/\mathcal{Y}_{\mathrm{s}} = \{y_1^{\mathrm{a}}, \ldots, y_{p_{\mathrm{a}}}^{\mathrm{a}}\}$, with $|\mathcal{Y}_{\mathrm{a}}| = p_{\mathrm{a}}$ denotes the set of attacked outputs.

The CPS (95) under cyber-attacks can be expressed by

$$\begin{aligned} \dot{x}(t) =& Ax(t) + B(u(t) + L_{\mathrm{a}} a_{\mathrm{u}}(t)), \\ y(t) =& Cx(t) + D_{\mathrm{a}} a_{\mathrm{y}}(t), \end{aligned} \tag{96}$$

where $a_{\mathrm{u}}(t) \in \mathbb{R}^{m_{\mathrm{a}}}$ is the actuator attack signal and $a_{\mathrm{y}}(t) \in \mathbb{R}^{p}$ is the sensor attack signal. Moreover, $B_{\mathrm{a}} = BL_{\mathrm{a}}$ and D_{a} are the actuator attack and the sensor attack signatures, respectively. The matrix $D_{\mathrm{a}} = \mathrm{diag}(d_1, d_2, \ldots, d_p) \in \mathbb{R}^{p \times p}$ is diagonal, where $d_r = 1$ if the r-th sensor measurement belongs to the set $\mathcal{Y}_{\mathrm{a}}$ for $r = 1, \ldots, p$, and $d_r = 0$ if $y_r \in \mathcal{Y}_{\mathrm{s}}$. Hence, one has $\mathrm{rank}(D_{\mathrm{a}}) = p_{\mathrm{a}}$. Furthermore, the matrix $L_{\mathrm{a}} = [l_{u_1^{\mathrm{a}}}, \ldots, l_{u_{m_{\mathrm{a}}}^{\mathrm{a}}}] \in \mathbb{R}^{m \times m_{\mathrm{a}}}$ denotes the input channels that are compromised by adversaries, where u_q^{a}-th element of $l_{u_q^{\mathrm{a}}} \in \mathbb{R}^{m}$ is equal to 1, and the rest of its entries are zero, for $q = 1, \ldots, m_{\mathrm{a}}$. Consequently, L_{a} is an injective operator, i.e., $\mathrm{rank}(L_{\mathrm{a}}) = m_{\mathrm{a}}$.

5.1.3 Various Types of Cyber-Attacks in the Linear CPS

Given the linearity of the CPS (96), and due to the superposition principle, one can separately consider and study the impact of cyber-attacks and control inputs on the CPS. Hence, we eliminate the effects of $u(t)$ from the CPS in the following form:

$$\begin{aligned} \dot{x}(t) =& Ax(t) + B_{\mathrm{a}} a_{\mathrm{u}}(t), \\ y(t) =& Cx(t) + D_{\mathrm{a}} a_{\mathrm{y}}(t). \end{aligned} \tag{97}$$

Let $Y(x(0), a_{\mathrm{u}}(t), a_{\mathrm{y}}(t))$ denote the output of the CPS (97) as a function of the initial condition $x(0)$,

the actuator attack signal $a_{\mathrm{u}}(t)$, and the sensor attack signal $a_{\mathrm{y}}(t)$, $\forall t \geq 0$. This chapter is concerned with cyber-attacks that their impacts cannot be observed in the output measurements of CPS. In the following, the above types of cyber-attacks are defined according to [2, 123–125].

Definition 5.1. *In the CPS (97), the following cyber-attacks are defined:*

(1) The actuator attack signal $a_u(t) \neq 0$ is a zero dynamics attack if $Y(x(0), a_u(t), 0) = 0$, $\forall t \geq 0$, where $x(0) \neq 0$, and adversaries only need to have access to input communication channels.

(2) The attack signal $a(t) = [a_u(t)^\top, a_y(t)^\top]^\top \neq 0$ is a covert attack if $Y(0, a_u(t), a_y(t)) = 0$, $\forall t \geq 0$.

(3) The actuator attack signal $a_u(t) \neq 0$ is a controllable attack if $Y(0, a_u(t), 0) = 0$, $\forall t \geq 0$, where adversaries need to compromise input communication channels.

Definition 5.2 ([21]). *The cyber-attack signal $a(t) = [a_u(t)^\top, a_y(t)^\top]^\top \neq 0$ is designated as perfectly undetectable if it satisfies $Y(0, a_u(t), a_y(t)) = 0$, $\forall t \geq 0$.*

Consequently, as per Definitions 5.1 and 5.2, there are two types of perfectly undetectable cyber-attacks. First, if one has $a_{\mathrm{u}}(t) \neq 0$ and $a_{\mathrm{y}}(t) \neq 0$ such that $Y(0, a_{\mathrm{u}}(t), a_{\mathrm{y}}(t)) = 0$, $\forall t \geq 0$, this is referred to as a *covert attack* in [2, 96, 123]. Second, if $a_{\mathrm{u}}(t) \neq 0$ and $a_{\mathrm{y}}(t) = 0$ such that $Y(0, a_{\mathrm{u}}(t), 0) = 0$, $\forall t \geq 0$, this is defined as a *controllable attack* in [124, 125], and a *zero stealthy attack* in [126].

However, since the cyber-attack that results in having $Y(0, a_{\mathrm{u}}(t), 0) = 0$ is related to the controllable weakly unobservable subspace of the system (see [125] and [98] for more details), we have adopted the convention from [124] and [125] and refer to this type of perfectly undetectable cyber-attacks as controllable attacks. Moreover, despite the fact that the given zero dynamics attack in Definition 5.1 is not perfectly undetectable (as per Definition 5.2), under certain initial conditions, it results in a zero output. Hence, in this chapter, in addition to perfectly undetectable cyber-attacks, we also investigate zero dynamics attacks.

5.1.4 Overview of Koopman Operator Theory for Nonlinear CPS

In this subsection, by adopting the results in [52–55], the Koopman operator theory for a nonlinear system is briefly studied. Let us consider a nonlinear system expressed by

$$\dot{x} = f(x), \tag{98}$$

where $x \in \mathbb{X} \subset \mathbb{R}^d$ is the state and $f : \mathbb{X} \to \mathbb{X}$ is a nonlinear vector field. Moreover, the flow map $\Phi(t, x_0)$ is the solution to (98) at the initial condition $x_0 \in \mathbb{X}$. Consider $\mathcal{F}$ as the space of all complex-valued scalar functions $\psi : \mathbb{X} \to \mathbb{C}$. Consequently, $\mathcal{K}^t : \mathcal{F} \to \mathcal{F}$ is defined as the Koopman operator which satisfies $(\mathcal{K}^t\psi)(\cdot) = \psi \circ \Phi(t, \cdot)$.

Due to the linearity of the Koopman operator [52, 54], it can be characterized by its eigenvalues and eigenfunctions. The eigenfunction of $\mathcal{K}^t$ can be defined as the function $\phi : \mathbb{X} \to \mathbb{C}$ that satisfies $\mathcal{K}^t\phi = e^{\lambda t}\phi$, where $\lambda \in \mathbb{C}$ is the Koopman eigenvalue. Moreover, it can be shown that the Koopman eigenfunction ϕ satisfies $\mathcal{L}_f\phi = \lambda\phi$ [55], where $\mathcal{L}_f = f.\nabla$ denotes the Lie derivative with respect to f and "$\cdot$" denotes the dot product. Consequently, the time-varying function $\tilde{\Psi}(t, x) = \mathcal{K}^t\psi$ is the solution to $\frac{\partial\tilde{\Psi}}{\partial t} = \mathcal{L}_f\tilde{\Psi}$, where $\tilde{\Psi}(0, x) = \psi(x_0)$.

It should be noted that the Koopman operator is infinite dimensional. Moreover, if ϕ_1 and ϕ_2 are eigenfunctions of the Koopman operator with their corresponding eigenvalues λ_1 and λ_2, respectively, it follows that $\phi_1^q\phi_2^l$ is an eigenfunction with the eigenvalue $q\lambda_1 + l\lambda_2$, for any $q, l \in \mathbb{N}$. In addition to the point spectrum, the Koopman operator could possess both residual and continuous parts of the spectrum [55, 127]. However, this work is restricted to the point spectra of the Koopman operator.

Consider ϕ_i as an eigenfunction of $\mathcal{K}^t$ with λ_i as its eigenvalue. Consequently, the vector-valued observable $r(x) \in \mathcal{F}^p$, where $p \in \mathbb{N}$, can be represented in the following form:

$$r(x) = \sum_{i=1}^{\infty} \phi_i(x) v_i^r, \tag{99}$$

where $v_i^r \in \mathbb{R}^p$, for $i = 1, 2, \ldots$ are the Koopman modes corresponding to $r(x)$ [54, 55]. It should be noted that the Koopman eigenfunctions and eigenvalues depend on the dynamics of the system (98) and the space $\mathcal{F}$. However, the Koopman modes v_i^r are specific to the observable $r(x)$.

5.1.5 Nonlinear CPS Model in the Koopman Canonical Form

Consider the following control affine nonlinear system

$$\dot{x} = f(x) + \sum_{j=1}^{m} g_j(x) u_j, \;\; y = h(x), \tag{100}$$

where $x \in \mathbb{X} \subset \mathbb{R}^d$ is the state, $u_j \in \mathbb{R}$, for $j = 1, 2, \ldots, m$, is the control input, $g_j : \mathbb{X} \to \mathbb{R}^d$ is the control nonlinear coupling term, and $h : \mathbb{X} \to \mathbb{R}^p$ denotes the nonlinear output function.

The following assumption holds throughout the chapter.

Assumption 5.1 ([55]). *There exists a finite subset of Koopman eigenfunctions $\phi_i(x)$, for $i = 1, 2, \ldots, n$ and $n > d$, such that*

$$x = \sum_{i=1}^{n} \phi_i(x) v_i^x, \quad h(x) = \sum_{i=1}^{n} \phi_i(x) v_i^h, \tag{101}$$

where $v_i^x \in \mathbb{C}^d$ and $v_i^h \in \mathbb{C}^p$.

Remark 5.1. *If Assumption 5.1 holds, the state x and $h(x)$ can be represented by a finite subset of Koopman eigenfunctions. However, if $x = \sum_{i=1}^{\infty} \phi_i(x) v_i^x$ and $h(x) = \sum_{i=1}^{\infty} \phi_i(x) v_i^h$, adopting a finite number of Koopman eigenfunctions results in a truncation error in approximating x and $h(x)$. The truncation error can be made arbitrarily small by choosing a greater number of Koopman eigenfunctions. Further discussions on the case where one requires an infinite number of Koopman eigenfunctions in Assumption 5.1 can be found in [55].*

In the following, equation (101) in the Assumption 5.1 is utilized to represent the Koopman Canonical Transform (KCT) and to express the nonlinear system (100) in terms of its Koopman eigenfunctions, Koopman eigenvalues, and Koopman modes.

If $\phi_i(x)$ is a complex-valued Koopman eigenfunction, i.e., $\phi_i(x) : \mathbb{X} \to \mathbb{C}$, we choose $\phi_{i+1}(x)$ as its complex conjugate. Hence, in Assumption 5.1, one has $2n_{\mathrm{c}}$ complex-valued Koopman eigenfunctions, where n_{c} is the number of distinct complex conjugate pairs $(\phi_i(x), \phi_{i+1}(x))$. Without loss of generality, suppose in Assumption 5.1 the first n_{r} Koopman eigenfunctions are real-valued, i.e., $\phi_i(x) : \mathbb{X} \to \mathbb{R}$ for $i = 1, \ldots, n_{\mathrm{r}}$, and the $2n_{\mathrm{c}}$ remaining eigenfunctions are complex-valued such that $n = n_{\mathrm{r}} + 2n_{\mathrm{c}}$. Consider the following change of coordinates [55]:

$$z(t) = \begin{bmatrix} z_1(t) \\ \vdots \\ z_n(t) \end{bmatrix} = T(x(t)), \tag{102}$$

where $T(x) = [\hat{\phi}_1(x), \ldots, \hat{\phi}_n(x)]^\top$, $\hat{\phi}_{i_{\mathrm{r}}}(x) = \phi_{i_{\mathrm{r}}}(x)$ for $i_{\mathrm{r}} = 1, \ldots, n_{\mathrm{r}}$, and $[\hat{\phi}_{n_{\mathrm{r}}+2i_{\mathrm{c}}-1}(x), \hat{\phi}_{n_{\mathrm{r}}+2i_{\mathrm{c}}}(x)]^\top =$

$[2\mathrm{Re}(\phi_{n_\mathrm{r}+2i_\mathrm{c}-1}(x)), -2\mathrm{Im}(\phi_{n_\mathrm{r}+2i_\mathrm{c}}(x))]^\top$ for $i_\mathrm{c} = 1, \ldots, n_\mathrm{c}$. Moreover, $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ denote the real and imaginary parts of a complex number, respectively.

The Lie derivative of $T(x)$ yields

$$\mathcal{L}_f T(x) = \Lambda T(x), \tag{103}$$

where $\Lambda \in \mathbb{R}^{n\times n}$ is a block diagonal matrix such that its i_r-th diagonal entry is $\Lambda_{i_\mathrm{r},i_\mathrm{r}} = \lambda_{i_\mathrm{r}}$ for $i_\mathrm{r} = 1, \ldots, n_\mathrm{r}$, and Λ has a block diagonal entry in the following form

$$\begin{bmatrix} \Lambda_{\hat{i},\hat{i}} & \Lambda_{\hat{i},\hat{i}+1} \\ \Lambda_{\hat{i}+1,\hat{i}} & \Lambda_{\hat{i}+1,\hat{i}+1} \end{bmatrix} = |\lambda| \begin{bmatrix} \cos(\arg\lambda_{\hat{i}}) & \sin(\arg\lambda_{\hat{i}}) \\ -\sin(\arg\lambda_{\hat{i}}) & \cos(\arg\lambda_{\hat{i}}) \end{bmatrix},$$

for $\hat{i} = n_\mathrm{r} + 2i_\mathrm{c} - 1$ and $i_\mathrm{c} = 1, \ldots, n_\mathrm{c}$, where $\arg\lambda_{\hat{i}}$ denotes the argument of $\lambda_{\hat{i}}$.

Consequently, through the change of coordinate (102), equation (100) is transformed into its Koopman canonical form (KCF) as follows:

$$\begin{cases} \dot{z} = & \Lambda z + \sum_{j=1}^{m} \tilde{g}_j(z) u_j, \\ y = & C^h z, \\ x = & C^x z, \end{cases} \tag{104}$$

where $C^x = [\tilde{v}_1^x \cdots \tilde{v}_n^x]$, $C^h = [\tilde{v}_1^h \cdots \tilde{v}_n^h]$, and $\tilde{g}_j(z) = \mathcal{L}_{g_j} T(x)|_{x=C^x z}$. Moreover, $\tilde{v}_{i_\mathrm{r}}^h = v_{i_\mathrm{r}}^h$ for $i_\mathrm{r} = 1, \ldots, n_\mathrm{r}$, and $[\tilde{v}_{\hat{i}}^h,\ \tilde{v}_{\hat{i}+1}^h] = [\mathrm{Re}(v_{\hat{i}}^h),\ \mathrm{Im}(v_{\hat{i}}^h)]$ for $\hat{i} = n_\mathrm{r} + 2i_\mathrm{c} - 1$ and $i_\mathrm{c} = 1, \ldots, n_\mathrm{c}$. Also, C^x has a similar structure as C^h.

5.1.6 Model of the Control Affine Nonlinear CPS Under Cyber-Attacks

In this chapter, we consider CPS under actuator and sensor cyber-attacks as depicted in Figure 5.1. The nonlinear CPS (100) under cyber-attacks can be expressed in the following form:

$$\begin{cases} \dot{x}^* = & f(x^*) + \sum_{j=1}^{m} g_j(x^*)(u_j + a_\mathrm{u}^j), \\ y^* = & h(x^*) + a_\mathrm{y}, \end{cases} \tag{105}$$

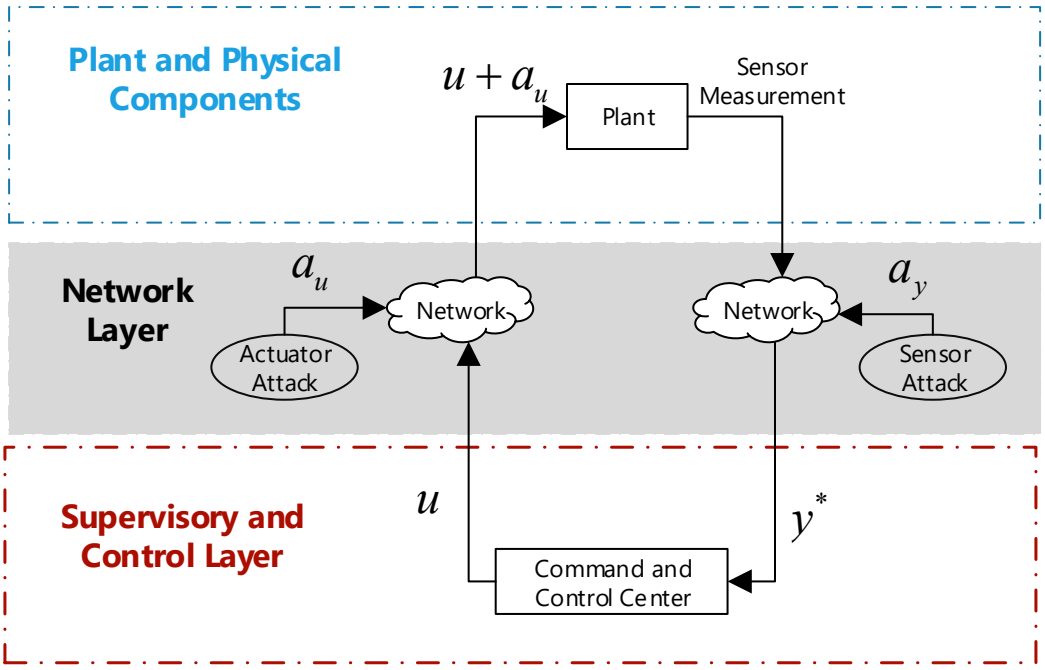

Figure 5.1: The CPS framework under cyber-attacks.

where $x^* \in \mathbb{X}$ is the state of the CPS in the presence of cyber-attacks, $y^* \in \mathbb{R}^p$ is the manipulated sensor measurements by adversaries, $a_{\mathrm{u}}^j \in \mathbb{R}$ is the actuator cyber-attack on the j-th input, and $a_{\mathrm{y}} \in \mathbb{R}^p$ is the sensor cyber-attack signal. Under the attack free conditions, i.e., $a_{\mathrm{u}} = 0$ and $a_{\mathrm{y}} = 0$, one has $x^*(t) = x(t)$, $\forall t \geq 0$. Consequently, the Koopman canonical form (104) of the CPS in presence of the actuator and sensor cyber-attacks is given by

$$\begin{cases} \dot{z}^* = & \Lambda z^* + \sum_{j=1}^{m} \tilde{g}_j(z^*)(u_j + a_{\mathrm{u}}^j), \\ y^* = & C^h z^* + a_{\mathrm{y}}, \end{cases} \tag{106}$$

where $x^* = C^x z^*$.

5.1.7 Objectives

We have five objectives in this chapter. Our first objective is to develop and study conditions under which adversaries cannot perform zero dynamics attacks, covert attacks, and controllable attacks that are provided in Definition 5.1. The latter is achieved by studying conditions under which the largest weakly unobservable and the largest controllable weakly unobservable subspace of the linear CPS are zero. As for our second objective, we formally define a security measure that determines the minimum number of input and output communication channels that should be secured in order to prevent adversaries from performing certain cyber-attacks that are provided in Definitions 5.1 and 5.2. Moreover, the proposed security measure is studied from a geometric control perspective. Our third objective is to formally define a measure of stealthiness for cyber-attacks in the nonlinear control affine CPS (105). The stated measure which is defined by means of the Koopman eigenfunctions and Koopman modes of the CPS can be used to categorize various types of cyber-attacks based on their levels of detectability. The fourth objective is to utilize the Koopman eigenfunctions, Koopman eigenvalues, and Koopman modes of the CPS (100) which are used in (104) to define a relative degree for nonlinear systems. By using the defined relative degree and the normal form representation of the CPS, one can discover its internal dynamics, i.e., the zero dynamics. Our fifth objective is to utilize the KCF of the CPS which can be approximated by means of the EDMD algorithm to propose strategies for executing zero dynamics and covert cyber-attacks in nonlinear CPS systems and to find sensor measurements that should be secured to prevent adversaries from performing zero dynamics and covert attacks.

5.2 Investigation of Weakly Unobservable and Controllable Subspaces for Linear CPS

In case of the covert attacks, adversaries design their sensor attack signals such that they cancel out the impact of actuator attacks from sensor readings [96]. Hence, the sensor attack signal is designed in the

following from:

$$\begin{aligned}\dot{x}_{\rm a}(t) =& Ax_{\rm a}(t) + B_{\rm a}a_{\rm u}(t),\\ y_{\rm a}(t) =& -Cx_{\rm a}(t),\end{aligned} \tag{107}$$

where $x_{\rm a}(t) \in \mathbb{R}^n$ and $a_{\rm y}(t) = y_{\rm a}(t)$. One can augment the dynamics in (97) and (107) into the following form:

$$\begin{aligned}\dot{\check{x}}(t) =& \check{A}\check{x}(t) + \check{B}_{\rm a}a_{\rm u}(t),\\ y(t) =& \check{C}\check{x}(t),\end{aligned} \tag{108}$$

where $\check{x}(t) = [x(t)^\top, x_{\rm a}(t)^\top]^\top$, $y(t) = Cx(t) + D_{\rm a}y_{\rm a}(t)$, $\check{A} = \text{diag}(A, A)$, $\check{B}_{\rm a} = [B_{\rm a}^\top, B_{\rm a}^\top]^\top$, and $\check{C} = [C, -D_{\rm a}C]$. In terms of the main advantage of the augmented system (108), only the actuator attack signal is an input to the system, and the sensor attack signal $a_{\rm y}(t)$ is expressed by using the dynamics given by (107). Let $\check{Y}(\check{x}(0), a_{\rm u}(t))$ represent the output of the augmented system (108) as a function of the initial condition $\check{x}(0)$ and the actuator attack signal $a_{\rm u}(t)$. In the following, it is shown how one can utilize the augmented system (108) in order to study cyber-attacks on the CPS (97). In particular, it is shown that covert attacks, controllable attacks, and zero dynamics attacks in CPS (97) can be equivalently studied in the augmented system (108).

Theorem 5.1. *In the augmented dynamics* (108)*, one has* $\check{Y}(\check{x}(0), a_u(t)) = 0$ *if and only if there exists a sensor attack signal* $a_y(t) \in \mathbb{R}^p$ *and* $\check{x}(0) = [x(0)^\top, x(0)^\top]^\top$ *such that* $Y(x(0), a_u(t), a_y(t)) = 0$ *holds true,* $\forall t \geq 0$.

Proof. **Necessary Condition:** Suppose $\check{Y}(\check{x}(0), a_{\rm u}(t)) = 0$ holds and for any $a_{\rm y}(t) \in \mathbb{R}^p$, one has $Y(x(0), a_{\rm u}(t), a_{\rm y}(t)) \neq 0$, where $\check{x}(0) = [x(0)^\top, x(0)^\top]^\top$. It follows from $\check{Y}(\check{x}(0), a_{\rm u}(t)) = 0$ that $y(t) = Cx(t) + D_{\rm a}y_{\rm a}(t) = 0$, $\forall t \geq 0$. Since $Y(x(0), a_{\rm u}(t), a_{\rm y}(t)) \neq 0$, from (97), one obtains $y(t) = Cx(t) + D_{\rm a}a_{\rm y}(t) \neq 0$. However, considering $a_{\rm y}(t) = y_{\rm a}(t)$ results in having $y(t) = Cx(t) + D_{\rm a}a_{\rm y}(t) = 0$, which contradicts the assumption.

Moreover, suppose $\check{Y}(\check{x}(0), a_{\rm u}(t)) = 0$ and $Y(x(0), a_{\rm u}(t), a_{\rm y}(t)) = 0$, where $\check{x}(0) = [x(0)^\top, x_{\rm a}(0)^\top]^\top$

such that $x(0) \neq x_{\mathrm{a}}(0)$. According to the definition of the augmented system (108), having $\check{Y}(\check{x}(0), a_{\mathrm{u}}(t)) = 0$ implies that one either has $Cx(t) = -D_{\mathrm{a}}y_{\mathrm{a}}(t) = D_{\mathrm{a}}Cx_{\mathrm{a}}(t)$ for $Cx(t) \neq 0$ or in the other case, $Cx(t) = 0$ and $D_{\mathrm{a}}y_{\mathrm{a}}(t) = D_{\mathrm{a}}Cx_{\mathrm{a}}(t) = 0$. Since in both (107) and (108) the input is $a_{\mathrm{u}}(t)$, one should have $x(0) = x_{\mathrm{a}}(0)$ for either case of $Cx(t) = D_{\mathrm{a}}Cx_{\mathrm{a}}(t) \neq 0$ or $Cx(t) = 0$ and $D_{\mathrm{a}}Cx_{\mathrm{a}}(t) = 0$ to hold, which contradicts the assumption.

Sufficient Condition: Assume that there exists a sensor attack signal $a_{\mathrm{y}}(t) \in \mathbb{R}^p$ such that $Y(x(0), a_{\mathrm{u}}(t), a_{\mathrm{y}}(t)) = 0$ and $\check{x}(0) = [x(0)^\top,\, x(0)^\top]^\top$. Moreover, due to the linearity of (108), one obtains

$$\check{Y}(\check{x}(0), a_{\mathrm{u}}(t)) = Y(x(0), a_{\mathrm{u}}(t), 0) - D_{\mathrm{a}}Y(x(0), a_{\mathrm{u}}(t), 0). \tag{109}$$

If $a(t) = [a_{\mathrm{u}}(t)^\top, a_{\mathrm{y}}(t)^\top]^\top$ is either a zero dynamics attack or a controllable attack, as per Definition 5.1, one has $a_{\mathrm{y}}(t) = 0$ and $Y(x(0), a_{\mathrm{u}}(t), 0) = 0$. Consequently, it follows from (109) that $\check{Y}(\check{x}(0), a_{\mathrm{u}}(t)) = 0, \forall t \geq 0$. Also, if $a(t) = [a_{\mathrm{u}}(t)^\top, a_{\mathrm{y}}(t)^\top]^\top$ is a covert attack, according to Definition 5.1, $Y(0, a_{\mathrm{u}}(t), a_{\mathrm{y}}(t)) = 0$ holds. Consequently, due to the definition of (107), one obtains $a_{\mathrm{y}}(t) = y_{\mathrm{a}}(t)$ and $y(t) = Cx(t) + D_{\mathrm{a}}y_{\mathrm{a}}(t) = 0, \forall t \geq 0$. This completes the proof of the theorem. □

Remark 5.2. *As the main implication of the Theorem 5.1, $a(t) = [a_u(t)^\top, a_y(t)^\top]^\top$ in the CPS (97) results in $Y(x(0), a_u(t), a_y(t)) = 0$ if and only if $\check{Y}(\check{x}(0), a_u(t)) = 0$, where $\check{x}(0) = [x(0)^\top,\, x(0)^\top]^\top$. Hence, if no zero dynamics attacks, covert attacks, and controllable attacks can be executed in the augmented system (108), the above cyber-attacks cannot be performed on the CPS (97) as well.*

5.2.1 Cyber-Attacks and the Weakly Unobservable Subspace

Considering Theorem 5.1 and Remark 5.2, in order to study zero dynamics attacks, covert attacks, and controllable attacks in the CPS (97), one can study these cyber-attacks in the augmented system (108). Hence, in the following, we study and derive conditions under which zero dynamics attacks, covert attacks, and controllable attacks cannot be performed in (108) from a geometric control theory perspective.

Definition 5.3 (Weakly Unobservable Subspace)**.** *Consider the triple $(\check{C}, \check{A}, \check{B}_a)$ in (108). A point $\check{x}(0) \in \mathbb{R}^{2n}$ is defined as weakly unobservable if there exists $a_u(t) \neq 0$ such that the output satisfies $y(t) = 0$, $\forall\, t \geq 0$. The set of all weakly unobservable points is called weakly unobservable subspace and is denoted*

by $\mathscr{V}$. Moreover, the largest weakly unobservable subspace is denoted by $\mathscr{V}^*$.

Given the Definition 5.3, and considering the results in [126, Theorem 1] and [124, Lemma 7], if $\mathscr{V}^* = 0$, no zero dynamics attacks, covert attacks, and controllable attacks can be executed in the augmented system (108). In the following, conditions under which $\mathscr{V}^* = 0$ are studied and proposed.

Definition 5.4 ([128]). *Consider $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Z} \subseteq \mathbb{R}^p$ as finite-dimensional inner product vector spaces and the matrix $Q \in \mathbb{R}^{p\times n}$. One has*

(1) $\mathcal{Z} = Q\mathcal{X} \coloneqq \{z : z = Qx,\ x \in \mathcal{X}\}$.

(2) $\mathcal{X} = Q^{-1}\mathcal{Z} \coloneqq \{x : z = Qx,\ z \in \mathcal{Z}\}$.

Theorem 5.2. *Let $Im(\check{B}_a) \neq 0$. In the augmented system (108), one has $\mathscr{V}^* = 0$ if for any $\mathscr{S} \subseteq Ker(\check{C})$, one has $\check{A}Ker(\check{C}) \cap (\mathscr{S} + Im(\check{B}_a)) = 0$.*

Proof. As described in [128, Algorithm 4.1.2] and [98], the largest weakly unobservable subspace of the system (108) can be computed in $2n$ steps by using the following algorithm:

$$\begin{aligned} \mathscr{V}_0 =& \text{Ker}(\check{C}), \\ \mathscr{V}_k =& \mathscr{V}_0 \cap \check{A}^{-1}(\mathscr{V}_{k-1} + \text{Im}(\check{B}_\text{a})). \end{aligned} \tag{110}$$

Since for any $\mathscr{S} \subseteq \text{Ker}(\check{C})$, one has $\check{A}\text{Ker}(\check{C}) \cap (\mathscr{S} + \text{Im}(\check{B}_\text{a})) = 0$, any $g \in \text{Ker}(\check{C})$ results in having $\check{A}g = z \notin \mathscr{S} + \text{Im}(\check{B}_\text{a})$. Hence, we have $\check{A}^{-1}(\mathscr{S} + \text{Im}(\check{B}_\text{a})) \cap \text{Ker}(\check{C}) = 0$, since otherwise, as per Definition 5.4, there exists $g \in \text{Ker}(\check{C})$ such that $\check{A}g = z \in \mathscr{S} + \text{Im}(\check{B}_\text{a})$, which contradicts having $\check{A}\text{Ker}(\check{C}) \cap (\mathscr{S} + \text{Im}(\check{B}_\text{a})) = 0$. Consequently, according to (110), $\mathscr{V}^* = \mathscr{V}_{2n} = 0$ since for any $\mathscr{V}_{2n-1} \subseteq \ker(\check{C})$, we have $\mathscr{V}_0 \cap \check{A}^{-1}(\mathscr{V}_{2n-1} + \text{Im}(\check{B}_\text{a})) = 0$. This completes the proof of the theorem. □

5.2.2 Perfectly Undetectable Cyber-Attacks and the Controllable Weakly Unobservable Subspace

In this subsection, conditions under which in the augmented system (108) perfectly undetectable cyber-attacks, i.e., covert and controllable attacks, cannot be performed are investigated. However, one needs to first study the following definitions.

Definition 5.5 (Strongly Reachable Subspace [98]). *The subspace $\mathscr{W} \subseteq \mathbb{R}^{2n}$ is the strongly reachable subspace of the triple $(\check{C}, \check{A}, \check{B}_a)$ in (108) if $\mathscr{V} = \mathscr{W}^{\perp}$ is the weakly unobservable subspace of $(\check{B}_a^{\top}, \check{A}^{\top}, \check{C}^{\top})$. Moreover, $\mathscr{W}^*$ denotes the smallest strongly reachable subspace.*

Definition 5.6 (Controllable Weakly Unobservable [98]). *The subspace $\mathscr{R} \subseteq \mathscr{V}^*$ is designated as the controllable weakly unobservable subspace of the triple $(\check{C}, \check{A}, \check{B}_a)$ if one has $\mathscr{R} \subseteq \mathscr{W}^*$. Moreover, $\mathscr{R}^* = \mathscr{V}^* \cap \mathscr{W}^*$ denotes the largest controllable weakly unobservable subspace.*

Definition 5.7 (Left-Invertibility [98]). *Let $\check{x}(0) = 0$. The augmented system $(\check{C}, \check{A}, \check{B}_a)$ in (108) is left-invertible if for any $y(t) = 0$ one has $a_u(t) = 0$, $\forall t \geq 0$.*

Lemma 5.1 ([98]). *Let $\check{\Sigma} = (\check{C}, \check{A}, \check{B}_a)$ denote the system in (108). The following statements are equivalent:*

(1) The system $\check{\Sigma}$ is left-invertible.

(2) $\mathscr{R}^* = 0$ and $\check{B}_a$ is injective.

(3) $\check{\mathscr{V}}^* \cap Im(\check{B}_a) = 0$ and $\check{B}_a$ is injective.

As shown in [125, Theorem 1], covert attacks and controllable attacks are related to the controllable weakly unobservable subspace of the system $(\check{C}, \check{A}, \check{B}_{\mathrm{a}})$, i.e., the subspace $\mathscr{R}^*$. Hence, since $\check{B}_{\mathrm{a}}$ is an injective map by definition, from Lemma 5.1 it follows that adversaries are capable of executing perfectly undetectable cyber-attacks if and only if for the triple $(\check{C}, \check{A}, \check{B}_{\mathrm{a}})$ one has $\mathscr{R}^* \neq 0$, or equivalently, the triple $(\check{C}, \check{A}, \check{B}_{\mathrm{a}})$ is not left-invertible.

Theorem 5.3. *The system $\check{\Sigma} = (\check{C}, \check{A}, \check{B}_a)$ in (108) is left-invertible if for any $\mathcal{S} \subseteq Ker(\check{C})$, one has $\check{A}(Im(\check{B}_a) \cap Ker(\check{C})) \cap (\mathcal{S} + Im(\check{B}_a)) = 0$.*

Proof. Considering Lemma 5.1 and since B_{a} has full column rank by definition, one needs to show that having $\check{A}(\mathrm{Im}(\check{B}_{\mathrm{a}}) \cap \mathrm{Ker}(\check{C})) \cap (\mathcal{S} + \mathrm{Im}(\check{B}_{\mathrm{a}})) = 0$ for every $\mathcal{S} \subseteq \mathrm{Ker}(\check{C})$ results in $\mathscr{V}^* \cap \mathrm{Im}(\check{B}_{\mathrm{a}}) = 0$.

Having $\check{A}(\mathrm{Im}(\check{B}_{\mathrm{a}}) \cap \mathrm{Ker}(\check{C})) \cap (\mathcal{S} + \mathrm{Im}(\check{B}_{\mathrm{a}})) = 0$ implies that for any $g \in \mathrm{Im}(\check{B}_{\mathrm{a}}) \cap \mathrm{Ker}(\check{C})$, one has $\check{A}g \notin \mathcal{S} + \mathrm{Im}(\check{B}_{\mathrm{a}})$. Hence, for any $\mathcal{S} \subseteq \mathrm{Ker}(\check{C})$, we have $\mathrm{Im}(\check{B}_{\mathrm{a}}) \cap \mathrm{Ker}(\check{C}) \cap \check{A}^{-1}(\mathcal{S} + \mathrm{Im}(\check{B}_{\mathrm{a}})) = 0$. Since

for $\mathscr{V}_{2n-1} \subseteq \ker(\check{C})$, we have $\mathscr{V}_0 \cap \mathrm{Im}(\check{B}_\mathrm{a}) \cap \check{A}^{-1}(\mathscr{V}_{2n-1} + \mathrm{Im}(\check{B}_\mathrm{a})) = 0$, one obtains

$$\mathscr{V}_{2n} \cap \mathrm{Im}(\check{B}_\mathrm{a}) = \mathscr{V}_0 \cap \mathrm{Im}(\check{B}_\mathrm{a}) \cap \check{A}^{-1}(\mathscr{V}_{2n-1} + \mathrm{Im}(\check{B}_\mathrm{a})) = 0.$$

This completes the proof of the theorem. □

5.3 Security Effort for Linear CPS

In this section, the security effort (SE) is formally defined as a measure that shows the minimum number of input and output communication channels that should be secured by CPS operators and should be kept attack free to prevent adversaries from executing cyber-attacks that are provided in Definitions 5.1 and 5.2. Specifically, it is shown how one can study the SE for a given CPS from a geometric control theory perspective.

5.3.1 Definition of the Security Effort (SE)

The SE is defined as the solution to the following optimization problem:

$$\begin{aligned} SE_\Sigma \coloneqq \min_{a_\mathrm{u}(\cdot), a_\mathrm{y}(\cdot)} \quad & m - \|a_\mathrm{u}(t)\|_0 + p - \|a_\mathrm{y}(t)\|_0 \\ \text{s.t.} \quad & \dot{x}(t) = Ax(t) + Ba_\mathrm{u}(t), \\ & y(t) = Cx(t) + a_\mathrm{y}(t), \\ & y(t) \neq 0,\ x(0) \in \mathbb{R}^n, \\ & a(t) \neq 0, \end{aligned} \tag{111}$$

where $a(t) = [a_\mathrm{u}(t)^\top, a_\mathrm{y}(t)^\top]^\top$.

If conditions in (111) are satisfied, adversaries cannot design a cyber-attack signal $a(t)$ that results in $Y(x(0), a_\mathrm{u}(t), a_\mathrm{y}(t)) = 0, \forall t \geq 0$. In other words, in problem (111), SE_Σ denotes the minimum number of actuators and sensors that should be secured so that the weakly unobservable subspace of the CPS in (97) is empty, and consequently, no zero dynamics attacks, covert attacks, and controllable attacks can be initiated and executed.

Considering that in order to perform zero dynamics cyber-attacks and perfectly undetectable cyber-attacks, i.e., covert attacks and controllable attacks, adversaries need to have access to at least one input communication channel and actuator, one has $0 < SE_{\Sigma} \leq m$. This implies that in the worst-case scenario, the CPS operators need to secure all the input communication channels to prevent zero dynamics attacks and perfectly undetectable cyber-attacks. However, similar to the problem of computing the security index in [21], computing SE_{Σ} is an NP-hard problem, which makes it computationally intensive to solve.

It follows from Theorem 5.1 and Definition 5.3 that if the weakly unobservable subspace of the augmented system (108) is zero, i.e., $\mathscr{V}^* = 0$, no zero dynamics attacks, covert attacks, and controllable attacks can be executed in both the CPS (97) and the augmented system (108). Consequently, according to Theorem 5.2, $\mathscr{V}^* = 0$ if for any $\mathscr{S} \subseteq \text{Ker}(\check{C})$, one has $\check{A}\text{Ker}(\check{C}) \cap (\mathscr{S} + \text{Im}(\check{B}_{\text{a}})) = 0$.

Consequently, an upper bound for the SE in problem (111) can be given in the following form:

$$\begin{aligned} \bar{SE}_{\Sigma} \coloneqq \min_{\text{rank}(B_{\text{a}}),\, \text{rank}(D_{\text{a}})} \quad & m - \text{rank}(B_{\text{a}}) + p - \text{rank}(D_{\text{a}}) \\ \text{s.t.} \quad & \check{A}\text{Ker}(\check{C}) \cap (\text{Ker}(\check{C}) + \text{Im}(\check{B}_{\text{a}})) = 0. \end{aligned} \tag{112}$$

Consequently, in Algorithm 2, a pseudo code for finding an upper bound on SE_{Σ} is proposed. Let $S = \{u_1, \ldots, u_{\text{m}}, y_1, \ldots, y_{\text{p}}\}$ denote the set of all actuators and sensors of the CPS as described in Section 5.1.2. In Algorithm 2, by utilizing the binary representation of the elements of S, its power set is created. Consequently, the sufficient condition $\check{A}\text{Ker}(\check{C}) \cap (\text{Ker}(\check{C}) + \text{Im}(\check{B}_{\text{a}})) = 0$ is considered for each subset of the power set to check if $\mathscr{V}^* = 0$ is satisfied and to compute $\bar{SE}_{\Sigma}$ as an upper bound for the SE, i.e., $SE_{\Sigma} \leq \bar{SE}_{\Sigma}$. Moreover, one of the outputs of Algorithm 2 is the set $\bar{S}_{\text{min}}$ which contains the actuators and sensors that should be secured to prevent adversaries from performing zero dynamics and perfectly undetectable cyber-attacks in the CPS, where $|\hat{S}_{\text{min}}| = \bar{SE}_{\Sigma}$.

5.3.2 Security Effort (SE) for Perfectly Undetectable Cyber-Attacks

The specified SE in optimization problems (111) and (112) are defined to prevent all zero dynamics attacks, covert attacks, and controllable attacks that belong to the weakly unobservable subspace of the CPS. However, as per Definition 5.1, in contrast to zero dynamics attacks, the execution of covert attacks and controllable attacks does not depend on the initial conditions $x(0)$ of the CPS. Moreover, according

Algorithm 2 Pseudo code to find an upper bound for SE_Σ

Input: $\check{A} = \text{diag}(A, A)$, $\check{B}_\text{a} = [B_\text{a}^\top,\ B_\text{a}^\top]^\top$, and $\check{C} = [C,\ -D_\text{a}C]$, $S = \{u_1,\ \ldots,\ u_\text{m},\ y_1,\ \ldots,\ y_\text{p}\}$
Output: $\bar{SE}_\Sigma$, $\bar{S}_\text{min}$

```
Initialize SĒ_s = m + p
Set l = |S|, where |·| denotes the cardinality of a set
for i = 1 : 2^l − 1 do
    Create the empty set Ŝ = {}
    for j = 1 : l do
        if the j-th bit of the binary representation of i is equal to 1 then
            Add j-th member of S to Ŝ
        end if
    end for
    Secure only actuators and sensors that belong to the set Ŝ, update B̌_a and Č accordingly, and set
    Q = ker(Č)
    if ĂQ ∩ (ker(Č) + Im(B̌_a)) = 0 and |Ŝ| ≤ SĒ_s then
        SĒ_s = |Ŝ|
        Ŝ* = Ŝ
    end if
end for
SĒ_Σ = min{SĒ_s, m} and S̄_min = Ŝ*
```

to Definition 5.2, covert attacks and controllable attacks are perfectly undetectable cyber-attacks. Hence, one may only be interested in preventing perfectly undetectable cyber-attacks in the CPS (97). Thus, in the following, SE for perfectly undetectable cyber-attacks is formally defined and investigated.

The SE for perfectly undetectable cyber-attacks in the CPS can be expressed as

$$\begin{aligned} \hat{SE}_\Sigma \coloneqq \min_{a_\text{u}(\cdot), a_\text{y}(\cdot)} \quad & m - \|a_\text{u}(t)\|_0 + p - \|a_\text{y}(t)\|_0 \\ \text{s.t.} \quad & \dot{x}(t) = Ax(t) + Ba_\text{u}(t), \\ & y(t) = Cx(t) + a_\text{y}(t), \\ & y(t) \neq 0,\ x(0) = 0, \\ & a(t) \neq 0, \end{aligned} \tag{113}$$

The only difference between SE_Σ in (111) and $\hat{SE}_\Sigma$ in (113) is that in (113) one has $x(0) = 0$, which implies that zero dynamics attacks are excluded in computing the $\hat{SE}_\Sigma$. Furthermore, given that $\mathscr{R}^* \subseteq \mathscr{V}^*$, one has $\hat{SE}_\Sigma \leq SE_\Sigma$. Moreover, according to Theorem 5.1, if conditions in (113) hold, no covert attacks

and controllable attacks can be executed in the CPS (97) and the augmented system (108). Also, as per Definition 5.7, the augmented system (108) is left-invertible. Hence, it can be inferred that the optimization problem (113) determines the minimum number of input and output communication channels that should be secured to make the CPS (97) left-invertible. Similar to the case of SE_Σ in (111), (113) is also an NP-hard problem to be solved.

In order to compute the upper bound of the SE for perfectly undetectable cyber-attacks, which is designated as $\hat{SE}_\Sigma$, results in Theorem 5.3 are utilized. It follows from Theorem 5.3 that $\check{\Sigma} = (\check{C}, \check{A}, \check{B}_\mathrm{a})$ is left-invertible if for any $\mathcal{S} \subseteq \mathrm{Ker}(\check{C})$, one has $\check{A}(\mathrm{Im}(\check{B}_\mathrm{a}) \cap \mathrm{Ker}(\check{C})) \cap (\mathcal{S} + \mathrm{Im}(\check{B}_\mathrm{a})) = 0$. Hence, the problem of computing $\hat{SE}_\Sigma$ can be rewritten in the following form:

$$\begin{aligned} \hat{SE}_\Sigma \coloneqq \min_{\mathrm{rank}(B_\mathrm{a}),\, \mathrm{rank}(D_\mathrm{a})} \quad & m - \mathrm{rank}(B_\mathrm{a}) + p - \mathrm{rank}(D_\mathrm{a}) \\ \text{s.t.} \quad & \check{A}(\mathrm{Im}(\check{B}_\mathrm{a}) \cap \mathrm{Ker}(\check{C})) \cap (\mathrm{Ker}(\check{C}) + \mathrm{Im}(\check{B}_\mathrm{a})) = 0. \end{aligned} \tag{114}$$

Therefore, one can modify Algorithm 2 to determine an upper bound for $\hat{SE}_\Sigma$, i.e., $\hat{SE}_\Sigma$. In order to compute $\hat{SE}_\Sigma$, one needs to set $\mathcal{Q} = \mathrm{Im}(\check{B}_\mathrm{a}) \cap \mathrm{Ker}(\check{C})$ in steps 10 of Algorithm 2. Moreover, the output of the algorithm is $\hat{SE}_\Sigma = \min\{\bar{SE}_\mathrm{s}, m\}$.

5.4 ϵ-Stealthy Cyber-Attacks in the Sense of Koopman Operator for Nonlinear CPS

The main focus of this section is to propose a data-driven method for implementation of two types of stealthy cyber-attacks, namely the zero dynamics and covert cyber-attacks, for the nonlinear CPS (105). Thus, similar to the case of linear CPS [2, 17], one needs to formally define ϵ-stealthy cyber-attacks for the nonlinear CPS.

Under Assumption 5.1, equations (100) and (104) yield

$$y(t) = h(x(t)) = \sum_{i=1}^{n} \hat{\phi}_i(x(t)) \tilde{v}_i^h. \tag{115}$$

Moreover, in presence of cyber-attacks, from equations (105) and (106) it follows that:

$$y^*(t) = \sum_{i=1}^{n} \hat{\phi}_i(x^*(t))\tilde{v}_i^h + a_{\mathrm{y}}. \tag{116}$$

The following assumption holds throughout the chapter.

Assumption 5.2. *The difference between the outputs in equations* (115) *and* (116) *is bounded such that* $\|y(t) - y^*(t)\| \leq c$, *where* $c > 0$.

Remark 5.3. *As per Assumption 5.2, in stealthy cyber-attacks, adversaries are assumed to design their attack signals such that* $y(t) - y^*(t)$ *remains bounded.*

Definition 5.8 (ϵ-Stealthy Cyber-Attacks)**.** *Under the Assumptions 5.1 and 5.2, a cyber-attack on the CPS* (105) *is* ϵ*-stealthy if*

$$\|\sum_{i=1}^{n}(\hat{\phi}_i(x(t)) - \hat{\phi}_i(x^*(t)))\tilde{v}_i^h - a_y\|_\infty \leq \epsilon,\ \forall t \geq 0 \tag{117}$$

where $\|\cdot\|_\infty$ *denotes the infinity norm, i.e., supremum norm. Moreover, a cyber-attack is designated as perfectly undetectable if it is* 0*-stealthy.*

5.4.1 Zero Dynamics of the Nonlinear CPS in the Sense of the Koopman Operator

In this subsection, by means of the Koopman eigenfunctions, Koopman eigenvalues, and Koopman modes, a relative degree is defined for the nonlinear CPS (100). Subsequently, internal dynamics, i.e., the zero dynamics, of the CPS are characterized.

Definition 5.9 (Relative Degree [129])**.** *Let* $m \geq p$. *The system* (100) *has a vector relative degree* $\{r_1, \ldots, r_p\}$ *at* $x = x_0$ *if*

(1) $\mathcal{L}_{g_j}\mathcal{L}_f^k h_q(x_0) = 0$, *for* $j = 1, \ldots, m$ *and* $k < r_q - 1$, *where* $h_q(x)$ *is the* q*-th entry (or component) of* $h(x)$, *for* $q = 1, \ldots, p$, *and*

(2) the following matrix has full row rank at $x = x_0$:

$$M(x) = \begin{bmatrix} \mathcal{L}_{g_1}\mathcal{L}_f^{r_1-1}h_1(x) & \cdots & \mathcal{L}_{g_m}\mathcal{L}_f^{r_1-1}h_1(x) \\ \mathcal{L}_{g_1}\mathcal{L}_f^{r_2-1}h_2(x) & \cdots & \mathcal{L}_{g_m}\mathcal{L}_f^{r_2-1}h_2(x) \\ \cdots & \cdots & \cdots \\ \mathcal{L}_{g_1}\mathcal{L}_f^{r_p-1}h_p(x) & \cdots & \mathcal{L}_{g_m}\mathcal{L}_f^{r_p-1}h_p(x) \end{bmatrix}.$$

Assumption 5.3 ([129]). *In the CPS (100), one has $m \geq p$ and the system has the vector relative degree $\{r_1, \ldots, r_p\}$ at $x = x_0$ as stated in Definition 5.9.*

Lemma 5.2. *For the CPS (100) and the transformed dynamics (104), one has $\mathcal{L}_{g_j}\mathcal{L}_f^i h_q(x) = C_q^h \Lambda^i \tilde{g}_j(z)$, for $x = C^x z$ and any integer $i \geq 0$, where C_q^h is the q-th row of C^h. Moreover, under Assumptions 5.1 and 5.3, the vector relative degree of the nonlinear CPS (100) in the sense of Definition 5.9 is equal to the vector relative degree of the transformed system (104) at each point $x_0 = C^x z_0$.*

Proof. The proof is omitted due to space limitations. □

Remark 5.4. *It follows from Lemma 5.2 that, at any point in the state-space, one can define the relative degree for the nonlinear CPS (100) in terms of the spectral properties of the Koopman operator.*

Remark 5.5. *Considering that $n > d$, from Lemma 5.2 one can conclude that when compared to (100), the Koopman canonical transformation in (102) has resulted in having $n - d$ additional zero dynamics or, equivalently, internal dynamics in the transformed system (104). Moreover, since C^x by definition is a surjective map, but not an injective one, it can be inferred that the null space of C^x having a dimension of $n - d$ has contributed to having the additional internal dynamics.*

Let us define $u = [u_1, \ldots, u_m]^\top \in \mathbb{U} \subset \mathbb{R}^m$, where $\mathbb{U}$ is the space of admissible control inputs. From (115) it follows that the zero dynamics of the CPS (100) are excited if there exists a certain $x_0 \neq 0$ such that u satisfies

$$[\tilde{v}_1^h \cdots \tilde{v}_n^h] \begin{bmatrix} \hat{\phi}_1(x(t)) \\ \vdots \\ \hat{\phi}_n(x(t)) \end{bmatrix} = 0, \ \forall t \geq 0. \tag{118}$$

However, it would be challenging to directly determine the pair (x_0, u) from the Koopman eigenfunctions, Koopman eigenvalues, and Koopman modes of the CPS in (104) to satisfy (118).

Consequently, in the following, the normal form (in the sense of [129]) for the CPS in its KCF (104) is derived to study the zero dynamics of the system. Let Assumption 5.3 hold and $r = r_1 + \cdots + r_p$. If $r < n$, one can define $\zeta_{i_q}^q = C_q^h \Lambda^{i_q - 1} z$, for every $i_q = 1, \ldots, r_q$ and $q = 1, \ldots, p$. Moreover, we choose $\eta_{i_\eta} = l_{i_\eta}(z) \in \mathbb{R}$, for $i_\eta = 1, \ldots, n - r$. The function $l_{i_\eta}(z)$ should be chosen such that the Jacobian matrix of the following map is nonsingular at $z = z_0$:

$$U(z) = \text{col}(C_1^h z, \ldots, C_1^h \Lambda^{r_1} z, \ldots, C_p^h \Lambda^{r_p - 1} z, l_1(z), \ldots, l_{n-r}(z)), \tag{119}$$

where $\text{col}(\cdot)$ denotes the column vector.

Consequently, the dynamics of (104) under new coordinates can be expressed by

$$\begin{aligned}
\dot{\zeta}_1^q =& \zeta_2^q, \\
\cdots \quad & \cdots \\
\dot{\zeta}_{r_q - 1}^q =& \zeta_{r_q}^q, \\
\dot{\zeta}_{r_q}^q =& a_q(\zeta, \eta) + \sum_{j=1}^{m} b_{qj}(\zeta, \eta) u_j, \\
\dot{\eta} =& Q(\zeta, \eta) + P(\zeta, \eta) u,
\end{aligned} \tag{120}$$

for $q = 1, \ldots, p$, where $\zeta = [\zeta_1^1, \ldots, \zeta_{r_1}^1, \ldots, \zeta_{r_p}^p]^\top$, $\eta = [\eta_1, \ldots, \eta_{n-r}]^\top$, $y = [\zeta_1^1, \ldots, \zeta_1^p]^\top$, $a_q(\zeta, \eta) = C_q^h \Lambda^{r_q} U^{-1}(\eta, \zeta)$, $b_{qj}(\zeta, \eta) = C_q^h \Lambda^{r_q - 1} \tilde{g}_j(U^{-1}(\zeta, \eta))$, and $Q(\zeta, \eta)$ and $P(\zeta, \eta)$ depend on the choice of $l_{i_\eta}(z)$.

Proposition 5.1. *Let Assumptions 5.1 and 5.3 hold. For the CPS (100) and its KCF (104), one has $y \equiv 0$ if $z_0 \in \bigcap_{k=1}^{r_q - 1} \ker(C_q^h \Lambda^k)$ such that $x_0 = C^x z_0 \neq 0$, for $q = 1, \ldots, p$, and $u = -[\hat{M}(U^{-1}(0, \eta))]^\dagger a(U^{-1}(0,$*

$\eta))$, *where*

$$\hat{M}(z) = \begin{bmatrix} C_1^h \Lambda^{r_1-1} \tilde{g}_1(z) & \cdots & C_1^h \Lambda^{r_1-1} \tilde{g}_m(z) \\ C_2^h \Lambda^{r_2-1} \tilde{g}_1(z) & \cdots & C_2^h \Lambda^{r_2-1} \tilde{g}_m(z) \\ \cdots & \cdots & \cdots \\ C_p^h \Lambda^{r_p-1} \tilde{g}_1(z) & \cdots & C_p^h \Lambda^{r_p-1} \tilde{g}_m(z) \end{bmatrix}, \; a(z) = \begin{bmatrix} C_1^h \Lambda^{r_1} z \\ C_2^h \Lambda^{r_2} z \\ \cdots \\ C_p^h \Lambda^{r_p} z \end{bmatrix},$$

and $[\cdot]^\dagger$ *denotes a pseudoinverse of the matrix.*

Proof. The proof is omitted due to space limitations. □

Remark 5.6. *In view of the Proposition 5.1, the internal dynamics, i.e., the zero dynamics, of the CPS (100) can be obtained and characterized by* $\dot{\eta} = Q(0,\eta) - P(0,\eta)[\hat{M}(U^{-1}(0,\eta))]^{-1} b(U^{-1}(0,\eta))$. *Also, the stated internal dynamics can be either stable or unstable.*

5.4.2 Zero Dynamics Cyber-Attacks in the Nonlinear CPS

In light of the zero dynamics of the CPS (100) that is discussed in the previous subsection, we now investigate a method that may be employed by adversaries to carry out a zero dynamics cyber-attack and exploit this type of vulnerabilities in nonlinear CPS. It is assumed that the adversaries have two main objectives. The first objective is to maintain their cyber-attacks stealthy in the sense of Definition 5.8. Their second objective is to cause the maximum possible damage to the CPS, which in the case of zero dynamics cyber-attacks, it can be achieved if the internal dynamics of the CPS (provided in Remark 5.6) are unstable.

Since the superposition principle does not hold for nonlinear systems, adversaries need to consider the impact of the control input of the CPS on the stealthiness of their cyber-attacks. Hence, in order to perform the zero dynamics cyber-attacks in the CPS (105) and its KCF (106), adversaries may eliminate the impact of the control input and design their actuator attack signals as follows

$$a_{\mathrm{u}} = -u - [\hat{M}(U^{-1}(0,\eta))]^{-1} b(U^{-1}(0,\eta)), \tag{121}$$

where $a_{\mathrm{u}} = [a_{\mathrm{u}}^1, \ldots, a_{\mathrm{u}}^m]^\top$. Moreover, if the initial conditions of the CPS satisfy the hypothesis of the Proposition 5.1, one obtains $y^* = a_y$ as the output of the CPS (105). Furthermore, in order to increase the

stealthiness level of the cyber-attack in (121), i.e., to decrease ϵ in (117), adversaries can use the dynamics of the KCF (104) and design their sensor attack signals in the following form:

$$\dot{\hat{z}} = \Lambda \hat{z} + \sum_{j=1}^{m} \tilde{g}_j(\hat{z}) u_j, \;\; \hat{y} = C^h \hat{z}, \tag{122}$$

where $a_{\rm y} = \hat{y}$ and $\hat{z} \in \mathbb{R}^n$.

Remark 5.7. *It should be noted that without performing the sensor cyber-attacks, i.e., $a_{\rm y} = 0$, since $\hat{\phi}_i(x^*(t))\tilde{v}_i^h = 0$, the actuator cyber-attack (121) will be ϵ-stealthy in the sense of Definition 5.8, where ϵ satisfies $\| \sum_{i=1}^{n} \hat{\phi}_i(x(t))\tilde{v}_i^h \|_\infty \leq \epsilon$, $\forall t \geq 0$. Hence, as opposed to the zero dynamics cyber-attacks in linear systems (see [24] for more details), in the case of zero dynamics cyber-attacks in nonlinear CPS, adversaries should perform the sensor cyber-attack $a_{\rm y} = \hat{y}$ to decrease ϵ.*

If the initial condition in (122) is set to $\hat{z}(0) = z(0)$, according to Definition 5.8, the zero dynamics cyber-attack which is carried out by utilizing $a_{\rm u}$ in (121) and the $a_{\rm y}$ given by (122) will be 0-stealthy, i.e., perfectly undetectable. Moreover, in terms of the stealthiness of the cyber-attack as per Definition 5.8, if $\hat{z}(0) \neq z(0)$, one obtains

$$\| \sum_{i=1}^{n} (\hat{\phi}_i(x(t)) - \hat{\phi}_i(\hat{x}(t)))\tilde{v}_i^h \|_\infty \leq \hat{\epsilon}, \; \forall t \geq 0 \tag{123}$$

where $\hat{x}(t) = C^x \hat{z}(t)$ and $\hat{\epsilon} > 0$.

Remark 5.8. *As a countermeasure against zero dynamic attacks, one may investigate adding sensors or modifying $y = h(x)$ such that internal dynamics of the CPS in its KCF (104) are stable. The latter can be achieved by studying Koopman eigenfunctions, Koopman eigenvalues, and Koopman modes and the transformation $z(t) = T(x)$, however, it is not within the scope of this chapter and is not addressed here.*

5.4.3 Covert Cyber-Attacks in the Nonlinear CPS

In the case of covert cyber-attacks for linear systems, adversaries compromise both input and output communications of the CPS [2, 96]. Moreover, actuator cyber-attack signals can be designed arbitrarily, whereas sensor cyber-attack signals are designed to eliminate the impact of actuator attacks from sensor measurements.

In order to execute a covert cyber-attack in the CPS (100), adversaries can design the actuator cyber-attack signal a_{u} to achieve their malicious goals and objectives. Consequently, the sensor cyber-attack signal is designed as $a_{\mathrm{y}} = -h(x^*) + \hat{y}$, where $\hat{y}$ is given by (122). Similar to the zero dynamics cyber-attacks, covert cyber-attacks will be 0-stealthy if $\hat{z}(0) = z(0)$, and as shown in (123), $\hat{\epsilon}$-stealthy, otherwise.

Remark 5.9. *From the proposed zero dynamics cyber-attacks and covert cyber-attacks in Sections 5.4.2 and 5.4.3, respectively, it can be concluded that the zero dynamics cyber-attacks are a special case of covert cyber-attacks. To be more precise, the proposed zero dynamics cyber-attack in Section 5.4.2 is a covert cyber-attack in which actuator cyber-attack signals are designed according to the Proposition 5.1.*

Corollary 5.1. *Let the q-th sensor measurement of the CPS (100) be secured such that $h_{-q}(x) = \sum_{i=1}^{n} \phi_i^q(x) \times v_i^{h_{-q}}$, where $h_{-q}(x)$ contains all the entries of $h(x)$ except for the q-th one. If $x \neq \sum_{i=1}^{n} \phi_i^q(x) v_i^x$, adversaries will not be able to perform zero dynamics and covert cyber-attacks in the CPS.*

Proof. Since the q-th sensor is secured, one has $x \neq \sum_{i=1}^{n} \phi_i^q(x) v_i^x$. Consequently, adversaries will not be able to eavesdrop all the necessary sensor measurements to develop an accurate model of (122) for performing stealthy cyber-attacks. It should be noted that the decrease in accuracy of the dynamics in (122) results in a lower level of stealthiness, i.e., a greater ϵ in (117). This completes the proof of the corollary. □

Remark 5.10. *One can utilize Corollary 5.1 to study the importance of each sensor measurement in transforming the CPS (100) into its KCF (104). The latter can be carried out by removing the q-th sensor measurement and investigating whether $h_{-q}(x) = \sum_{i=1}^{n} \phi_i^q(x) v_i^{h_{-q}}$ satisfies $x = \sum_{i=1}^{n} \phi_i^q(x) v_i^x$ (refer to the Assumption 5.1). Hence, if having the q-th sensor measurement is necessary for $x = \sum_{i=1}^{n} \phi_i^q(x) v_i^x$ to hold, as a countermeasure to covert and zero dynamics cyber-attacks, one should protect the privacy of the communication channels that correspond to the q-th sensor measurement.*

5.5 Data-Driven Approximation of the Dynamics and Cyber-Attacks in the Nonlinear CPS

As discussed in Sections 5.4.2 and 5.4.3, in the implementation of the proposed zero dynamics and covert cyber-attacks, it is assumed that adversaries know the characteristic dynamics of the KCF of the CPS (104).

However, this assumption may not be satisfied. Hence, in this section, a linear discrete-time approximation of the KCF of the CPS in (104) is provided by utilizing the EDMD algorithm [56–58, 130]. Consequently, the approximated data-driven representation of the KCF is used to implement the zero dynamics and covert cyber-attacks.

We assume that the adversaries have access to the following snapshot data matrices:

$$\tilde{X} = -[\gamma_1, \dots, \gamma_k],\ \tilde{Y} = [\gamma_1^+, \dots, \gamma_k^+],\ \tilde{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_k],$$

where $\gamma_i = [y_{i,n_d}^\top\ \bar{u}_{i,n_d-1}^\top\ y_{i,n_d-1}^\top \cdots \bar{u}_{i,0}^\top\ y_{i,0}^\top]^\top$, $\gamma_i^+ = [y_{i,n_d+1}^\top\ \bar{u}_{i,n_d}^\top\ y_{i,n_d}^\top \cdots \bar{u}_{i,1}^\top\ y_{i,1}^\top]^\top$, $\boldsymbol{u}_i = \bar{u}_{i,n_d}$, and $y_{i,j}$ is the vector of sensor measurements in the CPS (100) that have been resulted from the vector of control inputs $\bar{u}_{i,\hat{j}}$, for $j = 0, \dots, n_d+1$, and $\hat{j} = 0, \dots, n_d$. Moreover, n_d is the number of control inputs that are captured over different time windows to construct the above snapshot data matrices.

Following [56], let

$$(\hat{A}, \hat{B}) \in \arg\min_{A,B} \|Y_{\text{lift}} - AX_{\text{lift}} - B\tilde{U}\|_{\text{F}},$$

where $X_{\text{lift}} = [\Psi(\gamma_1), \dots, \Psi(\gamma_k)]$, $Y_{\text{lift}} = [\Psi(\gamma_1^+), \dots, \Psi(\gamma_k^+)]$, and $\Psi(x) = [\psi_1(x), \dots, \psi_n(x)]^\top$ is a given basis of the nonlinear lifting functions, and $\|\cdot\|_{\text{F}}$ denotes the Frobenius norm. Furthermore,

$$\hat{C} \in \arg\min_{C} \|Y - CX_{\text{lift}}\|_{\text{F}},$$

where $Y = [y_{1,n_d}, \dots, y_{k,n_d}]$.

Consequently, adversaries can obtain a linear approximation of the dynamics in (122) as expressed below

$$\hat{z}_{k+1} = \hat{A}\hat{z}_k + \hat{B}u_k,\ \ \hat{y}_k = \hat{C}\hat{z}_k, \tag{124}$$

where $\hat{z}_k \in \mathbb{R}^n$ and u_k is the sampled control input.

Hence, in the case of zero dynamics cyber-attacks, the actuator attack signal should be set to $a_k^{\text{u}} = -u_k + u_k^{\text{a}}$, where the initial conditions of the CPS should satisfy the conditions in the Proposition 5.1. Moreover, $u_k^{\text{a}} = b_0\sigma^k \neq 0$, where $b_0 \in \mathbb{R}^m$ and $\sigma \in \mathbb{R}$ are the input-zero direction and the zero dynamics of the triple $(\hat{C}, \hat{A}, \hat{B})$, respectively (see Chapter 1 in [97]). Also, since the proposed a_k^{u} excites the zero dynamics of a

linear approximation of the CPS (104), i.e., a_k^{u} excites the zero dynamics of the triple $(\hat{C}, \hat{A}, \hat{B})$, the sensor cyber-attack signal is designed as $a_k^{\mathrm{y}} = -y_k + \hat{y}_k$ to increase the stealthiness level of the cyber-attack, where y_k is the sampled output measurement of the CPS. As for the case of covert cyber-attacks, the actuator attack a_k^{u} can be any arbitrary signal and the sensor cyber-attack should be set to $a_k^{\mathrm{y}} = -y_k + \hat{y}_k$.

5.6 Numerical Case Studies

5.6.1 Linear CPS: the Quadruple-Tank Process

In this case study, we compute the SE by using (112) and (114) for a Quadruple-Tank Process (QTP) with a non-minimum phase zero. The characteristic matrices of the QTP are expressed as follows [104]:

$$A = \begin{bmatrix} -0.0158 & 0 & 0.0256 & 0 \\ 0 & -0.0109 & 0 & 0.0178 \\ 0 & 0 & -0.0256 & 0 \\ 0 & 0 & 0 & -0.0178 \end{bmatrix},$$
$$B = \begin{bmatrix} 0.0482 & 0 \\ 0 & 0.0349 \\ 0 & 0.0775 \\ 0.0559 & 0 \end{bmatrix}, C = \begin{bmatrix} 0.5 & 0 & 0 & 0 \\ 0. & 0.5 & 0 & 0 \end{bmatrix}. \tag{125}$$

The QTP in (125) is left-invertible, which as per Definition 5.7 it implies that no controllable attack can be performed on it. However, it is vulnerable to zero dynamics attacks and covert attacks. In the case where there exists no secure input and output communication channel, i.e., $B_{\mathrm{a}} = B$ and $D_{\mathrm{a}} = I_p$, the adversaries can execute both zero dynamics attacks and covert attacks as shown in Figures 5.2 and 5.3, respectively.

If only one actuator is secured, the adversaries cannot execute zero dynamics attacks, but they are still capable of performing covert attacks. Consequently, securing the first actuator and the first sensor will result in $\mathscr{V}^* = 0$. Hence, the SE for the QTP is $SE_\Sigma = 2$.

In order to prevent perfectly undetectable cyber-attacks, i.e., covert attacks and controllable attacks, one needs to compute $\hat{SE}_\Sigma$ given by (114). Consequently, $\hat{SE}_\Sigma = 2$ and securing the first actuator and the first

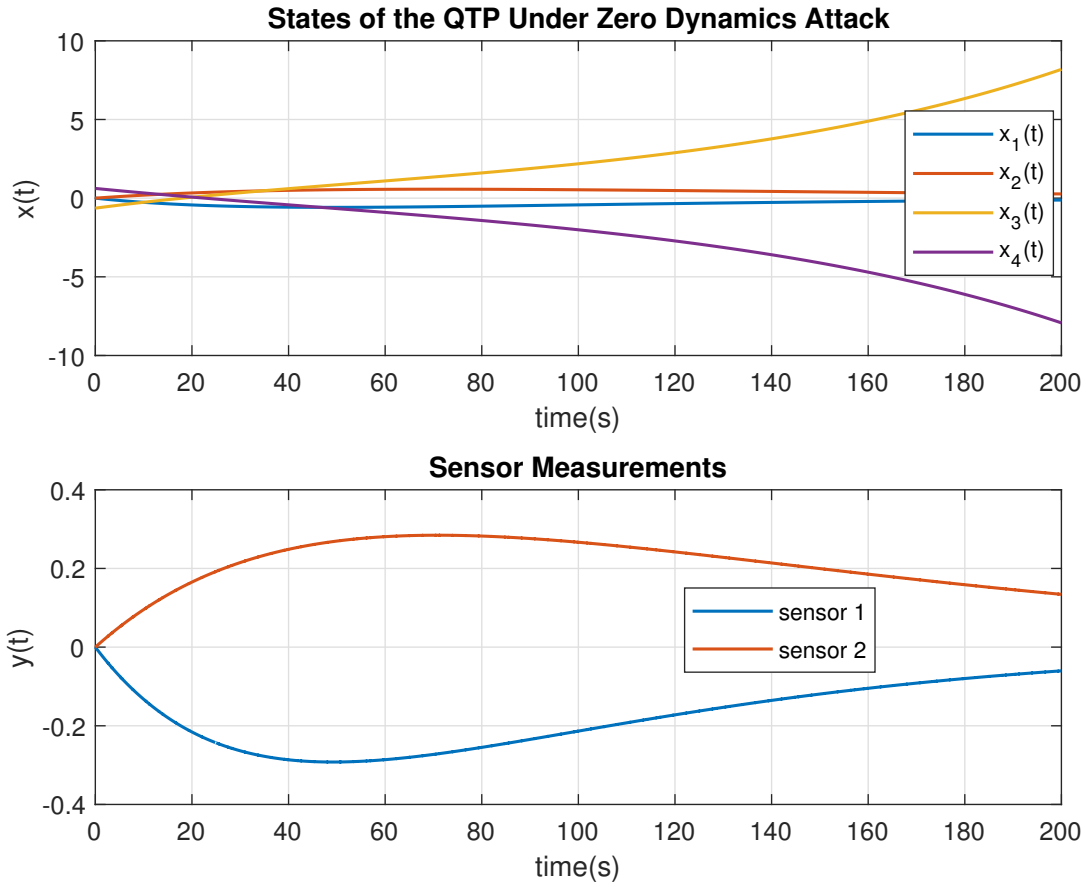

Figure 5.2: The QTP under zero dynamics attacks where the states become unbounded while the outputs show an attack-free behavior.

sensor results in having $\mathrm{Im}(\check{B}_{\mathrm{a}}) \cap \mathscr{V}^* = 0$. Thus, having one secure input and one secure output communication channel prevents adversaries from executing perfectly undetectable cyber-attacks in the QTP. As seen from Figure 5.4, once the first actuator and the first sensor are secured, the adversaries cannot perform covert attacks.

As it was mentioned earlier, adversaries need to compromise both input channels to perform zero dynamics attacks in the QTP. Moreover, as for the case of covert attacks, adversaries need to have access to at least 2 input and 1 output communication channels. Hence, if one considers both zero dynamics and perfectly undetectable cyber-attacks, the security index for the QTP is equal to 2. However, the system operators need to secure 1 input and 1 output communication channel to prevent both zero dynamics and

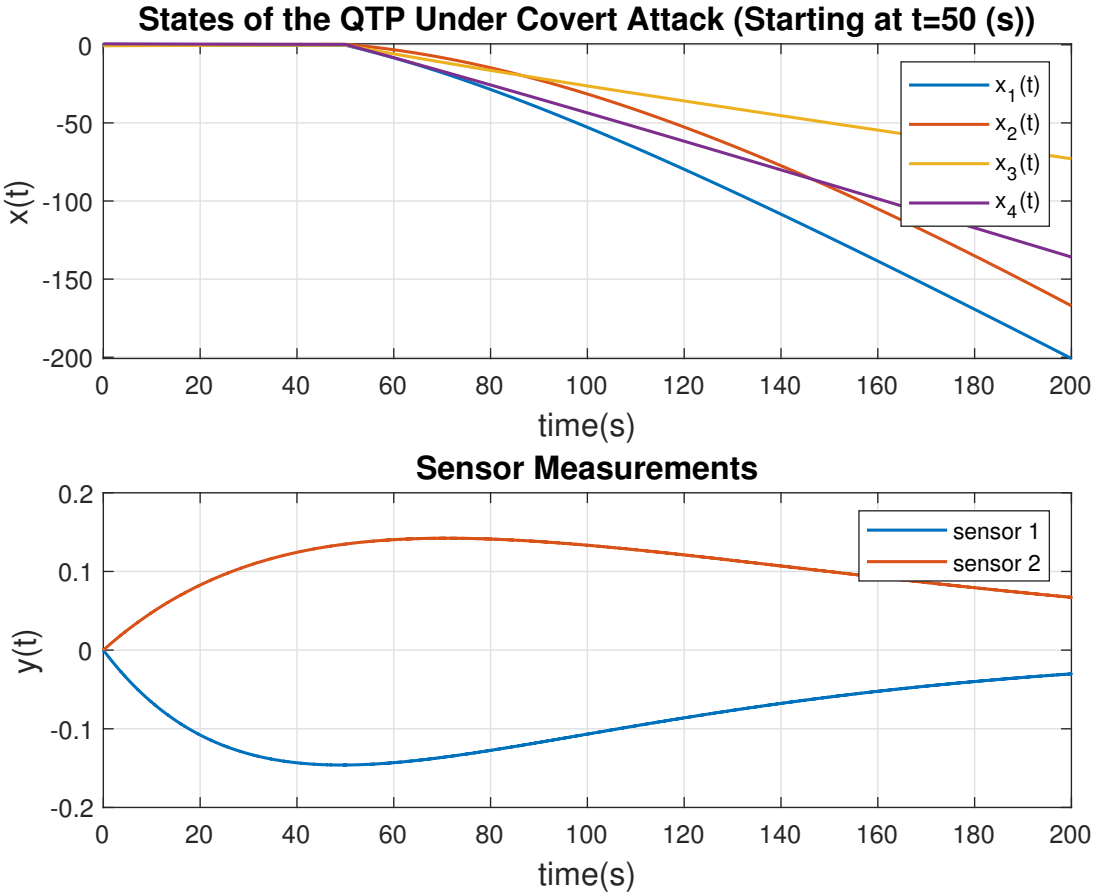

Figure 5.3: The QTP under covert attacks where the states become unbounded while the outputs show a normal attack-free behavior.

perfectly undetectable cyber-attacks in the QTP, i.e., $SE_{\Sigma} = 2$. Moreover, the security index for only perfectly undetectable cyber-attacks is equal to 3 and the system operators can prevent them by securing 1 input and 1 output communication channel, i.e., $\hat{SE}_{\Sigma} = 2$. Hence, in this case study, we have $SE_{\Sigma} = \hat{SE}_{\Sigma}$ while the security index for undetectable cyber-attacks is 2 and that for perfectly undetectable cyber-attacks is 3.

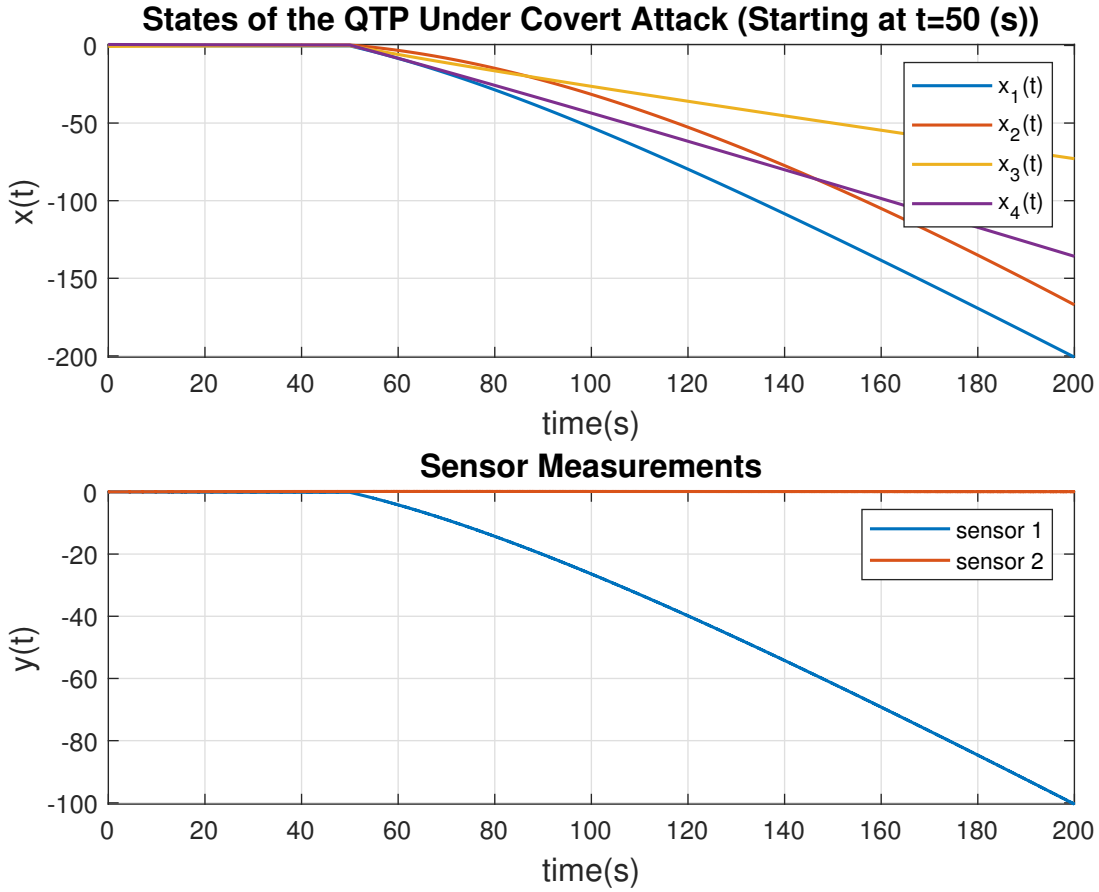

Figure 5.4: Preventing adversaries from executing a covert attack in the QTP by securing the first input and the first output communication channel given that the first output remains unbounded and detectable.

5.6.2 Stealthy Cyber-Attacks in Nonlinear CPS

In this case study, the effectiveness of our proposed methodologies and results in Sections 5.4 and 5.5 are illustrated. We consider the following nonlinear control affine system [54, 55]:

$$\begin{aligned} \dot{x} =& \begin{bmatrix} 0.3x_1 \\ 0.2(x_2 - x_1^2) \end{bmatrix} + \begin{bmatrix} 1 \\ x_1^2 \end{bmatrix} u_1 + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_2 \\ y =& h(x) = x_1^2 + x_2, \end{aligned} \tag{126}$$

where $x = [x_1,\ x_2]^\top$. Similar to [55], we consider the following KCT (102):

$$z(t) = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{bmatrix} = T(x),$$

where $\phi_1 = x_1$, $\phi_2 = x_2 + 0.5x_1^2$, and $\phi_3 = x_1^2$. Consequently, the KCF characteristic matrices in (104) are

$$\begin{aligned} \Lambda =& \text{diag}(0.3, 0.2, 0.6),\ C^x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -0.5 \end{bmatrix}, \\ C^h =& \begin{bmatrix} 0 & 1 & 0.5 \end{bmatrix},\ \tilde{g}_1(z) = \begin{bmatrix} 1 & z_1 + z_3 & 2z_1 \end{bmatrix}^\top, \\ \tilde{g}_2(z) =& \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^\top. \end{aligned}$$

States of the system (126) and its nominal output $y(t)$ along with the output of the KCF with the control input $u_1 = \ -\ 0.5x_1$ and $u_2 = -2x_2$ are depicted in Figure 5.5.

Since $C^h\tilde{g}_1(z) = 2z_1 + z_3$ and $C^h\tilde{g}_2(z) = 1$, it follows from Lemma 5.2 that at any point in the state-space the relative degree of the system (126) is equal to 1. Moreover, according to (119) and (120), we define

$$U(z) = \begin{bmatrix} 0 & 1 & 0.5 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} z,$$

where $\zeta_1 = C^h z$, $\eta_1 = z_1$, and $\eta_2 = z_3$. Moreover, (126) in its normal form can be expressed by

$$\begin{aligned} \dot{\zeta}_1 =& 0.2\zeta_1 + 0.2\eta_2 + (2\eta_1 + \eta_2)u_1 + u_2, \\ \dot{\eta}_1 =& 0.3\eta_1 + u_1, \\ \dot{\eta}_2 =& 0.6\eta_2 + 2\eta_1 u_1, \\ y =& \zeta_1. \end{aligned}$$

Consequently, as per Proposition 5.1, if $z(0) \in \ker(C^h)$, the control input $u_1 = 0$ and $u_2 = -0.2\eta_2$

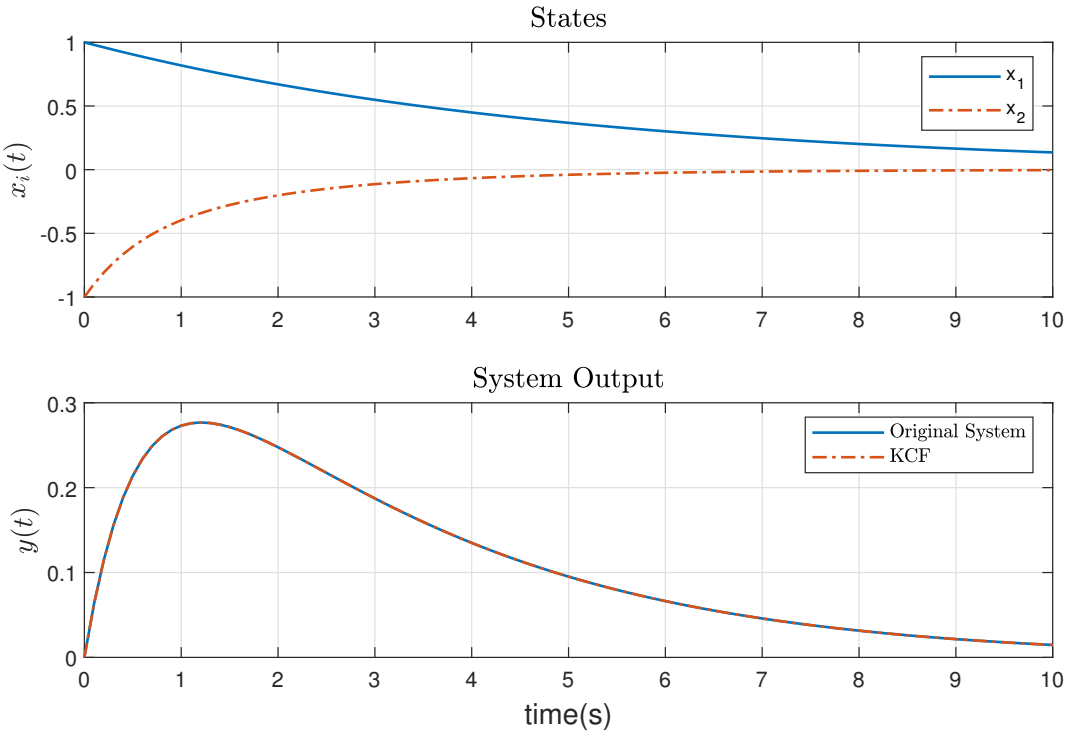

Figure 5.5: Response of the original system and its KCF.

results in $y \equiv 0$. Moreover, it can be easily seen that the internal dynamics of (126), given by $\dot{\eta}_1 = 0.3\eta_1$ and $\dot{\eta}_2 = 0.6\eta_2$, are unstable. If $z_1(0) = 1$, from the definition of ϕ_3 one obtains $z_3(0) = 1$, therefore, $z_2(0) = -0.5$ satisfies the condition $z(0) \in \ker(C^h)$. Moreover, from $x = C^x z$ we derive $x(0) = [1,\ -1]^\top$.

Under initial condition $z(0) = [1,\ -0.5,\ 1]^\top$, a zero dynamics attack is performed on the system (126), where the nominal control input is $u = [-0.5x_1,\ -2x_2]^\top$. The manipulated output of the system, i.e., $y^* = h(x^*) + a_y$, and the actual output without the sensor attack, i.e., $y^* = h(x^*)$, are shown in Figure 5.6, where a_y is designed based on (122). As it can be seen in Figure 5.6, the actual output of the system without sensor attacks is equal to zero while the the system is internally unstable and $|x_2|$ is increased over time. Also, in Figure 5.6, the received output measurements of the system which are manipulated by the adversary, i.e., $y^* = h(x^*) + a_y$, show the nominal attack free behavior of the system similar to Figure 5.5.

Moreover, the case of covert attack is shown in Figure 5.7, where the adversary's objectives are to make the system unstable and to maintain the attack stealthy. It can be observed from Figure 5.7 that as the result of the covert attack, the actual output of the system without sensor attacks, i.e., $y^* = h(x^*)$, is increased

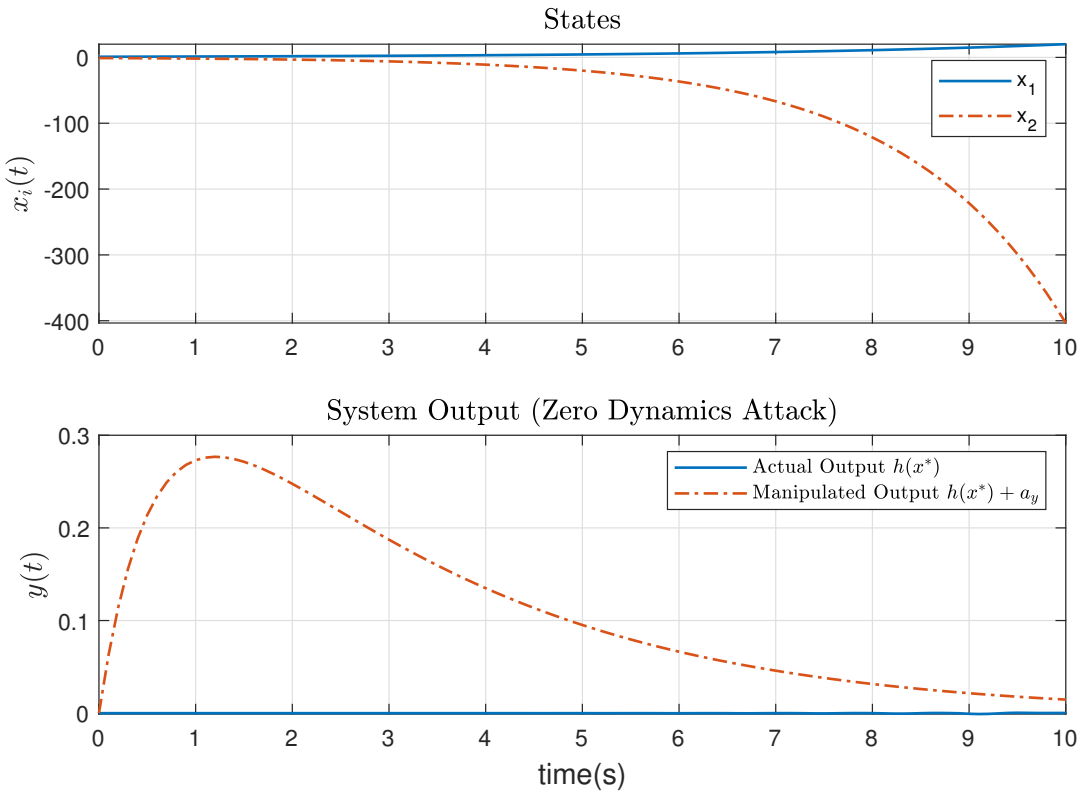

Figure 5.6: System under zero dynamics attacks.

over time which, in theory, can lead to an unbounded output measurement while the manipulated output is similar to the case of attack free system in Figure 5.5, i.e., the nominal output y. Hence, both objectives of the adversary in the performed covert attack in Figure 5.7 are achieved.

5.7 Conclusion

In this chapter, stealthy cyber-attacks in linear and nonlinear cyber-physical systems (CPS) were studied. The notion of security effort (SE) is developed and formally specified as a security measure for linear CPS. The SE metric denotes the minimum number of input and output communication channels that should be secured to prevent adversaries from executing zero dynamics attacks, covert attacks, and controllable attacks. Moreover, it is shown that SE can be specified to prevent only perfectly undetectable cyber-attacks in the CPS, namely covert attacks and controllable attacks. Since zero dynamics attacks, covert attacks, and

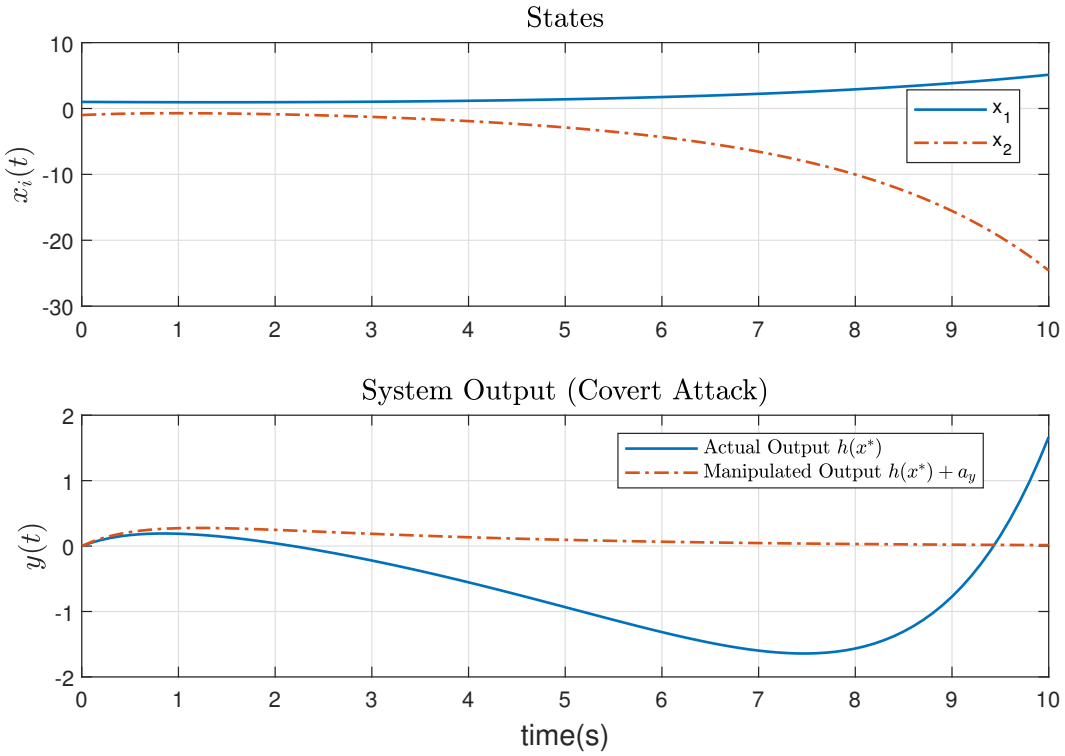

Figure 5.7: System under covert attacks.

controllable attacks belong to the weakly unobservable and controllable weakly unobservable subspaces of the CPS, conditions for making these subspaces equal to zero are developed and investigated. Hence, the conditions that are developed to make weakly unobservable and controllable weakly unobservable subspaces equal to zero are utilized to compute SE. As for nonlinear CPS, by utilizing the Koopman operator theory, data-driven stealthy cyber-attacks for a class of nonlinear CPS have been studied. The notion of ϵ-stealthy cyber-attacks for nonlinear CPS was defined which can be used as a measure of detectability for these systems. Moreover, the Koopman canonical form of the nonlinear control affine CPS was utilized to define relative degree for a given system using its Koopman eigenfunctions, Koopman eigenvalues, and Koopman modes. Consequently, a methodology was proposed to discover internal dynamics, i.e., the zero dynamics, of the nonlinear CPS. Furthermore, strategies for executing zero dynamics and covert cyber-attacks were proposed. Moreover, conditions for finding sensor measurements that should be secured to prevent prevent zero dynamics and covert cyber-attacks in nonlinear CPS were studied.

Chapter 6

On Cyber-Attacks in Multi-Agent Systems

In this chapter, four main problems for multi-agent systems (MAS) are investigated and addressed, namely the privacy preserving consensus control, executing controllability cyber-attacks in the MAS, performing undetectable cyber-attacks in MAS, and developing cyber-attack detection methodologies for these systems. The MAS require sharing their information with their neighboring agents to reach a consensus in a distributed manner. In this chapter, a transformation-based consensus control methodology is developed and implemented that can be utilized to reach a consensus among agents in a distributed manner without revealing their true information to their neighboring agents. The proposed methodology protects agents' privacy against eavesdropper adversaries and malicious agents capable of intercepting and accessing the agents' exchanged data. A unique isometric isomorphism is employed for each agent to map the true value of exchanged sensor measurements and state estimates. By leveraging the property of isometric isomorphism in preserving norms, it is shown that reaching a consensus among agents is equivalent to that can be accomplished by the transformed agents' dynamics. Moreover, this chapter aims at investigating a novel type of cyber-attacks that is injected to the MAS having an underlying directed graph. The cyber-attack, which is designated as the controllability attack, is injected by the malicious adversary into the communication links among the agents. The adversary, leveraging the compromised communication links disguises the cyber-attack signals and attempts to take control over the entire network of MAS. The adversary aims at achieving this by directly attacking only a subset of the multi-agents. Conditions under which the malicious hacker has control over the entire MAS network are provided. Two notions of security controllability

indices are proposed and developed. These notions are utilized as metrics to evaluate the controllability that each agent provides to the adversary for executing the malicious cyber-attack. Furthermore, the possibility of introducing zero dynamics cyber-attacks on the MAS through compromising the communication links is also investigated. Our next objective in this chapter is to study and develop conditions for a network of MAS where a malicious adversary can utilize vulnerabilities in order to ensure and maintain cyber-attacks undetectable. We classify these cyber-attacks as undetectable in the sense that their impact cannot be observed in the generated residuals. It is shown if an agent that is the root of a rooted spanning tree in the MAS graph is under a cyber-attack, the attack is undetectable by the entire network. Next we investigate if a non-root agent is compromised, then under certain conditions cyber-attacks can become detectable. Moreover, a novel cyber-attack that is designated as quasi-covert cyber-attack is introduced that can be used to eliminate detectable impacts of cyber-attacks to the entire network and maintain these attacks as undetected. Finally, an event-triggered based detector is proposed that can be used to detect the quasi-covert cyber-attacks. The work in this chapter has partly appeared in [13, 131, 132].

The main contributions of this chapter are stated below.

(1) A unique isometric isomorphisms is developed and designed and used for each agent so that adversaries require discovering all the utilized isometric isomorphisms to disclose information of the entire network.

(2) To preserve the privacy of agents when they are communicating with agents in their nearest neighborhood, a distributed consensus control is proposed that requires the transformed output measurements and dynamic controller states of the nearest neighboring agents to ensure reaching consensus.

(3) We introduce the notion of controllability attacks on communication channels of the MAS systems. The importance of these attacks by studying and developing conditions that would provide the adversary full control over the entire MAS system is developed and formalized.

(4) It is shown that the adversary is not capable of exciting zero dynamics of the directly attacked and healthy agents simultaneously.

(5) A definition is introduced and proposed that specifies characteristics of undetectable cyber-attacks on MAS. Then conditions on the graph topology and its Laplacian matrix along with detectors of MAS

are developed so that an adversary is capable of performing undetectable cyber-attacks. Moreover, if the above does not hold, we investigate under what conditions cyber-attacks are detectable on a certain team of agents.

(6) Quasi-covert cyber-attacks are introduced where malicious hackers can inject in order to maintain their attacks undetected provided only non-root agents are compromised.

(7) An event-triggered detector is proposed for quasi-covert cyber-attacks that given its event-based communication strategy is more secure in comparison with conventional communication protocols.

The remainder of the chapter is organized as follows. In Section 6.1, the basic concepts in graph theory that are required are presented and model of MAS systems along with certain assumptions and lemmas are provided. Model of MAS systems where the communication channels are under attack as well as the objectives of this chapter are introduced in Section 6.2. Our proposed privacy preserving consensus control methodology for MAS is investigated in Section 6.3. In Section 6.4, necessary and sufficient conditions for the adversary to gain full control over the MAS systems network are formulated and presented. The limitations on zero dynamics attacks that the adversary is capable of injecting by compromising the communication channels are investigated in Section 6.4.3. Undetectable cyber-attacks in MAS are formally defined and introduced in Section 6.5. Moreover, an event-triggered cyber-attack detection methodology is developed and investigated in Section 6.6. Illustrative numerical case studies to demonstrate the capabilities of our proposed methodologies are provided in Section 6.7.

6.1 Preliminaries

6.1.1 Graph Theory

A directed graph (digraph) $\mathcal{G}$ is defined with a set of vertices or nodes $\mathcal{V} = \{1, 2, \ldots, N\}$ and the set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ that denotes the edges of the digraph. The pair of distinct vertices $\mathcal{G} : (i, j) \in \mathcal{E}$ defines an edge. Graph $\mathcal{G}$ is called directed if $(i, j) \in \mathcal{E}$ does not imply $(j, i) \in \mathcal{E}$. The matrix $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{N \times N}$ denotes the adjacency matrix of $\mathcal{G}$, where $a_{ij} = 1$ when there exists a link from node j to i. The set $\mathcal{N}_i$ denotes the set of neighbors of the node i which contains those nodes that have an edge to i. Moreover, $|\mathcal{N}_i| = d_i$,

where $|\cdot|$ is the cardinality of the set. The Laplacian matrix of the graph $\mathcal{G}$ is defined as $L = D - \mathcal{A}$, where $D = \text{diag}(d_1, d_2, \ldots, d_N)$ denotes the in-degree matrix. A directed path between nodes i and j, that is denoted by $\mathcal{P}_{ij}$, is a sequence of edges that connects the node i to the node j and follows along the direction of edges.

6.1.2 Model of MAS

We consider a group of N agents with the i-th agent dynamics, denoted by Σ_i, expressed by

$$\Sigma_i : \begin{cases} \dot{x}_i(t) = & Ax_i(t) + Bu_i(t), \\ y_i(t) = & Cx_i(t), \quad i = 1, \ldots, N, \end{cases} \tag{127}$$

where $x_i(t) \in \mathbb{R}^n$ is the state of the i-th agent, $u_i(t) \in \mathbb{R}^m$ denotes the control input of agent i, and $y_i(t) \in \mathbb{R}^p$ represents the sensor measurement of the i-th agent. The matrices (A, B, C) are of appropriate dimensions and agents are assumed to be controllable and observable.

Definition 6.1. *Consensus is achieved among agents in* (127) *if for any initial condition the following holds:*

$$\lim_{t\to\infty} \|x_i(t) - x_j(t)\| = 0,$$

for all $i, j = 1, \ldots, N$*, where* $\|.\|$ *denotes the Euclidean norm.*

Lemma 6.1 ([93, 133])**.** *If the digraph* $\mathcal{G}$ *contains a directed spanning tree, the algebraic multiplicity of the eigenvalue* $\lambda_0 = 0$ *of the Laplacian matrix* L *associated with the digraph is one. Moreover,* $\mathbf{1}_N$ *and* $r = [r_1, \ldots, r_N]^\top \in \mathbb{R}^N$ *are the right and left eigenvectors associated with* λ_0*, respectively, where for scalar* r_i*,* $i = 1, \ldots, N$*, one has* $\sum_{i=1}^N r_i = 1$*.*

Lemma 6.2 ([134])**.** *Given the matrices* Q*,* W*,* M*, and* Z *with appropriate dimensions, the Kronecker product* $\otimes$ *satisfies the following conditions:*

$(i)\ (Q + W) \otimes M = Q \otimes M + W \otimes M;$

$(ii)\ (Q \otimes W)(M \otimes Z) = (QM) \otimes (WZ).$

Lemma 6.3 ([93,133]). *The algebraic multiplicity of the eigenvalue $\lambda_0 = 0$ of the Laplacian matrix L_0 that belongs to the graph $\mathcal{G}_0$ is one if $\mathcal{G}_0$ contains a directed spanning tree. Furthermore, the other eigenvalues have positive real parts and the right and left eigenvectors associated with λ_0 are denoted by $\mathbf{1}_N$ and $r_0 = [r_1, \ldots, r_N]^\top \in \mathbb{R}^N$, respectively, where $\sum_{j=1}^{N} r_j = 1$ with scalar r_j.*

Assumption 6.1. *The directed graph $\mathcal{G}$ contains at least one directed spanning tree. The set $V_r = \{i_r,\ i_r + 1, \ldots, i_r + N_r - 1\}$ contains agents that constitute as roots of directed spanning trees in $\mathcal{G}$ and N_r denotes the number of these agents.*

6.2 Problem Statement

6.2.1 Privacy Preserving Control in MAS

Since we assume that states of the MAS (127) are not measurable, i.e., $p < n$, to reach a consensus among agents, each agent requires an observer-based controller. The proposed observer-based consensus protocols in the literature, such as in [64, 66, 135, 136] require agents to transmit their observer states and output measurement information to their neighboring agents. Hence, each agent has access to its neighbors' sensitive information, which can be considered as private information. Moreover, adversaries that are capable of reading the transmitted data among agents will have access to this sensitive information and can exploit it for malicious purposes and cyber-attacks.

Models of Adversaries

In this work, we consider two types of adversaries, namely external eavesdroppers and internal honest-but-curious agents in the MAS, where both have access to the parameters of A, B, and C. Eavesdropper adversaries are capable of reading transmitted data among agents, but they cannot manipulate data or inject malicious cyber-attack signals into the communication links among the agents.

A group of agents in the MAS is considered as honest-but-curious if they are legitimate participants of the distributed consensus protocol and will not deviate from it, but attempt to learn all possible information about the neighboring agents from the received data [70]. For instance, consider a group of self-driving cars with vehicle-to-vehicle (V2V) communication capabilities which are on the route and need to reach an

agreement on their speed according to speed limitations of that area and road conditions. In this scenario, an honest-but-curious agent can discover location and the traveled route of other cars. Hence, an honest-but-curious agent is considered as an adversary in the sense that it violates data privacy of other agents.

Privacy and Control Objectives

The objectives of this part of the chapter are twofold. Our control objective is to design a distributed dynamic controller for MAS (127) that ensures reaching a consensus among the agents in sense of Definition 6.1. The second objective is to protect the privacy of agents by developing a transformation-based scheme for MAS and the proposed dynamic controller that can be utilized to preserve the privacy of agents against eavesdropper adversaries and honest-but-curious agents, while consensus is ensured. Preserving privacy of agents is achieved if agents do not share the true values of observer state and output measurement with their neighboring agents, and also the transformed dynamics are indistinguishable from the original ones.

6.2.2 Controllability Cyber-Attacks in MAS

To design a consensus control protocol for the MAS in (127), one needs to first estimate the states of the system since only a few are assumed to be measurable. Consider the following observer-based consensus protocol for the system (127) [64]:

$$\begin{aligned} \dot{\hat{x}}_i(t) =& A\hat{x}_i(t) + Bu_i(t) + H\sum_{j\in\mathcal{N}_i}(\zeta_{\mathrm{y}}(t) + C\zeta_{\mathrm{x}}(t)), \\ u_i(t) =& K\hat{x}_i(t), \end{aligned} \tag{128}$$

where $\hat{x}_i(t) \in \mathbb{R}^n$ denotes the state of the observer for the i-th agent, $\zeta_{\mathrm{y}}(t) = y_j(t) - y_i(t)$, $\zeta_{\mathrm{x}}(t) = \hat{x}_i(t) - \hat{x}_j(t)$, $H \in \mathbb{R}^{n\times p}$ is a full column rank observer gain matrix, and $K \in \mathbb{R}^{m\times n}$ is a control gain matrix that should be designed.

Cyber-Attack on the Communication Links

As described in (128), the agent $j \in \mathcal{N}_i$ transmits its observer state $\hat{x}_j(t)$ and output $y_j(t)$ to the agent i as the pair $p_{ji}(t) = (\hat{x}_j(t),\, y_j(t))$. Since this communication is carried out through a network link, it would

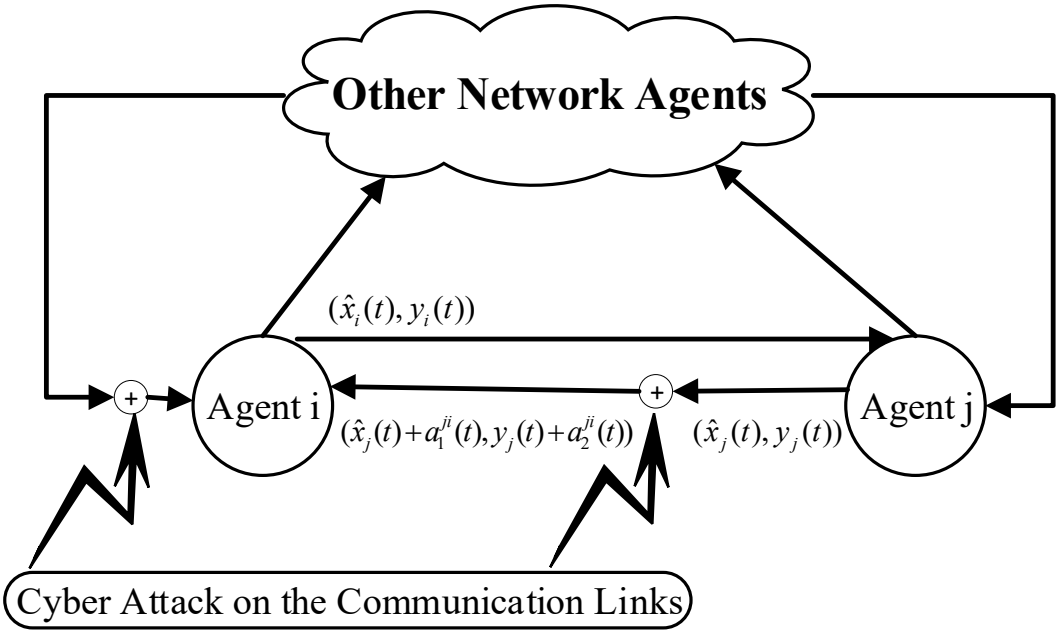

Figure 6.1: A communication link cyber-attack on the agent i. $a_1^{ji}(t)$ designates the cyber-attack on the transmitted states of the observer, and $a_2^{ji}(t)$ designates the cyber-attack on the transmitted output measurements.

be prone and vulnerable to cyber-attacks, as illustrated in Figure 6.1. The adversary disguises their injected signals as legitimate information from the neighboring agents of their target such that the targeted agent i only receives the cyber-attack signals. This cyber-attack can be considered as a man-in-the-middle type of attack [137].

Consequently, the malicious attacker adds signals $a_1^{ji}(t) = a_{\hat{x}}^{ji}(t) - \hat{x}_j(t)$ and $a_2^{ji}(t) = a_y^{ji}(t) - y_j(t)$ to $p_{ji}(t)$ so that the agent i receives $p_{ji}^{\mathrm{a}}(t) = (\hat{x}_j(t) + a_1^{ji}(t),\ y_j(t) + a_2^{ji}(t)) = (a_{\hat{x}}^{ji}(t),\ a_y^{ji}(t))$ from the agent j. Two cyber-attack signals $a_{\hat{x}}^{ji}(t) \in \mathbb{R}^n$ and $a_y^{ji}(t) \in \mathbb{R}^p$ are unknown and are to be designed based on the adversary's intentions.

Assumption 6.2. *The adversary is capable of executing the worst case scenario attack in which all the incoming communication links of a given agent are under attack.*

Remark 6.1. *Since the MAS have limited power resources and to make their communications more efficient*

they use the same communication protocols and encryption/decryption algorithms on all their communication channels [138]. Hence, if an adversary discovers a vulnerability for one channel of an agent, it is capable of attacking other channels as well.

Given the observer-based consensus protocol (128), the closed-loop equations of the system (127) and observer (128) given the communication link cyber-attacks can be reformulated as follows:

$$\dot{x}_i(t) = Ax_i(t) + BK\hat{x}_i(t), \tag{129}$$

$$\begin{aligned}\dot{\hat{x}}_i(t) = {} & A\hat{x}_i(t) + BK\hat{x}_i(t) + H\textstyle\sum_{j\in\mathcal{N}_i}(\zeta_{\mathrm{y}}(t) + q_i a_2^{ji}(t) \\ & + C(\zeta_{\mathrm{x}}(t) - q_i a_1^{ji}(t))),\end{aligned} \tag{130}$$

for $i = 1, ..., N$ with $q_i = 1$ if the communication links of the agent i are under attack, and $q_i = 0$, otherwise.

Objectives in Designing Controllability Cyber-Attacks

The objectives of this topic are threefold. The first objective is to investigate conditions on the MAS and its Laplacian matrix under which the adversary can gain full controllability over the system in (129). The adversary attempts to directly attack a subset of MAS agents and control the remaining agents as followers of the attacked agents. The second objective is to propose and investigate controllability measures that are based on graph of the MAS which is not fully controllable by the adversary and can be employed to inject attacks on agents that can be controlled through the directly attacked agents. And finally, the third objective is to study the possibility of executing zero dynamics attacks in the MAS governed by (129).

6.2.3 Undetectable Cyber-Attacks in MAS

In [135], for MAS in the form of (127) an observer-based consensus protocol was proposed as follows:

$$\begin{aligned}\dot{\hat{x}}_i(t) =& A\hat{x}_i(t) + Bu_i(t) - cF\zeta_i(t), \\ u_i(t) =& cK\epsilon_i(t),\end{aligned} \tag{131}$$

where $\hat{x}_i(t) \in \mathbb{R}^n$ denotes the observer state of the i-th agent, $\zeta_i(t) = \sum_{j\in\mathcal{N}_i}((y_j(t) - y_i(t)) + C(\hat{x}_i(t) - \hat{x}_j(t)))$, $\epsilon_i(t) = \sum_{j\in\mathcal{N}_i}(\hat{x}_i(t) - \hat{x}_j(t))$, $c \in \mathbb{R}$ is a scalar, $F \in \mathbb{R}^{n\times p}$ is the observer gain matrix, and

$K \in \mathbb{R}^{m\times n}$ denotes the control gain matrix to be selected. It should be noted that the observer (131) is a special case of that in [135] in which the MAS do not have a leader.

It has been shown in [135] that if the Assumption 6.1 holds, the observer and control parameters in (131) can be designed such that the closed-loop system reaches a consensus in sense of the Definition 6.1.

Lemma 6.4 ([93]). *Given the Hurwitz matrix $H \in \mathbb{R}^{n\times n}$, there exists a nonsingular matrix P_H that satisfies $P_H^{-1}HP_H = J_H$ and $\|e^{Ht}\| \leq \|P_H\|\|P_H^{-1}\|c_H e^{\lambda_H^m t}$, $\forall\, t \geq 0$, where J_H denotes the Jordan canonical form of H, $c_H > 0$ is a positive constant, and $\max Re(\lambda(H)) < \lambda_H^m < 0$.*

Cyber-Attacks on MAS and Residual Generation

The state estimator (131) for the i-th agent receives information from the agent $j \in \mathcal{N}_i$ that is represented as pair $p_{ji}(t) = (\hat{x}_j(t), y_j(t))$. Since the communication links among agents are vulnerable to cyber-attacks, the adversary is capable of injecting attack signals $a_{\hat{x}}^{ji}(t) \in \mathbb{R}^n$ and $a_y^{ji}(t) \in \mathbb{R}^p$ to $p_{ji}(t)$. Therefore, the agent i under cyber-attacks receives the manipulated pair $p_{ji}^a(t) = (\hat{x}_j(t) + a_{\hat{x}}^{ji}(t), y_j(t) + a_y^{ji}(t))$ from the agent j.

Remark 6.2. *cyber-attacks on MAS in this chapter can be considered as the "man-in-the-middle" type of attack [137]. In this cyber-attack type the adversary first blocks the received information from a group of neighboring agents of a compromised agent. Following this the adversaries inject their attack signals as legitimate information that are then received from a group of neighboring agents and transmitted to the targeted agent.*

Consequently, the closed-loop equations of the MAS (127) and the observer (128) under cyber-attacks can be represented in the following form:

$$\dot{x}_i(t) = Ax_i(t) + BcK\epsilon_i(t) - BcK\sum_{j\in\mathcal{N}_i} q_i a_{\hat{x}}^{ji}(t), \tag{132}$$

$$\begin{aligned}\dot{\hat{x}}_i(t) &= A\hat{x}_i(t) + BcK\epsilon_i(t) - cF\zeta_i(t) - BcK\sum_{j\in\mathcal{N}_i} q_i a_{\hat{x}}^{ji}(t)\\ &\quad -cF\sum_{j\in\mathcal{N}_i} q_i(a_y^{ji}(t) - Ca_{\hat{x}}^{ji}(t)), \quad i = 1,\ldots,N,\end{aligned} \tag{133}$$

where $q_i = 1$ if the incoming communication links to the agent i are compromised and under cyber-attack, and $q_i = 0$, otherwise.

To detect anomalies in the i-th agent or its neighboring agents one can utilize the received information from the neighboring agents and generate the following residual signals:

$$res_y^i(t) = \sum_{j\in\mathcal{N}_i} \|(y_j(t) - y_i(t)) + q_i a_y^{ji}(t)\|, \tag{134}$$

$$res_{\hat{x}}^i(t) = \sum_{j\in\mathcal{N}_i} \|(\hat{x}_i(t) - \hat{x}_j(t)) - q_i a_{\hat{x}}^{ji}(t)\|. \tag{135}$$

Objectives in Designing Undetectable Cyber-Attacks in MAS and Event-Triggered Detectors

We are pursuing three main objectives in this topic. First, we introduce a formal definition for undetectable cyber-attacks on MAS in the sense that residuals (134) and (135) converge to zero as time approaches to infinity. Following that we show that under certain conditions on the network the cyber-attacks on a given agent which happens to be the root of the communication graph can become undetectable. The second objective is to introduce a novel type of undetectable cyber-attacks that are designated as *quasi-covert cyber-attacks* in which if a group of non-root agents are compromised, still the adversary is capable of eliminating and hiding the impact of cyber-attacks on agents that could otherwise detect them. The final objective is to develop an event-triggered based detector with the goal of detecting quasi-covert cyber-attack signals that are developed in the second objective.

6.3 Proposed Methodology for Privacy Preserving Consensus Control

In this section, our proposed transformation-based scheme to preserve the privacy of the MAS is studied. Isometric isomorphisms are utilized to transform the original dynamics of each agent and its dynamic controller to a new basis. It is shown that by utilizing the proposed dynamic controller in this chapter, agents can reach a consensus and preserve their privacy. Moreover, design conditions for the proposed dynamic controller are provided.

6.3.1 Privacy Preserving Distributed Consensus Control

Let us consider the i-th agent in the MAS (127) and invertible linear maps (isomorphisms) $P_i : \mathbb{R}^n \to \mathbb{R}^n$ and $S_i : \mathbb{R}^p \to \mathbb{R}^p$ as similarity transformations. One can use P_i and S_i to transform the dynamics of the agent i into the following form:

$$\tilde{\Sigma}_i : \begin{cases} \dot{\tilde{x}}_i(t) = & \tilde{A}_i\tilde{x}_i(t) + \tilde{B}_i u_i(t), \\ \tilde{y}_i(t) = & \tilde{C}\tilde{x}_i(t), \quad i = 1, \dots, N, \end{cases} \tag{136}$$

where $\tilde{x}_i(t) = P_i x_i(t)$, $\tilde{A}_i = P_i A P_i^{-1}$, $\tilde{B}_i = P_i B$, $\tilde{C}_i = S_i C P_i^{-1}$.

Definition 6.2 (Isometric Isomorphism [139])**.** *Let $(X, \|\cdot\|_X)$ and $(Z, \|\cdot\|_Z)$ denote normed vector spaces over the field of real numbers. A linear isomorphism $U : X \to Z$ is an isometric isomorphism between X and Z if*

$$\|Ux\|_Z = \|x\|_X,$$

$\forall x \in X$. Moreover, U^{-1} is also an isometric isomorphism.

In this chapter, a dynamic controller is proposed for the MAS (127) that can be employed to reach a consensus among the agents and preserve their privacy. Let us consider distinct isometric isomorphisms P_q and S_q, for every $q = 0, \dots, N$, such that $P_0 = MQ$, where $M \in \mathbb{R}^{n\times n}$ and $Q \in \mathbb{R}^{n\times n}$ are invertible matrices. The proposed dynamic controller for the i-th agent, that is denoted by $\mathcal{C}_i$, is given by

$$\mathcal{C}_i : \begin{cases} \dot{\zeta}_i(t) = & \tilde{A}_i\zeta_i(t) + \tilde{B}_i u_i(t) + \tilde{H}_i\Pi_i(t), \\ u_i(t) = & K\sum\limits_{j\in\mathcal{N}_i} \bar{P}(z_i^{\mathrm{x}}(t) - z_j^{\mathrm{x}}(t)), \end{cases} \tag{137}$$

for $i = 1, \dots, N$, where $\zeta_i(t) \in \mathbb{R}^n$ is the state of the dynamic controller of the agent i, $\bar{P} = MP_1^{-1}$, $z_i^{\mathrm{x}}(t) = F_i\zeta_i(t)$, $F_i = P_1 Q P_i^{-1}$, $\Pi_i(t) = \sum_{j\in\mathcal{N}_i}(\bar{H}(z_i^{\mathrm{x}}(t) - z_j^{\mathrm{x}}(t)) + (\mathcal{J}_j z_j^{\mathrm{y}}(t) - \mathcal{J}_i z_i^{\mathrm{y}}(t)))$, $\bar{H} = S_0 C Q^{-1} P_1^{-1}$, $z_i^{\mathrm{y}}(t) = J_i y_i(t)$, $J_i = S_0 S_i$, $\mathcal{J}_i = S_0 S_i^{-1} S_0^{-1}$, and $\tilde{H}_i = P_i H$. One can consider $\hat{x}_i(t) = P_i^{-1}\zeta_i(t)$ as the true value of the i-th agent's state estimation. Moreover, matrices H and K are of appropriate dimensions and to be designed subsequently. Without loss of generality, we assume that the agent 1 is the root of the spanning tree contained in the digraph $\mathcal{G}$.

Remark 6.3. *Given the dynamics of the controller $\mathcal{C}_i$ in (137), the agent i receives $z_j^x(t)$ and $z_j^y(t)$ from its neighboring agents and transmits $z_i^x(t)$ and $z_i^y(t)$. Moreover, the matrices $(\bar{P},\, F_i,\, \bar{H},\, J_i,\, \mathcal{J}_i,\, \tilde{H}_i)$ have been computed off-line by a third party (e.g., an operator) and integrated into the dynamics of $\mathcal{C}_i$. Hence, each honest-but-curious agent i does not need to know S_q and P_q, for $q = 0, \ldots, N$, to compute $(\bar{P},\, F_i,\, \bar{H},\, J_i,\, \mathcal{J}_i,\, \tilde{H}_i)$.*

Remark 6.4. *Given that in this chapter we assume that each agent in the MAS (127) can be an honest-but-curious adversary, it is reasonable to limit the access of all agents to certain parameters of $\mathcal{C}_i$ and isometric isomorphisms. Hence, in the proposed consensus protocol (137), the i-th agent has access to $u_i(t)$ and $\Pi_i(t)$, and has knowledge of $(\zeta_i(t), y_i(t))$, $(z_j^x(t), z_j^y(t))$, and $(\bar{P},\, F_i,\, \bar{H},\, J_i,\, \mathcal{J}_i,\, \tilde{H}_i)$, for $i = 1, \ldots, N$ and $j \in \mathcal{N}_i$. The set of all information known to an honest-but-curious agent is defined as its "view of the protocol" [70]. Moreover, honest-but-curious agents and eavesdropper adversaries do not have knowledge of P_q and S_q, for $q = 0, \ldots, N$.*

Remark 6.5. *Considering that our control objective is to reach a consensus among the agents, internal dynamics of the controller $\mathcal{C}_i$ in (137) is similar to the agent's dynamics. However, to achieve control objectives other than reaching a consensus, one can design the controllers having internal dynamics that are different from the agent's dynamics.*

One can augment dynamics of the transformed MAS (136) and the controller (137) as given below

$$\dot{\mathcal{X}}_i(t) = \bar{A}_i \mathcal{X}_i(t) + \bar{K}_i \sum_{j \in \mathcal{N}_i} (\hat{\mathcal{X}}_i(t) - \hat{\mathcal{X}}_j(t)), \tag{138}$$

where $\mathcal{X}_i(t) = [\tilde{x}_i(t)^\top\, \zeta_i(t)^\top]^\top$, $\hat{\mathcal{X}}_i(t) = [\gamma_i(t)^\top\, \hat{\zeta}_i(t)^\top]^\top$, $\gamma_i(t) = P_0 x_i(t)$, $\hat{\zeta}_i(t) = P_0 P_i^{-1} \zeta_i(t)$, $\bar{A}_i = I_2 \otimes \tilde{A}_i$, and

$$\bar{K}_i = \begin{bmatrix} 0 & \tilde{B}_i K \\ -\tilde{H}_i \tilde{C}_0 & \tilde{B}_i K + \tilde{H}_i \tilde{C}_0 \end{bmatrix}.$$

Given that a distinct transformation P_i is employed for each agent, we have N *heterogeneous* augmented

dynamics in (138). Let us define

$$\check{A} = \begin{bmatrix} \bar{A}_1 & 0 & \cdots & 0 \\ 0 & \bar{A}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \bar{A}_N \end{bmatrix}, \; \check{K} = \begin{bmatrix} \bar{K}_1 & 0 & \cdots & 0 \\ 0 & \bar{K}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \bar{K}_N \end{bmatrix}. \tag{139}$$

Using (139), one obtains

$$\dot{\mathcal{X}}(t) = \check{A}\mathcal{X}(t) + \check{K}(L \otimes I_{2n})\hat{\mathcal{X}}(t), \tag{140}$$

where $\mathcal{X}(t) = [\mathcal{X}_1(t)^\top, \ldots, \mathcal{X}_N(t)^\top]^\top$, and $\hat{\mathcal{X}}(t) = [\hat{\mathcal{X}}_1(t)^\top, \ldots, \hat{\mathcal{X}}_N(t)^\top]^\top$.

Assumption 6.3. *Each agent handles its output measurement $y_i(t)$ as private information. Hence, eavesdropper adversaries are not capable of reading $y_i(t)$, for every $i = 1, \ldots, N$. Moreover, except for the agent i, the rest of the network do not have access to $y_i(t)$.*

Theorem 6.1. *Let Assumptions 6.1 and 6.3 hold. By using the controller $\mathcal{C}_i$ (137) as the consensus protocol in the MAS (127) and distinct isometric isomorphisms $P_{q_1} \neq P_{q_2}$ and $S_{q_1} \neq S_{q_2}$, for $q_1, q_2 = 0, \ldots, N$, $q_1 \neq q_2$, agents reach a consensus in the sense of Definition 6.1 if and only if matrices $\tilde{A}_0 + \lambda_z \tilde{B}_0 K$ and $\tilde{A}_0 + \lambda_z \tilde{H}_0 \tilde{C}_0$ are Hurwitz for $z = 2, \ldots, N$, where $\lambda_z \neq 0$ is the z-th eigenvalue of the Laplacian matrix L. Moreover, one has $\|x_i(t) - x_j(t)\| = \|\gamma_i(t) - \gamma_j(t)\|$, $\forall t \geq 0$.*

Proof. Let us define the disagreement vector between states of agents according to:

$$\delta(t) = \hat{\mathcal{X}}(t) - (\mathbf{1}_N r^\top \otimes I_{2n})\hat{\mathcal{X}}(t). \tag{141}$$

It follows that $\delta(t) = 0$ if and only if $\hat{\mathcal{X}}_1(t) = \hat{\mathcal{X}}_2(t) = \cdots = \hat{\mathcal{X}}_N(t)$. Moreover, from (140) one obtains

$$\dot{\hat{\mathcal{X}}}(t) = (I_N \otimes \bar{A}_0 + (L \otimes \bar{K}_0))\hat{\mathcal{X}}(t).$$

Consequently,

$$\dot{\delta}(t) = (I_N \otimes \bar{A}_0 + (L \otimes \bar{K}_0))\delta(t). \tag{142}$$

Since we assume that the digraph $\mathcal{G}$ contains a spanning tree and from Lemma 6.1, there exists a block diagonal matrix $\Delta \in \mathbb{R}^{N-1\times N-1}$ with diagonal entries equal to nonzero eigenvalues of L, and matrices $T \in \mathbb{R}^{N\times N}$, $Y \in \mathbb{R}^{N\times N-1}$, $W \in \mathbb{R}^{N-1\times N}$ such that $T = [\mathbf{1}_N\, Y]$ and:

$$T^{-1} = \begin{bmatrix} r^\top \\ W \end{bmatrix},\ T^{-1}\tilde{L}T = J = \begin{bmatrix} 0 & \mathbf{0}_{1\times N-1} \\ \mathbf{0}_{N-1} & \Delta \end{bmatrix},$$

where r is defined in Lemma 6.1 [66].

Consequently, we use T^{-1} to transform the dynamics of $\delta(t)$ such that $\varepsilon^1(t) = (T^{-1} \otimes I_{2n})\delta(t) = [\varepsilon_1^1(t)^\top,\, \varepsilon_{2:N}^1(t)^\top]^\top$. It follows from definition of $\delta(t)$ and $\varepsilon^1(t)$ that $\varepsilon_1^1(t) = 0$. Hence,

$$\dot{\varepsilon}_{2:N}^1(t) = (I_{N-1} \otimes \bar{A}_0 + (\Delta \otimes \bar{K}_0))\varepsilon_{2:N+1}^1(t). \tag{143}$$

Similar to [93], by using the Jordan canonical form of the matrix $I_{N-1} \otimes \bar{A}_0 + (\Delta \otimes \bar{K}_0)$, it can be shown that the matrices along the diagonal are similar to

$$\begin{bmatrix} \tilde{A}_0 + \lambda_q \tilde{B}_0 K & \lambda_z \tilde{B}_0 K \\ 0 & \tilde{A}_0 + \lambda_z \tilde{H}_0 \tilde{C}_0 \end{bmatrix}$$

for $z = 2, \ldots, N$, where λ_z is the z-th nonzero eigenvalue of the Laplacian matrix L. Hence, the stability of (143) is achieved if and only if $\tilde{A}_0 + \lambda_z \tilde{B}_0 K$ and $\tilde{A}_0 + \lambda_z \tilde{H}_0 \tilde{C}_0$ are Hurwitz. The stability of (143) implies

$$\lim_{t\to\infty} \|\check{x}_i(t) - \check{x}_j(t)\| = \lim_{t\to\infty} \|\hat{\mathcal{X}}_i(t) - \hat{\mathcal{X}}_j(t)\| = 0, \tag{144}$$

which follows the definition of isometric isomorphisms, where $\check{x}_i(t) = [x_i(t)^\top\, \hat{x}_i(t)^\top]^\top$. Moreover, from Definition 6.2 it can be inferred that $\|\check{x}_i(t) - \check{x}_j(t)\| = \|\hat{\mathcal{X}}_i(t) - \hat{\mathcal{X}}_j(t)\|$, $\forall t \geq 0$, which implies that the norm between agent states trajectories in the transformed space by the isometric isomorphism P_0 is equal to

that of in the original space. This completes the proof of the theorem. □

Remark 6.6. *From Definition 6.2 it can be seen that an isometric isomorphism preserves the norm between metric spaces, which has contributed to having* $\|\check{x}_i(t) - \check{x}_j(t)\| = \|\hat{\mathcal{X}}_i(t) - \hat{\mathcal{X}}_j(t)\|$, $\forall t \geq 0$, *as well as reaching a consensus in MAS as stated in Theorem 6.1. The latter implies that the energy of input signal in (137) considering the proposed transformations* P_i *and* S_i *is equal to that of without any transformation. Hence, our proposed consensus control methodology (137) does not increase the energy consumption of agents. However, reaching a consensus among agents also can be achieved by using non-isometric isomorphisms. Hence, if preserving the norm between agent states in the transformed space is not desirable, one can employ non-isometric isomorphisms for the MAS (127).*

6.3.2 Indistinguishibility of Dynamics

Considering the two types of adversaries in this chapter, namely eavesdropper adversaries and honest-but-curious agents, a privacy preserving consensus protocol was introduced in the last subsection. In Algorithm 3 below, the pseudo code of the communication protocol for the agent i and its controller $\mathcal{C}_i$ are provided. However, the next step is to show that the dynamics of the MAS Σ_i is indistinguishable from the dynamics of $\tilde{\Sigma}_i$ by adversaries. Towards this end, the following definition is adopted from [78] and is modified accordingly.

Algorithm 3 Pseudo code of the communication protocol for the i-th agent

Input:
$\Gamma_i = (\Sigma_i,\, y_i(t)),\, F_i,\, J_i,\, z_j^{\mathrm{x}}(t),\, z_j^{\mathrm{y}}(t), \forall\, j \in \mathcal{N}_i$

Output:
$z_i^{\mathrm{x}}(t),\, z_i^{\mathrm{y}}(t)$

communication protocol:

(1) Use F_i and J_i to encode $\zeta_i(t)$ and $y_i(t)$ into $z_i^{\mathrm{x}}(t) = F_i\zeta_i(t)$ and $z_i^{\mathrm{y}}(t) = J_i y_i(t)$, respectively;

(2) Transmit $z_i^{\mathrm{x}}(t)$ and $z_i^{\mathrm{y}}(t)$ to the agent q, where $i \in \mathcal{N}_q$, for $q \in \{1, \ldots, N\}$ and $q \neq i$;

(3) Use $\zeta_i(t)$, $y_i(t)$, and the received $z_j^{\mathrm{x}}(t)$ and $z_j^{\mathrm{y}}(t)$, for $j \in \mathcal{N}_i$, to update $u_i(t)$ using (137);

(4) Use the updated $u_i(t)$ in (127).

Definition 6.3. *The pairs* $\Gamma_i = (\Sigma_i,\, y_i(t))$ *and* $\tilde{\Gamma}_i = (\tilde{\Sigma}_i,\, \tilde{y}_i(t))$ *are indistinguishable from the perspective of eavesdropper adversaries and honest-but-curious agents, for* $i = 1, \ldots, N$*, if the exchanged data among the agents,* $z_i^x(t)$ *and* $z_i^y(t)$*, when* Γ_i *is considered as the input to the Algorithm 3, and when* $\tilde{\Gamma}_i$ *is considered as the input to the algorithm, can be made the same.*

Theorem 6.2. *By utilizing Algorithm 3,* $\Gamma_i = (\Sigma_i,\, y_i(t))$ *and* $\tilde{\Gamma}_i = (\tilde{\Sigma}_i,\, \tilde{y}_i(t))$*, for* $i = 1, \ldots, N$*, are indistinguishable from the perspective of eavesdropper adversaries and honest-but-curious agents in the sense of Definition 6.3.*

Proof. Similar to [78], indistinguishibility of $\Gamma_i = (\Sigma_i,\, y_i(t))$ and $\tilde{\Gamma}_i = (\tilde{\Sigma}_i,\, \tilde{y}_i(t))$ can be derived by running two instances of the Algorithm 3. First, consider $\Gamma_i = (\Sigma_i,\, y_i(t))$, F_i, and J_i as inputs, and second consider $\tilde{\Gamma}_i = (\tilde{\Sigma}_i,\, \tilde{y}_i(t))$, F_i, and $\bar{J}_i = S_0$ as inputs, for $i = 1, \ldots, N$. Following the steps provided in the Algorithm 3, it can easily be seen that for both inputs the resulting output would be $z_i^{\mathrm{x}}(t)$ and $z_i^{\mathrm{y}}(t)$. Hence, $\Gamma_i = (\Sigma_i,\, y_i(t))$ and $\tilde{\Gamma}_i = (\tilde{\Sigma}_i,\, \tilde{y}_i(t))$, for $i = 1, \ldots, N$, are indistinguishable by both types of adversaries. This completes the proof of the theorem. □

6.3.3 Designing Isometric Isomorphisms

In this subsection, two unitary transformations over the field of real numbers, which preserve the inner product, are presented for being employed as isometric isomorphisms for the MAS (127). These transformations are called Givens rotation and Householder transformation [140, 141].

Definition 6.4 (Givens rotation [140])**.** *A Givens rotation matrix* $G(\alpha_i, \beta_i, \theta_i) \in \mathbb{R}^{n\times n}$ *for the* i*-th agent is a rotation matrix with diagonal entries equal to* 1*, and* 0 *elsewhere, except for intersections of* α_i*-th and* β_i*-th rows and columns. The* $G(\alpha_i, \beta_i, \theta_i)v$ *rotates the vector* $v \in \mathbb{R}^n$ *by* θ_i *radians in the* (α_i, β_i) *plane in*

the counterclockwise scale. The Givens rotation matrix is represented in the following form:

$$G(\alpha_i, \beta_i, \theta_i) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c^i & \cdots & -s^i & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & s^i & \cdots & c^i & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}, \tag{145}$$

where $c^i = \cos\theta_i$, *and* $s^i = \sin\theta_i$.

Definition 6.5 (Householder transformation)**.** *[141]] Consider the matrix* $Q_i \in \mathbb{R}^{n\times n}$ *as a transformation over the field of real numbers for the* i*-th agent. The* Q_i *is a Householder transformation if it can be expressed according to*

$$Q_i = I_n - 2v_i v_i^\top, \tag{146}$$

where $v_i \in \mathbb{R}^n$ *with* $\|v\| = 1$.

Remark 6.7. *Both* $G(\alpha_i, \beta_i, \theta_i)$ *and* Q_i *are unitary transformations, which implies that they are isometric isomorphisms. Hence, we can utilize either Givens rotations or Householder transformations to design sets of distinct isometric isomorphisms* $V_p = \{P_0, \ldots, P_N\}$ *and* $V_s = \{S_0, \ldots, S_N\}$ *for the MAS* (127).

6.4 Controllability cyber-attacks

6.4.1 Conditions for Controllability

In this subsection, controllability of the MAS (129) and its observer that is provided in (130) from the adversary's point of view is studied. Let us define

$$\begin{aligned}\check{A} &= \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix}, \check{B} = \begin{bmatrix} 0 & B \\ 0 & B \end{bmatrix}, \check{H} = \begin{bmatrix} 0 & 0 \\ -H & H \end{bmatrix}, \\ \check{H}_{\mathrm{a}} &= \begin{bmatrix} 0 \\ H \end{bmatrix}, \check{K} = \begin{bmatrix} K & 0 \\ 0 & K \end{bmatrix}, \check{C} = \begin{bmatrix} C & 0 \\ 0 & C \end{bmatrix}.\end{aligned} \tag{147}$$

Using (147), the augmented dynamic of (129) and (130) can be derived as follows:

$$\begin{aligned}\dot{\check{x}}_i(t) =& (\check{A} + \check{B}\check{K})\check{x}_i(t) + \check{H}\check{C}\sum_{j\in\mathcal{N}_i}(\check{x}_i(t) - \check{x}_j(t)) \\ &+ \check{H}\check{C}\sum_{j\in\mathcal{N}_i} q_i\check{x}_j(t) + \check{H}_a q_i a_i(t),\end{aligned} \tag{148}$$

where $\check{x}_i(t) = [x_i(t)^\top\ \hat{x}_i(t)^\top]^\top$, and $a_i(t) = \sum_{j\in\mathcal{N}_i} a_y^{ji}(t) - Ca_{\hat{x}}^{ji}(t)$.

One can easily partition the agents into two groups, namely the first group contains agents that are directly under attack and the second group consists of agents that receive information from their neighboring agents without any manipulation by the adversary. Consequently, one has $x_{\mathrm{f}}(t) = [\check{x}_1(t)^\top, \check{x}_2(t)^\top, ..., \check{x}_{N_{\mathrm{f}}}(t)^\top]^\top$, which designates the state of those agents that are not directly under attack and act as followers. Second, $x_{\mathrm{a}}(t) = [\check{x}_{N_{\mathrm{f}}+1}(t)^\top, \check{x}_{N_{\mathrm{f}}+2}(t)^\top, ..., \check{x}_N(t)^\top]^\top$, which designates the directly attacked agents. The subscripts "f" and "a" are used to denote followers and attacked agents, respectively. N_{f} denotes the number of followers and N_{a} denotes the number of attacked agents, where $N = N_{\mathrm{f}} + N_{\mathrm{a}}$. Without loss of generality, we assume that the first N_{f} agents are not under attack. Consequently, the Laplacian matrix can be partitioned into the following form:

$$L = \begin{bmatrix} L_{\mathrm{f}} & l_{\mathrm{fa}} \\ l_{\mathrm{af}} & L_{\mathrm{a}} \end{bmatrix}, \tag{149}$$

where $L_{\mathrm{f}} \in \mathbb{R}^{N_{\mathrm{f}} \times N_{\mathrm{f}}}$ is a grounded Laplacian matrix [142], $L_{\mathrm{a}} \in \mathbb{R}^{N_{\mathrm{a}} \times N_{\mathrm{a}}}$, $l_{\mathrm{fa}} \in \mathbb{R}^{N_{\mathrm{f}} \times N_{\mathrm{a}}}$, and $l_{\mathrm{fa}} \in \mathbb{R}^{N_{\mathrm{a}} \times N_{\mathrm{f}}}$.

The dynamics of all N agents can be expressed as follows:

$$\dot{x}_{\mathrm{a}}(t) = \quad A_{\mathrm{a}} x_{\mathrm{a}}(t) + B_{\mathrm{a}} a(t), \tag{150}$$

$$\dot{x}_{\mathrm{f}}(t) = \quad A_{\mathrm{f}} x_{\mathrm{f}}(t) + A_{\mathrm{fa}} x_{\mathrm{a}}(t), \tag{151}$$

where $A_{\mathrm{a}} = I_{N_{\mathrm{a}}} \otimes (\check{A} + \check{B}\check{K}) + D_{\mathrm{a}} \otimes \check{H}\check{C}$, $A_{\mathrm{f}} = I_{N_{\mathrm{f}}} \otimes (\check{A} + \check{B}\check{K}) + L_{\mathrm{f}} \otimes \check{H}\check{C}$, $D_{\mathrm{a}} = \mathrm{diag}(d_{N_{\mathrm{f}}+1}, d_{N_{\mathrm{f}}+2}, ..., d_N)$, $B_{\mathrm{a}} = I_{N_{\mathrm{a}}} \otimes \check{H}_{\mathrm{a}}$, $A_{\mathrm{fa}} = l_{\mathrm{fa}} \otimes \check{H}\check{C}$, and $a(t) = [a_{N_{\mathrm{f}}+1}(t)^\top, a_{N_{\mathrm{f}}+2}(t)^\top, ..., a_N(t)^\top]^\top$. The dynamics of directly attacked agents (150) and the followers (151) can be augmented in the following form:

$$\begin{bmatrix} \dot{x}_{\mathrm{a}}(t) \\ \dot{x}_{\mathrm{f}}(t) \end{bmatrix} = \begin{bmatrix} A_{\mathrm{a}} & 0 \\ A_{\mathrm{fa}} & A_{\mathrm{f}} \end{bmatrix} \begin{bmatrix} x_{\mathrm{a}}(t) \\ x_{\mathrm{f}}(t) \end{bmatrix} + \begin{bmatrix} B_{\mathrm{a}} \\ 0 \end{bmatrix} a(t). \tag{152}$$

We are now in a position to state our first definition as well as the first result of this section.

Definition 6.6. *The MAS described in (152) is controllable by the adversary if, for every x_a^* and x_f^* and every finite $T > 0$, there exists an attack signal $a(t)$, $0 < t < T$, such that the MAS states do transition from $x_a(0) = 0$ and $x_f(0) = 0$ to $x_a(T) = x_a^*$ and $x_f(T) = x_f^*$, respectively.*

Theorem 6.3. *The adversary is capable of controlling the system (152) if the pairs $(\check{A} + \check{B}\check{K} + d_i\check{H}\check{C}, \check{H}_a)$ for $i = N_f + 1, ..., N$, and (A_f, A_{fa}) are controllable, rank$(\sum_{k=1}^{2nN-1} M_k Q_k^\top w_1)$ is either equal to $2nN_f$ if $N_f \leq N_a$ or equal to $2nN_a$ if $N_a < N_f$, where $Q = [Q_1, \ldots, Q_{(2nN-1)}]$, $Q_k = A_a^k B_a$ for $k = 1, \ldots, 2nN - 1$, $Q_0 = B_a$, $M = [M_1, \ldots, M_{(2nN-1)}]$, $M_k = \sum_{z=0}^{k-1} A_f^z A_{fa} Q_{k-1-z}$, columns of M_k are nonzero, and $w_1 \in \mathbb{R}^{(2nN_a)\times(2nN_a)}$ is a matrix that satisfies $w_1 \in ker(B_a^\top)$.*

Proof. Let us define

$$A^* = \begin{bmatrix} A_{\mathrm{a}} & 0 \\ A_{\mathrm{fa}} & A_{\mathrm{f}} \end{bmatrix}, \; B^* = \begin{bmatrix} B_{\mathrm{a}} \\ 0 \end{bmatrix}. \tag{153}$$

The controllability matrix of the system (152), $\mathcal{C}^* = \begin{bmatrix} B^*, & A^*B^*, & \ldots & (A^*)^{2nN-1}B^* \end{bmatrix}$, can be expressed

in the following form:

$$\mathcal{C}^* = \begin{bmatrix} Q_0 & Q_1 & \cdots & Q_{(2nN-1)} \\ 0 & M_1 & \cdots & M_{(2nN-1)} \end{bmatrix}.$$

For the system (152) to be controllable, $\mathcal{C}^*$ should be of full row rank. Hence, controllability is achieved if $[Q_0,\, Q]$ and M are right invertible and rows of Q and M, under some conditions that are provided below, are linearly independent.

From the definition of Q_0 and Q, one can conclude the right invertibility of $[Q_0,\, Q]$ is equivalent to the pair $(A_{\rm a}, B_{\rm a})$ being controllable. For this pair the matrix $D_{\rm a}$ is diagonal, therefore, $A_{\rm a} = \text{blockdiag}((\check{A} + \check{B}\check{K} + d_{N_{\rm f}+1}\check{H}\check{C}), ...,\, (\check{A} + \check{B}\check{K} + d_N\check{H}\check{C}))$ is a block diagonal matrix. The operator blockdiag$(\cdot)$ denotes a block diagonal matrix. In addition, $I_{N_{\rm a}} \otimes \check{H}_{\rm a} = \text{bockdiag}(\check{H}_{\rm a}, ..., \check{H}_{\rm a})$ is block diagonal. Hence, the controllability condition can be studied for each attacked agent separately.

The matrix $M = [M_1, \ldots, M_{(2nN-1)}]$ can be written as the product of two matrices, namely M^* and Q^*, i.e., $M = M^*Q^*$, where

$$M^* = \begin{bmatrix} A_{\rm fa} & A_{\rm f}A_{\rm fa} & \ldots & (A_{\rm f})^{2nN-2}A_{\rm fa} \end{bmatrix},$$

$$Q^* = \begin{bmatrix} B_{\rm a} & A_{\rm a}B_{\rm a} & A_{\rm a}^2B_{\rm a} & \cdots & A_{\rm a}^{(2nN-2)}B_{\rm a} \\ 0 & B_{\rm a} & A_{\rm a}B_{\rm a} & \cdots & A_{\rm a}^{(2nN-3)}B_{\rm a} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & B_{\rm a} & A_{\rm a}B_{\rm a} \\ 0 & 0 & \cdots & 0 & B_{\rm a} \end{bmatrix}.$$

The rows of the matrices $M_k = \sum_{z=0}^{k-1} A_{\rm f}^z A_{\rm fa} Q_{k-1-z}$, $k = 1, \ldots, 2nN-1$, are equal to the rows of M^* multiplied by the columns of Q^*. The matrices M_k not having any zero column is equivalent to them not having any basis of $\ker(M^*)$ in common with basis of $\text{Im}(Q^*)$. In other words, $\ker(M^*) \cap \text{Im}(Q^*) = 0$. This condition along with the fact that the number of rows of M^* is smaller than the dimensions of Q^*, in turn imply that $\text{rank}(M) = \text{rank}(M^*)$. Consequently, for M to be right invertible, M^* should be of full row rank, which is satisfied if the pair $(A_{\rm f}, A_{\rm fa})$ is controllable.

Considering $w_1 \in \ker(B_{\mathrm{a}}^\top)$ and an appropriate matrix w_2, one has

$$\begin{bmatrix} w_1^\top & 0 \\ 0 & w_2^\top \end{bmatrix} \begin{bmatrix} B_{\mathrm{a}} & Q \\ 0 & M \end{bmatrix} = \begin{bmatrix} 0 & w_1^\top Q \\ 0 & w_2^\top M \end{bmatrix}. \tag{154}$$

Rows of $w_1^\top Q$ and $w_2^\top M$ should not be linearly dependent to have a right invertible $\mathcal{C}^*$. This is satisfied if $w_1^\top Q \neq w_2^\top M$ for every w_2. This implies there does not exist any $w_2^\top$ such that the rows of $w_1^\top Q$ and $w_2^\top M$ are linearly dependent if $\ker(M) \nsubseteq \ker(w_1^\top Q)$. This condition is satisfied if $\ker(M) \cap \ker(w_1^\top Q) = 0$. For the latter condition to be satisfied the following matrix

$$\begin{aligned} S &= \begin{bmatrix} M_1 & \dots & M_{(2nN-1)} \end{bmatrix} \times \begin{bmatrix} Q_1^\top w_1 \\ \vdots \\ Q_{(2nN-1)}^\top w_1 \end{bmatrix} \\ &= \sum_{k=1}^{2nN-1} M_k Q_k^\top w_1, \end{aligned} \tag{155}$$

should be either full row rank if $N_{\mathrm{f}} \leq N_{\mathrm{a}}$ or full column rank if $N_{\mathrm{a}} < N_{\mathrm{f}}$. This completes the proof of the theorem. □

Remark 6.8. *Since the conditions in Theorem 6.3 are difficult to verify the adversary may not be able to gain control over the entire network as described in Definition 6.6. In such a scenario the adversary is capable of injecting its attack signals to the directly targeted agents and control the followers through them. In this type of attack, states of the directly attacked agents are used as control inputs to the followers.*

The definition of controllability of MAS followers is provided below.

Definition 6.7. *The followers (151) are controllable through the directly attacked agents by the adversary if for every x_f^* and every finite $T > 0$, there exists a proper $x_a(t)$, $0 < t < T$, such that the state transitions can be accomplished from $x_f(0) = 0$ to $x_f(T) = x_f^*$.*

Assumption 6.4. *The set of eigenvectors of L_f span $\mathbb{R}^{N_f}$.*

It should be noted that the grounded Laplacian matrix in case of a directed graph is not necessarily diagonalizable. For example, consider the Laplacian matrix $L = [1, 0, 0, -1; -1, 1, 0, 0; 0, -1, 1, 0; 0,$

0, -1, 1] and its corresponding grounded Laplacian matrix $L_{\mathrm{f}} = [1, 0, 0; -1, 1, 0; 0, -1, 1]$, where the agent 4 is directly under cyber-attack. The algebraic multiplicity of the eigenvalue of L_{f}, namely $\lambda = 1$, is 3, however, its geometric multiplicity is 1, implying that L_{f} is not diagonalizable. Since in the Theorem 6.4 provided below one requires L_{f} to be diagonalizable, the above Assumption 6.4 is given.

Proposition 6.1 ([143]). *The system $\dot{x}_i(t) = Ax_i(t) + Bu_i(t)$ is controllable if and only if $\forall\ v_k$, $k = 1, ..., n$, where v_k is the k-th eigenvector of A, $v_k \notin ker(B^T)$.*

Theorem 6.4. *Under Assumption 6.4 the adversary is capable of controlling the system (151) through the directly attacked agents (150) according to the Definition 6.7 if and only if the pairs $(\check{A}+\check{B}\check{K}+d_i\check{H}\check{C}, \check{H}_a)$, (L_f, l_{fa}), and $(\check{A} + \check{B}\check{K} + \lambda_j\check{H}\check{C}, \check{H}\check{C})$ are controllable for $i = N_f + 1, ..., N$ and $j = 1, ..., N_f$, where λ_j is the jth eigenvalue of L_f.*

Proof. In this attack scenario, the adversary uses $x_{\mathrm{a}}(t)$ as the control input to the followers. Hence, the adversary should be capable of setting $x_{\mathrm{a}}(t)$ to its desired value, which can be achieved if (150) is controllable. Consequently, the followers in (151) should be controllable through $x_{\mathrm{a}}(t)$. Since (150) is considered to be controllable, the adversary is capable of designing $a(t)$ such that $x_{\mathrm{a}}(t)$ tracks its desired trajectory (see Theorem 5.2.5 and Corollary 5.2.6 in [144]).

The controllability condition of the pair $(A_{\mathrm{a}}, B_{\mathrm{a}})$ was studied in Theorem 6.3. Controllability of $(A_{\mathrm{f}}, A_{\mathrm{fa}})$ indicates that the followers are controlled via the state of the attacked agents, $x_{\mathrm{a}}(t)$. In view of Assumption 6.4, there always exists an invertible matrix P, with its rows representing N_{f} right eigenvectors of L_{f}, such that $PL_{\mathrm{f}}P^{-1} = \mathrm{diag}(\lambda_1, ..., \lambda_{N_{\mathrm{f}}})$. Using the similarity transformation $P \otimes I_n$, (151) can be rewritten as

$$\dot{x}_{\mathrm{f}}^{\mathrm{p}}(t) = (P \otimes I_n)A_{\mathrm{f}}(P^{-1} \otimes I_n)x_{\mathrm{f}}^{\mathrm{p}}(t) + (P \otimes I_n)(l_{\mathrm{fa}} \otimes \check{H}\check{C})x_{\mathrm{a}}(t), \tag{156}$$

where $x_{\mathrm{f}}^{\mathrm{p}}(t) = (P \otimes I_n)x_{\mathrm{f}}(t)$. Since P is nonsingular, the controllability of $x_{\mathrm{f}}^{\mathrm{p}}(t)$ implies the controllability of $x_{\mathrm{f}}(t)$. The matrix $(P \otimes I_n)A_{\mathrm{f}}(P^{-1} \otimes I_n) = \mathrm{blockdiag}(\check{A} + \check{B}\check{K} + \lambda_1\check{H}\check{C}, ..., \check{A} + \check{B}\check{K} + \lambda_{N_{\mathrm{f}}}\check{H}\check{C})$ is block diagonal and $(P \otimes I_n)(l_{\mathrm{fa}} \otimes \check{H}\check{C}) = Pl_{\mathrm{fa}} \otimes \check{H}\check{C}$. Consequently, (156) can be expressed in the

following form:

$$\begin{aligned}\dot{x}_{\mathrm{f}}^{p}(t) =&\text{blockdiag}(\check{A}+\check{B}\check{K}+\lambda_1\check{H}\check{C},...,\ \check{A}+\check{B}\check{K}+\lambda_{N_{\mathrm{f}}}\check{H}\check{C})\\ &\times x_{\mathrm{f}}^{\mathrm{p}}(t)+(Pl_{\mathrm{fa}}\otimes\check{H}\check{C})x_{\mathrm{a}}(t).\end{aligned} \tag{157}$$

Since the rows of P are the right eigenvectors of L_{f}, in view of Proposition 6.1, the controllability of $(L_{\mathrm{f}},\ l_{\mathrm{fa}})$ can be interpreted as not having completely zero rows in the matrix Pl_{fa}. The vector $x_{\mathrm{f}}(t)$ contains the states of N_{f} followers, however, due to the similarity transformation, $x_{\mathrm{f}}^{\mathrm{p}}(t)$ contains a combination of these states, but still one has N_{f} modes that are the N_{f} different blocks of blockdiag$(\check{A}+\check{B}\check{K}+\lambda_1\check{H}\check{C},...,\ \check{A}+\check{B}\check{K}+\lambda_{N_{\mathrm{f}}}\check{H}\check{C})$. Next we provide and prove the necessary and sufficient conditions of our proposed methodology that are stated in this theorem.

Necessary Condition: Assume the j-th mode of (157) is controllable through $x_{\mathrm{a}}(t)$, while either $(\check{A}+\check{B}\check{K}+\lambda_j\check{H}\check{C},\check{H}\check{C})$ is not controllable or the j-th row of Pl_{fa} is zero. Due to block diagonal structure of (157), either the uncontrollability of $(\check{A}+\check{B}\check{K}+\lambda_j\check{H}\check{C},\check{H}\check{C})$ or the j-th row of Pl_{fa} being zero results in the uncontrollability of the mode j, which contradicts the assumption on this mode.

Sufficient Condition: Suppose that the mode j is uncontrollable, while $(\check{A}+\check{B}\check{K}+\lambda_j\check{H}\check{C},\check{H}\check{C})$ is controllable and the j-th row of Pl_{fa} is nonzero. However, from the block diagonal structure of (157), the mode j being uncontrollable implies that either $(\check{A}+\check{B}\check{K}+\lambda_j\check{H}\check{C},\check{H}\check{C})$ is uncontrollable or the j-th row of Pl_{fa} is zero, which is a contradiction. This completes the proof of the theorem. □

Remark 6.9. *As shown in Theorem 6.4, the problem of interest here is to show that there exists a proper $x_a(t)$ that satisfies the controllability objective provided in Definition 6.7. However, designing the attack signal $a(t)$ such that $x_a(t)$ follows the adversary's desired trajectory is not within the scope of this chapter and is not addressed here.*

Remark 6.10. *Generally speaking, the difference between the goals in Definitions 6.6 and 6.7 has resulted in different types of conditions that need to be satisfied in Theorems 6.3 and 6.4. In Theorem 6.3, the conditions are more restrictive, however they ensure controllability over the entire network for the adversary. Nevertheless, the main objective of the malicious hacker is to exert the maximum possible influence on the MAS given the available resources. Consequently, the adversary may not be able to control the entire*

network as studied in Theorem 6.3, whereas they can still compromise the system and lead the MAS to dangerous trajectories only if the conditions in Theorem 6.4 are satisfied. This result is illustrated through the numerical example that is provided in Section 6.7.2.

6.4.2 Cybersecurity Controllability Index

As shown in Theorem 6.4, the only condition on controllability of the MAS that connects the structure of the communication graph among the followers and the directly attacked agents is the controllability of $(L_{\mathrm{f}}, l_{\mathrm{fa}})$. By leveraging this controllability condition, we aim to define two security metrics for the MAS. These notions can be used to evaluate the security of the MAS with respect to their controllability by an adversarial intruder. In this subsection, we assume that all the conditions in Theorem 6.4, except for the controllability of $(L_{\mathrm{f}}, l_{\mathrm{fa}})$, hold true. Let us denote $\hat{L}_{\mathrm{f}} = PL_{\mathrm{f}}P^{-1} = \mathrm{diag}(\lambda_1, ..., \lambda_{N_{\mathrm{f}}})$ and $\hat{l}_{\mathrm{fa}} = Pl_{\mathrm{fa}}$, where rows of P are the N_{f} right eigenvectors of L_{f}.

Definition 6.8. *The security controllability index of the directly attacked agent i, designated by SCI_i, is defined by:*

$$SCI_i = rank(\mathcal{C}_i), \quad i = N_f + 1, ..., N, \tag{158}$$

where $\mathcal{C}_i = [(\hat{l}_{fa})_i, \hat{L}_f(\hat{l}_{fa})_i, \ldots, \hat{L}_f^{N_f-1}(\hat{l}_{fa})_i]$ denotes the controllability matrix (considered not to be ill-conditioned) and $(\hat{l}_{fa})_i$ denotes the i-th column of $\hat{l}_{fa}$.

The maximum value for SCI_i can be N_{f}, which if satisfied implies that all the followers can be manipulated and controlled via the agent i.

Definition 6.9. *The security controllability index (SCI) of the MAS is defined as*

$$SCI = rank(\mathcal{C}), \tag{159}$$

where $\mathcal{C} = [\hat{l}_{fa}, \hat{L}_f\hat{l}_{fa}, \ldots, \hat{L}_f^{N_f-1}\hat{l}_{fa}]$.

The problem for the adversary is to find the minimum number of directly attacked agents that gives the full control over the multi-agent network. More specifically, the adversary's goal is to minimize $|N_{\mathrm{a}}|$ such that $SCI = N_{\mathrm{f}}$. In the literature this problem is referred to as actuator placement problem [145]. Solving

the above minimization problem provides the adversary with the minimum required number of agents that the hacker needs to compromise and attack. A few methods that incorporate graph of the network to select agents for ensuring controllability over the MAS have been suggested in [146–150].

Remark 6.11. *Due to the possibility of existence of sufficiently small singular values and ill-conditioning of the matrices $\mathcal{C}$ and $\mathcal{C}_i$ for $i = N_f + 1, ..., N$ in (158) and (159), one may have nearly singular matrices. In such cases rank($\mathcal{C}$) and rank($\mathcal{C}_i$) can be computed by imposing a tolerance condition on computation of the rank such that if the singular value is smaller than a pre-specified tolerance level it is then considered to be zero.*

6.4.3 Zero Dynamics Attacks Through the Communication Links

Given an $s = s_{\mathrm{a}}$ and the dynamics of the directly attacked agents in (150), the zero dynamics of the MAS are those s_{a} in which the Rosenbrock system matrix

$$P_{\mathrm{a}}(s) = \begin{bmatrix} sI - A_{\mathrm{a}} & -(I_{N_{\mathrm{a}}} \otimes \check{H}_{\mathrm{a}}) \\ I_{N_{\mathrm{a}}} \otimes C & 0 \end{bmatrix} \tag{160}$$

is rank deficient, i.e., its rank falls below its normal rank. This implies that there exist nonzero $x_{\mathrm{a}0}$ and a_0 such that

$$\begin{bmatrix} s_{\mathrm{a}}I - A_{\mathrm{a}} & -(I_{N_{\mathrm{a}}} \otimes \check{H}_{\mathrm{a}}) \\ I_{N_{\mathrm{a}}} \otimes \check{C} & 0 \end{bmatrix} \begin{bmatrix} x_{\mathrm{a}0} \\ a_0 \end{bmatrix} = 0, \tag{161}$$

where $X_{\mathrm{a}}(t) = x_{\mathrm{a}0}e^{s_{\mathrm{a}}t}$ with $X_{\mathrm{a}}(t)$ defined as the solution to (150) and $a(t) = a_0 e^{s_{\mathrm{a}}t}$.

The zero dynamics of the followers (151) are defined as $s = s_{\mathrm{f}}$ and are associated with nonzero directional vectors $x_{\mathrm{f}0}$ and x_{af} such that the following is satisfied:

$$\begin{bmatrix} s_{\mathrm{f}}I - A_{\mathrm{f}} & -(l_{\mathrm{fa}} \otimes \check{H}\check{C}) \\ I_{N_{\mathrm{f}}} \otimes \check{C} & 0 \end{bmatrix} \begin{bmatrix} x_{\mathrm{f}0} \\ x_{\mathrm{af}} \end{bmatrix} = 0, \tag{162}$$

where $X_{\mathrm{f}}(t) = x_{\mathrm{f}0}e^{s_{\mathrm{f}}t}$ with $X_{\mathrm{f}}(t)$ as the solution to (151) and $X_{\mathrm{a}}(t) = x_{\mathrm{af}}e^{s_{\mathrm{f}}t}$.

Definition 6.10. *The zero dynamics s_a and s_f are excited in the systems* (150) *and* (151) *if their initial conditions and the attack signal satisfy the conditions in* (161) *and* (162)*, respectively.*

From (161) and (162) one can conclude that the differences that exist between the attacked agents and the followers can result in having different zero dynamics in these two groups. Moreover, in case of an attacker exciting the zero dynamics, the states should satisfy $x_i(t) \in \ker(C)$, for $i = 1, \ldots, N$, to have a zero output in the system [15].

Lemma 6.5. *The zero dynamics of the followers* (151) *and the directly attacked agents* (150) *are excited by the adversary in the sense of Definition 6.10 if* (162) *and* (161) *for $s_a = s_f$ hold true, while $(l_{fa} \otimes \check{H}\check{C})x_{af} \neq 0$ and $(I_{N_a} \otimes \check{H}_a)a_0 \neq 0$, respectively.*

Proof. Suppose the output of the system is zero with nonzero x_{af} and a_0 that are in the $\ker(l_{\text{fa}} \otimes \check{H}\check{C})$ and $\ker(I_{N_{\text{a}}} \otimes \check{H}_{\text{a}})$, respectively. This implies that the attack signal does not have an impact on exciting the zero dynamics. Therefore, in the case of zero dynamics attack by the adversary it is necessary for the attack signals to satisfy $x_{\text{af}} \notin \ker(l_{\text{fa}} \otimes \check{H}\check{C})$ and $a_0 \notin \ker((I_{N_{\text{a}}} \otimes \check{H}_{\text{a}}))$. This completes the proof of the lemma. □

Theorem 6.5. *The adversary is not capable of simultaneously exciting the zero dynamics of the directly attacked agents* (150) *and the followers* (151) *in the sense of Definition 6.10.*

Proof. Suppose the adversary excites the zero dynamics of directly attacked agents in (150) so that $X_{\text{a}}(t) = x_{\text{a0}}e^{s_{\text{a}}t}$ is in $\ker(I_{N_{\text{a}}} \otimes \check{C})$. Consequently, x_{a0} should be of the form $x_{\text{a0}} = I_{N_{\text{a}}} \otimes \check{x}_{\text{a0}}$, where $\check{x}_{\text{a0}} \in \ker(\check{C})$. Since $(l_{\text{fa}} \otimes \check{H}\check{C}) \times (I_{N_{\text{a}}} \otimes \check{x}_{\text{a0}}) = l_{\text{fa}} \otimes \check{H}\check{C}\check{x}_{\text{a0}} = 0$ one can conclude $x_{\text{af}} = X_{\text{a}}(0) = x_{\text{a0}} \in \ker(l_{\text{fa}} \otimes \check{H}\check{C})$, which based on Lemma 6.5 implies that the adversary is not capable of exciting the zero dynamics of the followers. Now let us assume (162) holds and the zero dynamics of the followers are excited by the adversary. Therefore, $(l_{\text{fa}} \otimes \check{H}\check{C})x_{\text{af}} \neq 0$ is satisfied. This implies that $X_{\text{a}}(0) = x_{\text{af}} \notin \ker(I_{N_{\text{a}}} \otimes \check{C})$. Hence, (161) does not hold and the zero dynamics of the directly attacked agents (150) cannot be excited by the adversary. This completes the proof of the theorem. □

6.5 Undetectable cyber-attacks in MAS

In this section, conditions to introduce and inject undetectable cyber-attacks are investigated and developed. Let us define $\check{A} = I_2 \otimes A$, $\check{K} = I_2 \otimes cK$, $\check{C} = I_2 \otimes C$, and

$$\begin{aligned} \check{B} &= \begin{bmatrix} 0 & B \\ 0 & B \end{bmatrix}, \ \check{F} = \begin{bmatrix} 0 & 0 \\ cF & -cF \end{bmatrix}, \\ \check{B}_{\mathrm{a}} &= \begin{bmatrix} -BcK & 0 \\ -BcK - cFC & cF \end{bmatrix}. \end{aligned} \tag{163}$$

By utilizing (147) and augmenting the dynamics (132) and (133) yields:

$$\begin{aligned} \dot{\check{x}}_i(t) =& \check{A}\check{x}_i(t) + (\check{B}\check{K} - \check{F}\check{C}) \sum_{j \in \mathcal{N}_i} (\check{x}_i(t) - \check{x}_j(t)) \\ &+ \check{B}_{\mathrm{a}}\check{a}_i(t), \end{aligned} \tag{164}$$

where $\check{x}_i(t) = [x_i(t)^\top, \hat{x}_i(t)^\top]^\top$, $\check{a}_i(t) = \sum_{j \in \mathcal{N}_i} q_i \check{a}_{ji}$, and $\check{a}_{ji}(t) = [a_{\hat{x}}^{ji}(t)^\top, a_y^{ji}(t)^\top]^\top$.

Definition 6.11. *The cyber-attacks on communication links in the MAS* (148) *are undetectable by the entire network if the MAS reaches a consensus under attack free conditions, i.e.,* $\check{a}_i(t) = 0$*, as per Definition 6.1, and in presence of cyber-attacks, i.e.,* $\check{a}_i(t) \neq 0$*,* $\forall\, t > 0$*, the following equations are satisfied:*

$$\lim_{t \to \infty} \| y_i(t) - y_j(t) - q_i a_y^{ji}(t) \| = 0, \tag{165}$$

$$\lim_{t \to \infty} \| \hat{x}_i(t) - \hat{x}_j(t) - q_i a_{\hat{x}}^{ji}(t) \| = 0, \tag{166}$$

$\forall\, i, j = 1, \ldots, N$.

Remark 6.12. *If* $\check{a}_i(t) = 0$ *and the MAS reaches a consensus, the residuals* (134) *and* (135) *converge to zero as* $t \to \infty$ *as expected since no cyber-attack is injected to the system. If in presence of cyber-attacks, i.e.,* $\check{a}_i(t) \neq 0$*, equations* (165) *and* (166) *are satisfied, it can be inferred that* $\lim_{t \to \infty} res_y^i(t) = 0$ *and* $\lim_{t \to \infty} res_{\hat{x}}^i(t) = 0$ *for* $i = 1, \ldots, N$*. Consequently, the cyber-attack is undetectable.*

Theorem 6.6. *Under Assumptions 6.1 and 6.2, let the cyber-attacks on the agent i in (148) be selected as $a_{\hat{x}}^{ji}(t) = a_0 - \hat{x}_j(t)$ and $a_y^{ji}(t) = Ca_0 - y_j(t)$ for $j \in \mathcal{N}_i$, where $a_0 \in \mathbb{R}^n$ is a constant vector. The above cyber-attacks are undetectable by all agents if $i \in V_r$, and $A + BcK$ and $A + \tilde{\lambda}_k^r cFC$ for $k = 2, \ldots, N+1$ are Hurwitz, where $\tilde{\lambda}_k^r \neq 0$ is the k-th eigenvalue of the matrix*

$$\tilde{L}_r = \begin{bmatrix} 0 & \mathbf{0}_{1\times N} \\ -D_r & L_r \end{bmatrix},$$

where $D_r = [\mathbf{0}_{i-1}^\top,\, d_i,\, \mathbf{0}_{N-i}^\top]^\top$, $L_r = L + Q_r\mathcal{A}$, $Q_r = diag(\mathbf{0}_{i-1}, q_i, \mathbf{0}_{N-i})$, and $\mathbf{0}_{m\times n} \in \mathbb{R}^{m\times n}$ denotes a matrix with all entries equal to 0.

Proof. By adding the cyber-attack signals to (148) and augmenting the dynamics of agents, one obtains

$$\dot{\check{x}}(t) = (I_N \otimes \check{A} + L_{\mathrm{r}} \otimes (\check{B}\check{K} - \check{F}\check{C}))\check{x}(t) + D_{\mathrm{r}} \otimes \check{B}\check{a}_0,$$

where $\check{x}(t) = [\check{x}_1(t)^\top,\, \check{x}_2(t)^\top, \ldots,\, \check{x}_N(t)^\top]^\top$, $\check{a}_0 = [a_0^\top, (Ca_0)^\top]^\top$. One can consider the adversary as a virtual agent that transmits its information a_0 and Ca_0 to the i-th agent. Since from the agent i there exists paths to the rest of $N-1$ agents the virtual adversary agent is the root of a directed spanning tree that is contained in the graph $\mathcal{G}_{\mathrm{r}}$ with $\tilde{L}_{\mathrm{r}}$ as its Laplacian matrix. Hence, the state space representation of the MAS augmented with the virtual adversary agent can be represented as follows:

$$\dot{\tilde{x}}_{\mathrm{r}}(t) = (\tilde{I}_{\mathrm{r}} \otimes \check{A} + \tilde{L}_{\mathrm{r}} \otimes (\check{B}\check{K} - \check{F}\check{C}))\tilde{x}_{\mathrm{r}}(t),$$

where $\tilde{x}_{\mathrm{r}}(t) = [\mathbf{1}_2^\top \otimes a_0^\top,\, \check{x}(t)^\top]^\top$, $\tilde{I}_{\mathrm{r}} = \begin{bmatrix} 0 & \mathbf{0}_{1\times N} \\ \mathbf{0}_N & I_N \end{bmatrix}$, and $\mathbf{1}_n \in \mathbb{R}^n$ denotes a vector with all its elements equal to 1. In similar manner as in [66], the disagreement vector $\delta_{\mathrm{r}}(t) = \tilde{x}_{\mathrm{r}}(t) - (\mathbf{1}_{N+1}\tilde{r}_{\mathrm{r}}^\top \otimes I_{2n})\tilde{x}_{\mathrm{r}}(t)$ is defined, where $\tilde{r}_{\mathrm{r}}$ is the left eigenvector of $\tilde{L}_{\mathrm{r}}$ as defined in the Lemma 6.3. Consequently, we obtain

$$\dot{\delta}_r(t) = (\tilde{I}_{\mathrm{r}} \otimes \check{A} + \tilde{L}_{\mathrm{r}} \otimes (\check{B}\check{K} - \check{F}\check{C}))\delta_{\mathrm{r}}(t). \tag{167}$$

It can be shown from the definition of $\delta_r(t)$ that $\delta_r(t) = 0$ if and only if $\mathbf{1}_2 \otimes a_0 = \check{x}_1(t) = \check{x}_2(t) =$

$\cdots = \check{x}_N(t)$. Therefore, if (167) is stable, one can conclude that all the agents reach a_0 corresponding to the consensus set point.

From Lemma 6.3 and considering that $\mathcal{G}_r$ contains a directed spanning tree it follows that there exist matrices $T \in \mathbb{R}^{N+1\times N+1}$, $Y \in \mathbb{R}^{N+1\times N}$, $W \in \mathbb{R}^{N\times N+1}$, and block diagonal matrix $\Delta \in \mathbb{R}^{N\times N}$ with diagonal entries equal to nonzero eigenvalues of $\tilde{L}_{\mathrm{r}}$ such that [66]:

$$T = [\mathbf{1}_{N+1}\, Y],\ T^{-1} = \begin{bmatrix} \tilde{r}^\top \\ W \end{bmatrix},\ T^{-1}\tilde{L}T = J = \begin{bmatrix} 0 & \mathbf{0}_{1\times N} \\ \mathbf{0}_N & \Delta \end{bmatrix}.$$

Let us define $\varepsilon(t) = (T^{-1} \otimes I_{2n})\delta_{\mathrm{r}}(t) = [\varepsilon_1(t)^\top, \varepsilon_{2:N+1}(t)^\top]^\top$. It follows from Lemma 6.3 and definition of $\delta_{\mathrm{r}}(t)$ that $\varepsilon_1(t) = \mathbf{0}_{2n}$ and

$$\dot{\varepsilon}_{2:N+1}(t) = (I_N \otimes \check{A} + \Delta \otimes (\check{B}\check{K} - \check{F}\check{C}))\varepsilon_{2:N+1}(t). \tag{168}$$

It is easy to show that the matrices $I_N \otimes \check{A} + \tilde{\lambda}_k^{\mathrm{r}} \otimes (\check{B}\check{K} - \check{F}\check{C})$ for $k = 2, \ldots, N+1$ having a similar structure to $\begin{bmatrix} A + \tilde{\lambda}_k^{\mathrm{r}} cFC & \mathbf{0}_{n\times n} \\ -\tilde{\lambda}_k^{\mathrm{r}} cFC & A + BcK \end{bmatrix}$.

Consequently, if $A + BcK$ and $A + \tilde{\lambda}_k^{\mathrm{r}} cFC$ for $k = 2, \ldots, N+1$ are Hurwitz, (167) is stable and states of all the agents reach a_0 as $t \to \infty$. Hence, according to the Definition 6.11, the cyber-attacks are undetectable from the entire network. □

6.5.1 Cyber-Attacks Injected to Non-Root Agents

In this subsection, cyber-attacks that are injected to the communication links of agents that do not belong to the set V_{r} are investigated.

Definition 6.12. *The agents in the network that are directly under cyber-attacks and their incoming communication channels are compromised are included in the set $V_{da} = \{i_a, i_a + 1, \ldots, i_a + N_{da} - 1\}$, where N_{da} is the number of directly attacked agents.*

Definition 6.13 ([151])**.** *The reachable subgraph of the vertex i, $R(i)$, is defined as the vertex subgraph that contains the node i and all reachable nodes from it.*

Definition 6.14. *The directed path between the vertices i and j, denoted by $\mathcal{P}_{ij}$, is a directed healthy path if none of the communication links on this path is compromised by adversaries.*

By utilizing the Definitions 6.12-6.14, agents in the network can be partitioned into two groups. For agents from the root nodes of the directed tree, i.e., the set $V_{\rm r}$, there exist directed healthy paths that constitute the first group. Agents in this group are designated as "*uncompromised*" where their states can be stacked into the vector $x_{\rm nc}(t) = [\check{x}_1(t)^\top,\, \check{x}_2(t)^\top, \ldots,\, \check{x}_{N_{\rm nc}}(t)^\top]^\top$. The second group consists of the remaining network agents. This group contains agents in the vertex subgraphs of directly attacked agents, i.e., the set $V_{\rm da}$, such that there does not exist any directed healthy path among agents in $V_{\rm r}$ and these nodes. The agents that belong to this group are designated as "*attacked*" agents and $x_{\rm a}(t) = [\check{x}_{N_{\rm nc}+1}(t)^\top,\, \check{x}_{N_{\rm nc}+2}(t)^\top, \ldots,\, \check{x}_N(t)^\top]^\top$ represents the states of this group. The subscripts "*nc*" and "*a*" are employed to denote the $N_{\rm nc}$ uncompromised agents and the $N_{\rm a}$ attacked agents, respectively. Without loss of any generality, we assume that the first $N_{\rm nc}$ agents are not under cyber-attacks. Consequently, the Laplacian matrix is partitioned into the following form:

$$L = \begin{bmatrix} L_{\rm nc} & l_{\rm nca} \\ l_{\rm anc} & L_{\rm a} \end{bmatrix}, \tag{169}$$

where $L_{\rm nc} \in \mathbb{R}^{N_{\rm nc}\times N_{\rm nc}}$, $L_{\rm a} \in \mathbb{R}^{N_{\rm a}\times N_{\rm a}}$, $l_{\rm nca} \in \mathbb{R}^{N_{\rm nc}\times N_{\rm a}}$, and $l_{\rm anc} \in \mathbb{R}^{N_{\rm a}\times N_{\rm nc}}$.

The state space representation of the entire MAS can now be described in the following form:

$$\dot{x}_{\rm a}(t) = A_{\rm a}x_{\rm a}(t) + A_{\rm anc}x_{\rm nc}(t) + B_{\rm a}a(t), \tag{170}$$

$$\dot{x}_{\rm nc}(t) = A_{\rm nc}x_{\rm nc}(t) + A_{\rm nca}x_{\rm a}(t), \tag{171}$$

where $A_{\rm a} = I_{N_{\rm a}} \otimes \check{A} + L_{\rm a} \otimes (\check{B}\check{K} - \check{F}\check{C})$, $A_{\rm anc} = l_{\rm anc} \otimes (\check{B}\check{K} - \check{F}\check{C})$, $A_{\rm nc} = I_{N_{\rm nc}} \otimes \check{A} + L_{\rm nc} \otimes (\check{B}\check{K} - \check{F}\check{C})$, $A_{\rm nca} = l_{\rm nca} \otimes (\check{B}\check{K} - \check{F}\check{C})$, $B_{\rm a} = Q_{\rm a} \otimes \check{B}_{\rm a}$, and $Q_{\rm a} = {\rm diag}(q_{N_{\rm nc}+1},\, q_{N_{\rm nc}+2}, \ldots,\, q_N) \in \mathbb{R}^{N_{\rm a}\times N_{\rm a}}$ is a diagonal matrix.

Definition 6.15. *The set of N agents is partitioned into two subsets, namely $V_{nc} = \{1, 2, \ldots, N_{nc}\}$ and $V_a = \{N_{nc}+1,\, N_{nc}+2, \ldots, N\}$, that contain uncompromised and attacked agents, respectively.*

Definition 6.16. *The set of uncompromised agents that receive information from the set of attacked agents are defined by $V_{nca} = \{i_{nca},\, i_{nca}+1, \ldots, i_{nca}+N_{nca}-1\}$, where N_{nca} is the number of uncompromised*

agents that receive information from the agents in V_a. In other words, $j \in V_{nca}$ if and only if $j \in V_{nc}$ and there exists $i \in V_a$ such that $i \in \mathcal{N}_j$.

Assumption 6.5. *None of agents in the set V_r is directly under attack, i.e., $V_r \cap V_{da} = 0$.*

Without loss of generality, let us assume that the first N_{da} agents in the set V_{a} create the set of directly attacked agents. We are now in a position to state our main result of this subsection.

Theorem 6.7. *Let the Assumptions 6.1-6.5 hold and the cyber-attacks on agents in the set V_{da} are designed as $a_{\hat{x}}^{ji}(t) = a_0 - \hat{x}_j(t)$ and $a_y^{ji}(t) = Ca_0 - y_j(t)$ for $j = 1, \ldots, N_{nc}$ and $j \in \mathcal{N}_i$, where $Ca_0 \neq \lim_{t\to\infty} y_{i_r}(t)$, $a_0 \neq \lim_{t\to\infty} x_{i_r}(t)$, and $a_0 \in \mathbb{R}^n$ is a constant vector. Consequently, the following can be stated:*

(1) The cyber-attacks are undetectable on the set of nodes V_a if $A + BcK$ and $A + \tilde{\lambda}_k cFC$ for $k = 2, \ldots, N_a + 1$ are Hurwitz, where $\tilde{\lambda}_k \neq 0$ denotes the k-th eigenvalue of the Laplacian matrix

$$\tilde{L}_a = \begin{bmatrix} 0 & \mathbf{0}_{1\times N_a} \\ -D_a & L_a \end{bmatrix},$$

with $D_a = [d_{i_a}^{nc}, d_{i_a+1}^{nc}, \ldots, d_{i_{a+N_{da}-1}}^{nc}, \mathbf{0}_{N_a-N_{da}}^{\top}]^{\top}$ and $d_i^{nc} = |\mathcal{N}_i \cap V_{nc}|$ for $i \in V_{da}$.

(2) The cyber-attacks are detectable by agents that belong to the set V_{nca}.

Proof. Substituting the cyber-attack signals into (170) the following can be expressed:

$$\dot{x}_{\text{a}}(t) = (I_{N_{\text{a}}} \otimes \check{A} + L_{\text{a}} \otimes (\check{B}\check{K} - \check{F}\check{C}))x_{\text{a}}(t) + D_{\text{a}} \otimes \check{B}_{\text{a}}\check{a}_0, \tag{172}$$

where $\check{a}_0 = [a_0^{\top}, (Ca_0)^{\top}]^{\top}$. It follows from (172) that the nodes which belong to the set V_{a} do not receive information from the nodes in the set V_{nc}. Since the cyber-attack signal a_0 is the same for all agents in the set V_{da}, one can consider the adversary as a virtual agent that is the root of a directed spanning tree contained in the graph $\mathcal{G}_{\text{a}}$ with $\tilde{L}_{\text{a}}$ as its Laplacian matrix. The virtual adversary agent transmits its information a_0 and Ca_0 to all the directly attacked agents. Consequently, the dynamics of the attacked agents augmented with

the virtual adversary agent can be derived as given below:

$$\dot{\tilde{x}}_{\mathrm{a}}(t) = (\tilde{I}_{\mathrm{a}} \otimes \check{A} + \tilde{L}_{\mathrm{a}} \otimes (\check{B}\check{K} - \check{F}\check{C}))\tilde{x}_{\mathrm{a}}(t),$$

where $\tilde{x}_{\mathrm{a}}(t) = [\mathbf{1}_2^\top \otimes a_0^\top,\, x_{\mathrm{a}}(t)^\top]^\top$ and $\tilde{I}_{\mathrm{a}} = \begin{bmatrix} 0 & \mathbf{0}_{1\times N_a} \\ \mathbf{0}_{N_a} & I_{N_{\mathrm{a}}} \end{bmatrix}$.

Following along the same steps as in the proof of the Theorem 6.6 it can be concluded that if $A + BcK$ and $A + \tilde{\lambda}_k cFC$, with $k = 2, \ldots, N_{\mathrm{a}} + 1$, are Hurwitz, the states of all agents in the set V_{a} reach a_0 as $t \to \infty$. Therefore, according to the Definition 6.11 the cyber-attacks are undetectable by agents that belong to the set V_{a}.

Suppose all agents in (170) and (171) reach a consensus. Following along the results previously derived it follows that states of the attacked agents, and consequently uncompromised agents, should reach a_0, which contradicts the assumption that $Ca_0 \neq \lim_{t\to\infty} y_{i_{\mathrm{r}}}(t)$ and $a_0 \neq \lim_{t\to\infty} x_{i_{\mathrm{r}}}(t)$. Therefore, the entire network cannot reach a consensus. The residuals for the agent $j \in V_{\mathrm{nc}}$ with $i \in V_{\mathrm{a}}$ and $i \in \mathcal{N}_j$ are nonzero as $t \to \infty$ since $\hat{x}_i(t) \neq \hat{x}_j(t)$ and $y_i(t) \neq y_j(t)$. This completes the proof of the theorem. □

Remark 6.13. *Suppose in Theorem 6.7 one has $Ca_0 = \lim_{t\to\infty} y_{i_r}(t)$ and $a_0 = \lim_{t\to\infty} x_{i_r}(t)$, which implies that the attack signals will not change the consensus set point and follow the control objective of the MAS. Hence, it is reasonable to assume that the control objective of adversaries differs from the group of agents, in other words $Ca_0 \neq \lim_{t\to\infty} y_{i_r}(t)$ and $a_0 \neq \lim_{t\to\infty} x_{i_r}(t)$.*

6.5.2 Quasi-Covert Cyber-Attack on the MAS Network

In the Theorem 6.7, it was shown that cyber-attack signals are detectable by agents that belong to the set V_{nc} and receive information from agents in the set V_{a}. Hence, if an adversary attacks the communication channels that connect agents in these two sets, they may be able to manipulate the transmitted information such that impacts of cyber-attacks are made hidden and eliminated. In this chapter, this attack methodology where impacts of cyber-attacks on the set V_{nc} are eliminated is denoted by the "*quasi-covert cyber-attacks*".

Assumption 6.6. *The adversary has knowledge on the parameters of the MAS* (127) *and the observer-based consensus protocol* (131).

To conceal impacts of cyber-attacks on the set $V_{\rm nc}$ the adversary needs to attack the outgoing communication links from the nodes in the set $V_{\rm a}$ to the agents that belong to $V_{\rm nc}$. These agents are included in the set $V_{\rm nca}$. To achieve this, under Assumption 6.6 the adversary is capable of running the following process:

$$\dot{x}_i^{\rm c}(t) = Ax_i^{\rm c}(t) + BcK\epsilon_i^{\rm c}(t), \tag{173}$$

$$\dot{\hat{x}}_i^{\rm c}(t) = A\hat{x}_i^{\rm c}(t) + BcK\epsilon_i^{\rm c} - cF\zeta_i^{\rm c}(t), \tag{174}$$

for $i = N_{\rm nc}+1,\, N_{\rm nc}+2, \ldots,\, N$, where $x_i^{\rm c}(t),\, \hat{x}_i^{\rm c}(t) \in \mathbb{R}^n$, for any arbitrary initial conditions $x_i^{\rm c}(0)$ and $\hat{x}_i^{\rm c}(0)$, and

$$\epsilon_i^{\rm c}(t) = \sum_{k\in\mathcal{N}_i\cap V_{\rm a}} (\hat{x}_i^{\rm c}(t) - \hat{x}_k^{\rm c}(t)) + \sum_{j\in\mathcal{N}_i\cap V_{\rm nc}} (\hat{x}_i^{\rm c}(t) - \hat{x}_j(t)),$$

$$\zeta_i^{\rm c}(t) = \sum_{k\in\mathcal{N}_i\cap V_{\rm a}} ((Cx_k^{\rm c}(t) - Cx_i^{c}(t)) + C(\hat{x}_i^{\rm c} - \hat{x}_k^{\rm c}(t))) + \sum_{j\in\mathcal{N}_i\cap V_{\rm nc}} ((y_j(t) - Cx_i^{c}(t)) + C(\hat{x}_i^{\rm c} - \hat{x}_j(t))).$$

Lemma 6.6. *Let the Assumptions 6.1-6.6 hold and agents in the set V_{da} are under cyber-attacks as specified in the Theorem 6.7. The adversary is capable of eliminating impacts of these cyber-attacks and make them undetectable on the set V_{nca} by adding cyber-attack signals $\hat{a}_{\hat{x}}^{ij}(t) = \hat{x}_i^{c}(t) - \hat{x}_i(t)$ and $\hat{a}_{y}^{ij}(t) = Cx_i^{c}(t) - y_i(t)$ for $i \in V_a$ and $j \in V_{nca}$ to the outgoing communication channels of agents in the set V_a to agents that belong to V_{nca}.*

Proof. Let us define $\check{x}_i^{\rm c}(t) = [x_i^{\rm c}(t)^\top,\, \hat{x}_i^{\rm c}(t)^\top]^\top$ for $i \in V_{\rm a}$ and $x_{\rm c}(t) = [x_{N_{\rm nc}+1}^{\rm c}(t)^\top,\, x_{N_{\rm nc}+2}^{\rm c}(t)^\top, \ldots,\, x_N^{\rm c}(t)^\top]^\top$. Consequently, the dynamics of $x^{\rm c}(t)$ can be derived as follows:

$$\dot{x}_{\rm c}(t) = A_{\rm a}x_{\rm c}(t) + A_{\rm anc}x_{\rm nc}(t). \tag{175}$$

By adding the cyber-attack signals $\hat{a}_{\hat{x}}^{ij}(t)$ and $\hat{a}_{y}^{ij}(t)$ to the outgoing communication channels of agents in the set $V_{\rm a}$, (171) can be reformulated in the following form:

$$\dot{x}_{\rm nc}(t) = A_{\rm nc}x_{\rm nc}(t) + A_{\rm nca}x_{\rm c}(t). \tag{176}$$

One can augment (175) and (176) as follows:

$$\dot{\bar{x}}(t) = (I_N \otimes \check{A} + L \otimes (\check{B}\check{K} - \check{F}\check{C}))\bar{x}(t), \tag{177}$$

where $\bar{x}(t) = [x_{\text{nc}}(t)^\top, x_{\text{c}}(t)^\top]^\top$. If the observer-based control protocol is designed such that the MAS reaches a consensus, then the augmented dynamics in (177) also reaches a consensus. Hence, according to the Definition 6.11, the cyber-attacks on nodes in V_{da} are undetectable by the entire MAS and their impacts on agents in V_{nc} are eliminated. □

6.6 Event-Triggered Cyber-Attack Detection Methodology

6.6.1 Event-Triggered Detector Module

Our objective in this section is to interrupt disclosure capabilities of the adversary by employing an event-triggered protocol of information exchange among our proposed detector modules.

The event-triggered detector for the i-th agent is designed as follows:

$$\begin{aligned}\dot{z}_i(t) =& A_{\text{z}} z_i(t) + B_{\text{z}} \hat{x}_i(t) + F_{\text{z}} \sum_{j \in \mathcal{N}_i} (e^{A_{\text{z}}(t - t^j_{k_j})} z_j(t^j_{k_j}) \\ & - e^{A_{\text{z}}(t - t^i_{k_i})} z_i(t^i_{k_i}) + q^{\text{z}}_i a^{ji}_{\text{z}}(t^j_{k_j})), \quad i = 1, \ldots, N,\end{aligned} \tag{178}$$

where $z_i(t) \in \mathbb{R}^n$ is the state of detector for agent i, $t^i_{k_i}$ denotes the time of the most recent triggering event of the agent i, $k_i \in \mathbb{N}$ indicates the k_i-th event on the agent i, $z_i(t^i_{k_i})$ denotes the latest broadcast state of the detector for the i-th agent, A_{z} is a diagonal Hurwitz matrix, and the matrices $(A_{\text{z}}, B_{\text{z}}, F_{\text{z}})$ are of appropriate dimensions that should be designed. Also $q^{\text{z}}_i = 1$ indicates that the i-th detector is under cyber-attacks, and $q^{\text{z}}_i = 0$ if it is not under attack, and $a^{ji}_{\text{z}}(t^j_{k_j}) \in \mathbb{R}^n$ denotes the cyber-attack signal on the received information from the neighboring agents. It is worth noting that given the detector (178) and the consensus-based observer (131) the agents receive output measurement information, observer states, and detector states from their neighboring agents.

Similar to [92], we define the state error for the i-th agent as

$$e_i^{\mathrm{z}}(t) = e^{A_{\mathrm{z}}(t-t^i_{k_i})} z_i(t^i_{k_i}) - z_i(t), \quad t \in [t^i_{k_i}, t^i_{k_i+1}). \tag{179}$$

Moreover, the triggering function on the agent i is defined as

$$f_i^{\mathrm{z}}(t, e_i^{\mathrm{z}}(t)) = \|e_i^{\mathrm{z}}(t)\| - c_{\mathrm{z}} e^{-\alpha t}, \tag{180}$$

where c_{z} and α are positive constants to be selected and designed. The triggering function (180) determines the occurrence of an event by the agent i. Hence, $f_i^{\mathrm{z}}(t, e_i^{\mathrm{z}}(t)) \geq 0$ implies that an event is triggered by the agent i. Consequently, this agent updates its detector's state such that $z_i(t^i_{k_i+1}) = z_i(t)$, and $t^i_{k_i+1} = t$. Subsequently, the updated state $z_i(t^i_{k_i+1})$ is broadcast to agents that agent i belongs to their neighboring set and they use the updated state in their detectors. Due to an event by the i-th agent, the state error (179) of this agent is reset to zero.

One can augment states of detectors into the vector $z(t) = [z_1(t)^\top, z_2(t)^\top, \ldots, z_N(t)^\top]^\top$ such that dynamics of detectors can be expressed as follows:

$$\begin{aligned} \dot{z}(t) =& (I_N \otimes A_{\mathrm{z}}) z(t) + (I_N \otimes B_{\mathrm{z}}) \hat{x}(t) - L \otimes F_{\mathrm{z}}(e_{\mathrm{z}}(t) \\ &+ z(t)) + Q_{\mathrm{z}} \otimes a_{\mathrm{z}}(t), \end{aligned} \tag{181}$$

where $\hat{x}(t) = [\hat{x}_1(t)^\top, \hat{x}_2(t)^\top, \ldots, \hat{x}_N(t)^\top]^\top$, $e_{\mathrm{z}}(t) = [e_1^{\mathrm{z}}(t)^\top, e_2^{\mathrm{z}}(t)^\top, \ldots, e_N^{\mathrm{z}}(t)^\top]^\top$, $Q_{\mathrm{z}} = \mathrm{diag}(q_1^{\mathrm{z}}, q_2^{\mathrm{z}}, \ldots, q_N^{\mathrm{z}})$, $a_{\mathrm{z}}(t) = [a_{\mathrm{z}}^1(t)^\top, a_{\mathrm{z}}^2(t)^\top, \ldots, a_{\mathrm{z}}^N(t)^\top]^\top$, and $a_{\mathrm{z}}^i(t) = \sum_{j \in \mathcal{N}_i} a_{\mathrm{z}}^{ji}(t^j_{k_j})$ for $i = 1, \ldots, N$.

Theorem 6.8. *Let the Assumption 6.1 hold. Consider the MAS (148) and the detector (178) under cyber-attack free conditions, with the triggering function parameters (180) satisfy $0 < c_z$ and $0 < \alpha < -\max Re(\lambda(\tilde{A}_z))$, where $\lambda(\tilde{A}_z)$ denotes the eigenvalue of $\tilde{A}_z = I_N \otimes A_z - \Delta_z \otimes F_z$, and Δ_z can be computed in a similar manner as described for Δ in the proof of Theorem 6.6, where the Laplacian matrix is now L. The detectors in (181) reach a consensus if and only if $A_z - \lambda_i F_z$, $i = 2, \ldots, N$, are Hurwitz, where $\lambda_i \neq 0$ is the i-th eigenvalue of L. Moreover, the detector (181) does not exhibit Zeno behavior under attack free conditions.*

Proof. Under cyber-attack free conditions the expression (181) can be rewritten as

$$\begin{aligned}\dot{z}(t) =&(I_N \otimes A_{\mathrm{z}} - L \otimes F_{\mathrm{z}})z(t) + (I_N \otimes B_{\mathrm{z}})\hat{x}(t)\\ &- (L \otimes F_{\mathrm{z}})e_{\mathrm{z}}(t),\end{aligned}$$

Following along the derivations in [93], and since by definition the triggering function (180) does not cross zero in the interval $t \in [t^i_{k_i}, t^i_{k_i+1})$, we have $\|e^{\mathrm{z}}_i(t)\| < c_{\mathrm{z}}e^{-\alpha t}$, which implies that $\|e_{\mathrm{z}}(t)\| < \sqrt{N}c_{\mathrm{z}}e^{-\alpha t}$. Therefore, it can be concluded that $\|e_{\mathrm{z}}(t)\| \to 0$ as $t \to \infty$.

In a similar manner as in the proof of Theorem 6.6, let us define the disagreement vector $\delta_{\mathrm{z}}(t) = z(t) - (\mathbf{1}_N r^\top \otimes I_n)z(t)$. Since under cyber-attack free conditions states of the observers $\hat{x}(t)$ reach a consensus and $\mathbf{1}_N$ is the right eigenvector corresponding to the zero for $I_N - \mathbf{1}_N r^\top$, it follows that $\dot{\delta}_{\mathrm{z}}(t)$ is independent of $\hat{x}(t)$ once the agents reach a consensus.

Let us define $\varepsilon_{\mathrm{z}}(t) = (T_{\mathrm{z}}^{-1} \otimes I_n)\delta_{\mathrm{z}}(t) = [\varepsilon^{\mathrm{z}}_1(t)^\top, \varepsilon^{\mathrm{z}}_{2:N}(t)^\top]^\top$. Similar to the proof of the Theorem 6.6 it can be shown that $\varepsilon^{\mathrm{z}}_1(t) = \mathbf{0}_n$ and

$$\dot{\varepsilon}^{\mathrm{z}}_{2:N}(t) = (I_N \otimes A_{\mathrm{z}} - \Delta_{\mathrm{z}} \otimes F_{\mathrm{z}})\varepsilon^{\mathrm{z}}_{2:N}(t) - (\Delta_{\mathrm{z}} W_{\mathrm{z}} \otimes F_{\mathrm{z}})e^{\mathrm{z}}_{2:N}(t), \tag{182}$$

where $e^{\mathrm{z}}_{2:N}(t) = [e^{\mathrm{z}}_2(t)^\top, \ldots, e^{\mathrm{z}}_N(t)^\top]^\top$, $T_{\mathrm{z}} \in \mathbb{R}^{N\times N}$, $Y_{\mathrm{z}} \in \mathbb{R}^{N\times N-1}$, $W_{\mathrm{z}} \in \mathbb{R}^{N-1\times N}$, and the block diagonal matrix $\Delta_{\mathrm{z}} \in \mathbb{R}^{N-1\times N-1}$ has diagonal entries that are equal to the nonzero eigenvalues of L such that $T_{\mathrm{z}} = [\mathbf{1}_N\, Y_{\mathrm{z}}]$, $T_{\mathrm{z}}^{-1} = \begin{bmatrix} r^\top \\ W_{\mathrm{z}} \end{bmatrix}$ and $T_{\mathrm{z}}^{-1}LT_{\mathrm{z}} = J_{\mathrm{z}} = \begin{bmatrix} 0 & \mathbf{0}_{1\times N-1} \\ \mathbf{0}_{N-1} & \Delta_{\mathrm{z}} \end{bmatrix}$.

Following along the proof of Theorem 6.6, it can be shown that (182) is stable if and only if $A_{\mathrm{z}} - \lambda_i F_{\mathrm{z}}$, for $i = 2, \ldots, N$, are Hurwitz.

To show that (181) does not exhibit the Zeno behavior one needs to show that the inter-event intervals are lower bounded. The solution to (182) can be derived as follows:

$$\varepsilon^{\mathrm{z}}_{2:N}(t) = e^{\tilde{A}_{\mathrm{z}}t}\varepsilon^{\mathrm{z}}_{2:N}(0) + \int_0^t e^{\tilde{A}_{\mathrm{z}}(t-\tau)}(\Delta_{\mathrm{z}} W_{\mathrm{z}} \otimes F_{\mathrm{z}})e^{\mathrm{z}}_{2:N}(\tau)\mathrm{d}\tau,$$

where $\tilde{A}_{\mathrm{z}} = I_N \otimes A_{\mathrm{z}} - \Delta_{\mathrm{z}} \otimes F_{\mathrm{z}}$. From the Lemma 6.4, it can be concluded that

$$\|\varepsilon_{\mathrm{z}}(t)\| = \|\varepsilon^{\mathrm{z}}_{2:N}(t)\| \leq \alpha_3 e^{\lambda^{\mathrm{m}}_{\tilde{A}_{\mathrm{z}}} t} + \alpha_2 e^{-\alpha t}, \tag{183}$$

where $\alpha_3 = \alpha_1 + \alpha_2$, $\alpha_1 = c_{\tilde{A}_{\mathrm{z}}} nN \|P_{\tilde{A}_{\mathrm{z}}}\| \|P^{-1}_{\tilde{A}_{\mathrm{z}}}\| \|\varepsilon_{\mathrm{z}}(0)\|$, $\alpha_2 = (c_{\tilde{A}_{\mathrm{z}}} nN \|P_{\tilde{A}_{\mathrm{z}}}\| \|P^{-1}_{\tilde{A}_{\mathrm{z}}}\| \|\Delta_{\mathrm{z}} W_{\mathrm{z}} \otimes F_{\mathrm{z}}\| \sqrt{N} c_{\mathrm{z}})/(|\alpha + \lambda^{\mathrm{m}}_{\tilde{A}_{\mathrm{z}}}|)$, $c_{\tilde{A}_{\mathrm{z}}} > 0$ is a positive constant, and $\max \mathrm{Re}(\lambda(\tilde{A}_{\mathrm{z}})) < \lambda^{\mathrm{m}}_{\tilde{A}_{\mathrm{z}}} < -\alpha < 0$. It follows from (183) that $\|\delta_{\mathrm{z}}(t)\| \leq \|T_{\mathrm{z}} \otimes I_n\| \|\varepsilon_{\mathrm{z}}(t)\| \leq \beta_1 e^{\lambda^{\mathrm{m}}_{\tilde{A}_{\mathrm{z}}} t} + \beta_2 e^{-\alpha t}$, where $\beta_1 = \|T_{\mathrm{z}} \otimes I_n\| \alpha_3$ and $\beta_2 = \|T_{\mathrm{z}} \otimes I_n\| \alpha_2$.

In the interval $t \in [t^i_{k_i},\, t^i_{k_i+1})$ the dynamics of $e_{\mathrm{z}}(t)$ can be expressed as

$$\dot{e}_{\mathrm{z}}(t) = (I_N \otimes A_{\mathrm{z}} - L \otimes F_{\mathrm{z}}) e_{\mathrm{z}}(t) - (I_N \otimes B_{\mathrm{z}}) \hat{x}(t) - (L \otimes F_{\mathrm{z}}) z(t).$$

Let us define $\delta_{\mathrm{e}}(t) = e_{\mathrm{z}}(t) - (\mathbf{1}_N r^\top \otimes I_n) e_{\mathrm{z}}(t)$, which is governed by

$$\begin{aligned} \dot{\delta}_{\mathrm{e}}(t) =& (I_N \otimes A_{\mathrm{z}} - L \otimes F_{\mathrm{z}}) \delta_{\mathrm{e}}(t) - ((I_N - \mathbf{1}_N r^\top) \otimes B_{\mathrm{z}}) \hat{x}(t) \\ & - (L \otimes F_{\mathrm{z}}) \delta_{\mathrm{z}}(t). \end{aligned}$$

Since it is assumed that the MAS is cyber-attack free, $\delta_{\mathrm{z}}(t)$ does not approach to zero unless $((I_N - \mathbf{1}_N r^\top) \otimes B_{\mathrm{z}}) \hat{x}(t)$ approaches to zero. Thus, there exists a bounded scalar $M > 0$ such that $\|((I_N - \mathbf{1}_N r^\top) \otimes B_{\mathrm{z}}) \hat{x}(t)\| \leq M \|\delta_{\mathrm{z}}(t)\|$. Moreover, given that $\|I_N - \mathbf{1}_N r^\top\| \geq \|I_N\|$, by definition one can conclude that $\|e_{\mathrm{z}}(t)\| \leq \|\delta_{\mathrm{e}}(t)\|$. Therefore, it implies that

$$\begin{aligned} \|\dot{e}_{\mathrm{z}}(t)\| \leq & \|\dot{\delta}_{\mathrm{e}}(t)\| \leq \|I_N \otimes A_{\mathrm{z}} - L \otimes F_{\mathrm{z}}\| \|I_N - \mathbf{1}_N r^\top\| \\ & \times \|e_{\mathrm{z}}(t)\| + \|L \otimes F_{\mathrm{z}}\| \|\delta_{\mathrm{z}}(t)\| \\ & + \|((I_N - \mathbf{1}_N r^\top) \otimes B_{\mathrm{z}}) \hat{x}(t)\| \leq g_{\mathrm{z}}(t), \end{aligned}$$

where $g_{\mathrm{z}}(t) \triangleq a_1 e^{\lambda^{\mathrm{m}}_{\tilde{A}_{\mathrm{z}}} t} + a_2 e^{-\alpha t}$, $a_1 = (\|L \otimes F_{\mathrm{z}}\| + M)\beta_1$, and $a_2 = (\|L \otimes F_{\mathrm{z}}\| + M)\beta_2 + \sqrt{N} \|I_N \otimes A_{\mathrm{z}} - L \otimes F_{\mathrm{z}}\| \|I_N - \mathbf{1}_N r^\top\|$.

Let t^* denote the latest triggering instant and consider $\tau^* = t - t^*$ as the time-interval between the two latest triggered events. Given that at the triggering instant $f^{\mathrm{z}}_i(t, e^{\mathrm{z}}_i(t)) = 0$, it can be concluded that in the i-th detector the next event cannot be triggered before $\|e^{\mathrm{z}}_i(t)\| = c_{\mathrm{z}} e^{-\alpha t}$, that implies $\|e_{\mathrm{z}}(t)\| = \| \int_{t^*}^{t} \dot{e}_{\mathrm{z}}(s) \mathrm{d}s\| \leq$

$\int_{t^*}^{t} g_z(s)(s)\mathrm{d}s = \sqrt{N}c_z e^{-\alpha t}$. Since $t^* \geq t$, we have $e^{-\alpha t} \leq e^{-\alpha t^*}$ and $e^{\lambda^{\mathrm{m}}_{\tilde{A}_z} t} \leq e^{\lambda^{\mathrm{m}}_{\tilde{A}_z} t^*}$. Consequently, τ^* is lower bounded by $\bar{\tau}$, which is the solution to the equation $(a_1 e^{\lambda^{\mathrm{m}}_{\tilde{A}_z} t^*} + a_2 e^{-\alpha t^*})\bar{\tau} = \sqrt{N}c_z e^{-\alpha(t^*+\bar{\tau})}$, that is equivalent to $(a_1 e^{(\lambda^{\mathrm{m}}_{\tilde{A}_z}+\alpha)t^*} + a_2)\bar{\tau} = \sqrt{N}c_z e^{-\alpha\bar{\tau}}$.

Since $0 < \alpha < |\lambda^{\mathrm{m}}_{\tilde{A}_z}| < -\max \mathrm{Re}(\lambda(\tilde{A}_z))$, there exists $\tilde{\tau}$ such that $\tau^* \geq \bar{\tau} \geq \tilde{\tau}$, where $\tilde{\tau}$ is the strictly positive solution to the equation $(a_1 + a_2)\tilde{\tau} = \sqrt{N}c_z e^{-\alpha\tilde{\tau}}$. Hence, $\tilde{\tau}$ is the lower bound on the inter-event times of the detector (178), which implies that there are no Zeno behavior. This completes the proof of the theorem. □

Definition 6.17. *A cyber-attack injected to the closed-loop MAS* (132) *and the observer* (133) *is detected if the residual signal*

$$res_z^i(t) = \sum_{j \in \mathcal{N}_i} \|z_j(t^j_{k_j}) - z_i(t^i_{k_i}) + q_{ji} a_z^{ji}(t)\|, \tag{184}$$

satisfies the inequality $\|res_z^i(t)\| > \eta_z$*, where* η_z *is the cyber-attack detection threshold.*

Assumption 6.7. *The adversaries do not have knowledge on the parameters of the event-triggered detector* (178)*. Hence, they are not capable of designing* $a_z^{ji}(t^j_{k_j})$ *such that* $z_j(t^j_{k_j}) - z_i(t^i_{k_i}) + q_{ji} a_z^{ji}(t^j_{k_j}) = 0$ *as* $t^j_{k_j} \to \infty$.

Remark 6.14. *Since agents communicate their state detectors according to the triggering function* (180)*, the adversary does not have access to these states continuously. Consequently, it is quite reasonable to assume that the adversary does not have knowledge of the exact values of the parameters in* (178)*.*

Corollary 6.1. *Consider the Assumptions 6.1-6.7 hold and the MAS* (148) *is under quasi-covert cyber-attacks as introduced in the Lemma 6.6. Given that the triggering function* (180) *has parameters that are provided in the Theorem 6.8, let the cyber-attack signal be denoted by* $a_z^{ji}(t^j_{k_j}) = a_{z0} - e^{A_z(t-t^j_{k_j})} z_j(t^j_{k_j})$ *for* $i \in V_{da}$ *and* $j \in V_{nc}$*, where* $a_{z0} \in \mathbb{R}^n$ *is a constant vector. Consequently, the generated residual* (184) *by the* k*-th agent is nonzero if* $k \in V_a$ *or* $k \in V_{nca}$*, and* B_z *is full column rank, and* $A_z - \lambda_q F_z$*, for* $q = 2, \ldots, N$*, are Hurwitz in detectors* (181)*, where* λ_q *is defined as in Theorem 6.8.*

Proof. Suppose B_z is a full column rank matrix and $A_z - \lambda_q F_z$, for $q = 2, \ldots, N$, are Hurwitz. Consequently, following along the steps as in proof of Theorem 6.7, the detectors that belong to the set V_a do

not reach a_{z0} as the consensus set point since dynamics of the virtual adversary agent does not have $B_z\hat{x}_i$. Hence, cyber-attacks are detectable by using the detectors of agents that belong to this set.

Since under cyber-attacks the agents observers states do not reach a consensus, $\dot{\delta}_z(t)$ in the proof of Theorem 6.8 depends on $\hat{x}(t)$, which implies that the detectors do not reach a consensus as well and the cyber-attacks are detectable on the set V_{nca}. This completes the proof of the corollary. □

6.7 Numerical Case Studies

6.7.1 Privacy Preserving Consensus Control for Formation Flying of Satellites

In this numerical case study, the effectiveness of the proposed privacy preserving dynamic controller $\mathcal{C}_i$ in reaching consensus among the MAS Σ_i is illustrated. Dynamics of agents and controllers are given by (127) and (137), respectively. In this case study, our objective is to achieve formation flying among a group of satellites such that they reach the same velocity along the x-axis, the y-axis, and the z-axis in three dimensional space. Moreover, satellites are required to maintain a prescribed distance from one another.

A group of 6 satellites is considered with the characteristic matrices given by [66]:

$$A = \begin{bmatrix} 0 & I_3 \\ A_1 & A_2 \end{bmatrix}, \; B = \begin{bmatrix} 0 \\ I_3 \end{bmatrix}, \; C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix},$$

$$A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 3\omega_0^2 & 0 \\ 0 & 0 & -\omega_0^2 \end{bmatrix}, \; A_2 = \begin{bmatrix} 0 & 2\omega_0 & 0 \\ 2\omega_0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where $x_i(t) = [r_i(t)^\top \; \dot{r}_i(t)^\top]^\top$, $r_i(t) = [x_i^{\text{x}}(t), x_i^{\text{y}}(t), x_i^{\text{z}}(t)]^\top \in \mathbb{R}^3$ denotes the position of the i-th satellite, $\dot{r}_i(t) = [\dot{x}_i^{\text{x}}(t) \; \dot{x}_i^{\text{y}}(t) \; \dot{x}_i^{\text{z}}(t)]^\top \in \mathbb{R}^3$ is the velocity vector of the satellite i, and $\omega_0 = 0.001$ is the circular orbit rate.

Moreover, our objective is to have $\|(r_i(t) - h_i) - (r_j(t) - h_j)\| = 0$ and $\|\dot{r}_i(t) - \dot{r}_j(t)\| = 0$ as $t \to \infty$, where $h_i - h_j \in \mathbb{R}^3$ denotes the prescribed constant separation between the i-th and the j-th satellites, for

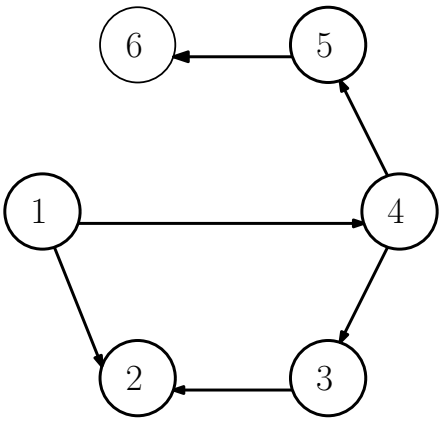

Figure 6.2: Communication graph of the satellites.

$j \in \mathcal{N}_i$. Consequently, we modify $\Pi_i(t)$ and $u_i(t)$ in (137) as given below

$$\begin{cases} \Pi_i(t) = & \sum\limits_{j \in \mathcal{N}_i} (\bar{H}(\tilde{z}_i^{\mathrm{x}}(t) - \tilde{z}_j^{\mathrm{x}}(t)) + (\mathcal{J}_j \tilde{z}_j^{\mathrm{y}}(t) - \mathcal{J}_i \tilde{z}_i^{\mathrm{y}}(t))), \\ u_i(t) = & K \sum\limits_{j \in \mathcal{N}_i} \bar{P}(\tilde{z}_i^{\mathrm{x}}(t) - \tilde{z}_j^{\mathrm{x}}(t)), \end{cases}$$

for $i = 1, \ldots, 6$, where $\tilde{z}_i^{\mathrm{x}}(t) = z_i^{\mathrm{x}}(t) - \tilde{h}_i^{\mathrm{x}}$, $\tilde{z}_i^{\mathrm{y}}(t) = z_i^{\mathrm{y}}(t) - \tilde{h}_i^{\mathrm{y}}$, $\tilde{h}_i^{\mathrm{x}} = P_i[h_i^\top 0_{1\times 3}^\top]^\top$, and $\tilde{h}_i^{\mathrm{y}} = J_i h_i$. Also, we have $h_1 = [20\,20\,0]^\top$, $h_2 = [-20\,-20\,0]^\top$, $h_3 = [-60\,-60\,0]^\top$, $h_4 = [-40\,-40\,0]^\top$, $h_5 = [40\,40\,0]^\top$, $h_6 = [80\,80\,0]^\top$.

The communication graph among the satellites is shown in Figure 6.2. Moreover, the Laplacian matrix associated with the communication graph is given by:

$$L = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}.$$

The nonzero eigenvalues of the matrix L are 1 and 2.

In this case study, due to its simplicity, we have used the Givens rotation provided in Definition 6.4 to

design the isometric isomorphisms. One can also use the Householder transformation to achieve control and privacy objectives stated in Subsection 6.2.1. We choose, for $q = 0, \ldots, 6$, $P_q = G_1(\alpha_q^{\mathrm{p}}, \beta_q^{\mathrm{p}}, \theta_q^{\mathrm{p}}) \in \mathbb{R}^{6\times 6}$ such that $\alpha_q^{\mathrm{p}} = 3$, $\beta_q^{\mathrm{p}} = 4$, and $\theta_q^{\mathrm{p}} = (q+1)\pi/13$. Moreover, let $S_q = G_2(\alpha_q^{\mathrm{s}}, \beta_q^{\mathrm{s}}, \theta_q^{\mathrm{s}}) \in \mathbb{R}^{3\times 3}$, where we select $\alpha_q^{\mathrm{s}} = 2$, $\beta_q^{\mathrm{s}} = 3$, and $\theta_q^{\mathrm{s}} = (q+1)\pi/17$.

The parameters of dynamic controller $\mathcal{C}_i$ are selected as

$$K = \begin{bmatrix} -22 & -4 & -6 & -4 & 4 & -10 \\ -18 & -18 & -12 & -2 & -6 & 6 \\ -22 & -8 & -22 & 2 & -4 & -2 \end{bmatrix}, \; H = \begin{bmatrix} -42 & -24 & -42 \\ -2 & -22 & -22 \\ -2 & -2 & -20 \\ -22 & -22.2 & -24 \\ -22 & -20 & -36 \\ -22 & 0 & -36 \end{bmatrix}.$$

In Figure 6.3, the output measurements of all satellites are shown. As seen from Figure 6.3, satellites reach a formation flying, while they utilize the controller $\mathcal{C}_i$ and share $\tilde{z}_i^{\mathrm{x}}(t)$ and $\tilde{z}_i^{\mathrm{y}}(t)$. Moreover, according to Theorem 6.2 their dynamics are indistinguishable by eavesdroppers and honest-but-curious agents. Relative positions of satellites in three dimensional space is shown in Figure 6.4. The transformed output measurements, $\tilde{y}_i(t)$ are shown in Figure 6.5 and Figure 6.6 which are different from true output measurements. Adversaries are capable of reading the transmitted signal $\tilde{z}_i^{\mathrm{y}}(t)$ shown in Figure 6.7. However, from Figure 6.7 it can be seen that by eavesdropping the signal $\tilde{z}_i^{\mathrm{y}}(t)$, adversaries cannot discover true positions of satellites.

6.7.2 Controllability Cyber-Attacks in MAS

In this numerical example, the controllability conditions that are provided in Theorems 6.3 and 6.4 are studied for a MAS system consisting of 6 agents. The agent dynamics and its observer are given by (127),

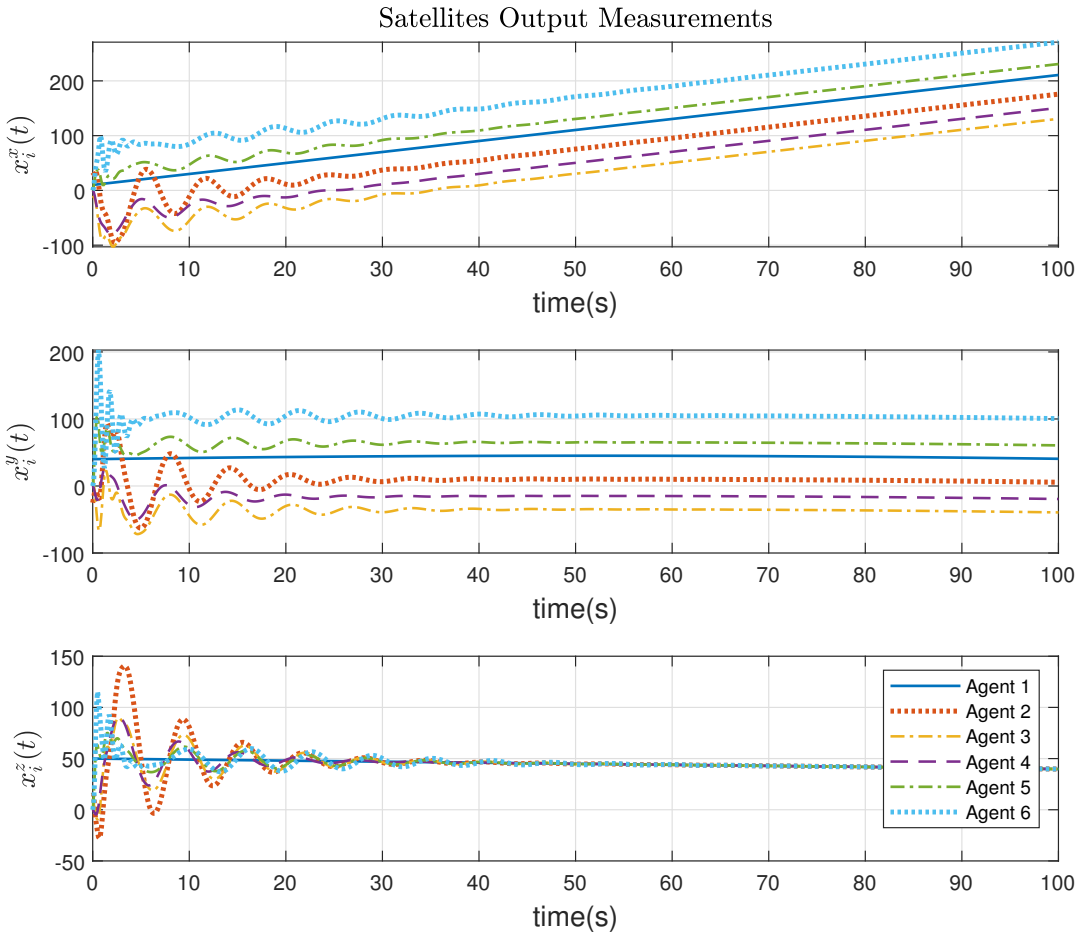

Figure 6.3: Output measurement of each satellite.

and (128), respectively, with the following matrices [66]:

$$A = \begin{bmatrix} -2 & 2 \\ -1 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$
$$H = \begin{bmatrix} 0 & 0.3 \\ -0.3 & 0 \end{bmatrix}, K = \begin{bmatrix} -1 & 2 \end{bmatrix}.$$

The communication graph among the agents is shown in Figure 6.8, and its corresponding Laplacian

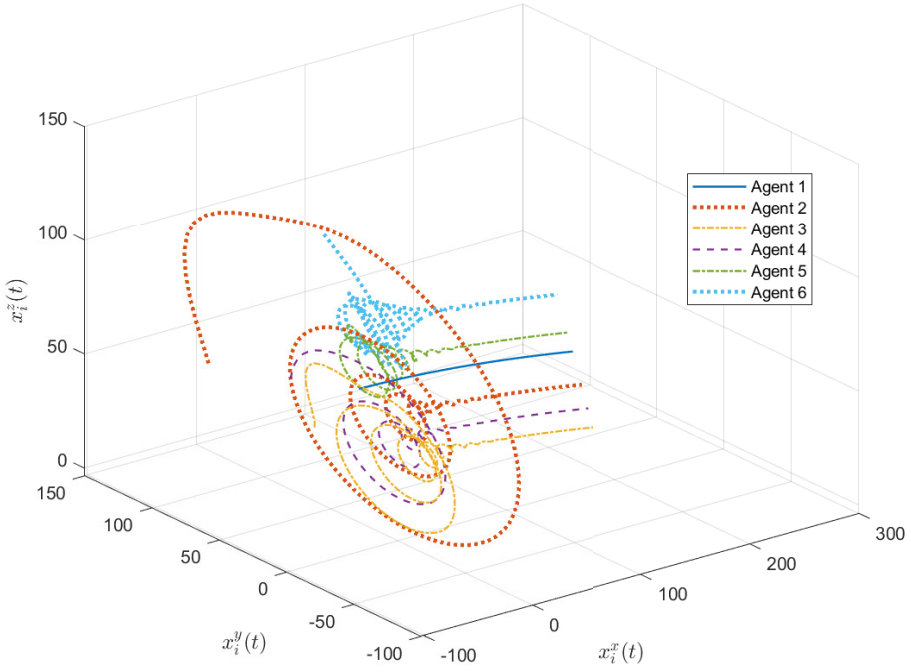

Figure 6.4: Relative positions at $t = [2, 100)$.

matrix is $L = [1, 0, 0, 0, 0, -1; 0, 2, 0, 0, -1, -1; 0, -1, 1, 0, 0, 0; 0, -1, -1, 2, 0, 0; 0, -1, 0, 0, 1, 0; 0, 0, 0, 0, -1, 1]$.

Let us assume that the incoming communication links of agents 4, 5, and 6 are under attack so that one obtains $L_{\text{f}} = [1, 0, 0; 0, 2, 0; 0, -1, 1]$ and $l_{\text{fa}} = [0, 0, -1; 0, -1, -1; 0, 0, 0]$, where the eigenvalues of L_{f} are $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 2$, corresponding to the right eigenvectors $[1, 0, 0]^T$, $[0, 0, 1]^T$, and $[0, 0.7071, -0.7071]^T$, respectively. Since the geometric multiplicity of each eigenvalue is equal to its algebraic multiplicity, conditions in Assumption 6.4 hold. In this example, the conditions in Theorem 6.3 are not satisfied and $\text{rank}(\mathcal{C}^*) = 5$. Hence, the adversary does not have control over the entire MAS system as provided in Definition 6.6, however, an adversary may still impact the followers as described in Definition 6.7 and Theorem 6.4.

Considering Theorem 6.4, the rank of the controllability matrices $(\check{A}+\check{B}\check{K}+d_i\check{H}\check{C}, \check{H}_a)$ for $i = 4, 5, 6$ are equal to 4, the rank of the controllability of the pair $(L_{\text{f}}, l_{\text{fa}})$ is 3, and the rank of the controllability of $(\check{A} + \check{B}\check{K} + \lambda_j\check{H}\check{C}, \check{H}\check{C})$ for $j = 1, 2, 3$ is equal to 4. Therefore, the adversary has the capability of

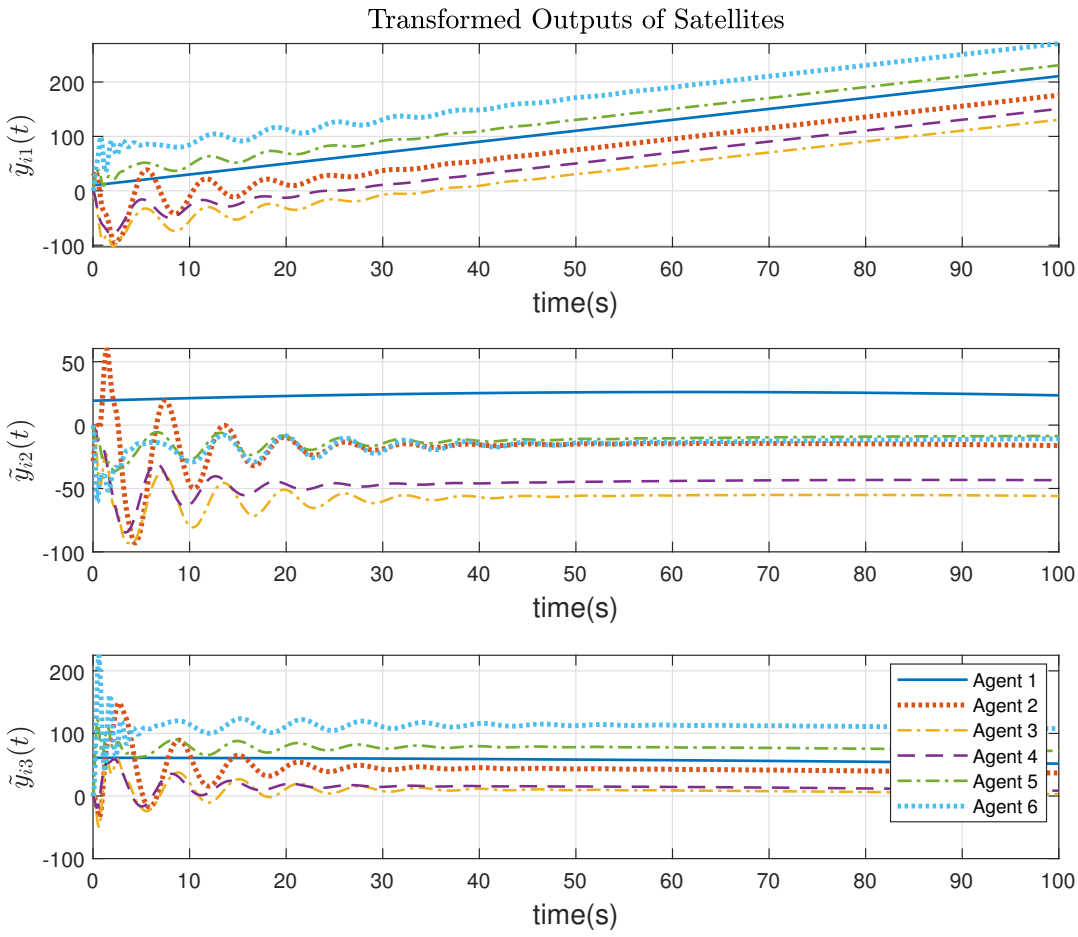

Figure 6.5: The transformed outputs, $\tilde{y}_i(t)$.

manipulating and controlling the three agents 1, 2, 3 by simultaneously attacking the agents 4, 5, and 6.

As shown in Figure 6.9, the six agents reach a consensus and their states converge, while at $t = 30$ (s) the adversary injects its attack signals to the agents 4, 5, and 6 and the remaining agents are controlled through the directly attacked agents. In Figure 6.9, to illustrate the capability of the adversary in controlling all the agents, the states of each agent are set to different values by choosing different attack signals for the directly attacked agents.

In Figure 6.10, it can be seen that the attack that has occurred at $t = 30$ (s) is designed such that the agents reach a new consensus that is desirable to the adversary. In this attack scenario, the directly attacked

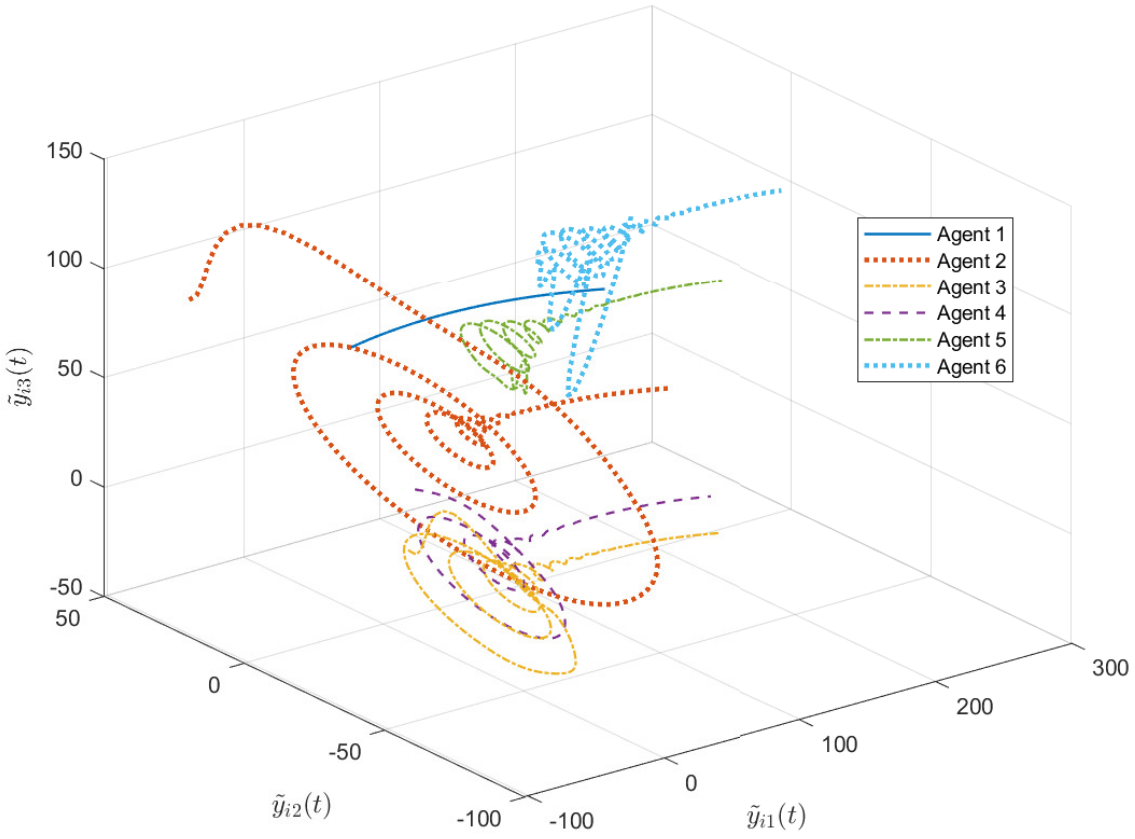

Figure 6.6: Transformed positions in 3D space at $t = [2, 100)$.

agents have the same attack signals so that they reach to the same point and the remaining agents follow them. This example illustrates that even without having full controllability over the MAS systems as stated in Definition 6.6, the adversary is capable of imposing a major impact on the trajectory and behavior of the agents.

The security controllability index for the directly attacked agents are $SCI_4 = 0$, $SCI_5 = 1$, $SCI_6 = 2$, and for the MAS system is $SCI = 3$. It follows that $SCI_4 = 0$, which implies that through agent 4 the adversary is not capable of controlling any of the followers. However, attacking the agents 5 and 6 do not provide controllability over the agent 4 to the adversary, and hence, in this case it will be attacked directly.

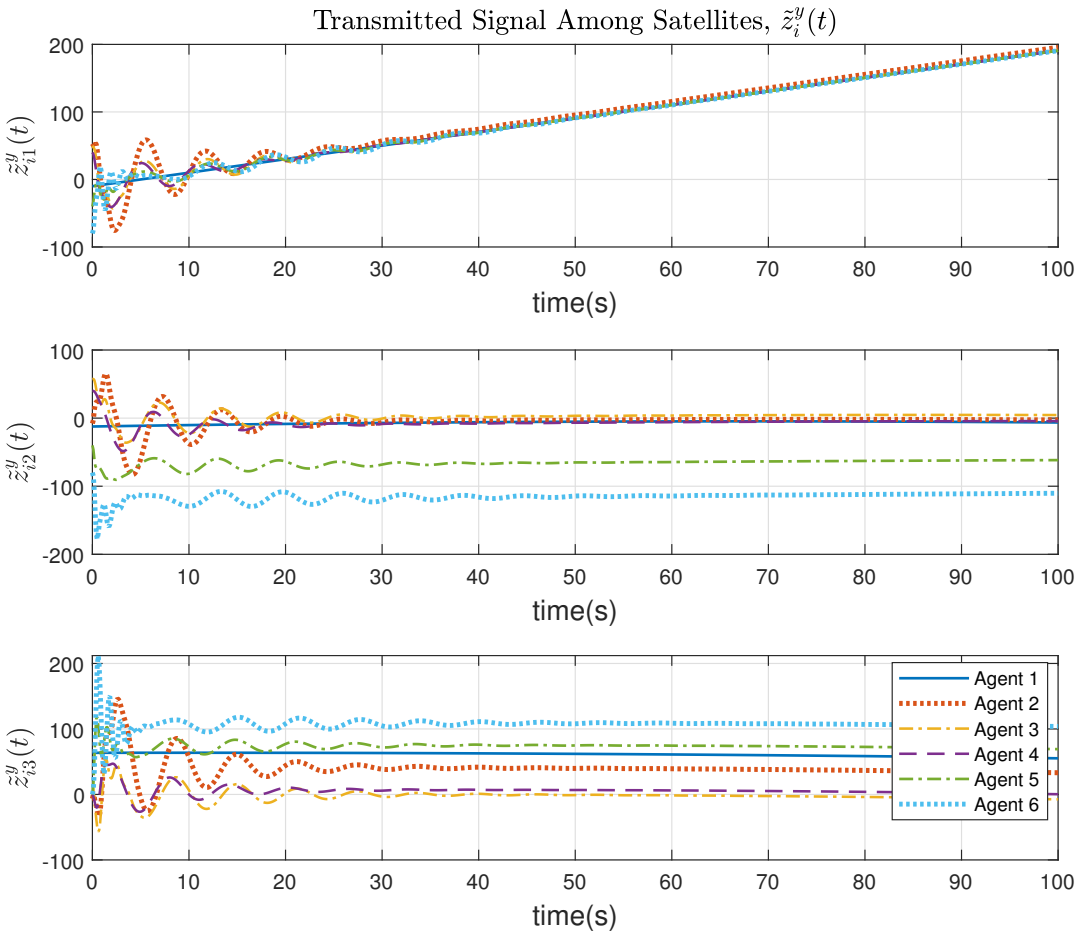

Figure 6.7: The transmitted signal $\tilde{z}_i^{\mathrm{y}}(t)$ among satellites.

6.8 Undetectable Cyber-Attacks in MAS and Event-Triggered Detector Module

In this case study, cyber-attacks that are introduced in the Theorem 6.7 and the quasi-covert cyber-attacks in the Lemma 6.6 are investigated. Moreover, the effectiveness of event triggered detector that was proposed in Section 6.6 in detecting the quasi-covert cyber-attacks is demonstrated and illustrated. A MAS consisting of 6 agents along with their observer-based consensus protocols having the dynamics as given in (127) and

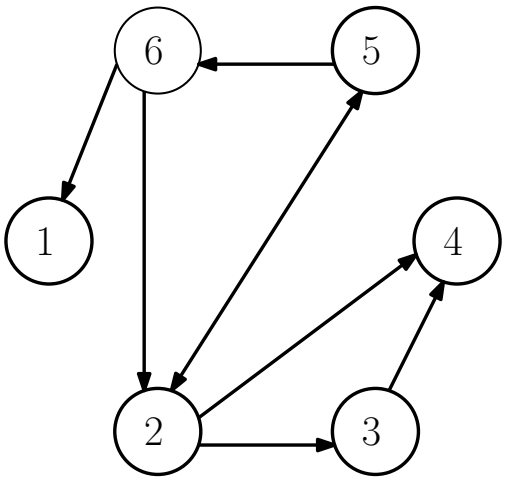

Figure 6.8: Communication graph of the MAS system.

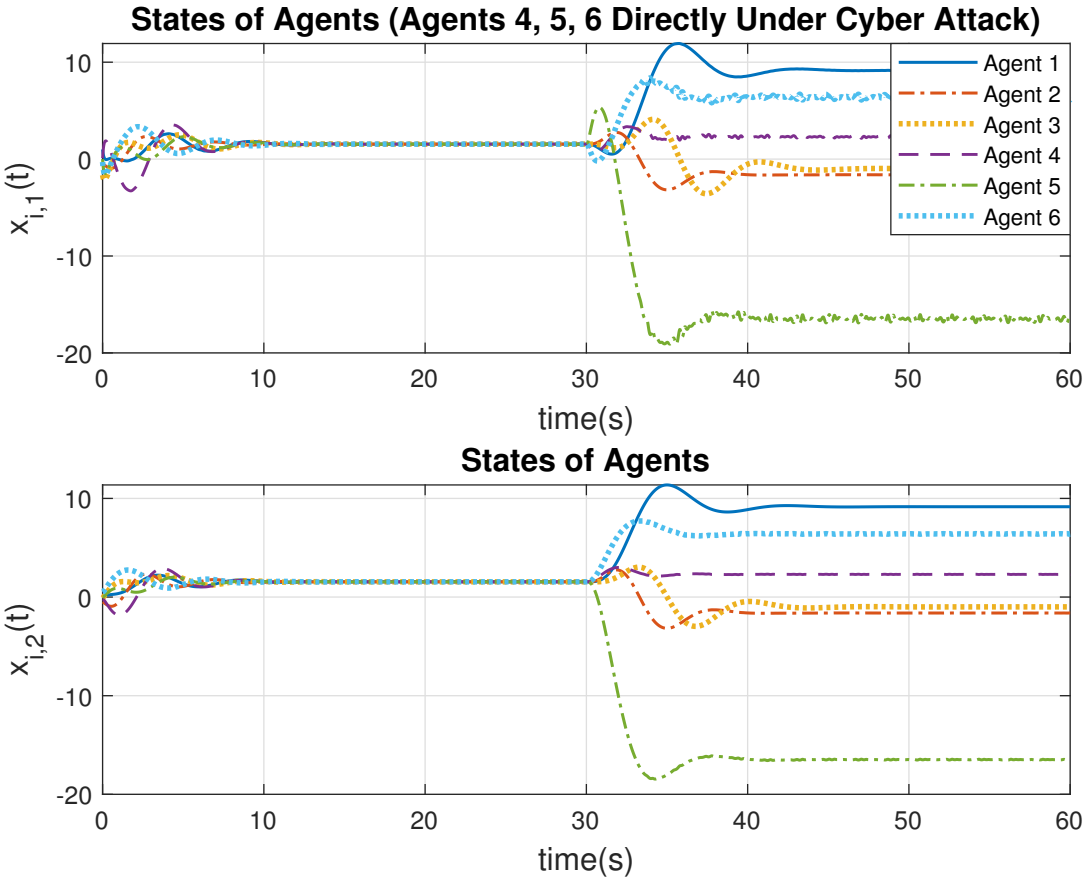

Figure 6.9: State trajectories of the six agents in presence of cyber-attack injected at $t = 30$ (s).

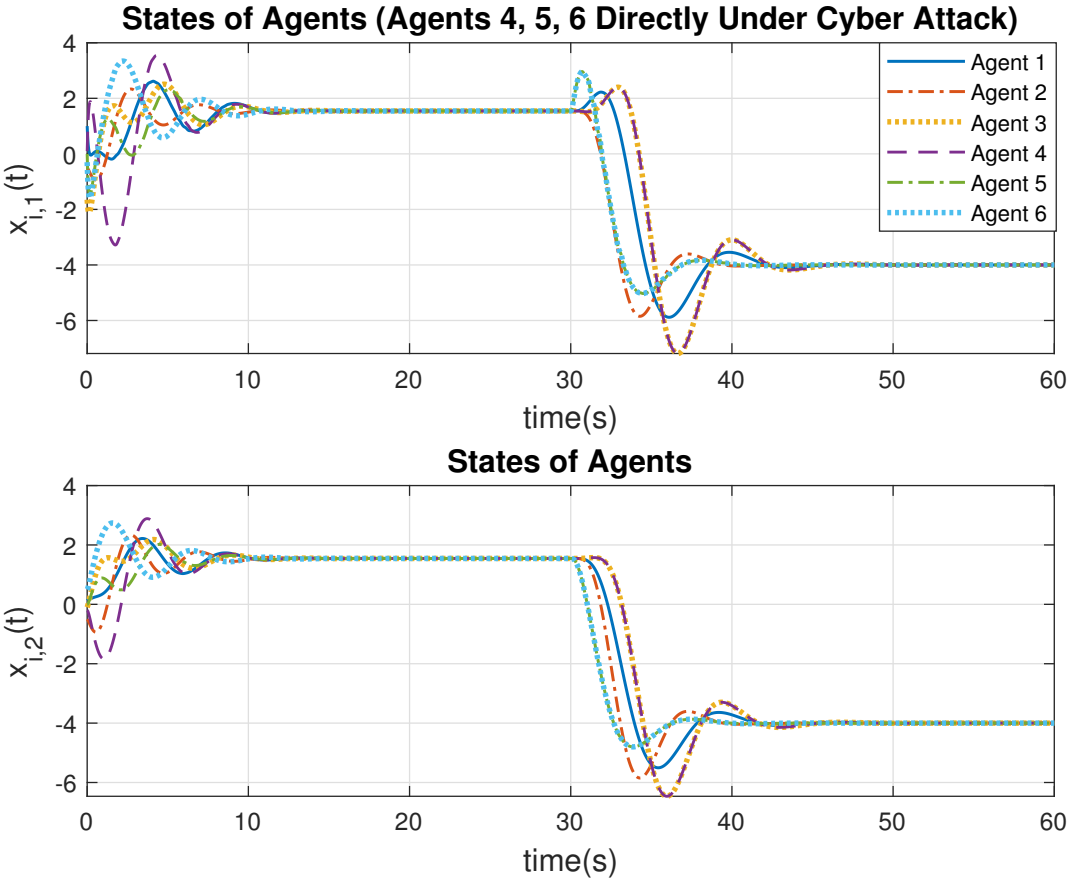

Figure 6.10: Change in the consensus set point by the adversary injecting cyber-attack signals at $t = 30$ (s).

(131), respectively, with the following parameters are studied [66] in:

$$A = \begin{bmatrix} -2 & 2 \\ -1 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$
$$F = \begin{bmatrix} 15 & 0 \\ 15 & 15 \end{bmatrix}, K = \begin{bmatrix} 2 & -10 \end{bmatrix}, c = -2.$$

The Laplacian matrix of the communication links among agents is given by L =[1, -1, 0, 0, 0, 0; -1, 1, 0, 0, 0, 0; -1, 0, 2, -1, 0, 0; 0, -1, 0, 2, 0, -1; 0, 0, -1, 0, 1, 0; 0, 0, 0, 0, -1, 1]. The detector parameters in (178) are $A_z = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}$, $B_z = I_2 \times 5$, $F_z = I_2 \times 10$, $c_z = 0.002$, and $\alpha = 0.2$.

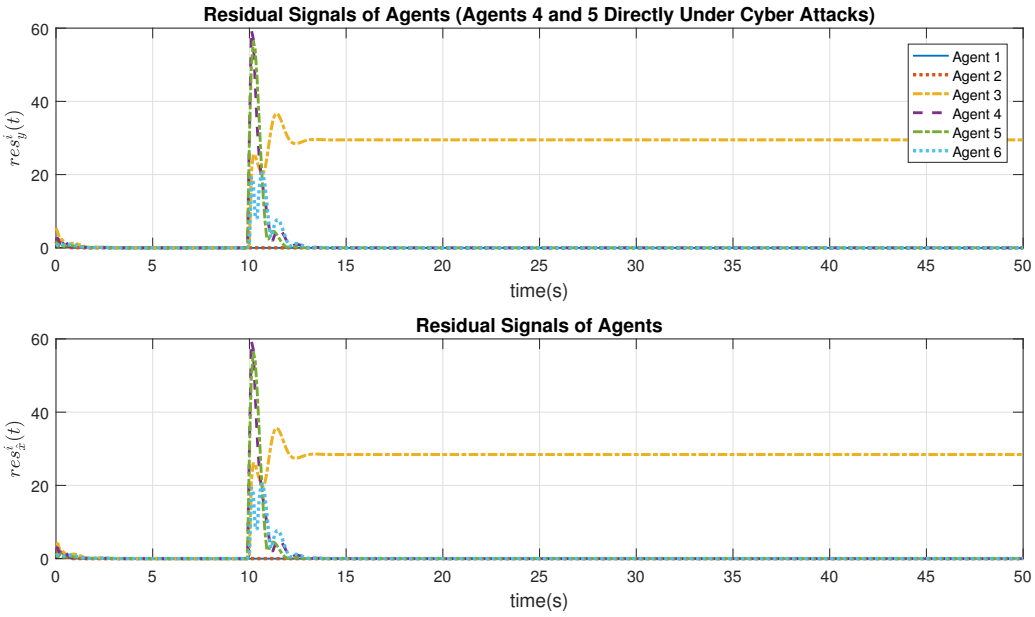

Figure 6.11: Residuals of agents while the cyber-attacks are injected at $t = 10$ (s).

Agents 1 and 2 are the roots of the spanning trees as contained in graph $\mathcal{G}$ of the network. In this case study, agents 4 and 5 are directly under cyber-attacks such that $V_{\text{da}} = \{4, 5\}$, $V_{\text{a}} = \{4, 5, 6\}$, $V_{\text{nc}} = \{1, 2, 3\}$, and $V_{\text{nca}} = \{3\}$.

In Figure 6.11, the residuals of agents in presence of cyber-attacks as introduced in the Theorem 6.7 are shown. After the occurrence of the cyber-attack at $t = 10$ (s) it is detected by the agent 3 that belongs to the set V_{nca}, whereas it is undetectable by the rest of network agents.

The impacts of the quasi-covert cyber-attack on MAS are illustrated in Figure 6.12. As can be observed in this figure the agent 3 states after the quasi-covert cyber-attack is injected at $t = 30$ (s) reach to those of the agents 1 and 2. Consequently, this cyber-attack is undetectable on the entire network as shown in Figure 6.13. By using the event-triggered detectors that are proposed in (178), the residuals (184) are generated and shown in Figure 6.14. It follows from this figure that all agents' residuals that belong to the set V_{a} are nonzero and exceed the threshold $\eta_{\text{z}} = 3$. Moreover, the agent 3 residual exceeds the threshold as well.

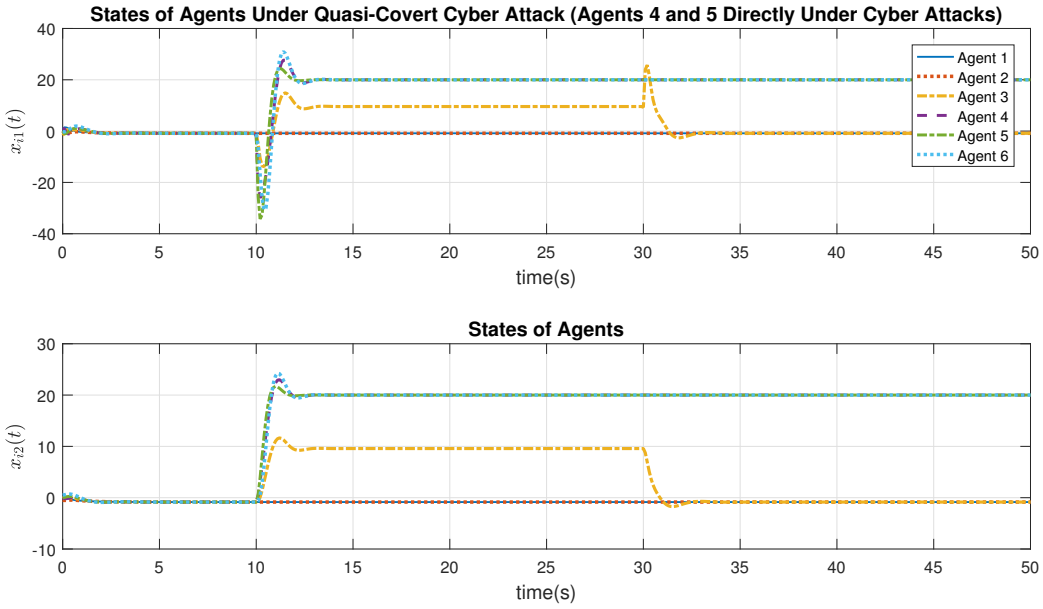

Figure 6.12: States of agents in presence of cyber-attacks at $t = 10$ (s) and quasi-covert cyber-attacks at $t = 30$ (s).

6.9 Conclusion

In this chapter, four main problems related to cyber-attacks in multi-agent systems (MAS) were studied. First, the problem of data privacy protection of MAS was investigated. By utilizing our proposed isometric isomorphisms, dynamics of agents are transformed into new bases. Each agent employs a unique isometric isomorphism. Moreover, a dynamic consensus protocol has been developed which uses the transformed sensor measurements and controller states of the neighboring agents to reach consensus in the MAS and preserve the privacy of agents. An algorithm is also provided to describe the communication protocol among the agents in order to ensure that the transformed dynamics of agents are indistinguishable by eavesdropper adversaries and honest-but-curious agents. As for the second problem, certain types of cyber-attacks on MAS systems were investigated and developed. In one cyber-attack scenario, the adversary targets the incoming communication links for a team of agents and disguises the attack signals as transmitted information among the agents. Therefore, there are two groups of agents, those that are first directly attacked, and those that can be considered as followers of the first group. The conditions under which the adversary has

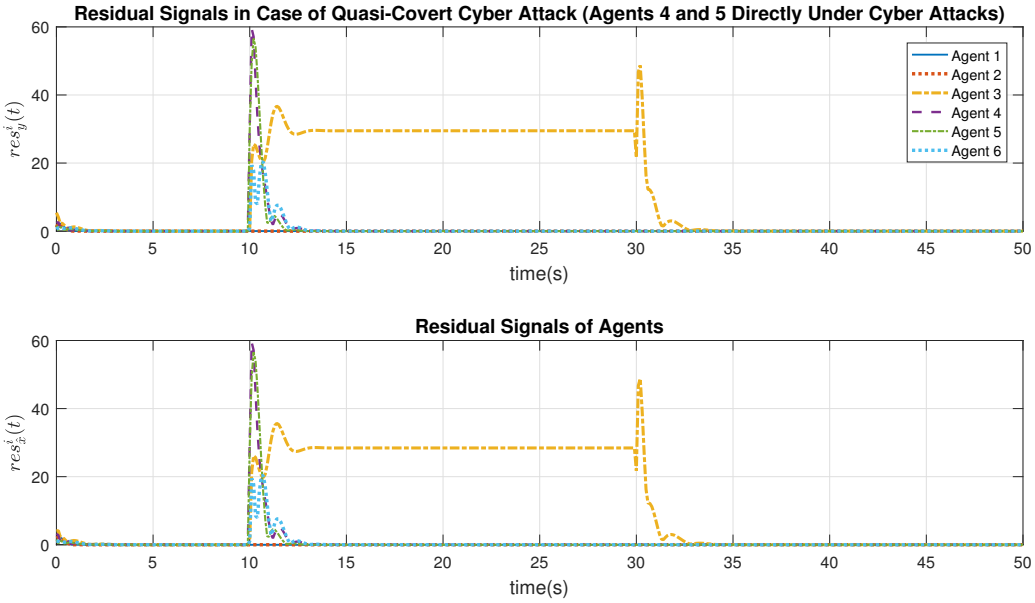

Figure 6.13: Residuals approach to zero after occurrence of the quasi-covert cyber-attacks at $t = 30$ (s).

full control over the two agent groups were investigated. The notions of security controllability for each of the directly attacked agents as well as the entire MAS system have been proposed and developed. These notions can be used to identify agents that allow the adversary high control authority over the MAS network. Moreover, it was shown that the adversary is not capable of simultaneously performing zero dynamics attacks on the directly attacked agents and the followers. In the third problem, detectability of cyber-attacks on the communication links of certain teams of MAS has been investigated. A definition that can be used to specify undetectable cyber-attacks on MAS was developed and proposed. It was shown that cyber-attacks on communication links of the root of a directed spanning tree graph are undetectable. Moreover, cyber-attacks on the non-root agents were investigated. It was shown that cyber-attacks on non-root agents can be detected by the set of agents provided that one can determine an uncompromised directed path from the root of the graph to these agents. Novel quasi-covert cyber-attacks were introduced that can be injected to maintain and ensure these attacks on non-root agents remain undetectable by the entire network. Finally, an event-triggered detector was proposed that is capable of detecting quasi-covert cyber-attacks.

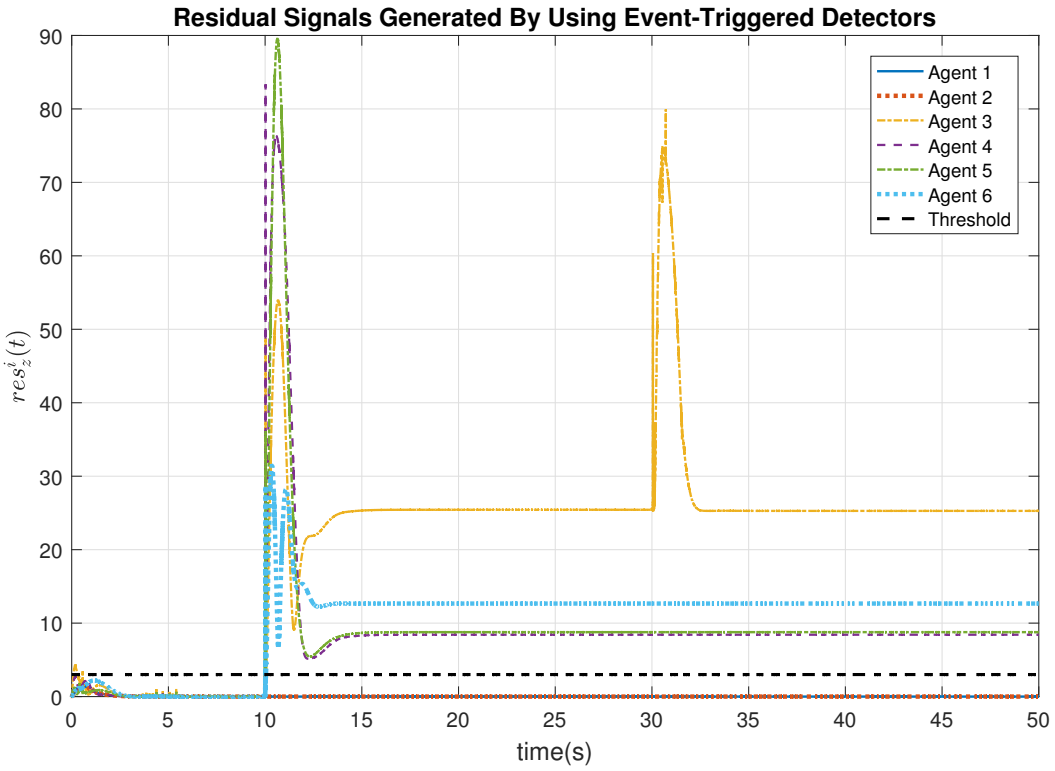

Figure 6.14: Residuals that are generated by using the event-triggered detectors and exceed the threshold in presence of quasi-covert cyber-attacks injected at $t = 30$ (s).

Chapter 7

Conclusions and Future Directions of Research

In this thesis, various cyber-attack detection methodologies for cyber-physical systems (CPS) were developed and studied from a system theoretic point of view. In Chapter 3, the focus was on addressing the challenge of simultaneously detecting and isolating machine-induced faults and intelligent malicious adversarial cyber-attacks within CPS. Two methodologies, centralized and distributed, were introduced based on the use of filters and unknown input observers (UIO) on both the plant and command and control (C&C) sides of the CPS. The proposed distributed methodology involves communication between UIO-based detectors of each subsystem and those in nearby subsystems. This enables each subsystem, under specific conditions, to identify and isolate both its own cyber-attacks and faults, as well as anomalies in neighboring subsystems. Through the centralized and distributed strategies, the simultaneous detection of machine-induced actuator and sensor faults, along with undetectable cyber-attacks like covert and zero dynamics attacks, and detectable cyber-attacks such as false data injection attacks, becomes achievable.

Chapter 4 explores the vulnerability of CPS to zero dynamics attacks, covert attacks, and controllable cyber-attacks. These stealthy cyber-attacks can cause damage to the CPS without being detected. Under certain assumptions, we derived conditions for the existence of these cyber-attacks based on nonzero Markov parameters and entries of the observability matrix. Moreover, based on the derived conditions, the required level of system knowledge and disruption resources for executing zero dynamics attacks, covert attacks, and

controllable attacks can be determined for CPS.

Consequently, a dynamic coding scheme was specifically introduced for zero dynamics and controllable attacks to maximize the security index of the CPS and increase their resiliency against the mentioned cyber threats. In presence of the dynamic coding schemes, securing just one actuator prevents adversaries from executing zero dynamics and controllable cyber-attacks. Also, challenges related to formulating necessary and sufficient conditions regarding disruption resources that adversaries need for executing covert cyber-attacks were tackled. These conditions help to identify the input and output communication channels necessary for performing covert attacks. Furthermore, an upper bound on the security index for covert attacks was defined which indicates the minimum number of actuators and sensors that need to be attacked for a successful covert cyber-attack. As a countermeasure, a dynamic coding scheme was developed that under certain design conditions and assuming the existence of secure input and output communication channels, prevents the occurrence of covert cyber-attacks.

In Chapter 5, the focus is on studying stealthy cyber-attacks within both linear and nonlinear CPS. The concept of security effort (SE) is introduced as a formal metric for linear CPS, representing the minimum number of input and output communication channels that should be secured to prevent adversaries from executing zero dynamics attacks, covert attacks, and controllable attacks. Since zero dynamics attacks and perfectly undetectable cyber-attacks belong to the weakly unobservable and controllable weakly unobservable subspaces of the CPS, the SE is defined and derived based on making these subspaces zero.

In the case of nonlinear CPS, the Koopman operator theory is employed to investigate data-driven stealthy cyber-attacks. The concept of ϵ-stealthy cyber-attacks is defined as a measure of detectability for nonlinear CPS. The Koopman canonical form of the nonlinear control affine CPS is utilized to determine the relative degree, enabling the discovery of internal dynamics, i.e., the zero dynamics, of the nonlinear CPS. Strategies for executing zero dynamics and covert cyber-attacks in nonlinear CPS are proposed, and conditions for securing sensor measurements to prevent these cyber-attacks are studied.

Chapter 6 addresses four key issues concerning cyber-attacks in multi-agent systems (MAS). First, the problem of protecting data privacy within MAS is explored. Isometric isomorphisms are utilized to transform the dynamics of agents into new bases, with each agent using a unique isometric isomorphism. A dynamic consensus protocol is developed, employing transformed sensor measurements and controller states

to achieve consensus while preserving agents' privacy. An algorithm is provided to ensure communication protocols maintain the indistinguishibility of transformed dynamics from eavesdropper adversaries and honest-but-curious agents.

The second problem investigates specific types of cyber-attacks on MAS. In one scenario, the adversary targets incoming communication links for a group of agents, disguising attack signals as transmitted information. Conditions for the adversary to have full control over the entire MAS by attacking a few number of agents are investigated. Moreover, notions of security controllability for individual agents and the entire MAS system are introduced. It is demonstrated that how the adversary can perform zero dynamics attacks on directly attacked agents and their followers.

The third problem concerns the detectability of cyber-attacks on communication links within MAS teams. A definition for undetectable cyber-attacks on MAS is proposed. Cyber-attacks on the root of a directed spanning tree graph are shown to be undetectable, while those on non-root agents can be detected if an uncompromised directed path from the root to these agents exists. Moreover, quasi-covert cyber-attacks are introduced as novel cyber-attacks in MAS. As for the fourth problem, an event-triggered detector is proposed to detect cyber-attacks in MAS. It is demonstrated that the proposed event-triggered detector can detect quasi-covert cyber-attacks in MAS.

7.1 Future Research Directions

The future research direction of this thesis can be summarized below.

(1) Considering that in real-world application, CPS contain nonlinearities, developing CAFDI methodologies for nonlinear CPS would be an interesting topic for further investigation.

(2) We have adopted a centralized design approach for the proposed dynamic coding schemes in this thesis. In order to make the coding scheme design and implementation more efficient, one needs to investigate developing distributed dynamic coding schemes that can be used in distributed CPS and MAS.

(3) Computation of the proposed security effort (SE) is an NP-hard problem. Hence, as for a future work, one may investigate a method to determine the SE in a generic manner that is computationally

efficient.

(4) Considering the availability of data for various systems in the modern world, developing a data-driven cyber-attack detection methodology for nonlinear CPS can be an interesting topic and a focus for future works.

(5) Our proposed privacy preserving consensus control methodology was developed for homogeneous MAS. Hence, as an extension of this work, one can focus on developing a privacy preserving consensus protocol for heterogeneous MAS.

(6) As for our proposed transformation method for privacy preserving control, we utilized Given's rotation and Householder transformation which are linear isometric isomorphisms. To extend this work, one can explore nonlinear isomorphisms for privacy preservation control which may lead to higher levels of data privacy protection.

(7) In Theorem 6.4, we have an assumption on the grounded Laplacian of the MAS. Relaxing this assumption is challenging and can be a focus of future research which can help to address the problem of controllability in MAS with directed graph that has applications in the general problem of actuator placement for large-scale CPS.

(8) A challenging topic for future investigation is to study and develop cyber-attack detection and monitoring systems for nonlinear MAS.

Bibliography

[1] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, no. Supplement C, pp. 135 – 148, 2015.

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

[3] H. Sandberg, V. Gupta, and K. H. Johansson, "Secure networked control systems," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 445–464, 2022.

[4] S. K. Khaitan and J. D. McCalley, "Design techniques and applications of cyberphysical systems: A survey," *IEEE Systems Journal*, vol. 9, no. 2, pp. 350–365, June 2015.

[5] A. A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *2008 The 28th International Conference on Distributed Computing Systems Workshops*, 2008, pp. 495–500.

[6] A. Cardenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, S. Sastry *et al.*, "Challenges for securing cyber physical systems," in *Workshop on future directions in cyber-physical systems security*, vol. 5, no. 1. Citeseer, 2009.

[7] A. Teixeira, "Toward cyber-secure and resilient networked control systems," Ph.D. dissertation, KTH Royal Institute of Technology, 2014.

[8] J. Slay and M. Miller, "Lessons learned from the maroochy water breach," in *International conference on critical infrastructure protection*. Springer, 2007, pp. 73–82.

[9] J. P. Farwell and R. Rohozinski, "Stuxnet and the future of cyber war," *Survival*, vol. 53, no. 1, pp. 23–40, 2011.

[10] R. Shirey, "Internet Security Glossary," https://www.rfc-editor.org/info/rfc2828, accessed: 2022-02-28.

[11] Y. Li, P. Zhang, L. Zhang, and B. Wang, "Active synchronous detection of deception attacks in microgrid control systems," *IEEE Transactions on Smart Grid*, vol. 8, no. 1, pp. 373–375, Jan 2017.

[12] F. Miao, M. Pajic, and G. J. Pappas, "Stochastic game approach for replay attack detection," in *52nd IEEE Conference on Decision and Control*, Dec 2013, pp. 1854–1859.

[13] M. Taheri, K. Khorasani, I. Shames, and N. Meskin, "Undetectable cyber attacks on communication links in multi-agent cyber-physical systems," in *2020 59th IEEE Conference on Decision and Control (CDC)*, 2020, pp. 3764–3771.

[14] Z. Zhao, Y. Yang, Y. Li, and R. Liu, "Security analysis for cyber-physical systems under undetectable attacks: A geometric approach," *International Journal of Robust and Nonlinear Control*.

[15] A. Baniamerian and K. Khorasani, "Security index of linear cyber-physical systems: A geometric perspective," in *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, April 2019, pp. 391–396.

[16] A. Baniamerian, K. Khorasani, and N. Meskin, "Determination of security index for linear cyber-physical systems subject to malicious cyber attacks," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 4507–4513.

[17] R. M. Ferrari and A. M. Teixeira, "A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks," *IEEE Transactions on Automatic Control*, vol. 66, no. 6, pp. 2558–2573, 2020.

[18] A. Teixeira, D. Pérez, H. Sandberg, and K. H. Johansson, "Attack models and scenarios for networked control systems," in *Proceedings of the 1st international conference on High Confidence Networked Systems*. ACM, 2012, pp. 55–64.

[19] R. Alisic and H. Sandberg, "Data-injection attacks using historical inputs and outputs," in *2021 European Control Conference (ECC)*, 2021, pp. 1399–1405.

[20] H. Sandberg and A. M. Teixeira, "From control system security indices to attack identifiability," in *2016 Science of Security for Cyber-Physical Systems Workshop (SOSCYPS)*, 2016, pp. 1–6.

[21] J. Milošević, A. Teixeira, K. H. Johansson, and H. Sandberg, "Actuator security indices based on perfect undetectability: Computation, robustness, and sensor placement," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3816–3831, 2020.

[22] S. Gracy, J. Milošević, and H. Sandberg, "Security index based on perfectly undetectable attacks: Graph-theoretic conditions," *Automatica*, vol. 134, p. 109925, 2021.

[23] S. Weerakkody, X. Liu, S. H. Son, and B. Sinopoli, "A graph-theoretic characterization of perfect attackability for secure design of distributed control systems," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 60–70, 2017.

[24] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2012, pp. 1806–1813.

[25] A. Baniamerian, K. Khorasani, and N. Meskin, "Monitoring and detection of malicious adversarial zero dynamics attacks in cyber-physical systems," in *2020 IEEE Conference on Control Technology and Applications (CCTA)*, 2020, pp. 726–731.

[26] M. Taheri, K. Khorasani, I. Shames, and N. Meskin, "Cyberattack and machine-induced fault detection and isolation methodologies for cyber-physical systems," *IEEE Transactions on Control Systems Technology*, pp. 1–16, 2023.

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

[28] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 93–109, 2015.

[29] 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, March 2017.

[30] S. Fang, K. H. Johansson, M. Skoglund, H. Sandberg, and H. Ishii, "Two-way coding in control systems under injection attacks: from attack detection to attack correction," in *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*, 2019, pp. 141–150.

[31] Y. Chen, S. Kar, and J. M. 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, 2016.

[32] M. Taheri, K. Khorasani, I. Shames, and N. Meskin, "Data-driven covert-attack strategies and countermeasures for cyber-physical systems," in *2021 60th IEEE Conference on Decision and Control (CDC)*, 2021, pp. 4170–4175.

[33] S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," in *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE, 2015, pp. 5820–5826.

[34] M. A. Massoumnia, "A geometric approach to the synthesis of failure detection filters," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 839–846, September 1986.

[35] S. H. Zad and M.-A. Massoumnia, "Generic solvability of the failure detection and identification problem," *Automatica*, vol. 35, no. 5, pp. 887 – 893, 1999.

[36] J. Chen, R. J. Patton, and H.-Y. Zhang, "Design of unknown input observers and robust fault detection filters," *International Journal of control*, vol. 63, no. 1, pp. 85–105, 1996.

[37] J. Wünnenberg and P. Frank, "Sensor fault detection via robust observers," in *System fault diagnostics, reliability and related knowledge-based approaches*. Springer, 1987, pp. 147–160.

[38] N. Tudoroiu and K. Khorasani, "Fault detection and diagnosis for satellite's attitude control system (acs) using an interactive multiple model (imm) approach," in *Proceedings of 2005 IEEE Conference on Control Applications, 2005. CCA 2005.*, Aug 2005, pp. 1287–1292.

[39] N. Meskin, E. Naderi, and K. Khorasani, "A multiple model-based approach for fault diagnosis of jet engines," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 1, pp. 254–262, Jan 2013.

[40] M. Davoodi, N. Meskin, and K. Khorasani, "Simultaneous fault detection and consensus control design for a network of multi-agent systems," *Automatica*, vol. 66, pp. 185 – 194, 2016.

[41] I. Shames, A. M. Teixeira, H. Sandberg, and K. H. Johansson, "Distributed fault detection for interconnected second-order systems," *Automatica*, vol. 47, no. 12, pp. 2757 – 2764, 2011.

[42] J. Gertler, "Fault detection and isolation using parity relations," *Control engineering practice*, vol. 5, no. 5, pp. 653–661, 1997.

[43] R. J. Patton and J. Chen, "A review of parity space approaches to fault diagnosis," *IFAC Proceedings Volumes*, vol. 24, no. 6, pp. 65–81, 1991.

[44] A. Barboni, H. Rezaee, F. Boem, and T. Parisini, "Detection of covert cyber-attacks in interconnected systems: A distributed model-based approach," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3728–3741, 2020.

[45] C. Schellenberger and P. Zhang, "Detection of covert attacks on cyber-physical systems by extending the system dynamics with an auxiliary system," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, Dec 2017, pp. 1374–1379.

[46] K. Zhang, C. Keliris, M. M. Polycarpou, and T. Parisini, "Discrimination between replay attacks and sensor faults for cyber-physical systems via event-triggered communication," *European Journal of Control*, vol. 62, pp. 47–56, 2021.

[47] S. Chockalingam, W. Pieters, A. Teixeira, and P. van Gelder, "Bayesian network model to distinguish between intentional attacks and accidental technical failures: a case study of floodgates," *Cybersecurity*, vol. 4, no. 1, pp. 1–19, 2021.

[48] H. Sandberg and A. M. Teixeira, "From control system security indices to attack identifiability," in *2016 Science of Security for Cyber-Physical Systems Workshop (SOSCYPS)*. IEEE, 2016, pp. 1–6.

[49] G. Park, C. Lee, and H. Shim, "On stealthiness of zero-dynamics attacks against uncertain nonlinear systems: A case study with quadruple-tank process," in *International Symposium on Mathematical Theory of Networks and Systems (ISMTNS)*, 2018, pp. 10–17.

[50] K. Zhang, C. Keliris, T. Parisini, and M. M. Polycarpou, "Stealthy integrity attacks for a class of nonlinear cyber-physical systems," *IEEE Transactions on Automatic Control*, 2021.

[51] B. O. Koopman, "Hamiltonian systems and transformation in hilbert space," *Proceedings of the national academy of sciences of the united states of america*, vol. 17, no. 5, p. 315, 1931.

[52] I. Mezić, "Analysis of fluid flows via spectral properties of the koopman operator," *Annual Review of Fluid Mechanics*, vol. 45, pp. 357–378, 2013.

[53] M. Budišić, R. Mohr, and I. Mezić, "Applied koopmanism," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 22, no. 4, p. 047510, 2012.

[54] D. Goswami and D. A. Paley, "Global bilinearization and controllability of control-affine nonlinear systems: A koopman spectral approach," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 6107–6112.

[55] A. Surana, "Koopman operator based observer synthesis for control-affine nonlinear systems," in *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 2016, pp. 6492–6499.

[56] M. Korda and I. Mezić, "Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control," *Automatica*, vol. 93, pp. 149–160, 2018.

[57] H. Arbabi, M. Korda, and I. Mezić, "A data-driven koopman model predictive control framework for nonlinear partial differential equations," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 6409–6414.

[58] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, "A data–driven approximation of the koopman operator: Extending dynamic mode decomposition," *Journal of Nonlinear Science*, vol. 25, no. 6, pp. 1307–1346, 2015.

[59] M. Korda and I. Mezić, "Optimal construction of koopman eigenfunctions for prediction and control," *IEEE Transactions on Automatic Control*, vol. 65, no. 12, pp. 5114–5129, 2020.

[60] M. Bakhtiaridoust, M. Yadegar, N. Meskin, and M. Noorizadeh, "Model-free geometric fault detection and isolation for nonlinear systems using koopman operator," *IEEE Access*, vol. 10, pp. 14 835–14 845, 2022.

[61] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, Sep. 2004.

[62] N. Meskin and K. Khorasani, "Actuator fault detection and isolation for a network of unmanned vehicles," *IEEE Transactions on Automatic Control*, vol. 54, no. 4, pp. 835–840, April 2009.

[63] E. Semsar-Kazerooni and K. Khorasani, "Multi-agent team cooperation: A game theory approach," *Automatica*, vol. 45, no. 10, pp. 2205–2213, 2009.

[64] J. Xu, L. Xie, T. Li, and K. Y. Lum, "Consensus of multi-agent systems with general linear dynamics via dynamic output feedback control," *IET Control Theory & Applications*, vol. 7, no. 1, pp. 108–115, 2013.

[65] M. Taheri, F. Sheikholeslam, M. Najafi, and M. Zekri, "Adaptive fuzzy wavelet network control of second order multi-agent systems with unknown nonlinear dynamics," *ISA Transactions*, vol. 69, pp. 89–101, 2017.

[66] Z. Li, Z. Duan, G. Chen, and L. Huang, "Consensus of multiagent systems and synchronization of complex networks: A unified viewpoint," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 1, pp. 213–224, 2009.

[67] W. Yu, F. Liang, X. He, W. G. Hatcher, C. Lu, J. Lin, and X. Yang, "A survey on the edge computing for the internet of things," *IEEE access*, vol. 6, pp. 6900–6919, 2017.

[68] A. Sultangazin, S. Diggavi, and P. Tabuada, "Protecting the privacy of networked multi-agent systems controlled over the cloud," in *2018 27th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 2018, pp. 1–7.

[69] T. Abdelzaher, N. Ayanian, T. Basar, S. Diggavi, J. Diesner, D. Ganesan, R. Govindan, S. Jha, T. Lepoint, B. Marlin, K. Nahrstedt, D. Nicol, R. Rajkumar, S. Russell, S. Seshia, F. Sha, P. Shenoy, M. Srivastava, G. Sukhatme, A. Swami, P. Tabuada, D. Towsley, N. Vaidya, and V. Veeravalli, "Will distributed computing revolutionize peace? the emergence of battlefield iot," in *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, 2018, pp. 1129–1138.

[70] A. Paverd, A. Martin, and I. Brown, "Modelling and automatically analysing privacy properties for honest-but-curious adversaries," *Tech. Rep.*, 2014.

[71] A. A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *2008 The 28th International Conference on Distributed Computing Systems Workshops*. IEEE, 2008, pp. 495–500.

[72] F. Farokhi, I. Shames, and N. Batterham, "Secure and private cloud-based control using semi-homomorphic encryption," *IFAC-PapersOnLine*, vol. 49, no. 22, pp. 163 – 168, 2016, 6th IFAC Workshop on Distributed Estimation and Control in Networked Systems NECSYS 2016.

[73] N. M. Freris and P. Patrinos, "Distributed computing over encrypted data," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2016, pp. 1116–1122.

[74] Y. Mo and R. M. Murray, "Privacy preserving average consensus," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 753–765, 2016.

[75] C. Murguia, F. Farokhi, and I. Shames, "Secure and private implementation of dynamic controllers using semihomomorphic encryption," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3950–3957, 2020.

[76] J. Tran, F. Farokhi, M. Cantoni, and I. Shames, "Implementing homomorphic encryption based secure feedback control," *Control Engineering Practice*, vol. 97, p. 104350, 2020.

[77] A. B. Alexandru, K. Gatsis, Y. Shoukry, S. A. Seshia, P. Tabuada, and G. J. Pappas, "Cloud-based quadratic optimization with partially homomorphic encryption," *IEEE Transactions on Automatic Control*, pp. 1–1, 2020.

[78] A. Sultangazin and P. Tabuada, "Symmetries and isomorphisms for privacy in control over the cloud," *IEEE Transactions on Automatic Control*, pp. 1–1, 2020.

[79] M. Kishida, "Encrypted average consensus with quantized control law," in *2018 IEEE Conference on Decision and Control (CDC)*, 2018, pp. 5850–5856.

[80] A. B. Alexandru, M. Schulze Darup, and G. J. Pappas, "Encrypted cooperative control revisited," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 7196–7202.

[81] M. Ruan, H. Gao, and Y. Wang, "Secure and privacy-preserving consensus," *IEEE Transactions on Automatic Control*, vol. 64, no. 10, pp. 4035–4049, 2019.

[82] M. Schulze Darup, A. Redder, and D. E. Quevedo, "Encrypted cooperative control based on structured feedback," *IEEE Control Systems Letters*, vol. 3, no. 1, pp. 37–42, 2019.

[83] D. Fiore and G. Russo, "Resilient consensus for multi-agent systems subject to differential privacy requirements," *Automatica*, vol. 106, pp. 18 – 26, 2019.

[84] L. Gao, S. Deng, and W. Ren, "Differentially private consensus with an event-triggered mechanism," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 1, pp. 60–71, 2018.

[85] A. Sultangazin and P. Tabuada, "Symmetries and privacy in control over the cloud: uncertainty sets and side knowledge*," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 7209–7214.

[86] Z. Feng, G. Hu, and G. Wen, "Distributed consensus tracking for multi-agent systems under two types of attacks," *International Journal of Robust and Nonlinear Control*, vol. 26, no. 5, pp. 896–918, 2016.

[87] W. He, X. Gao, W. Zhong, and F. Qian, "Secure impulsive synchronization control of multi-agent systems under deception attacks," *Information Sciences*, vol. 459, pp. 354–368, 2018.

[88] L. Ma, Z. Wang, and Y. Yuan, "Consensus control for nonlinear multi-agent systems subject to deception attacks," in *2016 22nd International Conference on Automation and Computing (ICAC)*. IEEE, 2016, pp. 21–26.

[89] A. Mustafa and H. Modares, "Attack analysis for discrete-time distributed multi-agent systems," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2019, pp. 230–237.

[90] A. Mustafa and H. Modares, "Attack analysis and resilient control design for discrete-time distributed multi-agent systems," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 369–376, April 2020.

[91] F. Boem, A. J. Gallo, G. Ferrari-Trecate, and T. Parisini, "A distributed attack detection method for multi-agent systems governed by consensus-based control," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 5961–5966.

[92] H. Zhang, G. Feng, H. Yan, and Q. Chen, "Observer-based output feedback event-triggered control for consensus of multi-agent systems," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 9, pp. 4885–4894, 2013.

[93] D. Yang, W. Ren, X. Liu, and W. Chen, "Decentralized event-triggered consensus for linear multi-agent systems under general directed graphs," *Automatica*, vol. 69, pp. 242–249, 2016.

[94] M. Davoodi, N. Meskin, and K. Khorasani, "Event-triggered multiobjective control and fault diagnosis: A unified framework," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 1, pp. 298–311, Feb 2017.

[95] S. Hajshirmohamadi, F. Sheikholeslam, M. Davoodi, and N. Meskin, "Event-triggered simultaneous fault detection and tracking control for multi-agent systems," *International Journal of Control*, vol. 92, no. 8, pp. 1928–1944, 2019.

[96] R. S. Smith, "A decoupled feedback structure for covertly appropriating networked control systems," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 90–95, 2011.

[97] J. Tokarzewski, *Finite zeros in discrete time control systems*. Springer, 2006, vol. 338.

[98] H. L. Trentelman, A. A. Stoorvogel, and M. Hautus, *Control theory for linear systems*. Springer Science & Business Media, 2012.

[99] M.-A. Massoumnia, G. C. Verghese, and A. S. Willsky, "Failure detection and identification," *IEEE Transactions on Automatic Control*, vol. 34, no. 3, pp. 316–321, 1989.

[100] M. Mola, N. Meskin, K. Khorasani, and A. Massoud, "Distributed event-triggered consensus-based control of dc microgrids in presence of dos cyber attacks," *IEEE Access*, vol. 9, pp. 54 009–54 021, 2021.

[101] A. J. Gallo, M. S. Turan, F. Boem, T. Parisini, and G. Ferrari-Trecate, "A distributed cyber-attack detection scheme with application to dc microgrids," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3800–3815, 2020.

[102] W. M. Wonham, "Linear multivariable control," in *Optimal control theory and its applications*. Springer, 1974, pp. 392–424.

[103] H. K. Khalil and J. W. Grizzle, *Nonlinear systems*. Prentice hall Upper Saddle River, NJ, 2002, vol. 3.

[104] 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.

[105] J. Miller, T. Dai, M. Sznaier, and B. Shafai, "Data-driven control of positive linear systems using linear programming," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, 2023, pp. 1588–1594.

[106] B. Shafai, A. Moradmand, and M. Siami, "Data-driven positive stabilization of linear systems," in *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)*, vol. 1, 2022, pp. 1031–1036.

[107] R. Gore and M. Kande, "Analysis of wide area monitoring system architectures," in *2015 IEEE International Conference on Industrial Technology (ICIT)*, 2015, pp. 1269–1274.

[108] S. Nabavi, J. Zhang, and A. Chakrabortty, "Distributed optimization algorithms for wide-area oscillation monitoring in power systems using interregional pmu-pdc architectures," *IEEE Transactions on Smart Grid*, vol. 6, no. 5, pp. 2529–2538, 2015.

[109] S. Nabavi and A. Chakrabortty, "Topology identification for dynamic equivalent models of large power system networks," in *2013 American Control Conference*. IEEE, 2013, pp. 1138–1143.

[110] S. Trip, M. Cucuzzella, C. De Persis, A. van der Schaft, and A. Ferrara, "Passivity-based design of sliding modes for optimal load frequency control," *IEEE Transactions on Control Systems Technology*, vol. 27, no. 5, pp. 1893–1906, 2019.

[111] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861 – 874, 2006, rOC Analysis in Pattern Recognition.

[112] M. Sain and J. Massey, "Invertibility of linear time-invariant dynamical systems," *IEEE Transactions on Automatic Control*, vol. 14, no. 2, pp. 141–149, 1969.

[113] A. MacFarlane and N. Karcanias, "Poles and zeros of linear multivariable systems: a survey of the algebraic, geometric and complex-variable theory," *International Journal of Control*, vol. 24, no. 1, pp. 33–74, 1976.

[114] J. Tokarzewski, "System zeros analysis via the moore-penrose pseudoinverse and svd of the first nonzero markov parameter," *IEEE Transactions on Automatic Control*, vol. 43, no. 9, pp. 1285–1291, 1998.

[115] A. Hajdasinski and A. A. H. Damen, *Realization of the Markov parameter sequences using the singular value decomposition of the Hankel matrix*. Technische Hogeschool Eindhoven, 1979.

[116] B. De Moor, J. Vandewalle, M. Moonen, L. Vandenberghe, and P. Van Mieghem, "A geometrical strategy for the identification of state space models of linear multivariable systems with singular value decomposition," *IFAC Proceedings Volumes*, vol. 21, no. 9, pp. 493–497, 1988.

[117] L. Ljung, "System identification," *Wiley encyclopedia of electrical and electronics engineering*, pp. 1–19, 1999.

[118] J. Dong and M. Verhaegen, "Identification of fault estimation filter from i/o data for systems with stable inversion," *IEEE Transactions on Automatic Control*, vol. 57, no. 6, pp. 1347–1361, 2011.

[119] O. Härkegård and S. T. Glad, "Resolving actuator redundancy—optimal control vs. control allocation," *Automatica*, vol. 41, no. 1, pp. 137–144, 2005.

[120] B. Boussaid, C. Aubrun, J. Jiang, and M. N. Abdelkrim, "Ftc approach with actuator saturation avoidance based on reference management," *International Journal of Robust and Nonlinear Control*, vol. 24, no. 17, pp. 2724–2740, 2014.

[121] M. Taheri, K. Khorasani, N. Meskin, and I. Shames, "Data-driven koopman operator based cyber-attacks for nonlinear control affine cyber-physical systems," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 6769–6775.

[122] M. Taheri, K. Khorasani, and N. Meskin, "The security requirement to prevent zero dynamics attacks and perfectly undetectable cyber-attacks in cyber-physical systems," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, 2023, pp. 7067–7072.

[123] 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.

[124] A. Baniamerian and K. Khorasani, "Security index of linear cyber-physical systems: A geometric perspective," in *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, 2019, pp. 391–396.

[125] A. Baniamerian, K. Khorasani, and N. Meskin, "Determination of security index for linear cyber-physical systems subject to malicious cyber attacks*," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, Dec 2019, pp. 4507–4513.

[126] Z. Zhao, Y. Yang, Y. Li, and R. Liu, "Security analysis for cyber-physical systems under undetectable attacks: A geometric approach," *International Journal of Robust and Nonlinear Control*, vol. 30, no. 11, pp. 4359–4370, 2020.

[127] R. Mohr and I. Mezić, "Construction of eigenfunctions for scalar-type operators via laplace averages with connections to the koopman operator," *arXiv preprint arXiv:1403.6559*, 2014.

[128] G. Basile and G. Marro, *Controlled and Conditioned Invariants in Linear System Theory*. Prentice Hall Englewood Cliffs, 1992.

[129] A. Isidori, *Nonlinear control systems: an introduction*. Springer, 1985.

[130] A. Mauroy, Y. Susuki, and I. Mezić, *Koopman operator in systems and control*. Springer, 2020.

[131] M. Taheri, K. Khorasani, I. Shames, and N. Meskin, "Towards privacy preserving consensus control in multi-agent cyber-physical systems subject to cyber attacks," in *2021 European Control Conference (ECC)*, 2021, pp. 939–945.

[132] M. Taheri, K. Khorasani, I. Shames, and N. Meskin, "Mitigation and resiliency of multi-agent systems subject to malicious cyber attacks on communication links," in *2020 IEEE Conference on Control Technology and Applications (CCTA)*, 2020, pp. 857–862.

[133] W. Ren and Y. Cao, *Distributed coordination of multi-agent networks: emergent problems, models, and issues*. Springer Science & Business Media, 2010.

[134] R. A. Horn and C. R. Johnson, "Topics in matrix analysis cambridge university press," *Cambridge, UK*, 1991.

[135] H. Zhang, F. L. Lewis, and A. Das, "Optimal design for synchronization of cooperative systems: state feedback, observer and output feedback," *IEEE Transactions on Automatic Control*, vol. 56, no. 8, pp. 1948–1952, 2011.

[136] J. Wang, Z. Liu, and X. Hu, "Consensus of high order linear multi-agent systems using output error feedback," in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. IEEE, 2009, pp. 3685–3690.

[137] Y. Desmedt, "Man-in-the-middle attack," *Encyclopedia of cryptography and security*, pp. 759–759, 2011.

[138] D. Juneja, A. Singh, and A. Jagga, "Kqml based communication protocol for multi agent systems," *International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)*, vol. 12, pp. pp. 158–162, 05 2015.

[139] C.-K. Li and W. So, "Isometric isomorphisms between normed spaces," *The Rocky Mountain journal of mathematics*, pp. 607–624, 1998.

[140] P. Ginzberg and C. Mavroyiakoumou, "The qrd and svd of matrices over a real algebra," *Linear Algebra and its Applications*, vol. 504, pp. 27 – 47, 2016.

[141] A. S. Householder, "Unitary triangularization of a nonsymmetric matrix," *Journal of the ACM (JACM)*, vol. 5, no. 4, pp. 339–342, 1958.

[142] M. Pirani and S. Sundaram, "Spectral properties of the grounded laplacian matrix with applications to consensus in the presence of stubborn agents," in *2014 American Control Conference*. IEEE, 2014, pp. 2160–2165.

[143] W. L. Brogan, *Modern control theory*. Pearson education india, 1991.

[144] J. C. Willems and J. W. Polderman, *Introduction to mathematical systems theory: a behavioral approach*. Springer Science & Business Media, 1997, vol. 26.

[145] T. H. Summers and J. Lygeros, "Optimal sensor and actuator placement in complex dynamical networks," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 3784–3789, 2014.

[146] A. Rahmani, M. Ji, M. Mesbahi, and M. Egerstedt, "Controllability of multi-agent systems from a graph-theoretic perspective," *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 162–186, 2009.

[147] M. Ji, A. Muhammad, and M. Egerstedt, "Leader-based multi-agent coordination: Controllability and optimal control," in *2006 American Control Conference*. IEEE, 2006, pp. 6–pp.

[148] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, "Control centrality and hierarchical structure in complex networks," *Plos one*, vol. 7, no. 9, p. e44459, 2012.

[149] A. Olshevsky, "Minimal controllability problems," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 3, pp. 249–258, 2014.

[150] K. Fitch and N. E. Leonard, "Optimal leader selection for controllability and robustness in multi-agent networks," in *2016 European Control Conference (ECC)*. IEEE, 2016, pp. 1550–1555.

[151] V. Ugrinovskii, "Conditions for detectability in distributed consensus-based observer networks," *IEEE Transactions on Automatic Control*, vol. 58, no. 10, pp. 2659–2664, 2013.