UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN

Ingegneria dell'Automazione

"Model-Based Distributed Strategies for Detection and Isolation of Covert Cyber-Attacks in Networked Systems"

Relatore: Prof. Angelo Cenedese
Relatore: Prof Thomas Parisini (Imperial College London, Department of Electrical and Electronic Engineering)
Co-relatore: Angelo Barboni (Imperial College London, Department of Electrical and Electronic Engineering)

Laureando: Tommaso Benciolini

ANNO ACCADEMICO 2019 –2020

21 Luglio 2020

Imperial College London

Università degli Studi di Padova

University of Padova
Department of Information Engineering
Master's Degree in Automation Engineering

Model-based Distributed Strategies for Detection and Isolation of Covert Cyber-Attacks in Networked Systems

Supervisor:
Prof. Angelo Cenedese

Supervisor:
Prof. Thomas Parisini

Co-supervisor:
Angelo Barboni

Candidate:
Tommaso Benciolini
ID 1197481

Academic year 2019/2020
Tuesday, July 21, 2020

"I have mentioned Mathematics as a way to settle in the mind a habit of reasoning closely and in train; not that I think it necessary that all men should be deep mathematicians, but that, having got the way of reasoning which that study necessarily brings the mind to, they might be able to transfer it to other parts of knowledge, as they shall have occasion. For, in all sorts of reasoning, every single argument should be managed as a mathematical demonstration; the connection and dependence of ideas should be followed till the mind is brought to the source on which it bottoms, and observes the coherence all along"

John Locke, "The Conduct of the Understanding"

"If you can't explain it simply, you don't understand it well enough"

Albert Einstein

To those who inspired and supported me

Abstract

This thesis project investigates the cyber-security problem for linear interconnected systems in a distributed fashion. Starting from existing results on the detectability of covert cyber-attacks in a single agent of the network, the work addresses the isolation task, proposing different algorithms based on the algebraic properties of the interconnection matrices of each local neighborhood. Moreover, the detection problem is extended to the scenario of multiple agents simultaneously attacked. The whole theoretical analysis focuses on large-scale systems subject to bounded process and measurement disturbances. All the proposed methodologies can be implemented by using only local information available at each subsystem, and are endowed with a suitable threshold to avoid false alarms due to action of noise. Finally, numerical simulations on a simple data center model are given, showing the effectiveness of the introduced techniques in detecting and isolating covert cyber-attacks.

Contents

Chapter 1

Introduction

This thesis project addresses the problem of cyber-security in the context of dynamical systems. Differently from the perspective typically adopted in the computer engineering community, for a control system engineer this task naturally falls in the framework of state estimation and plant monitoring. The aim is to derive some sensitive quantities, typically called residuals, which highlight possible deviations of the system state from its nominal behavior as a consequence of the action of a malicious agent.

With respect to the Fault Detection problem, from which cyber-security inherits many methodologies, this setting is far more challenging. Indeed, an intentional manipulation by an intelligent attacker might be designed so as to make it difficult for a monitoring unit to detect it. Interestingly, linear systems are particularly vulnerable to this eventuality. As a matter of fact, the properties of linear dynamical systems have been extensively discussed and characterized in literature, and a multitude of control and estimation techniques have been derived and formally supported from a theoretical perspective. All these results rely on properties following from the simple mathematical structure of the linear systems. Nonetheless, in the same fashion, such properties make it particularly easy for the attacker to perfectly compensate for its action so that the monitoring unit cannot detect it by looking at some residual quantities.

The security problem is particularly interesting in interconnected systems. Indeed, many critical infrastructures are nowadays designed as a network of agents mutually influencing each other through some kind of coupling, and their reliability against external attacks is of primary concern. However, especially when the structure being considered is a large-scale system, distributed methodologies for both control and monitoring are extremely important, since they reduce the need for communication, and improve the scalability of the procedures.

The thesis extends the results of a previous work dealing with the detection problem in interconnected systems [1], both by focusing on the isolation issue, and by introducing the detection analysis to the scenario of simultaneous multiple attacks. For each of these tasks, a detailed discussion on the structural

properties of the matrices of the network is presented. Note that characterizing the theoretical properties of the systems allowing the proposed methodologies to work is relevant for a design purpose. Indeed, whenever possible, one should avoid all those configurations which are "intrinsically vulnerable", intentionally giving up some links (if needed) to properly tune the trade-off between security and required performance.

The organization of the thesis is the following. In Chapter 2, the contributions of this work are collocated in the present state of the art. Chapter 3 is dedicated to an outline of the results in [1]. The novel contributions about the isolation issue and the detection of simultaneous multiple attacks are presented in Chapter 4 and 5, respectively. In Chapter 6 a model of a data center is proposed, in order to explain how the developed strategies work in practice through some numerical simulations. Finally, in Chapter 7 possible future research directions are suggested. Moreover, in the next section the mathematical notation used is summarized, and at the end of the text a complete list of used symbols is reported.

1.1 Mathematical notation

Given any finite set A, $|A|$ denotes its cardinality. $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{R}$ denote the sets of natural, integer, and real numbers, respectively. Given a natural number k, $k!$ denotes its factorial. $\lfloor \cdot \rfloor$ denotes the floor operator of a real number.

All vectors are understood as column-vector, and $v_{[i]}$ denotes the ith entry of vector v. Matrices are denoted by means of capital symbols. $\dim(\mathcal{V})$ denotes the dimension of a linear space $\mathcal{V}$. Moreover, 0 is a matrix whose entries are all zero, I is the identity matrix. Given a matrix M, $M^\top$ indicates its transpose, $\mathrm{Im}(M)$ is its image, and $\mathrm{rank}(M) \doteq \dim(\mathrm{Im}(M))$ is the rank of such a matrix. M^{-L} is the Moore-Penrose left-pseudoinverse and, if M is a square matrix with full rank, M^{-1} is its inverse. A square matrix M is called Schur-stable if all its eigenvalues lie within the open circle of radius one in the complex plane. Given a sequence of matrices $\{M_j : j \in \mathcal{I}\}$, $\mathrm{row}_{j\in\mathcal{I}}(M_j)$ denotes the horizontal concatenation of said matrices. $\|\cdot\| \doteq \|\cdot\|_2$ refers to the standard Euclidean 2-norm for vectors, and to the induced norm for matrices.

Given a signal $s(t)$, $s(k) \doteq s(kT_s)$ indicates its value at the kth sampling instant, where T_s is the sampling time. Moreover, $s^+ \doteq s(k+1)$ and $s^- \doteq s(k-1)$ represent its value at the next and previous sampling instant, respectively.

Given a random variable x, $\hat{x}$ is an estimate of x, and $\mathbb{E}[x]$ denotes its expected value. Moreover, the notation $x \sim \mathcal{P}$ is used to describe that x is characterized as a Poisson variable. Finally, $p_\lambda(\cdot)$ denotes the probability mass function of a Poisson variable of parameter $\lambda \in \mathbb{R}, \lambda > 0$.

Chapter 2

State of the art

The interest around the problem of cyber-security in the context of interconnected systems is growing in recent research. Indeed, several infrastructures become more and more essential in modern society, and they can be described as a network of dynamical systems mutually fulfilling a task by cooperating one another. Among them, we find water supply and distribution networks in general, power grids, telecommunication networks, transportation networks, and industrial processes. The malfunctioning of these infrastructure greatly affects both our lives and our economy. These systems are typically controlled over a communication network. The measurements are transmitted to some control center, then the control signal is forwarded to the actuators via the communication channel as well. Due to this information exchange, these systems are vulnerable to both faults and intentional manipulations performed by malicious agents.

A malicious agent might act in order to drive the system to non-optimal (and potentially dangerous) operating conditions, by affecting the communication channel of such cyber-systems [2]. Power networks, for instance, are operated through supervisory control and data acquisition (SCADA) systems, which can easily be attacked by external agents [3], [4]. More in general, a cyber-attack could result in immediate consequence (for example, a blackout of a power distribution network), or in long-term deterioration of the manipulated plant, due to the improper handling.

Generally speaking, when a communication channel is employed, a system is exposed to *man-in-the-middle* attack strategies. Indeed, if the communication link between the control logic unit and the actuators/sensors can be hacked (for instance, with wireless communication technology), an attacker can effectively affect both the input and the output signals, in order to accomplish different sorts of manipulations. For example, in [5] the authors distinguish the following classes of cyber-attacks:

a) *False Data Injection Attacks*: This is the simplest attack to be considered. The attacker's action is restricted to the alteration of the actuation and/or

measurement signals. In doing so, an attacker can modify the equilibrium of a network. On the one hand, this type of attack is not particularly difficult to implement, as no disclosure capability is required (i.e., the malicious agent does not need to eavesdrop the information sent through the communication link; however, model knowledge is required. On the other hand, a false data injection attack can be fairly easily detected by comparing the expected output with the received measurement.

b) *Replay Attacks*: If an attacker is able to more aggressively violate the integrity of the communication network (meaning that it can listen to the transmitted signals) for a certain time span, it can later on modify the transmitted information by replaying stored old data. In such a way, it can effectively disguise any changes to the operating conditions. Specifically, it has been shown that replay attacks may be undetectable to attack monitoring schemes [6], as the replayed data has both the same statistical properties of the non-attacked data, and it evolves following correct dynamics. Still, observe that a malicious agent willing to perform a replay attack does not need any knowledge of the system's dynamics.

c) *Covert Attacks*: A resourceful attacker can properly design the manipulation on the actuation and measurement signals in order to exactly compensate for its action on the received output, irrespective of its action on the system. Indeed, if the attacker is able to run a replica of the system, it is also able, from that, to deduce the effect of its action on the system; therefore, it can counterbalance for its action on the output signal, making it impossible for the control and monitoring architecture to identify its presence. In order to achieve this attack, the malicious agent needs to have knowledge of the system's dynamics. Covert attacks are the most difficult to detect, and are the focus of this thesis work.

A vast number of works regarding the problem of security of cyber-physical systems have been presented in literature, such as [7], [8], [9], [10], and [11]. It is worth highlighting that, given the typical complexity of these systems, *distributed* algorithms for detecting anomalies are of particular interest. Some approaches, in particular, have been inspired by the field of distributed fault detection and isolation (FDI), a research area dealing with the problem of detecting and, possibly, locating the source of faults resulting in unexpected trends in the behavior of a monitored system (see, for instance, [12] and [13]). However, extending FDI techniques to the context of cyber security is far from trivial, since an intelligent malicious agent may be able to act in such a way not to be detected by these monitoring strategies, since a cyber-attack can affect the behavior of a system in a richer way than typical classes of faults.

Differently from the typical framework of the computer science research community, cyber security in control literature is typically addressed by assuming a dynamical model of the system being attacked and of its interconnections is

available, and this is the setting of this thesis work. Starting from the results on the detection of cyber-attacks in interconnected systems presented in [1], reported in Chapter 3, the main contribution of this work is related to isolation (Chapter 4), and an introductory analysis on the scenario of simultaneous coordinated multiple attacks within the same network of dynamical systems (Chapter 5). All the proposed results extensively rely on topology of the network being considered, and on the structural properties of the interconnection matrices themselves.

Chapter 3

Detection strategy

This chapter outlines in detail the distributed detection strategy presented in [1]. With respect to the cited work, the strategy is here adapted to fit the discrete-time scenario and, finally, different thresholds for the considered quantity are derived.

The chapter firstly describes the subsystems and the network considered (Section 3.1. Secondly, the details about the covert attack are given (Section 3.2). Finally, the detection strategy is illustrated in depth (Section 3.3).

3.1 Subsystems description

Let consider a networked system composed of N subsystems, where the generic ith component is characterized as a linear time-invariant dynamical system in the form:

$$\mathcal{S}_i : \begin{cases} x_i^+ = A_i x_i + B_i \tilde{u}_i + \sum_{j \in \mathcal{N}_i} A_{ij} x_j + w_i \\ y_i = C_i x_i + v_i, \end{cases} \tag{3.1}$$

where $x_i \in \mathbb{R}^{n_i}$ is the subsystem state vector, $\tilde{u}_i \in \mathbb{R}^{m_i}$ is the control input vector, $y_i \in \mathbb{R}^{p_i}$ is the output vector, and $w_i \in \mathbb{R}^{n_i}$ and $v_i \in \mathbb{R}^{p_i}$ denote the process noise and measurement noise, respectively. Moreover, $A_i \in \mathbb{R}^{n_i \times n_i}$ is the state matrix, and $B_i \in \mathbb{R}^{n_i \times m_i}$ and $C_i \in \mathbb{R}^{p_i \times n_i}$ are the input and output matrix, respectively. The neighbor set $\mathcal{N}_i$ of $\mathcal{S}_i$ represents the index set of those subsystems $\mathcal{S}_j$ dynamically influencing the subsystem $\mathcal{S}_i$ through the interconnection matrix $A_{ij} \in \mathbb{R}^{n_i \times n_j}$. All the matrices are assumed to be constant over time. Finally, let define:

$$\Xi_i \doteq \operatorname*{row}_{j \in \mathcal{N}_i}(A_{ij}). \tag{3.2}$$

Concerning the structural properties of the subsystems composing the network, we make the following assumptions.

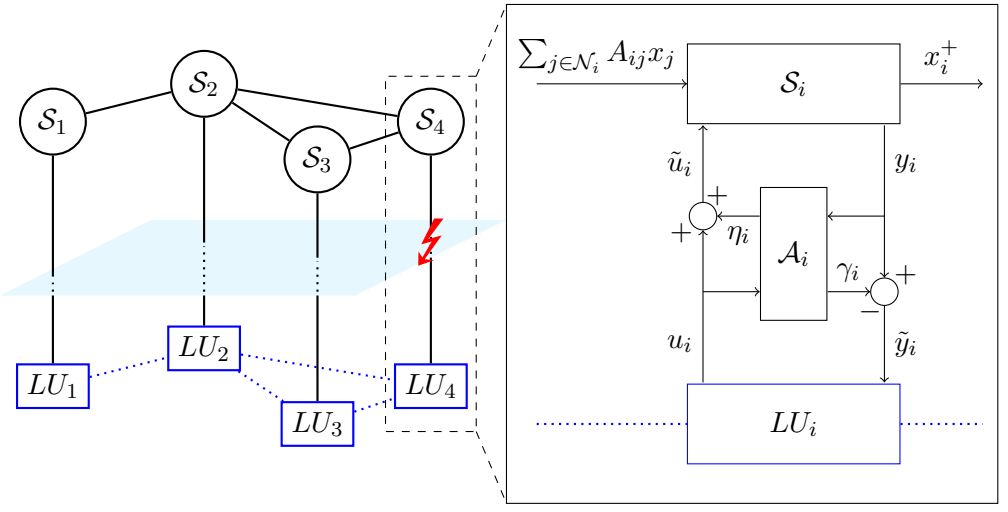


Figure 3.1: On the left, the network layout separated in physical and cyber layers; on the right, the diagram of the attacked subsystem.

Assumption 3.1.1. *$\forall i \in \{1 \dots N\}$, the pair (A_i, C_i) of subsystem $\mathcal{S}_i$ is observable and:*

$$\text{rank}(C_i \Xi_i) = \text{rank}(\Xi_i). \tag{3.3}$$

Assumption 3.1.2. *The topology of the network can be represented as an undirected graph, without self-loops, that is:*

$$\begin{aligned} i \notin &\mathcal{N}_i \\ i \in \mathcal{N}_j \Leftrightarrow &j \in \mathcal{N}_i. \end{aligned} \tag{3.4}$$

On the other hand, in general it is $A_{ij} \neq A_{ji}$.

Furthermore, the noises acting on each subsystem are assumed to satisfy the following condition.

Assumption 3.1.3. *$\forall i \in \{1 \dots N\}$, both the process and the measurement noises are bounded, i.e. there exists known positive constants $\bar{w}_i, \bar{v}_i \in \mathbb{R}_+$ such that:*

$$\|w_i(k)\| \leq \bar{w}_i, \|v_i(k)\| \leq \bar{v}_i, \forall k \in \mathbb{Z}, k \geq 0. \tag{3.5}$$

A scheme of the network is shown in Figure 3.1. Each subsystem is equipped with a local unit LU_i, consisting of a controller $\mathcal{C}_i$ and the detection architecture, which is thoroughly described in Section 3.3. The logic unit accesses the measurements $\tilde{y}_i$ and produces the control input u_i. The action of the logic unit LU_i is fully distributed, meaning that it is the result of the locally available information and variables of equation (3.1), whereas no knowledge of the overall

topology of the network is needed. Therefore, we assume the following condition to hold.

Assumption 3.1.4. *Only the local dynamics' matrices* (A_i, B_i, C_i)*, and the interconnection matrices* $A_{ij}, \forall j \in \mathcal{N}_i$*, are available to each* LU_i*.*

The link connecting each subsystem $\mathcal{S}_i$ with the associated logic unit LU_i is vulnerable. As a consequence, the signals entering the subsystem $\mathcal{S}_i$ might be different from those produced by the logic unit LU_i and viceversa. Specifically, in order to take into account the action of a malicious agent $\mathcal{A}_i$ altering both the actuation and the measurement signals, which will be discussed in Section 3.2, we will refer to the couple (u_i,y_i) as *legitimate* or *transmitted* signals, whereas their corrupted version ($\tilde{u}_i$,$\tilde{y}_i$) will be referred to as *attacked* or *received*.

If i denotes the index of the subsystem where the attacker is acting, the relation between the transmitted and the received signals is the following:

$$\begin{aligned} \tilde{u}_i &= u_i + \eta_i \\ \tilde{y}_i &= y_i - \gamma_i. \end{aligned} \tag{3.6}$$

Such cyber-attacks, in which a malicious agent can alter both the actuation and the measurement signals, are particularly difficult to detect. Indeed, a proper design of η_i, γ_i can make the attack effect on the output indistinguishable from the nominal behavior, independently of the manipulation on the subsystem $\mathcal{S}_i$. An exhaustive discussion on this important aspect can be found in the following section.

3.2 Covert attack

In this section, we depict the design of a covert attack model in state space form, with reference to [14].

Definition 3.2.1 (Covert Agent)**.** *The action of the malicious agent* $\mathcal{A}_i$ *is covert to subsystem* $\mathcal{S}_i$ *if the received measurement output* $\tilde{y}_i$ *is indistinguishable from the nominal subsystem response* y_i *in the attack-free scenario.*

In other words, an attack is covert as far as the attacker is able compensate for its action so that no abnormality can be deduced from the received output $\tilde{y}_i$ (in the following, this condition will be referred to as *covertness property*). Having said that, we trivially deduce that such attacks are *stealthy* by design. Moreover, from Definition 3.2.1 follows that any residual signal relying on the attacked output measurement $\tilde{y}_i$ necessarily satisfies the stealthiness condition [15, Definition 2].

An effective covert attack can be performed by replicating the dynamics of the targeted subsystem. Hence, the malicious agent $\mathcal{A}_i$ is modeled as dynamical

system in the form:

$$\tilde{\mathcal{S}}_i : \begin{cases} \tilde{x}_i^+ = \tilde{A}_i \tilde{x}_i + \tilde{B}_i \eta_i \\ \gamma_i = \tilde{C}_i \tilde{x}_i. \end{cases} \tag{3.7}$$

Specifically, the attacker design the signal η_i in order to fulfill its goal. For example, η_i might drive the subsystem $\mathcal{S}_i$ toward an undesired trajectory, or cause the subsystem state x_i to grow indefinitely. Obviously, the signal η_i is arbitrary, and its characteristics are in general unknown to a defender. On the other hand, the signal γ_i is computed with the purpose of compensating for the effect of η_i on the subsystem output y_i as in (3.6).

Remark 3.2.1. *Observe that, thanks to the superposition principle, the attacker $\mathcal{A}_i$ might run the replica $\tilde{\mathcal{S}}_i$ in open loop, meaning that it does not need any information on the current state x_i and legitimate input u_i of the subsystem $\mathcal{S}_i$ to hide its own effect on the output.*

As a result, model (3.7) is itself sufficient to describe a covert agent. On the other hand, it is reasonable to consider a more general scenario in which the attacker needs to implement its own controller $\tilde{\mathcal{C}}_i$ in order to achieve some desired dynamics:

$$\tilde{\mathcal{C}}_i : \begin{cases} \xi_i^+ = A_{\tilde{\mathcal{C}}_i} \xi_i + \Upsilon_i \begin{bmatrix} u_i \\ y_i \end{bmatrix} + R_{\tilde{\mathcal{C}}_i} \nu_i \\ \eta_i = C_{\tilde{\mathcal{C}}_i} \xi_i + K_{\tilde{\mathcal{C}}_i} \tilde{x}_i, \end{cases} \tag{3.8}$$

where ξ_i is the controller state, ν_i is used to determine the controller's reference, and $A_{\tilde{\mathcal{C}}_i}, \Upsilon_i, R_{\tilde{\mathcal{C}}_i}, C_{\tilde{\mathcal{C}}_i}$, and $K_{\tilde{\mathcal{C}}_i}$ are matrices of compatible dimensions. In particular, $K_{\tilde{\mathcal{C}}_i}$ provides a feedback from the state $\tilde{x}_i$ of $\tilde{\mathcal{S}}_i$, whereas Υ_i represents the *disclosure resources* as in [15], identifying information accessible by the attacker.

The attacker $\mathcal{A}_i$ can be represented in compact form by considering both (3.7) and (3.8), and by introducing a vector $\zeta_i \doteq \begin{bmatrix} \tilde{x}_i^\top & \xi_i^\top \end{bmatrix}^\top$ as follows:

$$\mathcal{A}_i : \begin{cases} \zeta_i^+ = \Phi_i \zeta_i + \begin{bmatrix} 0 \\ \Upsilon_i \end{bmatrix} \begin{bmatrix} u_i \\ y_i \end{bmatrix} + \begin{bmatrix} 0 \\ R_{\tilde{\mathcal{C}}_i} \end{bmatrix} \nu_i \\ \begin{bmatrix} \gamma_i \\ \eta_i \end{bmatrix} = \Gamma_i \zeta_i, \end{cases} \tag{3.9}$$

where:

$$\Phi_i = \left[\begin{array}{c|c} \tilde{A}_i + \tilde{B}_i K_{\tilde{\mathcal{C}}_i} & \tilde{B}_i C_{\tilde{\mathcal{C}}_i} \\ \hline 0 & A_{\tilde{\mathcal{C}}_i} \end{array}\right], \quad \Gamma_i = \left[\begin{array}{c|c} \tilde{C}_i & 0 \\ \hline K_{\tilde{\mathcal{C}}_i} & C_{\tilde{\mathcal{C}}_i} \end{array}\right]. \tag{3.10}$$

Γ_i is called *disruption resource*, as it defines which channels among actuation and measurement can be compromised with malicious signals. With this description, the attacker $\mathcal{A}_i$ is completely characterized by its model knowledge

$(\tilde{A}_i, \tilde{B}_i, \tilde{C}_i)$, its *infiltration resources* Υ_i and Γ_i, and its attack strategy defined by the controller $\tilde{\mathcal{C}}_i$ and the reference signal ν_i.

Clearly, (3.7) satisfies the covertness property if and only if the replica $\tilde{\mathcal{S}}_i$ is a realization of the same transfer function realized by the subsystem being attacked $\mathcal{S}_i$. To ensure this condition, the following is assumed to hold.

Assumption 3.2.1. *The malicious agent $\mathcal{A}_i$ has perfect knowledge of the subsystem being attacked, that is $(\tilde{A}_i, \tilde{B}_i, \tilde{C}_i) = (A_i, B_i, C_i)$. Conversely, it has no knowledge of the dynamic interconnections with neighboring subsystems.*

Remark 3.2.2. *The working framework fixed by Assumption 3.2.1 is a worst-case scenario. Indeed, as will be proved in the following, in such a way the residual quantities are not influenced by the attacker. In the case of an attacker with incomplete knowledge of the model, simpler detection strategy could be effectively implemented.*

An attacker willing to satisfy Assumption 3.2.1 needs to obtain the model information via some form of intelligence. This may happen both if the plant structure is known (see [16]), or if the information is leaked. Moreover, one can reasonably assume that an attacker who can write on some channels is able to read from them as well. Therefore, the model might be identified by eavesdropping on the measurement and actuation signals [17].

In the following, we formally prove that model (3.7) ensure the accomplishment of the covertness property. Let k_{ai} be the time instant in which the attacker begins its action, meaning that:

$$\eta_i, \gamma_i = 0, \forall k \in \mathbb{Z}, k < k_{ai}. \tag{3.11}$$

The following proposition states a sufficient condition for an attacker to be covert.

Proposition 3.2.1. *Under 3.2.1, there exists a signal γ_i such that, if A_i is Schur-stable, the attack is asymptotically covert. Furthermore, if the attacker state at k_{ai} is $\tilde{x}_i(k_{ai}) = 0$, the attack is covert $\forall k \in \mathbb{Z}$, $\forall A_i \in \mathbb{R}^{n_i \times n_i}$.*

Proof. Before the attacker starts its manipulation, from (3.11) we trivially have $\tilde{y}_i = y_i, \forall k \in \mathbb{Z}, k < k_{ai}$. Therefore, it is sufficient to prove the condition for $k \geq k_{ai}$.

From (3.1) and (3.6) we have:

$$\begin{aligned} y_i(k) =& C_i A_i^{k-k_{ai}} x_i(k_{ai}) + C_i \sum_{\tau=k_{ai}}^{k-1} A_i^{k-1-\tau} \Bigg[B_i \Big(u_i(\tau) + \eta_i(\tau) \Big) \\ &+ \sum_{j \in \mathcal{N}_i} A_{ij} x_j(\tau) + w_j(\tau) \Bigg] + v_j(k). \end{aligned} \tag{3.12}$$

On the other hand, for a given attacker action η_i, (3.7) gives:

$$\gamma_i(k) = \tilde{C}_i \tilde{A}_i^{k-k_{ai}} \tilde{x}_i(k_{ai}) + \tilde{C}_i \sum_{\tau=k_{ai}}^{k-1} \tilde{A}_i^{k-1-\tau} \tilde{B}_i \eta_i(\tau). \tag{3.13}$$

Under Assumption 3.2.1, and by exploiting again (3.6), the previous expression then give:

$$\begin{aligned} \tilde{y}_i(k) =& C_i A_i^{k-k_{ai}} \Big(x_i(k_{ai}) - \tilde{x}_i(k_{ai}) \Big) + C_i \sum_{\tau=k_{ai}}^{k-1} A_i^{k-1-\tau} \Big[B_i u_i(\tau) \\ &+ \sum_{j \in \mathcal{N}_i} A_{ij} x_j(\tau) + w_j(\tau) \Big] + v_j(k). \end{aligned} \tag{3.14}$$

If A_i is Schur-stable, from (3.14) we have that as $k \to \infty$, $\tilde{y}_i$ will converge to the output of the attack-free subsystem. Put it differently, if $A_i^{k-k_{ai}}$ is vanishing as $k \to \infty$, then the expression in (3.14) will converge to that in (3.12) where it is substituted an identically zero signal η_i. On the other hand, if $\tilde{x}_i(k_{ai}) = 0$, then $\tilde{y}_i$ in (3.14) is indistinguishable from the legitimate output $\forall k \in \mathbb{Z}$, irrespective of A_i. □

Remark 3.2.3. *Observe that Proposition 3.2.1 implicitly proves that the covert attack can be fulfilled with no knowledge of the neighbors or their interconnections. This greatly depends on the linearity of the dynamical system being considered. Namely, thanks to the principle of superposition of effects, the influence of the attacker signal η_i on the subsystem's output y_i does not depend on the other signals simultaneously contributing to generate the subsystem's state trajectory, (the legitimate input u_i, the influence of the neighbors $A_{ij}x_j$, and the process noise w_i). As a consequence, an attacker satisfying Assumption 3.2.1 can successfully design the signal γ_i to compensate for its effect on the output, since the latter is a function of the η_i signal and of the structure of the subsystem (A_i, B_i, C_i) only.*

Finally, it is worth highlighting that both the definition of covert attack and the results of Proposition 3.2.1 can equivalently be restated in terms of detection residuals, as will be discussed later on.

3.3 Detection architecture

This section extensively discusses the proposed detection architecture. Each Logic Unit is endowed with a local controller $\mathcal{C}_i$, a *decentralized* observer $\mathcal{O}_i^d$ (see Subsection 3.3.1), a *distributed* one $\mathcal{O}_i^c$ (Subsection 3.3.2), and a detection logic $\mathcal{D}_i$ (Subsection 3.3.3). A scheme of the Logic Unit is depicted in Figure 3.2.

The purpose of implementing both the observers is the following:

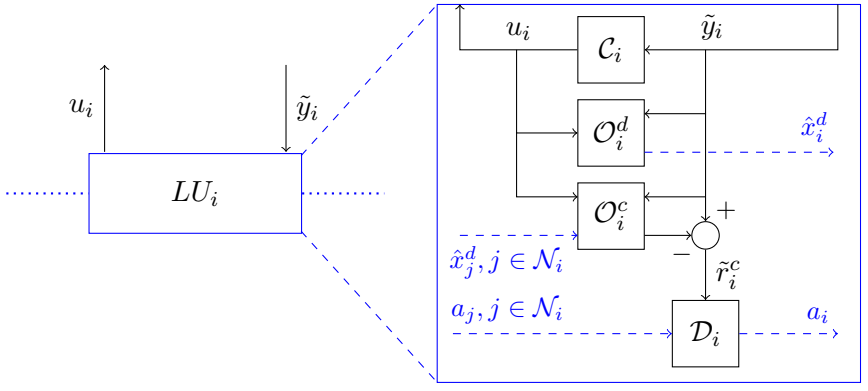


Figure 3.2: Detail of each logic unit LU_i, endowed with a local controller $\mathcal{C}_i$, a decentralized observer $\mathcal{O}_i^d$, a distributed observer $\mathcal{O}_i^c$, and a detector $\mathcal{D}_i$. The scheme distinguishes between signals withing the same logic unit (in black), and signals resulting from the communication with the neighboring subsystems' logic units (in blue).

- The decentralized observer produces a state estimate $\hat{x}_i^d$ based on the input u_i and on the locally measured output $\tilde{y}_i$, decoupled from the neighboring subsystems $\mathcal{S}_j, j \in \mathcal{N}_i$;
- The distributed observer dynamically computes an estimate $\hat{x}_i^c$ from the input u_i, from the locally measured output $\tilde{y}_i$, and from the decentralized estimates of the neighboring subsystems $\hat{x}_j^d, j \in \mathcal{N}_i$ (conveniently *communicated* from their logic unit LU_j).

In such a way, if i is the index of attacked subsystem $\mathcal{S}_i$, the distributed observers $\mathcal{O}_j^c$ of its neighboring subsystems $\mathcal{S}_j, j \in \mathcal{N}_i$ can detect possible inconsistencies between the true state x_i of $\mathcal{S}_i$ (to which each neighboring subsystem's state x_j is directly coupled) and the possibly wrong[1] estimate $\hat{x}_i^d$ which the logic unit LU_i produced from the attacked measurements $\tilde{y}_i$. Therefore, the attack is perfectly covert with respect to $\mathcal{S}_i$, its neighbors can reveal it.

In order to detect the aforementioned inconsistencies, each subsystem considers a residual signal $\tilde{r}_i^c$ and a suitable time-varying threshold $\bar{r}_i^c$. Both these quantities will be formally defined and discussed in detail in the following. In order to reveal stealthy attacks, the following *distributed* detection logic is implemented by a diagnoser $\mathcal{D}_i$:

- If $\|\tilde{r}_i^c\| > \bar{r}_i^c$, a binary alarm signal a_i is raised;
- Each logic unit LU_i broadcasts the alarm signal a_i to all the neighboring logic units $LU_j, j \in \mathcal{N}_i$, and receives the sequence $a_j, j \in \mathcal{N}_i$;

[1]Observe that the communication between logic units is assumed invulnerable. Therefore, the estimates $\hat{x}_i^d$ is correctly communicated, and it could only be wrong because it relies on attacked measurements. More details on this fact are given in Subsection 3.3.2.

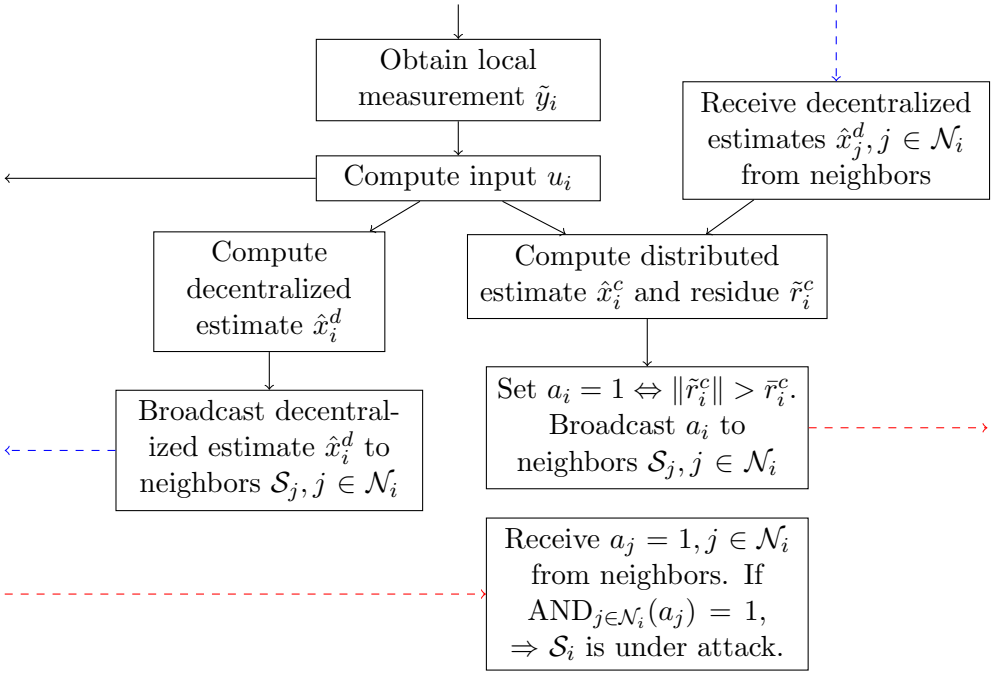


Figure 3.3: Detection algorithm performed at each sampling time in each LU.

- If $a_j = 1, \forall j \in \mathcal{N}_i$, then detector $\mathcal{D}_i$ decides $\mathcal{S}_i$ is under attack.

The detection logic is summarized in Figure 3.3.

Before proceeding with the analysis, let specify some quantity of interest. The state estimation errors for the decentralized and the distributed estimates, respectively, are defined as follow:

$$\begin{aligned} \epsilon_i^d &\doteq x_i - \hat{x}_i^d \\ \epsilon_i^c &\doteq x_i - \hat{x}_i^c. \end{aligned} \tag{3.15}$$

Observe that ϵ_i^d and ϵ_i^c represent the difference between the actual state of $\mathcal{S}_i$ and the state estimates of the corresponding observers. These quantities are named the *true* errors, and they cannot be computed in practice since the actual state is not directly accessible. Therefore, the output estimation errors are coherently defined as:

$$\begin{aligned} r_i^d &\doteq y_i - C_i\hat{x}_i^d = C_i\epsilon_i^d + v_i \\ r_i^c &\doteq y_i - C_i\hat{x}_i^c = C_i\epsilon_i^c + v_i. \end{aligned} \tag{3.16}$$

On the other hand, if subsystem $\mathcal{S}_i$ is under attack, these quantities are not available either. More conveniently, in this case, one can consider the *computed* state errors:

$$\begin{aligned} \tilde{\epsilon}_i^d &\doteq x_i - \tilde{x}_i - \hat{x}_i^d \\ \tilde{\epsilon}_i^c &\doteq x_i - \tilde{x}_i - \hat{x}_i^c, \end{aligned} \tag{3.17}$$

and, consistently, the computed output errors (the so-called *residuals*):

$$\begin{aligned} \tilde{r}_i^d &\doteq \tilde{y}_i - C_i\hat{x}_i^d = C_i\tilde{\epsilon}_i^d + v_i \\ \tilde{r}_i^c &\doteq \tilde{y}_i - C_i\hat{x}_i^c = C_i\tilde{\epsilon}_i^c + v_i. \end{aligned} \tag{3.18}$$

Similar to the conventions previously introduced, the latter quantities are called the received or attacked state and output errors. Observe that, when no attack is happening, the true and the computed errors trivially coincide.

In the next sections the technicalities of the detection strategy are outlined in detail. Subsection 3.3.1 and 3.3.2 are dedicated to the design and the analysis of the decentralized and distributed observers, respectively. Finally, in Subsection 3.3.3, the details of the attack detection logic are given.

3.3.1 Decentralized Observer $\mathcal{O}_i^d$

This section tackles the problem of designing a decentralized estimate $\hat{x}_i^d$ of the state x_i, not depending on the contribution from the neighboring subsystems. In this respect, an Unknown Input Observer (UIO) is implemented (see [18], here adapted to fit the discrete-time scenario), by regarding the influence of the neighbors as an unknown input acting on $\mathcal{S}_i$. This implementation of a UIO is derived from the distributed detection of anomalies literature (for instance, see [19] and [20]).

The implementation of the UIO is the following. Named $z_i \in \mathbb{R}^{n_i}$ the state of the observer, the decentralized estimate $\hat{x}_i^d$ is the output of the following dynamical system:

$$\mathcal{O}_i^d : \begin{cases} z_i^+ = F_i^d z_i + T_i B_i u_i + K_i \tilde{y}_i \\ \hat{x}_i^d = z_i + H_i \tilde{y}_i. \end{cases} \tag{3.19}$$

F_i^d, T_i, K_i, and H_i are gains with compatible dimensions, which will shortly be better characterized. Let observe that by defining $\boldsymbol{x}_i$ as:

$$\boldsymbol{x}_i^\top \doteq \underset{j\in\mathcal{N}_i}{\text{row}}(x_j^\top) \tag{3.20}$$

for each subsystem $\mathcal{S}_i$, and by exploiting (3.2), the influence of the neighbors on $\mathcal{S}_i$ on the right-hand side of (3.1) can be rewritten as follows:

$$\sum_{j\in\mathcal{N}_i} A_{ij}x_j = \Xi_i \boldsymbol{x}_i. \tag{3.21}$$

Having said that, it is required that [18, Theorem 1]:

a) The output matrix is such that:

$$\text{rank}(C_i\Xi_i) = \text{rank}(\Xi_i). \tag{3.22}$$

b) The pair $(\bar{A}_i, C_i)$ is detectable, where:

$$\bar{A}_i \doteq A_i - H_i C_i A_i. \tag{3.23}$$

Observe that condition (3.22) is ensured by Assumption 3.1.1. Under such conditions, by decomposing K_i as

$$K_i = K_i^{(1)} + K_i^{(2)}, \tag{3.24}$$

the gains appearing in (3.19) can be designed so that:

$$\begin{aligned} 0 =& (H_i C_i - I)\Xi_i && \text{(3.25a)} \\ T_i =& I - H_i C_i && \text{(3.25b)} \\ F_i^d =& \bar{A}_i - K_i^{(1)} C_i \text{ is Schur-stable} && \text{(3.25c)} \\ K_i^{(2)} =& F_i H_i. && \text{(3.25d)} \end{aligned}$$

Note that a particular solution to (3.25a) is:

$$H_i = \Xi_i [(C_i \Xi_i)^\top C_i \Xi_i]^{-1} (C_i \Xi_i). \tag{3.26}$$

The matrix F_i^d rules the rate of convergence to steady-state of the decentralized estimation error ϵ_i^d. Note that the ϵ_i^d approaches zero at steady-state only in the disturbance-free and attack-free scenario. More in general, the next proposition characterizes the decentralized estimation error trajectory for the attacked subsystem $\mathcal{S}_i$.

Proposition 3.3.1. *Let assume a malicious agent $\mathcal{A}_i$ modeled as in* (3.7) *is manipulating the ith subsystem $\mathcal{S}_i$ as in* (3.6)*. If Assumption 3.2.1 and the UIO condition* (3.25) *hold, then the decentralized true error dynamics of observer* (3.19) *is ruled by:*

$$\epsilon_i^{d+} = F_i^d \epsilon_i^d + T_i w_i - K_i^{(1)} v_i - H_i v_i^+ + (A_i - F_i^d)\tilde{x}_i + B_i \eta_i. \tag{3.27}$$

On the other hand, the decentralized computed error evolves according to:

$$\tilde{\epsilon}_i^{d+} = F_i^d \tilde{\epsilon}_i^d + T_i w_i - K_i^{(1)} v_i - H_i v_i^+, \tag{3.28}$$

and the attack is covert for $\mathcal{O}_i^d$.

Proof. Before proceeding with the computation, one might find it useful to remind that the actual subsystem $\mathcal{S}_i$ is driven by the attacked control input $\tilde{u}_i = u_i + \eta_i$, whereas the decentralized observer $\mathcal{O}_i^d$ is fed with the legitimate control input u_i and the attacked measurement signal $\tilde{y}_i = y_i - \gamma_i = C_i(x_i - \tilde{x}_i) + v_i$.

From definition (3.15), and models (3.1), (3.7), and (3.19), we have:

$$\begin{aligned}
\epsilon_i^{d+} =& x_i^+ - \hat{x}_i^{d+} \\
=& x_i^+ - z_i^+ - H_i[C_i(x_i^+ - \tilde{x}_i^+) + v_i^+] \\
=& (I - H_iC_i)x_i^+ - z_i^+ - H_iv_i^+ + H_iC_i\tilde{x}_i^+ \\
=& (I - H_iC_i)[A_ix_i + B_i(u_i + \eta_i) + \Xi_i\boldsymbol{x}_i + w_i] - [F_i^d z_i + T_iB_iu_i \\
&+ K_iC_i(x_i - \tilde{x}_i) + K_iv_i] - H_iv_i^+ + H_iC_i(A_i\tilde{x}_i + B_i\eta_i) \\
=& [(I - H_iC_i)A_i - K_iC_i]x_i + [(I - H_iC_i) - T_i]B_iu_i \\
&+ (I - H_iC_i)\Xi_i\boldsymbol{x}_i - F_i^d z_i + (I - H_iC_i)w_i - K_iv_i \\
&- H_iv_i^+ + (H_iC_iA_i + K_iC_i)\tilde{x}_i + B_i\eta_i.
\end{aligned} \tag{3.29}$$

Then, by substituting (3.23), (3.25a), and (3.25b), (3.29) can be rewritten as:

$$\begin{aligned}
\epsilon_i^{d+} =& [\bar{A}_i - K_iC_i]x_i - F_i^d z_i + T_iw_i - K_iv_i \\
&- H_iv_i^+ + (A_i - \bar{A}_i + K_iC_i)\tilde{x}_i + B_i\eta_i \\
=& [\bar{A}_i - K_i^{(1)}C_i - K_i^{(2)}C_i]x_i - F_i^d z_i + T_iw_i - K_i^{(1)}v_i - K_i^{(2)}v_i \\
&- H_iv_i^+ + (A_i - \bar{A}_i + K_i^{(1)}C_i + K_i^{(2)}C_i)\tilde{x}_i + B_i\eta_i
\end{aligned} \tag{3.30}$$

where the decomposition (3.24) was exploited. Finally, from (3.25c) and (3.25d), we have:

$$\begin{aligned}
\epsilon_i^{d+} =& [F_i^d - F_i^dH_iC_i]x_i - F_i^d z_i + T_iw_i - K_i^{(1)}v_i - F_i^dH_iv_i \\
&- H_iv_i^+ + (A_i - F_i^d + F_i^dH_iC_i)\tilde{x}_i + B_i\eta_i \\
=& F_i^d[x_i - z_i - H_i\tilde{y}_i] + T_iw_i - K_i^{(1)}v_i - H_iv_i^+ \\
&+ (A_i - F_i^d)\tilde{x}_i + B_i\eta_i \\
=& F_i^d\epsilon_i^d + T_iw_i - K_i^{(1)}v_i - H_iv_i^+ + (A_i - F_i^d)\tilde{x}_i + B_i\eta_i.
\end{aligned} \tag{3.31}$$

Moreover, from (3.27), (3.7), and (3.17), if Assumption 3.2.1 holds, one deduces:

$$\begin{aligned}
\tilde{\epsilon}_i^{d+} =& \epsilon_i^{d+} - \tilde{x}_i^+ \\
=& F_i^d\epsilon_i^d + T_iw_i - K_i^{(1)}v_i - H_iv_i^+ + (A_i - F_i^d)\tilde{x}_i \\
&+ B_i\eta_i - (A_i\tilde{x}_i + B_i\eta_i) \\
=& F_i^d(\epsilon_i^d - \tilde{x}_i) + T_iw_i - K_i^{(1)}v_i - H_iv_i^+ \\
=& F_i^d\tilde{\epsilon}_i^d + T_iw_i - K_i^{(1)}v_i - H_iv_i^+.
\end{aligned} \tag{3.32}$$

Since the decentralized computed error $\tilde{\epsilon}_i^d$ evolves independently of $\tilde{x}_i$ and η_i, that is the decentralized observer is insensitive to the malicious agent action, the attack is necessarily covert. □

Remark 3.3.1. *Note that by substituting $\eta_i, \gamma_i = 0, \forall k \in \mathbb{Z}$ in (3.27), we derive the decentralized true error dynamics in the attack-free scenario:*

$$\epsilon_i^{d+} = F_i^d \epsilon_i^d + T_i w_i - K_i^{(1)} v_i - H_i v_i^+. \tag{3.33}$$

As a consequence, we deduce that in the attack-free scenario the decentralized true error ϵ_i^d and the decentralized computed error $\tilde{\epsilon}_i^d$ coincide, coherently with their definition.

3.3.2 Distributed Observer $\mathcal{O}_i^c$

This section is dedicated to the design of the distributed observer $\mathcal{O}_i^c$. As previously mentioned, this observer's intent is to produce an estimate $\hat{x}_i^c$ of the state x_i from the legitimate input u_i, the (possibly attacked) measured output $\tilde{y}_i$, and the decentralized estimates of the neighboring subsystems $\hat{x}_j^d, j \in \mathcal{N}_i$. As it will be proved, the resulting distributed computed error $\tilde{\epsilon}_i^c$ (and, consequently, the residual quantity $\tilde{r}_i^c$) is insensitive to possible attacks at $\mathcal{S}_i$. Nonetheless, $\tilde{\epsilon}_i^c$ is sensitive to possible attacks in the neighboring subsystems, and this fact is the core of the proposed detection strategy.

Note that the considered attacks presented in Section 3.2 manipulate the actuation and the measurement signals as in (3.6). On the other hand, the communication channels between logic units are assumed safe, as specified in the following assumption.

Assumption 3.3.1. *The communication between logic units is ideal. As a consequence, the exchanged estimates $\hat{x}_j^d, j \in \mathcal{N}_i$ are not corrupted during the communication.*

Based on the subsystem dynamical equations (3.1), the distributed observer $\mathcal{O}_i^c$ is designed as a standard Luenberger observer, where the contribution due to the physical coupling with the neighboring subsystems is dealt as a known input derived from the communicated decentralized estimates $\hat{x}_j^d, j \in \mathcal{N}_i$. Precisely, the dynamics of such an observer is :

$$\mathcal{O}_i^c : \ \hat{x}_i^{c+} = A_i \hat{x}_i^c + B_i u_i + \sum_{j \in \mathcal{N}_i} A_{ij} \hat{x}_j^d + L_i(\tilde{y}_i - C_i \hat{x}_i^c). \tag{3.34}$$

$L_i \in \mathbb{R}^{n_i \times p_i}$ is the observer gain and could be designed in order to place some given eigenvalues of the matrix:

$$F_i^c \doteq A_i - L_i C_i \tag{3.35}$$

so that, at least in the ideal scenario of absence of noise and of attacks in the neighboring subsystems, the estimation error ϵ_i^c asymptotically converges to zero satisfying some given performance.

Alternatively, as the authors suggest in [1], an H_∞ approach (see [21]) can be employed to design the observer gain L_i in order to attenuate the effect of

the noises w_i and v_i, and of the decentralized true errors of the neighboring subsystems $\epsilon_j^d, j \in \mathcal{N}_i$ on the observer error ϵ_i^c.

Remark 3.3.2. *Observe that $\hat{x}_j^d$ is not affected by attacks in the neighboring subsystems $\mathcal{S}_h, h \in \mathcal{N}_j$. This is the reason why in* (3.34) *$A_{ij}\hat{x}_j^d$ was used instead of $A_{ij}\hat{x}_j^c$. This property will lay the basis for our detection strategy in the next section*

The next proposition analyzes the properties of the designed distributed observer $\mathcal{O}_i^c$.

Proposition 3.3.2. *Let consider a malicious agent, modeled as in* (3.7)*, affecting subsystem $\mathcal{S}_i$ as in* (3.6)*. If Assumption 3.2.1 holds, the true estimation error dynamics for observer* (3.34) *is:*

$$\epsilon_i^{c+} = F_i^c \epsilon_i^c + w_i - L_i v_i + \sum_{j \in \mathcal{N}_i} A_{ij}\epsilon_i^d + B_i\eta_i + L_i\gamma_i, \tag{3.36}$$

whereas for the computed estimation error is:

$$\tilde{\epsilon}_i^{c+} = F_i^c \tilde{\epsilon}_i^c + w_i - L_i v_i + \sum_{j \in \mathcal{N}_i} A_{ij}\epsilon_i^d, \tag{3.37}$$

and the attack is covert for $\mathcal{O}_i^c$.

Proof. Here again it is worth highlighting that the actual subsystem $\mathcal{S}_i$ is driven by the attacked control input $\tilde{u}_i = u_i + \eta_i$, whereas the decentralized observer $\mathcal{O}_i^d$ is fed with the legitimate control input u_i and the attacked measurement signal $\tilde{y}_i = y_i - \gamma_i = C_i x_i + v_i - \gamma_i$.

Therefore, from (3.15), (3.1), (3.34), and (3.35), a trivial computation gives:

$$\begin{aligned}
\epsilon_i^{c+} =& x_i^+ - \hat{x}_i^{c+} \\
=& A_i x_i + B_i(u_i + \eta_i) + \sum_{j \in \mathcal{N}_i} A_{ij} x_j + w_i \\
& - \left[A_i \hat{x}_i^c + B_i u_i + \sum_{j \in \mathcal{N}_i} A_{ij}\hat{x}_j^d + L_i(C_i x_i + v_i - \gamma_i - C_i \hat{x}_i^c) \right] \\
=& (A_i - L_i C_i)(x_i - \hat{x}_i^c) + w_i - L_i v_i + \sum_{j \in \mathcal{N}_i} A_{ij}(x_j - \hat{x}_j^d) \\
& + B_i\eta_i + L_i\gamma_i \\
=& F_i^c \epsilon_i^c + w_i - L_i v_i + \sum_{j \in \mathcal{N}_i} A_{ij}\epsilon_j^d + B_i\eta_i + L_i\gamma_i.
\end{aligned} \tag{3.38}$$

On the other hand, from (3.17), (3.7), and (3.36), if Assumption 3.2.1 holds, one

obtains:

$$\begin{aligned}\tilde{\epsilon}_i^{c+} =& \epsilon_i^{c+} - \tilde{x}_i^+ \\ =& F_i^c \epsilon_i^c + w_i - L_i v_i + \sum_{j \in \mathcal{N}_i} A_{ij} \epsilon_j^d + B_i \eta_i + L_i C_i \tilde{x}_i - (A_i \tilde{x}_i + B_i \eta_i) \\ =& F_i^c (\epsilon_i^c - \tilde{x}_i) + w_i - L_i v_i + \sum_{j \in \mathcal{N}_i} A_{ij} \epsilon_j^d \\ =& F_i^c \tilde{\epsilon}_i^c + w_i - L_i v_i + \sum_{j \in \mathcal{N}_i} A_{ij} \epsilon_j^d.\end{aligned} \tag{3.39}$$

Since the computed error follows the true error dynamics of the attack-free scenario, which can be deduced by substituting $\eta_i, \gamma_i = 0, \forall k \in \mathbb{Z}$ in (3.36), the attack is covert for $\mathcal{O}_i^c$. □

3.3.3 Attack detection scheme

In this subsection, the final details on the detection architecture are given. As Proposition 3.3.2 states, the distributed computed error $\tilde{\epsilon}_i^c$ and, consequently, the distributed computed residue $\tilde{r}_i^c$ are sensitive to the decentralized true errors $\epsilon_j^d, j \in \mathcal{N}_i$ of their neighbors. Therefore, in principle the alarm signal a_i might be raised whenever $\tilde{r}_i^c$ is different from zero. Nonetheless, this procedure would result in frequent false-alarms due to the presence of noise. For this reason, in the following, a suitable threshold is designed in order to take the disturbances action into account. Such a threshold results from the bounds (3.5) and accounts for the maximum noise contribution on $\tilde{r}_i^c$ in attack-free conditions.

In the following analysis, the next condition is assumed to be satisfied.

Assumption 3.3.2. *For any subsystem $\mathcal{S}_i$, there is only one attacker in its neighborhood $\mathcal{N}_i$.*

This assumption aims to rule out complex situations in which multiple attacks in different subsystems within the same neighborhood might be designed in order to compensate one another. Such scenario will be explored in Chapter 5. Nonetheless, this assumption is not unreasonable. Indeed, if the overall system is spread over a large area, it might be difficult for an attacker to target vast sections of it.

In order to derive a threshold on the distributed error, it is firstly needed to consider its decentralized counterparts. The next proposition derives an upper bound on the norm of this quantity.

Proposition 3.3.3. *Given* (3.5)*, and a bound on the decentralized true error at time $k = 0$, $\bar{x}_i(0)$, then, in attack-free conditions, the norm of the decentralized observer error $\left\|\epsilon_i^d\right\|$ is bounded by the positive function $\bar{\epsilon}_i^d$, which can be initialized as*

$$\bar{\epsilon}_i^d(0) = \bar{x}_i(0) + \|H_i\| \bar{v}_i, \tag{3.40}$$

and evolves according to:

$$\bar{\epsilon}_i^{d+} = \left\| F_i^d \right\| \bar{\epsilon}_i^d + \left(1 - \left\| F_i^d \right\| \right) \| H_i \| \, \bar{v}_i + Q_i^d, \tag{3.41}$$

where:

$$Q_i^d \doteq \| T_i \| \, \bar{w}_i + \| K_i \| \, \bar{v}_i. \tag{3.42}$$

Proof. By convolving the dynamical equation of the decentralized true error in attack-free conditions (3.33), we obtain:

$$\begin{aligned} \epsilon_i^d(k) =& (F_i^d)^k \epsilon_i^d(0) + \sum_{\tau=0}^{k-1} (F_i^d)^{k-1-\tau} \Big(T_i w_i(\tau) - K_i^{(1)} v_i(\tau) \Big) \\ & - \sum_{\tau=1}^{k} (F_i^d)^{k-\tau} H_i v_i(\tau). \end{aligned} \tag{3.43}$$

The last summation can be rearranged as follows:

$$\begin{aligned} \sum_{\tau=1}^{k} (F_i^d)^{k-\tau} H_i v_i(\tau) =& \sum_{\tau=0}^{k-1} (F_i^d)^{k-\tau} H_i v_i(\tau) + H_i v_i(k) - (F_i^d)^k H_i v_i(0) \\ =& \sum_{\tau=0}^{k-1} (F_i^d)^{k-1-\tau} F_i^d H_i v_i(\tau) + H_i v_i(k) - (F_i^d)^k H_i v_i(0) \\ =& \sum_{\tau=0}^{k-1} (F_i^d)^{k-1-\tau} K_i^{(2)} v_i(\tau) + H_i v_i(k) - (F_i^d)^k H_i v_i(0), \end{aligned} \tag{3.44}$$

where (3.25d) was applied. Therefore, by recalling the decomposition (3.24), the decentralized true error can be rewritten as:

$$\begin{aligned} \epsilon_i^d(k) =& (F_i^d)^k \Big(\epsilon_i^d(0) - H_i v_i(0) \Big) + H_i v_i(k) \\ & + \sum_{\tau=0}^{k-1} (F_i^d)^{k-1-\tau} \Big(T_i w_i(\tau) - K_i v_i(\tau) \Big). \end{aligned} \tag{3.45}$$

In order to derive a suitable threshold, let recall that for the induced euclidean matrix 2-norm, given any vector x, and any two matrices A and B of compatible dimensions, the following inequalities hold [22, Subsection 10.4.2]:

$$\| Ax \| \leq \| A \| \, \| x \| \tag{3.46a}$$

$$\| AB \| \leq \| A \| \, \| B \| . \tag{3.46b}$$

Therefore, by applying the triangle inequality multiple times to (3.45) we find:

$$\begin{aligned}\left\|\epsilon_i^d(k)\right\| \leq \left\|(F_i^d)^k\right\| \left(\left\|\epsilon_i^d(0)\right\| + \|H_i\| \, \|v_i(0)\| \right) + \|H_i\| \, \|v_i(k)\| \\ + \sum_{\tau=0}^{k-1} \left\|(F_i^d)^{k-1-\tau}\right\| \left(\|T_i\| \, \|w_i(\tau)\| + \|K_i\| \, \|v_i(\tau)\| \right).\end{aligned} \tag{3.47}$$

By recalling the fact that the noises are uniformly bounded (3.5), we obtain:

$$\begin{aligned}\left\|\epsilon_i^d(k)\right\| \leq \left\|(F_i^d)^k\right\| \left(\left\|\epsilon_i^d(0)\right\| + \|H_i\| \, \bar{v}_i \right) + \|H_i\| \, \bar{v}_i \\ + \sum_{\tau=0}^{k-1} \left\|(F_i^d)^{k-1-\tau}\right\| \left(\|T_i\| \, \bar{w}_i + \|K_i\| \, \bar{v}_i \right).\end{aligned} \tag{3.48}$$

Moreover, (3.46b) in particular gives, $\forall k \in \mathbb{Z}, k \geq 0$:

$$\left\|(F_i^d)^k\right\| \leq \left\|F_i^d\right\|^k. \tag{3.49}$$

From such property, and by defining Q_i^d as in (3.42), we find:

$$\begin{aligned}\left\|\epsilon_i^d(k)\right\| \leq \left\|F_i^d\right\|^k \left(\left\|\epsilon_i^d(0)\right\| + \|H_i\| \, \bar{v}_i \right) + \|H_i\| \, \bar{v}_i \\ + Q_i^d \sum_{\tau=0}^{k-1} \left\|F_i^d\right\|^{k-1-\tau}.\end{aligned} \tag{3.50}$$

Let $\bar{\epsilon}_i^d$ be the expression on right-hand side of the previous inequality. From simple manipulations, we have:

$$\begin{aligned}\bar{\epsilon}_i^d(k+1) &= \left\|F_i^d\right\|^{k+1} \left(\left\|\epsilon_i^d(0)\right\| + \|H_i\| \, \bar{v}_i \right) + \|H_i\| \, \bar{v}_i \\ &\quad + Q_i^d \sum_{\tau=0}^{k} \left\|F_i^d\right\|^{k-\tau} \\ &= \left\|F_i^d\right\| \left\|F_i^d\right\|^k \left(\left\|\epsilon_i^d(0)\right\| + \|H_i\| \, \bar{v}_i \right) + \|H_i\| \, \bar{v}_i \\ &\quad + Q_i^d + Q_i^d \left\|F_i^d\right\| \sum_{\tau=0}^{k-1} \left\|F_i^d\right\|^{k-1-\tau} \\ &= \left\|F_i^d\right\| \bar{\epsilon}_i^d(k) + \left(1 - \left\|F_i^d\right\|\right) \|H_i\| \, \bar{v}_i + Q_i^d.\end{aligned} \tag{3.51}$$

Finally, one can observe by inspection that the initialization (3.40) is coherent with the computation. □

In the next proposition, a similar analysis is proposed for the distributed error.

Proposition 3.3.4. *Given* (3.5)*, and a bound on the distributed true error at time* $k = 0$*,* $\bar{x}_i(0)$*, then, in attack-free conditions, the norm of the distributed observer error* $\|\epsilon_i^c\|$ *is bounded by the positive function* $\bar{\epsilon}_i^c$*, which can be initialized as* $\bar{\epsilon}_i^c(0) = \bar{x}_i(0)$*, and evolves according to:*

$$\bar{\epsilon}_i^{c+} = \|F_i^c\|\, \bar{\epsilon}_i^c + Q_i^c + \sum_{j \in \mathcal{N}_i} \|A_{ij}\| \left\|\epsilon_j^d\right\|, \tag{3.52}$$

where:

$$Q_i^c \doteq \bar{w}_i + \|L_i\|\, \bar{v}_i. \tag{3.53}$$

Proof. By convolving the dynamical equation of the distributed true error in attack-free conditions, which can be deduced by substituting $\eta_i, \gamma_i = 0, \forall k \in \mathbb{Z}$ in (3.36), one finds:

$$\epsilon_i^c(k) = (F_i^c)^k \epsilon_i^c(0) + \sum_{\tau=0}^{k-1} (F_i^c)^{k-1-\tau} \left(w_i(\tau) - L_i v_i(\tau) + \sum_{j \in \mathcal{N}_i} A_{ij} \epsilon_j^d(\tau) \right). \tag{3.54}$$

Therefore, by exploiting the triangle inequality, together with the properties (3.46a) and (3.46b), we obtain:

$$\begin{aligned} \|\epsilon_i^c(k)\| \leq \left\|(F_i^c)^k\right\| \|\epsilon_i^c(0)\| + \sum_{\tau=0}^{k-1} \left\|(F_i^c)^{k-1-\tau}\right\| \Bigg(\|w_i(\tau)\| \\ + \|L_i\|\, \|v_i(\tau)\| + \sum_{j \in \mathcal{N}_i} \|A_{ij}\| \left\|\epsilon_j^d(\tau)\right\| \Bigg). \end{aligned} \tag{3.55}$$

Moreover, from the fact that the noises are uniformly bounded (3.5), by defining Q_i^c as in (3.53), taken (3.49) into account, we obtain:

$$\|\epsilon_i^c(k)\| \leq \|F_i^c\|^k \|\epsilon_i^c(0)\| + \sum_{\tau=0}^{k-1} \|F_i^c\|^{k-1-\tau} \left(Q_i^c + \sum_{j \in \mathcal{N}_i} \|A_{ij}\| \left\|\epsilon_j^d(\tau)\right\| \right). \tag{3.56}$$

Let $\bar{\epsilon}_i^c$ be the expression on right-hand side of the previous inequality. Then,

from simple manipulations, we have:

$$\begin{aligned}\bar{\epsilon}_i^c(k+1) &= \|F_i^c\|^{k+1}\,\|\epsilon_i^c(0)\| + \sum_{\tau=0}^{k}\|F_i^c\|^{k-\tau}\left(Q_i^c + \sum_{j\in\mathcal{N}_i}\|A_{ij}\|\left\|\epsilon_j^d(\tau)\right\|\right)\\ &= \|F_i^c\|\,\|F_i^c\|^{k}\,\|\epsilon_i^c(0)\| + Q_i^c + \sum_{j\in\mathcal{N}_i}\|A_{ij}\|\left\|\epsilon_j^d(k)\right\|\\ &\quad + \|F_i^c\|\sum_{\tau=0}^{k-1}\|F_i^c\|^{k-1-\tau}\left(Q_i^c + \sum_{j\in\mathcal{N}_i}\|A_{ij}\|\left\|\epsilon_j^d(\tau)\right\|\right)\\ &= \|F_i^c\|\,\bar{\epsilon}_i^c(k) + Q_i^c + \sum_{j\in\mathcal{N}_i}\|A_{ij}\|\left\|\epsilon_j^d(k)\right\|.\end{aligned} \tag{3.57}$$

□

Observe that in order to obtain a bound on $\left\|\epsilon_j^d\right\|, j \in \mathcal{N}_i$, one can exploit (3.41). Moreover, a simple computation reveals that, in attack-free conditions, the norm of the distributed residue $\|\tilde{r}_i^c\|$ is bounded by the positive function $\bar{r}_i^c$, defined as:

$$\bar{r}_i^c \doteq \|C_i\|\,\bar{\epsilon}_i^c + \bar{v}_i. \tag{3.58}$$

Remark 3.3.3. *All the computed thresholds are time-varying quantities. Nonetheless, if the subsystems are stable, after a sufficiently long time the threshold converges to a steady-state value (when the transient due to the uncertainty on the initial condition vanishes).*

Finally, Theorem 3.3.1 asserts sufficient conditions for an attack in order to be detected.

Theorem 3.3.1 (Detectability)**.** *A covert cyber-attack starting at instant $k_{a,i}$ in $\mathcal{S}_i, i \in \mathcal{N}_j$, is detectable by $\mathcal{S}_j$ if $\exists \bar{k}_i > k_{a,i}$ such that:*

$$\left\|\sum_{\tau=k_{a,i}}^{\bar{k}_i-1}(F_j^c)^{k-1-\tau}A_{ji}\sum_{t=k_{a,i}}^{\tau-1}(F_i^d)^{\tau-1-t}\theta_i(t)\right\| > 2\bar{r}_j^c, \tag{3.59}$$

where:

$$\theta_i(k) \doteq (A_i - F_i^d)\tilde{x}_i(k) + B_i\eta_i(k). \tag{3.60}$$

Proof. To consider the attack effect, one needs to convolve (3.33) and (3.27) before and after $k_{a,i}$, respectively, obtaining (see (3.45)):

$$\begin{aligned}\epsilon_i^d(k) =&(F_i^d)^k\left(\epsilon_i^d(0) - H_iv_i(0)\right) + H_iv_i(k)\\ &+\sum_{\tau=0}^{k-1}(F_i^d)^{k-1-\tau}\left(T_iw_i(\tau) - K_iv_i(\tau)\right) + \sum_{\tau=k_{a,i}}^{k-1}(F_i^d)^{k-1-\tau}\theta_i(\tau).\end{aligned} \tag{3.61}$$

The first three terms consist in the attack-free error, which corresponds to the computed error $\tilde{\epsilon}_i^d$ as observed in Remark 3.3.1. The final summation is associated to the effect of the attack. Therefore, we can conveniently rewrite (3.61) as:

$$\epsilon_i^d(k) = \tilde{\epsilon}_i^d(k) + \sum_{\tau=k_{a,i}}^{k-1} (F_i^d)^{k-1-\tau}\theta_i(\tau). \tag{3.62}$$

On the other hand, by convolving (3.37) for subsystem $\mathcal{S}_j, j \in \mathcal{N}_i$, we obtain:

$$\begin{aligned}\tilde{\epsilon}_j^c(k) =&(F_j^c)^k\tilde{\epsilon}_j^c(0) + \sum_{\tau=0}^{k-1}(F_j^c)^{k-1-\tau}\left(w_j(\tau) - L_j v_j(\tau) + \sum_{l\in\mathcal{N}_j} A_{jl}\tilde{\epsilon}_l^d(\tau)\right)\\ &+ \sum_{\tau=0}^{k-1}(F_j^c)^{k-1-\tau}A_{ji}\sum_{t=k_{a,i}}^{\tau-1}(F_i^d)^{\tau-1-t}\theta_i(t),\end{aligned} \tag{3.63}$$

where again we can distinguish between the attack-free received error $\tilde{\epsilon}_{j,af}^c$, accounting for all the terms but the last double summation, and the attack contribution $\tilde{\epsilon}_{j,att}^c$, the last double summation itself.

By applying the inverse triangle inequality and the bound (3.58), we find:

$$\bar{r}_j \geq \left\|C_j(\tilde{\epsilon}_{j,att}^c + \tilde{\epsilon}_{j,af}^c)\right\| + \bar{v}_j \geq \left|\left\|C_j\tilde{\epsilon}_{j,att}^c\right\| - \left\|C_j\tilde{\epsilon}_{j,af}^c\right\|\right| + \bar{v}_j \tag{3.64}$$

and:

$$\left\|C_j\tilde{\epsilon}_{j,att}^c\right\| \leq \left\|C_j\tilde{\epsilon}_{j,af}^c\right\| + \bar{v}_j + \bar{r}_j \leq 2\bar{r}_j, \tag{3.65}$$

which holds $\forall k \in \mathbb{Z}, k > 0$. By negating this condition, we find (3.59). □

Remark 3.3.4. *Observe that an attack in $\mathcal{S}_i$ could not be detected by $\mathcal{S}_j$ because either the attacker action lies within the null space of the interconnection matrix A_{ji}, or because the attack "amplitude" is too low, that is the action is indistinguishable from the noise effect.*

Finally, observe that one may derive component-wise bounds to be used for the detection strategy, as shown in [1, Subsection IV-C]. This is particularly useful when the state components are not normalized, i.e. their magnitudes are on different scales. Indeed, in such a scenario, the norm-based thresholds derived in this subsection might be quite conservative.

Chapter 4

Isolation strategies

This chapter addresses the problem of isolation. This issue has been deeply discussed in the context of fault detection, see [12] and [13]. Indeed, once a fault or an attack has been detected, it might not be possible for the detection architecture itself to uniquely locate its source. Identifying specifically which subsystem is faulted or under attack is extremely important for practical purposes, for instance in order to employ an accommodation strategy such as [23].

At first, one might assume that the detection algorithm illustrated in Chapter 3 implicitly solves the isolation problem, since the single subsystems directly decide they are under attack. However, this argument is incorrect. Indeed, as it will be shown, depending on the topology of the network, there might be multiple subsystems simultaneously claiming to be under attack. The reason for this lies in the fact that the subsystems whose distributed residue $\tilde{r}^c$ violates the associated threshold have no information on which of their neighbours is under attack, and the only thing they can do is broadcast the alarm signal to all of them, whether they are the actually attacked subsystem or not. Therefore, there might be some layouts such that a subsystem receives a raised-up alarm signal from all of its neighbors, and consequently deduces it is under attack, even though this is not the case. In these circumstances, further solutions must be considered to overcome the issue, relying on the topology of the network and on structural properties of the interconnection matrices.

The organization or the chapter is the following. In Section 4.1, the topological configurations resulting in isolation ambiguities are formalized. Section 4.2 is dedicated to the presentation of three isolation strategies. Finally, Section 4.3, proposes a comparison on the condition required to implement each solution.

4.1 Problem formulation

All the techniques of this chapter are meant to isolate attacks detected by using the algorithm of Chapter 3. As a consequence, all the following results can effectively be applied only to those scenarios in which the "energy"of the attack

is sufficiently large. To this aim, the following requirement is taken for granted.

Assumption 4.1.1. *If a malicious agent acts, the attack is always such that condition* (3.59) *holds for all the involved subsystems.*

This assumption is extremely important. If it does not hold, not only the proposed strategies are not effective, but they could lead to a wrong resolution of the ambiguity as well.

Under such hypothesis, by considering the detection algorithm (see the flow chart in Figure 3.3), one easily deduces that a detector $\mathcal{D}_i$ decides subsystem $\mathcal{S}_i$ is under attack if and only if:

$$\forall j \in \mathcal{N}_i, \exists l \in \mathcal{N}_j \text{ s.t. } \mathcal{S}_l \text{ is under attack.} \tag{4.1}$$

Note that $\mathcal{S}_l$ might or might not be $\mathcal{S}_i$ itself. The problems arise in the latter case, that is when the detector incorrectly decides its subsystem is under attack due to ambiguity in the network topology.

Remark 4.1.1. *Throughout all this chapter, the analysis will focus on the framework of a single attacker within the same neighborhood, in line with Assumption 3.3.2. Moreover, concerning the topology, Assumption 3.1.2 is supposed to hold.*

If at most one attacker is perturbing a subsystem in each same neighborhood, then the topological configuration resulting in ambiguities can be completely characterized by means of the following condition.

Theorem 4.1.1 (Ambiguous topologies)**.** *Under Assumptions 3.3.2 and 3.1.2, a subsystem $\mathcal{S}_i$ incorrectly decides it is under attack if and only if the following condition holds:*

$$\mathcal{N}_i \subseteq \mathcal{N}_h, \tag{4.2}$$

where $\mathcal{S}_h, h \neq i$, is the (only) actually attacked subsystem.

Proof.

- **Sufficiency "$\Leftarrow$".**
  Let assume condition (4.2) is satisfied. Then, because of Assumption 3.1.2, we have:
  $$\mathcal{N}_i \subseteq \mathcal{N}_h \Rightarrow \forall j \in \mathcal{N}_i \subseteq \mathcal{N}_h, h \in \mathcal{N}_j. \tag{4.3}$$
  If $\mathcal{S}_h$ is the actually attacked subsystem, condition (4.1) is verified and detector $\mathcal{D}_i$ decides $\mathcal{S}_i$ it is under attack, even though it is not.

- **Necessity "$\Rightarrow$".**
  Let assume $\mathcal{D}_i$ (incorrectly) decides $\mathcal{S}_i$ is under attack. Therefore, from the characterization (4.1), and from Assumption 3.3.2, we have:
  $$\forall j \in \mathcal{N}_i, \exists l = h \in \mathcal{N}_j \text{ s.t. } \mathcal{S}_l = \mathcal{S}_h \text{ is under attack.} \tag{4.4}$$
  As a consequence, all the subsystems $\mathcal{S}_j$ neighbors of $\mathcal{S}_i$ must be neighbors of $\mathcal{S}_h$ as well, that is (4.2) holds.

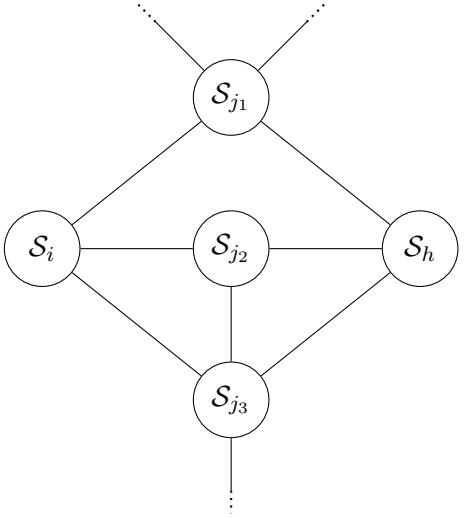


Figure 4.1: A simple example of topology where an isolation algorithm is needed.

□

Before concluding the analysis of this section, the next remark proposes a further observation based on the cardinality of the neighborhood, which might happen to be useful under some circumstances.

Remark 4.1.2. *It is worth noticing that, when condition* (4.2) *holds in a strict sense, that is* $\mathcal{N}_i \subset \mathcal{N}_h$*, the ambiguity can be easily resolved by looking at the alarm signals* a_l *of those subsystems* $\mathcal{S}_l, l \in \mathcal{N}_h \setminus \mathcal{N}_i$*. Indeed, if alarm signals* a_l *are off, then one can deduce* $\mathcal{S}_i$ *is under attack, since an attack in* $\mathcal{S}_h$ *would have caused* $a_j = 1, \forall j \in \mathcal{N}_i \Rightarrow a_l = 1, l \in \mathcal{N}_i \setminus \mathcal{N}_h$*. Obviously, in order for this to be done in practice, at least one of the subsystems* $\mathcal{S}_l, l \in \mathcal{N}_h \setminus \mathcal{N}_i$ *must be informed of the current ambiguity.*

4.2 Isolation strategies

Generally speaking, the main goal of an isolation strategy is to deduce a quantity sensitive to one only of the two (or more) possible candidate attacked subsystems, therefore allowing to discriminate whenever ambiguities might occur.

In light of Theorem 4.1.1 and of Remark 4.1.2, the main focus of this section are those configurations in which there are two subsystems $\mathcal{S}_i$ and $\mathcal{S}_h$ such that $\mathcal{N}_i = \mathcal{N}_h$, see the example depicted in Figure 4.1. Observe that such configurations can be identified by running a distributed algorithm. For instance, in the procedure illustrated in Algorithm 1 each subsystem receives the degree d_j (i.e. the cardinality of the neighborhood) of each of its neighboring subsystem $\mathcal{S}_j, j \in \mathcal{N}_i$. By comparing the degrees, it can find all the couples of neighbors

$(\mathcal{S}_j, \mathcal{S}_l), l \neq j$, with the same degree, which are potentially ambiguous subsystems. In such a case, $\mathcal{S}_i$ alerts both $\mathcal{S}_j$ and $\mathcal{S}_l$ of the presence of the other one by means of the sets $\mathcal{I}_{j,i}$ and $\mathcal{I}_{l,i}$, respectively. Therefore, each subsystem $\mathcal{S}_i$ receives from each neighbor $\mathcal{S}_j, j \in \mathcal{N}_i$, a list $\mathcal{I}_{ij}$, that is the index set of subsystems ambiguous to $\mathcal{S}_i$ from the perspective of $\mathcal{S}_j$. From these, $\mathcal{S}_i$ computes the intersection, obtaining $\mathcal{N}_{A,i}$. All the subsystems whose index is in $\mathcal{N}_{A,i}$ share exactly the same neighborhood of $\mathcal{S}_i$. Indeed, at the end of the procedure the index set $\mathcal{N}_{A,i}$ satisfies:

$$\mathcal{N}_{A,i} = \left\{ h : h \in \left(\bigcap_{j \in \mathcal{N}_i} \mathcal{N}_j \right), d_h = d_i \right\}. \tag{4.5}$$

As a consequence, all the subsystems $\mathcal{S}_h, h \in \mathcal{N}_{A,i}$, are so that $\mathcal{N}_i \subseteq \mathcal{N}_h$. Moreover, from $d_h = d_i$, the inclusion is actually an equality, therefore $\mathcal{N}_{A,i}$ accounts for all and only those subsystems $\mathcal{S}_h$ such that $\mathcal{N}_i = \mathcal{N}_h$.

Algorithm 1 Isolation ambiguities finder

#Propagate degree to neighbors
  send $d_i = |\mathcal{N}_i|$ to each $\mathcal{S}_j, j \in \mathcal{N}_i$
  receive $d_j, j \in \mathcal{N}_i$
#Collect candidate ambiguous subsystems in $\mathcal{I}_{j,i}$
  for $j \in \mathcal{N}_i$ **do**
    $\mathcal{I}_{j,i} \leftarrow \emptyset$
    for $l \in \mathcal{N}_i,\ l \neq j$ **do**
      if $d_j = d_l$ **then**
        $\mathcal{I}_{j,i} \leftarrow \mathcal{I}_{j,i} \cup \{l\}$
      end if
    end for
    send $\mathcal{I}_{j,i}$ to $\mathcal{S}_j$
  end for
#Obtain the set of ambiguous neighbors $\mathcal{N}_{A,i}$
  receive $\mathcal{I}_{i,j}, j \in \mathcal{N}_i$
  $\mathcal{N}_{A,i} \leftarrow \bigcap_{j \in \mathcal{N}_i} \mathcal{I}_{i,j}$

The whole analysis of this section relies on the following assumption.

Assumption 4.2.1. *Either subsystem $\mathcal{S}_i$ or $\mathcal{S}_h$ can be under attack, whereas all the common neighbors $\mathcal{S}_j, j \in \mathcal{N}_i = \mathcal{N}_h$ are not attacked.*

The reason behind this conjecture lies in the fact that all the isolation architecture is worth to be considered only when the ambiguity between $\mathcal{S}_i$ and $\mathcal{S}_h$ occurs. If other subsystems are under attack, the detection strategy of the previous chapter is itself sufficient to locate the attack.

In the following, three isolation techniques are presented: a UIO for a properly chosen merged subsystem (Subsection 4.2.1), a filtering of the Luenberger

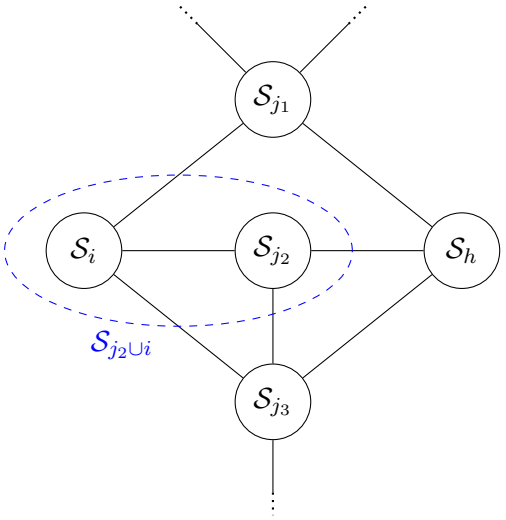


Figure 4.2: Scheme of the merged subsystem whose UIO $\mathcal{O}^d_{j_2 \cup i}$ can be used to solve the isolation problem.

residue (Subsection 4.2.2), and the filtering of the residue of a different type of observer, ad hoc designed to exploit the asymmetry in the interconnection matrices (Subsection 4.2.3).

4.2.1 Merged UIO

A first approach to resolve the ambiguity described in the previous section is to virtually merge one of the candidate attacked subsystems with one of the common neighbors, and to design and run a UIO for the merged subsystem.

Let $\mathcal{S}_i$ and $\mathcal{S}_h$ be the two candidate attacked subsystems, and let subsystem $\mathcal{S}_j, j \in \mathcal{N}_i = \mathcal{N}_h$, be a common neighbor. Let $\mathcal{S}_{j \cup i}$ be the merged subsystem, as illustrated in Figure 4.2. Such a subsystem is described by a state vector $x_{j\cup i} \doteq \left[\begin{array}{c|c} x_j^\top & x_i^\top \end{array}\right]^\top \in \mathbb{R}^{n_j+n_i}$, a input vector $u_{j\cup i} \doteq \left[\begin{array}{c|c} u_j^\top & u_i^\top \end{array}\right]^\top \in \mathbb{R}^{m_j+m_i}$, and an output vector $y_{j\cup i} \doteq \left[\begin{array}{c|c} y_j^\top & y_i^\top \end{array}\right]^\top \in \mathbb{R}^{p_j+p_i}$. Moreover, the attacked signals are defined analogously to (3.6), precisely:

$$\begin{aligned} \tilde{u}_{j\cup i} &= u_{j\cup i} + \left[\begin{array}{c} 0 \\ \hline \eta_i \end{array}\right] \\ \tilde{y}_{j\cup i} &= y_{j\cup i} - \left[\begin{array}{c} 0 \\ \hline \gamma_i \end{array}\right]. \end{aligned} \tag{4.6}$$

From (3.1), the dynamics of the merged subsystem $\mathcal{S}_{j\cup i}$ is obtained:

$$\mathcal{S}_{j\cup i} : \begin{cases} x^+_{j\cup i} = A_{j\cup i} x_{j\cup i} + B_{j\cup i}\tilde{u}_{j\cup i} + \Xi_{j\cup i}\boldsymbol{x}_{j\cup i} + w_{j\cup i} \\ y_{j\cup i} = C_{j\cup i} x_{j\cup i} + v_{j\cup i}, \end{cases} \tag{4.7}$$

where $w_{j\cup i} = \left[\begin{array}{c|c} w_j^\top & w_i^\top \end{array}\right]^\top$, $v_{j\cup i} = \left[\begin{array}{c|c} v_j^\top & v_i^\top \end{array}\right]^\top$, and:

$$\begin{aligned} A_{j\cup i} &\doteq \left[\begin{array}{c|c} A_j & A_{ji} \\ \hline A_{ij} & A_i \end{array}\right], \quad B_{j\cup i} \doteq \left[\begin{array}{c|c} B_j & 0 \\ \hline 0 & B_i \end{array}\right], \quad C_{j\cup i} \doteq \left[\begin{array}{c|c} C_j & 0 \\ \hline 0 & C_i \end{array}\right], \\ \Xi_{j\cup i} &\doteq \underset{\substack{l\in\mathcal{N}_j\cup\mathcal{N}_i \\ l\neq j,i}}{\text{row}} \left(\left[\begin{array}{c} A_{jl} \\ \hline A_{il} \end{array}\right] \right), \quad \boldsymbol{x}_{j\cup i}^\top \doteq \underset{\substack{l\in\mathcal{N}_j\cup\mathcal{N}_i \\ l\neq j,i}}{\text{row}} (x_l^\top). \end{aligned} \tag{4.8}$$

Furthermore, let subdivide the interconnection matrix of the merged subsystem as follows:

$$\left[\begin{array}{c} \check{\Xi}_j \\ \hline \check{\Xi}_i \end{array}\right] \doteq \Xi_{j\cup i} \tag{4.9}$$

so that, we have:

$$\text{Im}(\Xi_j) = \text{Im}\left(\left[\begin{array}{c|c} A_{ji} & \check{\Xi}_j \end{array}\right] \right) \tag{4.10a}$$

$$\text{Im}(\Xi_i) = \text{Im}\left(\left[\begin{array}{c|c} A_{ij} & \check{\Xi}_i \end{array}\right] \right). \tag{4.10b}$$

Observe that $\Xi_i \neq \left[\begin{array}{c|c} A_{ij} & \check{\Xi}_i \end{array}\right]$, since $\check{\Xi}_i$ is padded with zero blocks, precisely $0 \in \mathbb{R}^{n_i\times n_l}, \forall l \in \mathcal{N}_j \setminus \mathcal{N}_i, l \neq i$.

Then, the effectiveness of this isolation strategy can be understood by observing that the resulting UIO estimate of the state $x_{j\cup i}$ is insensitive to attacks at $\mathcal{S}_h$. Indeed, the influence of x_i on x_j is within the internal dynamics of the merged subsystem $\mathcal{S}_{j\cup i}$. Conversely, the other candidate attacked subsystem $\mathcal{S}_h$ is one of the neighbors of the merged subsystem $\mathcal{S}_{j\cup i}$, hence the coupling of x_j and x_h does not influence the computation the UIO estimate $\hat{x}_{j\cup i}^d$ of the state of $\mathcal{S}_{j\cup i}$. As a result, if particular conditions are met, the computed residual $\tilde{r}_{j\cup i}^d$ is sensitive to attacks in $\mathcal{S}_i$ and insensitive to attacks in $\mathcal{S}_h$, therefore it can be used to discriminate.

Specifically, let consider a UIO for the merged subsystem in the form:

$$\mathcal{O}_{j\cup i}^d : \begin{cases} z_{j\cup i}^+ = F_{j\cup i} z_{j\cup i} + T_{j\cup i} B_{j\cup i} u_{j\cup i} + K_{j\cup i} \tilde{y}_{j\cup i} \\ \hat{x}_{j\cup i}^d = z_{j\cup i} + H_{j\cup i} \tilde{y}_{j\cup i}, \end{cases} \tag{4.11}$$

where the gains are chosen so that equations (3.25) hold for the merged subsystem (4.7), that is:

$$0 = (H_{j\cup i} C_{j\cup i} - I)\Xi_{j\cup i} \tag{4.12a}$$

$$T_{j\cup i} = I - H_{j\cup i} C_{j\cup i} \tag{4.12b}$$

$$F_{j\cup i}^d = \bar{A}_{j\cup i} - K_{j\cup i}^{(1)} C_{j\cup i}, \text{ where } F_{j\cup i}^d \text{ is a Schur-stable matrix} \tag{4.12c}$$

$$K_{j\cup i}^{(2)} = F_{j\cup i} H_{j\cup i}, \tag{4.12d}$$

where

$$\bar{A}_{j\cup i} = A_{j\cup i} - H_{j\cup i} C_{j\cup i} A_{j\cup i}, \tag{4.13}$$

and

$$K_{j\cup i} = K^{(1)}_{j\cup i} + K^{(2)}_{j\cup i}. \tag{4.14}$$

Observe that a suitable threshold must be designed to take the action of noise into account. Nonetheless, the computation of Proposition 3.3.3 can be exploited, by properly substituting the merged subsystem's matrices and noise bounds, in place of the single subsystem's ones.

Remark 4.2.1. *In order to implement this strategy, the agent $\mathcal{S}_j$ needs to receive the following information:*

- ***offline**: the structure of the other subsystem $\mathcal{S}_i$ and of its interconnections (A_i, B_i, C_i, Ξ_i); this is needed to design the merged UIO gains $H_{j\cup i}$, $T_{j\cup i}$, $F^d_{j\cup i}$, $K_{j\cup i}$;*
- ***online**: the actuation and the measurement signals, $(u_i, \tilde{y}_i)$; this is needed to feed the UIO, and to dynamically compute the estimate $\hat{x}^d_{j\cup i}$ and the residue $\tilde{r}^d_{j\cup i}$.*

Moreover, if $F_{j\cup i}$ is designed in block-diagonal form (or at least triangular), one can compute only a portion of the vectors $z_{j\cup i}$ and $\hat{x}^d_{j\cup i}$ in order to reduce the computational burden, since at least a portion of $z_{j\cup i}$ evolves independently of the other.

On the other hand, designing a UIO for the merged subsystem $\mathcal{S}_{j\cup i}$ cannot always be effectively applied, and some considerations on the structural properties of the interconnections are needed.

Firstly, in order to implement the UIO for $\mathcal{S}_{j\cup i}$, condition (3.22) must hold for the merged subsystem. On the other hand, it can be easily proved that if such a condition holds for each single subsystem, it is also verified for the merged subsystem, as discussed in Appendix A.

Moreover, under particular conditions on the interconnection matrices, the attack in $\mathcal{S}_i$ might be covert for the UIO of the merged subsystem as well, causing the isolation strategy to fail. The following analysis aims to clarify the reason behind this important drawback.

For such an observer, along the lines of (3.17), the computed error is defined as:

$$\tilde{\epsilon}^d_{j\cup i} \doteq x_{j\cup i} - \begin{bmatrix} 0 \\ \hline \tilde{x}_i \end{bmatrix} - \hat{x}^d_{j\cup i}. \tag{4.15}$$

The next theorem provides sufficient conditions on the algebraic properties of the interconnections to prevent the merged subsystem's UIO $\mathcal{O}^d_{j\cup i}$ from detecting the attack in $\mathcal{S}_i$, hence limiting the effectiveness of this proposed strategy.

Theorem 4.2.1. *If the interconnections between subsystems are so that*

$$\dim\left(\operatorname{Im}(\check{\Xi}_j^\top)\cap\operatorname{Im}(\check{\Xi}_i^\top)\right)=0 \tag{4.16}$$

and:

$$\operatorname{Im}(A_{ji})\subseteq\operatorname{Im}(\check{\Xi}_j), \tag{4.17}$$

then any covert attack at $\mathcal{S}_i$ is covert for the the UIO $\mathcal{O}^d_{j\cup i}$ of the merged subsystem $\mathcal{S}_{j\cup i}$.

Proof. Before proceeding with the computation, one might find it useful to remind that the actual subsystem $\mathcal{S}_{j\cup i}$ is driven by the attacked control input $\tilde{u}_{j\cup i}$, whereas the decentralized observer $\mathcal{O}^d_{j\cup i}$ is fed with the legitimate control input $u_{j\cup i}$ and the attacked measurement signal $\tilde{y}_{j\cup i}$.

From (4.6), (3.7), (4.11), and definition (4.15), we have:

$$\begin{aligned}\tilde{\epsilon}^{d+}_{j\cup i}=&x^+_{j\cup i}-\begin{bmatrix}0\\ \hline \tilde{x}_i^+\end{bmatrix}-\hat{x}^{d+}_{j\cup i}\\ =&x^+_{j\cup i}-\begin{bmatrix}0\\ \hline \tilde{x}_i^+\end{bmatrix}-z^+_{j\cup i}-H_{j\cup i}\left[C_{j\cup i}\left(x^+_{j\cup i}-\begin{bmatrix}0\\ \hline \tilde{x}_i^+\end{bmatrix}\right)+v^+_{j\cup i}\right]\\ =&(I-H_{j\cup i}C_{j\cup i})\left(x^+_{j\cup i}-\begin{bmatrix}0\\ \hline \tilde{x}_i^+\end{bmatrix}\right)-z^+_{j\cup i}-H_{j\cup i}v^+_{j\cup i}.\end{aligned} \tag{4.18}$$

Let observe that, from (3.7), the following equality holds:

$$\begin{bmatrix}0\\ \hline \tilde{x}_i^+\end{bmatrix}=\begin{bmatrix}0\\ \hline A_i\tilde{x}_i+B_i\eta_i\end{bmatrix}=A_{j\cup i}\begin{bmatrix}0\\ \hline \tilde{x}_i\end{bmatrix}+B_{j\cup i}\begin{bmatrix}0\\ \hline \eta_i\end{bmatrix}-\begin{bmatrix}A_{ji}\tilde{x}_i\\ \hline 0\end{bmatrix}. \tag{4.19}$$

As a consequence, taken also (4.7) and (4.11) into account, it is obtained:

$$\begin{aligned}\tilde{\epsilon}^{d+}_{j\cup i}=&(I-H_{j\cup i}C_{j\cup i})\Bigg\{A_{j\cup i}x_{j\cup i}+B_{j\cup i}\left(u_{j\cup i}+\begin{bmatrix}0\\ \hline \eta_i\end{bmatrix}\right)+\Xi_{j\cup i}\boldsymbol{x}_{j\cup i}\\ &+w_{j\cup i}-A_{j\cup i}\begin{bmatrix}0\\ \hline \tilde{x}_i\end{bmatrix}-B_{j\cup i}\begin{bmatrix}0\\ \hline \eta_i\end{bmatrix}+\begin{bmatrix}A_{ji}\tilde{x}_i\\ \hline 0\end{bmatrix}\Bigg\}-\Bigg[F^d_{j\cup i}z_{j\cup i}\\ &+T_{j\cup i}B_{j\cup i}u_{j\cup i}+K_{j\cup i}C_{j\cup i}\left(x_{j\cup i}-\begin{bmatrix}0\\ \hline \tilde{x}_i\end{bmatrix}\right)+K_{j\cup i}v_{j\cup i}\Bigg]z-H_{j\cup i}v^+_{j\cup i}.\end{aligned} \tag{4.20}$$

The previous expression can be rearranged in the form:

$$\begin{aligned}\tilde{\epsilon}^{d+}_{j\cup i}=&[(I-H_{j\cup i}C_{j\cup i})A_{j\cup i}-K_{j\cup i}C_{j\cup i}]\left(x_{j\cup i}-\begin{bmatrix}0\\ \hline \tilde{x}_i\end{bmatrix}\right)\\ &+[(I-H_{j\cup i}C_{j\cup i})-T_{j\cup i}]B_{j\cup i}u_{j\cup i}+(I-H_{j\cup i}C_{j\cup i})\Xi_{j\cup i}\boldsymbol{x}_{j\cup i}\\ &+(I-H_{j\cup i}C_{j\cup i})w_{j\cup i}-F^d_{j\cup i}z_{j\cup i}-K_{j\cup i}v_{j\cup i}\\ &-H_{j\cup i}v^+_{j\cup i}+(I-H_{j\cup i}C_{j\cup i})\begin{bmatrix}A_{ji}\tilde{x}_i\\ \hline 0\end{bmatrix}.\end{aligned} \tag{4.21}$$

By applying (4.12a), (4.12b), and (4.13), we find:

$$\begin{aligned}\tilde{\epsilon}^{d+}_{j\cup i} =&[\bar{A}_{j\cup i} - K_{j\cup i}C_{j\cup i}]\left(x_{j\cup i} - \left[\frac{0}{\tilde{x}_i}\right]\right) + T_{j\cup i}w_{j\cup i} - F^d_{j\cup i}z_{j\cup i}\\ &- K_{j\cup i}v_{j\cup i} - H_{j\cup i}v^+_{j\cup i} + T_{j\cup i}\left[\frac{A_{ji}\tilde{x}_i}{0}\right]\\ =&[\bar{A}_{j\cup i} - (K^{(1)}_{j\cup i} - K^{(2)}_{j\cup i})C_{j\cup i}]\left(x_{j\cup i} - \left[\frac{0}{\tilde{x}_i}\right]\right) + T_{j\cup i}w_{j\cup i}\\ &- F^d_{j\cup i}z_{j\cup i} - (K^{(1)}_{j\cup i} + K^{(2)}_{j\cup i})v_{j\cup i} - H_{j\cup i}v^+_{j\cup i} + T_{j\cup i}\left[\frac{A_{ji}\tilde{x}_i}{0}\right],\end{aligned} \tag{4.22}$$

where (4.14) was employed. Finally, by recalling (4.12c) and (4.12d), we have:

$$\begin{aligned}\tilde{\epsilon}^{d+}_{j\cup i} =&F^d_{j\cup i}\left(x_{j\cup i} - \left[\frac{0}{\tilde{x}_i}\right]\right) - F^d_{j\cup i}\left(z_{j\cup i} + H_{j\cup i}\tilde{y}_{j\cup i}\right)\\ &+ T_{j\cup i}w_{j\cup i} - K^{(1)}_{j\cup i}v_{j\cup i} - H_{j\cup i}v^+_{j\cup i} + T_{j\cup i}\left[\frac{A_{ji}\tilde{x}_i}{0}\right]\\ =&F^d_{j\cup i}\tilde{\epsilon}^d_{j\cup i} + T_{j\cup i}w_{j\cup i} - K^{(1)}_{j\cup i}v_{j\cup i} - H_{j\cup i}v^+_{j\cup i} + T_{j\cup i}\left[\frac{A_{ji}\tilde{x}_i}{0}\right].\end{aligned} \tag{4.23}$$

Now, let consider the following block-partition of matrix $H_{j\cup i}$:

$$H_{j\cup i} = \left[\begin{array}{c|c} H_{jj} & H_{ji} \\ \hline H_{ij} & H_{ii}\end{array}\right], \tag{4.24}$$

where the blocks are coherent with the dimension of the state and output of $\mathcal{S}_j$ and $\mathcal{S}_i$, respectively. In this light, equation (4.12a) leads to:

$$\begin{aligned}0 =&(H_{j\cup i}C_{j\cup i} - I)\Xi_{j\cup i}\\ =&\left(\left[\begin{array}{c|c} H_{jj} & H_{ji} \\ \hline H_{ij} & H_{ii}\end{array}\right]\left[\begin{array}{c|c} C_j & 0 \\ \hline 0 & C_i\end{array}\right] - I\right)\Xi_{j\cup i}\\ =&\left[\begin{array}{c|c} H_{jj}C_j - I & H_{ji}C_i \\ \hline H_{ij}C_j & H_{ii}C_i - I\end{array}\right]\left[\begin{array}{c} \check{\Xi}_j \\ \hline \check{\Xi}_i\end{array}\right]\end{aligned} \tag{4.25}$$

that is:

$$(H_{jj}C_j - I)\check{\Xi}_j + H_{ji}C_i\check{\Xi}_i = 0 \tag{4.26a}$$

$$H_{ij}C_j\check{\Xi}_j + (H_{ii}C_i - I)\check{\Xi}_i = 0. \tag{4.26b}$$

In light of (4.16), the equalities (4.26a) and (4.26b) become:

$$(H_{jj}C_j - I)\check{\Xi}_j = 0 \tag{4.27a}$$

$$H_{ji}C_i\check{\Xi}_i = 0 \tag{4.27b}$$

$$H_{ij}C_j\check{\Xi}_j = 0 \tag{4.27c}$$

$$(H_{ii}C_i - I)\check{\Xi}_i = 0. \tag{4.27d}$$

Furthermore, because of (4.17), (4.27a) and (4.27c) give:

$$A_{ji} \in \ker \left(\left[\frac{H_{jj}C_j - I}{H_{ij}C_j} \right] \right) \Rightarrow \left[\frac{A_{ji}\tilde{x}_i}{0} \right] \in \ker(T_{j\cup i}), \tag{4.28}$$

where the last implication can be easily verified by inspection by recalling (4.12b) and the computation in (4.25). As a consequence, we have:

$$\tilde{\epsilon}^{d+}_{j\cup i} = F^d_{j\cup i}\tilde{\epsilon}^d_{j\cup i} + T_{j\cup i}w_{j\cup i} - K^{(1)}_{j\cup i}v_{j\cup i} - H_{j\cup i}v^+_{j\cup i}. \tag{4.29}$$

Since the computed error $\tilde{\epsilon}^d_{j\cup i}$ of the merged UIO $\mathcal{O}^d_{j\cup i}$ evolves independently of $\tilde{x}_i$ and η_i, it is insensitive to the malicious agent action, therefore the attack is necessarily covert. □

It is worth highlighting that Theorem 4.2.1 provides sufficient conditions only, that is under hypothesis (4.16) and (4.17) the proposed strategy fails in resolving the isolation problem, whereas, in practice, one is interested in finding sufficient conditions so that it is successful. Nonetheless, finding sufficient conditions on the interconnection matrices so that this approach is effective is far from trivial. In practice, once the UIO $\mathcal{O}^d_{j\cup i}$ has been designed, the necessary and sufficient condition for it to be insensitive to the attack in $\mathcal{S}_i$ is (4.28), which can easily be checked. Moreover, one might find it useful to observe that the least-squares solution (3.26) is not necessarily the only solution H in the UIO design. Furthermore, in order to solve the ambiguity by adopting this approach, one merged subsystem so that (4.28) does not hold is itself sufficient, as far as it includes either $\mathcal{S}_i$ or $\mathcal{S}_h$, and one of the common neighbors $\mathcal{S}_j, j \in \mathcal{N}_i = \mathcal{N}_h$.

Remark 4.2.2. *Technically, there might be pathological situations in which* (4.28) *does not hold, but the attack is designed so that:*

$$A_{ji}\tilde{x}_i \in \ker \left(\left[\frac{H_{jj}C_j - I}{H_{ij}C_j} \right] \right). \tag{4.30}$$

In such a case, the residue $\tilde{r}^d_{j\cup i}$ does not sense the attack and the strategy incorrectly leads to the conclusion that $\mathcal{S}_h$ is attacked. Firstly, let observe that this condition results in the mentioned wrong deduction only if it holds for all the time span of the attack, which is not very likely in practice. Moreover, a possible solution is to adopt multiple isolation strategies.

4.2.2 Filtered Luenberger Observer

In this subsection a further isolation strategy is proposed. The fundamental inconvenience of the detection strategy is that the communicated residue $\tilde{r}^c_j$, resulting from the Luenberger Observer $\mathcal{O}^c_j$ implemented in the logic unit LU_j of each common neighbor $\mathcal{S}_j, j \in \mathcal{N}_i = \mathcal{N}_h$, is sensitive to attacks in both the two ambiguous subsystems $\mathcal{S}_i$ and $\mathcal{S}_h$, see the scheme in Figure 4.1. Nonetheless,

under some conditions, it might still be possible to design a proper filter so that the filtered residue is sensitive to one only, breaking the symmetry, thus allowing for discrimination.

In the following, it is formally stated when, by designing an appropriate gain $g_j \in \mathbb{R}^{p_j}$, this filter can be successfully exploited. More precisely, the *filtered* residual being considered is:

$$\tilde{r}_j^g \doteq g_j^\top \tilde{r}_j^c. \tag{4.31}$$

Note that such a residual is only useful for isolation purposes, whereas it does not substitute the communicated residue $\tilde{r}_j^c$ in the detection architecture.

The details behind this technique are outlined in the following theorem, where the filtered residual is designed in order to obtain a quantity sensitive to attacks in $\mathcal{S}_i$ only. Obviously, a residual sensitive to attacks only in $\mathcal{S}_h$ would itself be effective.

Theorem 4.2.2. *Let assume the output matrix of subsystem $\mathcal{S}_j, j \in \mathcal{N}_i = \mathcal{N}_h$ is full column-rank, that is:*

$$\operatorname{rank}(C_j) = n_j. \tag{4.32}$$

Moreover, let $g_j \in \mathbb{R}^{p_j}$ be such that:

$$g_j^\top C_j A_{jh} = 0 \tag{4.33a}$$

$$g_j^\top C_j A_{ji} \neq 0. \tag{4.33b}$$

Then, is it possible to design a residue $\tilde{r}_j^g$ which is sensitive to attacks in $\mathcal{S}_i$ only.

Proof. Let consider the distributed observer (3.34). By convolving the computed error dynamics (3.37), one obtains:

$$\tilde{\epsilon}_j^c(k) = (F_j^c)^k \tilde{\epsilon}_j^c(0) + \sum_{\tau=0}^{k-1} (F_j^c)^{k-1-\tau} \left[w_j(\tau) - L_j v_j(\tau) + \sum_{l \in \mathcal{N}_j} A_{jl} \epsilon_l^d(\tau) \right] \tag{4.34}$$

Then, the distributed communicated residue $\tilde{r}_j^c$ of each common neighboring subsystem $\mathcal{S}_j, j \in \mathcal{N}_i = \mathcal{N}_h$, defined in (3.18), evolves as:

$$\begin{aligned} \tilde{r}_j^c(k) =& C_j (F_j^c)^k \tilde{\epsilon}_j^c(0) + C_j \sum_{\tau=0}^{k-1} (F_j^c)^{k-1-\tau} \Bigg[w_j(\tau) \\ & - L_j v_j(\tau) + \sum_{l \in \mathcal{N}_j} A_{jl} \epsilon_l^d(\tau) \Bigg] + v_j(k), \end{aligned} \tag{4.35}$$

and, from (4.31), the filtered residue's time evolution is then:

$$\begin{aligned} \tilde{r}_j^g(k) =& g_j^\top C_j (F_j^c)^k \tilde{\epsilon}_j^c(0) + g_j^\top C_j \sum_{\tau=0}^{k-1} (F_j^c)^{k-1-\tau} \Bigg[w_j(\tau) \\ & - L_j v_j(\tau) + \sum_{l \in \mathcal{N}_j} A_{jl} \epsilon_l^d(\tau) \Bigg] + g_j^\top v_j(k). \end{aligned} \tag{4.36}$$

Let L_j be designed so that F_j^c is a scalar matrix, that is:

$$F_j^c = A_j - L_j C_j = \lambda_j I, \tag{4.37}$$

with $\lambda_j \in \mathbb{R}, |\lambda_j| < 1$. Observe that this is always possible thanks to hypothesis (4.32). Indeed, given any desired F_j^c, it suffices to choose L_j as:

$$L_j = (A_j - F_j) C_j^{-L}, \tag{4.38}$$

being C_j^{-L} a Moore-Penrose left-inverse [22, Section 3.6].

Given (4.37), (4.36) can be rewritten as:

$$\begin{aligned} \tilde{r}_j^g(k) =& \lambda_j^k g_j^\top C_j \tilde{\epsilon}_j^c(0) + \sum_{\tau=0}^{k-1} \lambda_j^{k-1-\tau} \Bigg[g_j^\top C_j \Bigg(w_j(\tau) \\ & - L_j v_j(\tau) \Bigg) + \sum_{l \in \mathcal{N}_j} g_j^\top C_j A_{jl} \epsilon_l^d(\tau) \Bigg] + g_j^\top v_j(k). \end{aligned} \tag{4.39}$$

Then, as a consequence of (4.33), one easily deduces that the contribution of ϵ_h^d in (4.39) is canceled, whereas this is not the case for ϵ_i^d. Hence, that is $\tilde{r}_j^g$ is insensitive to the attacks in $\mathcal{S}_i$ only.

It is worth mentioning that if $\lambda_j = 0$, that is $\mathcal{O}_j^c$ is a dead-beat observer [24, Section 7.6], the argument is still effective because we have $(F_j^c)^0 = I$. □

Remark 4.2.3. *In line with Remark 4.2.2, pathological situations in which the attack in $\mathcal{S}_i$ is designed so that $\tilde{x}_i \in \ker(g_j^\top C_j A_{ji})$ might exist, despite* (4.33)*. In such situations, this isolation methodology leads to the wrong deduction that $\mathcal{S}_h$ is under attack. Despite very unlikely in practice, this scenario could be tackled by adopting multiple isolation architectures.*

In practice, due to the presence of noise, a suitable threshold on the norm of the filtered residue $\tilde{r}_j^g$ must be designed. Nonetheless, observe that if the filter is designed so that:

$$\|g_j\| = \left\| g_j^\top \right\| = 1, \tag{4.40}$$

then the same threshold of (3.58) can be exploited, as a consequence of the triangle inequality. Technically, a less conservative threshold might be designed, but the computation is here omitted as analogous to that of Proposition 4.2.3 in the next section.

Remark 4.2.4. *Note that this technique requires no additional burden in the information exchange, neither online, nor offline, with respect to that required for the detection strategy in Chapter 3. Indeed, the common neighbor $\mathcal{S}_j$ needs to have knowledge of the following information:*

- ***offline**: the structure of the interconnections A_{ji}, A_{jh}; this is needed to design the gain g_j;*

- ***online**: the decentralized estimate of each of its neighbors $\hat{x}_l^d, l \in \mathcal{N}_j$; this is needed to dynamically compute the residue $\tilde{r}_j^g$, together with its own actuation and measurement signals $(u_j, \tilde{y}_j)$.*

Finally, an important issue to be discussed is when a filter $g_j \in \mathbb{R}^{p_j}$ satisfying (4.33) (or the symmetric condition obtained by swapping $\mathcal{S}_i$ and $\mathcal{S}_h$) exists. The next proposition provides necessary and sufficient conditions about the existence of such a vector.

Proposition 4.2.1. *Under condition* (3.22)*, a vector $g_j \in \mathbb{R}^{n_j}$ satisfying either* (4.33) *or:*

$$g_j^\top C_j A_{jh} \neq 0 \tag{4.41a}$$

$$g_j^\top C_j A_{ji} = 0 \tag{4.41b}$$

exists if and only if:

$$\mathrm{Im}(A_{ji}) \neq \mathrm{Im}(A_{jh}), \tag{4.42}$$

where $j \in \mathcal{N}_i = \mathcal{N}_h$.

Proof. The existence of a vector $g_j \in \mathbb{R}^{p_j}$ satisfying either (4.33) or (4.41) is equivalent to the following condition:

$$\ker((C_j A_{ji})^\top) \neq \ker((C_j A_{jh})^\top). \tag{4.43}$$

By recalling the properties of the adjoint operator [24, Section A.13], the following identities hold:

$$\ker((C_j A_{ji})^\top) = (\mathrm{Im}(C_j A_{ji}))^\perp \tag{4.44a}$$

$$\ker((C_j A_{jh})^\top) = (\mathrm{Im}(C_j A_{jh}))^\perp. \tag{4.44b}$$

Furthermore, for any finite-dimension linear space $\mathcal{V}$, it holds:

$$(\mathcal{V}^\perp)^\perp = \mathcal{V}. \tag{4.45}$$

As a consequence of (4.44) and (4.45), (4.43) is equivalent to:

$$\mathrm{Im}(C_j A_{ji}) \neq \mathrm{Im}(C_j A_{jh}). \tag{4.46}$$

Finally, it suffices to prove that, under condition (3.22), (4.42) and (4.46) are equivalent.

- $\mathrm{Im}(A_{ji}) = \mathrm{Im}(A_{jh}) \Rightarrow \mathrm{Im}(C_j A_{ji}) = \mathrm{Im}(C_j A_{jh})$.
  From the fact that the image of a linear space through a linear mapping is unique, one trivially deduces that if the two range of the interconnection matrices coincide $\mathrm{Im}(A_{ij}) = \mathrm{Im}(A_{jh})$, then their image through the linear mapping represented by the matrix C_j with respect to some fixed basis must coincide as well, that is $\mathrm{Im}(C_j A_{ji}) = \mathrm{Im}(C_j A_{jh})$.

- $\mathrm{Im}(A_{ji}) \neq \mathrm{Im}(A_{jh}) \Rightarrow \mathrm{Im}(C_j A_{ji}) \neq \mathrm{Im}(C_j A_{jh})$.
  If $\mathrm{Im}(A_{ji}) \neq \mathrm{Im}(A_{jh})$, then it either holds that $\mathrm{Im}(A_{ji}) \not\subset \mathrm{Im}(A_{jh})$ or $\mathrm{Im}(A_{jh}) \not\subset \mathrm{Im}(A_{ji})$ (or both). Let assume $\mathrm{Im}(A_{ji}) \not\subset \mathrm{Im}(A_{jh})$. Then, $\exists v_i \in \mathbb{R}^{n_i}$ such that $\forall v_h \in \mathbb{R}^{n_h}$ we have $A_{ji} v_i \neq A_{jh} v_h$, that is:

$$\left[\begin{array}{c} v_i \\ \hline -v_h \end{array}\right] \notin \ker\left(\left[\begin{array}{c|c} A_{ji} & A_{jh} \end{array}\right] \right). \tag{4.47}$$

  Nonetheless, from Proposition A.0.1 (see Appendix A), we have:

$$\ker\left(\left[\begin{array}{c|c} A_{ji} & A_{jh} \end{array}\right] \right) = \ker\left(C_j \left[\begin{array}{c|c} A_{ji} & A_{jh} \end{array}\right] \right). \tag{4.48}$$

  Therefore, $\exists v_i$ s.t. $C_j A_{ji} v_i \neq C_j A_{jh} v_h, \forall v_h \in \mathbb{R}^{n_h}$, which trivially implies $\mathrm{Im}(C_j A_{ji}) \neq \mathrm{Im}(C_j A_{jh})$.

□

4.2.3 Filtered two-step Luenberger Observer

In this subsection the same idea of Subsection 4.2.2 is applied to a different type of observer, *ad hoc* designed in the following. Indeed, one of the main inconveniences behind the ambiguity arising when two subsystems $\mathcal{S}_i$ and $\mathcal{S}_h$ are so that $\mathcal{N}_i = \mathcal{N}_h$ is that the two subsystems are not mutually neighbors, and this prevents each one to be sensitive to attacks in the other. Were this the case, each one would be sensitive to attacks in the other.

The detection strategy presented in Chapter 3 relies on the fact that a neighbor $\mathcal{S}_j$ of the attacked subsystem $\mathcal{S}_i$ is able to detect inconsistency between the true value of the state x_i, to which x_j is physically coupled, and the wrong estimate broadcast from the logic unit LU_i. In absence of a physical coupling, one cannot directly reproduce this principle to virtually connect the two candidate attacked subsystems $\mathcal{S}_i$ and $\mathcal{S}_h$, so that $i \in \mathcal{N}_h, h \in \mathcal{N}_i$, thus breaking the symmetry $\mathcal{N}_i \neq \mathcal{N}_h$.

Nonetheless, the two subsystems are actually linked via a second-order relation through their common neighbors, therefore a similar principle can be applied. To this aim, a suitable two-step Luenberger observer is designed. For sake of simplicity, let define the set of the two-step neighbors of $\mathcal{S}_i$:

$$\bar{\mathcal{N}}_i \doteq \bigcup_{j \in \mathcal{N}_i} \mathcal{N}_j. \tag{4.49}$$

Moreover, let set $A_{ij} \doteq 0$ if $l \in \bar{\mathcal{N}}_i \setminus \mathcal{N}_j$.

The two-step Luenberger observer $\mathcal{O}_i^s$ updates the state estimate $\hat{x}_i^s$ according to the following dynamics:

$$\begin{aligned} \hat{x}_i^{s+} =& A_i \hat{x}_i^s + B_i u_i + \sum_{j \in \mathcal{N}_i} A_{ij} (A_j \hat{x}_j^{d-} + B_j u_j^-) \\ &+ \sum_{l \in \bar{\mathcal{N}}_i} M_{il} \hat{x}_l^{d-} + L_i (\tilde{y}_i - C_i \hat{x}_i^s), \end{aligned} \tag{4.50}$$

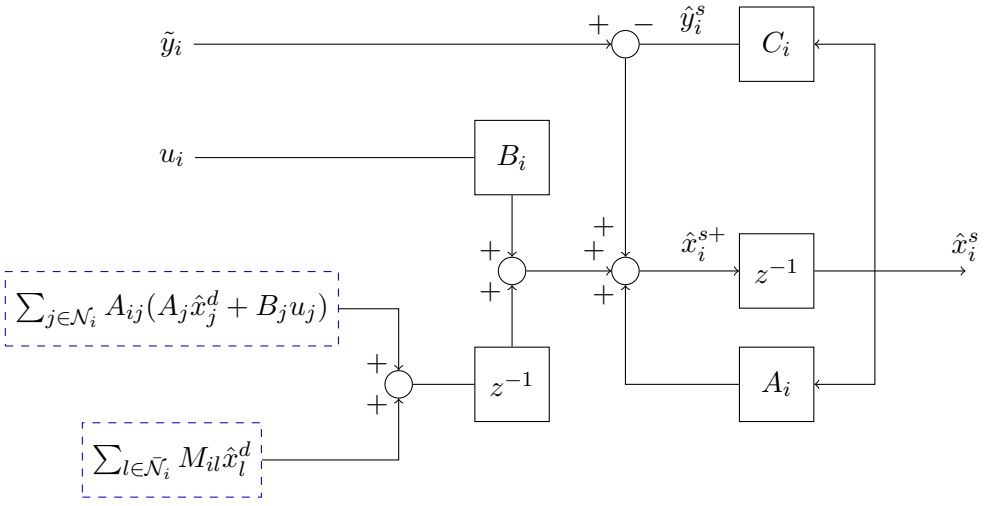


Figure 4.3: Scheme of the two-step Luenberger observer $\mathcal{O}_i^s$.

where, for each subsystem $\mathcal{S}_l, l \in \bar{\mathcal{N}}_i$, the second-order interconnection matrix is defined as:

$$M_{il} \doteq \sum_{j\in\mathcal{N}_i} A_{ij}A_{jl}. \tag{4.51}$$

The gain L_i can be chosen with the same criteria of that of the distributed observer presented in Subsection 3.3.2 and therefore the same symbol is here adopted. A scheme of the two-step Luenberger observer $\mathcal{O}_i^s$ is given in Figure 4.3. Observe that, in general, $\mathcal{N}_i \cap \bar{\mathcal{N}}_i \neq \emptyset$. This happens if (and only if) some of the neighboring subsystems $\mathcal{S}_j, j \in \mathcal{N}_i$, are connected and can be taken into account in order to simplify the setup.

The next proposition proves the effectiveness of the two-step Luenberger observer, and characterizes the dynamics of the computed estimation error, which, along the lines of (3.17), is defined as:

$$\tilde{\epsilon}_i^s \doteq x_i - \tilde{x}_i - \hat{x}_i^s. \tag{4.52}$$

Proposition 4.2.2. *Let assume a malicious agent $\mathcal{A}_i$ modeled as in* (3.7) *is manipulating the ith subsystem $\mathcal{S}_i$ as in* (3.6)*. If Assumption 3.2.1 holds, then the computed error dynamics of observer* (4.50) *is:*

$$\tilde{\epsilon}_i^{s+} = F_i^s\tilde{\epsilon}_i^s + w_i - L_i v_i + \sum_{j\in\mathcal{N}_i} A_{ji}w_j^- + \sum_{j\in\mathcal{N}_i} A_{ij}\epsilon_j^{d-} + \sum_{l\in\bar{\mathcal{N}}_i} M_{il}\epsilon_l^{d-}, \tag{4.53}$$

where:

$$F_i^s \doteq A_i - L_i C_i. \tag{4.54}$$

Proof. By rewriting (3.1) at the previous step, under Assumption 4.2.1, we have that for each neighboring subsystem $\mathcal{S}_j, j \in \mathcal{N}_i$, the following relation holds:

$$x_j = A_j x_j^- + B_j u_j^- + \sum_{l \in \mathcal{N}_j} A_{jl} x_l^- + w_j^-. \tag{4.55}$$

As a consequence, we can substitute it in (3.1) for $\mathcal{S}_i$, obtaining:

$$x_i^+ = A_i x_i + B_i \tilde{u}_i + \sum_{j \in \mathcal{N}_i} A_{ij} \left(A_j x_j^- + B_j u_j^- + \sum_{l \in \mathcal{N}_j} A_{jl} x_l^- + w_j^- \right) + w_i. \tag{4.56}$$

Form (4.56), (3.6), (3.7), and (4.50), the dynamics of the computed error can be derived:

$$\begin{aligned} \tilde{\epsilon}_i^{s+} =& x_i^+ - \tilde{x}_i^+ - \hat{x}_i^{s+} \\ =& A_i x_i + B_i(u_i + \eta_i) + \sum_{j \in \mathcal{N}_i} A_{ij} \Bigg(A_j x_j^- + B_j u_j^- + \sum_{l \in \mathcal{N}_j} A_{jl} x_l^- \\ & + w_j^- \Bigg) + w_i - (A_i \tilde{x}_i + B_i \eta_i) - \Bigg[A_i \hat{x}_i^s + B_i u_i + \sum_{j \in \mathcal{N}_i} A_{ij} (A_j \hat{x}_j^{d-} \\ & + B_j u_j^-) + \sum_{l \in \bar{\mathcal{N}}_i} M_{il} \hat{x}_l^{d-} + L_i (C_i x_i + v_i - C_i \tilde{x}_i - C_i \hat{x}_i^s) \Bigg]. \end{aligned} \tag{4.57}$$

After a simple computation, (4.57) results in:

$$\begin{aligned} \tilde{\epsilon}_i^{s+} =& (A_i - L_i C_i) \tilde{\epsilon}_i^s + w_i - L_i v_i + \sum_{j \in \mathcal{N}_i} A_{ij} w_j^- + \sum_{j \in \mathcal{N}_i} A_{ij} A_j x_j^- \\ & - \sum_{j \in \mathcal{N}_i} A_{ij} A_j \hat{x}_j^{d-} + \sum_{j \in \mathcal{N}_i} A_{ij} \sum_{l \in \mathcal{N}_j} A_{jl} x_l^- - \sum_{l \in \bar{\mathcal{N}}_i} M_{il} \hat{x}_l^{d-}. \end{aligned} \tag{4.58}$$

From (4.54), (3.15), and (4.51), the thesis easily follows. □

In such a way, the computed error $\tilde{\epsilon}_i^s$ is sensitive to attacks in both $\mathcal{S}_j, j \in \mathcal{N}_i$ (neighboring subsystems), and $\mathcal{S}_l, l \in \mathcal{N}_i$ (second-order neighboring subsystems). On the one hand, $\mathcal{S}_h$ is a second-order neighbor of $\mathcal{S}_i$; on the other hand, $\mathcal{S}_i$ is itself in $\bar{\mathcal{N}}_i$ due to Assumption 3.1.2. Therefore, in lines with the strategy of Subsection 4.2.2, this strategy solves the isolation problem by considering the residual quantity:

$$\tilde{r}_i^s \doteq g_i^\top (\tilde{y}_i - C_i \hat{x}_i^s), \tag{4.59}$$

where the filter $g_i \in \mathbb{R}^{p_i}$ is designed so that either:

$$g_i^\top C_i M_{ii} = 0 \tag{4.60a}$$

$$g_i^\top C_i M_{ih} \neq 0, \tag{4.60b}$$

or:

$$g_i^\top C_i M_{ii} \neq 0 \tag{4.61a}$$
$$g_i^\top C_i M_{ih} = 0, \tag{4.61b}$$

that is the residual $\tilde{r}_i^s$ is sensitive to attacks in only one between $\mathcal{S}_i$ and $\mathcal{S}_h$.

Remark 4.2.5. *In order to implement a two-step Luenberger observer $\mathcal{O}_i^s$, a significant additional communication effort is required. Indeed, the logic unit LU_i of subsystem $\mathcal{S}_i$ needs to have knowledge of the following information:*

- ***offline**: the couple $(A_j, B_j), \forall j \in \mathcal{N}_i$, and the structure of the interconnections $A_{jl}, \forall j \in \mathcal{N}_i, \forall l \in \bar{\mathcal{N}}_i$; this is needed to compute the matrices M_{il}, and to design the gain g_i;*
- ***online**: the decentralized estimate and the actuation signal of each of its neighbors $(\hat{x}_j^d, u_j), \forall j \in \mathcal{N}_i)$, and the decentralized estimate of each of its second-order neighbors $\hat{x}_l^d, \forall l \in \bar{\mathcal{N}}_i$, to produce the partial result to be stored in a buffer. This is needed to dynamically compute the estimate $\hat{x}_i^s$, together with its actuation and measurement signals $(u_i, \tilde{y}_i)$.*

Moreover, the next proposition provides a suitable threshold to be used when monitoring the norm of the residual $\tilde{r}_i^s$, in order to avoid false alarms caused by the presence of the noise.

Proposition 4.2.3. *Let suppose Assumption* (4.32) *holds, and let L_i be designed so that F_i^s is a scalar matrix as in* (4.37)*. Given* (3.5)*, and a bound on the distributed true error at time $k = 0$, $\bar{x}_i(0)$, in attack-free conditions, the norm of the filtered two-step Luenberger residue $\|\tilde{r}_i^s\|$ is bounded by the positive function $\bar{r}_i^s$, which can be initialized as:*

$$\begin{cases} \bar{r}_i^s(0) = \bar{x}_i(0) \\ \bar{r}_i^s(1) = |\lambda_i| \bar{x}_i(0) + \bar{w}_i + \|L_i\| \, \bar{v}_i, \end{cases} \tag{4.62}$$

and evolves according to:

$$\bar{r}_i^{s+} = |\lambda_i| \bar{r}_i^s + Q_i^s + s_i^-, \tag{4.63}$$

where:

$$Q_i^s \doteq \bar{w}_i + \|L_i\| \, \bar{v}_i + \sum_{j \in \mathcal{N}_i} \|A_{ji}\| \, \bar{w}_j, \tag{4.64}$$

and:

$$s_i(k) \doteq \sum_{j \in \mathcal{N}_i} \left\| g_i^\top C_i A_{ij} \right\| \bar{\epsilon}_j^d(k) + \sum_{l \in \bar{\mathcal{N}}_i} \left\| g_i^\top C_i M_{il} \right\| \bar{\epsilon}_l^d(k), \tag{4.65}$$

being $\bar{\epsilon}_j^d$ define as in Proposition 3.3.3.

Proof. If F_i^s is a scalar matrix with eigenvalues λ_i, by convolving (4.53), one obtains:

$$\begin{aligned}\tilde{\epsilon}_i^s(k) =&\lambda_i^{k-1}\tilde{\epsilon}_i^s(1) + \sum_{\tau=1}^{k-1}\lambda_i^{k-1-\tau}\Big[w_i(\tau) - L_i v_i(\tau) + \sum_{j\in\mathcal{N}_i} A_{ji}w_j(\tau-1)\\ &+ \sum_{j\in\mathcal{N}_i} A_{ij}\epsilon_j^d(\tau-1) + \sum_{l\in\bar{\mathcal{N}}_i} M_{il}\epsilon_l^d(\tau-1)\Big], \forall k\in\mathbb{Z}, k>1,\end{aligned} \tag{4.66}$$

where:

$$\tilde{\epsilon}_i^s(1) = \lambda_i\tilde{\epsilon}_i^s(0) + w_i(0) - L_i v_i(0). \tag{4.67}$$

Then, $\forall k\in\mathbb{Z}, k>1$, the residual defined in (4.59) is:

$$\begin{aligned}\tilde{r}_i^s(k) =&g_i^\top C_i\lambda_i^{k-1}\tilde{\epsilon}_i^s(1) + \sum_{\tau=1}^{k-1}\lambda_i^{k-1-\tau}\Big[g_i^\top C_i\Big(w_i(\tau) - L_i v_i(\tau)\\ &+ \sum_{j\in\mathcal{N}_i} A_{ji}w_j(\tau-1)\Big) + \sum_{j\in\mathcal{N}_i} g_i^\top C_i A_{ij}\epsilon_j^d(\tau-1)\\ &+ \sum_{l\in\bar{\mathcal{N}}_i} g_i^\top C_i M_{il}\epsilon_l^d(\tau-1)\Big].\end{aligned} \tag{4.68}$$

Therefore, by using the triangle inequality, the relations (3.46b) and (3.46a), and the bounds (3.5), we have:

$$\|\tilde{r}_i^s(k)\| \le|\lambda_i|^{k-1}\left\|g_i^\top C_i\right\|\|\tilde{\epsilon}_i^s(1)\| + \sum_{\tau=1}^{k-1}|\lambda_i|^{k-1-\tau}\Big[\left\|g_i^\top C_i\right\|Q_i^s + s_i(\tau-1)\Big], \tag{4.69}$$

where Q_i^s and $s_i(k)$ are defined as in (4.64) and (4.65), respectively. Let $\bar{r}_i^s$ be the expression in right-hand side of the previous equation. One can easily prove that:

$$\begin{aligned}\bar{r}_i^s(k+1) =&|\lambda_i|^k\left\|g_i^\top C_i\right\|\|\tilde{\epsilon}_i^s(1)\| + \sum_{\tau=1}^{k}|\lambda_i|^{k-\tau}\Big[\left\|g_i^\top C_i\right\|Q_i^s + s_i(\tau-1)\Big]\\ =&|\lambda_i||\lambda_i|^{k-1}\left\|g_i^\top C_i\right\|\|\tilde{\epsilon}_i^s(1)\| + Q_i^s + s_i(k-1)\\ &+ \lambda_i\sum_{\tau=1}^{k-1}|\lambda_i|^{k-1-\tau}\Big[\left\|g_i^\top C_i\right\|Q_i^s + s_i(\tau-1)\Big]\\ =&|\lambda_i|\bar{r}_i^s(k) + Q_i^s + s_i(k-1).\end{aligned} \tag{4.70}$$

Finally, by inspection initialization (4.62) can be verified to be feasible if the observer is initialized with the null initial condition $\hat{x}_i^s(0) = \hat{x}_i^s(1) = 0$, which is a reasonable choice in absence of further information. □

To conclude this section, the following proposition presents a discussion on the condition about the existence of such a vector g_i satisfying either (4.60) or (4.61). In light of Remark 4.2.5, the outline will be focused only on those configurations in which the strategy of Subsection 4.2.2 cannot be applied.

Proposition 4.2.4. *In general, the condition preventing the effectiveness of the filtered Luenberger residual technique, namely:*

$$\mathrm{Im}(A_{ji}) = \mathrm{Im}(A_{jh}), \forall j \in \mathcal{N}_i = \mathcal{N}_h, \tag{4.71}$$

does not prevent the filtered two-step Luenberger residual technique from successfully isolating the attack.

Proof. A vector $g_i \in \mathbb{R}^{n_i}$ satisfying either (4.60) exists if and only if:

$$\ker((C_i M_{ii})^\top) \neq \ker((C_i M_{ih})^\top). \tag{4.72}$$

After simple manipulations analogous to (4.44) and (4.45), (4.72) is found equivalent to:

$$\mathrm{Im}(C_i M_{ii}) \neq \mathrm{Im}(C_i M_{ih}). \tag{4.73}$$

Given the definition (4.51), we observe that:

$$\mathrm{Im}(M_{il}) \subseteq \mathrm{Im}(\Xi_i), \forall l \in \bar{\mathcal{N}}_i. \tag{4.74}$$

Therefore, taken (3.22) and Proposition A.0.1 into account, one obtains:

$$\mathrm{rank}\left(C_i \left[\begin{array}{c|c} M_{ii} & M_{ih} \end{array}\right]\right) = \mathrm{rank}\left(\left[\begin{array}{c|c} M_{ii} & M_{ih} \end{array}\right]\right). \tag{4.75}$$

As a consequence, (4.73) is equivalent to:

$$\mathrm{Im}(M_{ii}) \neq \mathrm{Im}(M_{ih}), \tag{4.76}$$

that is, via (4.51):

$$\mathrm{Im}\left(\sum_{j\in\mathcal{N}_i} A_{ij}A_{ji}\right) \neq \mathrm{Im}\left(\sum_{j\in\mathcal{N}_i} A_{ij}A_{jh}\right). \tag{4.77}$$

Finally, we observe that condition (4.71) implies:

$$\mathrm{Im}(A_{ij}A_{ji}) = \mathrm{Im}(A_{ij}A_{jh}), \forall j \in \mathcal{N}_i = \mathcal{N}_h. \tag{4.78}$$

Nonetheless, given any two matrices of the same size, in general it is:

$$\mathrm{Im}(A + B) \neq \mathrm{Im}(A) + \mathrm{Im}(B), \tag{4.79}$$

where the symbol + in the right-hand side of the equation is to be understood as the sum of linear subspaces. Therefore, chances are that (4.77) could hold despite of (4.71), that is there might be configurations in which this filtered two-step Luenberger observer allows for a discrimination to solve the ambiguity, whereas the filtered single Luenberger observer does not. An example to clarify this fact is given in Subsection 4.3.2. □

Remark 4.2.6. *Observe that this analysis proves that there might exist configurations in which, even though the two candidate attacked subsystems influence the same subspaces* $\mathrm{Im}(A_{ji}) = \mathrm{Im}(A_{jh})$ *of the same agents* $\mathcal{N}_i = \mathcal{N}_h$*, by exploiting the topology of the network* (4.51)*, it is still possible to discriminate and isolate potential attacks. In practice, if the considered subsystems have high dimensions and the interconnections have low rank, the matrices* M_{il} *are likely very sparse, hence* (4.77) *might easily hold (some of the* M_{il} *might even be zero).*

Remark 4.2.7. *Theoretically, other multi-step Luenberger observers could be designed in the same fashion, and chances are that they might overcome some of the configurations in which the ambiguity cannot be solved by implementing the two-step Luenberger observer. Nonetheless, at each step the required information exchange dramatically increases as well, making the strategy unfeasible in practice.*

Similarly, under particular circumstances it may be useful to implement $\mathcal{O}_j^s$*, the two-step Luenberger observer for a common neighbor* $\mathcal{S}_j, j \in \mathcal{N}_i = \mathcal{N}_h$*, and to filter its residual. Indeed, in such a case* $\tilde{\epsilon}_j^s$ *depends on the decentralized true errors* $\tilde{\epsilon}_i$ *and* $\tilde{\epsilon}_h$ *through the gains* $A_{ji}A_i$ *and* $A_{jh}A_h$*, respectively. Even tough* $\mathrm{Im}(A_{ji}) = \mathrm{Im}(A_{jh})$*, hypothetically it may happen that* $\mathrm{Im}(A_{ji}A_i) \neq \mathrm{Im}(A_{jh}A_h)$*, for low-rank update matrices* A_i *and* A_h*. Nonetheless, this is not very likely. For example, if none of the eigenvalues of* A_i *or* A_h *is zero, this cannot happen.*

4.3 Comparison on the requirements of the proposed isolation strategies

In this section, the conditions required to implement the techniques proposed in Section 4.2 are compared. Firstly, Subsection 4.3.1 compares the prerequisite for the existence of an effective UIO for a merged subsystem with those for the efficacy of the filtered Luenberger residual approach. Secondly, Subsection 4.3.2 provides a numerical example to clarify the considerations in Subsection 4.2.3, i.e. there are some configurations in which the two-step Luenberger observer can effectively solve the ambiguity, even though the other two isolation strategies proposed in this chapter fail.

4.3.1 UIO of the merged subsystem and filtered Luenberger residual

The intent of this subsection is to prove that the UIO of the merged subsystem and the filtered Luenberger residual can be applied under independently conditions.

For sake of simplicity, for each common neighbor $\mathcal{S}_j, j \in \mathcal{N}_i = \mathcal{N}_h$, let adopt the following notation:

$$\Xi_j = \left[\begin{array}{c|c} A_{ji} & \breve{\Xi}_j \end{array}\right] \doteq \left[\begin{array}{c|c|c} A_{ji} & A_{jh} & \breve{\breve{\Xi}}_j \end{array}\right]. \tag{4.80}$$

4.3. COMPARISON ON THE REQUIREMENTS OF THE PROPOSED ISOLATION STRATEGIES

Firstly, let consider a scenario such that, for a given common neighboring subsystem $\mathcal{S}_j, j \in \mathcal{N}_i = \mathcal{N}_h$, C_j is full column-rank, and (4.16) holds. If, furthermore:

$$\mathrm{Im}(A_{ji}) \subset \mathrm{Im}\left(\left[\begin{array}{c|c} A_{jh} & \breve{\breve{\Xi}}_j \end{array} \right] \right) \tag{4.81a}$$

$$\mathrm{Im}(A_{jh}) \subset \mathrm{Im}\left(\left[\begin{array}{c|c} A_{ji} & \breve{\breve{\Xi}}_j \end{array} \right] \right), \tag{4.81b}$$

then, the UIO of a merged subsystem is insensitive to both attacks. Nonetheless, (4.81a) and (4.81b) do not imply $\mathrm{Im}(A_{ji}) = \mathrm{Im}(A_{jh})$ in general, therefore the filtered Luenberger technique could still be viable.

Conversely, one might think that if the filtered Luenberger technique cannot be effectively implemented, neither can the UIO of a merged subsystem. Indeed, let assume the former cannot be implemented, that is $\mathrm{Im}(A_{ji}) = \mathrm{Im}(A_{jh})$. Recalling that $\mathcal{S}_i$ and $\mathcal{S}_h$ are not neighbors (otherwise there would be no ambiguity at all), we would then have (up to a permutation):

$$\Xi_{j\cup i} = \left[\begin{array}{c|c} A_{jh} & \breve{\breve{\Xi}}_j \\ \hline 0 & \breve{\Xi}_i \end{array} \right], \tag{4.82}$$

where the notation of (4.80) was adopted for $\mathcal{S}_j$. As a consequence, (4.12a) gives:

$$\begin{aligned} 0 =& (H_{j\cup i} C_{j\cup i} - I)\Xi_{j\cup i} \\ =& \left(\left[\begin{array}{c|c} H_{jj} & H_{ji} \\ \hline H_{ij} & H_{ii} \end{array} \right] \begin{bmatrix} C_j & 0 \\ 0 & C_i \end{bmatrix} - I \right) \Xi_{j\cup i} \\ =& \left[\begin{array}{c|c} H_{jj}C_j - I & H_{ji}C_i \\ \hline H_{ij}C_j & H_{ii}C_i - I \end{array} \right] \left[\begin{array}{c|c} A_{jh} & \breve{\breve{\Xi}}_j \\ \hline 0 & \breve{\Xi}_i \end{array} \right] \\ =& \left[\begin{array}{c|c} (H_{jj}C_j - I)A_{jh} & \star \\ \hline H_{ij}C_jA_{jh} & \star \end{array} \right]. \end{aligned} \tag{4.83}$$

Therefore, any solution $H_{j\cup i}$ of (4.12a), necessarily satisfies:

$$\mathrm{Im}(A_{ji}) = \mathrm{Im}(A_{jh}) \subseteq \ker\left(\left[\begin{array}{c} H_{jj}C_j - I \\ \hline H_{ij}C_j \end{array} \right] \right), \tag{4.84}$$

that is (4.28) holds, and the merged UIO is not effective. Nonetheless, the filtered Luenberger residual can be exploited only if the output matrix C_j is full column-rank, which is a stronger requirement than the rank condition for the existence of a UIO (3.22). As a result, there might exist configurations in which the merged UIO is effective, even though the filtering technique cannot be applied.

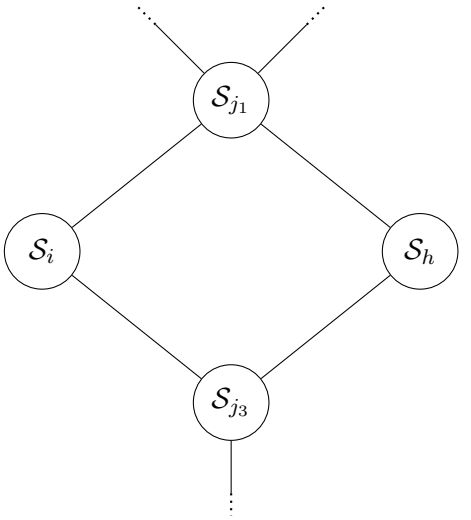


Figure 4.4: A simple example of topology where the two-step Luenberger observer strategy can effectively solve the isolation problem.

4.3.2 Numerical example of the capability of the two-step Luenberger observer

In this subsection, a simple numerical example is proposed to show that the two-step Luenberger observer technique can effectively be applied in some of those configurations in which both the merged UIO and the filtered Luenberger residue fail.

Let consider the topology depicted in Figure 4.4. Let assume C_{j_1} and C_{j_2} are full column-rank matrices. Moreover, assume the interconnections are:

$$\begin{aligned} A_{j_1 i} &= \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad A_{j_1 h} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \\ A_{j_2 i} &= \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \quad A_{j_2 h} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \\ A_{i j_1} &= \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad A_{i j_2} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}. \end{aligned} \tag{4.85}$$

Clearly, the following holds:

$$\begin{aligned} \mathrm{Im}(A_{j_1 i}) &= \mathrm{Im}(A_{j_1 h}) \\ \mathrm{Im}(A_{j_2 i}) &= \mathrm{Im}(A_{j_2 h}). \end{aligned} \tag{4.86}$$

As a consequence, the filtered Luenberger residue strategy is not effective; similarly, the merged UIO cannot be implemented, see Subsection 4.3.1.

Nonetheless, from (4.51) we have:

$$M_{ii} = A_{i j_1} A_{j_1 i} + A_{i j_2} A_{j_2 i} = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}, \tag{4.87}$$

and:

$$M_{ih} = A_{ij_1} A_{j_1 h} + A_{ij_2} A_{j_2 h} = \begin{bmatrix} 0 & 2 \\ 0 & 1 \end{bmatrix}. \tag{4.88}$$

Since it is:

$$\mathrm{Im}(M_{ii}) \neq \mathrm{Im}(M_{ih}), \tag{4.89}$$

then the two-step Luenberger Observer can effectively resolve the ambiguity, in line with Subsection 4.2.3.

Chapter 5

Simultaneous multiple attacks

This chapter addresses an introductory discussion on the scenarios of more capable and more resourceful attackers, who simultaneously manipulate multiple subsystems within the same neighborhood. By relaxing Assumption 3.3.2, the whole detection architecture is to be reconsidered, as the validity of the argument cannot be taken for granted.

Generally speaking, when two different subsystems in the same neighborhood are simultaneously attacked, the detection strategy fails only if specific conditions are met. As a consequence, if two malicious agents try to manipulate the network independently one other, the detection algorithm of Chapter 3 is likely to work as well.

The chapter consists of two sections. Section 5.1 explores the possibility of a coordinate attack in two subsystems sharing a neighbor so that there is a compensation of the effect on such a neighbor. Conversely, Section 5.2 addresses a study on the potential of an attacker simultaneously affecting two neighboring subsystems.

5.1 Coordinate attack in two subsystems sharing a neighbor

It might be interesting for an attacker to coordinately affect two subsystems within the same neighborhood in order to compensate the effect of the attacks one another. Indeed, the detection strategy presented in Chapter 3 relies on the fact that the neighboring subsystems of the attacked one are sensitive to the attack itself. Nonetheless, if two subsystems share a neighbor and are simultaneously attacked, chances are that their effects on the common neighbor might cancel out. Were this the case, the detection strategy would fail. Indeed, if the effect of the attacks is perfectly counterbalanced, then the common neighbor's communicated residue keeps following the nominal trajectory, therefore the alarm signal is not raised. As a consequence, both the attacked subsystems

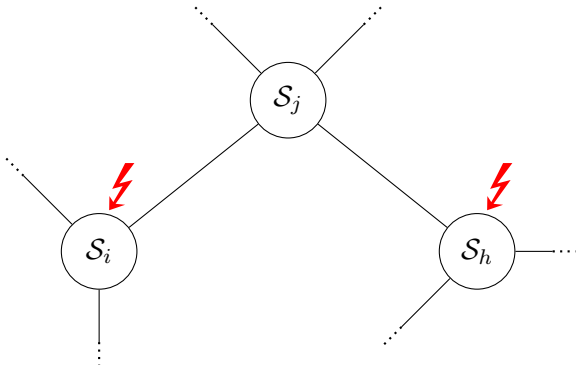


Figure 5.1: Detail of the neighborhood topology. $\mathcal{S}_i$ and $\mathcal{S}_h$ are attacked, while $\mathcal{S}_j$ is a common neighbor.

do not claim to be under attack. In the following, the precise details of this scenario are given.

Let $\mathcal{S}_i$ and $\mathcal{S}_h$ be two attacked subsystems, and let $\mathcal{S}_j$ be a common neighbor (see Figure 5.1). Let suppose the following hypothesis is granted.

Assumption 5.1.1. *The attacker is affecting $\mathcal{S}_i$ and $\mathcal{S}_h$ as in* (3.6)*, while all the others subsystems are safe.*

This assumption is considered for sake of simplicity. On the other hand, all the results of this section can be easily extended to more complex situations in which such hypothesis is relaxed.

The goal of an attacker willing to avoid the alarm signal a_j of $\mathcal{S}_j$ from being raised is to design $\eta_i \in \mathbb{R}^{m_i}$ and $\eta_h \in \mathbb{R}^{m_h}$ (and, consequently, $\gamma_i \in \mathbb{R}^{p_i}$ and $\gamma_h \in \mathbb{R}^{p_h}$) in order to prevent the norm of the computed distributed residue $\left\|\tilde{r}_j^c\right\|$ from exceeding the associated threshold $\bar{r}_j^c$.

In detail, from (3.37), we have:

$$\tilde{\epsilon}_j^{c+} = F_j^c \tilde{\epsilon}_j^c + w_j - L_j v_j + \sum_{l \in \mathcal{N}_j} A_{jl} \epsilon_l^d. \tag{5.1}$$

As a consequence, the residue is insensitive to the attacks in $\mathcal{S}_i$ and $\mathcal{S}_h$ if they are designed so that:

$$A_{ji} \epsilon_i^d + A_{jh} \epsilon_h^d = 0. \tag{5.2}$$

Let $\mathcal{E}_l \subseteq \mathbb{R}^{n_l}$ be the reachable space of the attacker $\mathcal{A}_l$, namely the set of all possible values where it can move the decentralized true error ϵ_l^d of subsystem $\mathcal{S}_l$, in absence of noise and for the zero initial condition. A requirement for condition (5.2) is:

$$\dim \left(\operatorname*{Im}_{A_{ji}}(\mathcal{E}_i) \cap \operatorname*{Im}_{A_{jh}}(\mathcal{E}_h) \right) > 0, \tag{5.3}$$

i.e. the influence of the two subsystems on the common neighbor $\mathcal{S}_j$ can (at least partially) overlap.

In order to complete the analysis, let take a look at the structure of the attacker reachable set $\mathcal{E}_l$. From (3.27), by assuming $w_l = 0, v_l = 0$, we have:

$$\epsilon_l^{d+} = F_l^d \epsilon_l^d + B_l \eta_l + (A_l - F_l^d)\tilde{x}_l. \tag{5.4}$$

Therefore, one would assert that the attacker reachable set $\mathcal{E}_l$ satisfies:

$$\mathcal{E}_l \supseteq \mathrm{Im}\left(\left[\begin{array}{c|c|c|c} B_l & (F_l^d)B_l & \dots & (F_l^d)^{n_l-1}B_l \end{array} \right] \right) = \mathcal{R}_{(F_l^d, B_l)}, \tag{5.5}$$

where $\mathcal{R}_{(F_l^d,B_l)}$ is the reachable set of the pair (F_l^d, B_l), meaning that for any reachable value by using η_l only, there exists a sequence of η_l and $\tilde{x}_l$ which drives the decentralized true error to that value (for example, by choosing $\tilde{x}_i$ identically zero). Nonetheless, this argument is incorrect, since $\tilde{x}_l$ is a function of η_l itself and cannot be chosen independently of it. As a consequence, (5.5) is not verified, in general.

The exact characterization of $\mathcal{E}_l$ is provided by the next theorem.

Theorem 5.1.1. *Let assume a malicious agent $\mathcal{A}_l$ is affecting subsystem $\mathcal{S}_l$ as in* (3.6)*. In absence of noise and with zero initial condition, the reachable set within which an attacker can move the decentralized true error ϵ_l^d is:*

$$\mathcal{E}_l = \mathrm{Im}\left(\left[\begin{array}{c|c|c|c} B_l & A_l B_l & \dots & A_l^{n_l-1}B_l \end{array} \right] \right) = \mathcal{R}_{(A_l,B_l)}, \tag{5.6}$$

that is it coincides with the reachable space of the couple (A_l, B_l).

Proof. Let assume $\tilde{x}_l(0)$. The attacker state $\tilde{x}_l$ can be so written as a function of the attacker control input η_l:

$$\tilde{x}_l(k) = \sum_{\tau=0}^{k-1} A_l^{k-1-\tau} B_l \eta_l(\tau). \tag{5.7}$$

On the other hand, by neglecting the noise, and by assuming zero initial condition, from (5.4) and (5.7), the decentralized true error is obtained:

$$\begin{aligned} \epsilon_l^d(k) = &\sum_{\tau=0}^{k-1} (F_l^d)^{k-1-\tau} \left[B_l \eta_l(\tau) + (A_l - F_l^d) \sum_{t=0}^{\tau-1} A_l^{\tau-1-t} B_l \eta_l(t) \right] \\ &\sum_{\tau=0}^{k-1} (F_l^d)^{k-1-\tau} B_l \eta_l(\tau) + \sum_{\tau=0}^{k-1} \sum_{t=0}^{\tau-1} (F_l^d)^{k-1-\tau} (A_l - F_l^d) A_l^{\tau-1-t} B_l \eta_l(t). \end{aligned} \tag{5.8}$$

Let observe that double summation can be so rearranged:

$$\begin{aligned}
&\sum_{\tau=0}^{k-1}\sum_{t=0}^{\tau-1}(F_l^d)^{k-1-\tau}(A_l - F_l^d)A_l^{\tau-1-t}B_l\eta_l(t) \\
=&\sum_{t=0}^{k-2}\left(\sum_{\tau=t+1}^{k-1}(F_l^d)^{k-1-\tau}(A_l - F_l^d)A_l^{\tau-1-t}\right)B_l\eta_l(t) \\
=&\sum_{t=0}^{k-2}(A_l^{k-1-t} - (F_l^d)^{k-1-t})B_l\eta_l(t),
\end{aligned} \tag{5.9}$$

where Proposition B.0.1 of Appendix B was exploited. Moreover, we can extend the summation without altering the overall value:

$$\sum_{t=0}^{k-2}(A_l^{k-1-t} - (F_l^d)^{k-1-t})B_l\eta_l(t) = \sum_{t=0}^{k-1}(A_l^{k-1-t} - (F_l^d)^{k-1-t})B_l\eta_l(t), \tag{5.10}$$

as the summand is zero for $t = k-1$.

As a consequence, we have:

$$\begin{aligned}
\epsilon_l^d(k) =&\sum_{\tau=0}^{k-1}(F_l^d)^{k-1-\tau}B_l\eta_l(\tau) + \sum_{\tau=0}^{k-1}(A_l^{k-1-\tau} - (F_l^d)^{k-1-\tau})B_l\eta_l(\tau) \\
=&\sum_{\tau=0}^{k-1}A_l^{k-1-\tau}B_l\eta_l(\tau),
\end{aligned} \tag{5.11}$$

and the thesis easily follows from standard considerations from the reachability analysis, see [24, Section 5.1]. □

Remark 5.1.1. *Equation* (5.11) *confirms and formally states the intuitive idea that the influence of the attacker $\mathcal{A}_l$ on the decentralized true error ϵ_l^d is the attacker state $\tilde{x}_l$ itself. As a consequence, one deduce that an attacker aiming to perform this coordinate attack strategy by affecting $\mathcal{S}_i$ and $\mathcal{S}_h$ does not need any knowledge of matrices F_i^d, F_h^d, since the couples $(A_i, B_i), (A_h, B_h)$ are themselves sufficient to determine the attacker reachable spaces $\mathcal{E}_i, \mathcal{E}_h$ and the actuation signals η_i, η_h. Observe this is a crucial consideration for the attacker, since matrices F_i^d, F_h^d are only known via software and they cannot be identified by eavesdropping on the actuation and measurement signals. On the other hand, the attacker needs to have knowledge of the interconnection matrices A_{ji}, A_{jh}.*

The problem of finding an effective detection strategy for this type of attack remains open. We just observe that the use of a filtered residue as in the isolation case (see Subsection 4.2.2) would not be successful, since according to (5.2) the attacker effect on the common neighbors would necessarily lie within the common range of the interconnection matrices A_{ji}, A_{jh}, causing the strategy to fail in lines with Section 4.2.2.

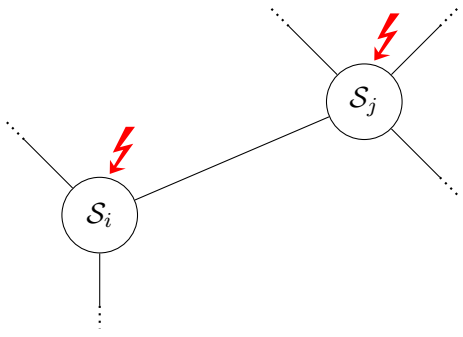


Figure 5.2: Detail of the neighborhood topology. $\mathcal{S}_i$ and $\mathcal{S}_j$ are neighbors, both under attack.

5.2 Coordinate attack in two neighboring subsystems

In this section a further kind of coordinate attack is examined. Specifically, an attacker might affect two neighboring subsystem so that neither of them is capable of recognizing the attack on the other one, therefore making the detection strategy unsuccessful. In the following, firstly an outline of the conditions to develop this coordinate attack is presented, secondly some simple detection strategies are briefly proposed.

Let $\mathcal{S}_i$ and $\mathcal{S}_h$ be two attacked neighboring subsystems, as depicted in Figure 5.2. In order to simplify the discussion, let make the following assumption.

Assumption 5.2.1. *The attacker is affecting $\mathcal{S}_i$ and $\mathcal{S}_j$ as in* (3.6)*, whereas all the others subsystems are safe.*

The next theorem provides sufficient conditions on the design of the attacker signals so that the attack is covert for both the subsystems. Note that the attack at $\mathcal{S}_j$ is not modeled as in Section 3.2 (whereas the one affecting $\mathcal{S}_i$ is).

Theorem 5.2.1. *Let assume a malicious agent is affecting $\mathcal{S}_i$ and $\mathcal{S}_j$, neighboring subsystems, as in* (3.6)*. Let the attacker's manipulation in $\mathcal{S}_i$ be modeled as in* (3.7)*. If the attacker designs $\eta_j \in \mathbb{R}^{m_j}$ and $\gamma_j \in \mathbb{R}^{p_j}$ satisfying:*

$$H_j\gamma_j^+ = (T_jL_j - K_j^{(1)})\gamma_j + T_jA_{ji}\epsilon_i^d \tag{5.12a}$$

$$B_j\eta_j = -A_{ji}\epsilon_i^d - L_j\gamma_j, \tag{5.12b}$$

then the attack cannot be detected by using the architecture presented in Chapter 3.

Proof. In order for the coordinate attack to be undetectable, the attacker signals η_j and γ_j must be designed to compensate for the influence of the decentralized error ϵ_i^d on the distributed error $\tilde{\epsilon}_j^c$.

Observe that, since the attack in $\mathcal{S}_j$ is not modeled as in (3.7), (3.37) does not hold. Nonetheless, from (3.36), we have:

$$\epsilon_j^{c+} = F_j^c \epsilon_j^c + w_j - L_j v_j + \sum_{l \in \mathcal{N}_j} A_{jl} \epsilon_l^d + B_j \eta_j + L_j \gamma_j. \tag{5.13}$$

Therefore, the attacker signals must satisfy:

$$A_{ji} \epsilon_i^d + B_j \eta_j + L_j \gamma_j = 0. \tag{5.14}$$

On the other hand, if η_j and γ_j are not designed according to (3.7), one needs to ensure that the attack cannot be sensed by the UIO $\mathcal{O}_j^d$. Observe that neither (3.27) holds. Nonetheless, from (3.6), (3.15), (3.19), and (3.1), we have:

$$\begin{aligned}
\epsilon_j^{d+} =& x_j^+ - \hat{x}_j^{d+} \\
=& x_j^+ - z_j^+ - H_j(C_j x_j^+ + v_j^+ - \gamma_j^+) \\
=& (I - H_j C_j)[A_j x_j + B_j(u_j + \eta_j) + \Xi_j \boldsymbol{x}_j + w_j] \\
& - [F_j z_j + T_j B_j u_j + K_j(C_j x_j + v_j - \gamma_j)] - H_j v_j^+ + H_j \gamma_j^+.
\end{aligned} \tag{5.15}$$

By exploiting (3.23), (3.25a), and (3.25b), one obtains:

$$\begin{aligned}
\epsilon_j^{d+} =& [\bar{A}_j - K_j C_j] x_j - F_j z_j + T_j w_j - K_j v_j \\
& - H_j v_j^+ + T_j B_j \eta_j + K_j \gamma_j + H_j \gamma_j^+ \\
=& [\bar{A}_j - K_j^{(1)} C_j - K_j^{(2)} C_j] x_j - F_j z_j + T_j w_j - K_j^{(1)} v_j - K_j^{(2)} v_j \\
& - H_j v_j^+ + T_j B_j \eta_j + K_j^{(1)} \gamma_j + K_j^{(2)} \gamma_j + H_j \gamma_j^+,
\end{aligned} \tag{5.16}$$

where (3.24) was used. Finally, from (3.25c), (3.25d),and (3.15), we have:

$$\begin{aligned}
\epsilon_j^{d+} =& F_j(x_j - z_j - H_j(C_j x_j + v_j - \gamma_j)] + T_j w_j - K_j^{(1)} v_j \\
& - H_j v_j^+ + T_j B_j \eta_j + K_j^{(1)} \gamma_j + H_j \gamma_j^+ \\
=& F_j \epsilon_j^d + T_j w_j - K_j^{(1)} v_j - H_j v_j^+ + T_j B_j \eta_j + K_j^{(1)} \gamma_j + H_j \gamma_j^+.
\end{aligned} \tag{5.17}$$

Therefore, it is needed:

$$T_j B_j \eta_j + K_j^{(1)} \gamma_j + H_j \gamma_j^+ = 0. \tag{5.18}$$

The thesis follows by observing that imposing (5.14) and (5.18) is equivalent to require to (5.12a) and (5.12b). Observe that the attack at $\mathcal{S}_j$ is covert for the decentralized observer $\mathcal{O}_i^c$ as well. □

Remark 5.2.1. *Observe that the attacker could design ϵ_i^d within $\mathcal{E}_i^R$ (see (5.6)) in a proper way to take advantage when solving* (5.12a) *and* (5.12b)*. On the other hand, it needs to have perfect knowledge of the gains L_j, H_j, T_j, and $K_j^{(1)}$, which in general are not uniquely identifiable given the system structural matrices and interconnections $(A_j, B_j, C_j, A_{jl}, \forall l \in \mathcal{N}_j)$.*

Remark 5.2.2. *Whenever C_j is full-column rank, a feasible solution to* (3.25a) *is $H_j = C_j^{-L}$. Nonetheless, this forces T_j to be zero, making* (5.12a) *and* (5.12b) *significantly easier to solve (specifically, γ_j identically zero is always a solution). Therefore, whenever possible, the choice $H_j = C_j^{-L}$ should be avoided.*

To conclude the dissertation, the existence of a solution (η_j, γ_j) to (5.12a) and (5.12b) is now to be discussed. Moreover, the conditions for the existence of such a solution can be exploited as a detection mechanism, as it will shortly be explained.

Firstly, let observe that (5.12a) is in the form:

$$Ex^+ = Ax + Bu, \tag{5.19}$$

i.e. it is a (discrete-time) linear time-invariant descriptor, a well-known class of dynamical system, whose properties are extensively analyzed in literature [25], [26]. In particular, (5.18) admits a solution for any sufficiently smooth input if and only if the matrix pencil $H_j + \lambda(T_j L_j - K_j^{(1)})$ is regular, that is:

$$\operatorname{rank}(H_j + \lambda(T_j L_j - K_j^{(1)})) = n_j \tag{5.20}$$

for all except a finite number of $\lambda \in \mathbb{C}$. As a consequence, one could intentionally design the UIO gains so that $\det(H_j + \lambda(T_j L_j - K_j^{(1)}))$ identically vanishes.

Conversely, given a certain decentralized true error ϵ_i^d and a particular solution $\gamma_i^\star$ of (5.12a), the existence of a signal $\eta_j^\star$ solving (5.12b) can be trivially discussed by recalling the Rouché-Capelli theorem. From that, we can assess that a solution $\eta_j^\star$ can be found if and only if:

$$A_{ji}\epsilon_i^d(k) + L_j\gamma_j^\star(k) \in \operatorname{Im}(B_j), \forall k \in \mathbb{Z}, k \geq k_{a,i}. \tag{5.21}$$

Nonetheless, an exhaustive discussion on the necessary and sufficient conditions on the involved matrices such that (5.21) admits a solution is far from trivial. We just observe that, if (5.12b) does not admit the trivial solution $\gamma_j^\star$ identically zero, that is $\exists \bar{k} \in \mathbb{Z}, \bar{k} \geq k_{a,i}$ such that $\epsilon_i^d(\bar{k}) \neq \ker(T_j A_{ji})$, then a sufficient condition preventing the existence of a solution $\eta_j^\star$ to (5.12b) is:

$$\dim\left(\operatorname{Im}(B_j) \cap (\operatorname{Im}(A_{ji}) + \operatorname{Im}(L_j))\right) = 0. \tag{5.22}$$

Finally, a different approach could be considered. Let assume that the local unit LU_j implements two distributed observers, $\mathcal{O}_j^{c,(A)}$ and $\mathcal{O}_j^{c,(B)}$, by allocating different eigenvalues through the gains $L_j^{(A)}$ and $L_j^{(B)}$, respectively. The purpose of this second observer is that the attacker must now satisfy the constraint (5.14) for both $L_j^{(A)}$ and $L_j^{(B)}$. Depending on the design, it might happen that the addition of such a condition results in the impossibility for the attacker to design

the signals η_j, γ_j so that all the requirements for the undetectability are met. In other words, the attack cannot simultaneously be covert for both $\mathcal{O}_j^d$, $\mathcal{O}_j^{c,(A)}$, and $\mathcal{O}_j^{c,(B)}$. For example, a sufficient condition for this to happen is that the gains $L_j^{(A)}$ and $L_j^{(B)}$ designed so that:

$$\dim(\mathrm{Im}(L_j^{(A)}) \cap \mathrm{Im}(L_j^{(B)})) = 0. \tag{5.23}$$

Similarly, a second UIO observer could be designed, in the same fashion.

Remark 5.2.3. *Observe that the implementation of an additional observer is entirely within the logic unit and requires no additional burden in the communication. On the other hand, a less trivial discussion should investigate whether different gains can be synthesized so that the constraints on the ranges are satisfied.*

Chapter 6

Data center model

In this chapter a practical context is considered. Firstly, in Section 6.1 a networked system representing a data center is derived. Secondly, Section 6.2 describes some numerical simulations aimed to clarify how the detection and isolation methodologies presented in Chapters 3, 4, and 5 can be implemented in practice.

6.1 Model derivation

In this section, a simple model of a data center is presented. Observe that the aim of the following analysis is to obtain a suitable model to be used for the simulations, allowing to show the effectiveness of the results on detectability and isolation in practice, whereas the derivation of a complete and realistic characterization of a data center is far beyond the scope of this thesis.

Let consider a data center composed of N computational units, each modeled as a discrete-time dynamical system. Let suppose each subsystem $\mathcal{S}_i$ is characterized by a two-dimensional state vector $x_i \doteq \begin{bmatrix} x_{i[1]} & x_{i[2]} \end{bmatrix}^\top$, where $x_{i[1]} \in \mathbb{R}$ is a temperature variable, and $x_{i[2]} \in \mathbb{R}$ expresses the computational effort of processor $\mathcal{S}_i$. Moreover, each processor is monitored by a local unit LU_i, which controls the amount of power $u_i \in \mathbb{R}$ driven into the computational unit itself, according to a feedback policy based on the output measurements $y_i \doteq \begin{bmatrix} y_{i[1]} & y_{i[2]} \end{bmatrix}^\top$.

Concerning the state variables, $x_{i[1]}$ is defined as the difference between the temperature of the computational unit $\mathcal{S}_i$ and the temperature of the room where the data center is located. The latter is assumed to be constant and monitored by a Heating, Ventilation, and Air Cooling System, as recommended in [27], since a long-time exposure of computer equipment to high temperatures greatly reduces reliability, longevity of components, and is likely to cause unplanned downtime. Having said that, the dynamics of $x_{i[1]}$ is modeled as an alternative version of the model in [28], namely:

$$x_{i[1]}^+ = \alpha_i x_{i[1]} + \beta_i x_{i[2]} + \varphi_i u_i + w_{i[1]}, \tag{6.1}$$

where one sees that the temperature evolves according to a first-order dynamics ruled by the parameter $0 \leq \alpha_i \leq 1$, and it is influenced both by the current computational effort $x_{i[2]}$, and by the amount of power entering the processor, through the coefficients $\beta_i > 0$ and $\varphi_i > 0$, respectively. Finally, $w_{i[1]}$ is a bounded noise, accounting for possible fluctuations in the temperature due to model uncertainties.

Remark 6.1.1. *Despite the fact that the temperature is naturally described by a continuous-time dynamics, a discrete-time one was chosen in order to adopt the perspective of a monitoring unit, whose computation are naturally performed in a discrete-time setting.*

On the other hand, the computational load $x_{i[2]}$ is assumed to be mainly due to the arrival of new queries, and the following model is employed:

$$x_{i[2]}^{+} = x_{i[2]} + \mu_i u_i + w_{i[2]}, \tag{6.2}$$

meaning that the computational load proportionally decreases ($\mu_i < 0$) with the power entering the processor, and increases due to the arrival of new queries which are modeled as a truncated Poisson process, and taken into account within the noise $w_{i[2]}$. Moreover, $w_{i[2]}$ also models the quantization noise which is introduced by assuming $x_{i[2]} \in \mathbb{R}$, where it naturally takes values in $\mathbb{N}$. Without loss of generality, one can refer to the ratio between the current computational load and the maximum capacity. Hence, it is assumed $0 \leq x_{i[2]} \leq 1$, and all the related variables are scaled accordingly.

Regarding the output variables, $y_{i[1]}$ is the result of an indirect measurement of the first state component $x_{i[1]}$, thus corrupted by the measurement noise $v_{i[1]}$. Conversely, given its physical interpretation, the computational load variable is directly known via software $y_{i[2]} = x_{i[2]}$, unaffected by measurement noise.

In the same fashion of $x_{i[2]}$, the control signal expresses the amount of supplied power with respect to the maximum possible amount, resulting in $0 \leq u_i \leq 1$. The control input is selected to dynamically tune the trade-off between a large number of enqueued queries, with the risk of rejecting the new incoming ones as a consequence of saturation ($x_{i[2]} = 1$), and a too large increase in the temperature, which possibly affects the performance [29]. More precisely, each local unit LU_i designs the control action u_i according to a Model Predictive Control (MPC) policy based on the local measurements of the two state variables, $y_{i[1]}$ and $y_{i[2]}$, respectively.

Moreover, the computational units are connected one another. Specifically, the influence among state components can be:

- **physical**: when two processors are located close one another, each represents a source of heat for the other. Therefore, their temperatures mutually influence according to some parameter $\delta_{ij}, \delta_{ji} < \alpha_i, \alpha_j$, primarily depending on the distance between the two processors.

- **via software**: each computational unit is connected to some other processors which it shares a certain number ($0 \leq \rho_{ji} \leq 1$) of its new incoming queries with, in order to keep the computational burden balanced, both for efficiency and for temperature's convenience. A simple consideration concerning the principle of conservation of the number of queries gives:

$$\rho_{ii} + \sum_{j \in \mathcal{N}_i} \rho_{ji} = 1, \forall i = 1...N. \tag{6.3}$$

The set $\mathcal{N}_i$ includes the indices of all the subsystems $\mathcal{S}_j$ connected to $\mathcal{S}_i$, either physically or via software. Nonetheless, the interconnections are independent one another, meaning that the redistribution of the queries does not necessarily happen only with the physical neighbors, nor with all of them. Moreover, all the links between subsystems are assumed to be reciprocal.

For sake of simplicity, let assume the following approximation holds.

Assumption 6.1.1. *The processors are physically arranged in a way so that each computational unit is physically connected with two others only.*

Such an hypothesis takes into account the fact that, if two processors are not sufficiently close, their mutual influence as source of heat is negligible in practice.

On the other hand, the virtual links for exchanging the queries can be arbitrarily designed, depending on the different purpose. A sort of trade-off arises. Indeed, a large number of links results in a faster balancing of the computational load, see Appendix C. On the contrary, one may intentionally decide to give up some of the virtual links in order to gain in security, specifically in the isolation capability. This interesting aspect is discussed in detail in Subsection 6.2.2.

Taken the interconnections into account, equations (6.1) and (6.2) can be rewritten in the compact matrix form of a discrete-time linear time-invariant dynamical system:

$$\mathcal{S}_i : \begin{cases} x_i^+ = A_i x_i + B_i u_i + \sum_{j \in \mathcal{N}_i} A_{ij} x_j + w_i \\ y_i = C_i x_i + v_i, \end{cases} \tag{6.4}$$

where:

$$A_i \doteq \begin{bmatrix} \alpha_i & \beta_i \\ 0 & \rho_{ii} \end{bmatrix}, \quad B_i \doteq \begin{bmatrix} \varphi_i \\ \mu_i \end{bmatrix}, \quad C_i \doteq I, \quad A_{ij} \doteq \begin{bmatrix} \delta_{ij} & 0 \\ 0 & \rho_{ij} \end{bmatrix}. \tag{6.5}$$

Furthermore, the system is subject to the following set of linear constraints:

$$0 \leq x_{i[2]} \leq 1 \tag{6.6a}$$

$$0 \leq u_i \leq 1, \tag{6.6b}$$

which are easily taken into account by the Model Predictive Controller. Given the fact that both the model and the set of constraints are linear, the Explicit Model Predictive Control [30] was employed to speed up the computation by pre-computing offline a piece-wise solution. All the numerical values used in the simulations are given in Appendix C.

In such a network, the attacker targets a subsystem and performs a covert-cyber attack as depicted in Section 3.2. The goal of such an attack is to increase the temperature of the processor, by entering more power of that required by the controller. Moreover, by manipulating the measurement output as in (3.6), the attack prevents the local unit LU_i from detecting its action.

Regarding the detection and isolation architecture, all the techniques outlined in Chapters 3, 4, and 5 are considered, and their efficacy will be discussed in detail in Section 6.2. The only fact worth mentioning is that, since the second component of the noise is known to have a positive expected value, the Luenberger observer should be modified. Specifically, one can substitute equation (3.34) with the following unbiased estimator:

$$\mathcal{O}_i^c: \ \hat{x}_i^{c+} = A_i\hat{x}_i^c + B_iu_i + \sum_{j\in\mathcal{N}_i} A_{ij}\hat{x}_j^d + L_i(\tilde{y}_i - C_i\hat{x}_i^c) + \begin{bmatrix} 0 \\ \mathbb{E}[w_{i[2]}] \end{bmatrix}, \tag{6.7}$$

where $\mathbb{E}[w_{i[2]}] \in \mathbb{R}$ is the expected value of $w_{i[2]}$. Let assume $w_{i[2]}$ is a Poisson distribution of parameter $\lambda_{w_{i[2]}}$ [31], scaled by the factor $\frac{1}{w_{i[2]f}}$, truncated at $\bar{w}_{i[2]}$, that is:

$$w_{i[2]} \sim \min\left(\bar{w}_{i[2]}, \frac{1}{w_{i[2]f}}\mathcal{P}(\lambda_{w_{i[2]}})\right). \tag{6.8}$$

As a result, the expected value can be so computed:

$$\mathbb{E}[w_{i[2]}] = \sum_{k=0}^{\lfloor \bar{w}_{i[2]}\cdot w_{i[2]f}\rfloor} \frac{k}{\bar{w}_{i[2]}} p_{\lambda_{w_{i[2]}}}(k) + \bar{w}_{i[2]} \sum_{k=\lfloor \bar{w}_{i[2]}\cdot w_{i[2]f}\rfloor+1}^{+\infty} p_{\lambda_{w_{i[2]}}}(k), \tag{6.9}$$

where $p_\lambda(k)$ is the probability mass function of the Poisson distribution:

$$p_\lambda(k) = \frac{\lambda^k e^{-k}}{k!}. \tag{6.10}$$

Observe that the same correction must be adopted in the implementation of the two-step Luenberger Observer $\mathcal{O}_i^s$.

Finally, observe that the two output measurements do not share the same physical dimension. As a consequence, before computing the residual quantities, both of them are normalized, so that one can conveniently compute the norm of the residual as a dimensionless quantity. Note that this is also useful since the two output components (and, consequently, the residual components) are reasonably quite different in magnitude.

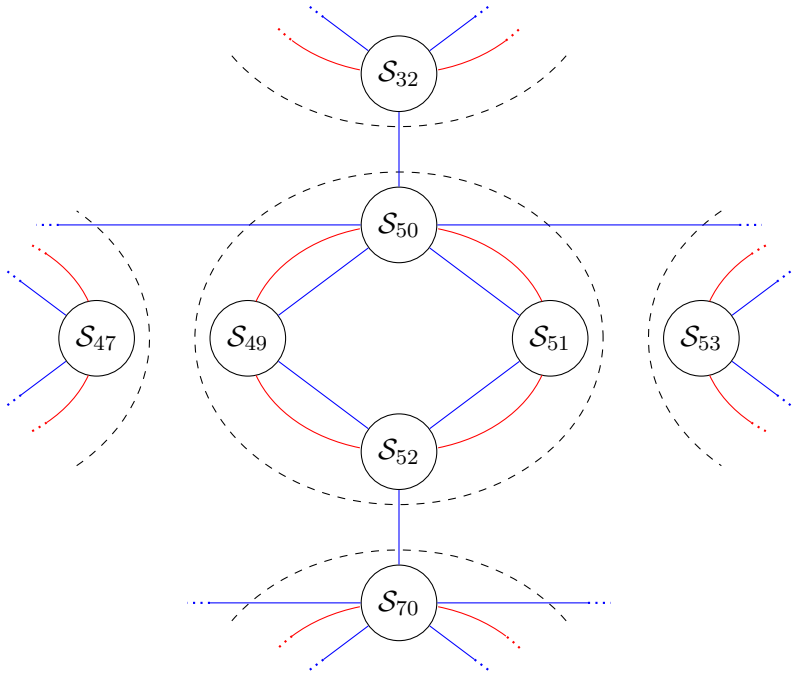


Figure 6.1: Detail of the interconnection graph, specifying the temperature links (red, curved), and the logical links (blue, straight).

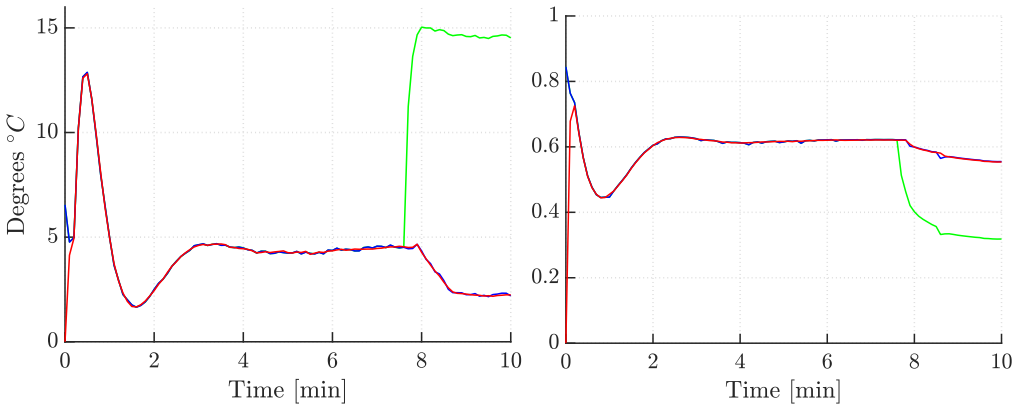
6.2 Simulations

This section is dedicated to a presentation of the numerical simulations. The aim of the simulation is to explain in practice how the detection and isolation strategies work.

It is considered a network of $N = 100$ computational units, organized in group of 4 each. The groups represent the physical proximity, and each subsystem is assumed to be physically connected to two of the the others subsystems within the same group. Moreover, the groups are disposed in a square grid composed of 5 columns and 5 rows, and this structure is assumed to be the same for all the simulations (see Figure 6.1). On the other hand, the virtual links will be changed in the simulations, to show the effectiveness of the proposed methodologies.

6.2.1 Detection

This first simulation shows the effectiveness of the detection strategy outlined in Chapter 3. With reference to the topology depicted in Figure 6.1, at $t = 7.5$ min, a malicious agent performs a covert cyber-attack in subsystem $\mathcal{S}_{52}$. As a consequence, the state components leave the nominal trajectory, but both the decentralized and the distributed estimate are insensitive to the attack, as showed in



(a) First component (temperature). (b) Second component (computational load).

Figure 6.2: State components of $\mathcal{S}_{52}$: true value x_{52} (green), decentralized estimate $\hat{x}^d_{52}$ (blue), and distributed estimate $\hat{x}^c_{52}$ (red).

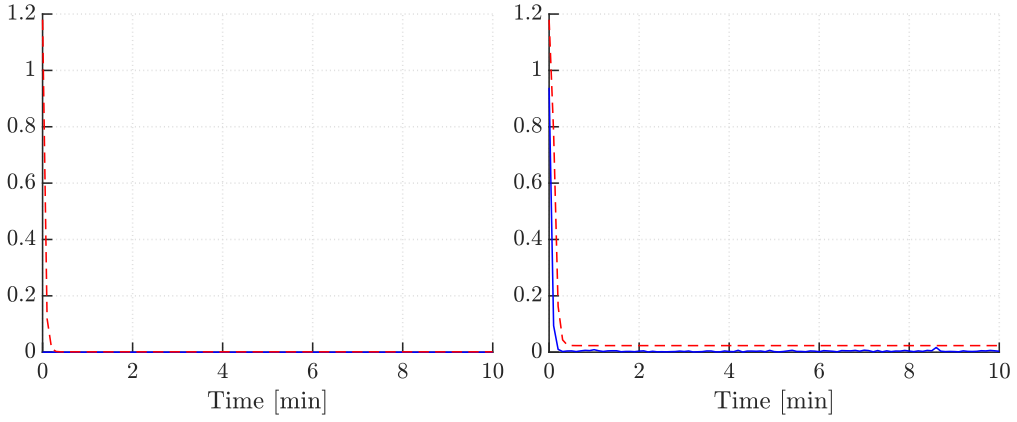
Figure 6.2. As a consequence, both the decentralized and the distributed residuals are steadily below the associated threshold, see Figures 6.3(a) and 6.3(b), respectively. Observe that the decentralized residual $\tilde{r}^d_{52}$ is remarkably smaller in amplitude with respect to the distributed residual $\tilde{r}^c_{52}$. This depends on the fact that, whenever the output matrix C_i is full-column rank (as it is the case for all the considered subsystems), then a suitable UIO gain H_i is the left inverse of the output matrix itself, $H_i = C_i^{-L}$. From this, many of the UIO gains are computed in a way so that the UIO estimate basically boils down to:

$$\hat{x}^d_i = C_i^{-L} \tilde{y}_i. \tag{6.11}$$

As a consequence, the decentralized state estimation error ϵ^d_i depends on the measurement noise v_i only. Furthermore, in this particular scenario, the measurement noise only affects the first component of the output $\tilde{y}_{i[1]}$, therefore the second component of the decentralized residual $\tilde{r}^d_{i[2]}$ is mathematically zero.

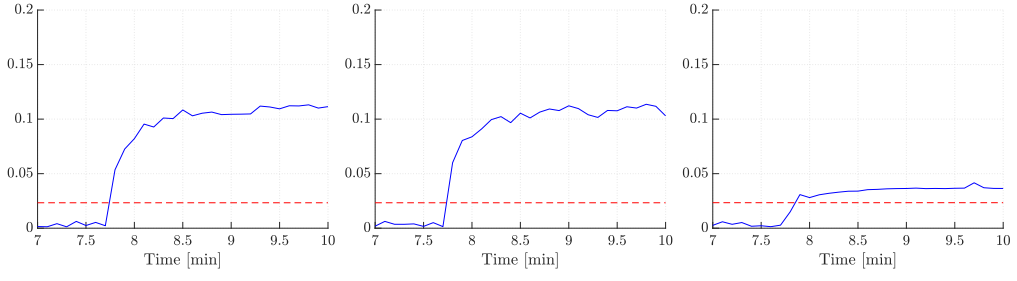
On the other hand, as proved in Subsection 3.3.3, the distributed residuals of its neighbors are sensitive to the attack. Indeed, by time $t = 7.9$ min, the norms of the distributed residual of all neighbors $\mathcal{S}_{49}$, $\mathcal{S}_{51}$, and $\mathcal{S}_{70}$ have crossed the threshold (Figures 6.4(a), 6.4(b), and 6.4(c), respectively). Observe that the residual $\tilde{r}^c_{70}$ is smaller in magnitude with respect to $\tilde{r}^c_{49}$ and $\tilde{r}^c_{51}$. This can be easily understood from the fact that subsystems $\mathcal{S}_{52}$ and $\mathcal{S}_{70}$ are linked only through the second state component, whereas $\mathcal{S}_{49}$ and $\mathcal{S}_{51}$ are coupled to $\mathcal{S}_{52}$ both physically and via software, hence the attack in $\mathcal{S}_{52}$ affects all the state components of $\mathcal{S}_{49}$ and $\mathcal{S}_{51}$.

Observe that it takes the attack a while to be detected. This is due to the fact that the attacker input is itself constrained. Indeed, given the physical interpretation of the actuation signal u_i, condition (6.6b) holds for the attack



(a) Decentralized residual $\left\|\tilde{r}_{52}^{d}\right\|$. (b) Distributed residual $\|\tilde{r}_{52}^{c}\|$.

Figure 6.3: Norm of the residual quantities of $\mathcal{S}_{52}$.



(a) $\|\tilde{r}_{49}^{c}\|$. (b) $\|\tilde{r}_{51}^{c}\|$. (c) $\|\tilde{r}_{70}^{c}\|$.

Figure 6.4: Norm of the distributed residuals of the neighbors of $\mathcal{S}_{52}$.

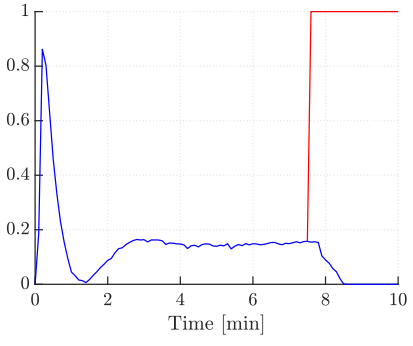


Figure 6.5: Input signals of $\mathcal{S}_{52}$: legitimate u_{52} (blue) and attacked $\tilde{u}_{52}$ (red).

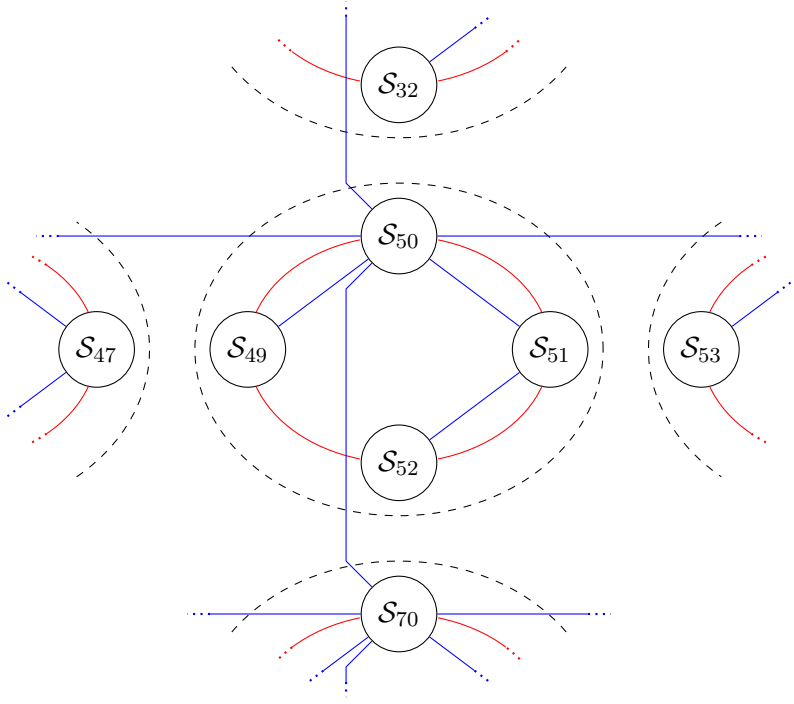


Figure 6.6: Detail of the interconnection graph, specifying the temperature links (red, curved), and the logical links (blue, straight).

control input signal η_i in the following form:

$$-u_i \leq \eta_i \leq 1 - u_i. \tag{6.12}$$

As a consequence, the attacker control input is itself constrained and its influence takes some time to grow and to cause the residuals to cross the threshold. The legitimate and attacked input signals are depicted in Figure 6.5.

Finally, observe that in such a case, the detection strategy correctly works since the only raised-up alarm signals are a_{49}, a_{51} and a_{70}, and this condition uniquely identifies $\mathcal{S}_{52}$, since $\mathcal{N}_{52} = \{49, 51, 70\}$. Nonetheless, we have $\mathcal{N}_{49} = \mathcal{N}_{51} = \{50, 52\}$, therefore an isolation strategy is needed to distinguish among $\mathcal{S}_{49}$ and $\mathcal{S}_{51}$. In Subsection 6.2.2, the theoretical results of Chapter 4 are employed to resolve such an ambiguity.

6.2.2 Isolation

In this second subsection, the isolation architecture is tested. As highlighted in Subsection 6.2.1, the configuration of Figure 6.1 results in an ambiguity whenever $\|\tilde{r}^c_{50}\|$ and $\|\tilde{r}^c_{52}\|$ are the only residual quantities to cross the associated thresholds, causing a_{50} and a_{52} to be raised up. In this condition, it is not possible to decide whether $\mathcal{S}_{49}$ or $\mathcal{S}_{51}$ is under attack.

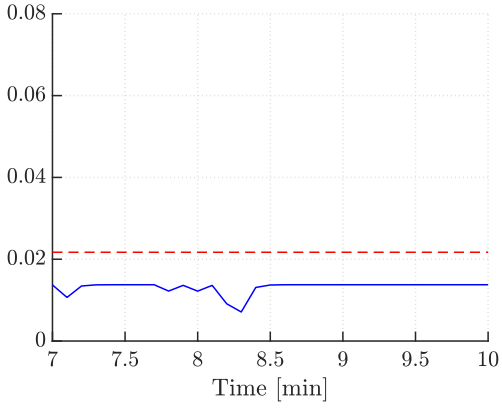
To overcome the issue, the isolation algorithms developed in Chapter 4 can be used. Nonetheless, since the subsystems are linked both physically and via software, the interconnection matrices A_{ij} are full-rank, preventing all the isolation strategies from working (see Subsection 4.2). On the one hand, the physical link is understood as a consequence of the physical proximity between computational units, and it is assumed to be fixed. On the other hand, one can easily give up some of the software links between subsystems in order to reduce the rank of the interconnection matrices, earning the possibility to successfully implement the isolation architectures. Observe that this comes at the price of (possibly) reducing the rate of convergence to the balance in the computational load (see Appendix C). Nonetheless, if the network is still sufficiently connected, the trade-off is acceptable, especially because this is only needed to break the symmetry.

For this reason, let assume the network topology is modified as in Figure 6.6. Observe that the software network is still represented by a connected graph. By considering the new network topology, one can effectively implement the UIO of the merged subsystem (Subsection 4.2.1) and the filtered Luenberger residual (Subsection 4.2.2). Figure 6.7 shows the residual $\tilde{r}^d_{51\cup52}$ of the UIO of the merged subsystem $\mathcal{S}_{51\cup52}$, and the filtered Luenberger residual $\tilde{r}^g_{52}$ of $\mathcal{S}_{52}$, both in the case of an attack in $\mathcal{S}_{49}$ (Figures 6.7(a) and 6.7(b)), and if $\mathcal{S}_{51}$ is attacked (Figures 6.7(c) and 6.7(d)). Both the residuals $\tilde{r}^d_{51\cup52}$ and $\tilde{r}^g_{52}$ are insensitive to attacks in $\mathcal{S}_{49}$, but sensitive to attacks in $\mathcal{S}_{51}$, and can be used to discriminate. Therefore, by looking at the above Figures 6.7(a) and 6.7(b) one deduces that $\mathcal{S}_{51}$ is not attacked, therefore $\mathcal{S}_{49}$ must be. Symmetrically, by looking at the below Figures 6.7(c) and 6.7(d), it is understood that $\mathcal{S}_{51}$ is under attack.

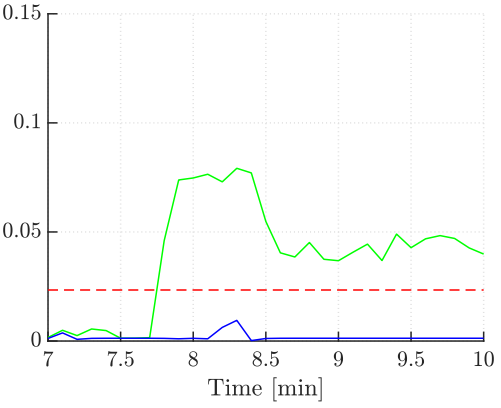
Note that the filtered two-step Luenberger residual approach presented in Subsection 4.2.3 cannot be adopted given the topology in Figure 6.6, since all the second-order interconnection matrices involving $\mathcal{S}_{49}$ and $\mathcal{S}_{51}$ would be full-rank.

To effectively test this technique, let consider the topology in Figure 6.8. Given the symmetry of the interconnections, each subsystem is necessarily a second order neighbor of itself. On the contrary, concerning the crossed interconnections, $\mathcal{S}_{49}$ and $\mathcal{S}_{51}$ are still second-order neighbors in the temperature perspective (red graph). Nonetheless, there is no path of length two connecting the two subsystems via software (blue graph), therefore the second-order interconnection matrix from $\mathcal{S}_{49}$ to $\mathcal{S}_{51}$ is not full-rank, and the technique can be adopted. Figure 6.9 shows the filtered residual $\tilde{r}^s_{52}$ of the two-step Luenberger Observer $\mathcal{O}^s_{51}$, designed to be sensitive to attacks in $\mathcal{S}_{51}$ only. As one can observe, if $\mathcal{S}_{49}$ is attacked (Figure 6.9(a)), $\tilde{r}^s_{51}$ continues fluctuating due to the presence of noise, with no significant trend after $t = 7.5$ min. Conversely, if $\mathcal{S}_{51}$ is attacked (Figure 6.9(b)), the residual clearly increases in norm. Still, the norm of the residual only barely crosses the threshold, and only for a few instants.

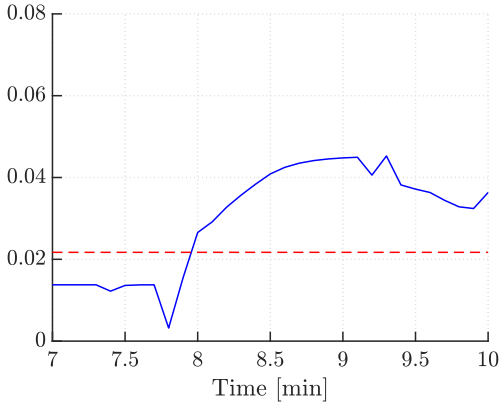
The reason for this can be found in light of equation (6.12), that is the at-



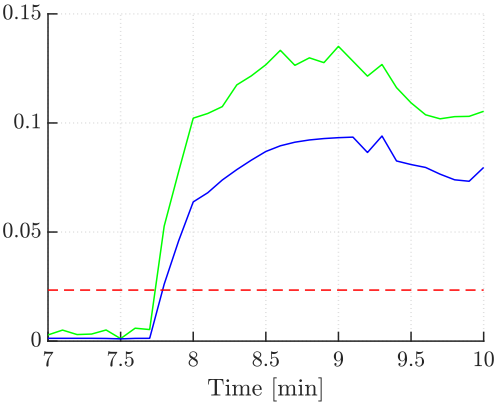
(a) UIO residual of $\mathcal{S}_{51\cup52}$, $\mathcal{S}_{49}$ is attacked.



(b) Luenberger residuals of $\mathcal{S}_{52}$, $\mathcal{S}_{49}$ is attacked.



(c) UIO residual of $\mathcal{S}_{51\cup52}$, $\mathcal{S}_{51}$ is attacked.



(d) Luenberger residuals of $\mathcal{S}_{52}$, $\mathcal{S}_{51}$ is attacked.

Figure 6.7: Isolation residual quantities. On the left, UIO residual of the merged subsystem $\mathcal{S}_{51\cup52}$ (blue) and associated threshold (red); on the right, original Luenberger residual $\tilde{r}^c_{52}$ of subsystem $\mathcal{S}_{52}$ (green), filtered Luenberger residual $\tilde{r}^g_{52}$ (blue), and associated threshold (red). All the quantities are depicted both if $\mathcal{S}_{49}$ is attacked (above), and if $\mathcal{S}_{51}$ is attacked (below).

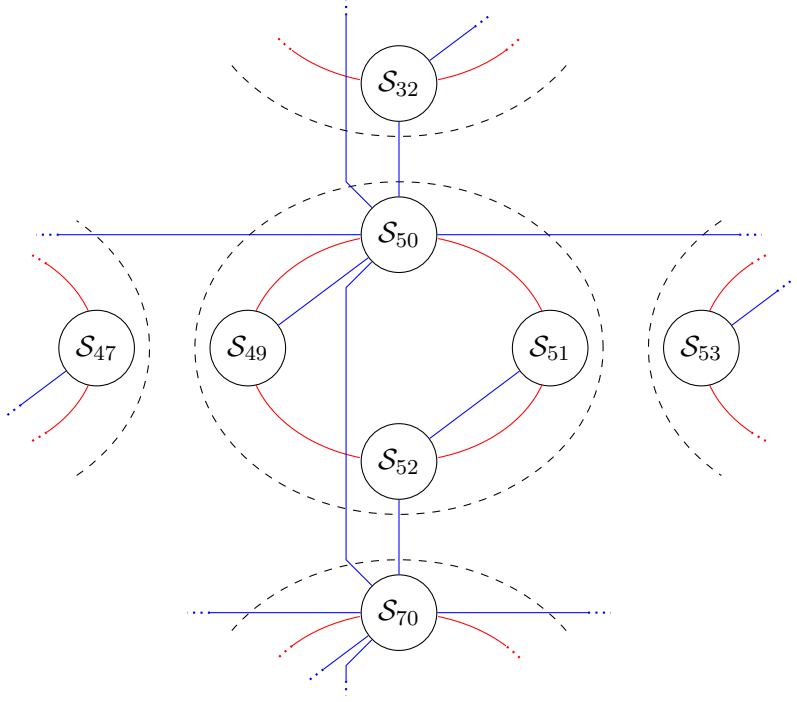
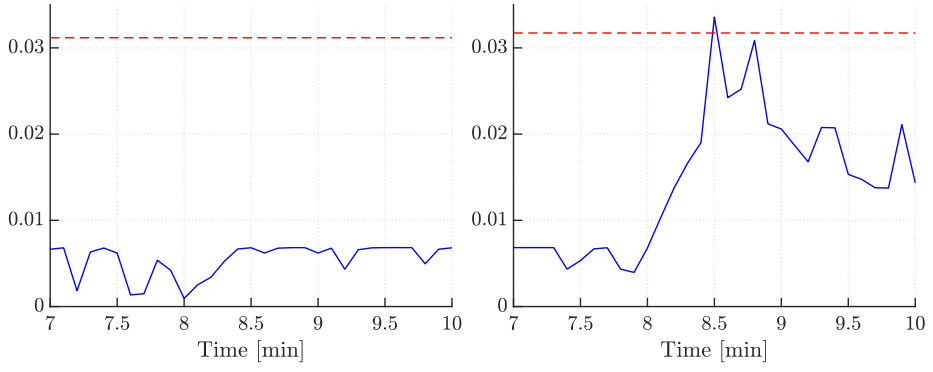


Figure 6.8: Detail of the interconnection graph, specifying the temperature links (red, curved), and the logical links (blue, straight).



(a) $\mathcal{S}_{49}$ is attacked.

(b) $\mathcal{S}_{51}$ is attacked.

Figure 6.9: Norm of the filtered residual $\|\tilde{r}^s_{51}\|$ of the two-step Luenberger Observer $\mathcal{O}^s_{51}$, sensitive to attacks in $\mathcal{S}_{51}$ only.

tacker action is limited by the constraints on the input. Moreover, this highlights an important practical limitation of this techinque. The entries of the second-order interconnection matrices M_{il} are likely small in absolute value, being the result of a multiplication of first-order interconnection matrices $A_{ij}A_{jl}$. Therefore, the contribution $M_{il}\epsilon_l^d$ of the decentralized true error ϵ_l^d on the filtered two-step Luenberger residual $\tilde{r}_i^s$ might be very attenuated. On the contrary, the contribution of the noises w_i and v_i is not attenuated and this can result in very conservative thresholds. The evidence of this fact can be found by observing that in both Figures 6.9(a) and 6.9(b) the nominal fluctuation of the residual is far below the threshold.

6.2.3 Coordinated multiple attacks

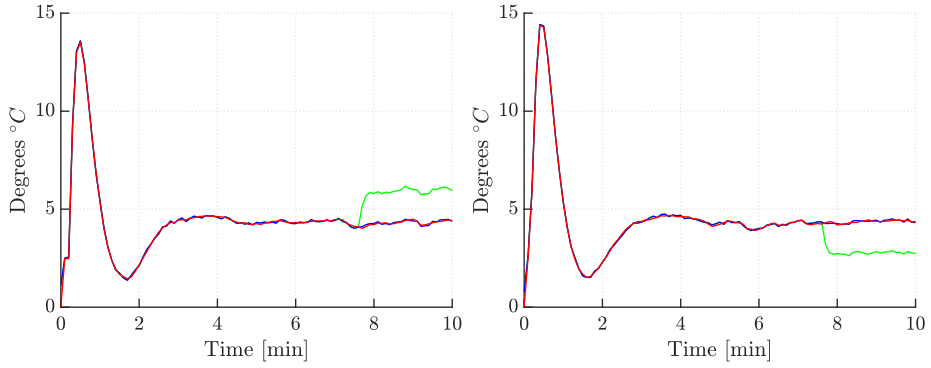
This last subsection focuses on the coordinated multiple attacks outlined in Chapter 5. The topology adopted is again the one of Figure 6.1.

Firstly, let assume that a malicious agent is affecting both subsystems $\mathcal{S}_{49}$ and $\mathcal{S}_{51}$ in order to compensate the effect on a common neighbor, as in Section 5.1. Observe that, given the symmetries of the topology, the compensation takes place in both $\mathcal{S}_{50}$ and $\mathcal{S}_{52}$. The state components and the input signals of the attacked subsystems are illustrated in Figure 6.10.

On the other hand, Figure 6.11 shows the distributed residual of the neighboring subsystems $\mathcal{S}_{50}$ (Figure 6.11(a)) and $\mathcal{S}_{52}$ (Figure 6.11(b)). As one can see, after $t = 7.5$ min none of the residual quantity reveals any significant behavior with respect to the nominal fluctuation, that is the coordinate attack is perfectly covert.

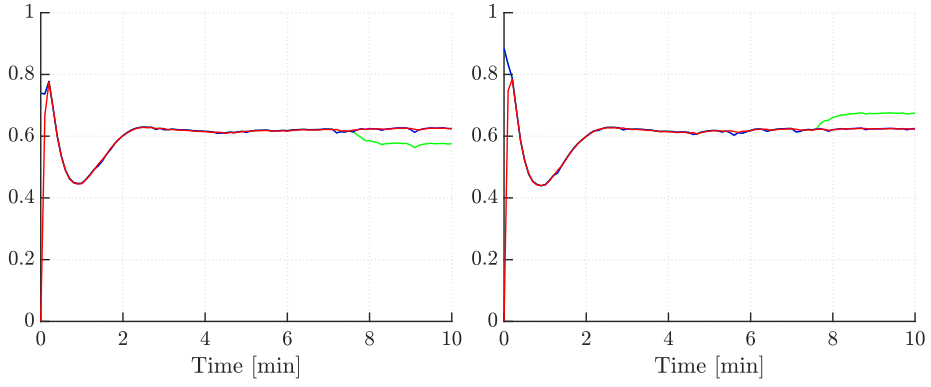
Finally, the last simulation is dedicated to the coordinated attack presented in Section 5.2. Specifically, let assume a malicious agent is performing a covert cyber-attack in $\mathcal{S}_{51}$, while trying to compensate the effect in $\mathcal{S}_{50}$. As discussed in Section 5.2, a low-rank input matrix B_i (as it is the case for all the considered subsystems) is likely enough to prevent this attack strategy from working.

Indeed, as one can see in Figure 6.12, the attempt of compensation by the attacker is unsuccessful. Indeed, despite significantly reduced than in the case of no compensation ($\tilde{r}_{52}^c$, see Figure 6.12(b)), the norm of the distributed residual $\tilde{r}_{50}^c$ still crosses the associated threshold, as depicted in Figure 6.12(a). As a consequence, the attack is correctly detected.



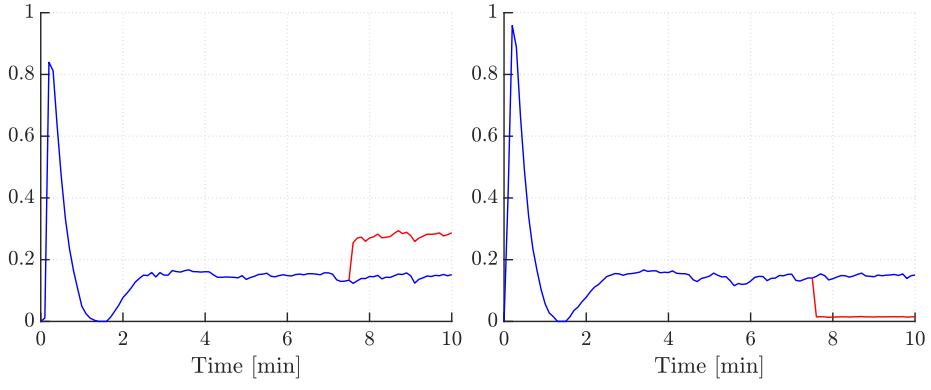
(a) $x_{49[1]}$.

(b) $x_{51[1]}$.



(c) $x_{49[2]}$.

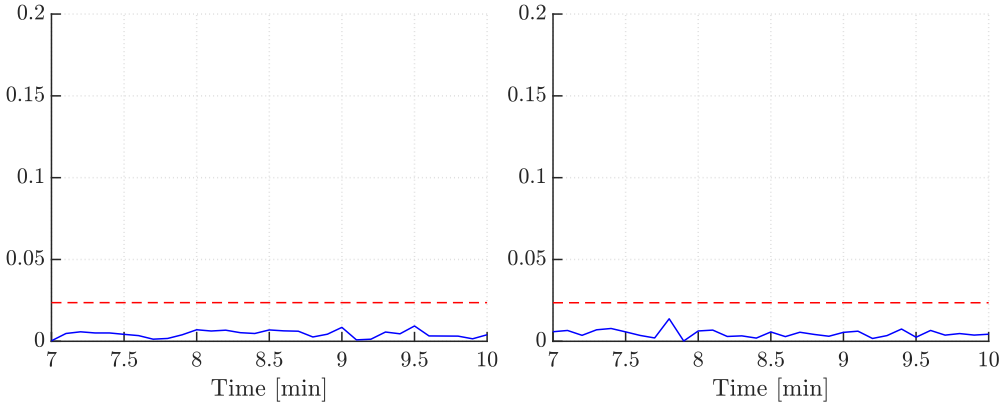
(d) $x_{51[2]}$.



(e) u_{49}.

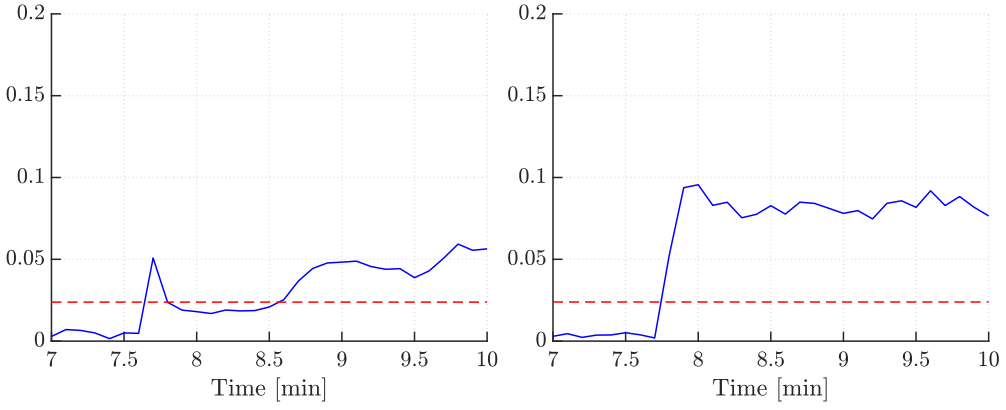
(f) u_{51}.

Figure 6.10: State components and input of the attacked subsystems $\mathcal{S}_{49}$ (left) and $\mathcal{S}_{51}$ (right). Concerning the state components, it is showed the true value x_i (green), the decentralized estimate $\hat{x}_i^d$ (blue), and the distributed estimate $\hat{x}_i^c$ (red). Regarding the input, the legitimate u_i (blue) and the attacked $\tilde{u}_i$ are depicted.



(a) $\|\tilde{r}^c_{50}\|$. (b) $\|\tilde{r}^c_{52}\|$.

Figure 6.11: Norm of the distributed residuals of subsystems $\mathcal{S}_{50}$ and $\mathcal{S}_{52}$, when subsystems $\mathcal{S}_{49}$ and $\mathcal{S}_{51}$ are coordinately attacked.



(a) $\|\tilde{r}^c_{50}\|$ (b) $\|\tilde{r}^c_{52}\|$

Figure 6.12: Norm of the distributed residuals of subsystems $\mathcal{S}_{50}$ and $\mathcal{S}_{52}$, when subsystems $\mathcal{S}_{50}$ and $\mathcal{S}_{51}$ are coordinately attacked.

Chapter 7

Conclusions and future work

This thesis revisited previous works on the detection of covert cyber-attacks in interconnected systems. The novel results here developed allow the isolation of a vast majority of all possible attacks. As a consequence, an interesting research direction is the advancement of accommodation strategies which can counterbalance for the deviation introduced by the attacker.

On the other hand, an introductory analysis of the scenario of simultaneous multiple attacks was provided. An important open issue is the development of a detection algorithm for the detection of the attacks described in Section 5.1. Observe that all the attacks considered in this thesis are man-in-the-middle, meaning that the attacker acts on the actuation and measurement signals, whereas the communication channels between monitoring units are assumed to be safe. Still, it might be interesting to consider an attacker that develops a man-in-the-middle attack in a subsystem and, simultaneously, affects the communication channels of the associated monitoring unit. These scenarios are left as future research work.

CHAPTER 7. CONCLUSIONS AND FUTURE WORK

Appendix A

Discussion on the existence of a UIO for a merged subsystem

Before proving that if the UIO can be designed for each single subsystem, let state and prove a preliminary result.

Proposition A.0.1. *Given any subset of columns* Ξ_A *of* $\Xi = \left[\begin{array}{c|c} \Xi_A & \Xi_B \end{array} \right]$, *the condition:*

$$\operatorname{rank}(C\Xi) = \operatorname{rank}(\Xi) \tag{A.1}$$

implies:

i) $\operatorname{rank}(C\Xi_A) = \operatorname{rank}(\Xi_A)$*;*

ii) $\ker(C\Xi_A) = \ker(\Xi_A)$.

Proof.

i). (A.1) is equivalent to:

$$\dim\Big(\operatorname{Im}(\Xi) \cap \ker(C)\Big) = 0. \tag{A.2}$$

Being $\operatorname{Im}(\Xi_A) \subseteq \operatorname{Im}(\Xi)$, in particular we have:

$$\dim\Big(\operatorname{Im}(\Xi_A) \cap \ker(C)\Big) = 0 \tag{A.3}$$

which is equivalent to i).

ii). In general, we have:

$$\ker(\Xi_A) \subseteq \ker(C\Xi_A). \tag{A.4}$$

APPENDIX A. DISCUSSION ON THE EXISTENCE OF A UIO FOR A MERGED SUBSYSTEM

Let $\bar{n}_A$ be the number of columns in Ξ_A. From the rank-nullity theorem, we have:

$$\dim(\ker(C\Xi_A)) = \bar{n}_A - \operatorname{rank}(C\Xi_A) \tag{A.5a}$$
$$\dim(\ker(\Xi_A)) = \bar{n}_A - \operatorname{rank}(\Xi_A). \tag{A.5b}$$

Therefore, taken i) into account, we find:

$$\dim(\ker(\Xi_A)) = \dim(\ker(C\Xi_A)). \tag{A.6}$$

ii) trivially follows from (A.4) and (A.6).

□

The next proposition formally proves the expected result.

Proposition A.0.2. *If the rank condition* (3.22) *holds for each single subsystem, then it also hold for the merged subsystem.*

Proof. Let observe that, given (4.9), any vector $v \in \operatorname{Im}(\Xi_{j\cup i})$ can be decomposed as $v = \left[\begin{array}{c|c} v_j^\top & v_i^\top \end{array}\right]^\top$, with $v_j \in \operatorname{Im}(\check{\Xi}_j), v_i \in \operatorname{Im}(\check{\Xi}_i)$.

Let assume, by contradiction, there exists a vector $v \in \operatorname{Im}(\Xi_{j\cup i}), v \neq 0$, such that:

$$C_{j\cup i}v = \left[\begin{array}{c|c} C_j & 0 \\ \hline 0 & C_i \end{array}\right] \left[\begin{array}{c} v_j \\ \hline v_i \end{array}\right] = \left[\begin{array}{c} C_j v_j \\ \hline C_i v_i \end{array}\right] = 0. \tag{A.7}$$

If $v \neq 0$, then either $v_j \neq 0$ or $v_i \neq 0$ (or both of them). Let assume $v_j \neq 0$. Then, we have $v_j \in \operatorname{Im}(\check{\Xi}_j) \subseteq \operatorname{Im}(\Xi_j)$, $0 = \operatorname{rank}(C_j v_j) < \operatorname{rank}(v_j) = 1$, and this, in view of Proposition A.0.1, contradicts the hypothesis $\operatorname{rank}(C_j\Xi_j) = \operatorname{rank}(\Xi_j)$, that is the UIO could not be designed for subsystem $\mathcal{S}_j$ in the first place. Therefore, such a vector v must not exist.

As a consequence, we have:

$$\dim\Big(\operatorname{Im}(\Xi_{j\cup i}) \cap \ker(C_{j\cup i})\Big) = 0, \tag{A.8}$$

or, equivalently, $\operatorname{rank}(C_{j\cup i}\Xi_{j\cup i}) = \operatorname{rank}(\Xi_{j\cup i})$, that is condition (3.22) holds for the merged subsystem $\mathcal{S}_{j\cup i}$. □

Appendix B

Preparatory result on the attacker reachable space

Proposition B.0.1. *$\forall t = 0...k-2$, given any two squared matrices of the same dimensions A, F, the following equality holds:*

$$\sum_{\tau=t+1}^{k-1} F^{k-1-\tau}(A-F)A^{\tau-1-t} = A^{k-1-t} - F^{k-1-t}. \tag{B.1}$$

Proof. The proof easily follows by applying the principle of Mathematical Induction.

Base case.
Let consider the case $t = k-2$. By inspection, the left-hand side term of equation (B.1) is:

$$\begin{aligned}\sum_{\tau=k-2+1}^{k-1} F^{k-1-\tau}(A-F)A^{\tau-1-(k-2)} &= \sum_{\tau=k-1}^{k-1} F^{k-1-\tau}(A-F)A^{\tau-(k-1)} \\ &= F^{k-1-\tau}(A-F)A^{\tau-(k-1)}\Big|_{\tau=k-1} \\ &= F^{k-1-(k-1)}(A-F)A^{(k-1)-(k-1)} \\ &= A-F.\end{aligned} \tag{B.2}$$

Moreover, the right-hand side is:

$$\left(A^{k-1-t} - F^{k-1-t}\right)\Big|_{t=k-2} = A-F. \tag{B.3}$$

Inductive step.
Let assume (B.1) holds $\forall k-2 \geq t > \bar{t}$. By considering the case $t = \bar{t}$, we have:

$$\begin{aligned}
\sum_{\tau=\bar{t}+1}^{k-1} F^{k-1-\tau}(A-F)A^{\tau-1-\bar{t}} =& \sum_{\tau=(\bar{t}+1)+1}^{k-1} F^{k-1-\tau}(A-F)A^{\tau-1-\bar{t}} \\
&+ \left(F^{k-1-\tau}(A-F)A^{\tau-1-\bar{t}}\right)\Big|_{\tau=\bar{t}+1} \\
=& \left(\sum_{\tau=(\bar{t}+1)+1}^{k-1} F^{k-1-\tau}(A-F)A^{\tau-1-(\bar{t}+1)}\right)A \\
&+ F^{k-2-\bar{t}}(A-F)A^0 \\
=& \left(A^{k-1-(\bar{t}+1)} - F^{k-1-(\bar{t}+1)}\right)A \\
&+ F^{k-2-\bar{t}}(A-F) \\
=& \left(A^{k-2-\bar{t}} - F^{k-2-\bar{t}}\right)A + F^{k-2-\bar{t}}(A-F) \\
=& A^{k-1-\bar{t}} - F^{k-1-\bar{t}}.
\end{aligned} \tag{B.4}$$

□

Appendix C

Numerical values used in the simulations

In this section all the numerical values of all the coefficients of the model used in the simulations are provided. Note that all the subsystems are characterized by the same parameters.

- $N = 100$.
- $T_s = 0.1$ min.
- α_i: drawing inspiration from [32], by considering that in this example both the distance and the body masses are on different scales, a guess of the continuous-time eigenvalue $\tau = 0.1$ min was obtained, leading to $\alpha_i = e^{-\frac{T_s}{\tau}}$.
- β_i: from the graphs and the considerations provided in [33], it was assumed $\beta_i = 1$ °C (note that $x_{i[1]}$ is in °C, and $x_{i[2]}$ is a dimensionless quantity).
- φ_i: from [32], [33], and further considerations on the size of the components, it was assumed $\varphi_i = 8.0327$ °C (note that $x_{i[1]}$ is in °C, and u_i is a dimensionless quantity).
- δ_{ij}: all the (non-zero) temperature links are assumed to be equal, with coefficients $\delta_{ij} = 0.3\alpha_i$.
- $w_{i[1]}$: it is simulated as a uniformly distributed noise in $[-0.1, 0.1]$ °C.
- $v_{i[1]}$: it is simulated as a uniformly distributed noise in $[-0.01, 0.01]$ °C.
- μ_i: the parameter basically expresses the ratio between the computational capacity deriving from a certain amount of power used, and the size of the buffer. In the simulations, the parameter is set to $\mu_i = -0.1285$.

- $w_{i[2]}$: in line with definition (6.8), it is simulated as a scaled and truncated Poisson variable. Concerning the parameters, the Poisson is ruled by $\lambda_{w_{i[2]}} = 10$, the scale factor is $w_{i[2]f} = 400$, and the saturation bound is $\bar{w}_{i[2]} = 0.02$
- regarding the MPC, it was designed by neglecting the neighbors temperature influence, and by considering an unstable system (that is $\rho_{ii} = 1$, since at the consensus the exchange of queries for each subsystem is perfectly balanced), subject to the linear constraints (6.6). Moreover, the reference signal was the vector $\begin{bmatrix} 0 & 0.6 \end{bmatrix}^\top$, and the weights chosen were 10^{-3} for the input, and 1 and 50 for the two state components, respectively.

Finally, the choice of the virtual link parameters ρ_{ij} deserves a separated discussion. First of all, they depend on the given topology of the virtual network, meaning that they ρ_{ij} can be non-zero only if there exists a virtual link between $\mathcal{S}_i$ and $\mathcal{S}_j$. Once the links are fixed, one can decide to fix the parameters value depending on different purposes.

The idea of redistribution of quantities among different agents composing a network has extensively been extensively discussed in the theory of consensus, see [34]. Among the notable results of such a theory, it is proved that the connectivity of the network (specifically, the number of interconnections between nodes) directly influences the maximum achievable rate of convergence to the consensus, which in our case is the perfect balancing of the number of queries among all processors in the network. Moreover, it is deduced that the consensus can be reached by imposing a time-invariant row-stochastic weights matrix P, where $P \in \mathbb{R}^{N \times N}$ is the matrix whose entry in position (i, j) is the coefficient ρ_{ij}. Equivalently, one need to impose that the coefficients ρ_{ij} are so that:

$$\rho_{ii} + \sum_{j \in \mathcal{N}_i} \rho_{ij} = 1. \tag{C.1}$$

On the other hand, given the constraint on the conservation of the number of queries (6.3), the P matrix must also be designed column-stochastic.

In the simulations, for all the different topologies of the network (see Section 6.2), the Metropolis rule [35] was employed. Such an algorithm for the weight assignment is well-known and frequently adopted because it allows for a distributed design of a symmetric double-stochastic P matrix, resulting in a balanced achievement of the average consensus. Specifically, given the fact that every subsystem $\mathcal{S}_i$ needs to have knowledge of the degree d_j of each of its neighbors $\mathcal{S}_j, j \in \mathcal{N}_i$, the coefficients are chosen as follows:

$$\rho_{ij} = \begin{cases} \frac{1}{1+\max(d_i, d_j)} & j \in \mathcal{N}_i \\ 1 - \sum_{j \in \mathcal{N}_i} \rho_{ij} & i = j \\ 0 & \text{otherwise.} \end{cases} \tag{C.2}$$

It is worth observing that the knowledge of the degree of the neighbors, which is a second-order information, is also a requirement for the implementation of the algorithm identifying possible ambiguities in the network, see Algorithm 1 in Section 4.2.

APPENDIX C. NUMERICAL VALUES USED IN THE SIMULATIONS

List of Symbols

α_i	temperature eigenvalue of the data center model
$\bar{\epsilon}_i^c$	bound on the norm of the distributed true error ϵ_i^c
$\bar{\epsilon}_i^d$	bound on the norm of the decentralized true error ϵ_i^d
$\bar{\mathcal{N}}_i$	index set of second-order neighbors of $\mathcal{S}_i$
$\bar{r}_i^c$	bound on the norm of the distributed residue r_i^c
$\bar{r}_i^s$	bound on the norm of the filtered two-step residue r_i^s
$\bar{v}_i$	bound on the norm of the measurement noise v_i
$\bar{w}_i$	bound on the norm of the process noise w_i
$\bar{w}_{i[2]}$	bound on the absolute value of $w_{i[2]}$ in the data center model
$\bar{x}_i(0)$	bound on the initial condition of the state of $\mathcal{S}_i$
β_i	computational load to temperature gain in the data center model
$\boldsymbol{x}_i$	state vector of the neighbors of $\mathcal{S}_i$
$\check{\check{\Xi}}_i$	partition of $\check{\Xi}_i$
$\check{\Xi}_i$	partition of Ξ_i
δ_{ij}	temperature dynamical coupling from $\mathcal{S}_j$ to $\mathcal{S}_i$ in the data center model
ϵ_i^c	distributed true error of observer $\mathcal{O}_i^c$
ϵ_i^d	decentralized true error of observer $\mathcal{O}_i^d$
η_i	control input vector of $\tilde{\mathcal{S}}_i$
Γ_i	state-to-output matrix of the full attacker model
γ_i	output vector of $\tilde{\mathcal{S}}_i$
$\hat{x}_i^c$	distributed estimate computed by observer $\mathcal{O}_i^c$

LIST OF SYMBOLS

$\hat{x}_i^d$	decentralized estimate computed by observer $\mathcal{O}_i^d$
$\hat{x}_i^s$	distributed estimate computed by two-step observer $\mathcal{O}_i^s$
$\lambda_{w_{i[2]}}$	expected value of the Poisson component in $w_{i[2]}$
$\mathbb{C}$	set of complex numbers
$\mathbb{N}$	set of natural numbers
$\mathbb{R}$	set of real numbers
$\mathbb{Z}$	set of integer numbers
$\mathcal{A}_i$	malicious agent locally acting on subsystem $\mathcal{S}_i$
$\mathcal{C}_i$	local controller of $\mathcal{S}_i$
$\mathcal{D}_i$	Detector of $\mathcal{S}_i$
$\mathcal{I}_{i,j}$	index set of subsystems ambiguous to $\mathcal{S}_i$ from the perspective of $\mathcal{S}_j$
$\mathcal{N}_i$	index set of neighbors of $\mathcal{S}_i$
$\mathcal{N}_{A,i}$	index set of subsystems ambiguous to $\mathcal{S}_i$
$\mathcal{O}_i^c$	distributed observer of $\mathcal{S}_i$
$\mathcal{O}_i^d$	decentralized observer of $\mathcal{S}_i$
$\mathcal{O}_i^s$	distributed two-step observer of $\mathcal{S}_i$
$\mathcal{S}_i$	ith subsystem
μ_i	power to computational load gain in the data center model
ν_i	reference vector of $\tilde{\mathcal{C}}_i$
Φ_i	state-to-state matrix of the full attacker model $\mathcal{A}_i$
ρ_{ij}	logical dynamical coupling from $\mathcal{S}_j$ to $\mathcal{S}_i$ in the data center model
θ_i	attacker influence on the decentralized true error ϵ_i^d
$\tilde{\epsilon}_i^c$	distributed computed error of observer $\mathcal{O}_i^c$
$\tilde{\epsilon}_i^d$	decentralized computed error of observer $\mathcal{O}_i^d$
$\tilde{\epsilon}_i^s$	distributed computed error of the two-step observer $\mathcal{O}_i^s$
$\tilde{\mathcal{C}}_i$	attacker controller
$\tilde{\mathcal{S}}_i$	attacker replica of $\mathcal{S}_i$

LIST OF SYMBOLS

$\tilde{r}_i^c$	distributed computed residue of observer $\mathcal{O}_i^c$
$\tilde{r}_i^d$	decentralized computed residue of observer $\mathcal{O}_i^d$
$\tilde{r}_i^g$	filtered distributed computed residue of observer $\mathcal{O}_i^c$
$\tilde{r}_i^s$	filtered two-step residue of observer $\mathcal{O}_i^s$
$\tilde{u}_i$	attacked control input vector of $\mathcal{S}_i$
$\tilde{x}_i$	state vector of $\tilde{\mathcal{S}}_i$
$\tilde{y}_i$	corrupted measurement vector of $\mathcal{S}_i$
Υ_i	disclosure information of attacker $\mathcal{A}_i$
φ_i	power to temperature gain in the data center model
Ξ_i	interconnection matrix of neighbors of $\mathcal{S}_i$
ξ_i	state vector of $\tilde{\mathcal{C}}_i$
ζ_i	state vector of the full attacker model $\mathcal{A}_i$
A_i	state-to-state matrix of $\mathcal{S}_i$
a_i	alarm signal of $\mathcal{S}_i$
$A_{\tilde{\mathcal{C}}_i}$	state-to-state matrix of the attacker controller $\tilde{\mathcal{C}}_i$
A_{ij}	dynamic interconnection from of S_j to S_i
B_i	input-to-state matrix of $\mathcal{S}_i$
C_i	state-to-output matrix of $\mathcal{S}_i$
$C_{\tilde{\mathcal{C}}_i}$	state-to-output matrix of the attacker controller $\tilde{\mathcal{C}}_i$
d_i	degree of the ith node in the network
F_i^c	dynamical matrix of error ϵ_i^c
F_i^d	dynamical matrix of error ϵ_i^d
F_i^s	dynamical matrix of error ϵ_i^s
g_i	filter for the residues $\tilde{r}_i^c$ or $\tilde{r}_i^s$
H_i	output-to-estimate matrix of $\mathcal{O}_i^d$
k	sampling instant
K_i	output-to-state matrix of $\mathcal{O}_i^d$

LIST OF SYMBOLS

$K_i^{(1)}$	first component of K_i
$K_i^{(2)}$	second component of K_i
$K_{\tilde{\mathcal{C}}_i}$	feedback from the state of $\tilde{\mathcal{C}}_i$
k_{ai}	initial instant of the attack in $\mathcal{S}_i$
L_i	gain of the observer $\mathcal{O}_i^c$
LU_i	logic unit monitoring subsystem $\mathcal{S}_i$
m_i	dimension of the control input vector u_i
N	number of subsystems in the network
n_i	dimension of the state vector x_i
P	weights matrix of the logical links in the data center model
p_i	dimension of the output vector y_i
Q_i^c	bound on the norm of the noise contribution to the true error ϵ_i^c
Q_i^d	bound on the norm of the noise contribution to the true error ϵ_i^d
Q_i^s	bound on the norm of the noise contribution to the true error ϵ_i^s
r_i^c	distributed residue of observer $\mathcal{O}_i^c$
r_i^d	decentralized residue of observer $\mathcal{O}_i^d$
$R_{\tilde{\mathcal{C}}_i}$	reference-to-state matrix of the attacker controller $\tilde{\mathcal{C}}_i$
s_i	bound on the norm of the neighbors' error contribution to ϵ_i^s
T_i	input-to-state matrix of observer $\mathcal{O}_i^d$
T_s	sampling time
u_i	control input vector of $\mathcal{S}_i$
v_i	measurement noise of $\mathcal{S}_i$
w_i	process noise of $\mathcal{S}_i$
$w_{i[2]f}$	scale factor of the Poisson component of $w_{i[2]}$ in the data center model
x_i	state vector of $\mathcal{S}_i$
y_i	output vector of $\mathcal{S}_i$
z_i	state vector of observer $\mathcal{O}_i^d$

Bibliography

[1] 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*, pages 1–1, 2020.

[2] R. M. Lee, M. J. Assante, and T. Conway. Analysis of the cyber attack on the ukrainian power grid. *SANS Industrial Control Systems*, 2016.

[3] Siobhan Gorman. Electricity grid in u.s. penetrated by spies. *The Wall Street Journal*, April 8th, 2009.

[4] Roger Anderson. Final report on the blackout in the united states and canada: Causes and recommendations final report on the blackout in the united states and canada: Causes and recommendations, 08 2004.

[5] 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*, pages 1–1, 2020.

[6] 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*, 35(1):93–109, 2015.

[7] 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)*, pages 1806–1813, 2012.

[8] H. Sandberg, S. Amin, and K. H. Johansson. Cyberphysical security in networked control systems: An introduction to the issue. *IEEE Control Systems Magazine*, 35(1):20–23, 2015.

[9] André Teixeira, Iman Shames, Henrik Sandberg, and Karl Henrik Johansson. A secure control framework for resource-limited adversaries. *Automatica*, 51:135 – 148, 2015.

[10] A. Teixeira, H. Sandberg, and K. H. Johansson. Networked control systems under cyber attacks with applications to power networks. In *Proceedings of the 2010 American Control Conference*, pages 3690–3696, 2010.

[11] Francesco Bullo and Fabio Pasqualetti. *Secure control systems: a control-theoretic approach to cyber-physical security.* PhD thesis, University of California, Santa Barbara, 2012.

[12] F. Boem, R. M. G. Ferrari, C. Keliris, T. Parisini, and M. M. Polycarpou. A distributed networked approach for fault detection of large-scale systems. *IEEE Transactions on Automatic Control*, 62(1):18–33, 2017.

[13] F. Boem, S. Riverso, G. Ferrari-Trecate, and T. Parisini. Plug-and-play fault detection and isolation for large-scale nonlinear systems with stochastic uncertainties. *IEEE Transactions on Automatic Control*, 64(1):4–19, 2019.

[14] R. S. Smith. Covert misappropriation of networked control systems: Presenting a feedback structure. *IEEE Control Systems Magazine*, 35(1):82–92, 2015.

[15] André Teixeira, Iman Shames, Henrik Sandberg, and Karl Henrik Johansson. A secure control framework for resource-limited adversaries. *Automatica*, 51:135 – 148, 2015.

[16] R. Langner. Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security Privacy*, 9(3):49–51, 2011.

[17] Alan Sá, Luiz Carmo, and Raphael Machado. A controller design for mitigation of passive system identification attacks in networked control systems. *Journal of Internet Services and Applications*, 9, 12 2018.

[18] Jie Chen, Ron Patton, and HONG-YUE ZHANG. Design of unknown input observers and robust fault detection filters. *International Journal of Control*, 63:85–105, 02 2007.

[19] Iman Shames, André M.H. Teixeira, Henrik Sandberg, and Karl H. Johansson. Distributed fault detection for interconnected second-order systems. *Automatica*, 47(12):2757 – 2764, 2011.

[20] A. J. Gallo, M. S. Turan, P. Nahata, F. Boem, T. Parisini, and G. Ferrari-Trecate. Distributed cyber-attack detection in the secondary control of dc microgrids. In *2018 European Control Conference (ECC)*, pages 344–349, 2018.

[21] A. J. van der Schaft. L/sub 2/-gain analysis of nonlinear systems and nonlinear state-feedback h/sub infinity / control. *IEEE Transactions on Automatic Control*, 37(6):770–784, 1992.

[22] K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook.* Technical University of Denmark, October 2008. Version 20081110.

[23] Angelo Barboni and Thomas Parisini. Towards distributed accommodation of covert attacks in interconnected systems, 2020.

[24] G. Marchesini and E. Fornasini. *Appunti di teoria dei sistemi*. Libreria Progetto, 1983.

[25] E. Yip and R. Sincovec. Solvability, controllability, and observability of continuous descriptor systems. *IEEE Transactions on Automatic Control*, 26(3):702–707, 1981.

[26] David G. Luenberger. Time-invariant descriptor systems. *Automatica*, 14(5):473 – 480, 1978.

[27] Rick Grundy. Recommended data center temperature and humidity, November 2005.

[28] F. Zanini, D. Atienza, L. Benini, and G. De Micheli. Multicore thermal management with model predictive control. In *2009 European Conference on Circuit Theory and Design*, pages 711–714, 2009.

[29] O. Semenov, A. Vassighi, and M. Sachdev. Impact of self-heating effect on long-term reliability and performance degradation in cmos circuits. *IEEE Transactions on Device and Materials Reliability*, 6(1):17–27, 2006.

[30] MathWorks. Explicit mpc. https://it.mathworks.com/help/mpc/ug/explicit-mpc.html.

[31] L. Finesso. *Lezioni di probabilità*. Progetto Libreria, 2017.

[32] Kevin Skadron, Mircea R. Stan, Karthik Sankaranarayanan, Wei Huang, Sivakumar Velusamy, and David Tarjan. Temperature-aware microarchitecture: Modeling and implementation. *ACM Trans. Archit. Code Optim.*, 1(1):94–125, March 2004.

[33] Seongmoo Heo, K. Barr, and K. Asanovic. Reducing power density through activity migration. In *Proceedings of the 2003 International Symposium on Low Power Electronics and Design, 2003. ISLPED '03.*, pages 217–222, 2003.

[34] Mehran Mesbahi and Magnus Egerstedt. *Graph Theoretic Methods in Multiagent Networks*. Princeton University Press, stu - student edition edition, 2010.

[35] L. Xiao, Stephen Boyd, and Sanjay Lall. Distributed average consensus with time-varying metropolis weights. *Automatica*, 01 2006.

BIBLIOGRAPHY

Acknowledgments

I was supposed to work on my whole thesis project at the Imperial College London, thanks to the Erasmus+ programme. However, the COVID-19 pandemic forced me back in Padova ahead of schedule, after just a third of my expected time in London had elapsed. Despite much shorter than planned, my stay at the Imperial College was great, and I want to thank professor Parisini and all the PhD students who warmly welcomed me and extended my horizons with their perspectives. In particular, I really want to thank Angelo for the invaluable support he gave me during these months, both in person and remotely. He taught me a lot in spite of the inconvenient condition we were in, and we also had a lot of fun.

My experience at the University of Padova was excellent. Not only I found a positive and stimulating environment, but I am particularly pleased to say I also received an incredible support by several professors, who inspired me and helped me develop my ambitious plans over the years. Among them, professors Beghi, Cenedese, and Valcher are especially worth my gratitude for the enthusiastic help provided in finding my way before, through, and after my university path.

Moreover, I am grateful to my family, who represented an irreplaceable landmark in my university career, and always encouraged me to do my best. Without such an advice, I would not have achieved all the goals I am proud of.

Furthermore, I want to thank all the friends who shared this years with me, either studying together, having fun, or discussing about life and the future. Without them, everything would have made much less sense.

Really last, but not least, a special thanks goes to Francesca. Despite being apart for the last two years, she never missed the chance to let me feel her commitment. Beyond this, I found the determination she showed in adapting to a new world really inspiring.