Detection of Covert Cyber-Attacks in Interconnected Systems: A Distributed Model-Based Approach

Angelo Barboni, *Student Member, IEEE*, Hamed Rezaee, *Member, IEEE*, Francesca Boem, *Member, IEEE*, and Thomas Parisini, *Fellow, IEEE*

***Abstract*—Distributed detection of covert attacks for linear large-scale interconnected systems is addressed in this paper. Existing results consider the problem in centralized settings. This work focuses on large-scale systems subject to bounded process and measurement disturbances, where a single subsystem is under a covert attack. A detection methodology is proposed, where each subsystem can detect the presence of covert attacks in neighboring subsystems in a distributed manner. The detection strategy is based on the design of two model-based observers for each subsystem using only local information. An extensive detectability analysis is provided and simulation results on a power network benchmark are given, showing the effectiveness of the proposed methodology for the detection of covert cyber-attacks.**

I. Introduction

CRITICAL infrastructures such as, for example, electric power systems, water distribution networks, telecommunication networks, transportation systems, and industrial processes are nowadays large-scale systems that are interconnected not only on the physical layer but through a communication infrastructure thus increasing the vulnerability to external cyber-attacks. Security concerns related to these systems include both physical security and cyber-security, as well as combined cyber-physical threats. Indeed, in recent years, the security challenge has become a vital technological issue, especially after the occurrence of incidents involving industrial plants and critical infrastructures (see [1], [2]).

Due to the complexity of these systems and the computational and communication constraints, the development of *distributed* methodologies for monitoring and detection of malicious cyber-attacks has become a necessity. Recently developed comprehensive techniques for distributed fault diagnosis (see, for instance the recent works [3], [4] and the references cited therein) may not be fully effective in detecting cyber-attacks [5], as they are typically carried out by intelligent and active agents. This difficulty has inspired a large stream of research efforts (see, for example the seminal works [6]–[11], the more recent ones [12]–[14], and the surveys [15], [16], as well as the references cited therein).

This paper deals with a distributed methodology towards the detection of a particularly harmful class of stealthy cyber-attacks, namely the so-called *covert attacks* [17]. The proposed approach is specifically designed for spatially-distributed networked large-scale interconnected systems. In the remaining part of this section, after providing a glimpse on the state of the art, the specific contributions will be illustrated and the organization of the paper will be outlined.

A. A Glimpse on the State of the Art

As mentioned before, the problem of detecting and isolating cyber-attacks plays a central role in secure control systems. In this respect, some approaches in the literature related to security of cyber-physical systems stem from prior research in the field of fault detection and isolation (FDI), a well-established research area whose aim is to detect (and possibly identify the source of) faulty modes of behavior of the monitored system. In this connection, several contributions proposing distributed FDI techniques are available (see, for instance [18]–[25]), but extending these approaches to successfully detect a large class of malicious cyber-attacks has not yet happened, to the best of our knowledge. The main complexities arise from the inherent limitations in the presence of attacks that affect the system behavior in a much different way as compared with typical classes of faults and malfunctions.

Differently from studies on cyber-security in the computer science research community, most techniques in the control literature on attack detection and isolation take advantage of a dynamic model of the interconnected system to detect whether the communicated information in the control-loop has been corrupted by malicious attacks [26]. As already anticipated, this paper focuses on a model-based distributed attack-detection methodology and only quite a limited number of related works can be found in the literature (see [27]–[31]). Specifically, in [27], [28], a distributed methodology is presented to detect attacks for interconnected subsystems in which the communication infrastructure is assumed to be secure, in [29], the knowledge of the model of the entire system is required. In the recent conference paper [30], only attacks on the communication network between controllers and monitoring units are considered in a DC micro-grid application scenario, while in [31], the performance of distributed and decentralized detectors is analyzed in a statistical framework.

This work has been partially supported by the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1), by the European Union's Horizon 2020 research and innovation programme under grant agreement No 739551 (KIOS CoE) and by the Italian Ministry for Research in the framework of the 2017 Program for Research Projects of National Interest (PRIN), Grant no. 2017YKXYXJ.

A. Barboni is with the Dept. of Electrical and Electronic Engineering, Imperial College London, London, UK (e-mail: a.barboni16@imperial.ac.uk).

H. Rezaee is with the Dept. of Electrical and Electronic Engineering, Imperial College London, London, UK (e-mail: h.rezaee@imperial.ac.uk).

F. Boem is with the Dept. of Electronic and Electrical Engineering at University College London, London, UK (f.boem@ucl.ac.uk).

T. Parisini is with the Dept. of Electrical and Electronic Engineering, Imperial College London, London, UK. He is also with the Dept. of Engineering and Architecture, University of Trieste, 34127 Trieste, Italy, and with the KIOS Research and Innovation Center of Excellence, University of Cyprus, CY-1678 Nicosia, Cyprus (e-mail: t.parisini@imperial.ac.uk).

The family of covert cyber-attacks considered in the paper may have a detrimental impact on the physical layer: a covert agent injects some undesired control actions in the networked actuation channels while "canceling" its effects on the measurements. In this way, under the assumption of perfect knowledge of the system model by the attacker, the state of the system can be arbitrarily driven to potentially unsafe state trajectories without any trace in the monitoring units. In fact, due to the attack, the sensing layer communicates measurements which are consistent with the normal behavior, thus making the attack undetectable.

A few works have considered this scenario: for instance, in [32], an intelligent type of covert attacks is presented using system identification tools; in [33], the problem of covert attack detection in cyber-physical systems is investigated and a random modulation is introduced on the system actuation side to cause errors in the attacker's model. In the very recent work [34], resiliency versus covert attacks is formulated as an $\mathcal{H}_2$ optimal control problem. However, the literature in the area of detection and isolation of covert attacks is still limited with many open research problems worth investigation. In particular, to the best of the authors' knowledge, the problem of distributed model-based detection of covert attacks on large-scale networked systems has still not been addressed.

B. Objectives and Contributions

In this paper[1], a distributed covert attack detection architecture is proposed in which each locally controlled subsystem is equipped with two local state observers that use different information. The first observer is designed using a local model of its respective subsystem and uses both information provided by local sensors and information communicated from neighboring subsystems (for this reason, this observer is called *distributed*). The second observer is an unknown-input one and uses only locally available information and measurements (hence this observer is called *decentralized*). On the basis of the local estimates provided by the observers, an attack detection strategy is devised that, under suitable conditions, allows the detection of covert attacks not otherwise possible by a fully decentralized approach or by traditional distributed observation methods.

The main specific contributions are:

- Definition of a state-space characterization of the covert property of *man-in-the-middle* local attacks in the context of large-scale interconnected systems.
- Design of a distributed observer-based estimation technique for detecting covert attacks.
- Sufficient detectability conditions and convergence analysis, in the case where the measurements and the process are affected by bounded disturbances.
- Validation of the proposed distributed detection technique via simulation on a power network benchmark problem.

C. Main Notations

The following notation is used throughout the paper. $\mathbb{R}$ denotes the set of real numbers. I is an identity matrix with compatible dimensions. $\hat{v}$ is the estimated value of the variable v. $\mathcal{L}_2$ is the space of signals with bounded energy. For a vector v, $v_{[l]}$ denotes its l-th component. $\|\cdot\|_2$ stands for the Euclidean norm of a matrix. $\|\cdot\|_{\mathcal{L}_2}$ denotes the $\mathcal{L}_2$ norm of a signal. $\chi(t)$ stands for a step function. $\|\cdot\|_\infty$ stands for the $\mathcal{H}_\infty$ norm of a transfer function. $\mathrm{diag}(\cdot)$ describes a block diagonal matrix composed of a set of matrices. We say a matrix $M > 0$ (or $M < 0$) if it is symmetric positive (negative) definite. We denote by $|M|$ the entry-wise absolute value of a matrix M. Moreover, we define a concatenation operation over a finite indexed family of matrices $(M_i \in \mathbb{R}^{p\times *})_{i\in\mathcal{I}}$ with index set $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$ as $\mathrm{row}_{i\in\mathcal{I}}(M_i) \doteq (M_{i_1}|\ldots|M_{i_n})$.

[1]Early results for the disturbance-free case have been presented in [35].

D. Paper Organization

The paper is organized as follows. In the next section, the problem dealt with is formulated in detail, including the description of the covert attacks, the architecture of the estimation scheme and the detection decision strategy. Section III illustrates the design of the two local observers and provides the convergence analysis, and Section IV presents the attack detection methodology and the related detectability analysis. Section V reports extensive simulation results on a power network benchmark problem and concluding remarks are given in Section VI.

II. Problem Statement

Consider a large-scale system (LSS) composed of N interconnected subsystems, with the ith subsystem described as

$$\mathcal{S}_i : \begin{cases} \dot{x}_i = A_i x_i + B_i \tilde{u}_i + \sum_{j\in\mathcal{N}_i} A_{ij} x_j + w_i \\ y_i = C_i x_i + v_i \end{cases} \tag{1}$$

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 external disturbance vectors. The set $\mathcal{N}_i$ of neighbors of $\mathcal{S}_i$ is defined as the index set of those systems $\mathcal{S}_j$ whose states x_j appear as an argument in the state equation of $\mathcal{S}_i$. Moreover, $A_i \in \mathbb{R}^{n_i\times n_i}$ denotes the state matrix, $B_i \in \mathbb{R}^{n_i\times m_i}$ is the input matrix, $C_i \in \mathbb{R}^{p_i\times n_i}$ is the output matrix, and $A_{ij} \in \mathbb{R}^{n_i\times n_j}$ describes the dynamic interconnection influence of $\mathcal{S}_j$ on $\mathcal{S}_i$.

Remark 1: The dynamic interconnection characterized by index set $\mathcal{N}_i$ and constant interconnection matrices A_{ij} does not change over time and typically has a precise physical meaning, i.e. the interconnected state variables could be – for instance – currents, forces, flows, etc., depending of the type of system being modeled. ◁

Assumption 1: $\forall i \in \{1,\ldots,N\}$, the pair (A_i, C_i) is observable. ◁

Assumption 2: $\forall i \in \{1,\ldots,N\}$ and $\forall t$ there exist known positive constants $\bar{w}_i$ and $\bar{v}_i$ such that $\|w_i\| < \bar{w}_i$ and $\|v_i\| < \bar{v}_i$. Moreover, w_i, v_i, $\dot{v}_i \in \mathcal{L}_2$. ◁

The proposed detection architecture is shown in Fig. 1. Each subsystem is equipped with a local unit LU_i composed of a given controller $\mathcal{C}_i$ and a detector $\mathcal{D}_i$. The local measurements

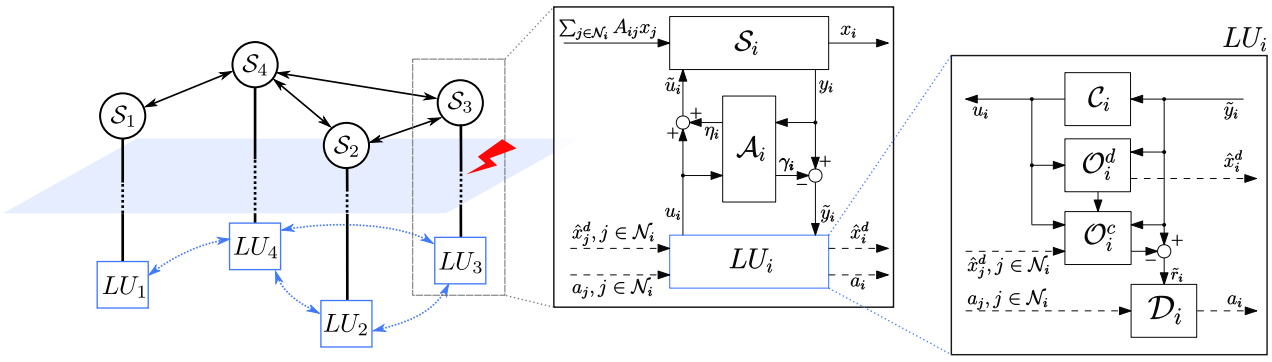


Fig. 1. Top-down architecture of the considered system. From left to right, the general layout can be seen, with the separation of physical and cyber layers. In the middle, the diagram of the attacked subsystem equipped with a local unit LU_i; finally, on the right, the detection architecture is further specialized in the two observers and the detection logic block $\mathcal{D}_i$.

available to LU_i are represented by $\tilde{y}_i \in \mathbb{R}^{p_i}$, while the control input computed by $\mathcal{C}_i$ is u_i.

By "local" we mean that each unit does not need any information about the overall topology of the LSS, but only exploits the model information and variables appearing in (1).

The variable $\tilde{y}_i$ denotes the measurement received by LU_i via a possibly vulnerable link (see Fig. 1). Due to the action of the attacker, $\tilde{y}_i$ can be different from y_i. Because of this possible discrepancy, we denote u_i and y_i as the *legitimate* or *transmitted* signals, and $\tilde{u}_i$ and $\tilde{y}_i$ as the *attacked* or *received* ones, respectively.

If i denotes the index of the subsystem under attack, we assume that the attacker $\mathcal{A}_i$ performs a man-in-the-middle attack and injects undesirable signals γ_i and η_i in the tapped link between the plant and the local unit such that

$$\begin{aligned} \tilde{y}_i &= y_i - \gamma_i \\ \tilde{u}_i &= u_i + \eta_i. \end{aligned} \tag{2}$$

The main difficulty in detecting such cyber-attacks is that γ_i and η_i can be designed by the attacker such that the attack effect is covert and not distinguishable from the nominal behavior. This important aspect is explained in more detail in the following subsections.

A. Covert Attack Model

In this section, we present a state space model for a covert attacker along the lines of [17].

Definition 1 (Covert agent): The malicious agent $\mathcal{A}_i$ is covert to subsystem $\mathcal{S}_i$ if the attacked measurement output $\tilde{y}_i$ is indistinguishable from the legitimate subsystem response y_i. ◁

An attacker is covert if it can hide its effect on the system such that the measured output is compatible with an attack-free behavior (we sometimes refer to this as *covertness property*). In this respect, we point out that covert attacks are stealthy by design. Since by Definition 1 the attacked measurements are indistinguishable from the nominal response, it follows that any residual signal relying on them necessarily satisfies the stealthiness condition in [9, Definition 2].

A covert strategy can be fulfilled by replicating the dynamics of the targeted system. Hence, the malicious agent $\mathcal{A}_i$ is modeled as a dynamical system

$$\tilde{\mathcal{S}}_i : \begin{cases} \dot{\tilde{x}}_i = \tilde{A}_i\tilde{x}_i + \tilde{B}_i\eta_i \\ \gamma_i = \tilde{C}_i\tilde{x}_i\,. \end{cases} \tag{3}$$

In particular, η_i is a signal that is chosen by the attacker to potentially steer the system towards some undesired trajectory. Because such a signal is arbitrary, its characteristics are in general unknown to a defender. As a result, the model (3) is in principle sufficient for describing a covert agent. In addition, the attacker may need to implement its own controller $\tilde{\mathcal{C}}_i$ in order to achieve some desired dynamics:

$$\tilde{\mathcal{C}}_i : \begin{cases} \dot{\xi}_i = A_{Ci}\xi_i + \Upsilon_i \begin{bmatrix} u_i \\ y_i \end{bmatrix} + R_{Ci}\rho_i \\ \eta_i = C_{Ci}\xi_i + K_{Ci}\tilde{x}_i\,. \end{cases} \tag{4}$$

In (4), $\xi_i \in \mathbb{R}^{\nu_i}$ is the controller's state and $\rho_i \in \mathbb{R}^{r_i}$ is used to determine the controller's reference.

By choosing ρ_i, the attacker can more easily control the system to achieve its own objective, for instance causing instability or to track a reference different from the nominal one. Moreover, A_{Ci}, Υ_i, R_i, C_{Ci}, and R_{Ci} are matrices of compatible dimensions, and K_{Ci} provides a feedback from the state of $\tilde{\mathcal{S}}_i$, and Υ_i represents the *disclosure resources* (as in [9], from which we borrow the jargon of this section), identifying accessible information by the attacker.

Using (3) and (4), the attacker can be represented in compact form by introducing a vector $\zeta_i \doteq \begin{bmatrix} \tilde{x}_i^\top & \xi_i^\top \end{bmatrix}^\top \in \mathbb{R}^{n_i+\nu_i}$ as follows:

$$\mathcal{A}_i : \begin{cases} \dot{\zeta}_i = \Phi_i\zeta_i + \begin{bmatrix} 0 \\ \Upsilon_i \end{bmatrix} \begin{bmatrix} u_i \\ y_i \end{bmatrix} + \begin{bmatrix} 0 \\ R_i \end{bmatrix} \rho_i \\ \begin{bmatrix} \gamma_i \\ \eta_i \end{bmatrix} = \Gamma_i\zeta_i\,, \end{cases} \tag{5}$$

where

$$\Phi_i = \begin{bmatrix} \tilde{A}_i + \tilde{B}_iK_{Ci} & \tilde{B}_iC_{Ci} \\ 0 & A_{Ci} \end{bmatrix}, \quad \Gamma_i = \begin{bmatrix} \tilde{C}_i & 0 \\ K_{Ci} & C_{Ci} \end{bmatrix},$$

and Γ_i plays the role of the *disruption* resources, as it defines which channels among actuation and measurement the attacker can be compromised with malicious signals. With this description, the attacker $\mathcal{A}_i$ on $\mathcal{S}_i$ 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 $\tilde{\mathcal{C}}_i$ and ρ_i.

For example, ρ_i can be a reference signal to an unsafe or disrupting operating point of some equipment. By designing $\tilde{\mathcal{C}}_i$, the attacker can inject η_i such that $\mathcal{S}_i$ is driven to the said point and can compensate the misbehavior using γ_i as in (2). We also note that, depending on Υ_i, the attacker could account for the unknown reference (see [17] for more detail on the issue) as it would know the values of u_i.

Model (5) by itself does not satisfy the covertness property. To do so, $\tilde{\mathcal{S}}_i$ needs to be a realization of the same transfer function realized by $\mathcal{S}_i$. This can be easily achieved if the following assumption holds.

Assumption 3: The attacker has *perfect knowledge* of $(\tilde{A}_i, \tilde{B}_i, \tilde{C}_i) = (A_i, B_i, C_i)$, while has no knowledge of the dynamic interconnection with neighboring subsystems. ◁

Remark 2: By considering an *omniscient local attacker*, with Assumption 3, in our analysis we consider the worst-case scenario, where the attacker is the most difficult to detect. In fact, as it is shown later on, in this case local residuals are not influenced by the attacker, and this is consistent with the results in [17]. By proving that the proposed detection strategy works in the *perfect knowledge* case, we also cover less tight cases: an attacker with incomplete information is not fully covert and therefore easier to reveal by residual analysis. ◁

Assumption 3 holds in practice when model information can be obtained via some form of intelligence, either because the components used in a plant are known (like in the case of the Stuxnet worm [36]) or because such information is leaked. In addition, it is fair to assume that an attacker who can write on some channels can also read from those, and therefore the model can be identified by eavesdropping on the measurement and actuation signals [37].

For what concerns its resources, the attacker has to be able to disrupt all the measurements and actuation channels of a single subsystem, whilst no disclosure resources are needed.

Let $T_{ai} \geq 0$ be the time instant when the attack occurs (i.e. $\gamma_i = \eta_i = 0$ for $t < T_{ai}$). We present sufficient conditions for an attacker to be covert. Covertness can be seen as an asymptotic property if we focus on the steady-state response, but here we are also interested in addressing the transient behavior, given that our analysis is in the time domain.

Proposition 1: Under Assumption 3, there exists a γ_i such that the attack is covert as $t \to \infty$, if A_i is Hurwitz. Moreover, if the attacker sets $\tilde{x}_i(T_{ai}) = 0$, the attack is covert for all time instants $\forall A_i \in \mathbb{R}^{n_i \times n_i}$. □

Proof: Before occurrence of the attack, for $0 < t < T_{ai}$, $y_i = \tilde{y}_i$. Let us analyze the covert property for $t \geq T_{ai}$. By considering (1)–(4), the attacked subsystem's output can be written as

$$\begin{aligned} y_i(t) = C_i e^{A_i(t-T_{ai})} x_i(T_{ai}) + C_i \int_{T_{ai}}^{t} e^{A_i(t-\tau)} \\ \left[B_i(u_i(\tau) + \eta_i(\tau)) + \sum_{j \in \mathcal{N}_i} A_{ij} x_j(\tau) + w_i(\tau) \right] d\tau \\ + v_i(t). \end{aligned} \tag{6}$$

In this condition, given a choice of η_i, the effect of γ_i can be computed as

$$\begin{aligned} \gamma_i(t) = & \tilde{C}_i e^{\tilde{A}_i(t-T_{ai})} \tilde{x}_i(T_{ai}) \\ & + \tilde{C}_i \int_{T_{ai}}^{t} e^{\tilde{A}_i(t-\tau)} \tilde{B}_i \eta_i(\tau) d\tau, \end{aligned} \tag{7}$$

by using (2) and Assumption 3, one can observe that

$$\begin{aligned} \tilde{y}_i(t) = & C_i e^{A_i(t-T_{ai})} (x_i(T_{ai}) - \tilde{x}_i(T_{ai})) \\ & + C_i \int_{T_{ai}}^{t} e^{A_i(t-\tau)} \left[B_i u_i(\tau) + \sum_{j \in \mathcal{N}_i} A_{ij} x_j(\tau) \right. \\ & \left. + w_i(\tau) \right] d\tau + v_i(t). \end{aligned} \tag{8}$$

From (8), it follows that for $t \to \infty$, $\tilde{y}_i$ will be the same as the output of the attack-free subsystem (the legitimate output). In other words, by considering (6)–(8), for $t \to \infty$, $\tilde{y}_i$ will be identical to y_i when $\eta_i = 0$ (no attacks) if the first exponential term is vanishing. Moreover, if the attacker sets the initial conditions of (3) as $\tilde{x}_i(T_{ai}) = 0$, $\tilde{y}_i$ is equal to y_i when no attack is underway and the first exponential term is identically zero. Hence, the proof is completed. ■

Remark 3: Note that in Proposition 1, the attacker is covert without any knowledge about the neighbors or their interconnection. In fact, a purely local model (3) is used along with Assumption 3; this is sufficient to successfully carry out a covert attack on the subsystem. ◁

It is worth noting that the results stated in Proposition 1 are related to the ones given in [17] for the centralized case in the frequency domain but are more general in that we consider a distributed framework and the transient behaviors due to unknown initial conditions are taken into account.

Finally, we emphasize that both the definition of covert attacks and the results of Proposition 1 can equivalently be restated in terms of detection residuals, as will be discuss later.

B. Detector Architecture

We describe in more detail the design principles of the detector shown in Fig. 1. The proposed architecture is based on two observers for each local unit LU_i: a decentralized observer $\mathcal{O}_i^d$ (described in Subsection III-A) and a distributed one $\mathcal{O}_i^c$ (described in Subsection III-B). More specifically, $\mathcal{O}_i^d$ is designed such that its state estimate $\hat{x}_i^d$ is decoupled from the neighboring subsystems $\mathcal{S}_j, \forall j \in \mathcal{N}_i$, while $\mathcal{O}_i^c$ computes a state estimate $\hat{x}_i^c$ that depends on *communicated* neighboring estimates $\hat{x}_j^d, j \in \mathcal{N}_i$. By exploiting the cooperation of a decentralized decoupled estimation strategy and a distributed one, it is possible for the observers to reveal possible inconsistencies in the measurements from neighboring subsystems.

In this way, a perfectly covert attack in $\mathcal{N}_i$ can be revealed by detectors in all neighboring LU_j.

For every subsystem $\mathcal{S}_i$, we design a residual signal $\tilde{r}_i^c$ and a (time-varying) threshold $\bar{r}_i$, whose definition and properties will be discussed later. In order to reveal stealthy attacks, the following *distributed* detection logic is implemented by the diagnoser $\mathcal{D}_i$ in Fig. 1:

- If $\|\tilde{r}_i^c\| > \bar{r}_i$, then a binary alarm signal $a_i = 1$ is raised.
- Each $\mathcal{S}_i$ broadcasts $a_i \in \{0, 1\}$ to its neighbors $\mathcal{S}_j, \forall j \in \mathcal{N}_i$. Conversely, it also receives a set of signals $a_j, j \in \mathcal{N}_i$, from the neighbors.
- If for any i, $a_j = 1, \forall j \in \mathcal{N}_i$, then detector $\mathcal{D}_i$ decides that $\mathcal{S}_i$ is under attack.

III. Observer Design

According to the definitions of $\hat{x}_i^d$ and $\hat{x}_i^c$ of the previous section, the output and state estimation errors for the distributed and decentralized observers are defined as follows:

$$\begin{aligned} r_i^d &= y_i - C_i\hat{x}_i^d = C_i\epsilon_i^d + v_i, \qquad \epsilon_i^d \doteq x_i - \hat{x}_i^d, \\ r_i^c &= y_i - C_i\hat{x}_i^c = C_i\epsilon_i^c + v_i, \qquad \epsilon_i^c \doteq x_i - \hat{x}_i^c. \end{aligned} \tag{9}$$

It should be noted 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. We refer to these quantities as the *true* errors, and they cannot be computed in practice since the actual state of any subsystem is not directly accessible. However, the related residuals can be computed in the attack-free scenario according to the relations in the left part of (9).

On the other hand, when $\mathcal{S}_i$ is under attack (and if $\tilde{C}_i = C_i$, see Assumption 3), since $\tilde{y}_i \neq y_i$, the residuals computed by the subsystem are as follows:

$$\begin{aligned} \tilde{r}_i^d &= \tilde{y}_i - C_i\hat{x}_i^d = C_i\tilde{\epsilon}_i^d + v_i, \qquad \tilde{\epsilon}_i^d \doteq x_i - \tilde{x}_i - \hat{x}_i^d, \\ \tilde{r}_i^c &= \tilde{y}_i - C_i\hat{x}_i^c = C_i\tilde{\epsilon}_i^c + v_i, \qquad \tilde{\epsilon}_i^c \doteq x_i - \tilde{x}_i - \hat{x}_i^c. \end{aligned} \tag{10}$$

Similar to the conventions introduced in Section II, we refer to (10) as the *received* or *attacked* output and state error. Also note that when no attack is under way, (9) and (10) coincide.

Design details are presented in the following subsections. First, the decentralized observation strategy is introduced and then the distributed observer based on coupling among the subsystems is proposed.

Finally, we introduce the following assumption that will be instrumental to the design of the observers as illustrated in the next two subsections.

Assumption 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. ◁

A. Decentralized Observation Strategy

In order to obtain a state estimate $\hat{x}_i^d$ which is independent of the states of the neighboring subsystems, we implement an Unknown Input Observer (UIO) [38] for each subsystem i, where the interconnection among the subsystems is considered as an unknown input. It should be noted that the use of UIOs for distributed detection of anomalies is not new (for instance, see [24] and [30]). However, in this work, we combine it for the first time (to the best of the authors' knowledge) with a distributed observer, and derive conditions under which covert attacks in neighboring systems can be revealed.

Based on a UIO, the estimate $\hat{x}_i^d$ can be obtained from the following dynamical system:

$$\mathcal{O}_i^d : \begin{cases} \dot{z}_i = F_i z_i + T_i B_i u_i + K_i \tilde{y}_i \\ \hat{x}_i^d = z_i + H_i \tilde{y}_i, \end{cases} \tag{11}$$

where F_i, T_i, K_i, and H_i are matrices with compatible dimensions later designed. First, let us define Ξ_i and $\mathbf{x}_i$ as

$$\Xi_i \doteq \mathrm{row}_{j\in\mathcal{N}_i}(A_{ij}),\ \mathbf{x}_i^\top \doteq \mathrm{row}_{j\in\mathcal{N}_i}(x_j^\top),$$

implying that for the ith subsystem, the effect of the neighbors' interconnection can be restated in a vector form as follows:

$$\sum_{j\in\mathcal{N}_i} A_{ij}x_j = \Xi_i \mathbf{x}_i\,.$$

Based on these definitions, the following conditions on the observer (11) are required [38, Theorem 1]:

a) $\mathrm{rank}(C_i\Xi_i) = \mathrm{rank}(\Xi_i)$
b) The pair $(C_i, \bar{A}_i)$ is detectable, where

$$\bar{A}_i = A_i - H_i C_i A_i.$$

Under these conditions, by decomposing K_i as $K_i = K_i^{(1)} + K_i^{(2)}$, it is possible to design F_i, T_i, K_i, and H_i such that

$$0 = (H_i C_i - I)\Xi_i, \tag{12a}$$

$$T_i = I - H_i C_i, \tag{12b}$$

$$F_i = \bar{A}_i - K_i^{(1)} C_i \text{ is Hurwitz}, \tag{12c}$$

$$K_i^{(2)} = F_i H_i. \tag{12d}$$

Under Condition a), we can compute the matrix $H_i = \Xi_i[(C_i\Xi_i)^\top C_i\Xi_i]^{-1}(C_i\Xi_i)^\top$ from (12a) which decouples the unknown inputs, whereas Condition b) implies that F_i can be obtained from (12c). By considering (11) and (12), we can derive the dynamical equations of the estimation error ϵ_i^d as follows:

$$\dot{\epsilon}_i^d = F_i\epsilon_i^d + T_i w_i - K_i^{(1)} v_i - H_i \dot{v}_i\,. \tag{13}$$

From (13) and the fact that F_i is Hurwitz, it follows that for a disturbance-free subsystem, the estimation error ϵ_i^d converges to zero. For a subsystem with bounded disturbances, the estimation error is bounded. It should be noted that (13) holds when the subsystem is not under attack, i.e., when the actuation and measurement channels are not corrupted.

Proposition 2: Let the ith subsystem be under the attack modeled in (5) and (2) and let Assumption 3 hold. Under the UIO conditions (12), the estimation error dynamics for the observer (11) are

$$\begin{aligned} \dot{\epsilon}_i^d = F_i\epsilon_i^d + T_i w_i - K_i^{(1)} v_i - H_i \dot{v}_i \\ + (A_i - F_i)\tilde{x}_i + B_i \eta_i, \end{aligned} \tag{14}$$

while the attacked estimation error is

$$\dot{\tilde{\epsilon}}_i^d = F_i\tilde{\epsilon}_i^d + T_i w_i - K_i^{(1)} v_i - H_i \dot{v}_i\,. \tag{15}$$

Furthermore, the attack is covert for the observer (11). □

Proof: See Appendix. ■

Remark 4: In Proposition 2, it is shown that for the proposed covert attack the received error $\tilde{\epsilon}_i^d$ has the same dynamics of the attack-free one ϵ_i^d, and by using Eqs. (10), we can state the following:

$$|\|\epsilon_i^d\| - \|\tilde{x}_i\|| \leq \|\epsilon_i^d - \tilde{x}_i\| = \|\tilde{\epsilon}_i^d\|.$$

By using triangle inequality, this leads to

$$\|\epsilon_i^d\| \geq \|\tilde{x}_i\| - \|\tilde{\epsilon}_i^d\|.$$

As a result, a covert attacker can maliciously increase the lower bound on the *true* error of the attacked subsystem by increasing the norm of its own internal state. ◁

B. Distributed Observation Strategy

By considering the interconnection model and by using the information received from neighboring subsystems, a distributed observation strategy is developed for each subsystem to estimate the value of its own state vector.

Assumption 5: We assume ideal communication between subsystems. As such, the exchanged estimates $\hat{x}_j^d$, $j \in \mathcal{N}_i$ are not corrupted during communication. ◁

By considering the subsystems dynamical equations given in (1), the distributed state-observer $\mathcal{O}_i^c$ is described by the following:

$$\dot{\hat{x}}_i^c = A_i\hat{x}_i^c + B_iu_i + \sum_{j\in\mathcal{N}_i} A_{ij}\hat{x}_j^d + L_i(y_i - \hat{y}_i^c), \tag{16}$$

where $L_i \in \mathbb{R}^{n_i\times p_i}$ is the observer gain to be designed later and $\hat{y}_i^c = C_i\hat{x}_i^c$.

Remark 5: It should be noted that, in (16), we have used $A_{ij}\hat{x}_j^d$ instead of $A_{ij}\hat{x}_j^c$, because the value of $\hat{x}_j^d$ is not affected by attacks in neighboring subsystems $k, k \in \mathcal{N}_j$. This property will lay the basis for our detection strategy in the next section. ◁

To design the observer gain L_i, an $\mathcal{H}_\infty$ optimization approach is employed. The gain L_i is designed such that the effect of the exogenous signals vector $\varpi_i = \begin{bmatrix} w_i^\top & v_i^\top & \sum_{j\in\mathcal{N}_i}\epsilon_j^{d\top}A_{ij}^\top \end{bmatrix}^\top$ is attenuated on the observer error ϵ_i^c. To achieve this goal, the induced norm of the $\mathcal{L}_2$ norm of ϵ_i^c and the $\mathcal{L}_2$ norm of ϖ_i is minimized as follows:

$$\min_{L_i}\sup_{\varpi}\frac{\|\epsilon_i^c\|_2}{\|\varpi_i\|_2} = \min_{L_i}\|T_{\epsilon_i^c\varpi_i}\|_\infty,$$

where $T_{\epsilon_i^c\varpi_i}$ is the transfer function from ϖ_i to ϵ_i^c, i.e.,

$$\min_{L_i}\lambda_i \text{ s.t. } \left(\|\epsilon_i^c\|_2 - \lambda_i\|\varpi_i\|_2\right) < 0, \tag{17}$$

where $\lambda_i > 0$. It should be noted that Hurwitz stability of the decentralized observer $\mathcal{O}_i^d$, introduced in the previous subsection, guarantees the $\mathcal{L}_2$-boundedness of ϵ_j^d. Thus, ϵ_j^d can be considered as an exogenous signal in ϖ_i.

Before presentation of the main results, let us introduce the following lemma.

Lemma 1: [39] The $\mathcal{H}_\infty$ performance (17) is satisfied if for a Lyapunov candidate V_i, $J_i = \dot{V}_i + \epsilon_i^{c\top}\epsilon_i^c - \lambda_i^2\varpi_i^\top\varpi_i$ is negative definite. □

Theorem 1: Consider the LSS described in (1) and the observer introduced in (16). The estimation errors $\epsilon_i^c, i \in \{1, 2, \ldots, N\}$, converge to zero and the $\mathcal{H}_\infty$ performance (17) is achieved if L_i satisfies $\forall i$ the following linear matrix inequality (LMI) for some P_i and S_i:

$$\begin{aligned}&\min_{L_i}\lambda_i \text{ s.t.}\\ &W_i = \begin{bmatrix} \Pi_i & P_i & -S_i & P_i \\ P_i & -\lambda_i^2 I & 0 & 0 \\ -S_i^\top & 0 & -\lambda_i^2 I & 0 \\ P_i & 0 & 0 & -\lambda_i^2 I \end{bmatrix} < 0,\end{aligned} \tag{18}$$

where $P_i \in \mathbb{R}^{n_i\times n_i}$ is a symmetric positive definite matrix, $S_i \in \mathbb{R}^{n_i\times p_i}$ which yields $L_i = P_i^{-1}S_i$, and $\Pi_i = P_iA_i - S_iC_i + A_i^\top P_i - C_i^\top S_i^\top + I$. □

Proof: Since $\hat{y}_i^c = C_i\hat{x}_i^c$, (16) can be restated as

$$\dot{\hat{x}}_i^c = (A_i - L_iC_i)\hat{x}_i^c + B_iu_i + \sum_{j\in\mathcal{N}_i} A_{ij}\hat{x}_j^d + L_iy_i.$$

By considering (1) and since $\epsilon_i^c = x_i - \hat{x}_i^c$, the error dynamics can be written as

$$\begin{aligned}\dot{\epsilon}_i^c =& A_ix_i + B_iu_i + \sum_{j\in\mathcal{N}_i} A_{ij}x_j + w_i \\ &- (A_i - L_iC_i)\hat{x}_i^c - B_iu_i - \sum_{j\in\mathcal{N}_i} A_{ij}\hat{x}_j^d \\ &- L_iy_i.\end{aligned} \tag{19}$$

Since $y_i = C_ix_i + v_i$, after some manipulation from (19), it follows that

$$\dot{\epsilon}_i^c = (A_i - L_iC_i)\epsilon_i^c + \sum_{j\in\mathcal{N}_i} A_{ij}\epsilon_j^d + w_i - L_iv_i. \tag{20}$$

According to Lemma 1, to satisfy a desirable $\mathcal{H}_\infty$ performance, we should have

$$\begin{aligned}J_i =& \dot{V}_i + \epsilon_i^{c\top}\epsilon_i^c - \lambda_i^2 v_i^\top v_i - \lambda_i^2 w_i^\top w_i \\ &- \lambda_i^2\sum_{j\in\mathcal{N}_i}\epsilon_j^{d\top}A_{ij}^\top A_{ij}\epsilon_j^d < 0.\end{aligned}$$

By defining $V_i = \epsilon_i^{c\top}P_i\epsilon_i^c$ and by considering the time derivative of V_i along (20), J_i can be obtained as

$$\begin{aligned}J_i =& \epsilon_i^{c\top}P_i(A_i - L_iC_i)\epsilon_i^c + \epsilon_i^{c\top}(A_i - L_iC_i)^\top P_i\epsilon_i^c \\ &+ \epsilon_i^{c\top}P_i\sum_{j\in\mathcal{N}_i}A_{ij}\epsilon_j^d + \sum_{j\in\mathcal{N}_i}\epsilon_j^{d\top}A_{ij}^\top P_i\epsilon_i^c \\ &+ \epsilon_i^{c\top}P_iw_i + w_i^\top P_i\epsilon_i^c - \epsilon_i^{c\top}P_iL_iv_i - v_i^\top L_i^\top P_i\epsilon_i^c \\ &+ \epsilon_i^{c\top}\epsilon_i^c - \lambda_i^2 v_i^\top v_i - \lambda_i^2 w_i^\top w_i - \lambda_i^2\sum_{j\in\mathcal{N}_i}\epsilon_j^{d\top}A_{ij}^\top A_{ij}\epsilon_i^d.\end{aligned}$$

Let $S_i = P_iL_i$, then J_i can be simplified as follows:

$$J_i = \begin{bmatrix}\epsilon_i^{c\top} & \varpi_i^\top\end{bmatrix}\begin{bmatrix} \Pi_i & P_i & -S_i & P_i \\ P_i & -\lambda_i^2 I & 0 & 0 \\ -S_i^\top & 0 & -\lambda_i^2 I & 0 \\ P_i & 0 & 0 & -\lambda_i^2 I \end{bmatrix}\begin{bmatrix}\epsilon_i^c \\ \varpi_i\end{bmatrix}.$$

In this condition, we have $J_i < 0$ if the LMI (18) is satisfied. Therefore, the proof is completed. ■

Remark 6: Note that since the pair (A_i, C_i) is observable, for any symmetric positive definite $Q_i \in \mathbb{R}^{n_i\times n_i}$, there exists

an L_i such that the Lyapunov equation $(A_i - L_iC_i)^\top P_i + P_i(A_i - L_iC_i) = -Q_i$ has a solution $\bar{P}_i$ implying that the LMI $\Pi_i < 0$ always has a solution $\bar{S}_i$ and $\bar{P}_i$. As a result, the Schur complement of the block Π_i of the matrix W_i is negative definite for some λ_i (see [40] for the theory of the Schur complement), and therefore $W_i < 0$ always has solutions. ◁

Remark 7: The $\mathcal{H}_\infty$ optimization technique proposed in Theorem 1 is also useful to design $K_i^{(1)}$ for the decentralized observer (11) such that the effect of the exogenous signals vector $\acute{\varpi}_i = \begin{bmatrix} w_i^\top T_i^\top & v_i^\top & -\dot{v}_i^\top H_i^\top \end{bmatrix}^\top$ is attenuated on the observer error ϵ_i^d. Therefore, following a logic similar to the proof of Theorem 1, $K_i^{(1)}$ can be obtained from the following optimization problem:

$$\min_{K_i^{(1)}} \acute{\lambda}_i \text{ s.t.}$$

$$\acute{W}_i = \begin{bmatrix} \acute{\Pi}_i & \acute{P}_i & -\acute{S}_i & \acute{P}_i \\ \acute{P}_i & -\acute{\lambda}_i^2 I & 0 & 0 \\ -\acute{S}_i^\top & 0 & -\acute{\lambda}_i^2 I & 0 \\ \acute{P}_i & 0 & 0 & -\acute{\lambda}_i^2 I \end{bmatrix} < 0$$

where $\acute{P}_i \in \mathbb{R}^{n_i \times n_i}$ is a symmetric positive definite matrix, $\acute{S}_i \in \mathbb{R}^{n_i \times p_i}$ which yields $K_i^{(1)} = \acute{P}_i^{-1}\acute{S}_i$, and $\acute{\Pi}_i = \acute{P}_i\bar{A}_i - \acute{S}_iC_i + \bar{A}_i^\top\acute{P}_i - C_i^\top\acute{S}_i^\top + I$. ◁

Proposition 3: Let the ith subsystem be under the attack modeled in (5) and (2) and Assumption 3 hold. The actual estimation error dynamics for observer (16) are

$$\begin{aligned}\dot{\epsilon}_i^c =& (A_i - L_iC_i)\epsilon_i^c + w_i - L_iv_i \\ &+ \sum_{j\in\mathcal{N}_i} A_{ij}\epsilon_j^d + B_i\eta_i + L_i\gamma_i,\end{aligned} \tag{21}$$

while the computed attacked estimation error is

$$\dot{\tilde{\epsilon}}_i^c = (A_i - L_iC_i)\tilde{\epsilon}_i^c + \sum_{j\in\mathcal{N}_i} A_{ij}\epsilon_j^d + w_i - L_iv_i. \tag{22}$$

Furthermore, the attack is covert for the observer (16). □

Proof: The proof is readily obtained by combining the observer formulation (16) with the attacker model in (2) and (3). Since (22) and (20) are identical, the attack is covert. ■

IV. Attack Detection Scheme

As anticipated in Subsection II-B, we monitor the behavior of the residual $\tilde{r}_i^c$ defined in (10) to trigger an attack alarm when the residual crosses a suitable threshold to be defined in order to take disturbances into account. It follows directly from Proposition 3 that the received error $\tilde{\epsilon}_i^c$ (and hence $\tilde{r}_i^c$) is sensitive to the *true* error in its neighbors.

A. Observers' Errors in Attack-Free Conditions

Since we are considering the possible presence of measurement and process disturbances, the proposed strategy requires the design of an appropriate threshold for the detection residuals such that the alarm binary variable is triggered *avoiding false-alarms*. The threshold can be obtained by considering the received errors in attack-free conditions.

Before proceeding with the analysis, we assume the following in order to rule out the more complex situation where two attackers in the same neighborhood may cooperate to compensate each other.

Assumption 6: For any subsystem $\mathcal{S}_i$, there is only one attacker in its neighborhood $\mathcal{N}_i$. ◁

Assumption 6 is in place only for the sake of analyzing the detectability property of the proposed scheme and it is not needed in general: i.e. there might be cases where detection is still possible with multiple attackers although the analysis becomes more complex. From a practical point of view, if the overall system is spread over a large area, it may be difficult for an attacker to target vast sections of it, especially since local control loops are targeted.

In order to simplify equations, we make use of the logarithmic norm $\mu(M)$ of a matrix M. This approach is relevant when deriving bounds as it can be shown that (see [41])

$$\mu(M) = \min\{\alpha : \|e^{Mt}\| \le e^{\alpha t}, t \ge 0\}.$$

Throughout this section, we will use the following inequality:

$$\|e^{Mt}\| \le e^{\mu(M)t}. \tag{23}$$

In the following result, we derive an upper bound for the estimation error of the decentralized observer.

Proposition 4: In attack-free conditions, the norm of the UIO error is bounded by a positive function $\bar{\epsilon}_i^d(t)$

$$\|\epsilon_i^d(t)\| \le \bar{\epsilon}_i^d(t),$$

where

$$\bar{\epsilon}_i^d(t) = h_ie^{\mu(F_i)t} + \|H_i\|\bar{v}_i - \frac{\|T_i\|\bar{w}_i + \|K_i\|\bar{v}_i}{\mu(F_i)},$$

with $h_i \doteq \|\epsilon_i^d(0)\| + \|H_i\|\bar{v}_i + \frac{\|T_i\|\bar{w}_i + \|K_i\|\bar{v}_i}{\mu(F_i)}$, and $\bar{v}_i$ and $\bar{w}_i$ defined in Assumption 2. □

Proof: Along the lines of [30], we integrate (13), obtaining

$$\begin{aligned}\epsilon_i^d(t) = e^{F_it}(\epsilon_i^d(0) + H_iv_i(0)) - H_iv_i(t) + \\ \int_0^t e^{F_i(t-s)}[T_iw_i(s) - K_iv_i(s)]ds,\end{aligned} \tag{24}$$

which can be bounded as follows:

$$\begin{aligned}\|\epsilon_i^d(t)\| &\le e^{\mu(F_i)t}(\|\epsilon_i^d(0)\| + \|H_i\|\bar{v}_i) + \|H_i\|\bar{v}_i \\ &\quad + (\|T_i\|\bar{w}_i + \|K_i\|\bar{v}_i)\int_0^t e^{\mu(F_i)(t-s)}ds \\ &= e^{\mu(F_i)t}(\|\epsilon_i^d(0)\| + \|H_i\|\bar{v}_i) \\ &\quad + \frac{\|T_i\|\bar{w}_i + \|K_i\|\bar{v}_i}{\mu(F_i)}\left(e^{\mu(F_i)t} - 1\right) + \|H_i\|\bar{v}_i \\ &= h_ie^{\mu(F_i)t} + \|H_i\|\bar{v}_i - \frac{\|T_i\|\bar{w}_i + \|K_i\|\bar{v}_i}{\mu(F_i)}.\end{aligned} \tag{25}$$

■

Remark 8: Derivation of (25) is correct only if $\mu(F_i) \neq 0$. Conditions in which this holds can be found in [41], however they easily hold for Hurwitz matrices. ◁

In the following proposition, we derive a threshold for the distributed detection residual r_i^c.

Proposition 5: In attack-free conditions, if $\mu(F_i^c) < 0$, the received residual is bounded by

$$\|\tilde{r}_i^c\| \leq \bar{r}_i(t) \doteq \|C_i\|\bar{\epsilon}_i(t) + \bar{v}_i,$$

where $\bar{\epsilon}_i \doteq \bar{\epsilon}_{z,i} + \bar{\epsilon}_{w,i} + \bar{\epsilon}_{k,i}$ and

$$\bar{\epsilon}_{z,i}(t) = \|\tilde{\epsilon}_i^c(0)\|e^{\mu(F_i^c)t} \tag{26}$$

$$\bar{\epsilon}_{w,i}(t) = (\bar{w}_i + \|L_i\|\bar{v}_i)\frac{e^{\mu(F_i^c)t} - 1}{\mu(F_i^c)} \tag{27}$$

$$\begin{aligned}\bar{\epsilon}_{k,i}(t) = \sum_{j\in\mathcal{N}_i}\|A_{ij}\|\Bigg[\big(\|H_j\|\bar{v}_j - \frac{\|T_j\|\bar{w}_j + \|K_j\|\bar{v}_j}{\mu(F_j)}\big) \\ \cdot\frac{e^{\mu(F_i^c)t} - 1}{\mu(F_i^c)} + h_j\frac{e^{\mu(F_j)t} - e^{\mu(F_i^c)t}}{\mu(F_j) - \mu(F_i^c)}\Bigg],\end{aligned} \tag{28}$$

with $F_i^c \doteq A_i - L_iC_i$. If $\mu(F_j) = \mu(F_i^c)$ for some i, j, we have instead

$$\begin{aligned}\bar{\epsilon}_{k,i}(t) = \sum_{j\in\mathcal{N}_i}\|A_{ij}\|\Bigg[\big(\|H_j\|\bar{v}_j - \frac{\|T_j\|\bar{w}_j + \|K_j\|\bar{v}_j}{\mu(F_j)}\big) \\ \cdot\frac{e^{\mu(F_i^c)t} - 1}{\mu(F_i^c)} + h_jte^{\mu(F_i^c)t}\Bigg].\end{aligned} \tag{29}$$

□

Proof: By considering (23) and integrating (22), we obtain

$$\begin{aligned}\|\tilde{\epsilon}_i^c(t)\| &= \Bigg\|e^{F_i^ct}\tilde{\epsilon}_i^c(0) + \sum_{j\in\mathcal{N}_i}\int_0^t e^{F_i^c(t-s)}A_{ij}\epsilon_j^d(s)ds \\ &\quad + \int_0^t e^{F_i^c(t-s)}(w_i(s) - L_iv_i(s))ds\Bigg\| \\ &\leq \|\tilde{\epsilon}_i^c(0)\|e^{\mu(F_i^c)t} \\ &\quad + \sum_{j\in\mathcal{N}_i}\|A_{ij}\|\int_0^t e^{\mu(F_i^c)(t-s)}\|\bar{\epsilon}_j^d(s)\|ds \\ &\quad + (\bar{w}_i + \|L_i\|\bar{v}_i)\frac{e^{\mu(F_i^c)t} - 1}{\mu(F_i^c)}.\end{aligned} \tag{30}$$

We can recognize (26) and (27) as the first and third terms of (30), respectively. After expanding $\|\epsilon_j^d(s)\|$ as per Proposition 4, the solution of the second integral yields $\bar{\epsilon}_{k,i}$ in (28). The thesis follows from norm properties. The special case (29) is obtained by expanding $\|\epsilon_j^d(s)\|$, which cancels the outer exponential and leads to the integration of a constant. ■

Remark 9: The considerations made in Remark 8 also apply to the computations in this theorem. In the limit case when $\mu(F_j) \approx \mu(F_i^c)$ for some i, j, it can be shown that

$$\lim_{\mu(F_j)\to\mu(F_i^c)}\frac{e^{\mu(F_j)t} - e^{\mu(F_i^c)t}}{\mu(F_j) - \mu(F_i^c)} = e^{\mu(F_i^c)t}.$$

◁

B. Detectability Analysis

In this section, we obtain some important results about attack detectability and detection time with the proposed distributed detection methodology. We consider a generic $\mathcal{S}_i$ and a single covert attack in one of its neighbors $k \in \mathcal{N}_i$, according to Assumption 6.

Theorem 2 (Detectability): A covert attack starting at time T_{ai} in $\mathcal{S}_k, k \in \mathcal{N}_i$, is detectable by $\mathcal{S}_i$ if $\exists\bar{t}_i > T_{ai}$ such that

$$\left\|\int_{T_{ai}}^{\bar{t}_i} e^{F_i^c(t-\tau)}A_{ik}\int_{T_{ai}}^{\tau} e^{F_k(\tau-s)}\alpha_k(s)dsd\tau\right\| > 2\bar{\epsilon}_i, \tag{31}$$

where $\alpha_k(s) = (A_k - F_k)\tilde{x}_k(s) + B_k\eta_k(s)$. □

Proof: To consider the attack effect, we integrate (13) and (14) before and after time T_{ai}, respectively. This leads to

$$\begin{aligned}\epsilon_k^d(t) &= e^{F_kt}\epsilon_k^d(0) \\ &+ \int_0^t e^{F_k(t-s)}\left(T_kw_k(s) - K_k^{(1)}v_k(s) - H_k\dot{v}_k(s)\right)ds \\ &+ \int_{T_{ai}}^t e^{F_k(t-s)}\alpha_k(s)ds.\end{aligned} \tag{32}$$

The first two terms consist in the attack-free error ϵ_k^d which corresponds to the received error $\tilde{\epsilon}_k^d$ in virtue of Proposition 2. Also, notice that this error expression has been expanded in (24) in Proposition 4. We can conveniently rewrite (32) as

$$\epsilon_k^d = \tilde{\epsilon}_k^d + \epsilon_k'^d,$$

where $\epsilon_k'^d(t) \doteq \int_0^t e^{F_k(t-s)}\alpha_k(s)\chi(s - T_{ai})ds$, and by integrating (22), we obtain

$$\begin{aligned}\tilde{\epsilon}_i^c(t) &= e^{F_i^ct}\tilde{\epsilon}_i^c(0) + \int_0^t e^{F_i^c(t-s)}(w_i(s) - L_iv_i(s))ds \\ &\quad + \sum_{j\in\mathcal{N}_i}\int_0^t e^{F_i^c(t-s)}A_{ij}\epsilon_j^d(s)ds \\ &\quad + \int_{T_{ai}}^t e^{F_i^c(t-\tau)}A_{ik}\int_{T_{ai}}^{\tau} e^{F_k(\tau-s)}\alpha_k(s)dsd\tau,\end{aligned} \tag{33}$$

where again the first three terms correspond to the attack-free received error $\tilde{\epsilon}_{i,\text{af}}^c$, and the last term is the attack contribution. Let us denote for brevity this last term with $\varphi_{i,k}$.

By applying the inverse triangle inequality and the bounds of Proposition 5, we obtain

$$\begin{aligned}\bar{\epsilon}_i \geq \|\varphi_{i,k} + \tilde{\epsilon}_{i,\text{af}}^c\| \geq \big|\|\varphi_{i,k}\| - \|\tilde{\epsilon}_{i,\text{af}}^c\|\big| \\ \|\varphi_{i,k}\| \leq \bar{\epsilon}_i + \|\tilde{\epsilon}_{i,\text{af}}^c\| \leq 2\bar{\epsilon}_i,\end{aligned}$$

which holds $\forall t$. By negating this condition we finally obtain (31). ■

Remark 10: In [35], it is pointed out that reachability of the pair (F_i^c, A_{ik}), is a necessary condition for attack detectability. However, this condition is implied by (31). With this, we remark the importance of interconnections on the attack detectability properties.

Corollary 1: A covert attack starting at time T_{ai} in $\mathcal{S}_k, k \in \mathcal{N}_i$, is detectable by $\mathcal{S}_i$ if $\exists\bar{t}_i > T_{ai}$ such that

$$\left\|C_i\int_{T_{ai}}^{\bar{t}} e^{F_i^c(t-\tau)}A_{ik}\int_{T_{ai}}^{\tau} e^{F_k(\tau-s)}\alpha_k(s)dsd\tau\right\| > 2\bar{r}_i. \tag{34}$$

□

Proof: Eq. (34) is obtained by definition of residuals in (10) and (9) and by following the same steps of the proof of Theorem 2. The last inverse triangle inequality is:

$$\|C_i\varphi_{i,k}\| \leq \bar{r}_i + \|C_i\tilde{\epsilon}_{i,af}^c + v_i\| \leq 2\bar{r}_i.$$

The thesis follows by negating the condition above. ■

Remark 11: Note that, in fact, Assumption 6 ensures that the summation of integrals in (33) contains the attack signal η_j only once. This is done only to avoid pathological cases where a particularly resourceful attacker designs multiple attacks such that their dynamic effect is mutually canceled in the dynamics (22). Such a strategy, however, requires a considerable amount of resources, *non local* model knowledge, and timing. We stress that the analysis of the observer errors under attack does not rely on such assumption, which is effectively used when deriving bounds on (33). ◁

C. Component-Wise Bounds

Using the same arguments of Subsection IV-A, it is possible to obtain component-wise bounds for the local residual vector that lead to less conservative detection thresholds than the one based on the norm. Unfortunately, considering entry-wise absolute values does not allow to obtain closed-form expression as those shown in the previous subsections. In the following, we show the component-wise counterparts of Propositions 4 and 5, and Corollary 1.

Lemma 2: In attack-free conditions, the UIO error is bounded component-by-component by

$$\begin{aligned}|\epsilon_i^d(t)| \leq \left|e^{F_i t}\right| \left(|\epsilon_i^d(0)| + |H_i|\bar{v}_i\right) + |H_i|\bar{v}_i + \\ + \int_0^t \left|e^{F_i(t-s)}\right| ds \left(|T_i|\bar{w}_i + |K_i|\bar{v}_i\right).\end{aligned} \tag{35}$$

□

Proposition 6: In attack-free conditions, the received residual is bounded as follows:

$$|\tilde{r}_i^c(t)| \leq \bar{r}_i'(t),$$

where

$$\begin{aligned}\bar{r}_i'(t) = &\left|C_i e^{F_i^c t}\right| |\tilde{\epsilon}_i^c(0)| \\ &+ \sum_{j\in\mathcal{N}_i} \int_0^t \left|C_i e^{F_i^c(t-s)}\right| |A_{ij}||\epsilon_j^d(s)| ds \\ &+ \int_0^t \left|C_i e^{F_i^c(t-s)}\right| ds(\bar{w}_i + |L_i|\bar{v}_i) + \bar{v}_i.\end{aligned} \tag{36}$$

□

Theorem 3 (Component-wise detectability): A covert attack starting at time T_a in $\mathcal{S}_k, k \in \mathcal{N}_i$, is detectable from $\mathcal{S}_i$ if for at least one component l of $\tilde{r}_i^c$, $\exists \bar{t}_i > T_a$ such that

$$\left| C_i \int_{T_a}^{\bar{t}_i} e^{F_i^c(t-\tau)} A_{ik} \int_{T_a}^{\tau} e^{F_k(\tau-s)} \alpha_k(s) ds d\tau \right|_{[l]} > 2\bar{r}_{i[l]}'(t). \tag{37}$$

□

Remark 12: Equations (34) and (37) provide an implicit characterization of the attack signals η_i that are detected *surely*. Conditions in Theorem 2 and 3 are only sufficient, i.e. they only provide a guaranteed detection threshold, but nothing can be said if such a threshold is not crossed. ◁

D. Settling Time

We complement the results on this section by briefly discussing the convergence properties of the obtained bounds. The scalar bounds introduced in Subsection IV-A may be quite conservative in the case when the state vector (or disturbances) are not normalized, i.e. they have components whose magnitudes are on different scales. On the other hand, all the exponentials presented in this section are related to transients in the state estimates, and not to transients in attack detection. Since it is always true that for any vector $x \in \mathbb{R}^n$

$$\|x\| \geq |x_{[i]}|, \forall i \in 1, \ldots, n,$$

we can argue that if the residual norm has converged within a certain tolerance level, then also each one of its components has. Therefore, if the computation of (35)–(37) is problematic, then the respective steady state values could be computed offline and a constant threshold could be employed. If that is the case, a lower bound $\bar{T}_i$ on the convergence time of the detection residuals is needed, in order to activate the detection logic only afterwards. It is possible to use (26)–(28) to obtain such lower bound for a given tolerance level $\bar{\delta}_i$.

Proposition 7: Given $\bar{\delta}_i$, $\forall t > \bar{T}_i$ the vanishing part $\bar{r}_i^z(t)$ of $\bar{r}_i(t)$ is no greater than $\bar{\delta}_i$, where

$$\bar{T}_i = \frac{\ln q_i - \ln \bar{\delta}_i}{|\mu(F_i^c)|},$$

with

$$\begin{aligned}q_i = &\|C_i\| \left(\|\tilde{\epsilon}_i^c(0)\| + \frac{(\bar{w}_i + \|L_i\|\bar{v}_i)}{\mu(F_i^c)} \right. \\ &+ \sum_{j\in\mathcal{N}_i} \left[\frac{\|A_{ij}\|}{\mu(F_i^c)} \left(\|H_j\|\bar{v}_j - \frac{\|T_j\|\bar{w}_j + \|K_j\|\bar{v}_j}{\mu(F_j)} \right) \right. \\ &\left. \left. + \left| \frac{h_j}{\mu(F_j) - \mu(F_i^c)} \right| \right] \right).\end{aligned} \tag{38}$$

□

Proof: Eq. (38) can be obtained by grouping the exponential parts of (26) and (27), whereas for (28), we wish to remove dependency from $e^{\mu(F_j)t}$. This can be done by considering

$$h_j \frac{e^{\mu(F_j)t} - e^{\mu(F_i^c)t}}{\mu(F_j) - \mu(F_i^c)} < \left| \frac{h_j}{|\mu(F_j) - \mu(F_i^c)|} \right| e^{\mu(F_i^c)t}.$$

Therefore, we can choose $\bar{\delta}_i$ and find a solution to:

$$\bar{r}_i^z(t) < q_i e^{\mu(F_i^c)t} \leq \bar{\delta}_i,$$

which is satisfied for

$$t \geq \frac{\ln q_i - \ln \bar{\delta}_i}{|\mu(F_i^c)|},$$

where we have explicitly considered the fact that the logarithmic norm is negative for Hurwitz matrices. ■

V. Simulation Results

In order to show the effectiveness of the proposed methodology, we address a covert attack scenario in the context of the Power Network System benchmark proposed in [42]. To emphasize the independence of our detector from the

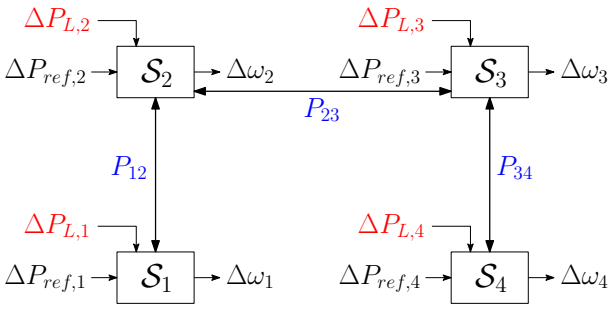


Fig. 2. Interconnection diagram of the considered benchmark [42].

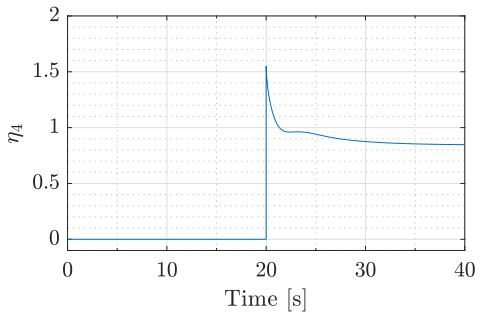


Fig. 3. Attack signal injected by the attacker in $\mathcal{S}_4$.

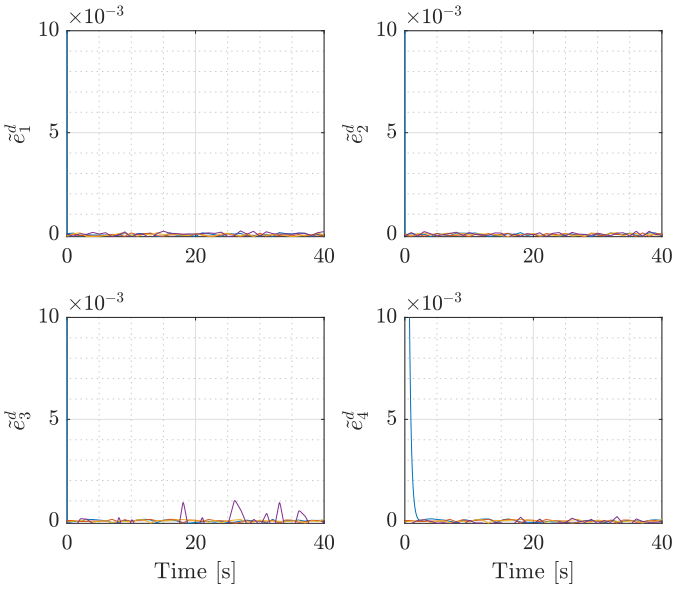


Fig. 4. *Received* errors of local (decentralized) estimators in each $\mathcal{S}_i$. There is no noticeable change in the trends before and after the attack.

controller design, we use a pre-designed distributed model-predictive controller from the PnPMPC toolbox [43], on top of which we implement the proposed detection architecture.

The scheme of the considered interconnected system is shown in Fig. 2, where each subsystem $\Sigma_{[i]}$ represents a different power generation area interconnected through a tie-line. Each distributed controller, accounting for desired input and state constrains, is in charge of the Automatic Control Generation layer in its respective zone, with the aim of keeping the subsystem around its nominal values. We refer the reader to [43] for further details on the system's model, choice of parameters, and control algorithm. However, for the reader's convenience, we recall that the power system is linearized around its operating point, and therefore all quantities should be regarded as deviations from a desired equilibrium. The state of each subsystem is defined as $x_i^\top = [\Delta\theta_i,\ \Delta\omega_i,\ \Delta P_{m_i},\ \Delta P_{v_i}]$, where $\Delta\theta_i$ and $\Delta\omega_i$ are deviations of rotor's angular displacement and speed, ΔP_{m_i} is the deviation from the nominal mechanical power, and ΔP_{v_i} represents the deviation of the steam valve position from its nominal value.

In the simulation scenario considered in this paper, we refer to a regulation task, rather than the original set-point tracking one, because in this way we can avoid discrepancies between our problem formulation and the benchmark one, which includes also exogenous load references. We do not lose generality in doing this, as we still employ the same controller to achieve a meaningful and realistic control objective. Furthermore, only for the control task, we adopt the assumption of fully accessible state, in order to guarantee its convergence analytical properties. For the diagnosers, instead, we consider a sensor channel with bounded disturbances.

The disturbances are random variables independently uniformly distributed in the following interval, for each component in each subsystem $\mathcal{S}_i$:

$$w_{i[k]}, v_{i[k]} \in [-10^{-4}, 10^{-4}], k \in [1, \dots, n_i]. \tag{39}$$

To design the detector, by solving the LMI presented in Theorem 1 a set of stabilizing matrices L_i is obtained. With regards to the threshold, we opt for the bounds briefly presented in Subsection IV-C mainly because the state variables of the subsystems differ by orders of magnitude, thus the norm bounds may not be particularly sensitive to deviations in some of the smaller components.

We simulate the system for a total time span of 40 s. At $T_{a4} = 20$ s, a malicious agent covertly attacks Subsystem $\mathcal{S}_4$ and tries to force a deviation on ΔP_{m_4}. The attack reference signal is designed such that this deviation amounts approximately to 0.6 p.u. (*per unit*):

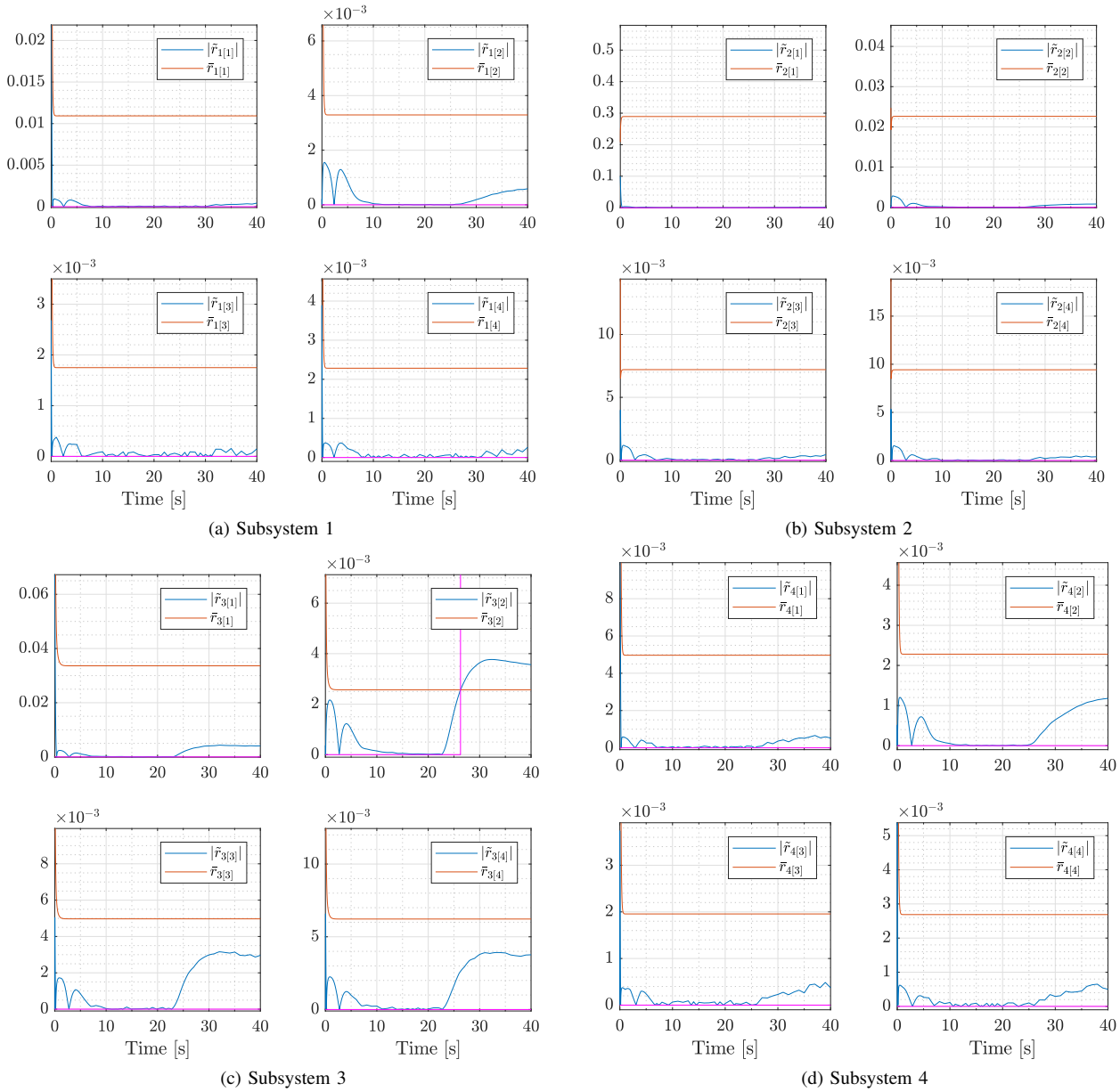
$$\rho_4(t) = 0.6\chi(t - T_{a4}).$$

We consider the case where the attacker's objective is to introduce some form of deviation from a desired state, rather than controlling it along a certain trajectory or set point.

The attacker implements a purely algebraic controller $\tilde{\mathcal{C}}_i$ with $K_{C4} = [0.9045,\ 6.1340,\ 0.2579,\ 0.1241]$. This results in the attack signal η_4 depicted in Figure 3.

First of all, we show that indeed the attack is covert for local estimators. In particular, in Fig. 4 we plot the errors *received* by each subsystem: as expected from (15), they do not show any visible trace of the attack, hence they cannot be used for the purpose of detection. This justifies the use of the architecture presented in the paper.

The results of the simulation are shown in Fig. 5. As presented in the previous sections, the residual signal $\tilde{r}_i^c$ ($\tilde{r}_i$ for brevity in the figures) is sensitive to attacks in the neighborhood of its corresponding subsystem. This is evident from Fig. 5c (the only neighbor of Subsystem 4, according to the layout in Fig. 2), where we see the threshold being



(a) Subsystem 1

(b) Subsystem 2

(c) Subsystem 3

(d) Subsystem 4

Fig. 5. Simulation results for the considered benchmark. In each plot, the 4 components of the subsystems states are drawn . The residual signal, the detection bounds, and the alarm signals a_i are marked in orange, blue, and magenta, respectively.

trespassed for the second component of the state at approximately $t = 26$ s, where the signal a_3 is triggered. According to the detection strategy summarized in Subsection II-B, the detector broadcasts this signal to $\mathcal{S}_2$ and $\mathcal{S}_4$. Since $\mathcal{S}_2$, receives $\{a_1 = 0, a_3 = 1\}$ the local decision of being under attack is not made. Conversely, $\mathcal{S}_4$ receives $\{a_3 = 1\}$ from its only neighbor, and therefore it decides to be under attack.

Since the considered system is weakly coupled ($\|A_{ij}\| \approx 10^{-1}$) and the state variables are in the *per unit* system, the influence of $\tilde{\epsilon}_4^d$ on $\tilde{r}_3^c$ is small (e.g. $\sim 4 \cdot 10^{-3}$ for the second state component in Fig. 5c). As a result, this specific system with the proposed detection architecture cannot tolerate high levels of disturbances in order to maintain acceptable effectiveness. We note, however, that the bounds (39) are within the same order as or larger than those used in other instances of the considered benchmark [44]. Furthermore, since the state components are in the *per unit* system, (39) have to be considered relative to the equipment's rated values and not in absolute terms.

VI. Concluding Remarks

In this paper, we propose a distributed method for the detection of covert attacks in interconnected large-scale LTI systems subject to bounded disturbances. We design a novel local detection scheme based on pairs of decentralized and distributed observers, in order to reveal local covert attacks. A rigorous analysis is provided dealing with estimation errors, detectability conditions, and detection-time upper bounds and extensive simulation results are given using a widely used Power Systems Benchmark.

Future research efforts will be devoted to considering cooperating attackers, the effects of imperfect model knowledge,

the generalization to the case of distributed nonlinear systems, and resilient control.

ACKNOWLEDGMENTS

We acknowledge Stefano Riverso and Giancarlo Ferrari-Trecate for the useful discussion on the benchmark used in the presented simulations.

APPENDIX

Proof of Proposition 2

Proof: To prove the proposition, note that the *actual* subsystem is driven by the control input $\tilde{u}_i$, whereas the observer estimates are computed using u_i and $\tilde{y}_i$. For the sake of notation simplicity, we omit in the following subscript i.

$$\begin{aligned}
\dot{\epsilon}^d &= \dot{x} - \dot{\hat{x}}^d = \dot{x} - \dot{z} - H\dot{\tilde{y}} = Ax + B(u+\eta) + \Xi\mathbf{x} + w \\
&\quad - [Fz + TBu + K(y-\gamma) + H(C\dot{x} + \dot{v} - C\dot{\tilde{x}})] \\
&= Ax + B(u+\eta) + \Xi\mathbf{x} + w - [Fz + TBu + K(y-\gamma) \\
&\quad + HC(Ax + B(u+\eta) + \Xi\mathbf{x} + w - A\tilde{x} - B\eta) \\
&\quad + H\dot{v}] \\
&= \bar{A}\epsilon^d + (I - HC - T)Bu + (I - HC)(\Xi\mathbf{x} + w) \\
&\quad + HCA\tilde{x} + B\eta - Fz + \bar{A}(z + H(y-\gamma)) \\
&\quad - K(y-\gamma) - H\dot{v} \\
&= (\bar{A} - K^{(1)}C)\epsilon^d + (I - HC - T)Bu + (I - HC)\Xi\mathbf{x} \\
&\quad + (I - HC)w + HCA\tilde{x} + B\eta + (\bar{A} - F)z \\
&\quad + (\bar{A}H - K)(y-\gamma) - H\dot{v} \\
&\quad + K^{(1)}[y - v - C(z + H(y-\gamma))] \\
&= (\bar{A} - K^{(1)}C)\epsilon^d + [I - HC - T]Bu + (I - HC)\Xi\mathbf{x} \\
&\quad + [(\bar{A} - K^{(1)}C) - F]z + (I - HC)w \\
&\quad + HCA\tilde{x} + B\eta - [(\bar{A} - K^{(1)}C)H - K]\gamma \\
&\quad + [(\bar{A} - K^{(1)}C)H - K^{(2)}]y - K^{(1)}v - H\dot{v},
\end{aligned}$$

where we defined $\bar{A} = A - HCA$. If conditions (12) hold, (14) is obtained. To prove (15), the same steps above can be repeated using (10). Finally, since the error dynamics under attack (15) is the same as the attack-free case (13), we conclude that the attack is covert. ■

REFERENCES

[1] R. M. Lee, M. J. Assante, and T. Conway, "Analysis of the cyber attack on the Ukrainian power grid," *SANS Industrial Control Systems*, pp. 1–23, 2016.

[2] J. Weiss, "Aurora generator test," *Handbook of SCADA/Control Systems Security*, p. 107, 2016.

[3] 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*, vol. 62, no. 1, pp. 18–33, 2017.

[4] 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*, vol. 64, no. 1, pp. 4–19, 2019.

[5] A. A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *Proceedings of the International Conference on Distributed Computing Systems*, Beijing, China, June 2008, pp. 495–500.

[6] Y. Mo and B. Sinopoli, "False Data Injection Attacks in Cyber Physical Systems," in *Proceedings of the First Workshop on Secure Control Systems*, 2010.

[7] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, October 2012, pp. 1806–1813.

[8] H. Sandberg, S. Amin, and K. H. Johansson, "Cyberphysical security in networked control systems: An introduction to the issue," *IEEE Control Systems*, vol. 35, no. 1, pp. 20–23, 2015.

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

[10] F. Pasqualetti, "Secure control systems: A control-theoretic approach to cyber-physical security," Ph.D. dissertation, University California, Santa Barbara, 2012.

[11] S. Weerakkody, "Detecting Integrity Attacks on Control Systems using Robust Physical Watermarking," in *Proceedings of the 53rd IEEE Conference on Decision and Control*, 2014, pp. 3757–3764.

[12] P. Cheng, L. Shi, and B. Sinopoli, "Guest editorial special issue on secure control of cyber-physical systems," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 1–3, 2017.

[13] S. Weerakkody, X. Liu, S. H. Son, and B. Sinopoli, "A graph-theoretic characterization of perfect attackability for secure design of distributed control systems," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 60–70, 2017.

[14] R. Anguluri, V. Katewa, and F. Pasqualetti, "A probabilistic approach to design switching attacks against interconnected systems," in *Proceedings of the American Control Conference*, July 2019, pp. 4430–4435.

[15] D. I. Urbina, J. A. Giraldo, A. A. Cardenas, N. O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, and H. Sandberg, "Limiting the Impact of Stealthy Attacks on Industrial Control Systems," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 1092–1105.

[16] S. M. Dibaji, M. Pirani, D. B. Flamholz, A. M. Annaswamy, K. H. Johansson, and A. Chakrabortty, "A systems and control perspective of cps security," *Annual Reviews in Control*, vol. 47, pp. 394–411, 2019.

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

[18] F. Arrichiello, A. Marino, and F. Pierri, "Observer-based decentralized fault detection and isolation strategy for networked multirobot systems," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 4, pp. 1465–1476, 2015.

[19] M. Blanke, M. Kinnaert, J. Lunze, and M. Staroswiecki, "Distributed fault diagnosis and fault-tolerant control," in *Diagnosis and Fault-Tolerant Control*. Springer, 2016, pp. 467–518.

[20] F. Boem, L. Sabattini, and C. Secchi, "Decentralized fault diagnosis for heterogeneous multi-agent systems," in *Control and Fault-Tolerant Systems (SysTol), 2016 3rd Conference on*. IEEE, 2016, pp. 771–776.

[21] M. Davoodi, N. Meskin, and K. Khorasani, "Simultaneous fault detection and consensus control design for a network of multi-agent systems," *Automatica*, vol. 66, pp. 185–194, 2016.

[22] R. M. Ferrari, T. Parisini, and M. M. Polycarpou, "Distributed fault detection and isolation of large-scale discrete-time nonlinear systems: An adaptive approximation approach," *IEEE Transactions on Automatic Control*, vol. 57, no. 2, pp. 275–290, 2012.

[23] S. Riverso, F. Boem, G. Ferrari-Trecate, and T. Parisini, "Plug-and-play fault detection and control-reconfiguration for a class of nonlinear large-scale constrained systems," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 3963–3978, 2016.

[24] I. Shames, A. M. Teixeira, H. Sandberg, and K. H. Johansson, "Distributed fault detection for interconnected second-order systems," *Automatica*, vol. 47, no. 12, pp. 2757–2764, 2011.

[25] S. Stanković, N. Ilić, Ž. Djurović, M. Stanković, and K. Johansson, "Consensus based overlapping decentralized fault detection and isolation," in *Control and Fault-Tolerant Systems (SysTol), 2010 Conference on*. IEEE, 2010, pp. 570–575.

[26] A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *Distributed Computing Systems Workshops, 2008. ICDCS'08. 28th International Conference on*. IEEE, 2008, pp. 495–500.

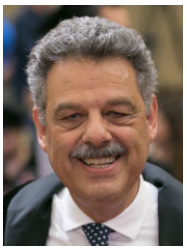
[27] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[28] ——, "A divide-and-conquer approach to distributed attack identification," in *Decision and Control (CDC), 2015 IEEE 54th Annual Conference on*. IEEE, 2015, pp. 5801–5807.
[29] A. Teixeira, H. Sandberg, and K. H. Johansson, "Networked control systems under cyber attacks with applications to power networks," in *American Control Conference (ACC), 2010*. IEEE, 2010, pp. 3690–3696.
[30] 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 *Proceedings of the European Control Conference*, Limassol, Cyprus, June 2018.
[31] R. Anguluri, V. Katewa, and F. Pasqualetti, "Attack detection in stochastic interconnected systems: Centralized vs decentralized detectors," in *Proceedings of the 57th IEEE Conference on Decision and Control*, December 2018, pp. 4541–4546.
[32] A. O. de Sá, L. F. R. d. C. Carmo, and R. C. S. Machado, "Covert attacks in cyber-physical control systems," *IEEE Transactions on Industrial Informatics*, vol. 13, pp. 1641–1651, August 2017.
[33] A. Hoehn and P. Zhang, "Detection of covert attacks and zero dynamics attacks in cyber-physical systems," in *Proceedings of the American Control Conference*, Boston, MA, USA, July 2016, pp. 302–307.
[34] M. I. Müller, J. Milošević, H. Sandberg, and C. R. Rojas, "A risk-theoretical approach to $\mathcal{H}_2$-optimal control under covert attacks," in *Proceedings of the 57th IEEE Conference on Decision and Control*, December 2018, pp. 4553–4558.
[35] A. Barboni, H. Rezaee, F. Boem, and T. Parisini, "Distributed detection of covert attacks for interconnected systems," in *Proceedings of the European Control Conference*, Naples, Italy, June 2019, pp. 2240–2245.
[36] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49–51, 2011.
[37] A. O. de Sa, L. F. R. da Costa Carmo, and R. C. S. Machado, "A controller design for mitigation of passive system identification attacks in networked control systems," *Journal of Internet Services and Applications*, vol. 9, no. 1, pp. 1–19, Feb. 2018.
[38] J. Chen, R. J. Patton, and H.-Y. Zhang, "Design of unknown input observers and robust fault detection filters," *International Journal of Control*, vol. 63, no. 1, 1996.
[39] A. J. van der Schaft, "$\mathcal{L}_2$ gain analysis of nonlinear systems and nonlinear state-feedback $\mathcal{H}_\infty$," *IEEE Transactions on Automatic Control*, vol. 37, no. 6, pp. 770–784, 1992.
[40] S. P. Boyd, *Linear Matrix Inequalities in System and Control Theory*. SIAM, 1994.
[41] T. Ström, "On logarithmic norms," *SIAM Journal on Numerical Analysis*, vol. 12, no. 5, pp. 741–753, 1975.
[42] S. Riverso, M. Farina, and G. Ferrari-Trecate, "Plug-and-play decentralized model predictive control for linear systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 10, pp. 2608–2614, 2013.
[43] S. Riverso, A. Battocchio, and G. Ferrari-Trecate, "PnPMPC: a toolbox for MatLab," 2012. [Online]. Available: http://sisdin.unipv.it/pnpmpc/pnpmpc.php
[44] S. Riverso, D. Rubini, and G. Ferrari-Trecate, "Distributed bounded-error state estimation based on practical robust positive invariance," *International Journal of Control*, vol. 88, no. 11, pp. 2277–2290, 2015.

Angelo Barboni (S'19) received the B.Sc. and the M.Sc. degrees (cum laude) in Electrical and Control Engineering in 2012 and 2015, respectively, both from the University of Trieste, Italy. He is currently a PhD candidate at Imperial College London, UK, with the Control and Power Research Group and the EPSRC – HiPEDS Centre for Doctoral Training (CDT). His current research interests span distributed estimation, fault diagnosis, cyber-physical systems, and security.

Hamed Rezaee (S'10, M'18) received the B.Sc., M.Sc., and Ph.D. degrees in control engineering from the Department of Electrical Engineering at Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2009, 2011, and 2016, respectively. He is currently a Research Associate in the Department of Electrical and Electronic Engineering at Imperial College London, London, UK, with a research focus on resilient control and monitoring in cyber-physical systems, multiagent systems, and consensus problems.

Francesca Boem received the M.Sc. degree (cum laude) in Management Engineering in 2009 and the Ph.D. degree in Information Engineering in 2013, both from the University of Trieste, Italy. She was Post-Doc at the University of Trieste with the Machine Learning Group from 2013 to 2014. From 2014 to 2018, she was Research Associate at the Department of Electrical and Electronic Engineering, Imperial College London, with the Control and Power Research Group. Since April 2018 Dr. Boem is a Lecturer in the Department of Electronic and Electrical Engineering at University College London (UCL). Since 2015 she has been part of the team at Imperial College which has been awarded the flagship EU H2020-WIDESPREAD-TEAMING project for the development of the EU KIOS Research and Innovation Centre of Excellence, a strategic partnership between University of Cyprus and Imperial College London. In February 2018 Dr. Boem has been awarded the Imperial College Research Fellowship. Her current research interests include distributed fault diagnosis and fault-tolerant control methods for large-scale networked systems and distributed estimation methods for sensor networks. Dr. Boem is member of the IFAC Technical Committee 6.4 ("Fault Detection, Supervision & Safety of Technical Processes - SAFEPROCES") and Associate Editor for the IEEE Systems Journal, the IEEE Control System Society Conference Editorial Board and for the EUCA Conference Editorial Board.



Thomas Parisini (F'11) received the Ph.D. degree in Electronic Engineering and Computer Science in 1993 from the University of Genoa. He was with Politecnico di Milano and since 2010 he holds the Chair of Industrial Control and is Director of Research at Imperial College London. He is a Deputy Director of the KIOS Research and Innovation Centre of Excellence, University of Cyprus. Since 2001 he is also Danieli Endowed Chair of Automation Engineering with University of Trieste. In 2009-2012 he was Deputy Rector of University of Trieste. In 2018 he received an *Honorary Doctorate* from University of Aalborg, Denmark. He authored or co-authored more than 320 research papers in archival journals, book chapters, and international conference proceedings. His research interests include neural-network approximations for optimal control problems, distributed methods for cyber-attack detection and cyber-secure control of large-scale systems, fault diagnosis for nonlinear and distributed systems, nonlinear model predictive control systems and nonlinear estimation. He is a co-recipient of the IFAC Best Application Paper Prize of the Journal of Process Control, Elsevier, for the three-year period 2011-2013 and of the 2004 Outstanding Paper Award of the IEEE Trans. on Neural Networks. He is also a recipient of the 2007 IEEE Distinguished Member Award. In 2016, he was awarded as Principal Investigator at Imperial of the H2020 European Union flagship Teaming Project KIOS Research and Innovation Centre of Excellence led by University of Cyprus. In 2012, he was awarded an ABB Research Grant dealing with energy-autonomous sensor networks for self-monitoring industrial environments. Thomas Parisini currently serves as 2020 President-Elect of the IEEE Control Systems Society and has served as Vice-President for Publications Activities. During 2009-2016 he was the Editor-in-Chief of the IEEE Trans. on Control Systems Technology. Since 2017, he is Editor for Control Applications of Automatica and since 2018 he is the Editor in Chief of the European Journal of Control. He is also the Chair of the IFAC Technical Committee on Fault Detection, Supervision & Safety of Technical Processes - SAFEPROCESS. He was the Chair of the IEEE Control Systems Society Conference Editorial Board and a Distinguished Lecturer of the IEEE Control Systems Society. He was an elected member of the Board of Governors of the IEEE Control Systems Society and of the European Control Association (EUCA) and a member of the board of evaluators of the 7th Framework ICT Research Program of the European Union. Thomas Parisini is currently serving as an Associate Editor of the Int. J. of Control and served as Associate Editor of the IEEE Trans. on Automatic Control, of the IEEE Trans. on Neural Networks, of Automatica, and of the Int. J. of Robust and Nonlinear Control. Among other activities, he was the Program Chair of the 2008 IEEE Conference on Decision and Control and General Co-Chair of the 2013 IEEE Conference on Decision and Control. Prof. Parisini is a Fellow of the IEEE and of the IFAC.