Distributed Detection of Covert Attacks for Interconnected Systems

Angelo Barboni, Hamed Rezaee, Francesca Boem, and Thomas Parisini

Abstract— **The problem of covert attacks detection in a network of interconnected subsystems is addressed in this paper. Existing approaches in the area of covert attacks detection have been devoted to centralized systems and are mainly based on mismatch between the attacker's model and the actual plant. Instead, in this paper, we consider a large-scale system where the attacker has full knowledge on the subsystems models. By using the information received from neighboring subsystems and by exploiting the mismatch between a distributed Luenberger observer and a decentralized unknown input observer, we propose a local detection strategy allowing each subsystem to detect anomalies in its neighborhood. The effectiveness of the proposed strategy is shown in a numerical example.**

Index Terms— **Attack detection, covert attack, cyber-physical systems, large-scale systems.**

I. Introduction

The context of cyber-physical systems (CPSs) defined as collaboration of cybernetic and physical components has emerged as an important field of research in many scientific communities in recent years. The increasing use of accessible and networked platforms and protocols such as the Internet Protocol Suite has increased the vulnerability of CPSs to external cyber-attacks. These attacks can deteriorate the performance of physical systems and ultimately lead to failures or unsafe behavior, as shown in practice by [1] or [2]. Although fault detection and isolation in dynamical system is a rather mature research area, the fact that attacks are carried by intelligent and active agents makes this task more challenging [3], and for this reason, their detection has become a relevant research topic in control and estimation theory in the last decade [4]–[7].

On the other hand, the design of cyber-attacks that have detrimental impact on the physical layer has also been an objective of research, where the stealthiness property – i.e. the capability of a malicious signal to appear as "noise" to a detection mechanism – has drawn great attention, and has allowed to pinpoint vulnerabilities of common algorithms used in the control engineering practice. One of these stealthy attacks is the *covert attack* [8]. The main feature of a covert agent is that it injects some undesired control actions in the actuation channels while canceling its effects in the measurements. The result is that, under the assumption of perfect model knowledge by the attacker, the true state of the attacked system can be driven to arbitrary potentially unsafe trajectories without any trace in the sensing equipment, which in turn provides measurements compatible with the normal behavior making the attack undetectable. A few studies have already been devoted to this scenario. For instance, in [9] an intelligent type of covert attack is presented using system identification tools. The problem of covert attack detection in CPSs was investigated also in [10], where a random modulation is introduced on the system actuation side to cause errors in the attacker's model.

However, research in the area of detection and isolation of covert attacks is still in its infancy, and many problems in this area are still open to investigation. This paper is devoted to the detection of covert attacks for interconnected system. The proposed strategy is based on two local observers for each individual subsystem using only locally available and neighboring information, respectively. By comparing the performance of these two observers, we show that it is possible to detect anomalies in a distributed fashion, whereas local strategies are not effective for this type of attacks.

In summary, compared with the existing results in the literature, the main contributions of this paper are the following:

- We design a distributed strategy that allows to detect covert attacks that would be undetectable by traditional model-based techniques.
- We consider the worst case scenario where an attacker has full knowledge on the subsystems' models, and we provide quantitative results about covertness of the attacks in the proposed detection architecture.

In this work, we show some early results which are limited to the noise-free case but are sufficient to characterize our architecture. Further research will be devoted to addressing factors such as disturbances and isolation issues.

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. $h(t)$ stands for a step function. $\mathrm{diag}(\cdot)$ describes a block diagonal matrix composed of a set of matrices.

The rest of the paper is organized as follows. In the next section, the problem statement is presented. The proposed detection strategy is presented in Section III. Simulation results are given in Section IV, and conclusions reside in Section V.

This work is supported by the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1) and the European Union's Horizon 2020 Research and Innovation Program under the grant agreement No. 739551 (KIOS CoE).

A. Barboni, H. Rezaee, and T. Parisini are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK (e-mail: {a.barboni16, h.rezaee, t.parisini}@imperial.ac.uk).

T. Parisini is also with the Department 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.

F. Boem is with the Department of Electronic and Electrical Engineering, University College London, London, UK (f.boem16@ucl.ac.uk).

II. Problem Statement

We consider a large-scale system composed of N interconnected linear time-invariant (LTI) subsystems $\mathcal{S}_i$, $i \in \{1, \dots, N\}$, each modeled by the equations

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

where $x_i \in \mathbb{R}^{n_i}$, $\tilde{u}_i \in \mathbb{R}^{m_i}$, and $y_i \in \mathbb{R}^{p_i}$ are the local states, inputs, and outputs, respectively. The index set of the neighbors of $\mathcal{S}_i$ is $\mathcal{N}_i$.

Definition 1: We say that subsystem j is a *neighbor* of subsystem i if any component of x_j affects the dynamics of subsystem i.

We say that a quantity is *local to subsystem* i if it is directly related to this subsystem or to one of its neighbors. With this notion, we say that the local dynamics' matrices A_i, B_i, C_i, and the interconnection matrices A_{ij}, $\forall j \in \mathcal{N}_i$ are known locally for each subsystem $\mathcal{S}_i$.

Assumption 1: The pairs (A_i, C_i) are observable for $i \in \{1, \dots, N\}$. $\triangleleft$

Each subsystem is equipped with a local unit LU_i comprising a pair of observers $\mathcal{O}_i^d$ and $\mathcal{O}_i^c$, and a local controller $\mathcal{C}_i$, which receive the local measurements $\tilde{y}_i$ and compute a control action u_i, respectively. The local observer $\mathcal{O}_i^d$ is allowed to communicate with neighboring observers $\mathcal{O}_j^d$, $j \in \mathcal{N}_i$ to exchange local estimates $\hat{x}_i^c$ that will be later formally defined. The overall architecture is shown in Fig. 1.

Remark 1: The proposed scheme does not assume any framework with regard to the control strategy, and in fact it just requires knowledge of the plant measurements and the controller signals, regardless of how they are obtained. Similarly, the controller $\mathcal{C}_i$ does not necessarily depend on the estimates of $\mathcal{O}_i^d$.

Assumption 2: We assume that the information exchanged between local units is not corrupted. $\triangleleft$

This is a realistic scenario when the communication between detectors is realized, for example, over a closed or encrypted network. This could be possible when the detection architecture is built on top of existing plant equipment and has dedicated communication.

The signal $\tilde{y}_i \in \mathbb{R}^{p_i}$ is the measurement signal as received by the observers, which can in general be different from y_i obtained at the sensors. These signals are exchanged over a vulnerable link that can be therefore tapped by an attacker (see Fig.1). For this reason, we denote u_i and y_i as the *legitimate* (*nominal*) or *transmitted* signals, and $\tilde{u}_i$ and $\tilde{y}_i$ as the *attacked* or *received* counterparts.

Let i now denote the index of the subsystem where an attacker $\mathcal{A}$ performs a man-in-the-middle (MITM) attack, injecting malicious signals γ and μ in the tapped link between the plant and the local unit. Then, for the link between the pair $(\mathcal{S}_i, LU_i)$ we have that

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

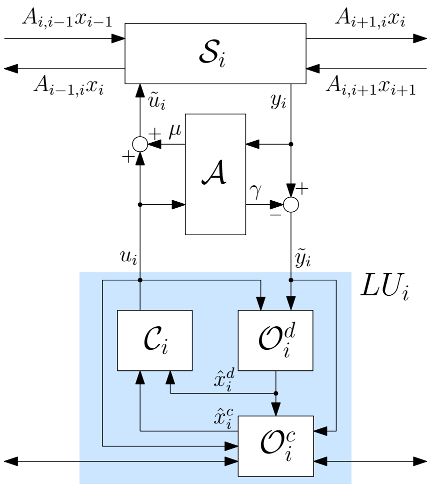

Fig. 1. Architecture of the attacked subsystem.

For the scope of this paper, we consider the following assumption.

Assumption 3: Only one subsystem is under attack in the overall system. $\triangleleft$

The malicious agent $\mathcal{A}$ is modeled as a dynamical system, comprising two parts: the model

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

which is used to replicate $\mathcal{S}_i$ (ignoring the interconnection with neighbors), and a malicious controller $\tilde{\mathcal{C}}_i$ which is in general unknown and implements the attacker's objective. Such controller is responsible for computing the attack action μ, which can depend on the attack objective ρ, the attacker's state $\tilde{x}_i$, time t, and potentially available data u_i and y_i.

$$\mu = f(t, \tilde{x}_i, \rho, u_i, y_i)$$

A. Covert attacks

Broadly speaking, we say that a malicious agent $\mathcal{A}$ is covert if is able to fully compensate in the output channel its action on the actuator side. This is true when the output signal after the attack is equal to the output signal in the attack-free case. Let $T_a > 0$ be the time instant in which the attack begins (therefore $\gamma(t) = \mu(t) = 0$ for $t < T_a$), and let us state the following assumption, which represents the worst case scenario for a defender.

Assumption 4: The attacker has *perfect knowledge* of the subsystem model, i.e. $(\tilde{A}_i, \tilde{B}_i, \tilde{C}_i) = (A_i, B_i, C_i)$. However, the attacker ignores the interconnection topology and the neighbors' effects. $\triangleleft$

Our approach to covert attacks in the distributed setting was inspired by [8], where they are formulated in the frequency domain for centralized systems. We provide the following definition of a covert agent, and prove afterwards that the scheme presented in this section is in fact covert.

Definition 2 (Covert agent): The malicious agent $\mathcal{A}$ is covert if the attacked measurement output $\tilde{y}_i$ cannot be distinguished from the legitimate system response.

Proposition 1: If Assumption 4 holds, then the attack scheme defined by (2) and (3) is covert asymptotically. Furthermore, if the attacker sets the initial conditions of the model (3) as $\tilde{x}_i(T_a) = 0$, then the attack is covert $\forall t$.

Proof: By (2), we clearly have that before the attack, for $0 < t < T_a$, the nominal and the attacked measurements are the same, and let $x_i(T_a)$ be the plant's state when the attack begins. The response of the attacked subsystem $\mathcal{S}_i$ can be written as:

$$\begin{aligned} y_i(t) = C_i e^{A_i(t-T_a)} x_i(T_a) + C_i \int_{T_a}^{t} e^{A_i(t-\tau)} \\ \left[B_i(u_i(\tau) + \mu(\tau)) + \sum_{j \in \mathcal{N}_i} A_{ij} x_j(\tau) \right] d\tau, \end{aligned} \tag{4}$$

and the compensation signal produced by the attacker is

$$\gamma(t) = \tilde{C}_i e^{\tilde{A}_i t} \tilde{x}_i(T_a) + \tilde{C}_i \int_{T_a}^{t} e^{\tilde{A}_i(t-\tau)} \tilde{B}_i \mu(t) d\tau. \tag{5}$$

After subtracting (5) from (4) and using Assumption 4, we obtain the attacked output response

$$\begin{aligned} \tilde{y}_i(t) = C_i e^{A_i(t-T_a)} (x_i(T_a) - \tilde{x}_i(T_a)) + \\ C_i \int_{T_a}^{t} e^{A_i(t-\tau)} \left[B_i u_i(t) + \sum_{j \in \mathcal{N}_i} A_{ij} x_j(\tau) \right] d\tau. \end{aligned} \tag{6}$$

Notice that (6) has in fact the same expression of the attack-free time response, with the exception of the initial conditions, and therefore the attacked and the legitimate measured outputs have the same asymptotic response. If the attacker sets the initial conditions of the model (3) as $\tilde{x}_i(T_a) = 0$, then (6) matches the legitimate output response exactly, and the time responses for $0 < t < T_a$ and $t \geq T_a$ can be composed with no discontinuity. ■

Remark 2: Even if the attacker model (3) does not include any neighboring states, the compensation signal in (5) is sufficient to make the attack covert. This implies that the attacker does not need any intelligence on the system connectivity to perform a covert attack.

III. Observer-based detection architecture

In this section, we propose an observer-based architecture for detecting covert attacks in a distributed scenario. The underlying idea is that even though an attacker can perfectly deceive the information flow between the plant and the local unit LU_i, the local state x_i affects the neighboring systems. We design two local observers such that we can distinguish attacks by analyzing their individual performance. The first observer depends on the coupling effects of each subsystem with its neighbors and uses information received from those neighbors; the second one is based on a decoupled design, thus is independent on the interconnection variables. We denote these two observers by $\mathcal{O}_i^c$ and $\mathcal{O}_i^d$ computing the estimates $\hat{x}_i^c$ and $\hat{x}_i^d$, respectively.

We define the following decoupled and coupled estimation errors, as well as the residuals related to them:

$$\begin{aligned} \epsilon_i^d &\doteq x_i - \hat{x}_i^d, \quad r_i^d = y_i - C_i \hat{x}_i^d = C_i \epsilon_i^d, \\ \epsilon_i^c &\doteq x_i - \hat{x}_i^c, \quad r_i^c = y_i - C_i \hat{x}_i^c = C_i \epsilon_i^c. \end{aligned} \tag{7}$$

The errors ϵ_i^d and ϵ_i^c represent the difference between the *actual* state of $\mathcal{S}_i$ and the corresponding estimates, and they are not available in general.

Furthermore, in the presence of an attacker, the estimates will depend on $\tilde{y}_i \neq y_i$, and we therefore define the *attacked* (*received*) estimation errors and residuals:

$$\begin{aligned} \tilde{\epsilon}_i^d &\doteq x_i - \tilde{x}_i - \hat{x}_i^d, \quad \tilde{r}_i^d = \tilde{y}_i - C_i \hat{x}_i^d = C_i \tilde{\epsilon}_i^d, \\ \tilde{\epsilon}_i^c &\doteq x_i - \tilde{x}_i - \hat{x}_i^c, \quad \tilde{r}_i^c = \tilde{y}_i - C_i \hat{x}_i^c = C_i \tilde{\epsilon}_i^c. \end{aligned} \tag{8}$$

Notice that only the attacked residuals in (8) are computable as only $\tilde{y}_i$ is available to the observers.

A. Decoupled Observation Strategy

In order to provide a decoupled estimation of the state, we treat the neighbors' interconnections as disturbances and design an unknown input observer (UIO) which guarantees a convergent estimator error ϵ_i^d which does not depend on the neighbors [11]. UIOs have been employed for fault detection in distributed scenarios (for example in [12]), and for detecting integrity attacks within the communication between agents [13].

The UIO has the form

$$\begin{cases} \dot{z}_i = F_i z_i + T_i B_i u_i + K_i \tilde{y}_i \\ \hat{x}_i^d = z_i + H_i \tilde{y}_i \,, \end{cases} \tag{9}$$

where $K_i = K_i^{(1)} + K_i^{(2)}$.

By rewriting the interconnection summation in matrix form, and by verifying that the conditions in [11] hold, we have that (9) is an UIO for the system (1), and the estimation error evolves according to the dynamics

$$\dot{\epsilon}_i^d = F_i \epsilon_i^d \,, \tag{10}$$

where F is chosen to be Hurwitz by appropriately choosing $K_i^{(1)}$, and therefore $\epsilon_i^d \to 0$, i.e. $\hat{x}_i^d \to x_i$. It is important to note that (10) holds for the attack-free system, when the actuation and measurement channels are not corrupted, that is to say $\epsilon_i^d = \tilde{\epsilon}_i^d$.

Proposition 2: When the ith subsystem is under attack, if the UIO conditions and Assumption 4 are satisfied, the following equations hold. The error dynamic of the observer (9) is

$$\dot{\epsilon}_i^d = F_i \epsilon_i^d + H_i C_i A_i \tilde{x}_i + B_i \mu + K_i^{(1)} \gamma, \tag{11}$$

while the attacked estimation error defined in (8) is given by

$$\dot{\tilde{\epsilon}}_i^d = F_i \tilde{\epsilon}_i^d \,. \tag{12}$$

Then, since the error dynamics (12) is the same as in the attack-free case, the attack model (3) *is covert for a UIO.*

Proof: See Appendix. ■

Remark 3: Since $\tilde{\epsilon}_i^d = \epsilon_i^d - \tilde{x}_i \to 0$, the actual error converges to the attacker's model state, and the state estimate converges to the difference $x_i - \tilde{x}_i$.

B. Observer Design Based on Subsystems Coupling Information

In this subsection, we develop an observer-based strategy where the observers $\mathcal{O}_i^c$ in each subsystem provide an estimate of its local state vector by taking into account the information received from its neighboring subsystems. Thus, we propose the following Luenberger-like distributed state observer for the ith subsystem:

$$\dot{\hat{x}}_i^c = A_i\hat{x}_i^c + B_iu_i + \sum_{j\in\mathcal{N}_i} A_{ij}\hat{x}_j^d + L_i(y_i - \hat{y}_i^c). \tag{13}$$

It should be noted that with this design, (13) depends on the dynamic estimates from all the neighboring decoupled observers $\mathcal{O}_j^d$. Thus, to design $L_i \in \mathbb{R}^{n_i\times p_i}$ in the distributed continuous-time observer (13) and ensure stability, the following proposition is stated.

Proposition 3: Consider the large-scale system described in (1). The observer (13) guarantees convergence to zero of the estimation errors $\epsilon_i^c = x_i - \hat{x}_i^c, i \in \{1, 2, \ldots, N\}$, if F_i in (9) and $A_i - L_iC_i$ are Hurwitz.

Proof: By considering $\hat{y}_i^c = C_i\hat{x}_i^c$, (13) can be restated as follows:

$$\dot{\hat{x}}_i^c = (A_i - L_iC_i)\hat{x}_i^c + B_iu_i + \sum_{j\in\mathcal{N}_i} A_{ij}\hat{x}_j^d + L_iy_i. \tag{14}$$

Now, using (14) and (1), the evolution of the estimation error ϵ_i^c can be stated as follows:

$$\begin{aligned}\dot{\epsilon}_i^c = A_ix_i + B_iu_i + \sum_{j\in\mathcal{N}_i} A_{ij}x_j - (A_i - L_iC_i)\hat{x}_i^c\\ -B_iu_i - \sum_{j\in\mathcal{N}_i} A_{ij}\hat{x}_j^d - L_iy_i.\end{aligned} \tag{15}$$

Since $y_i = C_ix_i$, after some manipulation, from (15) it follows that

$$\dot{\epsilon}_i^c = (A_i - L_iC_i)\epsilon_i^c + \sum_{j\in\mathcal{N}_i} A_{ij}\epsilon_j^d. \tag{16}$$

If we define

$$\begin{aligned}\epsilon &= \begin{bmatrix}\epsilon_1^{c\top} & \epsilon_2^{c\top} & \ldots & \epsilon_N^{c\top} & \epsilon_1^{d\top} & \epsilon_2^{d\top} & \ldots & \epsilon_N^{d\top}\end{bmatrix}^\top\\ \acute{L} &= \mathrm{diag}(L_1C_1, L_2C_2, \ldots, L_NC_N, 0, 0, \ldots, 0)\\ F &= \mathrm{diag}(0, 0, \ldots, 0, F_1, F_2, \ldots, F_N)\\ D &= \mathrm{diag}(A_1, A_2, \ldots, A_N),\end{aligned}$$

by considering (10), for all the interconnected subsystems, we can say that

$$\dot{\epsilon} = U\epsilon - \acute{L}\epsilon + F\epsilon,$$

where

$$U = \begin{bmatrix} D & A - D\\ 0 & 0\end{bmatrix},$$

and where $A = [A_{ij}], A_{ii} = A_i$. Since $U - \acute{L} + F$ is a triangular block matrix with Hurwitz diagonal entries, we have that $U - \acute{L} + F$ is Hurwitz. Thus, $\lim_{t\to\infty}\epsilon = 0$, and the proof is completed. ∎

In the following Proposition, we analyze the effects that a covert attack has on such observer, and what can be said about the error dynamics.

Proposition 4: When the ith subsystem is under attack, if Assumption 4 is satisfied, the following equations hold. The error dynamics of the observer (13) is

$$\dot{\epsilon}_i^c = (A_i - L_iC_i)\epsilon_i^c + B_i\mu + L_i\gamma + \sum_{j\in\mathcal{N}_i} A_{ij}\epsilon_j^d, \tag{17}$$

where ϵ_j^d is given by (10), (11). Conversely, the attacked estimation error is given by

$$\dot{\tilde{\epsilon}}_i^c = (A_i - L_iC_i)\tilde{\epsilon}_i^c + \sum_{j\in\mathcal{N}_i} A_{ij}\epsilon_j^d, \tag{18}$$

Then, since the error dynamics (18) is the same in as the attack-free case, the attack model (3) is covert for the observer (13).

Proof: The error equation (15) can be rewritten as

$$\begin{aligned}\dot{\epsilon}_i^c = A_ix_i + B_i\tilde{u}_i + \sum_{j\in\mathcal{N}_i} A_{ij}x_j - (A_i - L_iC_i)\hat{x}_i^c\\ -B_iu_i - \sum_{j\in\mathcal{N}_i} A_{ij}\hat{x}_j^d - L_i\tilde{y}_i,\end{aligned}$$

which under attack injection (2), leads to (17). From the definition of $\tilde{\epsilon}_i^c$ in (8), by similar algebraic operations, one obtains (18). ∎

C. Attack detection scheme and detectability analysis

The detection method is based on the discrepancy between the decoupled and coupled state estimates $\|\hat{x}_i^d - \hat{x}_i^c\|$. The decoupled estimate $\hat{x}_i^d$ does not depend on the neighboring dynamics, but only on local input and output measurements, while being affected *only* locally by a possible covert agent. On the other hand, the coupled estimate $\hat{x}_i^c$ depends on the state estimates $\hat{x}_j^d$ communicated by the neighbors, and as evident from (17) and (18) the estimation error depends on the *true* error ϵ_j^d, regardless of the attack presence.

Before providing the main result of this subsection, we provide the sketch of a possible detection logic. Our detection strategy relies on the fact that if a subsystem is under attack, all its neighbors should provide an indication of such event. The design of thresholds to be triggered in this case is out of the scope of this paper, however suppose that such local thresholds exist, then each Subsystem j can broadcast to its neighbors an alarm signal $a_j \in \{1, 0\}$. Therefore, if Subsystem i is under attack it will receive a set of signals $\{a_j = 1, j \in \mathcal{N}_i\}$, which Subsystem i can locally use to decide it is under attack, under the given assumptions.

We now state the following result which motivates the choice of the the difference $\|\hat{x}_i^d - \hat{x}_i^c\|$ as a sensitive indicator for detecting this type of attack.

Proposition 5: Consider the architecture presented in Section II, let subsystem i be under a covert attack, and let the pair $(A_j - L_jC_j, A_{ji})$ be reachable $\forall j \in \mathcal{N}_i$. By Assumption 3 only subsystem i can be under attack. The covert attack in subsystem i can be detected if and only if

$$\|\hat{x}_j^c - \hat{x}_j^d\| \nrightarrow 0,\ \forall j \in \mathcal{N}_i. \tag{19}$$

Proof: *(Sufficiency)* We proceed by contradiction. Let (19) hold and the subsystem i be attack-free. If there is no attack in $\mathcal{S}_i$, the observers (9) and (13) both converge to 0. Then, we can find an arbitrary small $\varepsilon > 0$ such that for all t greater than some finite $\bar{\tau}$,

$$\varepsilon > \|\epsilon_j^d(t)\| + \|\epsilon_j^c(t)\| \geq \|\epsilon_j^d(t) - \epsilon_j^c(t)\|, \; \forall j \in \mathcal{N}_i$$

and finally by adding and subtracting $x_j(t)$, we obtain that $\|\hat{x}_j^c(t) - \hat{x}_j^d(t)\| < \varepsilon$, which contradicts (19), thus we conclude that if (19) holds, then subsystem i is under attack.

(Necessity) We use contradiction again, and assume that subsystem i is under attack but (19) is not verified, thus $\exists j \in \mathcal{N}_i : \|\hat{x}_j^c - \hat{x}_j^d\| \to 0$, namely

$$\forall \varepsilon > 0, \exists \bar{\tau} > 0 : \|\hat{x}_j^d(t) - \hat{x}_j^c(t)\| < \varepsilon, \; \forall t > \bar{\tau}.$$

We have that

$$\varepsilon > \|\hat{x}_j^d(t) - \hat{x}_j^c(t)\| = \|\tilde{\epsilon}_j^c(t) - \tilde{\epsilon}_j^d(t)\| \geq \big|\|\tilde{\epsilon}_j^c(t)\| - \|\tilde{\epsilon}_j^d(t)\|\big|.$$

By Proposition 2 we know that $\|\tilde{\epsilon}_j^d\| \to 0$ by design if subsystem i is under attack, then we can choose an arbitrarily small $\varepsilon' > 0$ such that for $t > \bar{\tau}$

$$\|\tilde{\epsilon}_j^c(t)\| < \|\tilde{\epsilon}_j^d(t)\| + \varepsilon < \varepsilon' + \varepsilon.$$

Equation (18) can be rewritten as

$$\dot{\tilde{\epsilon}}_j^c = (A_j - L_j C_j)\tilde{\epsilon}_j^c + \sum_{k \in \mathcal{N}_j \setminus \{i\}} A_{jk}\epsilon_k^d + A_{ji}\epsilon_i^d.$$

The terms in the summation are all converging to 0 because they are states of stable autonomous linear systems and they are not under attack. If the pair $(A_j - L_j C_j, A_{ji})$ is reachable, there is no non-zero input that leaves the system at 0. Also, we have shown in Remark 3 that $\epsilon_i^d - \tilde{x}_i \to 0$, where the attacker's state $\tilde{x}_i$ represents therefore the attack-induced error. Since the attacker can arbitrarily steer its own state and aims at deteriorating the actual system performance, namely

$$\|\tilde{x}_i(t)\| > \delta > 0, \; t \in [T_a, T_a + \Delta],$$

with $\Delta > 0$, there will be a time $t > T_a > \bar{\tau}$ when for any arbitrary choice of ε' and ε such that $\|\tilde{\epsilon}_j^c(t)\| = \varepsilon' + \varepsilon$, thus, we reach contradiction. This proves that if the system is under attack, then condition (19) holds. ■

The convergence result in Proposition 5 and integration of the error equations (12) and (18) can be used to design local thresholds, however we leave this task to our future work.

IV. Simulation Results

In this section, we use a simple academic example to show and validate the performance of the architecture presented in the paper. Without loss of generality, we consider a network of 5 identical subsystems each one described by

$$A_i = \begin{bmatrix} -2 & -1 \\ 1 & 0 \end{bmatrix}, \qquad A_{ij} = \begin{bmatrix} 0.1 & 0.3 \\ 0 & 0 \end{bmatrix},$$
$$B_i = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top, \qquad C_i = \begin{bmatrix} 1 & 1 \end{bmatrix},$$

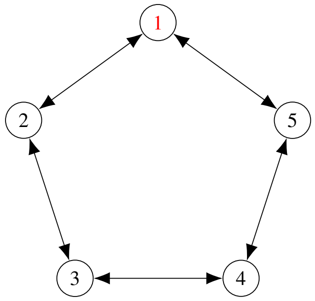

Fig. 2. Coupling topology of the subsystems.

where the coupling among the subsystems is considered as Fig. 2. We have chosen $K_i^{(1)} = L_i^\top = \begin{bmatrix} 0.5317 & 2.5317 \end{bmatrix}$ that satisfy Proposition 3 by means of pole placement.

Let subsystem 1 be under a covert attack, as specified in Proposition 1. Each controller tries to achieve a reference value of $y_i = 1$, by employing the control input $u_i = -[4 \;\; 7]\big(\hat{x}_i^d - [0 \;\; 1]^\top h(t-10)\big) - \sum_{j \in \mathcal{N}_i} A_{ij}\hat{x}_j^d$. At $t = 20$, the attacker tries to divert the output of subsystem 1 to a different target by means of the covert attack strategy explained in Section II.

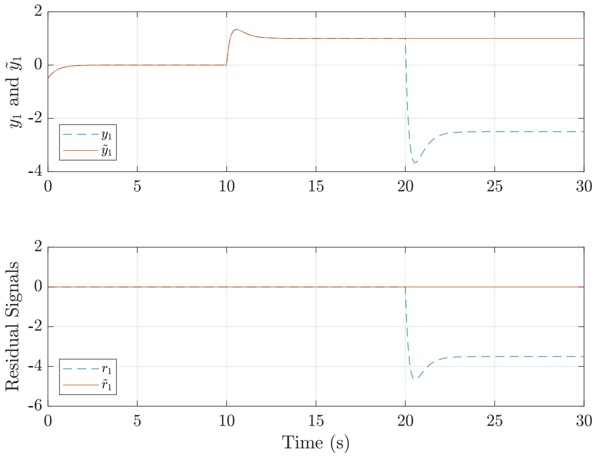

Fig. 3. Effect of the attacker in changing the real output (dashed line) of subsystem 1 while remaining hidden in the measurement output (solid line).

Under these conditions, as shown in Fig. 3-a, the received measurement output goes to reference while the real one does not. Also notice how the reference change does not cause any change in the residual signals and in the estimation mismatch $\|\hat{x}_i^d - \hat{x}_i^c\|$ in Fig. 4: the transient phase depends only on the initial conditions of the estimators. Moreover, as depicted in Fig. 3-b, the analysis of local residuals $\tilde{r}_i$ does not allow to detect the covert agent, which is stealthy at all times as shown in Proposition 1.

In order to clarify the logic described in the previous section, let us comment on this specific example, with reference to Fig. 4. We note that estimate differences in Subsystems 2 and 5 do not converge to 0, therefore they trigger local alarms $a_2 = 1$ and $a_5 = 1$. These alarms are sent to Subsystems 1 and 3, and to 1 and 5, respectively. At this point, Subsystem 1 receives $a_j = 1, \forall j \in \mathcal{N}_1$, and can therefore detect its own attacked condition.

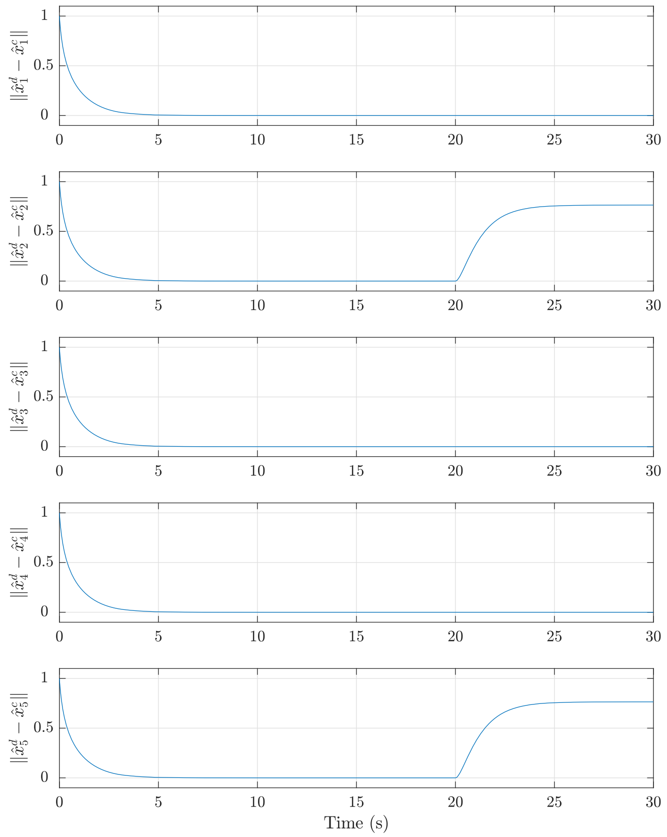

Fig. 4. Attack detection signals $\|\hat{x}_i^d - \hat{x}_i^c\|, i \in \{1, 2, \ldots, N\}$.

V. Conclusion and Future Work

Covert attacks detection in large-scale LTI systems has been addressed in this paper. We designed an architecture to detect covert attacks by exploiting the physical interconnections and designing a set of distributed observers either coupled or decoupled with neighboring subsystems. Accordingly, conditions for their detection have been derived based on convergence analysis of properly designed residuals. This study was an initial step towards distributed detection of covert attacks in interconnected systems. Our future work will be devoted to detection in the presence of disturbances, as well as threshold design and isolation strategies.

APPENDIX

Proof of Proposition 2

Let us drop subscript i for clarity, then by applying (2) to the error definition (7), we can proceed with the following manipulations.

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

where we defined $\bar{A} = A - HCA$ and $\Xi\mathbf{x}$ as the interconnection terms. If the UIO conditions [11] are verified, then we obtain (11). To obtain (12), the same computations can be done starting from (8), and we obtain that $\dot{\tilde{\epsilon}} = F\tilde{\epsilon}$, which is the same error dynamic equation as the attack-free case, thus proving that the attack strategy 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] 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), pp. 495–500, June 2008.

[4] 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*, pp. 1092–1105, 2016.

[5] 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.

[6] F. Pasqualetti, F. Dorfler, F. Bullo, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, pp. 2715–2729, November 2013.

[7] S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," in *Proceedings of the 54th IEEE Conference on Decision and Control*, (Osaka, Japan), pp. 5820–5826, December 2015.

[8] R. S. Smith, "Covert misappropriation of networked control systems: Presenting a feedback structure," *IEEE Control Systems*, vol. 35, pp. 82–92, February 2015.

[9] 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.

[10] 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), pp. 302–307, July 2016.

[11] 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.

[12] I. Shames, A. M. H. Teixeira, H. Sandberg, and K. H. Johansson, "Distributed fault detection for interconnected second-order systems," *Automatica*, vol. 47, pp. 2757–2764, December 2011.

[13] 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.