A Distributed Approach for the Detection of Covert Attacks in Interconnected Systems with Stochastic Uncertainties

Angelo Barboni, Alexander J. Gallo, Francesca Boem, Thomas Parisini

Abstract— **The design of a distributed architecture for the detection of covert attacks in interconnected Cyber-Physical Systems is addressed in this paper, in the presence of stochastic uncertainties. By exploiting communication between neighbors, the proposed scheme allows for the detection of covert attacks that are locally stealthy. The proposed methodology adopts a decentralized filter, jointly estimating the local state and the aggregate effect of the physical interconnections, and uses the communicated estimates to obtain an attack-sensitive residual. We derive some theoretical detection properties for the proposed architecture, and present numerical simulations.**

I. Introduction

Cyber-Physical Systems (CPSs) describe a class of large-scale systems, where the physical components are integrated with cyber resources, such as communication, control, and monitoring infrastructures. They are an ever more common class of systems, following the increased penetration of information technology (IT) for monitoring and coordination purposes in industrial plants and infrastructure systems.

Among the systems that can be described as CPSs, many are safety critical, as their inadequate provision of service may have severe consequences. This has led to a growing interest in the literature on the subject of *secure control*, as demonstrated by the recent special issue [1], as well as the surveys [2], [3], and the works cited therein.

Several works in the literature rely on centralized architectures for monitoring CPS [4], [5], [6]. Implementation of these architectures present some disadvantages, however, as it may require excessive computational power and communication resources. Hence, *distributed* methods for attack detection have been developed, of which [7], [8], [9], [10], [11] are notable examples. These often draw upon the work done in the Fault Detection and Isolation (FDI) literature (see for instance [12], [13], [14]).

It has been been shown in [4], [10], [15] that malicious agents can covertly misappropriate control systems by carefully designing the signals they inject in the available communication channels. In [10], the authors leveraged the physical interconnections between subsystems to the defender's advantage. Specifically, an architecture based on the combination of two observers permits to reveal misbehavior in neighboring subsystems that is instead concealed in the attacked one. In this paper, the problem is formulated in a similar way, however, here we consider:

- A discrete-time linear stochastic model for each subsystem instead of a continuous-time one.
- The proposed distributed detection architecture is based on different estimation models: a minimum-variance unbiased estimator jointly estimates the local states and the aggregate effect of the neighbors' interconnection.
- A detection method based on the statistical analysis of a properly designed residual signal is proposed, and its detectability properties are studied.

The distributed detection of attacks in stochastic systems is also considered in [8]. However, the authors do not focus on covert attacks, and do not build a distributed estimation architecture to achieve this, but rather perform hypothesis testing on appropriately processed output measurements.

The problem of unknown-input decoupling in the estimation of stochastic systems has drawn great attention in the past, and milestone contributions in the area include [16]. A more general problem is solved in [17], where unknown inputs also affect the measurement channel, while [18] improves on previous results by designing a two-step filter that also optimally estimates the input.

In this work, we adopt the filter presented in [18] to compute a distributed estimate of the local state, by decoupling it from the effect of the subsystem's neighbors.

The rest of the paper is structured as follows: in Section II, we formulate the considered problem. In Section III, the decoupled distributed filter is presented, and the properties of the state and unknown-input estimates are analyzed in Sections IV and V. Following this, a novel detection strategy is proposed in Section VI, where a suitable statistical test is defined, and some of its properties are provided. Finally, some numerical simulations are presented in Section VII.

Notation: For a vector $v \in \mathbb{R}^n$, $v_{[i]}$ denotes its i-th component. The identity matrix of dimension n will be defined as $\mathbf{I}_n$, and $\mathbf{0}_{n\times m} \in \mathbb{R}^{n\times m}$ is used to define a matrix of all zeros; when clear from the context, the indices will be omitted. We use notation $\mathrm{col}_{j\in\mathcal{J}}[x_j]$ and $\mathrm{row}_{j\in\mathcal{J}}[x_j]$ for the column or row concatenation of vectors x_j, j belonging to a set of indices $\mathcal{J} \subset \mathbb{N}$. The same notation is

This work has been partially supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 739551 (KIOS CoE). This work has also been conducted thanks to the support of the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1).

A. Barboni and A. J. Gallo is with the Department of Electrical and Electronic Engineering at the Imperial College London, UK. Email: {angelo.barboni16,alexander.gallo12}@ic.ac.uk

F. Boem is with the Department of Electronic and Electrical Engineering, University College London, UK. Email: f.boem@ucl.ac.uk

T. Parisini is with the Department of Electrical and Electronic Engineering at the Imperial College London, UK, the KIOS Research and Innovation Centre of Excellence, University of Cyprus, and the Department of Engineering and Architecture at University of Trieste, Italy. Email: t.parisini@gmail.com

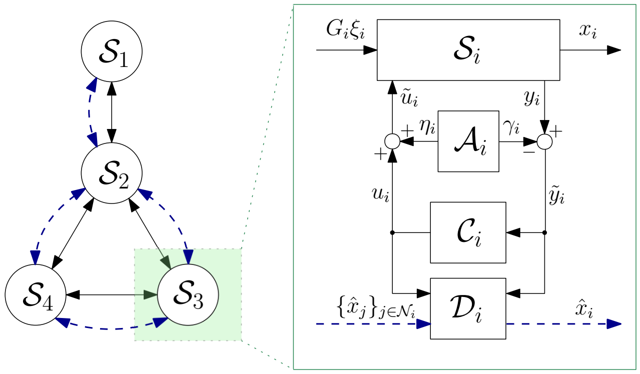

Fig. 1: The network of interconnected subsystems and the internal structure of subsystem $\mathcal{S}_i$ (under attack), equipped with a controller $\mathcal{C}_i$ and a diagnoser $\mathcal{D}_i$.

also used with matrices. Additionally, $\text{diag}_{j\in\mathcal{J}}[M_j]$ denotes block diagonal concatenation of matrices M_j, $j \in \mathcal{J}$. Given a random variable $x(k)$, $\mathbf{E}\left[x(k)\right]$ denotes its expected value. Furthermore, $\text{Cov}(x, y)$ denotes the covariance matrix of two random variables x and y, and $\text{Var}(x) = \text{Cov}(x, x)$ is the covariance matrix of a random variable x.

II. Problem Formulation

A. Model of CPS

Consider a CPS composed of N subsystems $\mathcal{S}_i$, which are interconnected through both physical and communication links. We consider the topology of the graphs defined by these links to be the same, i.e. a communication link between two subsystems is present if there is also a physical link. We define the set of neighbors of $\mathcal{S}_i$ as $\mathcal{N}_i \doteq \{j \in \{1 \dots N\} | \frac{\partial x_i}{\partial x_j} \neq 0\}$. The dynamics of $\mathcal{S}_i$ is:

$$
\begin{aligned}
x_i(k+1) &= A_i x_i(k) + B_i \tilde{u}_i(k) + \sum_{j\in\mathcal{N}_i} A_{ij} x_j(k) + w_i(k) \\
y_i(k) &= C_i x_i(k) + v_i(k)\,,
\end{aligned}
\tag{1}
$$

where $x_i \in \mathbb{R}^{n_i}$ is the state, $\tilde{u}_i \in \mathbb{R}^{m_i}$ is the control input, $y_i \in \mathbb{R}^{p_i}$ is the output measurement; $w_i \sim \mathcal{N}(0, W_i)$ and $v_i \sim \mathcal{N}(0, V_i)$ are process and measurement i.i.d. Gaussian noise, with known variance matrices $W_i \geq \mathbf{0}$ and $V_i > \mathbf{0}$. Furthermore, we assume the initial condition $x_i(0) \sim \mathcal{N}(\bar{x}_i^0, \Pi_i^0)$, with $\Pi_i^0 > 0$ and $\bar{x}_i^0$ known, and independent from $w_i(k)$ and $v_i(k)$ for all k. Matrices A_i, A_{ij}, B_i, and C_i are supposed to be known by the local diagnoser. As shown in the schematic diagram in Figure 1, each subsystem is locally equipped with a controller $\mathcal{C}_i$ and a diagnoser $\mathcal{D}_i$, the latter of which exchanges information with their neighbors.

B. Attack model

From time $k = k_a$, we consider a covert attack on subsystem $\mathcal{S}_i$. The attack is modeled as (see [4]):

$$
\begin{aligned}
x_i^a(k+1) &= A_i^a x_i^a(k) + B_i^a \eta_i(k) \\
\gamma_i(k) &= C_i^a x_i^a(k)
\end{aligned}\,,
\tag{2}
$$

where η_i is the attacker's control input. We stress that η_i is unknown to $\mathcal{D}_i$; furthermore, it is arbitrarily defined by the attacker to steer x_i away from its desired nominal trajectory. The signals η_i and γ_i are injected in the control and measurement channels, respectively, as follows:

$$
\begin{aligned}
\tilde{y}_i &\doteq y_i - \gamma_i \\
\tilde{u}_i &\doteq u_i + \eta_i
\end{aligned}\,,
\tag{3}
$$

where u_i is the input as computed by the controller $\mathcal{C}_i$. By mimicking the dynamics of $\mathcal{S}_i$ in (2), the attacker can compensate the effects of η_i through γ_i, thus making their action undetectable from the sole measurements observation.

Assumption 1: The attacker has *perfect knowledge* of the subsystem model, i.e. $(A_i^a, B_i^a, C_i^a) = (A_i, B_i, C_i)$. $\triangleleft$

The result of this attack is that the dynamics x_i^a is superimposed to that of $\mathcal{S}_i$. In the following result – a discretized version of Proposition 1 in [10] – the state of $\mathcal{S}_i$ is decomposed in a *healthy* and an *attacked* component.

Proposition 1: Consider attack strategy (2), and let Assumption 1 hold. If the attacker's initial state is $x_i^a(k_a) = \mathbf{0}$, the output received by the diagnoser unit $\mathcal{D}_i$ is:

$$
\tilde{y}_i(k) = y_i^h(k) = C_i x_i^h(k) + v_i(k)\,, \quad \forall k \geq k_a.
$$

where y_i^h is the output of the subsystem as if it were not affected by the attack.

Proof: Throughout this work, for reasons of space, proofs will be omitted. ■

Remark 1: Proposition 1 provides a sufficient condition for *stealthiness* of the covert attack, and implies that, whatever the local estimate of the state, the residual error based on local measurements is not affected by the attack, as has been shown in the literature [4], [10]. $\triangleleft$

In this paper, we address the following:

Problem 1: Given a subsystem $\mathcal{S}_i$ with dynamics as in (1) and an attack as in (2) from time k_a, and let Assumption 1 hold. Design a diagnoser $\mathcal{D}_i$ to detect the attack. $\triangleleft$

III. Distributed Detection Filters

Each local diagnosis unit $\mathcal{D}_i$ is equipped with a decentralized estimator, based on the filter proposed in [18] in the context of centralized estimation. Each diagnosis unit $\mathcal{D}_i$ then exchanges information with $\mathcal{D}_j, j \in \mathcal{N}_i$, in order to compute the detection residual introduced in Section V. The state estimator is unbiased and guarantees minimum variance of the estimation error regardless of the presence of an *unknown input* [18].

We design the diagnostic unit $\mathcal{D}_i$ such that the computed local estimates are decoupled from the neighbors of $\mathcal{S}_i$. To achieve this, the interconnection terms can be treated as unknown inputs, and rewritten as:

$$
\begin{aligned}
\sum_{j\in\mathcal{N}_i} A_{ij} x_j &= \underset{j\in\mathcal{N}_i}{\text{row}} \left[A_{ij}\right] \underset{j\in\mathcal{N}_i}{\text{col}} \left[x_i\right] \\
&= E_i \zeta_i = G_i \bar{E}_i \zeta_i = G_i \xi_i
\end{aligned}
\tag{4}
$$

where G_i has full column rank, $\bar{E}_i$ is a matrix of weights and defines a vector of unknown inputs $\xi_i \in \mathbb{R}^{g_i}$, which can be interpreted as the *aggregate* effect of all neighbors' physical interconnection on dynamics (1). The following further assumptions are needed for all subsystems $\mathcal{S}_i$:

Assumption 2: Matrices C_i are such that $\operatorname{rank}(C_i G_i) = \operatorname{rank}(G_i) = g_i$. ◁

Assumption 3: The pair (A_i, C_i) is observable. ◁

Given the structure of (1), the filter design in [18] can be exploited to obtain unbiased minimum-variance local state and disturbance estimates $\hat{x}_i$ and $\hat{\xi}_i$, respectively. We obtain the following estimates of the local state and of the aggregate interconnections (note the presence of $\tilde{y}_i$, and the use of u_i):

$$\begin{aligned}\hat{x}_i(k) = \bar{A}_i(k)\left[A_i\hat{x}_i(k-1) + B_i u_i(k-1)\right] + \\ + \bar{L}_i\tilde{y}_i(k)\end{aligned} \tag{5a}$$

$$\begin{aligned}\hat{\xi}_i(k-1|k) = M_i(k)\,[\tilde{y}_i(k) + \\ -C_i\left(A_i\hat{x}_i(k-1) + B_i u_i(k-1)\right)]\end{aligned} \tag{5b}$$

where

$$\begin{aligned}\bar{A}_i(k) &\doteq \left(\mathbf{I} - K_i(k)C_i\right)\left(\mathbf{I} - G_i M_i(k) C_i\right)\\ \bar{L}_i(k) &\doteq \left(K_i(k) + \left(\mathbf{I} - K_i(k)C_i\right)G_i M_i(k)\right).\end{aligned}$$

Note that these two matrices are related to each other as:

$$\bar{A}_i(k) = \mathbf{I} - \bar{L}_i(k)C_i.$$

Remark 2: Note that (5b) depends on delayed information, as the estimate $\hat{\xi}_i(k-1|k)$ is only available at time k, once measurement $\tilde{y}_i(k)$ is available. This is to be expected, since the ξ_i dynamically affects the state, i.e. the effects of $\xi_i(k-1)$ can only be seen from $\tilde{y}(k)$. ◁

We now repeat Theorem 12 in [18], which gives the theoretical properties of the estimates (5b) and (5a):

Lemma 1 ([18, Thm.12]): Consider the joint input and state estimator in (5), where $M_i(k)$ satisfies:

$$M_i(k)C_i G_i = \mathbf{I}_{g_i}, \quad \forall k \geq 0.$$

If $M_i(k)$ and $K_i(k)$ are designed as in [18], (5b) and (5a) are unbiased estimates of $\xi_i(k-1)$ and $x_i(k)$, minimizing the mean square error over the class of all linear unbiased estimates based on $\bar{x}_i^0$ and $y_i(\kappa), 0 \leq \kappa \leq k$.

Remark 3: Assumption 2 is a sufficient condition for the existence of an estimate which is decoupled from an unknown input ξ_i, both in the stochastic [16] and in the deterministic case [19]. On the other hand, the decomposition $G_i\bar{E}_i$ is needed for the input estimation, as at most $\operatorname{rank}(E_i)$ components can be estimated. By means of the decomposition in (4), ξ_i aggregates the independent components of the interconnection that influence x_i. ◁

In the following, we analyze the specific properties of both the state and unknown-input estimates $\hat{x}_i$ and $\hat{\xi}_i$.

IV. Local State Estimation

Let us start by considering the system in healthy conditions, by analyzing the estimation and residual errors under healthy mode of behavior:

$$\epsilon_i^h(k|k) \doteq x_i^h(k) - \hat{x}_i(k|k) \tag{6a}$$

$$r_i(k|k) \doteq \tilde{y}_i(k) - C_i\hat{x}_i(k|k), \tag{6b}$$

where the superscript h has been added to highlight that the estimation error is considered in nominal conditions. We then analyze the estimation error *under attack*.

Remark 4: Note that, since $\tilde{y}_i = y_i^h$, from Proposition 1, the estimates in (5) only use information from the state which is *not affected* by the attack. As such, it is unnecessary to include the superscript h when analyzing $\hat{x}_i$, as well as dealing with r_i. Conversely, distinguishing between *healthy* and *attacked* information is crucial for error analysis. ◁

Hence, the estimation error dynamics can be derived as:

$$\begin{aligned}\epsilon_i^h(k) = \bar{A}_i(k)&\left[A_i\epsilon_i^h(k-1) + G_i\xi_i(k-1) + w_i(k-1)\right]\\ &- \bar{L}_i(k)v_i(k)\\ =\bar{A}_i(k)&\left[A_i\epsilon_i^h(k-1) + w_i(k-1)\right] - \bar{L}_i(k)v_i(k)\end{aligned} \tag{7}$$

where the interconnection term $G_i\xi_i(k-1)$ is removed thanks to definition of $M_i(k)$ satisfying Lemma 1, as

$$\bar{A}_i(k)G_i = \left(\mathbf{I} - K_i(k)C_i\right)\left(\mathbf{I} - G_i M_i(k)C_i\right)G_i = \mathbf{0}.$$

The influence of the physical interconnections of $\mathcal{S}_i$ is therefore decoupled from the estimation error $\epsilon_i^h(k)$.

As the state x_i is not directly available, the residual error r_i must be used to analyze detection properties. By exploiting the decomposition of ϵ_i in healthy and attacked parts, and using the definition of the residual (6b) and the estimation error under nominal conditions as given in (7), we obtain:

$$\begin{aligned}r_i(k) &= \tilde{y}_i(k) - C_i\hat{x}_i(k) = C_i\epsilon_i^h(k) + v_i(k)\\ &= C_i\bar{A}_i(k)\left[A_i\epsilon_i^h(k-1) + w_i(k-1)\right]\\ &\quad + (\mathbf{I} - C_i\bar{L}_i(k))v_i(k).\end{aligned} \tag{8}$$

Proposition 2: Let an attacker carry out a covert attack as defined in (2) for time $k \geq k_a$, with $x_i^a(k_a) = \mathbf{0}$, and let Assumption 1 hold. The residual $r_i(k)$ is not affected by the covert attack and hence cannot be used to detect it.

Let the estimation error be defined as $\epsilon_i \doteq x_i - \hat{x}_i$. Although a covert attack (2) on $\mathcal{S}_i$ does not influence the local residual r_i, the same cannot be said about the estimation error. This will be exploited further in Sections V and VI to define a residual and a suitable statistical test that enables the detection of covert attacks.

A. State estimation error statistics

We analyze the mean and variance of the residual terms, in order to define a suitable detection strategy. We initialize $\hat{x}_i(0) = \bar{x}_i^0, \forall i \in \mathcal{N}$, and we note that $\epsilon_i(k) = \epsilon_i^h(k), \forall k \leq k_a$ holds in healthy conditions. Given the estimates' unbiasedness property defined in Lemma 1, the mean of the estimation error before the attack occurs is:

$$\mathbf{E}\left[\epsilon_i(k)\right] = \mathbf{0}, \forall k \leq k_a,$$

while $\mathbf{E}\left[\epsilon_i^h(k)\right] = \mathbf{0}$, for all $k \geq 0$. Similarly, the expected value of the residual is $\mathbf{E}\left[r_i(k)\right] = \mathbf{0}, \forall k \geq 0$.

We derive the variance matrix $\Pi_i(k) \doteq \mathrm{Var}(\epsilon_i(k))$ for the estimation error, initializing it as $\Pi_i(0) = \Pi_i^0$:

$$\begin{aligned} \Pi_i(k) = {} & \bar{A}_i(k) A_i \Pi_i(k-1) A_i^\top \bar{A}_i(k)^\top \\ & + \bar{A}_i(k) W_i \bar{A}_i(k)^\top + \bar{L}_i(k) V_i \bar{L}_i(k)^\top \end{aligned} \tag{9}$$

where the covariance terms $\mathrm{Cov}(\epsilon_i(k-1), w_i(k-1))$ and $\mathrm{Cov}(\epsilon_i(k-1), v_i(k))$ have been omitted, as $\epsilon_i^h(k-1)$ is uncorrelated to $w_i(k-1)$ and $v_i(k)$.

For $k > k_a$, i.e. after the occurrence of the attack, the estimation error is $\epsilon_i = x_i^a + x_i^h - \hat{x}_i = x_i^a + \epsilon_i^h$, and as such its mean is given by:

$$\mathbf{E}\left[\epsilon_i(k)\right] = \mathbf{E}\left[x_i^a(k)\right] + \mathbf{E}\left[\epsilon_i^h(k)\right] = x_i^a(k), \forall k > k_a. \tag{10}$$

As the attack strategy in (2) is considered to be deterministic, it will not affect the variance $\Pi_i(k)$. Furthermore, although the estimation error mean is affected by the attack, the expected value of r_i does not change, in line with Proposition 2.

V. Estimation of Coupling Effects

As covert attacks cannot be detected using only local estimates, we exploit the communication between $\mathcal{D}_i$ and its neighbors to detect them in $\mathcal{S}_j, j \in \mathcal{N}_i$. Specifically, we analyze the error between the unknown input estimate (5b) and that computed from the received estimates $\hat{x}_j(k)$ computed by $\mathcal{D}_j$ $\hat{\xi}_i(k-1|k)$. The corresponding error is:

$$\begin{aligned} \rho_i(k-1|k) & \doteq \xi_i(k-1) - \hat{\xi}_i(k-1|k) \\ & = -M_i(k) C_i (A_i \epsilon_i(k-1) - w_i(k-1)) \\ & \quad - M_i(k) v_i(k), \end{aligned} \tag{11}$$

which holds by virtue of Lemma 1. This estimation error therefore depends only on local noise and uncertainties, as $\epsilon_i(k)$ is decoupled from the neighboring subsystems.

Given Lemma 1, the estimate $\hat{x}_i(k)$ is unbiased by construction. Thus it is easy to see that

$$\mathbf{E}\left[\rho_i(k-1|k)\right] = \mathbf{0}. \tag{12}$$

As far as the variance is concerned, from the definitions of the variance matrix (9), it follows that it is possible to evaluate $\mathrm{Var}(\rho_i(k-1|k))$ as:

$$\begin{aligned} \Delta_i(k-1|k) & \doteq \mathrm{Var}(\rho_i(k-1|k)) \\ & = M_i(k) C_i A_i \Pi_i(k-1) A_i^\top C_i^\top M_i^\top(k) + \\ & \quad + M_i(k) C_i W_i C_i^\top M_i^\top(k) + M_i(k) V_i M_i^\top(k). \end{aligned} \tag{13}$$

As $\rho_i(k-1|k)$ is unavailable to $\mathcal{D}_i$, it cannot be used to detect an attack in $\mathcal{S}_j, j \in \mathcal{N}_i$. Instead, supposing that $\mathcal{D}_i$ receives the estimates $\hat{x}_j(k)$ from the neighbors' diagnosis units $\mathcal{D}_j, \forall j \in \mathcal{N}_i$, it is possible to locally define

$$\hat{\rho}_i(k-1|k) \doteq \hat{\xi}_i(k-1) - \bar{E}_i \operatorname*{col}_{j \in \mathcal{N}_i} \left[\hat{x}_j(k-1)\right] \tag{14}$$

that can be regarded as a distributed estimate of the unknown-input estimation error, which may be used for detection. From (14) and (11), we obtain:

$$\hat{\rho}_i(k-1|k) = \bar{E}_i \operatorname*{col}_{j \in \mathcal{N}_i} \left[\epsilon_j(k-1)\right] - \rho_i(k-1|k). \tag{15}$$

Proposition 3: Let Lemma 1 hold. When there are no attacked subsystems $\mathcal{S}_j, j \in \mathcal{N}_i$, the residual $\hat{\rho}_i(k-1|k)$ follows a Gaussian distribution with mean and variance $\mu_i(k)$ and $\Sigma_i(k)$, respectively, where

$$\mu_i(k) = 0, \tag{16a}$$

$$\Sigma_i(k) = \bar{E}_i \operatorname*{diag}_{j \in \mathcal{N}_i} \left[\Pi_j(k-1)\right] \bar{E}_i^\top + \Delta_i(k-1|k). \tag{16b}$$

Proof: Let us examine the expected value of $\hat{\rho}_i(k-1|k)$. Let Subsystem $\mathcal{S}_l, l \in \mathcal{N}_i$ be under attack starting from $k = k_a > 0$; then, it follows from (10), (12), and (15) that the mean $\mu_i(k) \doteq \mathbf{E}\left[\hat{\rho}_i(k-1|k)\right]$ is given by:

$$\mu_i(k) = \begin{cases} 0, & k \leq k_a, \\ \zeta_{i,l}^a(k-1) & k > k_a. \end{cases} \tag{17}$$

where

$$\zeta_{i,l}^a(k-1) \doteq E_{i,[l]} x_l^a(k-1).$$

Here, with some abuse of notation, $\bar{E}_{i,[l]} \in \mathbb{R}^{g_i \times n_l}$ defines the block of row matrix $\bar{E}_i$ corresponding to $\mathcal{S}_l$.

For what concerns the variance $\Sigma_i(k) \doteq \mathrm{Var}(\hat{\rho}_i(k-1|k))$, from the definition of $\mathrm{Var}\left(\hat{\rho}_i\right)$ and (15) it follows that:

$$\Sigma_i(k) = \bar{E}_i \operatorname*{diag}_{j \in \mathcal{N}_i} \left[\Pi_j(k-1)\right] \bar{E}_i^\top + \Delta_i(k-1|k),$$

where the covariance terms satisfy $\mathrm{Cov}(\epsilon_j, \hat{\rho}_i) = \mathbf{0}$, since the estimator error ϵ_i is independent of neighboring states by construction, for all subsystems $\mathcal{S}_i, i \in \mathcal{N}$. ■

It is important to recall that since $x_i^a(k)$ is deterministic, it will not influence the variance of either the estimation error or the residual. Hence, we focus on the estimation error mean. Also note that the residual variance $\Sigma_i(k)$ can be computed locally at subsystem $\mathcal{S}_i$, provided that the neighbors' process and measurement covariance matrices W_j and V_j, and models (A_j, C_j, G_j) are known to $\mathcal{D}_i$.

VI. Detection strategy

In this section we exploit the known statistical properties of the residual $\hat{\rho}_i$, to design a statistical test apt at raising an alarm when suitable conditions are satisfied.

We consider a residual sequence of finite length ω_i, containing samples of $\hat{\rho}_i(k-1|k)$ from $k - \omega_i + 1$ to k:

$$\left\{\hat{\rho}_i(\kappa-1|\kappa)\right\}_{\kappa=k-\omega_i+1}^{k}.$$

The following composite hypothesis test can be formulated. The null hypothesis $\mathcal{H}_i^0$ represents the healthy case when no subsystem $S_j, j \in \mathcal{N}_i$ is under attack, whereas the alternative hypothesis $\mathcal{H}_i^1$ holds otherwise.

Problem 2 (Covert Attack Detection): The detection logic in $\mathcal{D}_i$ accepts one of the following hypotheses:

$$\begin{aligned} & \mathcal{H}_i^0 : \hat{\rho}_i(k-1|k) = \hat{\rho}_i^h(k-1|k), \\ & \mathcal{H}_i^1 : \hat{\rho}_i(k-1|k) = \hat{\rho}_i^h(k-1|k) + \zeta_{i,l}^a(k-1), \end{aligned} \tag{18}$$

given the estimation residual $\hat{\rho}_i(k-1|k)$ defined in (14). Again, the superscript h denotes the component not affected by the attack, $\zeta^a_{i,l}(k)$ is considered to be unknown, and $\hat{\rho}^h_i$ follows the statistic properties in (16). $\triangleleft$

Proposition 4: If $M_i(k)$ is defined according to Lemma 1, and $V_i > 0$, then matrix $\Sigma_i(k)$ is invertible for all $k \geq 0$.

Problem 2 is equivalent to detecting an unknown signal embedded in white Gaussian noise, and as such a solution can be found by means of a Generalized Likelihood Ratio test (see for instance [20]). Hypothesis $\mathcal{H}^1_i$ is accepted when

$$\frac{p\left(\hat{\rho}_i(k-1|k)\left|\hat{\zeta}^a_{i,l}(k-\omega_i),\ldots,\hat{\zeta}^a_{i,l}(k-1),\mathcal{H}^1_i\right.\right)}{p\left(\hat{\rho}(k-1|k)\left|\mathcal{H}^0_i\right.\right)} \tag{19}$$

is greater than a threshold to be defined, where $\hat{\zeta}^a_{i,l}$ is a maximum likelihood estimate of $\zeta^a_{i,l}$. Because of whiteness of $\hat{\rho}^h_i$, $\hat{\zeta}^a_{i,l}(k-1) = \hat{\rho}_i(k-1|k)$ is such an estimate.

Let us define the statistic $T(\hat{\rho}_i,k)$ as the logarithm of (19) and $\theta_i(k)$ as a detection threshold; then we obtain the following detection test:

$$\underbrace{\sum_{\kappa=k-w+1}^{k} \hat{\rho}_i(\kappa-1|\kappa)^\top \Sigma_i(\kappa)^{-1}\hat{\rho}_i(\kappa-1|\kappa)}_{T(\hat{\rho}_i,k)} > \theta_i(k), \tag{20}$$

where it is sufficient for any component of (20) to satisfy the inequality for detection to occur. The probabilities of false alarm and detection are defined as the following:

$$\begin{aligned} \mathrm{P}^f_i(k) &\doteq \Pr\{T(\hat{\rho}_i,k) > \theta_i(k); \mathcal{H}^0_i\}, \\ \mathrm{P}^d_i(k) &\doteq \Pr\{T(\hat{\rho}_i,k) > \theta_i(k); \mathcal{H}^1_i\}. \end{aligned} \tag{21}$$

Since $\Sigma_i(k)$ is symmetric positive definite it is possible to find $U_i(k)$ orthogonal such that $\Sigma_i(k) = U_i(k)\Lambda_i(k)U_i^\top(k)$, with $\Lambda_i(k)$ diagonal. Thus, a transformation

$$\hat{z}_i(k) \doteq U_i(k)\hat{\rho}_i(k)$$

can be defined, where the components of $\hat{z}_i$ are mutually uncorrelated and each q-th component has variance $\lambda_{i[q]}(k)$. Therefore, for $q \in [1, g_i]$, we have that, for threshold $\bar{\theta}_{i[q]}(k)$:

$$T'(\hat{z}_{i[q]},k) = \sum_{\kappa=k-\omega_i+1}^{k} \hat{z}^2_{i[q]}(\kappa-1|\kappa) > \bar{\theta}_{i[q]}(k). \tag{22}$$

Since $\hat{z}_i$ is linearly related to $\hat{\rho}_i$, and in light of (17), $T'(\hat{z}_{i[q]},k)$ follows the distribution:

$$T'(\hat{z}_{i[q]},k) \sim \begin{cases} \chi^2_{\omega_i}(0) & \text{if } \mathcal{H}^0, \\ \chi^2_{\omega_i}(\nu_q) & \text{if } \mathcal{H}^1, \end{cases} \tag{23}$$

where $\chi^2_k(\nu_q)$ is a chi-squared distribution with degree of freedom ω_i and non-centrality parameter

$$\nu_q = \sum_{\kappa=k-\omega_i}^{k-1} \frac{1}{\lambda_{i[q]}(\kappa)}\left(U_{i[q]}(\kappa)\zeta^a_{i,l[q]}(\kappa)\right)^2, \tag{24}$$

where $U_{i[q]}(k)$ denotes the q-th row of matrix $U_i(k)$. Let us define the tail probability of the *normalized* χ^2 distribution as $\Phi(u) \doteq 1 - \Pr\{T'(\hat{z}_{i[q]},k) < u\}$. Then, it is possible to compute the probabilities in (21) for each component q as:

$$\mathrm{P}^f_{i[q]}(k) = \Phi\left(\frac{1}{\sqrt{2\omega_i}}\left(\frac{\bar{\theta}_{i[q]}}{\lambda_{i[q]}} - \omega_i\right)\right) \tag{25a}$$

$$\mathrm{P}^d_{i[q]}(k) = \Phi\left(\frac{\sqrt{2k}\Phi^{-1}(\mathrm{P}^f_{i[q]}(k)) - \nu_q}{\sqrt{4\nu_q + 2\omega_i}}\right). \tag{25b}$$

Remark 5: Note that $T'(\hat{z}_{i[q]},k)$ represents the energy of the attack received by $\mathcal{S}_i$. From (25b) it can be seen that the probability of detection decreases as the attack energy decreases. Furthermore, as $\nu_q \to 0$, the probability of detection approaches that of false alarm. More precisely, ν_q depends on the energy of the attacked state x^a_l as scaled by the corresponding interconnection weight.

Also, note that the presence of the input estimate variance $\lambda_{i[q]}(k)$ reduces the effect of the attack on ν_q. $\triangleleft$

Eqs. (25a) and (25b) hold component-wise. It is possible to find an expression for the probability of false alarm $\mathrm{P}^f_i(k)$ of detector $\mathcal{D}_i$ by observing that the probability of at least one false alarm is the complementary to the probability of no false alarms. Thus, recalling that the components of $\hat{z}_i$ are independent by construction, we have that:

$$\mathrm{P}^f_i(k) = 1 - \prod_{q=1}^{g_i}\left(1 - \mathrm{P}^f_{i[q]}(k)\right). \tag{26}$$

If we assume the same probability $\mathrm{P}^f_{i[q]}(k)$ for each component q, it is possible to invert (26) and (25a). This allows to compute individual thresholds $\bar{\theta}_{i[q]}$, given a desired cumulative probability $\mathrm{P}^f_i(k)$. The overall probability of detection can be found in the same way, although it depends on ν_q.

VII. Numerical Simulations

A. Simulation setup

We consider a CPS composed of $N = 4$ subsystems, interconnected as in Figure 1. We consider the linearized model of multiple pendula coupled through a spring, as presented in [21, Ex. 1.36], where each subsystem is described by:

$$m_i l_i^2 \ddot{\delta}_i = m_i g l_i \delta_i + u_i + \sum_{j \in \mathcal{N}_i} k_{ij} a_i^2(\delta_j - \delta_i), \tag{27}$$

where δ_i, m_i, l_i are respectively the displacement angle, mass, and length of the pendulum; g is the gravitational constant; k_{ij} is the spring coefficient, with $k_{ij} = k_{ji}$, and a_i is the height at which the spring is attached to pendulum i. The parameter values used in the numerical simulation can be found in Table I.

Choosing $x_i \doteq [\delta_i, \dot{\delta}_i]^\top$, and defining a decentralized state feedback control law $u_i \doteq \mathcal{K}_i x_i$, we discretize the pendulum's dynamics subsystem-by-subsystem with Euler's approximation with sampling time $T_s = 0.01$ s, preserving the topology and the interconnection structure of the CPS. For all subsystems, we assume that all states are measurable, i.e. $C_i = \mathbf{I}_2$. The process and measurement noise variance matrices are $W_i = 10^{-3}\mathbf{I}$ and $V_i = 10^{-3}\mathbf{I}$. We run the simulation for 100 s.

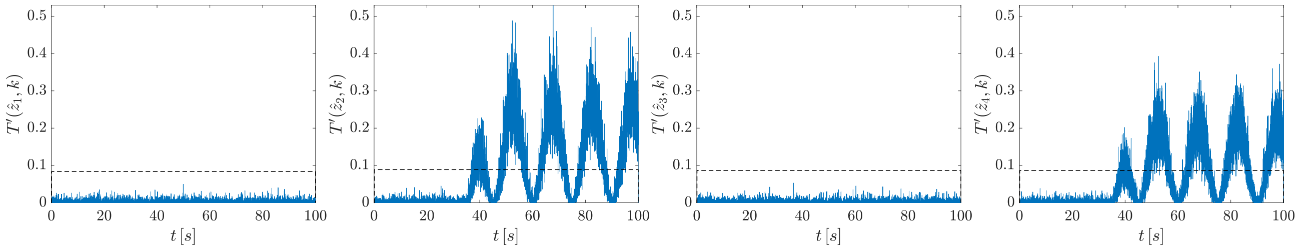

Fig. 2: Comparison of the statistic $T'(\hat{z}_i,k)$, in blue, against the detection threshold $\bar{\theta}_i(k)$, in dashed-black.

TABLE I: Subsystem and interconnection parameters

m_i	l_i	a_i	k_{12}	k_{23}	k_{24}	k_{34}
0.5 kg	0.1 m	0.06 m	27	40	35	53

From (27), and considering the Euler approximation for the discretization of each subsystem, it is possible to choose $G_i = [0, 1]^\top$. Note that $\xi_i \in \mathbb{R}$, for all subsystems.

B. Attack scenario and detection

Starting from time $k_a = 35$ s, an attacker is able to inject

$$\eta_3(k) \doteq 0.5\left(1 - e^{-0.3(k\cdot T_s - k_a)}\right)\sin\left(\frac{2}{30}\pi k \cdot T_s\right),$$

where the attenuation term $\left(1 - e^{-0.3(k\cdot T_s - k_a)}\right)$ is added to reduce the transient behavior of the attack. In Figure 2, we show the effectiveness of our detection technique, by comparing the statistic $T'(\hat{z}_i,k)$, computed by using a window of size $\omega_i = 20$, to the threshold $\bar{\theta}_i(k)$, defined for all subsystems such that the probability of false alarm $P_i^f = 0.05$.

At time $k = 36.78$ s, detector $\mathcal{D}_2$ detects the presence of an attack in $\mathcal{N}_2$, while $\mathcal{D}_4$ detects the attack in $\mathcal{N}_4$ at time $k = 37.16$ s. As expected, the diagnosers for subsystems $\mathcal{S}_1$ and $\mathcal{S}_3$ do not detect an attack.

VIII. Concluding Remarks

We have proposed a distributed method capable of detecting local covert attacks in interconnected CPSs with stochastic uncertainties. The proposed method is based on the joint estimation of local states and the neighbors' cumulative effect; communication among subsystems enables definition of a suitable residual signal and a related statistical test.

Future work will include studying additional detectability properties of the proposed approach and comparison to other techniques for solving Problem 2, as well as investigation into the architecture's robustness to other types of attacks.

References

[1] 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.

[2] 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.

[3] D. I. Urbina, J. Giraldo, A. A. Cardenas, J. Valente, M. Faisal, N. O. Tippenhauer, J. Ruths, R. Candell, and H. Sandberg, *Survey and new directions for physics-based attack detection in control systems*. US Department of Commerce, National Institute of Standards and Technology, 2016.

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

[5] 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.

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

[7] A. Teixeira, H. Sandberg, and K. H. Johansson, "Networked control systems under cyber attacks with applications to power networks," in *American Control Conference*, 2010, pp. 3690–3696.

[8] R. Anguluri, V. Katewa, and F. Pasqualetti, "Attack detection in stochastic interconnected systems: Centralized vs decentralized detectors," in *57th IEEE Conference on Decision and Control (CDC)*, 2018, pp. 4541–4546.

[9] F. Pasqualetti, F. Dörfler, and F. Bullo, "A divide-and-conquer approach to distributed attack identification," in *2015 IEEE 54th Annual Conference on Decision and Control*, 2015, pp. 5801–5807.

[10] A. Barboni, H. Rezaee, F. Boem, and T. Parisini, "Distributed detection of covert attacks for interconnected systems," in *European Control Conference (ECC)*, 2019, Accepted.

[11] 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 *European Control Conference (ECC)*, 2018, pp. 344–349.

[12] F. Boem, R. M. G. Ferrari, C. Keliris, T. Parisini, and M. M. Polycarpou, "A distributed networked approach for fault detection of large-scale systems," *IEEE Trans. on Automatic Control*, vol. 62, no. 1, pp. 18–33, 2017.

[13] 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.

[14] 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.

[15] 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 *57th IEEE Conference on Decision and Control (CDC)*, 2018, pp. 4553–4558.

[16] P. K. Kitanidis, "Unbiased minimum-variance linear state estimation," *Automatica*, vol. 23, no. 6, pp. 775–778, 1987.

[17] M. Hou and R. Patton, "Optimal filtering for systems with unknown inputs," *IEEE transactions on Automatic Control*, vol. 43, no. 3, pp. 445–449, 1998.

[18] S. Gillijns and B. De Moor, "Unbiased minimum-variance input and state estimation for linear discrete-time systems," *Automatica*, vol. 43, no. 1, pp. 111–116, 2007.

[19] J. Chen, R. J. Patton, and H.-Y. Zhang, "Design of unknown input observers and robust fault detection filters," *International Journal of control*, vol. 63, no. 1, pp. 85–105, 1996.

[20] S. M. Kay, *Fundamentals of statistical signal processing*. Prentice Hall, 1993.

[21] D. D. Siljak, *Decentralized control of complex systems*, ser. Mathematics in science and engineering. Academic Press, 1990, vol. 184.