Distributed cyber-attack isolation for large-scale interconnected systems

Alexander J. Gallo, Francesca Boem, Thomas Parisini

***Abstract*— This work addresses the problem of cyber-attack isolation within a distributed diagnosis architecture for large-scale interconnected systems. Considering a distributed control architecture, malicious agents are capable of compromising the data exchanged between distributed controllers. Building on a distributed detection strategy existent in literature, in this paper we propose a distributed isolation algorithm to identify the attacked communication link. After presenting the isolation algorithm, we give a necessary and a sufficient condition for isolation to occur, relating to the structure of the physical interconnection matrices. We demonstrate the effectiveness of the proposed technique through numerical simulations.**

I. Introduction

With the integration of an ever increasing number of cyber resources in control systems, such as distributed computing processors, wireless communication networks, and low cost sensors, significant work has been put into the study of cyber-physical systems (CPS) [1]. This evolution of industrial systems and infrastructure presents many benefits, enabling a more effective and efficient regulation of complex systems. Conversely, introducing cyber-resources, particularly communication networks, into control systems exposes them to cyber-attacks. Recently, a number of cyber-security threats have demonstrated the disruption these attacks may cause the attacked systems and society as a whole, e.g. [2], [3].

Thus to ensure safe operation, security must be included in the design of control systems [4]. Specifically, as delineated in [5], there are three important features that must be included in cyber-secure control systems, namely: cyber-attack *detection*, *isolation*, and *control reconfiguration*. "Detection" is the problem of evaluating whether the system, and all its components (i.e. the plant, the actuators, the sensors, etc.), are behaving nominally or whether they are under attack; "isolation" relates to understanding where a cyber-attack is present; finally, "control reconfiguration" is the problem of automatically redesigning the control system such that, if possible some level of performance may be maintained, or else to ensure graceful degradation of the system. In this paper, we focus primarily on the problem of isolation.

This work has been partially supported by European Union's Horizon 2020 research and innovation program 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. J. Gallo is with the Delft Center for Systems and Controls, at the Delft Technical University, the Netherlands a.j.gallo@tudelft.nl

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

A common feature in modern engineering systems is that they are large-scale, and often composed of physically coupled interconnected subsystems. It is well known that this class of systems requires distributed methods to be implemented, to address both the cost and constraint of communication networks associated to centralized control architectures, and to manage the high computational burden required to centrally compute complex control laws. This, clearly, also applies to architectures for cyber-security for large-scale interconnected cyber-physical systems, leading to recent research interest in the subject [6], [7], [8], [9], [10]. Of these works, however, few focus on the problem of distributed cyber-attack isolation [11]. It is worth noting, that there is a consistent literature on distributed fault detection and isolation (FDI), which addresses the problem of isolating the presence of faults in large-scale and complex systems (e.g. [12], [13], [14], [15], as well as the references in [16]).

In this work, we present a distributed cyber-attack isolation algorithm, extending the monitoring architecture in [6], analyzing cyber-attacks affecting the communication between neighboring distributed controllers. Specifically, we show how isolation may be implemented, while giving structural constraints limiting the isolation of the cyber-attack. Our contributions are threefold: an isolation algorithm, complementing the detection algorithm in [6]; the definition of a sufficient condition on the structure of the physical coupling between subsystems that allow for the construction of an isolation algorithm; the definition of a necessary condition to be satisfied by the physical coupling, such that isolation may occur.

The structure of this paper is the following: in Section II we formally introduce the problem of distributed attack isolation, given a large-scale interconnected system; in Section III we summarize the detection architecture presented in [6], and provide an overview of its detection properties; in Section IV we give the main results of this paper: after presenting an isolation algorithm, we provide a necessary and a sufficient condition on the structure of the physical coupling to verify whether isolation may occur or not; finally, in Section V we present numerical validation of our theoretical results.

Notation: Throughout this paper we use the following notation. $\mathbb{N}_0$ is the set of non-negative integers. I_n represents the n-dimensional identity matrix, while $0_{n\times m}$ is a matrix of zeros in $\mathbb{R}^{n\times m}$. When clear from context, I and 0 are used. For a matrix A, $A^\dagger$ denotes its right pseudo-inverse. Given a matrix X, $\sigma(X)$ is its spectrum, and $\rho(X)$ its spectral radius. For a vector x_i, with i an index in a set $\mathcal{N}$, $x_{i,k}$ denotes its k-th component. The operator $|\cdot|$ applied to a set determines its cardinality, while used with matrices or

vectors it defines their component-by-component absolute value. In this paper inequalities are considered component-by-component, i.e. for two matrices A and B with the same dimensions, $A \geq B$ indicates the element-wise inequality; the same is considered for vectors. With $\mathrm{col}(\cdot)$, $\mathrm{diag}(\cdot)$, and $\ker(\cdot)$ we define the column concatenation of vectors or matrices, the block-diagonal concatenation of matrices, and the null-space of a matrix, respectively.

II. Problem formulation

A. Large-scale system modeling

We consider a large-scale interconnected system composed of N physically-coupled subsystems, $\mathcal{S}_i, i \in \mathcal{N} \doteq \{1, \dots, N\}$. Each subsystem has dynamics:

$$\mathcal{S}_i : \begin{cases} x_i^+ = A_i x_i + B_i u_i + \xi_i + w_i \\ y_i = C_i x_i + v_i \end{cases} \tag{1}$$

with x_i^+ symbolizing $x_i(k+1), k \in \mathbb{N}_0$, x_i the state of the subsystem, u_i its control input, w_i process disturbance, y_i the measurement output, v_i the measurement disturbance, and ξ_i modeling the aggregate physical coupling between $\mathcal{S}_i$ and a set $\mathcal{N}_i \subset \mathcal{N}$; this set of subsystems is called the set of "neighbors" of $\mathcal{S}_i$, and is defined as $\mathcal{N}_i \doteq \{j \in \mathcal{N} | \partial x_i / \partial x_j \neq 0\}$, i.e. the set of indices of those subsystems that dynamically affect the state of $\mathcal{S}_i$. The aggregate physical coupling term is defined as $\xi_i = \sum_{j \in \mathcal{N}_i} A_{ij} x_j$, with A_{ij} the so-called "coupling" matrix between $\mathcal{S}_j$ and $\mathcal{S}_i$. All matrices in (1) are given to be of appropriate dimensions.

Assumption 1: For all subsystems $\mathcal{S}_i$, the following hold:

- (A_{ii}, B_i) is stabilizable;
- (C_i, A_{ii}) is detectable;
- the disturbances $w_i(k)$ and $v_i(k)$ are such that

$$|w_i(k)| \leq \bar{w}_i \,, \quad |v_i(k)| \leq \bar{v}_i \tag{2}$$

for all $k \in \mathbb{N}_0$. ◁

B. Distributed control architecture

Each subsystem is controlled by a controller $\mathcal{C}_i$, which is designed to be *distributed*, i.e. to rely only on information which is *local* to $\mathcal{S}_i$, or that of its neighbors $\mathcal{S}_j, j \in \mathcal{N}_i$. We formalize this by defining the regulating input u_i as $u_i = \kappa_i(y_i, \gamma_i^c)$, where $\kappa_i(\cdot, \cdot)$ is some operator defining the control policy, and $\gamma_i^c \doteq \underset{j \in \mathcal{N}_i^c}{\mathrm{col}} \gamma_{ji}^c$ is the combination of all signals γ_{ji}^c that $\mathcal{C}_i$ *receives* from the neighboring controllers over some communication network. The set $\mathcal{N}_i^c$ represents the collection of all subsystems which transmit data to $\mathcal{S}_i$.

Assumption 2: The communication network is such that $\mathcal{N}_i \subseteq \mathcal{N}_i^c$ holds for all subsystems $\mathcal{S}_i, i \in \mathcal{N}$. ◁

The inclusion of communication resources in the control system exposes it to the possibility of cyber-attacks. We model this in the definition of the received signal γ_{ji}^c as:

$$\gamma_{ji}^c(k) = y_j(k) + \beta_{ji}(k - K_{ji}^a)\alpha_{ji}(k), \tag{3}$$

where $\beta_{ji}(k - K_{ji}^a)\alpha_{ji}(k)$ models the effect of an attack on γ_{ji}^c, and K_{ji}^a is the first instant an attacker injects a signal on the communication between $\mathcal{S}_j$ and $\mathcal{S}_i$. The attack vector $\alpha_{ji}(\cdot)$ is defined by the malicious agent to disrupt the nominal[1] operations of the system, and $\beta_{ji}(\cdot)$ is a so-called activation function, borrowing terminology from the fault detection and isolation literature [17].

C. Distributed diagnosis architecture

The potential exposure to malicious alteration of the behavior motivates the introduction of a distributed diagnostic module $\mathcal{D}_i$, which is tasked with detecting whether the information $\gamma_{ji}^c, j \in \mathcal{N}_i^c$ is nominal or not, relying on the following assumption.

Assumption 3 (Trusted information set): For each subsystem $\mathcal{S}_i$, the set of information that is trusted by diagnoser $\mathcal{D}_i$ is defined as:

$$\mathcal{I}_i^{\mathrm{T}} = \{\mathcal{M}_i, u_i, y_i\} \tag{4}$$

with $\mathcal{M}_i$ defining the knowledge of the overall system available to $\mathcal{D}_i$. ◁

Remark 1: In Assumption 3, we consider input and measurement vectors u_i and y_i to be trusted by $\mathcal{D}_i$. ◁

The problem of distributed attack detection and isolation can be formalized, borrowing from the fault detection and isolation literature, as follows:

Problem 1 (Distributed attack detection): Design a diagnostic unit $\mathcal{D}_i$ capable of verifying whether the received information γ_i^c is nominal or under attack. △

Problem 2 (Distributed attack isolation): Given detection of an attack active on γ_i^c by $\mathcal{D}_i$, isolate the communication link (j, i) over which an attack $\alpha_{ji}(k) \neq 0$. △

III. Distributed detection architecture

Before moving forward with the presentation of our results on cyber-attack isolation, let us briefly summarize the detection architecture we take into consideration. We consider that each subsystem $\mathcal{S}_i$ is equipped with a distributed diagnosis unit $\mathcal{D}_i$ presented in [6].

A. Structure of diagnoser $\mathcal{D}_i$

The distributed diagnosis module $\mathcal{D}_i$ we adopt in this paper is made up of two parallel modules, $\mathcal{O}_i$ and $\mathcal{O}_{ji}$, each of which is composed of the following elements [6]: 1) a state estimator; 2) a residual generator; 3) a detection test. Although these elements are common to both modules, their design is dependent on the information available to each. In the following we briefly describe the two modules.

1) Local state estimation – $\mathcal{O}_i$: The module $\mathcal{O}_i$ exploits a distributed Luenberger-like observer of the form

$$\hat{\mathcal{S}}_i : \begin{cases} \hat{x}_i^+ = A_{ii}\hat{x}_i + B_i u_i + \hat{\xi}_i + L_i(y_i - C_i\hat{x}_i) \\ \hat{y}_i = C_i \hat{x}_i \end{cases} \tag{5}$$

where L_i is such that $A_{Li} \doteq A_{ii} - L_i C_i$ is Schur stable, and $\hat{\xi}_i$ is the estimate of the coupling between $\mathcal{S}_i$ and its neighbors, computed from received data vector γ_i as $\hat{\xi}_i = \sum_{j \in \mathcal{N}_i} A_{ij} C_j^\dagger \gamma_{ji}^c$.

[1]Throughout this paper, by "nominal" we intend the operations of the system as not exposed to attacks.

Assumption 4: The matrices C_i are such that $\ker C_i \subseteq \ker A_{ji}$ for all subsystems $\mathcal{S}_i$, and all $j \in \mathcal{N}_i$. $\triangleleft$

Given the definition of L_i such that $\rho(A_{Li}) < 1$ and Assumption 4, the dynamics of the estimation error $\epsilon_i \doteq x_i - \hat{x}_i$ are asymptotically stable, and therefore, given Assumption 1, $\epsilon_i(k)$ is bounded in nominal conditions. Furthermore, given knowledge on $\bar{w}_i$ and $\bar{v}_i$, a bound $\bar{\epsilon}_i$ can be explicitly found such that $|\epsilon_i(k)| \leq \bar{\epsilon}_i(k), \forall k \in \{0, \dots, K^a_{ji}\}$. Thus, computing the residual as $r_i \doteq y_i - \hat{y}_i$, the module $\mathcal{O}_i$ is equipped with the following detection test:

$$|r_i(k)| > \bar{r}_i(k) \tag{6}$$

where $\bar{r}_i(k)$ is an appropriately defined threshold, guaranteeing that $|r_i(k)| \leq \bar{r}_i(k)$ for all $k \in \{0, \dots, K^a_{ji}\}$. Then, if (6) holds for at least one component of $r_i(k)$, an attack is detected, solving Problem 1.

2) Estimation of neighboring states – $\mathcal{O}_{ji}$: The module $\mathcal{O}_{ji}$ exploits an unknown-input observer $\hat{\mathcal{S}}_{ji}$ to estimate the state of $\mathcal{S}_j, j \in \mathcal{N}_i^c$ without requiring further information to be transmitted. In rough terms, $\mathcal{O}_{ji}$ exploits trusted knowledge of the dynamics of $\mathcal{S}_j, j \in \mathcal{N}_i$, included in $\mathcal{M}_i$ in the trusted information set $\mathcal{I}_i^T$, to verify whether the time-varying behavior of γ^c_{ji} is consistent with its nominal dynamics. Specifically, the dynamics in (1) are rewritten as

$$\mathcal{S}_j : \begin{cases} x_j^+ = A_{jj}x_j + E_j d_j + w_j \\ y_j = C_j x_j + v_j \end{cases} \tag{7}$$

introducing an unknown input vector d_j, which is defined together with full column rank matrix E_j such that $E_j d_j = B_j u_j + \xi_j$. Thus, the UIO takes the form

$$\hat{\mathcal{S}}_{ji} : \begin{cases} z_{ji}^+ = F_{ji} z_{ji} + \widehat{K}_{ji} \gamma^c_{ji} \\ \hat{x}_{ji} = z_{ji} + H_{ji} \gamma^c_{ji} \end{cases} \tag{8}$$

where z_{ji} is the internal state of the observer, $\hat{x}_{ji}$ is the state estimate, and H_{ji}, F_{ji} and $\widehat{K}_{ji}$ are appropriately designed such that the estimation error $\epsilon_{ji} = x_j - \hat{x}_{ji}$ is decoupled from d_j, as presented in [18]. To guarantee existence of the observer matrices, the following is assumed.

Assumption 5: For all $\mathcal{S}_i, i \in \mathcal{N}$, the dynamics (1) satisfy:

- C_i is such that $\operatorname{rank}(C_i E_i) = \operatorname{rank} E_i$;
- the system defined by the tuple (C_i, A_{ii}, E_i) is strongly observable [19, Def 7.15]. $\triangleleft$

Given its design, $\epsilon_{ji}(k)$ is bounded for $k \in \mathbb{N}_0$, and therefore, given knowledge of $\bar{w}_j$ and $\bar{v}_j$, a bound $\bar{\epsilon}_{ji}(k)$ can be defined such that $|\epsilon_{ji}(k)| \leq \bar{\epsilon}_{ji}(k)$ holds for all $k \in \{0, \dots, K^a_{ji}\}$. Therefore, a similar bound on the residual $r_{ji} = \gamma^c_{ji} - C_j \hat{x}_{ji}$ can be found, and the following detection test may be used for detection by the module $\mathcal{O}_{ji}$

$$|r_{ji}(k)| > \bar{r}_{ji}(k), \tag{9}$$

for the information received by $\mathcal{S}_i$ from all subsystems $\mathcal{S}_j, j \in \mathcal{N}_i^c$. Thus, if (9) holds for at least one component of $r_{ji}(k)$, an attack is detected, solving Problem 1.

B. Properties of $\mathcal{D}_i$

Having presented the structure of $\mathcal{D}_i$ as introduced in [6], let us now summarize its properties briefly. Firstly, note that the diagnoser $\mathcal{D}_i$ has the combined properties of $\mathcal{O}_i$ and $\mathcal{O}_{ji}$. Indeed, it is sufficient that either (6) or (9) hold for $\mathcal{D}_i$ to detect an attack. This implies that any attack guaranteed to be detected by either $\mathcal{O}_i$ or $\mathcal{O}_{ji}$ is also guaranteed to be detected by $\mathcal{D}_i$. On the other hand, similarly, for an attack to be stealthy it must be designed to be undetected by both $\mathcal{O}_i$ and $\mathcal{O}_{ji}$.

Here we focus on attacks designed to be *stealthy* [20], i.e. defined by a malicious agent such that the detection module $\mathcal{D}_i$ cannot detect their effect on γ_i^c. For a detailed analysis of the properties of $\mathcal{D}_i$, we refer the interested reader to [6].

1) Attacks stealthy to $\mathcal{O}_i$: Given that $\mathcal{O}_i$ is designed using limited knowledge of the large-scale interconnected system's dynamics, namely only those of $\mathcal{S}_i$, this module relies on the physical coupling between subsystems to provide the necessary analytical redundancy for detection. Specifically, it was shown in [6] that any attack satisfying

$$\left(\operatorname*{diag}_{j \in \mathcal{N}_i} C_j^\dagger \right) \alpha_i \in \ker \left[\operatorname*{row}_{j \in \mathcal{N}_i} A_{ij} \right] \tag{10}$$

is stealthy to $\mathcal{O}_i$, with $\alpha_i \doteq \operatorname*{col}_{j \in \mathcal{N}_i} \alpha_{ji}$. This *limitation* was shown in [21] to be structural, i.e. to be common to any distributed diagnoser exploiting the local dynamics of $\mathcal{S}_i$ and the physical coupling ξ_i to detect an attack on γ_i^c.

2) Attacks stealthy to $\mathcal{O}_{ji}$: Let us now focus on the properties of $\mathcal{O}_{ji}$. Given the subsystem dynamics described by (7) and the observer structure in (8), the subsystem can be seen as stand-alone from the rest of the large-scale interconnected system, as the coupling with its neighbors is contained in the unknown input d_j, as well as, in turn, each UIO $\hat{\mathcal{S}}_{ji}$. Hence, $\mathcal{O}_{ji}$ is vulnerable to all those attacks that are stealthy to a centralized diagnoser, as highlighted in [20], including zero-dynamics, covert, and replay attacks.

Remark 2: It is important to note that, given its design in (5), $\mathcal{O}_i$ can only detect attacks on $\gamma^c_{ji}, j \in \mathcal{N}_i$. Thus, for all $j \in \mathcal{N}_i^c \backslash \mathcal{N}_i$, any attack stealthy to $\mathcal{O}_{ji}$ is stealthy to $\mathcal{D}_i$. $\triangleleft$

IV. Distributed cyber-attack isolation

Having shown the main detection characteristics of $\mathcal{D}_i$, we are now ready to present the main results of this paper, the cyber-attack isolation by the distributed diagnoser. Given the design of $\mathcal{O}_i$ and $\mathcal{O}_{ji}$, and their properties as described above, we can see that the distributed isolation of cyber-attacks by $\mathcal{D}_i$ changes depending on whether (6) or (9) hold. In the following, we treat each case separately.

A. Isolation via $\mathcal{O}_{ji}$

Let us start by noting that if $\mathcal{O}_{ji}$ detects an attack, it is also isolated. This is shown in the following proposition.

Proposition 1: Consider a subsystem $\mathcal{S}_i$, receiving communication signal γ^c_{ji} from each of its neighbors $\mathcal{S}_j, j \in \mathcal{N}_i^c$, and monitored by the distributed diagnoser $\mathcal{D}_i$. If an attack

$\alpha_{ji}(k) \neq 0$ is such that (9) holds for $\mathcal{O}_{ji}$, then the attack is isolated to the communication link (j,i). □

Proof: Due to space constraints proofs are omitted. ■

B. Isolation by $\mathcal{O}_i$

Differently from the above, when an attack is detected by $\mathcal{O}_i$, but not by $\mathcal{O}_{ji}$, the possibility of isolating an attack depends on the structural characteristics of the physical coupling matrices A_{ij}. Here, we consider the attack function α_{ji} to be designed by a malicious agent to be stealthy to $\mathcal{O}_{ji}$. Indeed, if this were not the case, there may be some time $K_{ji}^D > K_i^D$ for which (9) holds, and thus $\mathcal{O}_{ji}$ isolates the attack α_{ji}, where K_i^D is the first time instant for which (6) holds for at least one component of $|r_i(K_i^D)|$.

We here consider the malicious agent implements a covert attack [20]. Suppose that an attacker injecting an attack signal α_{ji} on the communication link (j,i) has perfect knowledge of the dynamic model of $\mathcal{S}_j$, given by the tuple (C_j, A_{jj}, E_j). Then, an attack strategy is said to be "covert" if α_{ji} is the result of the following dynamics:

$$\begin{cases} x_{ji}^{a+} = A_{jj}x_{ji}^a + E_j d_{ji}^a \\ \alpha_{ji} = C_j x_{ji}^a \end{cases} \tag{11}$$

with initial condition $x_{ji}^a(K_{ji}^a) = 0$, and where d_{ji}^a is defined by the attacker. It has been shown both in centralized and distributed scenarios that these attacks are stealthy to diagnosers with similar structure to $\mathcal{O}_{ji}$ [20], [7], [6]. As such, for an attack α_{ji} defined as in (11), there is no time K_{ji}^D for which (9) holds. Therefore, covert attacks are an appropriate class of attacks to analyze the isolation performance of $\mathcal{O}_i$.

We base our isolation algorithm on strategies of unknown-input reconstruction, such as those defined in [22], [23]. Specifically, for time $k \geq K_i^D$, a set of N_i unknown-input estimators are constructed to estimate $\alpha_{ji}(k)$ for all $j \in \mathcal{N}_i$.

Remark 3: Given we are analyzing the isolation properties of $\mathcal{O}_i$, we can only isolate attacks on links $(j,i), j \in \mathcal{N}_i$, as a direct consequence of Remark 2, thus not isolating attacks on $(j,i), j \in \mathcal{N}_i^c \backslash \mathcal{N}_i$. ◁

Let us now introduce the input reconstruction method used to estimate α_{ji}, critical for the isolation algorithm. It is well known that it is only possible to estimate $n_{ij} \doteq \operatorname{rank} A_{ij}$ components of α_{ji}. Thus, introduce $\bar{A}_{ij}$ and $\bar{\alpha}_{ji}$ satisfying $\bar{A}_{ij}\bar{\alpha}_{ji} = A_{ij}\alpha_{ji}$, with $\bar{\alpha}_{ji} \in \mathbb{R}^{n_{ji}}$, $\bar{A}_{ij} \in \mathbb{R}^{n_i \times n_{ij}}$ and $\operatorname{rank} \bar{A}_{ij} = n_{ij}$. Note that, given Assumption 5, the matrix $\bar{A}_{ij}$ is such that $\operatorname{rank}(C_i\bar{A}_{ij}) = n_{ij}$, and it can be seen as a basis for $\operatorname{Im} A_{ij}$. Before we move on with the presentation of the estimate $\hat{\bar{\alpha}}_{ji}$, we introduce the following from [23].

Lemma 1 (Lemma 1, [23]): Suppose X and Y are $n \times m$ and $p \times n$ matrices, respectively. Then $\operatorname{rank}(YX) = \operatorname{rank} X$ if and only if exist nonsingular matrices P and S such that

$$P^{-1}X = \begin{bmatrix} X_1 \\ 0 \end{bmatrix}, \quad S^{-1}YP = \begin{bmatrix} Y_1 & 0 \\ 0 & Y_2 \end{bmatrix} \tag{12}$$

where X_1 and Y_1 have the same number of rows, with X_1 full row rank and Y_1 invertible. □

Given Assumption 5 and Lemma 1, it is possible to define P_{ij} and S_{ij} such that:

$$P_{ij}^{-1}\bar{A}_{ij} = \begin{bmatrix} \bar{A}_{ij,1}^{(ij)} \\ 0 \end{bmatrix}, \quad S_{ij}^{-1}C_iP_{ij} = \begin{bmatrix} C_{i,1}^{(ij)} & 0 \\ 0 & C_{i,2}^{(ij)} \end{bmatrix},$$

with $\bar{A}_{ij,1}^{(ij)} \in \mathbb{R}^{n_{ij} \times n_{ij}}$ and invertible, given $\bar{A}_{ij}$ is full column rank, and $C_{i,1} \in \mathbb{R}^{n_{ij} \times n_{ij}}$. The transformation matrix P_{ij} is such that

$$P_{ij}^{-1}A_{Li}P_{ij} = \begin{bmatrix} A_{Li,11}^{(ij)} & A_{Li,12}^{(ij)} \\ A_{Li,21}^{(ij)} & A_{Li,22}^{(ij)} \end{bmatrix}.$$

Note that the pair $(C_{i,2}^{(ij)}, A_{Li,22}^{(ij)})$ is observable, given that the system defined by the tuple (C_i, A_{Li}, A_{ij}) is strongly observable for all $j \in \mathcal{N}_i$, following Assumption 5 and appropriate construction of L_i in (5).

Let us define the following transformations: $\epsilon_i = P_{ij}\begin{bmatrix} \eta_{i,1}^{(ij)\top} & \eta_{i,2}^{(ij)\top} \end{bmatrix}^\top$, $r_i = S_{ij}\begin{bmatrix} \rho_{i,1}^{(ij)\top} & \rho_{i,2}^{(ij)\top} \end{bmatrix}^\top$, with $\eta_{i,1}^{(ij)} \in \mathbb{R}^{n_{ij}}$ and $\rho_{i,1}^{(ij)} \in \mathbb{R}^{n_{ij}}$. An estimate $\hat{\bar{\alpha}}_{ji}$ of $\bar{\alpha}_{ji}$ can be computed as:

$$\begin{aligned} \hat{\bar{\alpha}}_{ji}(k|k+1) = \bar{A}_{ij,1}^{(ij)\ -1} \Big(& C_{i,1}^{(ij)-1}\rho_{i,1}^{(ij)}(k+1) \\ & - A_{Li,11}^{(ij)}C_{i,1}^{(ij)-1}\rho_{i,1}^{(ij)}(k) - A_{Li,12}^{(ij)}\hat{\eta}_{i,2}^{(ij)}(k) \Big) \end{aligned} \tag{13}$$

where $\hat{\eta}_{i,2}^{(ij)}$ is an estimate of $\eta_{i,2}^{(ij)}$, asymptotically convergent to $\eta_{i,2}^{(ij)}$ which can be implemented, without loss of generality, as a Luenberger-like observer, given the observability of $(C_{i,2}^{(ij)}, A_{Li,22}^{(ij)})$. Specifically, $\hat{\eta}_{i,2}^{(ij)}$ takes the form

$$\begin{aligned} \hat{\eta}_{i,2}{}^{(ij)+} = & A_{Li,22}^{(ij)}\hat{\eta}_{i,2}^{(ij)} + A_{Li,21}^{(ij)}C_{i,1}^{-1}\rho_{i,1}^{(ij)} \\ & + \mathcal{L}_i^{(ij)}\left(\rho_{i,2}^{(ij)} - C_{i,2}^{(ij)}\hat{\eta}_{i,2}^{(ij)}\right) + \sum_{k \in \mathcal{N}_i^j} \tilde{A}_{ik,2}^{(ij)}C_j^\dagger \hat{\bar{\alpha}}_{ki} \end{aligned} \tag{14}$$

where $\mathcal{N}_i^j \doteq \mathcal{N}_i \backslash \{j\}$, $\mathcal{L}_i^{(ij)}$ is such that $(A_{Li,22}^{(ij)} - \mathcal{L}_i^{(ij)}C_{i,2}^{(ij)})$ is Schur stable, $\tilde{A}_{ik,2}^{(ij)} \in \mathbb{R}^{(n_i - n_{ij}) \times n_j}$ is defined as $P_{ij}^{-1}A_{ik} = \begin{bmatrix} \tilde{A}_{ik,1}^{(ij)\top} & \tilde{A}_{ik,2}^{(ij)\top} \end{bmatrix}^\top$, and $\hat{\bar{\alpha}}_{ki}(k|k+1)$ is the attack estimated assuming (k,i) is the attacked link.

Remark 4: Note that with (13) we introduce a single time-step delay. This is a feature of input reconstruction algorithms, given the dynamic relationship between α_{ji} and ϵ_i. Furthermore, the estimate of α_{ji} depends on the properties of the interconnection matrices A_{ij}. ◁

To perform isolation, $\mathcal{D}_i$ computes N_i estimates of $\bar{\alpha}_{ji}, j \in \mathcal{N}_i$, and then compares them to some appropriately defined thresholds $\theta_{ji}, j \in \mathcal{N}_i$. If the condition

$$|\hat{\bar{\alpha}}_{ji}(k|k+1)| > \theta_{ji}(k) \tag{15}$$

holds for some $j \in \mathcal{N}_i$, for at least one component of $|\hat{\bar{\alpha}}_{ji}(k|k+1)|$, the link (j,i) is said to be isolated. The threshold $\theta_{ji}(k)$ is defined by exploiting Assumption 1, asymptotic convergence of $\hat{\eta}_{i,2}^{(ij)}$ to $\eta_{i,2}^{(ij)}$, and the triangle inequality: by defining $\bar{\eta}_{i,2}^{(ij)}$ as the bound on the estimation

Algorithm 1: Attack isolation

for $k \geq K_i^D$ **do**
  Receive $\gamma^c_{ji,j\in\mathcal{N}_i}$ from neighbors $\mathcal{S}_j$, $j \in \mathcal{N}_i$;
  Update $\hat{x}_i(k)$, $\bar{r}_i(k)$, $\hat{x}_{ji}(k)$, $\bar{r}_{ji}(t), \forall j \in \mathcal{N}_i$;
  Evaluate (6) and (9)
  if $|r_{ji}(t)| > \bar{r}_{ji}(t)$ for at least one $j \in \mathcal{N}_i^c$ **then**
    Attack isolated to link (j, i)
  else
    for $j \in \mathcal{N}_i$ **do**
      Update $\hat{\bar{\alpha}}_{ji}(k-1|k)$ and $\theta_{ji}(k-1)$
      if $|\hat{\bar{\alpha}}_{ji}(k-1|k)| > \theta_{ji}(k-1)$ **then**
        Attack isolated to link (j, i)
      end if
    end for
  end if
end for

error $\tilde{\eta}^{(ij)}_{i,2} \doteq \eta^{(ij)}_{i,2} - \hat{\eta}^{(ij)}_{i,2}$ satisfying $|\tilde{\eta}^{(ij)}_{i,2}(k)| \leq \bar{\eta}^{(ij)}_{i,2}(k)$, θ_{ji} is computed as:

$$\begin{aligned}\theta_{ji}(k) = |\bar{A}^{(ij)-1}_{ij,1}| \Big(&|C^{(ij)-1}_{i,1}|\bar{\rho}^{(ij)}_{i,1}(k+1) \\ &+|A^{(ij)}_{Li,11}C^{(ij)-1}_{i,1}|\bar{\rho}^{(ij)}_{i,1}(k) + |A^{(ij)}_{Li,12}|\bar{\eta}^{(ij)}_{i,2}(k)\Big)\end{aligned} \tag{16}$$

with $\bar{\rho}^{(ij)}_i = S^{-1}_{ji}\bar{r}_i$. The error bound $\bar{\eta}^{(ij)}_{i,2}$, on the other hand, can be found through a procedure similar to that in [24].The overall isolation strategy is summarized in Algorithm 1.

Remark 5: Note that, for isolation purposes, the estimate $\hat{\bar{\alpha}}_{ji}$ is itself used as a residual, using it directly in the isolation test (15). Indeed, on the one hand α_{ji} is not directly available, and thus $\tilde{\alpha}_{ji} \doteq \bar{\alpha}_{ji} - \hat{\bar{\alpha}}_{ji}$ cannot be computed; on the other $\bar{\alpha}_{ji} = 0$ in nominal conditions, and therefore $\hat{\bar{\alpha}}_{ji}$ is suitable for isolation. ◁

C. Analysis of isolation properties

Having presented the isolation logic within $\mathcal{D}_i$, let us now present its properties. As hinted at, the isolation of an attack by $\mathcal{O}_i$ depends on the physical interconnection matrices A_{ij}, as it is the error $\xi_i - \hat{\xi}_i$ that leads to $|r_i(k)| > \bar{r}_i(k)$, and therefore detection.

Theorem 1: Consider a subsystem $\mathcal{S}_i$, equipped with a diagnoser $\mathcal{D}_i$. Suppose that, for any $j, k \in \mathcal{N}_i, j \neq k$, the physical coupling matrices are such that $\mathrm{Im}A_{ij}$ and $\mathrm{Im}A_{ik}$ are orthogonal. Then, a cyber attack $\alpha_{li} \neq 0$ may be isolated to the communication link (l, i), for $l \in \mathcal{N}_i$, by $\mathcal{D}_i$ implementing Algorithm 1. □

The result presented in Theorem 1 is a sufficient structural condition on the physical coupling between subsystems such that Algorithm 1 is an appropriate isolation strategy. In the following, we give an equivalent necessary condition.

Theorem 2: Consider a subsystem $\mathcal{S}_i$, equipped with diagnoser $\mathcal{D}_i$ defined in Section III. Suppose that from time K^a_{ij} an attack $\alpha_{ji} \neq 0$ is present on the communication link $(j, i) \in \mathcal{E}$. The attack may be isolated by $\mathcal{D}_i$ only if

$$\mathrm{Im}A_{ik} \neq \mathrm{Im}A_{il} \tag{17}$$

holds for all $k, l \in \mathcal{N}_i, k \neq l$. □

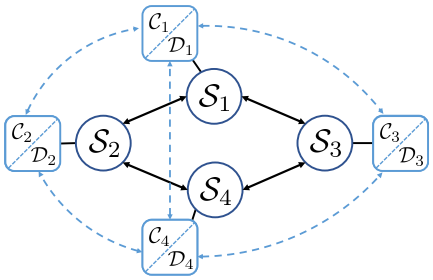

Fig. 1: Network of $N = 4$ physically coupled subsystems. The physical coupling are represented as the black solid arrows, are shown as blue dashed arrows.

V. Simulation results

Let us now show the effectiveness of the proposed method in simulation. Here, we limit ourselves to illustrating the detection and isolation properties of $\mathcal{O}_i$ within $\mathcal{D}_i$, as the isolation properties of $\mathcal{O}_{ij}$ are straightforward, and have been shown in [6]. We consider a network of $N = 4$ subsystems, with physical and communication coupling as in Figure 1. Each subsystem has dynamics as in (1), with matrices:

$$\begin{aligned} A_{ii} &\doteq \begin{bmatrix} a_{ii,11} & a_{ii,12} & 0 \\ a_{i,21} & a_{ii,22} & 0 \\ -1 & 0 & 1 \end{bmatrix} \quad B_i \doteq \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \\ A_{ij} &\doteq \begin{bmatrix} a_{ij,1} & 0 & 0 \\ 0 & a_{ij,2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad C_i \doteq I_3. \end{aligned} \tag{18}$$

where the values of A_{ii} are chosen randomly such that A_{ii} is Schur stable, and the coupling parameters can be found in Table I. Each subsystem is regulated by a control input $u_i = \begin{bmatrix} u_{i,1}, u_{i,2} \end{bmatrix}^\top$, where $u_{i,1} = K_i x_i$ is a decentralized input guaranteeing stability of the overall system through appropriate design of K_i, and $u_{i,2} = x^{ref}_{i,1}$ is a reference value for the first component of x_i. Specifically, $x^{ref}_{i,1} = u^{ref}_i + \delta_i$ where u^{ref}_i is a local reference, while δ_i is the result of the following consensus protocol $\delta^+_i = \delta_i + \sum_{j\in\mathcal{N}^c_i} a^c_{ij}\left(\gamma^c_{ji,1} - y_{i,1}\right)$. The state $x_{i,3}$ is an internal integrator state, necessary to ensure $x_{i,1}$ tracks $x^{ref}_{i,1}$. The process and measurement disturbances are assumed to be random processes with uniform distribution between $\pm\bar{w}_i$ and $\pm\bar{v}_i$, $\bar{w}_i \doteq \begin{bmatrix} 0.02 & 0.02 & 0 \end{bmatrix}^\top$, $\bar{v}_i \doteq \begin{bmatrix} 0.05 & 0.05 & 0 \end{bmatrix}^\top$, where the last element of both $\bar{w}_i$ and $\bar{v}_i$ are taken to be zero given that $x_{i,3}$ is an internal integrator state.We consider a covert attack to be present on the communication link $(3, 1)$ from time $t = 4\,\mathrm{s}$. Specifically, we assume the attacker is capable of simulating the exact dynamics of $\mathcal{S}_3$, changing the

TABLE I: Physical coupling

$a_{12,1}$	0.2376	$a_{12,2}$	0
$a_{13,1}$	0	$a_{13,2}$	0.3113
$a_{24,1}$	0	$a_{24,2}$	0.3031
$a_{34,1}$	0.3988	$a_{34,2}$	0

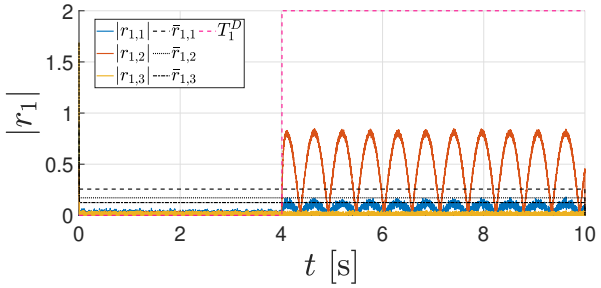

Fig. 2: Component-by-component comparison of the residual $|r_1|$ to the detection threshold $\bar{r}_1$. To visualize the detection time K_1^D, we add, in dashed pink, a flag which is "low" for $k < K_1^D$ and "high" for $k \geq K_1^D$.

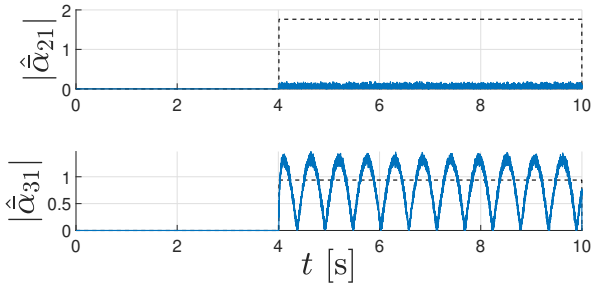

Fig. 3: Comparison of the estimates $\hat{\alpha}_{j1}, j \in \mathcal{N}_1$ to the isolation threshold $\theta_{j1}, j \in \mathcal{N}_1$. Once $|\hat{\alpha}_{31}| > \theta_{31}$, at time $t = 4.029$s, the attack is isolated to the link $(3, 1)$.

reference u_3^{ref} from a constant to a sinusoidal waveform. In Figure 2, we show how the detection module $\mathcal{O}_i$ described in Section III detects the attack by time $t = 4.016$s. After detection, $\mathcal{D}_1$ exploits the residual r_1 to estimate the attack, estimating $\hat{\alpha}_{ji}$ for $j \in \mathcal{N}_1 \doteq \{2, 3\}$, and computes the isolation threshold $\theta_{ji}(k)$, as defined in (16). Thus, as in Algorithm 1, the diagnoser $\mathcal{D}_i$ evaluates the detection test (15), isolating the attack to communication link $(3, 1)$ by time $t = 4.029$s, as shown in Figure 3. The fact that $\hat{\alpha}_{21} \neq 0$ is due to the effect of process and measurement noise. Note that $\mathcal{D}_1$ does not compute an estimate for $j = 4 \in \mathcal{N}_i^c \backslash \mathcal{N}_i$, as $\mathcal{O}_i$ cannot detect any attacks on $(4, 1)$, by construction.

VI. Conclusion

We consider the presence of cyber-attacks on the communication links between controllers in a large-scale interconnected system regulated via a distributed control architecture. We present a distributed cyber-attack isolation strategy based on unknown-input reconstruction to identify the communication link over which an attack is present. Furthermore, we provide a sufficient and a necessary condition that must be satisfied by the structure of the physical coupling between subsystems to guarantee isolability.

As future work, we will investigate how to mitigate the effect of cyber attacks in large-scale interconnected systems, either through attack accommodation or system reconfiguration, as well as evaluating more realistic assumptions on the communication network between distributed controllers.

References

[1] S. Zanero, "Cyber-physical systems," *Computer*, vol. 50, no. 4, pp. 14–16, 2017.

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

[3] B. Sobczak, *Denial of Service attack caused grid cyber disruption: DOE*. Environment & Energy Publishing, 2019.

[4] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems." in *HotSec*, 2008.

[5] S. Weerakkody and B. Sinopoli, "Challenges and opportunities: Cyber-physical security in the smart grid," in *Smart Grid Control*. Springer, 2019, pp. 257–273.

[6] A. J. Gallo, M. S. Turan, F. Boem, T. Parisini, and G. Ferrari-Trecate, "A distributed cyber-attack detection scheme with application to dc microgrids," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3800–3815, 2020.

[7] A. Barboni, H. Rezaee, F. Boem, and T. Parisini, "Detection of covert cyber-attacks in interconnected systems: A distributed model-based approach," *IEEE Transactions on Automatic Control*, 2020.

[8] S. Dibaji, M. Pirani, A. Annaswamy, K. Johansson, and A. Chakrabortty, "Secure control of wide-area power systems: Confidentiality and integrity threats," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 7269–7274.

[9] M. Deghat, V. Ugrinovskii, I. Shames, and C. Langbort, "Detection and mitigation of biasing attacks on distributed estimation networks," *Automatica*, vol. 99, pp. 369–381, 2019.

[10] R. Anguluri, V. Katewa, and F. Pasqualetti, "Centralized versus decentralized detection of attacks in stochastic interconnected systems," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3903–3910, 2020.

[11] 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.

[12] F. Boem, R. M. Ferrari, and T. Parisini, "Distributed fault detection and isolation of continuous-time non-linear systems," *European Journal of Control*, vol. 17, no. 5, pp. 603–620, 2011.

[13] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Distributed fault detection and isolation resilient to network model uncertainties," *IEEE Transactions on Cybernetics*, vol. 44, no. 11, pp. 2024–2037, 2014.

[14] V. Reppa, M. M. Polycarpou, and C. G. Panayiotou, "Decentralized isolation of multiple sensor faults in large-scale interconnected nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 60, no. 6, pp. 1582–1596, 2015.

[15] 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.

[16] M. Blanke, M. Kinnaert, J. Lunze, M. Staroswiecki, and J. Schröder, *Diagnosis and fault-tolerant control*. Springer, 2006, vol. 2.

[17] X. Zhang, M. M. Polycarpou, and T. Parisini, "A robust detection and isolation scheme for abrupt and incipient faults in nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 47, no. 4, pp. 576–593, 2002.

[18] 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.

[19] H. L. Trentelman, A. A. Stoorvogel, and M. Hautus, *Control theory for linear systems*. Springer Science & Business Media, 2012.

[20] 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.

[21] A. J. Gallo, A. Barboni, and T. Parisini, "On detectability of cyber-attacks for large-scale interconnected systems," in *21st IFAC World Congress*, 2020, pp. –.

[22] F. J. Bejarano, "Partial unknown input reconstruction for linear systems," *Automatica*, vol. 47, no. 8, pp. 1751–1756, 2011.

[23] M. Corless and J. Tu, "State and input estimation for a class of uncertain systems," *Automatica*, vol. 34, no. 6, pp. 757–764, 1998.

[24] F. Boem, A. J. Gallo, G. Ferrari-Trecate, and T. Parisini, "A distributed attack detection method for multi-agent systems governed by consensus-based control," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 5961–5966.