On detectability of cyber-attacks for large-scale interconnected systems [⋆]

Alexander J. Gallo [*] Angelo Barboni [*] Thomas Parisini [*,**]

[*] *Imperial College London, London, UK*
({alexander.gallo12,a.barboni16}@ic.ac.uk).
[**] *University of Trieste, Trieste, Italy, and KIOS Research and Innovation Center of Excellence, Nicosia, Cyprus*
(t.parisini@imperial.ac.uk)

Abstract: The paper deals with the key problem of detecting cyber-attacks in the context of large-scale systems (LSS). As these systems grow in size and complexity, cyber-attacks may target limited parts of the LSS, leading to tackling the problem in a decentralized way. We analyze the properties of distributed detection schemes under both local and interconnection attacks, and show that they are vulnerable to attacks that exploit the structure of the interconnections between subsystems. We also provide conditions and strategies that may be adopted to make the distributed control architecture regulating the LSS robust to this class of cyber-attacks.

Keywords: Secure networked control systems, distributed control and estimation, fault detection and diagnosis

1. INTRODUCTION

Many large-scale critical infrastructure and industrial control systems can be usefully described by the class of large-scale dynamic systems (Lunze, 1992), in turn modeled as the interconnection of smaller subsystems. Indeed, this partition into subsystems may be done either to satisfy certain implementation objectives (such as a reduction of computational or communication costs associated with a centralized controller design), or because the plant itself is distributed over a large area. This has lead over the past decades to formulate *distributed architectures* to operate LSS (Šiljak and Zečević, 2005), where each subsystem is equipped with a *local* controller.

In order to achieve global objectives, controllers can share information over a communication network, allowing for coordination. In addition, these architectures increasingly rely on the integration of networked sensors and other so-called *cyber* infrastructure, leading to an increased concern in their cyber-security (Cárdenas et al., 2008). It has been ascertained in the security literature (Teixeira et al. (2015); Dibaji et al. (2019) and references therein) that malicious attackers can detrimentally affect the performance of a system by injecting disruptive signals into the plant's actuator, sensor, or communication channels, as real instances of attacks have proven (see Lee et al. (2016); Sobczak (2019) for recent evidence). Given this risk, together with the critical nature of LSS, it is deemed necessary to ensure their safety and security. Specifically, implementation of distributed monitoring architectures has attracted recent attention in literature (Anguluri et al., 2018; Gallo et al., 2018; Barboni et al., 2019), for which the distributed fault detection and isolation literature has been foundational (see for example Shames et al. (2011); Blanke et al. (2016); Boem et al. (2017a) to name but a few).

In this paper, we address the distributed detection of attacks on the communication network between controllers, named *interconnection attacks*, and highlight the structural limitations caused by the interconnection between subsystems. Furthermore, we include *local* attacks in our analysis, as in certain cases they can be modeled as interconnection attacks. The property of detectability of interconnection attacks is shown to be related to that of input observability of neighbors' states through interconnections, which can be seen as input to a subsystem's dynamics. Comprehensive results on input observability can be found in Hou and Patton (1998), or more recently Boukhobza et al. (2009) tackled a similar problem in a distributed context from a graph theoretic point of view.

Other works relevant to detectability of attacks are Pasqualetti et al. (2013) and Weerakkody et al. (2017). In the cited works, attackability is studied from a centralized point of view, and properties are formulated in terms of left-invertibility, which may not hold for subsystems with weakly coupled interconnections in LSS. Hence, we offer the following contributions:

a. a generalized model of attacks which are local to one subsystem but target either the local control loop or the communication links;

[⋆] The support of the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1) is gratefully acknowledged. This work has been supported in part by European Unions Horizon 2020 Research and Innovation Programme under Grant 739551 (KIOS CoE), and in part by the Italian Ministry for Research in the framework of the 2017 Program for Research Projects of National Interest (PRIN) under Grant 2017YKXYXJ.

Copyright lies with the authors

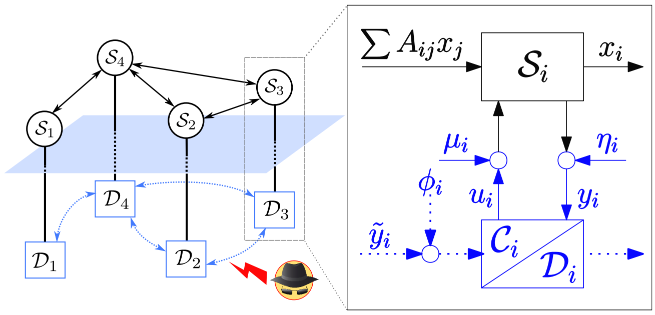

Fig. 1. **[On the left:]** schematic of a LSS partitioned into subsystems, with physical interconnections in black, and communication network links in blue. **[On the right:]** $\mathcal{S}_i$ with its control/monitoring layer; we follow the same color convention and show in detail the influence of attack signals. Dashed lines represent the communication network between controllers, while solid lines are local links.

b. analysis of the structural vulnerabilities of distributed attack detection schemes, following from model in a.;
c. as a consequence of the structural limitations of distributed architectures, a definition of robustness is given to the controller, and the following strategies proposed to achieve it:
   i. the design of a control law dependent only on the information verifiable by the monitoring scheme;
   ii. the augmentation of the local information in a way that makes it possible to detect an attack.

Notation

Given a matrix $A \in \mathbb{R}^{p\times n}$, $\operatorname{rank} A = p$, $A^{\dagger}$ is its the Moore-Penrose pseudo-inverse. $\mathbf{I}$ is the identity matrix of dimensions appropriate to the context. For a given matrix M, $[M]$ is used to define its related *structural matrix*, where specific entries are fixed at zeros, i.e., if $[M]_{(ij)} = 0$ then $M_{(ij)} = 0$ also, while all others are set as free parameters. For some matrix Q, $\ker Q^{\perp}$ denotes the space orthogonal to $\ker Q$, the null space defined by Q. This decomposition is *complete*, hence any vector $x = \bar{x} + x^{\perp}$, where $\bar{x} \in \ker Q$ and $x^{\perp} \in \ker Q^{\perp}$. Throughout the paper the superscripts a and h respectively indicate the attacked and nominal components of some signal x, i.e. in $x = x^h + x^a$, x^h denotes its healthy component and x^a the attacked one. With $\operatorname{col}[\cdot], \operatorname{row}[\cdot], \operatorname{diag}[\cdot]$ we intend column and row concatenation of vectors or matrices, and block-diagonal concatenation of matrices.

2. PROBLEM STATEMENT

2.1 Modeling Large-Scale Systems

We consider a large-scale system as a network of N subsystems $\mathcal{S}_i$, each of which is interconnected with a set $\mathcal{N}_i \subseteq \mathcal{N} \triangleq \{1,\dots,N\}$, called the set of *neighbors* of $\mathcal{S}_i$. The dynamics of the state of the LSS can be partitioned into N equations, each describing the dynamics of $\mathcal{S}_i$ as:

$$\begin{aligned} \dot{x}_i &= A_{ii}x_i + B_i u_i + \sum_{j\in\mathcal{N}_i} A_{ij}x_j \\ y_i &= C_i x_i, \end{aligned} \tag{1}$$

where $x_i \in \mathbb{R}^{n_i}$, $u_i \in \mathbb{R}^{m_i}$, and $y_i \in \mathbb{R}^{p_i}$ are respectively the state, the control input, and the output of $\mathcal{S}_i$. The term $\sum_{j\in\mathcal{N}_i} A_{ij}x_j$ models the interconnection between different subsystems. Matrices A_{ii}, B_i, A_{ij}, and C_i are of appropriate dimensions. Furthermore, without loss of generality, $\operatorname{rank} C_i = p_i$.

Assumption 1. The pair (A_{ii}, C_i), $\forall i \in \mathcal{N}$ is observable. ◁

Assumption 2. For some full column rank $\tilde{A}_{ij}$ such that $\operatorname{Im}(\tilde{A}_{ij}) = \operatorname{Im}(\operatorname{row}[A]_{ij\in\mathcal{N}_i})$, there are no zero dynamics from $\tilde{A}_{ij}$ to C_i. ◁

Remark 3. In this paper we focus on the structural properties of a distributed diagnostic tool. As such no uncertainties are included in (1). ◁

We consider that controllers $\mathcal{C}_i$ are designed to implement a distributed control algorithm, for which $u_i, i \in \mathcal{N}$ is computed both from locally measured outputs y_i, and information received from neighboring controllers $\mathcal{C}_j, j \in \mathcal{N}_i$. In order to implement such an algorithm, it is necessary to introduce a peer-to-peer communication network between the controllers, as represented in Figure 1.

Assumption 4. The topology of the network of physical interconnections between subsystems is mirrored by that of the communication network, and the networks are undirected. ◁

We consider two scenarios:

a) Subsystems share their local measurements y_i.
b) Subsystems share locally computed state estimates $\hat{x}_i$.

The state estimate $\hat{x}_i$ of x_i is such that the estimation error $\epsilon_i \triangleq x_i - \hat{x}_i \to 0$ asymptotically as $t \to \infty$. How to compute such an estimate is out of the scope this paper. We introduce the following vector to model the information *received* by $\mathcal{C}_i$:

$$\tilde{y}_i \triangleq \tilde{C}_i \tilde{x}_i + \tilde{v}_i, \tag{2}$$

where $\tilde{x}_i \triangleq \operatorname{col}[x]_{j\in\mathcal{N}_i}$ is a vector grouping all the states of the neighboring subsystems $\mathcal{S}_j, j \in \mathcal{N}_i$, acting as an external input to $\mathcal{S}_i$. Notation (2) allows to consider the two information exchange scenarios in a unified way. Indeed, defining $\tilde{C}_i \triangleq \operatorname{diag}[\breve{C}]_{j\in\mathcal{N}_i}$, where in case a), $\breve{C}_j \triangleq C_j$ and $\tilde{v}_i \triangleq 0$, while for b), $\breve{C}_j \triangleq \mathbf{I}$ and $\tilde{v}_i \triangleq \operatorname{col}[\epsilon]_{j\in\mathcal{N}_i}$. Note that $\tilde{C}_i$ is block-diagonal. In the following, with some abuse of notation, we refer to its j-th block as $\tilde{C}_{i,(j)}$.

Assumption 5. If the following condition does not hold:

$$\ker C_j \subseteq \ker A_{ij}, \forall i \in \mathcal{N}, j \in \mathcal{N}_i, \tag{3}$$

controllers $\mathcal{C}_i, i \in \mathcal{N}$ transmit their local estimate $\hat{x}_i$. ◁

Remark 6. We note here that Assumption 5 is required to allow for the possibility of obtaining a local estimate $\hat{x}_i$ such that $\epsilon_i \to 0$ as $t \to \infty$ without making the estimate independent of the interconnections. ◁

To simplify notation, from (2) we can define an augmented output vector $\mathbf{y}_i$, containing both local measurements and transmitted signals:

$$\begin{aligned} \mathbf{y}_i &= \begin{bmatrix} C_i & 0 \\ 0 & \tilde{C}_i \end{bmatrix} \begin{bmatrix} x_i \\ \tilde{x}_i \end{bmatrix} + \begin{bmatrix} 0 \\ \tilde{v}_i \end{bmatrix} \\ &= \mathbf{C}_i \mathbf{x}_i + \mathbf{v}_i. \end{aligned} \tag{4}$$

2.2 Modeling cyber-attacks

Definition 7. (Local attack). Attacks between the controller $\mathcal{C}_i$ and plant $\mathcal{S}_i$. These attacks alter the control input signals u_i and the output measurements y_i. ◁

Definition 8. (Interconnection attack). Attacks on information transmitted between controllers $\mathcal{C}_j$, or some subset of the neighboring subsystems $j \in \widehat{\mathcal{N}_i} \subseteq \mathcal{N}_i$, and $\mathcal{C}_i$. The attacker alters the signals contained in $\tilde{y}_i$. ◁

These two classes of attacks presuppose different architectures, and different attack resources. Specifically, *local* attacks presume that the control loop of $\mathcal{S}_i$ is closed over a communication network, and the attacker is capable of corrupting the signals exchanged between $\mathcal{C}_i$ and the networked sensors/actuators. On the other hand, in *interconnection* attacks, it is the communication network over which the controllers exchange information that is compromised. In Figure 1 we show how these different classes of attacks may affect the system. Note that we do not consider the possibility of attacks on the physical interconnections of the LSS.

Following the framework described in Teixeira et al. (2015), we model cyber-attacks in communication networks as additive signals. Specifically, the state dynamics, measurement and received communication outputs in (1) and (2) are defined as:

$$\begin{aligned} \dot{x}_i &= A_{ii}x_i + B_i u_i + \sum_{j\in\mathcal{N}_i} A_{ij}x_j + \mathbf{B}_i^a\mathbf{u}_i^a \\ \mathbf{y}_i &= \mathbf{C}_i\mathbf{x}_i + \mathbf{v}_i + \mathbf{D}_i^a\mathbf{u}_i^a, \end{aligned} \tag{5}$$

where

$$\mathbf{u}_i^a \triangleq \left[\, \phi_i^\top \;\; \mu_i^\top \;\; \eta_i^\top \,\right]^\top, \qquad \mathbf{y}_i \triangleq \left[\, y_i^\top \;\; \tilde{y}_i^\top \,\right]^\top,$$
$$\mathbf{B}_i^a \triangleq [\, 0 \;\; B_i^a \;\; 0 \,], \qquad \mathbf{D}_i^a \triangleq \begin{bmatrix} 0 & 0 & D_i^a \\ \mathbf{I} & 0 & 0 \end{bmatrix}.$$

Moreover ϕ_i models an interconnection attack on $\tilde{y}_i$ received by $\mathcal{C}_i$, μ_i a local attack on the actuators, and η_i one on the sensor measurements, as represented in Figure 1. We consider both B_i^a and D_i^a to be such that $\mathrm{Im}(B_i^a) \subseteq \mathrm{Im}(B_i)$ and $\mathrm{Im}(D_i^a) \subseteq \mathrm{Im}(C_i)$. Note that in (5) we have combined the modeling framework necessary to implement either *local* or *interconnection* attacks. Appropriate definition of $\mathbf{u}_i^a$ allows for modeling a number of different attack scenarios (Teixeira et al. (2015)). We denote the starting time of the attack as T_a.

Assumption 9. For $t \geq T_a$, an attack input function $\mathbf{u}_i^a \neq 0$, for a single $i \in \mathcal{N}$. Furthermore, it is defined either as

$$\mathbf{u}_i^a \triangleq \left[\phi_i^\top, 0^\top, 0^\top\right]^\top$$

for *interconnection* attacks, or as

$$\mathbf{u}_i^a \triangleq \left[0^\top, \mu_i^\top, \eta_i^\top\right]^\top$$

for *local* attacks. ◁

Remark 10. Assumption 9 implies that at any given time either (u_i, y_i) or $\tilde{y}_i$ can be *trusted*, i.e. are not corrupted by a malicious attacker. ◁

2.3 Distributed attack detection architectures

Given the possible vulnerability of the LSS's communication networks, it is necessary to equip it with a monitoring layer. Specifically, as shown in Figure 1, we consider the addition of local model-based diagnostic modules $\mathcal{D}_i$ capable of detecting local and interconnection attacks.

We choose to consider distributed architectures, as they are capable of integrating properties of scalability, low computational complexity, and privacy amongst subsystems within the diagnosis layer of the LSS. As such, we are interested in the analysis of its properties, and specifically of those limitations that arise following the partition of the LSS into subsystems, while having constraints on the information available to each local diagnoser. Let us define the set of information used in the design of $\mathcal{D}_i$:

Definition 11. (Information set). The *information set* $\mathcal{I}_i \triangleq \{\mathcal{M}_i, \gamma_i, \upsilon_i\}$ represents the information available to the diagnoser $\mathcal{D}_i$, where $\mathcal{M}_i$ is a tuple defining the dynamics, and γ_i and υ_i are output and input signals. ◁

Definition 12. (Reduced-information distributed architecture). We define an architecture as "reduced-information" if the information set of $\mathcal{D}_i$ is

$$\mathcal{I}_i \triangleq \{(A_{ii}, B_i, A_{ij\in\mathcal{N}_i}, \mathbf{C}_i), \mathbf{y}_i, u_i\}. \tag{6}$$

When $\mathcal{D}_i$ is designed according to this definition, it is said to be "reduced-information" itself. ◁

As $\mathcal{D}_i$ does not have direct access to x_i and $x_{j\in\mathcal{N}_i}$, to achieve detection the diagnoser defines an appropriate residual r_i given $\mathcal{I}_i$. This residual is such that in nominal conditions $r_i \to 0$ as $t \to \infty$, while if under attack this is detected once r_i^a is larger than some suitably chosen threshold.

Differently from other works in literature, we are not here focused on the detectability properties of the diagnoser with respect to local attacks. Rather, we are interested in the analysis of stealthiness of interconnection attacks with respect to the structure of the partitioned LSS. To clarify the explanation throughout the remainder of the paper, let us define the following classes of attacks [1]:

Definition 13. (Stealthy attacks). An attack is *stealthy* if no monitor detects the attack, i.e. for an attack on $\mathcal{S}_k$, $\mathbf{u}_k^a \neq 0$, $r_i^a = 0$, for any $k \in \mathcal{N}$ and all $i \in \mathcal{N}$. ◁

Definition 14. (Locally stealthy attack). An attack is *locally stealthy* if $[\mu_i^\top, \eta_i^\top]^\top \neq 0$ implies $r_i^a = 0$. ◁

Definition 15. (Structurally stealthy attacks). An interconnection attack on $\mathcal{S}_i$ is said to be *structurally stealthy* to $\mathcal{D}_i$ if, given $\mathcal{I}_i$, $\phi_i \neq 0$ leads to $r_i^a = 0$. ◁

3. DETECTABILITY LIMITATIONS OF REDUCED-INFORMATION $\mathcal{D}_i$

In this section, we highlight the limitations of $\mathcal{D}_i$ with respect to the partitioned structure of the LSS, given a reduced-information distributed architecture. We focus our attention to those attacks that are not locally detectable, namely, interconnection attacks and locally stealthy attacks. The first are analyzed as they influence $\tilde{y}_i$ as *received* by $\mathcal{C}_i$, not visible to $\mathcal{D}_j$, $j \in \mathcal{N}_i$, which are transmitting the information. Of the latter, we focus on local covert attacks, inspired by covert attacks in Smith (2015), introduced in a distributed scenario by Barboni

[1] The classes of attacks defined here can be seen as subsets of *undetectable attacks* in Pasqualetti et al. (2013).

et al. (2019), and shown to be locally stealthy to $\mathcal{D}_i$, as summarized in the following Lemma:

Lemma 16. (Barboni et al. (2019)). Given a local covert attack active for time $t \geq T_a$, where μ_i is freely chosen by the attacker to achieve some objective, and η_i is defined as the output of the following linear time invariant system with initial condition $x_j^a(T_a) = 0$:

$$\begin{aligned} \dot{x}_i^a &= A_{ii} x_i^a + B_i^a \mu_i \\ \eta_i &= -C_i x_i^a \end{aligned} , \tag{7}$$

a diagnostic module $\mathcal{D}_i$ is not capable of detecting it. □

Although Lemma 16 holds for all local diagnosers $\mathcal{D}_i$, in Barboni et al. (2019) it was shown that, through suitable design of the residual generator, it is possible for diagnosers of neighboring subsystems $\mathcal{D}_j$ to detect local covert attacks. As such, they share similar properties to interconnection attacks.

Proposition 17. Given a local covert attack on $\mathcal{S}_j$ such that Lemma 16 holds, it is equivalent to $\mu_j = \eta_j = 0$ and an interconnection attack

$$\phi_{i,(j)} = -\breve{C}_j x_j^a \tag{8}$$

for all $\mathcal{D}_i, i \in \mathcal{N}_j$. □

In (8), with abuse of notation, $\phi_{i,(j)}$ are the components of ϕ_i corresponding to the signal received by $\mathcal{C}_i$ from $\mathcal{C}_j$.

Remark 18. Although in this paper we limit our analysis to local covert attacks, for both space restrictions and ease of explanation, the method exploited in Proposition 17 can be applied to any locally stealthy attack. ◁

As such in the remainder, we consider that for some unknown $j \in \mathcal{N}_i$ for $t \geq T_a$, $\phi_{i,(j)} \neq 0$, modeling both interconnection and local covert attacks on $\mathcal{S}_j$. We now introduce the main result of this paper.

Theorem 19. Given a subsystem $\mathcal{S}_i$ with dynamics as in (5), equipped with a detector $\mathcal{D}_i$ designed with information set $\mathcal{I}_i$, suppose that, starting at some time $T_a > 0$, the subsystem is exposed to an interconnection attack $\phi_{i,(j)} \neq 0$, for some $j \in \mathcal{N}_i$. Such an attack is *structurally stealthy* to local diagnoser $\mathcal{D}_i$ if and only if

$$\breve{C}_j^\dagger \phi_{i,(j)} \in \ker A_{ij}. \tag{9}$$

□

Proof. [Sufficiency]: In order to detect an attack, $\mathcal{D}_i$ must generate a residual such that, given an attack on $\tilde{y}_{i,(j)}$, if $\breve{C}_j x_j \neq \tilde{y}_{i,(j)}$, then $r_i \not\to 0$.

As the control input u_i is known to $\mathcal{D}_i$ (given $\mathcal{I}_i$ in (6)) and as it cannot be altered maliciously (Assumption 9), its effect on y_i will be disregarded. An appropriate estimator can be defined such that, nominally, $r_i \to 0$ is satisfied. Consider the map from x_j to y_i:

$$y_i(t) = C_i \int_0^t e^{A_{ii}(t-\tau)} A_{ij} x_j(\tau) d\tau.$$

Given that x_j and $\mathcal{M}_j$ are unknown to $\mathcal{D}_i$, the neighbor's state must be estimated (statically) in order for $r_i \to 0$ to be satisfied in nominal conditions: such a static estimate is $\hat{x}_j(\tilde{y}_{i,(j)}) = \breve{C}_j^\dagger \tilde{y}_{i,(j)}$. Since $\operatorname{rank} \breve{C}_j = p_i \leq n_i$, this introduces error $\varepsilon_{i,(j)}(\tilde{y}_{i,(j)}) \triangleq x_j - \breve{C}_j^\dagger \tilde{y}_{i,(j)} \neq 0$. However, this does not affect the residual generator in nominal behavior, as

$$\left(\mathbf{I} - \breve{C}_j^\dagger \breve{C}_j\right) x_j \in \ker \breve{C}_j \subseteq \ker A_{ij},$$

given Assumption 5, and therefore $r_i \to 0$, even though $\varepsilon_{i,(j)} \neq 0$. Hence, $\hat{x}_j(\tilde{y}_{i,(j)})$ is an appropriate estimate of x_j. Following the results on *input observability* in Hou and Patton (1998) any component $\bar{x}_j \neq 0$, $\bar{x}_j \in \ker(A_{ij})$ does not affect y_i, as x_j is in fact not input observable from y_i. To prove sufficiency, suppose that $\phi_{i,(j)}$ is designed by the attacker such that (9) holds. Due to linearity, the estimated output affected by the attack, $\hat{y}_i^a$, is:

$$\hat{y}_i^a(t) = C_i \int_{T_a}^t e^{A_{ii}(t-\tau)} A_{ij} \breve{C}_j^\dagger \phi_{i,(j)}(\tau) d\tau = 0. \tag{10}$$

Hence, for $t \geq T_a$, $r_i^a = 0$, as $\hat{y}_i^a = 0, \forall t \geq T_a$, and as such the attack is stealthy.

[Necessity]: We prove necessity by contradiction. Let us suppose that $\breve{C}_j^\dagger \phi_{i,(j)} \notin \ker A_{ij}$, but that the attack is indeed structurally stealthy. Decompose $\breve{C}_j^\dagger \phi_{i,(j)}$ into $\breve{C}_j^\dagger \bar{\phi}_{i,(j)}$ and $\breve{C}_j^\dagger \phi_{i,(j)}^\perp$ such that:

$$\breve{C}_j^\dagger \phi_{i,(j)} = \breve{C}_j^\dagger \bar{\phi}_{i,(j)} + \breve{C}_j^\dagger \phi_{i,(j)}^\perp,$$

where $\breve{C}_j^\dagger \bar{\phi}_{i,(j)} \in \ker A_{ij}$ and $\breve{C}_j^\dagger \phi_{i,(j)}^\perp \in \ker A_{ij}^\perp$. The residual generator comparing y_i and $\hat{y}_i(\tilde{y}_{i,(j)})$, as $t \to \infty$:

$$r_i(t) \to C_i \int_{T_a}^t e^{A_{ii}(t-\tau)} A_{ij} \breve{C}_j^\dagger \phi_{i,(j)}^\perp(\tau) d\tau.$$

Given that $\breve{C}_j^\dagger$ is full column rank by design, and that condition (9) is assumed not to hold, $\breve{C}_j^\dagger \phi_{i,(j)}^\perp \neq 0$. Furthermore, following Assumption 2, there are no zero dynamics between $\tilde{A}_{ij}$ and C_i. Hence, $r_i^a \neq 0$, which contradicts the hypothesis. ■

The result of Theorem 19 shows some similarities with respect to the class of zero-dynamics attacks in literature. Indeed, as $\tilde{x}_i$ can be considered as an exogenous input to $\mathcal{S}_i$, ϕ_i can in turn be seen as an attack on the measurements of the inputs, and the class of attacks defined in Theorem 19 can be interpreted as a subset of zero-dynamics attacks. However, rather significantly, the attacks satisfying the stealthiness condition in Theorem 19 exploit the fact that A_{ij} is not full column rank, rather than requiring knowledge of the dynamics of $\mathcal{S}_j$ to exploit its transmission zeros. The above result is also relevant because attacks in this scenario are acting upon $\mathcal{S}_i$ through the computation of u_i by $\mathcal{C}_i$, which proves fundamental when considering the design of control architectures robust to such attacks.

If we consider an attacker with knowledge of the structure of the interconnection between subsystems, rather than A_{ij} itself, a special case of Theorem 19 is derived. Such a scenario is relevant as not only said structure might be easier to obtain, but also because, in fact, subsystems in LSS are often described as being weakly coupled (Lunze, 1992). As a consequence, the interconnection matrices are often not structurally full rank, and the following holds.

Corollary 20. If $\operatorname{rank}[A_{ij}] < n_j$, it is sufficient that

$$\breve{C}_j^\dagger \phi_{i,(j)} \in \ker [A_{ij}] \tag{11}$$

be satisfied for attack to be structurally stealthy. □

Note furthermore that, with the appropriate model knowledge and disruption resources, it would be possible for an interconnection attack to satisfy condition (9) even if $\mathrm{rank}\, A_{ij} = n_j$. Indeed, given an attack on a number of interconnections such that $\sum_{j\in\widehat{\mathcal{N}}_i} n_j > n_i$ is satisfied, and given knowledge of matrices A_{ij}, for $j \in \widehat{\mathcal{N}}_i$, an attacker would be able to satisfy Theorem 19. A similar result was presented in Boem et al. (2017b), where the properties of a specific distributed detection architecture were analyzed.

4. CONTROLLER ROBUSTNESS TO STRUCTURALLY STEALTHY ATTACKS

In this section, we are interested in analyzing the interaction between structurally stealthy attacks and distributed control architectures. By affecting only the received measurements $\tilde{y}_i$, interconnection attacks ϕ_i do not have a direct impact on the dynamics of $\mathcal{S}_i$. However, through appropriate definition of ϕ_i, the control signals u_i may be altered, thus maneuvering the state of $\mathcal{S}_i$ to an operating point undesired by the operator. It is therefore important to define control laws such that u_i is *robust* to structurally stealthy interconnection attacks.

Definition 21. Consider a LSS with an interconnection attack ϕ_i that satisfies Theorem 19, for some $\phi_{i,(j)}$, $j \in \mathcal{N}_i$. Let $\mathcal{C}_i$ compute $u_i \triangleq \kappa_i(\mathbf{y}_i)$, for some operator κ_i [2]. We say that the control architecture is robust if $\forall i \in \mathcal{N}$,

$$\kappa_i\left(\left[\,0^\top\ \phi_i^\top\,\right]^\top\right) = 0. \qquad (12)$$

for all attacks $\phi_{i,(j)} \neq 0$ stealthy to $\mathcal{D}_i$. ◁

In order to ensure control architecture $\mathcal{C}_i$ is robust, there are two strategies that can be pursued:

- ensure $\kappa_i([0^\top, \tilde{y}_i^\top]^\top) = 0$ for all $\tilde{y}_i$ such that $\breve{C}_j^\dagger \tilde{y}_{i,(j)} \in \ker A_{ij}$;
- augment information set $\mathcal{I}_i$ so that the class of stealthy interconnection attacks is reduced.

This section is dedicated to the former, while Section 5 will address the latter. Before proceeding, we analyze the special case in which $\phi_{i,(j)}$ models a local covert attack on $\mathcal{S}_j$. Indeed, although in Proposition 17 we have shown that they can be interpreted as interconnection attacks, they affect the state of $\mathcal{S}_j$ directly, and could therefore have an impact on the behavior of the LSS in two ways: i. by propagating the "attacked state" to other subsystems; ii. by having neighbors' controllers react to behavior that is not nominal. We address these cases separately.

Proposition 22. If $\mathcal{S}_j$ is subject to a local covert attack, controllers $\mathcal{C}_i$, $i \in \mathcal{N}$, are robust to it. □

Proof. In Proposition 17, we assumed the state of a subsystem under a covert attack to be nominal. This can be seen as the definition of a new nominal state and a shift of the attack action entirely onto the information transmission between subsystems. With respect to the actual nominal state from the perspective of $\mathcal{C}_j$, however, we have that $\tilde{y}_{i,(j)} = y_j^h$, which in fact satisfies (12), as $\phi_{i,(j)} = 0$. ■

Proposition 23. If $\mathcal{S}_j$ is subject to a local covert attack satisfying (9), the large-scale system composed of subsystems $\mathcal{S}_k$, $k \in \mathcal{N}\backslash\{j\}$ is not influenced by the attack. □

Proof. Let $\mathbf{x} \triangleq \mathrm{col}[x]_{k\in\mathcal{N}\backslash\{j\}}, \mathbf{y} \triangleq \mathrm{col}[y]_{k\in\mathcal{N}\backslash\{j\}}, \mathbf{u} \triangleq \mathrm{col}[u]_{k\in\mathcal{N}\backslash\{j\}}, \mathbf{u}^a \triangleq \mathrm{col}[u^a]_{k\in\mathcal{N}\backslash\{j\}}$. Hence, the dynamics of the large-scale system as a whole is, from (5):

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{A}_{kj}x_j + \mathbf{B}^a\mathbf{u}^a \\ \mathbf{y} &= \mathbf{C}\mathbf{x},\end{aligned} \qquad (13)$$

where $\mathbf{A}$ is a block-matrix where each ik-th entry is A_{ik}, $\mathbf{B} = \mathrm{diag}[B]_{j\in\mathcal{N}\backslash\{j\}}$, $\mathbf{B}^a = \mathrm{diag}[B^a]_{j\in\mathcal{N}\backslash\{j\}}$, $\mathbf{C} = \mathrm{diag}[C]_{j\in\mathcal{N}\backslash\{j\}}$, and $\mathbf{A}_{kj}$ groups the interconnection terms A_{ij} for all $\mathcal{S}_i$ such that $j \in \mathcal{N}_i$. We decompose $\mathbf{x} = \mathbf{x}^h + \mathbf{x}^a$, where $\mathbf{x}^a$ is the state affected by all attacks on the LSS. From Proposition 22, $\mathbf{u}^a = 0$, as for all $i \in \mathcal{N}_j$, $\tilde{y}_{i,(j)} = y_j^h$. Hence, to prove $\mathbf{x}^a = 0$, i.e. that the state of the LSS once $\mathcal{S}_j$ is subtracted is not affected by the attack, we must show that $\mathbf{A}_{kj}x_j^a = 0$. Recall condition (9)

$$\breve{C}_j^\dagger \phi_{i,(j)} \in \ker A_{ij},$$

and that $\phi_{i,(j)} = -\breve{C}_j x_j^a$ from Proposition 17. Therefore $\breve{C}_j^\dagger \breve{C}_j x_j^a \in \ker A_{ij}$, implying that $x_j^a \in \ker A_{ij}, \forall i \in \mathcal{N}_j$. Hence, $x_j^a \in \ker \mathbf{A}_{kj}$. As such, x_j^a will not influence the dynamics of any of its neighbors. ■

Having established that local covert attacks that are structurally stealthy do not propagate to the rest of the LSS, we evaluate the case in which there is an interconnection attack affecting $\tilde{y}_i$. Consequently, the control inputs computed by $\mathcal{C}_i$ should depend only on measurements $\tilde{y}_{i,(j)}$ associated to those components of x_j which directly influence $\mathcal{S}_i$ through physical interconnection. This translates into ensuring that the arguments of the control function lie in $\ker(A_{ij}\breve{C}^\dagger)^\perp$, as shown in the next result. Specifically, we introduce matrix $\Gamma_i \triangleq \mathrm{diag}[\Gamma]_{ij\in\mathcal{N}_i}$, with Γ_{ij} suitably designed for each $j \in \mathcal{N}_i$ so that the following holds:

Proposition 24. If each controller is such that

$$u_i = \kappa_i([y_i^\top, (\Gamma_i\tilde{y}_i)^\top]^\top), \qquad (14)$$

with each Γ_{ij} designed such that

$$\mathrm{Im}\Gamma_{ij} \cap \ker A_{ij}\breve{C}_j^\dagger = \emptyset, \qquad (15)$$

$\mathcal{C}_i$ is robust to all structurally stealthy attacks. □

Proof. By setting u_i as in (15), it is clear that it is no longer required to detect $\phi_{i,(j)}$, but rather we are restricting the sensitivity of our residual to $\Gamma_{ij}\phi_{i,(j)}$. As such, rather than (9), we are interested in verifying whether an attack function $\phi_{i,(j)}$ such that

$$\breve{C}_j^\dagger \Gamma_{ij}\phi_{i,(j)} \in \ker A_{ij}$$

holds. By setting Γ_{ij} such that (15) holds, no attack function such that $\Gamma_{ij}\phi_{i,(j)} \neq 0$ exists such that

$$A_{ij}\breve{C}_j^\dagger \Gamma_{ij}\phi_{i,(j)} = 0.$$

As such, any attack function such that $\kappa_i([0^\top, \Gamma_i\phi_i]^\top) \neq 0$ will not be stealthy to $\mathcal{D}_i$, and controllers are therefore robust to structurally stealthy attacks. ■

A possible definition of Γ_{ij} such that Proposition 24 holds is to set it as a basis of $\ker(A_{ij}\breve{C}_j^\dagger)^\perp$. This is indeed quite conservative, requiring $\mathcal{C}_i$ to only exploit measurements of neighbors' state components physically coupled with $\mathcal{S}_i$.

[2] With $\kappa_i(\cdot)$ we do not limit the control law to any specific class. Rather, we intend to show dependence of u_i on $\mathbf{y}_i$.

5. AUGMENTING $\mathcal{I}_i$ FOR DETECTABILITY OF $\mathcal{D}_i$

A robust control law as presented in the previous section might be infeasible or in contrast with design constraints. As a result, in this section, we show how the addition of information on dynamics of neighboring subsystems $\mathcal{S}_j, j \in \mathcal{N}_i$ to $\mathcal{I}_i$ may assist with the design of robust controllers. Indeed, under certain conditions, this permits the estimation of x_j directly, and to not rely on the input observability of $\tilde{x}_i$.

To illustrate the idea, we start by assuming that all of the model information $\mathcal{M}_j, j \in \mathcal{N}_i$ is available to $\mathcal{D}_i$, showing how a residual generator may be capable of detecting structurally stealthy interconnection attacks. Given Assumption 1, it is possible to augment residual generator $\mathcal{D}_i$ to include additional knowledge of the dynamics of $\mathcal{S}_j$.

By extending $\mathcal{D}_i$ to detect additive attacks on $\tilde{y}_{i,(j)}$, attackers must satisfy further constraints in order to be stealthy. Specifically, the following holds:

Proposition 25. Attack $\phi_{i,(j)}$ is stealthy to $\mathcal{D}_i$ if and only if it is both structurally stealthy given A_{ij}, and can be seen as a zero-dynamics attack on $\mathcal{S}_j$ through $\tilde{y}_{i,(j)}$. □

Proof. If $\mathcal{D}_i$ is composed of r_i and $r_{i,(j)}, j \in \mathcal{N}_i$, it holds that for an attack $\phi_{i,(j)} \neq 0$ must satisfy both

$$r_i^a = 0,$$
$$r_{i,(j)}^a = 0,$$

as if either of these do not hold $\mathcal{D}_i$ will detect an attack. Following Theorem 19, $\phi_{i,(j)}$ must be a structurally stealthy attack for $r_i \to 0$.

As attack signal $\phi_{i,(j)}$ is an additive signal to $\tilde{y}_{i,(j)} = y_j$, being a zero-dynamic attack is known to be a necessary and sufficient condition for it to be stealthy to a residual generator exploiting $\mathcal{M}_j$ (Pasqualetti et al., 2013). ■

In general, there are conditions on the dynamics of $\mathcal{S}_j$ that must be satisfied for $\mathcal{D}_i$ to compute convergent estimates. While out of the scope of this paper to discuss such conditions in a general way, for completeness we refer the reader to Gallo et al. (2018). In that work, distributed diagnosers exploiting model knowledge of neighbors are designed to detect interconnection attacks in islanded DC microgrids.

6. CONCLUSIONS

Analysis of stealthy attacks in partitioned LSS have been presented, within the framework of interconnection attacks. We have derived conditions under which such attacks are not detectable, and have shown how controllers can be made robust to them, thus preventing the impact to be propagated to the entire LSS. Detectability, furthermore, can be recovered by properly extending the information set locally available to each diagnoser. Future studies will be focused on defining such a set so that it is minimal while ensuring estimation convergence.

REFERENCES

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

Barboni, A., Rezaee, H., Boem, F., and Parisini, T. (2019). Distributed detection of covert attacks for interconnected systems. In *18th European Control Conference (ECC)*, 2240–2245. IEEE.

Blanke, M., Kinnaert, M., Lunze, J., and Staroswiecki, M. (2016). *Distributed Fault Diagnosis and Fault-Tolerant Control*. Springer.

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

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

Boukhobza, T., Hamelin, F., Martinez-Martinez, S., and Sauter, D. (2009). Structural analysis of the partial state and input observability for structured linear systems: Application to distributed systems. *European Journal of Control*, 15(5), 503–516.

Cárdenas, A.A., Amin, S., and Sastry, S. (2008). Research challenges for the security of control systems. In *HotSec*.

Dibaji, S.M., Pirani, M., Flamholz, D.B., Annaswamy, A.M., Johansson, K.H., and Chakrabortty, A. (2019). A systems and control perspective of CPS security. *Annual Reviews in Control*, 47, 394–411.

Gallo, A.J., Turan, M.S., Nahata, P., Boem, F., Parisini, T., and Ferrari-Trecate, G. (2018). Distributed cyber-attack detection in the secondary control of DC microgrids. In *17th European Control Conference (ECC)*, 344–349.

Hou, M. and Patton, R.J. (1998). Input observability and input reconstruction. *Automatica*, 34(6), 789–794.

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

Lunze, J. (1992). *Feedback control of large-scale systems*. Prentice Hall.

Pasqualetti, F., Dörfler, F., and Bullo, F. (2013). Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, 58(11), 2715–2729.

Shames, I., Teixeira, A.M., Sandberg, H., and Johansson, K.H. (2011). Distributed fault detection for interconnected second-order systems. *Automatica*, 47(12), 2757–2764.

Šiljak, D. and Zečević, A. (2005). Control of large-scale systems: Beyond decentralized feedback. *Annual Reviews in Control*, 29(2), 169–179.

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

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

Teixeira, A., Shames, I., Sandberg, H., and Johansson, K.H. (2015). A secure control framework for resource-limited adversaries. *Automatica*, 51, 135–148.

Weerakkody, S., Liu, X., and Sinopoli, B. (2017). Robust structural analysis and design of distributed control systems to prevent zero dynamics attacks. In *56th Conference on Decision and Control (CDC)*, 1356–1361.