A Distributed Attack Detection Method for Multi-Agent Systems Governed by Consensus-Based Control

Francesca Boem, Alexander J. Gallo, Giancarlo Ferrari-Trecate, Thomas Parisini

Abstract— **The paper considers the problem of detecting cyber-attacks occurring in communication networks for distributed control schemes. A distributed methodology is proposed to detect the presence of malicious attacks aimed at compromising the stability of large-scale interconnected systems and multi-agent systems governed by consensus-based controllers. Only knowledge of the local model is required. The detectability properties of the proposed method are analyzed. A class of undetectable attacks is identified. Preliminary simulation results show the effectiveness of the proposed approach.**

I. Introduction

In modern engineering systems, physical processes, computational resources and communication networks are integrated. Examples of systems of this kind, often called Cyber-Physical Systems [1], range from Large-Scale Systems (LSS) such as critical infrastructure systems, including electrical grids, water distribution systems, traffic networks, industrial control systems, etc., to autonomous vehicles and unmanned aerial vehicles (UAVs), which are described in the literature as multi-agent systems [2]. A common feature of these systems is that they can be described as the result of interaction, either physical or cyber, of multiple subsystems. When dealing with these kind of systems, centralized architectures for control and monitoring are neither feasible nor reliable due to communication and computational constraints. This has led to the development of distributed [3] and decentralized [4] architectures, in which local agents compute local regulation strategies to achieve global properties. One of the global properties which may be desirable is the ability of different subsystems to achieve consensus on the value of some state variables [5], [6]. Examples of systems which are regulated by this type of control strategy can be found in voltage and current regulation in DC microgrids [7], [8], [9], in cooperative control of UAVs and in formation control of autonomous vehicles [10]. In order to achieve these objectives, it is necessary for the local agents to communicate with each other. This, however, could make these systems vulnerable to the presence of attacks [11], [12]. Security of control systems is a topic which has attracted considerable interest in the literature in recent years [13], [14], [15], [16], [17], [18]. Most of the works about security of control systems consider centralized architectures. In the context of consensus protocols, [11] studies methods for cyber-attack detection, but requires knowledge of the global model to perform monitoring and detection. To the best of the authors‘ knowledge, the present paper is the first contribution proposing a distributed method to detect attacks in the communication network for the distributed control of interconnected systems.

Some distributed and decentralized methods for fault detection have been recently proposed [19], [20], [21], [22], [23], [24]. Most of the works for multi-agent systems assume that each agent knows the entire topology of the network connecting the agents (see [25] as example), thus requiring the knowledge of the global model. On the other hand, the proposed approach only requires information which is locally available or communicated by neighboring agents and the knowledge of the local model.

The main contribution of this paper is the design of a fully distributed model-based detection architecture for attacks in the communication network between subsystems which are physically interconnected, and which are regulated by a distributed consensus protocol. The proposed detection architecture can be performed locally by each subsystem through a local diagnoser. Furthermore, we analyze the conditions under which the detectability of the attacks is guaranteed, and, on the other hand, what properties the attack must have for it not to be detectable by the proposed strategy.

In this paper, we re-apprise the threshold-based distributed fault detection schemes proposed in [26] in the context of detection of cyber-attacks. However, we adapt these ideas to a novel scenario. Compared to our previous works on distributed fault detection, in this paper we consider a framework where each subsystem is governed by a two-layer controller, where primary control is devoted to stabilization and secondary control is based on a consensus dynamics for driving some outputs to a common value. Differently to previous papers, here cyber-attacks are modeled as faults on the communication links. Furthermore, in this paper

This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 739551 (KIOS CoE). This work has been conducted as part of the research project *Stability and Control of Power Networks with Energy Storage* (STABLE-NET) funded by the RCUK Energy Programme (contract no: EP/L014343/1), and has received support from the Swiss National Science Foundation under the COFLEX project (grant number 200021-169906).

F. Boem is with the Dept. of Electrical and Electronic Engineering at the Imperial College London, UK, and with the KIOS Research and Innovation Centre of Excellence, University of Cyprus. (f.boem@imperial.ac.uk)

A.J. Gallo is with the Dept. of Electrical and Electronic Engineering at the Imperial College London, UK. (alexander.gallo12@imperial.ac.uk)

G. Ferrari-Trecate is with the Automatic Control Laboratory, cole Polytechnique de Lausanne (EPFL), Swizerland. (giancarlo.ferraritrecate@epfl.ch)

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

we develop a threshold-based local detection scheme and analyze detectability of cyber-attacks even when a subsystem is simultaneously affected by multiple attacks. In this novel scenario, the detectability analysis offers a framework to analyze when a combination of attacks is stealthy. Finally, some preliminary results on the design of the consensus layer for counteracting stealthy attacks are provided.

II. Problem Formulation

Consider a LSS composed of N linear interconnected subsystems Σ_i, $i \in \mathcal{N} = \{1, ..., N\}$, modeled by the following discrete-time equations:

$$
\begin{aligned}
x^+_{[i]} &= A_{ii}x_{[i]} + \sum_{j\in\mathcal{N}_i} A_{ij}x_{[j]} + B_iu_{[i]} + M_id_{[i]} + G_iv_{[i]} + w_{[i]} \\
y_{[i]} &= x_{[i]} + \rho_{[i]},
\end{aligned}
\tag{1}
$$

where $x_{[i]} \in \mathbb{R}^{n_i}$ and $y_{[i]} \in \mathbb{R}^{n_i}$ are the local state and output vectors, and $x^+_{[i]}$ represents the state at time $t+1$; $u_{[i]} \in \mathbb{R}^{m_i}$ and $v_{[i]} \in \mathbb{R}^{p_i}$ are the primary and secondary control inputs; $d_{[i]} \in \mathbb{R}^{o_i}$ is a known exogenous input variable. Matrix $A_{ii} \in \mathbb{R}^{n_i\times n_i}$ is the local state transition matrix; $B_i \in \mathbb{R}^{n_i\times m_i}$, $G_i \in \mathbb{R}^{n_i\times l_i}$ and $M_i \in \mathbb{R}^{n_i\times o_i}$ are the primary, secondary and exogenous input transition matrices, respectively; $A_{ij} \in \mathbb{R}^{n_i\times n_j}$ describes the physical coupling between subsystems i and $j \in \mathcal{N}_i$, where the set $\mathcal{N}_i$ collects the parent subsystems of Σ_i: $\mathcal{N}_i = \{j \in \mathcal{N} | \partial x^+_{[i]}/\partial x_{[j]} \neq 0\}$. In (1), $w_{[i]}$ represents process noise and uncertainties, including those due to interconnections. We assume that all states variables are accessible but the measurements $y_{[i]}$ are affected by a noise term $\rho_{[i]}$. The following assumption is then required.

Assumption 1: The process uncertainty and the measurement noise functions $w_{[i]}(\cdot)$ and $\rho_{[i]}(\cdot)$ are unknown and bounded component-by-component by some known bounds: $w_{[i]}(t) \leq \bar{w}_{[i]}(t)\,, \quad \rho_{[i]}(t) \leq \bar{\rho}_{[i]}(t)\,, \quad \forall t \geq 0.$ ◁

We consider a *hierarchical control architecture* with the double objective of maintaining local stability of (1) and achieving consensus of some state variables among the subsystems of the LSS. This formulation is motivated by different application examples, such as multi-agent systems [10] and microgrids [7], [9].

Primary control input $u_{[i]}$ is a decentralized output feedback with gain matrix $K_i \in \mathbb{R}^{m_i\times n_i}$ designed so that $(A_{ii} + B_iK_i)$ is Schur stable. It can be computed as

$$
u_{[i]} = K_iy_{[i]}. \tag{2}
$$

Secondary control is based on a linear consensus protocol. To this end, we define a directed graph $\mathcal{G}$, with vertex set $\mathcal{V}$, and edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. In this paper, the topology of the digraph $\mathcal{G}$ corresponds to the underlying physical interconnections of the LSS. Hence, each node in $\mathcal{V}$ represents a subsystem of the LSS, and edge $(i, j) \in \mathcal{E}$ if and only if $j \in \mathcal{N}_i$. We assume the communication network has the same topology of $\mathcal{G}$ and define the secondary input as

$$
v^+_{[i]} = -\sum_{j\in\mathcal{N}_i} \alpha_{ij} \left(y_{[i]} - \tilde{y}_{[j]}\right) \tag{3}
$$

where $\alpha_{ij} \in \mathbb{R}^{p_i\times n_i}$ is the consensus-weighting matrix. It is defined such that the block matrix $\alpha = [\alpha_{ij}], \forall i, j \in \mathcal{N}, \alpha \in \mathbb{R}^{\bar{N}\times\bar{N}}$, $\bar{N} = \sum_{i=1}^{N} n_i$, is a consensus matrix, i.e. it is row-stochastic and primitive [27], in order to guarantee that there exists an equilibrium realizing the consensus of the output variables. This implies that $\mathcal{G}$ is strongly connected. Vector $\tilde{y}_{[j]} \in \mathbb{R}^{n_j}$ in (3) denotes the output $y_{[j]}$ received by the subsystem i from the neighboring subsystem j through a communication link of the communication infrastructure of the secondary control system. The communication link corresponding to the arc (i, j) is modeled as

$$
\tilde{y}_{[j]}(t) = y_{[j]}(t) + \beta_j(t - T_{a_{[j]}})\phi_{[j]}(t)\,, \tag{4}
$$

where the term $\beta_j(t - T_{a_{[j]}})\phi_{[j]}(t)$ denotes the effect on the received vector $\tilde{y}_{[j]}(t)$ of a possible attack. Specifically, the activation function $\beta_j(t - T_{a_{[j]}})$ is a function of time such that $\beta_j(t - T_{a_{[j]}}) = 0$, $t < T_{a_{[j]}}$ $\beta_j(t - T_{a_{[j]}}) = 1$ for $t \geq T_{a_{[j]}}$ (that is, before $T_{a_{[j]}}$ the attacker either does not have access to the communication link, or has not begun the attack). The function $\phi_{[j]}(t) : \mathbb{R}^+ \to \mathbb{R}^{n_j}$ denotes the actual effect of the attack on the communication link which, in turn, affects the consensus protocol (possibly harming it).

Assumption 2: Possible attacks affect only communication links of the secondary control system transmitting information between neighboring subsystems according to Eqs. (3) and (4). It is assumed that local measurements used in the primary control system cannot be accessed by an attacker (see Eqs. (1) and (2)). ◁

III. Attack Detection Architecture

In this section, we design a distributed attack detection architecture in which *only the knowledge of the local model is needed* and aiming at *locally* detecting the presence of attacks in the communication links between neighboring subsystems in the secondary consensus control scheme.

A. Distributed State Estimator

Each subsystem is equipped with a local diagnoser computing a local estimate of its state based on the knowledge of the local system dynamics (1), as well as the measurements that it receives from its neighbors. The local estimation model evolves according to the following equations:

$$
\hat{\Sigma}_i : \quad
\begin{aligned}
\hat{x}^+_{[i]} &= A_{ii}\hat{x}_{[i]} + \sum_{j\in\mathcal{N}_i} A_{ij}\tilde{y}_{[j]} + B_iu_{[i]} + M_id_{[i]} \\
&\quad + G_iv_{[i]} + L_i\left(y_{[i]} - \hat{y}_{[i]}\right) \\
\hat{y}_{[i]} &= \hat{x}_{[i]}
\end{aligned}
\tag{5}
$$

where matrix $L_i \in \mathbb{R}^{n_i\times n_i}$ is such that matrix $A_{L_i} = A_{ii} - L_i$ is Schur stable.

B. Detection threshold

We now design the local attack detection threshold. To do this, we first analyze the estimation error, defined as $\epsilon_{[i]} = x_{[i]} - \hat{x}_{[i]}$. The dynamics of the estimation error under healthy conditions, i.e. for time $t < T_{a_{[j]}}, \forall j \in \mathcal{N}_i$, are given using (1) and (5) as:

$$\epsilon_{[i]}^{+} = A_{L_i}\epsilon_{[i]} + \sum_{j\in\mathcal{N}_i} A_{ij}\rho_{[j]} + w_{[i]} - L_i\rho_{[i]} \tag{6}$$

Note that system (6) is BIBO stable, and that its solution may be written as the following:

$$\epsilon_{[i]}(t) = A_{L_i}^{t}\epsilon_{[i]}(0) + \sum_{k=0}^{t-1} A_{L_I}^{t-k-1}\eta_{[i]}(k)$$

where $\eta_{[i]}(t) = \sum_{j\in\mathcal{N}_i} A_{ij}\rho_{[j]}(t) + w_{[i]}(t) - L_i\rho_{[i]}(t)$. Since A_{L_i} is Schur stable, there exist constants $\gamma_i > 0$ and $\lambda_i \in [0, 1)$ such that

$$\|A_{L_i}^{t}\| \leq \gamma_i\lambda_i^{t},$$

where operator $\|\cdot\|$ is the matrix norm. Given Assumption 1 and exploiting the triangle inequality, it is possible to define a threshold bounding the estimation error's absolute value:

$$\bar{\epsilon}_{[i]}(t) = \gamma_i\lambda_i^{t}\bar{\epsilon}_{[i]}(0) + \sum_{k=0}^{t-1} \gamma_i\lambda_i^{t-k-1}\bar{\eta}_{[i]}(k) \tag{7}$$

where $\bar{\eta}_{[i]}(t) = \sum_{j\in\mathcal{N}_i} \|A_{ij}\|\bar{\rho}_{[j]}(t) + \bar{w}_{[i]}(t) + \|L_i\|\bar{\rho}_{[i]}(t)$, and $\bar{\epsilon}_{[i]}(0)$ is properly defined. As the threshold in (7) is generated using the triangle inequality, it guarantees the absence of false alarms. It is worth noting that the update of the estimation error bound requires only local communication from neighboring subsystems.

We now formulate the attack detection threshold bounding the residual $r_{[i]}(t) = y_{[i]}(t) - \hat{y}_{[i]}(t) = \epsilon_{[i]}(t) + \rho_{[i]}(t), \forall t \geq 0$. Hence the time-varying detection threshold is defined as

$$\bar{r}_{[i]}(t) = \bar{\epsilon}_{[i]}(t) + \bar{\rho}_{[i]}(t)\,. \tag{8}$$

The detection threshold $\bar{r}_{[i]}(t)$ given in (8) guarantees that

$$|r_{[i]}(t)| \leq \bar{r}_{[i]}(t) \tag{9}$$

when no attacks affect the secondary control system.

Therefore, the local detection algorithm relies on checking whether condition (9) is (locally) satisfied. It is sufficient that

$$|r_{[i,k]}(t)| > \bar{r}_{[i,k]}(t)$$

for at least one component $k \in \{1, ..., n_i\}$ at time $t = T_d$ to affirm that an attack has occurred in at least one of the output measurements received from the neighboring subsystems.

C. Detectability

In this section, the detectability properties of the proposed methodology are analyzed. The proofs are omitted due to space constraints. We start by considering the presence of a single attack $\phi_{[\hat{j}]}(\cdot)$ affecting the measurements received by subsystem i from one of its neighbors $\hat{j} \in \mathcal{N}_i$.

Proposition 1: Consider subsystem Σ_i, with dynamics as in (1), the distributed estimator (5), the detection threshold (8), and an attack with activation function $\beta_{\hat{j}}(t - T_a) = 1,\, t \geq T_a$. It is sufficient that

$$\exists\tau > T_a : \left|\sum_{k=0}^{\tau-1}(A_{L_i})^{\tau-1-k}A_{i\hat{j}}\phi_{[\hat{j}]}(k)\right| > 2\bar{r}_{[i]}(\tau). \tag{10}$$

for the attack sequence $\left[\phi_{[\hat{j}]}(T_a), \dots, \phi_{[\hat{j}]}(\tau)\right]$ to be detectable by the attack detection logic based on threshold (8). □

In Proposition 1, a class of attacks which may be detectable by their cumulative effect is defined in a non-closed form. The following – possibly conservative – sufficient condition provides a more explicit condition for detectability depending on the effect of the attack at a specific time instant.

Proposition 2: Consider subsystem Σ_i, with dynamics as in (1), the distributed estimator (5), the detection threshold (8), and an attack with activation function $\beta_{\hat{j}}(t - T_a) = 1,\, t \geq T_a$. It is sufficient that

$$\exists\tau \geq T_a : \left|\phi_{[\hat{j}]}(\tau)\right| > 2\bar{r}_{[i]}(\tau) \tag{11}$$

for the attack input $\phi_{[\hat{j}]}(\tau)$ to be detectable by the proposed detection logic at time $t = \tau$.

□

We now analyze the case in which the attacker has access to multiple communication lines entering Σ_i, and thus alters measurements $\tilde{y}_{[j]}(t), j \in \widehat{\mathcal{N}}_i \subseteq \mathcal{N}_i$, where $\widehat{\mathcal{N}}_i$ is the set of neighboring subsystems whose communication links have been attacked. We denote with $\phi_{[j\in\widehat{\mathcal{N}}_i]}$ the combination of attacks $\phi_{[j]}$ affecting nodes $j \in \widehat{\mathcal{N}}_i$.

Proposition 3: Consider subsystem Σ_i, with dynamics as in (1), distributed estimator (5), detection threshold (8), and an attack on communication channels $j \in \widehat{\mathcal{N}}_i$ with activation functions $\beta_j(t - T_a) = 1,\, t \geq T_a$. It is sufficient that

$$\exists\tau \geq T_a : \left|\sum_{k=0}^{\tau-1}(A_{L_i})^{\tau-1-k}\sum_{j\in\widehat{\mathcal{N}}_i} A_{ij}\phi_{[j]}(k)\right| > 2\bar{r}_{[i]}(\tau). \tag{12}$$

for the attack sequences $\left[\phi_{[j]}(T_a), \dots, \phi_{[j]}(\tau)\right],\, j \in \widehat{\mathcal{N}}_i$ to be detectable by the detection logic based on (8). □

Finally, similarly as the case of an attack on a single communication channel, a more explicit detectability condition is given below for the multiple-attack case.

Proposition 4: Consider subsystem Σ_i, with dynamics as in (1), the distributed estimator (5), the detection threshold (8), and an attack on multiple communication channels with activation functions $\beta_j(t - T_a) = 1,\, t \geq T_a, \forall j \in \widehat{\mathcal{N}}_i$. Suppose that

$$\exists\tau \geq T_a : \left|\sum_{j\in\widehat{\mathcal{N}}_i} A_{ij}\phi_{[j]}(\tau)\right| > 2\bar{r}_{[i]}(\tau) \tag{13}$$

Then, the attack input $\phi_{[j]}(\tau),\, j \in \widehat{\mathcal{N}}_i$ is detected by the attack detection logic at time τ.

□

D. Class of stealthy attacks

In the following proposition, a condition on the attack functions $\phi_{[j]}, j \in \widehat{\mathcal{N}}_i$ is given so that it is guaranteed that their effect is not detectable by the proposed method. It is worth noting that this attack assumes that the attacker perfectly knows the coupling matrices A_{ij}.

Theorem 1: Consider subsystem Σ_i, with dynamics as in (1), the distributed estimator (5), the detection threshold (8), and an attack on multiple communication channels with activation functions $\beta_j(t - T_{a_{[j]}}) = 1$, $t \geq T_{a_{[j]}}$, $j \in \widehat{\mathcal{N}}_i$. If

$$\left|\sum_{j\in\widehat{\mathcal{N}}_i} A_{ij}\phi_{[j]}(\tau)\right| = 0, \forall \tau \geq T_{a_{[j]}} \tag{14}$$

then the attacks are not detectable by the attack detection logic based on the threshold (8).

□

The proof is omitted due to space constraints.

The class of attacks satisfying (14), even if not detectable, causes a non-zero effect on the dynamics of the local state as they affect the consensus input (3) as follows:

$$v_{[i]}^{+} = -\sum_{j\in\mathcal{N}_i} \alpha_{ij}(y_i - y_j) + \sum_{j\in\widehat{\mathcal{N}}_i} \alpha_{ij}\phi_{[j]}$$

and, in general $\alpha_{ij} \neq A_{ij}$. It is hence possible for attack functions to be such that (9) is satisfied $\forall t \geq 0$, whilst resulting in a non-zero effect on the secondary input value.

In the next section, we analyze the effect that attack functions which are guaranteed not detectable have on subsystem dynamics, and we provide some notes on how the consensus protocol may be designed such that this effect is mitigated.

E. Remarks on the design of the consensus layer for counteracting stealthy attacks

We analyze the effect on Σ_i caused by an attack function $\bar{\phi}_{[j\in\widehat{\mathcal{N}}_i]}$ designed to be undetectable, i.e. to satisfy (14). We consider two trajectories of system (1) affected by the same noises: $\tilde{x}_{[i]}$ represents a nominal trajectory without attacks, while $\tilde{x}^a_{[i]}$ is affected by malicious attacks $\bar{\phi}_{[j\in\widehat{\mathcal{N}}_i]}$ after $t > T_{a_{[j]}}$. We note that $\tilde{x}_{[i]}(t) = \tilde{x}^a_{[i]}(t)$, for $0 \leq t \leq T_{a_{[j]}}$. We introduce a mismatch vector, $\tilde{\epsilon}_{[i]} = \tilde{x}^a_{[i]} - \tilde{x}_{[i]}(t)$, which describes the effect of the attack on the state trajectory. Therefore the dynamics of $\tilde{\epsilon}_{[i]}$ are given by:

$$\tilde{\epsilon}_{[i]}^{+} = A_{ii}\tilde{\epsilon}_{[i]} + G_i\tilde{\epsilon}_{v_{[i]}} \tag{15}$$

where $\tilde{\epsilon}_{v_{[i]}} = \tilde{v}^a_{[i]} - \tilde{v}_{[i]}$ represents the difference between the consensus inputs with and without attacks, and its dynamics are given by (3) as

$$\tilde{\epsilon}_{v_{[i]}}^{+} = -\tilde{\epsilon}_{[i]} + \sum_{j\in\widehat{\mathcal{N}}_i} \alpha_{ij}\bar{\phi}_j \tag{16}$$

Note that it is assumed that values $w_{[i]}(t)$ and $\rho_{[i]}(t)$ are the same for $\tilde{x}^a_{[i]}$ and $\tilde{x}_{[i]}$ for all $t \geq 0$. The satisfaction of (14) is equivalent to

$$\mathrm{col}(\bar{\phi}_{[j]}(t)|j \in \widehat{\mathcal{N}}_i) \in \ker\left(\left[A_{ij_1}, \ldots, A_{ij_{|\widehat{\mathcal{N}}_i|}}\right]\right)$$

where $\mathrm{col}(\bar{\phi}_{[j]}(t)|j \in \widehat{\mathcal{N}}_i) \in \mathbb{R}^{n_i|\widehat{\mathcal{N}}_i|}$ is a column vector collecting the attack functions acting on the i-th subsystem at time t. Hence, since from (16) we know that the attack functions affect $\tilde{\epsilon}_{[i]}$ through weight matrices $\alpha_{ij}, j \in \widehat{\mathcal{N}}_i$, we would like to design the consensus weights in such a way that if $\bar{\phi}_{[j\in\widehat{\mathcal{N}}_i]}$ satisfies (14), it also holds that

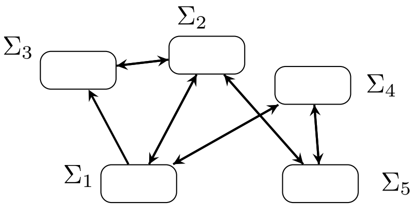


Fig. 1: Example large-scale system. Σ_i are the subsystems of the LSS, and the arrows represent both physical and communication links between the subsystems.

$$\mathrm{col}(\bar{\phi}_{[j]}|j \in \widehat{\mathcal{N}}_i) \in ker([\alpha_{ij_1}, \ldots, \alpha_{ij_{|\widehat{\mathcal{N}}_i|}}])$$

Hence, it is advisable to develop a design procedure for the consensus weights α_{ij} such that

$$\ker\left([\alpha_{ij_1}, \ldots, \alpha_{ij_{|\mathcal{N}_i|}}]\right) \subseteq \ker\left(\left[A_{ij_1}, \ldots, A_{ij_{|\mathcal{N}_i|}}\right]\right)$$

holds for all $i \in \mathcal{N}$. In the simulations section, we show a preliminary result in a specific simulation example where the consensus weights are designed to make the system resilient to undetected attacks.

IV. Simulation Results

Consider a large-scale system composed of $N = 5$ subsystems, which are interconnected as in Figure 1. Each subsystem evolves according to model (1), where A_{ii}, A_{ij}, B_i, G_i are

$$A_{ii} = \begin{bmatrix} 0 & 1 \\ -1.4 & 0.3 \end{bmatrix}, \ B_i = \begin{bmatrix} 0.4 \\ 0 \end{bmatrix}, \ G_i = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix},$$

$$A_{ij} = \begin{bmatrix} -1.1 & 0 \\ 0 & -0.9 \end{bmatrix},$$

$$\alpha_{12} = [0.5I_2], \ \alpha_{14} = [0.5I_2], \ \alpha_{21} = [0.5I_2], \ \alpha_{23} = [0.1I_2],$$

$$\alpha_{25} = [0.4I_2], \ \alpha_{31} = [0.6I_2], \ \alpha_{32} = [0.4I_2], \ \alpha_{41} = [0.4I_2],$$

$$\alpha_{45} = [0.6I_2], \alpha_{52} = [0.5I_2], \ \alpha_{54} = [0.5I_2],$$

state interconnections can be seen in Figure 1, and $M_i = 0$. Both the process disturbance $w_{[i]}(t)$ and the measurement noise $\rho_{[i]}(t)$ are modeled as uniform distributions bounded by $\bar{w}_{[i]} = [0.03, 0.01]^{\top}, \forall i \in \mathcal{N}, \forall t \geq 0$, and $\bar{\rho}_{[i]} = [0.03, 0.01]^{\top}, \forall i, \forall t \geq 0$, respectively. The initial conditions $x_{[i]}(0)$ are as follows: $x_{[1]}(0) = [0.4464, -0.4382]^{\top}, x_{[2]}(0) = [0.3678, -0.3929]^{\top}, x_{[3]}(0) = [-0.2886, 0.2462]^{\top}, x_{[4]} = [0.4183, 0.4615]^{\top}, x_{[5]} = [-0.3842, 0.0556]^{\top}$.

A. Detectability analysis

We start by presenting the results obtained by the proposed distributed detection strategy in the case of an attacker which has access to a single communication line linking the subsystems. Specifically, we consider that the attacker is able to modify output signal $y_{[1]}(t)$ transmitted from Σ_1 to Σ_2. We firstly show the effect of a constant attack occurring abruptly at time $T_{a_{[1]}} = 35$. The attack function

TABLE I: Results of 200 simulations with abrupt attacks

c	$Avg(T_d)$	$\sigma(T_d)$	Number of times detection failed
0.05	109	42	102
0.075	49	13	0
0.1	39	3	0
0.125	37	1	0
0.15	36	0.5	0
0.175	36	0.2	0
≥ 0.2	36	0	0

TABLE II: Results of 200 simulations with gradual attacks

c	$Avg(T_d)$	$\sigma(T_d)$	Number of times detection failed
0.05	112	44	107
0.075	50	14	0
0.1	39	3	0
0.125	37	1	0
0.15	36	0.5	0
0.175	36	0.2	0
≥ 0.2	36	0	0

is $\phi_{[1]}(t) = [c, c]^\top, \forall t \geq 35$, where c is a constant value. In Table I we show the average value and the standard deviation of the detection time T_d for different values of c. The samples are obtained using 200 simulations over an interval $t \in \{0, \ldots, 200\}$ for each value of c. In Figure 2 we plot $|r_{[i,h]}(t)|$ and the detection threshold $\bar{r}_{[i]}(t)$, for $i = 1, \ldots, 5$ and $c = 0.125$, and we can see that the attack is immediately detected by subsystem Σ_2 (red line).

Secondly, we show how the detection time changes if the attack function affects the communicated measurements in an incipient way, rather than abruptly. The attack function is therefore modeled as $\phi_{[1]}(t) = (1 - \gamma^{(T_{a_{[1]}} - t)})[c, c]^\top$. The results are summarized in Table II, with $\gamma = 1.001$, for different values of c.

B. Stealthy attack

We now show the case in which the attacker takes control of multiple communication links connecting subsystem 2 with its neighbors, specifically $\widehat{\mathcal{N}}_2 = \{1, 3\}$, and through perfect knowledge of A_{21} and A_{23} is able to design a stealthy attack. Hence functions $\phi_{[1]}(t)$ and $\phi_{[3]}(t)$ are designed such that $col(\phi_{[j]} | j \in \{1, 3\}) \in ker([A_{21}, A_{23}])$. We show the result of this attack in Figure 3 and Figure 4a. As expected, attack detection fails, as the attack functions are designed to satisfy (14), and the attacker is able to move the states away from the consensus value. Finally, we show how it is possible to design the consensus weight matrices α_{ij} so that stealthy attacks do not influence the consensus. To ensure that any stealthy attack satisfies $\bar{\phi}_i \in ker\left[\alpha_{ij_1}, \ldots, \alpha_{ij_{|\mathcal{N}_i|}}\right]$, we design the weight matrices as

$$\left[\alpha_{ij_1}, \ldots, \alpha_{ij_{|\mathcal{N}_i|}}\right] = W_i \left[A_{ij_1}, \ldots, A_{ij_{|\mathcal{N}_i|}}\right]. \qquad (17)$$

Matrix W_i must be such that the resulting $\alpha = [\alpha_{ij}], \forall i, j \in \mathcal{N}$ is row-stochastic and primitive. In the example, to ensure row-stochasticity of α, we design W_i as a diagonal matrix, having each element on the diagonal as

$$W_{i,(hh)} = \frac{1}{\sum_{j \in \mathcal{N}_i} \sum_l A_{ij,(hl)}}$$

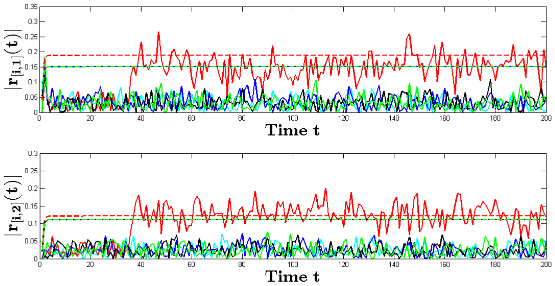


Fig. 2: Time evolution of each component of the absolute values of the residuals, together with the corresponding detection threshold. For each subsystem $i = 1, \ldots, 5$, $|r_{[i,h]}(t)|$ are represented by solid lines, while $\bar{r}_{[i,h]}(t)$ by dashed lines. Abrupt attack $\phi_{[1]}(t) = [0.125, 0.125]^\top, \forall t \geq 35$ in the communication link from Σ_1 to Σ_2. Detection occurs once $|r_{[2,h]}(t)| > \bar{r}_{[2,h]}(t)$ is satisfied for at least one component.

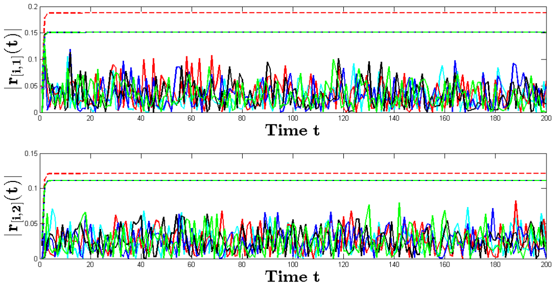


Fig. 3: Detection failure in the case of a stealthy attack. Time evolution of each component of the absolute values of the residuals, together with the corresponding detection threshold. For each subsystem $i = 1, \ldots, 5$, $\left|r_{[i,h]}(t)\right|$ are represented by solid lines, while $\bar{r}_{[i,h]}(t)$ by dashed lines.

where $A_{ij,(hl)}$ is the (h, l)-th component of A_{ij}. Moreover, since in this example the communication graph is strongly connected, then we can say that the obtained α is primitive. Using this procedure, we obtain

$$\alpha_{12} = [0.5I_2], \ \alpha_{14} = [0.5I_2], \ \alpha_{21} = [0.\bar{3}I_2], \ \alpha_{23} = [0.\bar{3}I_2],$$

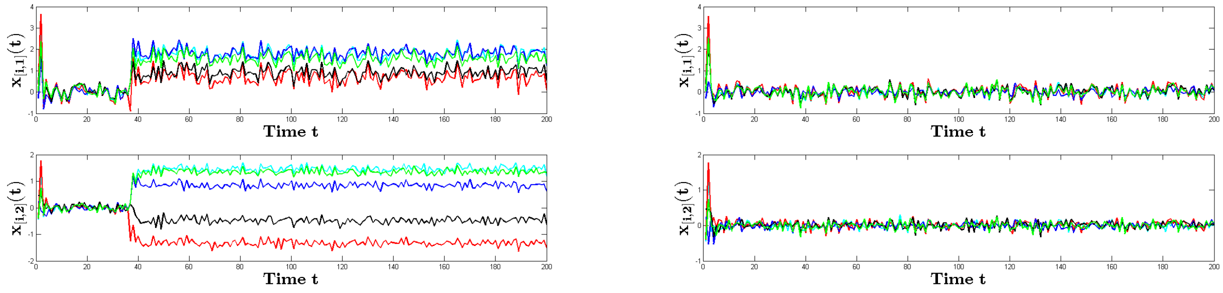
$$\alpha_{25} = [0.\bar{3}I_2], \ \alpha_{31} = [0.5I_2], \ \alpha_{32} = [0.5I_2], \ \alpha_{41} = [0.5I_2],$$

$$\alpha_{45} = [0.5I_2], \ \alpha_{52} = [0.5I_2], \ \alpha_{54} = [0.5I_2],$$

In Figure 4b we show how applying this design of α_{ij} to the consensus controllers, the effect of the stealthy attack functions $\phi_{[1]}(t)$ and $\phi_{[3]}(t)$ is nullified, whilst maintaining stability and preserving consensus equilibrium of the system.

V. Concluding Remarks

This paper presents some preliminary results for the problem of detecting cyber-attacks in the communication network between interconnected subsystems governed by consensus-based control. A distributed attack detection method is proposed and the detectability properties are analyzed both in theory and in a simulation example. The theoretical results are conservative but are a first step in the direction of designing an effective distributed attack detection scheme.



(a) (b)

Fig. 4: Evolution of states of LSS under stealthy attack with (b) and without (a) the design of the weight matrices of the consensus α_{ij} as in (17). Through the proposed design of the weight matrices of the consensus, the effect of the attack is nullified, and therefore states remain at consensus equilibrium.

A class of stealthy attacks is identified and some conditions on the design of the consensus-control layer to cancel the effect of these attacks on the system dynamics are provided.

As future work, we aim at considering the attack isolation problem in order to identify the communication links involved by the attack. Furthermore, we will investigate the opportunity to reconfigure the control law after attack detection to make the system resilient to attacks. Moreover, we will further investigate the idea to reduce the effect of stealthy attacks on the dynamics of the interconnected subsystems, while guaranteeing the convergence conditions for the consensus-based secondary control scheme.

Finally, in future work, the application of the proposed approach will be validated on a micro-grid use-case.

References

[1] R. Baheti and H. Gill, "Cyber-physical systems," *The impact of control technology*, vol. 12, pp. 161–166, 2011.

[2] J. Shamma, *Cooperative control of distributed multi-agent systems*. John Wiley & Sons, 2008.

[3] R. Scattolini, "Architectures for distributed and hierarchical model predictive control–a review," *Journal of Process Control*, vol. 19, no. 5, pp. 723–731, 2009.

[4] N. Sandell, P. Varaiya, M. Athans, and M. Safonov, "Survey of decentralized control methods for large scale systems," *IEEE Transactions on automatic Control*, vol. 23, no. 2, pp. 108–128, 1978.

[5] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on automatic control*, vol. 49, no. 9, pp. 1520–1533, 2004.

[6] G. Xie and L. Wang, "Consensus control for a class of networks of dynamic agents," *International Journal of Robust and Nonlinear Control*, vol. 17, no. 10-11, pp. 941–959, 2007.

[7] L. Meng, T. Dragicevic, J. Roldán-Pérez, J. C. Vasquez, and J. M. Guerrero, "Modeling and sensitivity study of consensus algorithm-based distributed hierarchical control for dc microgrids," *IEEE Transactions on Smart Grid*, vol. 7, no. 3, pp. 1504–1515, 2016.

[8] J. Zhao and F. Dörfler, "Distributed control and optimization in dc microgrids," *Automatica*, vol. 61, pp. 18–26, 2015.

[9] M. Tucci and G. Ferrari-Trecate, "Plug-and-play control and consensus algorithms for current sharing in dc microgrids," *IFAC-PapersOnLine*, 2017, 20th IFAC World Congress.

[10] W. Ren and R. W. Beard, *Distributed consensus in multi-vehicle cooperative control*. Springer, 2008.

[11] F. Pasqualetti, A. Bicchi, and F. Bullo, "Consensus computation in unreliable networks: A system theoretic approach," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 90–104, 2012.

[12] 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.

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

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

[15] F. Pasqualetti, "Secure control systems: A control-theoretic approach to cyber-physical security," Ph.D. dissertation, Citeseer, 2012.

[16] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems–part ii: Centralized and distributed monitor design," *arXiv preprint arXiv:1202.6049*, 2012.

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

[18] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems*, vol. 35, no. 1, pp. 93–109, 2015.

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

[20] 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.

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

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

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

[24] 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.

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

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

[27] F. Bullo, *Lectures on Network Systems*. Version 0.95, 2017, with contributions by J. Cortes, F. Dorfler, and S. Martinez. [Online]. Available: http://motion.me.ucsb.edu/book-lns