2015 IEEE 54th Annual Conference on Decision and Control (CDC)
December 15-18, 2015. Osaka, Japan

A Divide-and-Conquer Approach to Distributed Attack Identification

Fabio Pasqualetti, Florian Dörfler and Francesco Bullo

Abstract— **Identifying attacks is key to ensure security in cyber-physical systems. In this paper we remark upon the computational complexity of the attack identification problem by showing how conventional approximation techniques may fail to identify attacks. Then, we propose decentralized and distributed monitors for attack identification with performance guarantees and low computational complexity. The proposed monitors rely on a geometric control framework, yet they require only local knowledge of the system dynamics and parameters. We exploit a *divide-and-conquer* approach, where first the system is partitioned into disjoint regions, then corrupted regions are identified via distributed computation, and finally corrupted components are isolated within regions.**

I. Introduction

Cyber-physical systems are the core of many technological domains, including health care and biomedicine, telecommunications, and energy management. Due to their importance, cyber-physical systems are not only prone to sensor and actuator failures as legacy control systems, but also to intentional attacks against control and communications modules. Attacks can have major consequences, ranging from economic losses to instabilities and services disruption [1], [2], [3].

Detection and identification of attacks is necessary to design effective security mechanisms. Fundamental limitations in the detectability and identifiability of attacks have recently been characterized for different system dynamics, attack models, and monitoring systems. For instance, in [4], [5], [6], [7] it is shown how attackers with access to sufficiently many system resources can always avoid detection and identification, as well as attackers with more limited resources and full knowledge of the system dynamics and state. Conversely, if the monitoring resources and information outbalance the attack capabilities, the attack locations and strategy can be promptly reconstructed. Moreover, while detecting attacks is computationally *easy* in both centralized and distributed settings [4], [8], identifying the attack location and strategy is computationally *hard* [4].

Despite its importance, few solution have been proposed for the identification of attacks. A complete, yet computationally intensive, solution to the attack identification problem is described in [4] by using unknown-input observers and geometric control techniques [9]. Convex relaxation methods are employed in [10] to derive an efficient (yet incomplete and without guarantees) identification algorithm for the case of attacks corrupting the system measurements. In [11] it is shown that certain instances of the identification problem can in fact be solved efficiently. Finally, decentralized state estimation and attack identification is discussed in [12], [13]. In these decentralized approaches local control centers have sensory information and system data from within their region, and signals from other regions are either directly measured or naively treated as unknown inputs.

This paper concerns the computational complexity of the attack identification problem, and its main contributions are as follows. First, we highlight the complexity of the attack identification problem by converting it to an equivalent cardinality minimization problem. We show how and why common convex relaxation techniques for static problems aiming at circumventing the computational complexity may fail to identify attacks in dynamic systems (Section II). Our examples highlight that, in large-scale systems, different output and state attacks may achieve the same cost in relaxed optimization problems, thereby impeding their identification.

The inherent computational complexity and shortcomings of relaxation methods motivate our second contribution: we present a fully decentralized and low-complexity identification method and characterize its performance (Section III-B). Our decentralized method relies on geographically distributed control centers, which have local knowledge of the system parameters. We show that the performance of our decentralized method depends only on the system structure and parameters, and not on the attack strategy. Hence, our method also provides guidelines to design secure systems.

As third and main contribution, we propose a distributed identification method based on the *divide-and-conquer* principle (Section III-C). Our distributed method is based on local state estimation, cooperation with neighboring control centers, regional attack detection, and finally regional attack identification. Analogously to our decentralized method, our distributed algorithm requires only local model information and communication, and it achieves guaranteed identification of a class of attacks. Our distributed method overcomes the performance of its decentralized counterpart, at the expense of communication and a more involved algorithmic structure.

II. The centralized identification problem

In this section we present our setup for the attack identification problem, and we recall some results and fundamental limitations of centralized identification methods.

This material is based upon work supported in part by NSF award ECCS-1405330 and in part by ONR award N00014-14-1-0816. Fabio Pasqualetti is with the Mechanical Engineering Department, University of California at Riverside, fabiopas@engr.ucr.edu. Florian Dörfler is with the Automatic Control Laboratory, Swiss Federal Institute of Technology (ETH) Zürich dorfler@ethz.ch. Francesco Bullo is with the Mechanical Engineering Department, University of California at Santa Barbara, bullo@engineering.ucsb.edu.

A. Centralized setup and notation

We represent a cyber-physical system under attack with the continuous-time, linear, and time-invariant system[1]

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{1}$$

where $x : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$, $y : \mathbb{R}_{\geq 0} \to \mathbb{R}^p$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$. The inputs $Bu : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ and $Du : \mathbb{R}_{\geq 0} \to \mathbb{R}^p$ are assumed to be unknown. Besides reflecting the genuine failure of systems components, these unknown inputs model the effect of attacks against cyber and physical components. We assume that each state and output variable can be independently compromised. Accordingly, we partition the input matrices and attack signals as $B = [I_{n\times n}\ 0_{n\times p}]$, $D = [0_{p\times n}\ I_{p\times p}]$, and $u = [u_x^\mathsf{T},\ u_y^\mathsf{T}]^\mathsf{T}$, where u_x and u_y are referred to as *state attack* and *output attack*, respectively. As shown in [4], many interesting cyber-physical systems and attacks can be modeled by system (1) with unknown inputs.

The attack signal depends upon the attack strategy. In particular, if the *attack set* (or attacked variables) is $K \subseteq \{1, \ldots, n+p\}$, with $|K| = k$, then only the entries of u indexed by K are nonzero over time, that is, for each $i \in K$, there exists a time t such that $u_i(t) \neq 0$ and $u_j(t) = 0$ for all $j \notin K$ and at all times. To underline this sparsity relation, we use u_K to denote the attack signal, that is the subvector of u indexed by K. Analogously, the pair (B_K, D_K), where B_K and D_K are the submatrices of B and D with columns in K, are referred to as the *attack signature*. Hence, $Bu = B_K u_K$, and $Du = D_K u_K$. Finally, we assume that the cardinality k of the attack set, or an upper bound, is known.

B. Identifiability of cyber-physical attacks

Informally, an attack K is unidentifiable if it cannot be distinguished (from knowledge of the measurements and the system parameters) from another attack R corrupting equally many or fewer variables. Here, we confine ourselves to comparing the attack set K with other attack sets R with $|R| \leq |K|$ since sufficiently large attack sets can always be designed to be unidentifiable, for instance, by corrupting sufficiently many sensors [14].

More formally, let $y(x_0, u, t)$ be the output sequence generated from the state x_0 under the attack signal u. We adopt the following definition of identifiability of attacks [4]:

Definition 1: ***(Identifiability of cyber-physical attacks)*** For the system (1) with initial state x_0, the attack $(B_K u_K, D_K u_K)$ is *unidentifiable* if and only if $y(x_0, u_K, t) = y(x_1, u_R, t)$ for some initial state $x_1 \in \mathbb{R}^n$, for some attack $(B_R u_R, D_R u_R)$ with $|R| \leq |K|$ and $R \neq K$, and for all $t \in \mathbb{R}_{\geq 0}$.

In [4, Theorem 3.4], we provided the following equivalent system-theoretic characterization of identifiability:

[1]The results stated in this paper for continuous-time systems hold also in discrete time and for singular descriptor systems, see Remark 1. Moreover, due to linearity of (1), known inputs do not affect our results and are not included in the model.

Theorem 2.1: ***(Algebraic test for identifiability of cyber-physical attacks)*** For the system (1) and an attack set K, the following statements are equivalent:

(i) the attack set K is unidentifiable; and

(ii) there is an attack set R, with $|R| \leq |K|$ and $R \neq K$, some $s \in \mathbb{C}$, and vectors $g_K \in \mathbb{C}^{|K|}$, $g_R \in \mathbb{C}^{|R|}$, and $x \in \mathbb{C}^n$, with $x \neq 0$, such that

$$\begin{aligned} (sI - A)x - \begin{bmatrix} B_K & B_R \end{bmatrix} \begin{bmatrix} g_K \\ g_R \end{bmatrix} &= 0, \\ Cx + \begin{bmatrix} D_K & D_R \end{bmatrix} \begin{bmatrix} g_K \\ g_R \end{bmatrix} &= 0. \end{aligned} \tag{2}$$

Condition (2) shows the equivalence between unidentifiable attacks of cardinality k and the existence of *invariant zeros* for the system $(A, B_{\bar{K}}, C, D_{\bar{K}})$ with $|\bar{K}| \leq 2k$ [9].

C. Centralized identification: complexity and pitfalls

The attack identification problem is concerned with identifying the attack set K from measurements y and knowledge of the system parameters (A, C). The identification problem can be reformulated as the following cardinality minimization problem [4, Lemma 4.4]: given a system with state transition matrix $A \in \mathbb{R}^{n\times n}$, measurement matrix $C \in \mathbb{R}^{p\times n}$, and measurement signal $y : \mathbb{R}_{\geq 0} \to \mathbb{R}^p$, find the minimum cardinality input signals $v_x : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ and $v_y : \mathbb{R}_{\geq 0} \to \mathbb{R}^p$ and an initial condition $\xi_0 \in \mathbb{R}^n$ that explain the measurements y, that is,

$$\begin{aligned} \min_{v_x,\, v_y,\, \xi_0} \quad & \|v_x\|_{\mathcal{L}_0} + \|v_y\|_{\mathcal{L}_0} \\ \text{subject to} \quad & \dot{\xi}(t) = A\xi(t) + v_x(t), \\ & y(t) = C\xi(t) + v_y(t), \\ & \xi(0) = \xi_0 \in \mathbb{R}^n\,. \end{aligned} \tag{3}$$

Here we use the shorthands $\text{supp}(x) = \{i \in \{1, \ldots, n\} : x_i \neq 0\}$ for a vector $x \in \mathbb{R}^n$ and $\|v\|_{\mathcal{L}_0} = |\cup_{t\in\mathbb{R}_{\geq 0}} \text{supp}(v(t))|$ for a vector-valued signal $v : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$.

The optimization problem (3) is generally combinatorial and belongs to the class of *NP-hard* problems [4, Corollary 4.5]. As a consequence of this inherent complexity, existing complete solutions to identify the attack set K require a combinatorial procedure, since, a priori, K is one of the $\binom{n+p}{|K|}$ possible attack sets. In [4, Section 4.D], the authors provided a solution based on the implementation of $\binom{n+p}{|K|}$ residual filters [9] each determining whether a predefined set coincides with the attack set. The solution in [4] is *complete*, but does not scale to large attack sets.

In the case of discrete-time systems subject to output attacks, the attack identification problem can be solved efficiently if the monitoring system has access to a substantial amount of resources. The particular assumption is that the pair (A, C) remains observable after removing any set of $2|K|$ rows of C (that is, any set of $2|K|$ sensors) [11, Propositions 3.2 and 3.3]. If this strong observability assumption is not met, or in case of state attacks on (regular or singular) systems, a natural approach is to apply convex relaxation approaches to the optimization problem (3). Cardinality minimization problems of the form $\min_{v\in\mathbb{R}^n} \text{supp}(y - Av)$

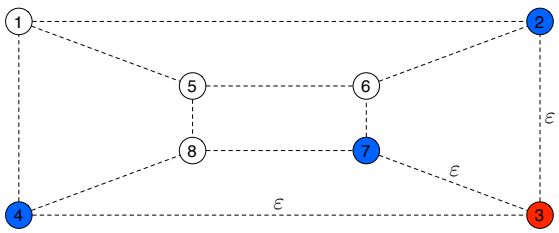

Fig. 1. A consensus system (A, B, C), where the state variable 3 is corrupted by the attacker, and the state variables 2, 4, and 7 are directly measured. Due to the sparsity pattern of (A, B, C) any attack of cardinality one is *generically* detectable and identifiable; see [4], [16] for further details.

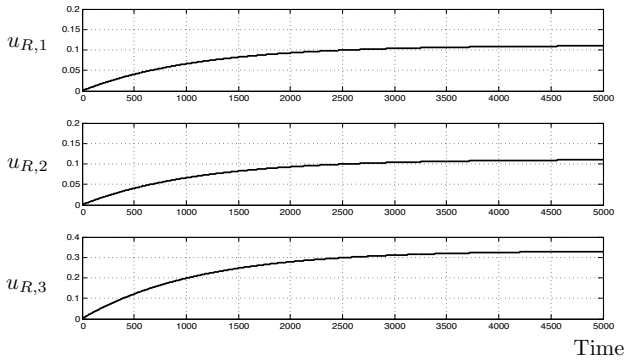

Fig. 2. In Fig. 2 we plot the attack mode u_R for the attack set $R = \{2, 4, 7\}$ to generate the same output as the attack set $K = \{3\}$ with attack mode $u_K = 1$. Although $|R| > |K|$, we have that $|u_{R,i}(t)| < |u_K(t)|/3$ for $i \in \{1, 2, 3\}$.

can often be efficiently solved via the ℓ_1 regularization $\min_{v\in\mathbb{R}^n} \|y - Av\|_{\ell_1}$ [15]. This procedure can be adapted to problem (3) after converting it into an algebraic problem, for instance by taking subsequent derivatives of the output y, or by discretizing the continuous-time system (1) and recording several measurements. As shown in [10], for discrete-time systems the ℓ_1 regularization performs reasonably well in the presence of output attacks. However, in the presence of state attacks such an ℓ_1 relaxation may perform poorly. In what follows we present a intuition explaining why this approach may fail, particularly in large-scale systems.

Example 1: ***(Ineffectiveness of regularization methods for sufficiently distant attacks)*** Consider a consensus system with underlying network graph (reflecting the sparsity pattern of A) illustrated in Fig. 1. In our model (1), the system matrix A is parameterized by $0 < \varepsilon \ll 1$, and given by the Laplacian

$$A = \begin{bmatrix} -0.8 & 0.1 & 0 & 0.2 & 0.5 & 0 & 0 & 0 \\ 0.1 & -0.4-\varepsilon & \varepsilon & 0 & 0 & 0.3 & 0 & 0 \\ 0 & 3\varepsilon & -9\varepsilon & 0 & 0 & 0 & 6\varepsilon & 0 \\ 0.1 & 0 & \varepsilon & -0.5-\varepsilon & 0 & 0 & 0 & 0.4 \\ 0.1 & 0 & 0 & 0 & -0.6 & 0.2 & 0 & 0.3 \\ 0 & 0.4 & 0 & 0 & 0.1 & -0.6 & 0.1 & 0 \\ 0 & 0 & 3\varepsilon & 0 & 0 & 0.4 & -0.6-3\varepsilon & 0.2 \\ 0 & 0 & 0 & 0.3 & 0.2 & 0 & 0.2 & -0.7 \end{bmatrix}.$$

Let the measurement matrix and the attack signature be

$$C = \begin{bmatrix} 0\,1\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,1\,0 \end{bmatrix}, \; B_K = \begin{bmatrix} 0\,0\,1\,0\,0\,0\,0\,0 \end{bmatrix}^{\mathsf{T}},$$

and define the transfer matrix $G_K(s) = C(sI - A)^{-1}B_K$. It can be verified that the state attack $K = \{3\}$ is identifiable.

Consider also the state attack $R = \{2, 4, 7\}$ with signature

$$B_R^{\mathsf{T}} = \begin{bmatrix} 0\,1\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,1\,0 \end{bmatrix} = C,$$

and define the transfer matrix $G_R(s) = C(sI - A)^{-1}B_R$. Let $U_K(s)$ and $U_R(s)$ be the Laplace transforms of $u_K(t)$ and $u_R(t)$, respectively. Notice that $G_R(s)$ is right-invertible [9]. Thus, letting G_R^{-r} denote the right-inverse of G_R,

$$Y(s) = G_K(s)U_K(s) = G_R(s)\left(G_R^{-r}(s)G_K(s)U_K(s)\right).$$

In other words, the measurements $Y(s)$ generated by the attack $U_K(s)$ can equivalently be generated by the attack

$$U_R(s) = G_R^{-r}(s)G_K(s)U_K(s).$$

Notice that $3 = \|u_R\|_{\mathcal{L}_0} > \|u_K\|_{\mathcal{L}_0} = 1$, that is, the attack set K achieves a lower cost than R in the problem (3).

Consider now the numerical realization with $\varepsilon = 0.0001$, $x(0) = 0$, and $u_K(t) = 1$ for all $t \in \mathbb{R}_{\geq 0}$. The corresponding attack signal $u_R = [u_{R,1}\ u_{R,2}\ u_{R,3}]$ is shown in Fig. 2. Observe that

$$\|u_K(t)\|_{\ell_p} > \|u_R(t)\|_{\ell_p}$$

holds point-wise in time for all integers $p \geq 1$. We also have

$$\|u_K(t)\|_{\mathcal{L}_q/\ell_p} > \|u_R(t)\|_{\mathcal{L}_p/\ell_q}$$

for any integers $p, q \geq 1$ and with the $\mathcal{L}_q/\ell_p$ signal norm

$$\|u_K(t)\|_{\mathcal{L}_q/\ell_p} = \left(\int_0^\infty \|u_K\|_p^q d\tau\right)^{1/q}.$$

Hence, the attack set R achieves a lower cost than K for any version of the optimization problem (3) penalizing a ℓ_p cost point-wise in time or a $\mathcal{L}_q/\ell_p$ cost over a time interval. On the other hand, we have $\|u_R\|_{\mathcal{L}_0} > \|u_K\|_{\mathcal{L}_0}$. We conclude that, in general, the identification problem cannot be solved by a point-wise ℓ_p or $\mathcal{L}_q/\ell_p$ regularization for any $p, q \geq 1$. Finally, we remark that for any choice of network parameters, a value of ε can be found such that a point-wise ℓ_p or a $\mathcal{L}_q/\ell_p$ regularization procedure fails at identifying the attack set.□

We emphasize that Example 1 is not of pathological nature, but large-scale stable systems often exhibit this behavior independently of the system parameters for attacks which are "sufficiently distant" from the sensors.

III. The Distributed Identification Problem

The obstacles and pitfalls in the centralized attack identification problem motivate our study of *divide-and-conquer* methods. In this section, we design distributed attack identification algorithms with performance guarantees, requiring low computational cost and local knowledge of the system.

A. Distributed setup and notation

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the undirected graph associated with the matrix A, where the vertex set $\mathcal{V} = \{1, \ldots, n\}$ corresponds to the system states, and the set of edges $\mathcal{E} = \{(i, j) : a_{ij} \neq 0\}$ is induced by the sparsity pattern of A; see also [4, Section IV]. Assume that $\mathcal{V}$ is partitioned into N disjoint subsets as $\mathcal{V} = \mathcal{V}_1 \cup \cdots \cup \mathcal{V}_N$, with $|\mathcal{V}_i| = n_i$, and let $\mathcal{G}^i = (\mathcal{V}_i, \mathcal{E}_i)$ be the i-th subgraph of $\mathcal{G}$ with vertices $\mathcal{V}_i$ and

edges $\mathcal{E}_i = \mathcal{E} \cap (\mathcal{V}_i \times \mathcal{V}_i)$. According to this partition, and possibly after relabeling the states, the system matrix A in (1) can be written as

$$A = \begin{bmatrix} A_1 & \cdots & A_{1N} \\ \vdots & \vdots & \vdots \\ A_{N1} & \cdots & A_N \end{bmatrix} = A_D + A_C,$$

where $A_i \in \mathbb{R}^{n_i \times n_i}$, $A_{ij} \in \mathbb{R}^{n_i \times n_j}$, A_D is block-diagonal, and $A_C = A - A_D$. Notice that, if $A_D = \text{blkdiag}(A_1, \ldots, A_N)$, then A_D represents the isolated subsystems and A_C describes the interconnection structure among the subsystems. Additionally, if the original system is sparse, then several blocks in A_C vanish.

We assume the following on the subsystem decomposition:

(A1) the matrix $C = \text{blkdiag}(C_1, \ldots, C_N)$ is block-diagonal with $C_i \in \mathbb{R}^{p_i \times n_i}$, and each pair (A_i, C_i) is observable.

Let $\mathcal{N}_i = \{j \in \{1, \ldots, N\} \backslash \{i\} \, : \, \|A_{ij}\| \neq 0 \text{ or } \|A_{ji}\| \neq 0\}$ be the neighbors of subsystem i, and let $\mathcal{N}_i^k$ be the set of neighbors at distance k from i, with subsystem i excluded. Each subsystem $\mathcal{G}^i$ has a *control center* with the following capabilities:

(A2) the i-th control center knows the matrices A_i and C_i, as well as the neighboring matrices A_{ij}, $j \in \mathcal{N}_i$; and

(A3) the i-th control center can transmit an estimate of its state to the j-th control center if $j \in \mathcal{N}_i$.

Given the above structure, the system (1) can be written as the interconnection of N subsystems of the form

$$\begin{aligned} \dot{x}_i(t) &= A_i x_i(t) + \sum_{j \in \mathcal{N}_i} A_{ij} x_j(t) + B_{K_i} u_{K_i}(t), \\ y_i(t) &= C_i x_i(t) + D_{K_i} u_{K_i}(t), \;\; i \in \{1, \ldots, N\}. \end{aligned} \tag{4}$$

Here $K_i = (K \cap V_i) \cup K_i^{\text{p}}$ is the attack set in region $\mathcal{G}^i$, and K_i^{p} is the set of corrupted measurements in region $\mathcal{G}^i$. Clearly, if the inter-subsystem signals $A_{ij}x_j$ are known or directly measured, then the regional attack identification problem within each subsystem reduces to the centralized problem. We will not make this assumption since it is restrictive and precludes the case that the inter-subsystem signals $A_{ij}x_j$ themselves are corrupted by an attacker.

B. Fully decoupled attack identification

As a first low-complexity identification method we consider the fully decoupled case (no cooperation among control centers). In the spirit of fully decentralized state estimation [12], the neighboring states x_j affecting x_i are treated as unknown inputs f_i to the i-th subsystem, and equation (4) becomes

$$\begin{aligned} \dot{x}_i(t) &= A_i x_i(t) + B_i^{\text{b}} f_i(t) + B_{K_i} u_{K_i}(t), \\ y_i(t) &= C_i x_i(t) + D_{K_i} u_{K_i}(t), \;\; i \in \{1, \ldots, N\}, \end{aligned} \tag{5}$$

where $B_i^{\text{b}} = [A_{i1} \cdots A_{i,i-1} \, A_{i,i+1} \cdots A_{iN}]$. We refer to (5) as to the *i-th decoupled system*, and we let $V_i^{\text{b}} \subseteq \mathcal{V}_i$ be the set of *boundary nodes* of (5), that is, the nodes $j \in \mathcal{V}_i$ with $A_{jk} \neq 0$ for some $k \in \{1, \ldots, n\} \setminus \mathcal{V}_i$. Due to partitioning, control centers are able to perform attack identification only on local subsystems.

Under certain identifiability assumptions (see Theorem 3.1 below), the i-th control center can uniquely identify the attack set K_i by again implementing a set of residual generators (e.g., see [17]), where the residual output is insensitive to the attack matrices B_{R_i} and D_{R_i} (for every attack set R_i of cardinality $|K_i|$), and to the boundary inputs B_i^{b}, which are here considered as unknown inputs. Observe that the i-th control center needs to construct $\binom{n_i}{|K_i|}$ residual filters, as opposed to $\binom{n}{|K|}$, with $|K| = \sum_{i=1}^N |K_i|$, in the centralized case. The reduction of the combinatorial logic comes at the expenses of having a restricted set of identifiable attacks due to unknown boundary inputs B_i^{b}.

Theorem 3.1: ***(Fully decoupled attack identification)*** For the partitioned system (5) and an attack set K, the following statements are equivalent:

(i) the attack set K in (5) is unidentifiable by the fully decoupled identification algorithm; and

(ii) for some region $i \in \{1, \ldots, N\}$ with $K_i \neq \emptyset$, there exists an attack set R_i, with $|R_i| \leq |K_i|$ and $R_i \neq K_i$, so that the system $(A, [B_i^{\text{b}} \; B_{K_i} \; B_{R_i}], C, [D_{K_i} \; D_{R_i}])$ has invariant zeros.

Proof: Let $y_i(x_{i,0}, u_{K_i}, f_i, t)$ denote the output of system (5) at time t, with initial value $x_{i,0}$, attack input u_{K_i}, and boundary input f_i. Notice that the boundary input f_i is considered arbitrary and unknown by the fully decoupled attack identification method. The attack set K_i is undistinguishable from R_i if and only if $y(x_{i,0}, u_{K_i}, f_i, t) = y(x_{i,1}, u_{R_i}, h_i, t)$ at all times t, for some initial conditions $x_{i,0}$ and $x_{i,1}$, attack inputs u_{K_i} and u_{R_i}, and boundary inputs f_i and h_i. Due to linearity of the system, K_i is unidentifiable if and only if $y(x_{i,0} - x_{i,1}, u_{K_i} - u_{R_i}, f_i - h_i, t) = 0$ at all times, which can be guaranteed if and only if the quadruple $(A, [B_i^{\text{b}} \; B_{K_i} \; B_{R_i}], C, [D_{K_i} \; D_{R_i}])$ features invariant zeros [9]. In the absence of invariant zeros, identifiability of attacks is ensured as in [4] for the centralized case. ■

By comparing Theorems 2.1 and 3.1 we conclude that, with the fully decoupled identification procedure, the i-th control center cannot distinguish between an unknown input from a safe subsystem, an unknown input from a corrupted subsystem, and a boundary attack with the same input direction.

*Corollary 3.2 (***Limitation of decoupled algorithm):** The following statements hold for the partitioned system (5) with the fully decoupled identification algorithm:

(L1) any (boundary) attack set $K_i \subseteq V_i^{\text{b}}$ is not identifiable by the i-th control center (in fact K_i is not detectable[2]), and

(L2) any (external) attack set $K \setminus K_i$ is not identifiable by the i-th control center (in fact K_i is not detectable).

C. Cooperative attack identification

In this section we improve upon the fully decoupled method presented in Subsection III-B and propose an identifi-

[2]An attack is detectable if it can be distinguished from the zero attack [4].

cation method based on a *divide-and-conquer* procedure with cooperation. In particular, the identification method proposed in this section differs from the fully decoupled identification method because control centers are allowed to exchange their estimates, which reduces the uncertainty due to unknown boundary inputs. In developing our method, we implicitly assume that communication among control centers is secure.

Our cooperative identification method is informally described as follows. First, control centers independently estimate the state of their local region subject to unknown inputs from the neighboring regions. Because of the presence of unknown inputs, the estimation computed by a control center is correct modulo some *uncertainty subspace*. Control centers exchange their estimate and the corresponding uncertainty subspaces. Second, control centers check the compatibility between their estimate and those received from the neighboring regions. Third, if the received estimates are not compatible with local estimates, then the system is recognized under attack. Finally, control centers implement a local attack identification procedure by leveraging local system parameters and estimates, and estimates received from their neighbors.

We next detail our cooperative identification method.

(S1: local state estimation) Each control center estimates the state of its own region by means of an *unknown-input observer* for the i-th subsystem subject to the unknown input $B_i^{\mathrm{b}} f_i$. We adopt the technique in [9, Section 4.3.1], which exploits the derivatives of the output signal to reconstruct the system state instantaneously. In the presence of process and measurement noise, different methods should be used.

Assume that the state x_i can be reconstructed modulo some subspace $\mathcal{F}_i$.[3] Let $F_i = \mathrm{Basis}(\mathcal{F}_i)$ be the uncertainty matrix, and partition the state accordingly as

$$x_i = \hat{x}_i + \tilde{x}_i, \tag{6}$$

where $\hat{x}_i(t) \perp \mathcal{F}_i$ is the portion of the state that can be estimated by the i-th control center in the presence of the unknown input $B_i^{\mathrm{b}} f_i$, and $\tilde{x}_i(t) \in \mathcal{F}_i$. Let $z_i(t)$ be the *estimate* at time t of $\hat{x}_i$. Notice that, if the i-th region is not corrupted, then $z_i(t) = \hat{x}_i(t)$, whereas it may be $z_i(t) \neq \hat{x}_i(t)$ when $K_i \neq \emptyset$.

(S2: communication) Control centers transmit their estimate $\hat{x}_i$ and uncertainty matrix F_i to every neighboring control center.

(S3: regional attack detection) Observe that

$$A_{ij} x_j = A_{ij} \hat{x}_j + A_{ij} \tilde{x}_j,$$

where $\hat{x}_j$ and $\tilde{x}_j$ are defined as in (6). After carrying out step (S1), since the matrices A_{ij} are known to the i-th control center due to Assumption (A6), only the inputs $A_{ij}\tilde{x}_j$ are unknown to the i-th control center, while the inputs $A_{ij}\hat{x}_j$ are known to the i-th center due to communication. Let

$$B_i^{\mathrm{b}} F_i = \begin{bmatrix} A_{i1}F_1 & \cdots & A_{i,i-1}F_{i-1} & A_{i,i+1}F_{i+1} & \cdots & A_{iN}F_N \end{bmatrix},$$

and rewrite the signal $B_i^{\mathrm{b}}\tilde{x}$ as $B_i^{\mathrm{b}}\tilde{x} = B_i^{\mathrm{b}} F_i f_i$, for some unknown signal f_i. The dynamics of the i-th subsystem read as

$$\dot{x}_i(t) = A_i x_i(t) + B_i^{\mathrm{b}} \hat{x}(t) + B_i^{\mathrm{b}} F_i f_i(t) + B_{K_i} u_{K_i}(t),$$

where $\hat{x}$ is the vector of $\hat{x}_i$ for all $i \in \{1, \ldots, N\}$.

Next, we construct a residual generator that is insensitive to the input $B_i^{\mathrm{b}} F_i f_i$ (e.g., see [17]) and makes use of the state estimates $B_i^{\mathrm{b}} z$ transmitted to control center i by its neighbors:

$$\begin{aligned} \dot{w}_i(t) &= (A_i + L_i C_i) w_i(t) - L_i y_i(t) + B_i^{\mathrm{b}} z(t), \\ r_i(t) &= M w_i(t) - H y_i(t). \end{aligned} \tag{7}$$

(S4: cooperative attack identification) Neighboring control centers exchange the zero/nonzero status of the previously computed residuals, identify corrupted regions, and independently identify attacks in each attacked region. Our cooperative identification procedure for the i-th control center is summarized in Algorithm 1. We make the following technical assumptions:

(A4) corrupted regions have one neighbor at distance 2, that is, $|\mathcal{N}_i^2| \geq 1$ for all regions i with $K_i \neq \emptyset$, and

(A5) corrupted regions are separated by 3 non-corrupted regions, that is, $K_j = \emptyset$ for all $j \in \mathcal{N}_i^3$ and i with $K_i \neq \emptyset$.

Assumption (A4) requires a sufficiently large number of clusters, while assumption (A5) restricts our procedure to localized attacks. The next theorem characterizes the effectiveness of our cooperative identification procedure.

Theorem 3.3: ***(Cooperative attack identification)*** For the partitioned system (4), the attack set K is identifiable by the cooperative identification algorithm if the following conditions hold:

(i) every system $(A_i, B_i^{\mathrm{b}}, C_i)$ has no invariant zeros, and

(ii) every system $(A_i, [B_i^{\mathrm{b}} F_i \;\; B_{K_i} \;\; B_{R_i}], C_i, [D_{K_i} \;\; D_{R_i}])$ has no invariant zeros for every attack set R_i with $|R_i| \leq |K_i|$.

In Theorem 3.3, conditions (i) with assumptions (A4) and (A5) ensure *regional identifiability*, that is, the possibility to identify corrupted regions from local residuals and communication with neighboring regions. Condition (ii) ensures *regional detectability* as (7) and *local identifiability*, that is,

Algorithm 1: *Cooperative attack identification*

Input : Matrices A_i, A_{ij} for $j \in \mathcal{N}_i$;
Require : Conditions (i), (ii), and (iii) in Theorem 3.3;
Output : Attack set K_i;

Compute the uncertainty subspace $\mathcal{F}_i = \mathrm{Im}(F_i)$;
Transmit F_i to control centers $\mathcal{N}_i$;
while *True* **do**
  Estimate state $\hat{x}_i$ (state x modulo $\mathcal{F}_i$);
  Transmit $\hat{x}_i$ to $\mathcal{N}_i$, and receive $\hat{x}_j$ from $\mathcal{N}_i$;
  Compute residual r_i as in (7);
  Transmit r_i to $\mathcal{N}_i$, and receive r_j from $\mathcal{N}_i$;
  if $r_i \neq 0$ *or* $r_j \neq 0$ *for all* $j \in \mathcal{N}_i$ **then**
    Identify K_i in local subsystem;
    return K_i

[3]For nonsingular systems without feedthrough matrix, $\mathcal{F}_i$ is as small as the largest (A_i, B_i^{b})-controlled invariant subspace contained in $\mathrm{Ker}(C_i)$ [9].

attack identifiability within each corrupted region from local measurements and communication with neighboring regions. We defer the proof of Theorem 3.3 to VI-A.

We conclude this section with the following observations. First, the cooperative identification procedure is implemented only on the corrupted regions (line 7 in Algorithm 1). Thus, the combinatorial complexity of our distributed identification procedure is $\sum_{i=1}^{\ell} \binom{n_i+p_i}{|K_i|}$, where ℓ is the number of corrupted regions. Hence, the distributed identification procedure greatly reduces the combinatorial complexity of the centralized procedure presented in [4] that requires the implementation of $\binom{n+p}{|K|}$ filters. Second, the conditions in Theorem 3.3 for cooperative identification improve upon those in Theorem 3.1 for fully decoupled identification; see Section IV for an example. Third, our cooperative identification procedure is effective when attacks are localized in some regions, and regions under attack are sufficiently far from each other. Under these assumptions, our cooperative identification overcomes the limitations described in Example 1, because it does not rely on the magnitude of the measurements, and has provable performance guarantees.

*Remark 1 (***Extension to descriptor systems):** The proposed methods can be extended to descriptor systems

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{8}$$

where $E \in \mathbb{R}^{n\times n}$ is a singular matrix. Descriptor models are encountered in cyber-physcial systems with conservation laws such as power, gas, or water networks [4]. The following assumptions guarantee existence of non-impulsive solutions [18]:

(A6) the pair (E, A) is regular, that is, the determinant $|sE - A|$ is nonzero for almost all values $s \in \mathbb{C}$;

(A7) the initial condition $x(0) \in \mathbb{R}^n$ is consistent, that is, $(Ax(0) + Bu(0)) \in \mathrm{Im}(E)$; and

(A8) the input u is smooth.

The characterization of identifiability (2) extends to descriptor systems [4, Theorem 3.4], residual filters can be designed as in [4, Section IV.D], and the state estimation step (S1) can be extended akin to [19, Appendix]. □

IV. Illustrative example

We now present an example showing that, contrary to the limitations of the naive fully decoupled approach (see Corollary 3.2), boundary attacks $K_i \subseteq V_i^{\mathrm{b}}$ can be identified by our cooperative attack identification method.

Consider the sensor network in Fig. 3, where the state of the blue nodes $\{2, 5, 7, 12, 13, 15, 19, 22, 23\}$ is measured and the state of the red node $\{3\}$ is corrupted by an attacker. Assume that the network evolves according to linear time-invariant dynamics. Assume further that the network has been partitioned into the three areas $\mathcal{V}_1 = \{1, \dots, 8\}$, $\mathcal{V}_2 = \{9, \dots, 16\}$, and $\mathcal{V}_3 = \{17, \dots, 24\}$. Since $\{3, 4\}$ are the boundary nodes for the first area, the attack set $K = 3$ is not identifiable via the fully decoupled procedure in Section III-B; see Corollary 3.2. It can be verified that the conditions in Theorem 3.3 are verified for generic network parameters [4, Section III.B], and that the attack can be identified via our cooperative identification procedure. We conclude that our cooperative identification algorithm outperforms the naive decoupled identification algorithm.

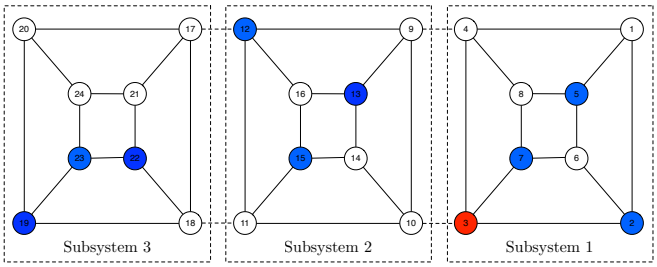

Fig. 3. This figure shows a network composed of three subsystems. A control center is assigned to each subsystem. Each control center knows the local dynamics. The state of the blue nodes $\{2, 5, 7, 12, 13, 15\}$ is continuously measured by the corresponding control center, and the state of the red node $\{3\}$ is corrupted by an attacker. The decoupled identification procedure presented in Subsection III-B fails at detecting the attack. Instead, our cooperative identification procedure identifies the corrupted agent.

V. Conclusion

The problem of identifying attacks in cyber-physical systems requires a substantial computational effort. This paper shows how standard relaxation techniques may fail to identify state attacks in cyber-physical systems, and proposes two distributed algorithms with performance guarantees for attack identification by a set of geographically deployed control centers. The algorithms require local measurements, local knowledge of the system, and communication with neighboring control centers. This paper provides promising results on the distributed attack identification problem, highlights its challenges and limitations, and foster the adoption of geometric control techniques for the solution of distributed control and estimation problems.

VI. Appendix

A. Proof of Theorem 3.3

We start with some preliminary results.

Lemma 6.1: ***(Residual of isolated non-corrupted regions)*** If $K_i = \emptyset$ and $K_j = \emptyset$ for all $j \in \mathcal{N}_i$, then the residual $r_i(t)$ in (7) is identically zero.

Proof: Consider a region j with $K_j = \emptyset$. Notice that the state estimation z_j satisfies $z_j = \hat{x}_j$ (see paragraph **(S1: local state estimation)**). Because $K_i = \emptyset$, from (7) we conclude that the residual r_i is driven only by the input $B_i^{\mathrm{b}} F_i f_i$. To conclude, notice that the residual generator (7) is constructed so that r_i is insensitive to the signature $(B_i^{\mathrm{b}} F_i, 0)$. ■

Lemma 6.2: ***(Residual of isolated corrupted regions)*** For the partitioned system (4), let $K_i \neq \emptyset$. If

(i) $K_j = \emptyset$ for all $j \in \mathcal{N}_i^2$,

(ii) every system $(A_j, B_j^{\mathrm{b}}, C_j)$ has no invariant zeros for all $j \in \mathcal{N}_i$, and

(iii) the system $(A_i, [B_i^{\mathrm{b}} F_i \; B_{K_i} \; B_{R_i}], C_i, [D_{K_i} \; D_{R_i}])$ has no invariant zeros for every attack set $R_i \neq K_i$ with $|R_i| \leq |K_i|$,

then

(i) $r_i(t) \neq 0$ at some time $t \in \mathbb{R}_{\geq 0}$, and

(ii) either $r_j(t) = 0$ for all $j \in \mathcal{N}_i$ at all times $t \in \mathbb{R}_{\geq 0}$, or $r_j(t) \neq 0$ for all $j \in \mathcal{N}_i$ at some times $t \in \mathbb{R}_{\geq 0}$.

Proof: The estimation computed by a control center is correct if its area is not under attack (see paragraph **(S1: local state estimation)**). In other words, since $K_j = \emptyset$ for all $j \in \mathcal{N}_i$, it follows $B_i^{\mathsf{b}} z = B_i^{\mathsf{b}} \hat{x}$ in (7). Because $(A_i, [B_i^{\mathsf{b}} F_i \; B_{K_i}], C_i, [D_{K_i} \;\; D_{R_i}])$ has no invariant zeros, the set K_i is locally identifiable via local measurements and transmitted estimates, and statement (i) follows; see also Theorem 2.1 and [4]. To show the second statement, observe that only two cases are possible: either $\hat{x}_i = z_i$, or $\hat{x}_i \neq z_i$, where $\hat{x}_i$ is defined in (6), and z_i is the estimate of $\hat{x}_i$ computed by the i-th control center. For instance, if $\mathrm{Im}(B_{K_i}) \subseteq \mathrm{Im}(B_i^{\mathsf{b}})$, that is, the attack set K_i lies on the boundary of the i-th area, then $\hat{x}_i(t) = z_i(t)$.

In the first case, $\hat{x}_i = z_i$, all residuals r_j, $j \in \mathcal{N}_i$, are identically zero. In fact, since $K_\ell = \emptyset$ for all $\ell \in \mathcal{N}_i^2$, it follows that $\hat{x}_p = z_p$ for all $p \in \mathcal{N}_j$ and $j \in \mathcal{N}_i$, so that the residual r_j in (7) is identically zero, as it is insensitive to the unknown inputs B_p^{b}. Consider now the second case: $\hat{x}_i \neq z_i$. Notice that $B_j^{\mathsf{b}} F_j f_j + B_j^{\mathsf{b}}(\hat{x} - z) \in \mathrm{Im}(B_j^{\mathsf{b}})$. Since $(A_j, B_j^{\mathsf{b}}, C_j)$ has no invariant zeros, every residual r_j, with $j \in \mathcal{N}_i$, cannot be identically zero. ∎

We are now ready to prove Theorem 3.3.

Proof: Consider the i-th region, and let $K_i \neq \emptyset$. Due to conditions (i) and (ii) in Theorem 3.3, assumptions (A4) and (A5), and Lemma 6.2 we conclude that:

(C1) the residual r_i is not identically zero, and

(C2) either $r_j(t) = 0$ for all $j \in \mathcal{N}_i$ at all times $t \in \mathbb{R}_{\geq 0}$, or $r_j(t) \neq 0$ for all $j \in \mathcal{N}_i$ at some times $t \in \mathbb{R}_{\geq 0}$.

Consider the region $p \in \mathcal{N}_i^2 \setminus \mathcal{N}_i$. Due to assumption (A5) and Lemma 6.1 we conclude that:

(C3) r_p is identically zero.

Consider the region $j \in \mathcal{N}_i$. Due assumption (A4) and the facts (C1) and (C3), we conclude that:

(C4) there exists $j_1, j_2 \in \mathcal{N}_j$ such that r_{j_1} is identically zero, while r_{j_2} is not identically zero (take $j_1 = p$ and $j_2 = i$.).

Corrupted regions are uniquely identified as the regions satisfying (C1) and (C2). See Figure 4 for an example. Due to condition (ii) in Theorem 3.3 each set K_i is locally identifiable (see also Theorem 2.1), and the statement follows. Note that (A4) is necessary. To see this, in Fig. 4 assume that region 7 is corrupted, and that the residual computed by the 8-th control center is nonzero. Since the residual computed by the 7-th control center is also nonzero, the 8-th control center cannot determine if region 7 or 8 is corrupted. ∎

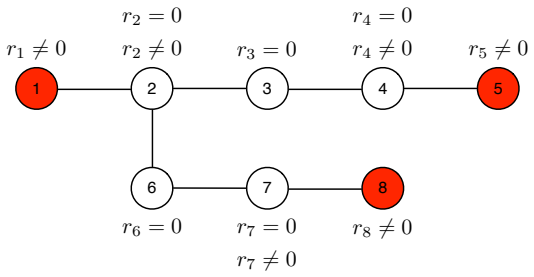

Fig. 4. An example of regional identification. The zero/nonzero pattern of the residuals computed by the control centers is reported: compromised regions $\{1, 5, 8\}$ have nonzero residuals; non-compromised regions $\{3, 6\}$ have zero residuals, because neighboring regions $\{2, 4, 7\}$ are not compromised; regions 2, 4, and 7 may have zero or nonzero residuals, depending on the strategy of the attacker in region 1, 5, and 8, respectively. Non-compromised regions $\{2, 3, 4, 6, 7\}$ are identified by Algorithm 1 as those with zero residual, and those with zero and nonzero neighboring residuals. Compromised regions $\{1, 5, 8\}$ are identified by exclusion.

References

[1] J. Slay and M. Miller, "Lessons learned from the Maroochy water breach," *Critical Infrastructure Protection*, vol. 253, pp. 73–82, 2007.

[2] S. Kuvshinkova, "SQL Slammer worm lessons learned for consideration by the electricity sector," *North American Electric Reliability Council*, 2003.

[3] J. P. Farwell and R. Rohozinski, "Stuxnet and the future of cyber war," *Survival*, vol. 53, no. 1, pp. 23–40, 2011.

[4] 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.

[5] Y. Liu, M. K. Reiter, and P. Ning, "False data injection attacks against state estimation in electric power grids," in *ACM Conference on Computer and Communications Security*, Chicago, IL, USA, Nov. 2009, pp. 21–32.

[6] R. Smith, "A decoupled feedback structure for covertly appropriating network control systems," in *IFAC World Congress*, Milan, Italy, Aug. 2011, pp. 90–95.

[7] L. Xie, Y. Mo, and B. Sinopoli, "False data injection attacks in electricity markets," in *IEEE Int. Conf. on Smart Grid Communications*, Gaithersburg, MD, USA, Oct. 2010, pp. 226–231.

[8] F. Dörfler, F. Pasqualetti, and F. Bullo, "Continuous-time distributed observers with discrete communication," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 2, pp. 296–304, 2013.

[9] G. Basile and G. Marro, *Controlled and Conditioned Invariants in Linear System Theory*. Prentice Hall, 1991.

[10] F. Hamza, P. Tabuada, and S. Diggavi, "Secure state-estimation for dynamical systems under active adversaries," in *Allerton Conf. on Communications, Control and Computing*, Monticello, IL, USA, Sep. 2011, pp. 337–344.

[11] Y. Shoukry and P. Tabuada, "Event-triggered state observers for sparse sensor noise/attacks," *arXiv preprint arXiv:1309.3511*, 2013.

[12] M. Saif and Y. Guan, "Decentralized state estimation in large-scale interconnected dynamical systems," *Automatica*, vol. 28, no. 1, pp. 215–219, 1992.

[13] H. Nishino and H. Ishii, "Distributed detection of cyber attacks and faults for power systems," pp. 11 932–11 937, Aug. 2014.

[14] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Allerton Conf. on Communications, Control and Computing*, Monticello, IL, USA, Sep. 2010, pp. 911–918.

[15] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[16] 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.

[17] F. Pasqualetti, F. Dörfler, and F. Bullo, "Cyber-physical attacks in power networks: Models, fundamental limitations and monitor design," Sep. 2011, available at `http://arxiv.org/abs/1103.2795`.

[18] T. Geerts, "Invariant subspaces and invertibility properties for singular systems: The general case," *Linear Algebra and its Applications*, vol. 183, pp. 61–88, 1993.

[19] 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.