Attack Detection and Identification in Cyber-Physical Systems – Part II: Centralized and Distributed Monitor Design

Fabio Pasqualetti, Florian Dörfler, and Francesco Bullo

***Abstract*—Cyber-physical systems integrate computation, communication, and physical capabilities to interact with the physical world and humans. Besides failures of components, cyber-physical systems are prone to malicious attacks so that specific analysis tools and monitoring mechanisms need to be developed to enforce system security and reliability. This paper builds upon the results presented in our companion paper [1] and proposes centralized and distributed monitors for attack detection and identification. First, we design optimal centralized attack detection and identification monitors. Optimality refers to the ability of detecting (respectively identifying) *every* detectable (respectively identifiable) attack. Second, we design an optimal distributed attack detection filter based upon a waveform relaxation technique. Third, we show that the attack identification problem is computationally hard, and we design a sub-optimal distributed attack identification procedure with performance guarantees. Finally, we illustrate the robustness of our monitors to system noise and unmodeled dynamics through a simulation study.**

I. Introduction

Cyber-physical systems need to remain functional and operate reliably in presence of unforeseen failures and, possibly, external attacks. Besides failures and attacks on the physical infrastructure, cyber-physical systems are also prone to cyber attacks against their data management, control, and communication layer [2], [3], [4], [5].

In several cyber-physical systems, including water and gas distribution networks, electric power systems, and dynamic Leontief econometric models, the physical dynamics include both differential equations as well as algebraic constraints. In [1] we model cyber-physical systems under attack by means of linear continuous-time differential-algebraic systems; we analyze the fundamental limitations of attack detection and identification, and we characterize the vulnerabilities of these systems by graph-theoretic methods. In this paper we design monitors for attack detection and identification for the cyber-physical model presented in [1].

Related work. Concerns about security of control, communication, and computation systems are not recent as testified by the numerous works in the fields of fault-tolerance control and information security. However, as discussed in [1], cyber-physical systems feature vulnerabilities beyond fault-tolerance control and information security methods.

This material is based upon work supported in part by NSF grant CNS-1135819 and by the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the U.S. Army Research Office.

Fabio Pasqualetti, Florian Dörfler, and Francesco Bullo are with the Center for Control, Dynamical Systems and Computation, University of California at Santa Barbara, {fabiopas,dorfler,bullo}@engineering.ucsb.edu

Attack detection and identification monitors have recently been proposed. In [6], [7] monitoring procedures are designed for the specific case of state attacks against discrete-time nonsingular systems. In [8] an algorithm to detect output attacks against discrete-time nonsingular systems is described and characterized. In [9] a detection scheme for replay attacks is proposed. Fault detection and identification schemes for linear differential-algebraic power network models are presented in [10], [11] and in the conference version of this paper [12]. We remark that the designs in [10], [11] consider particular known faults rather than unknown and carefully orchestrated cyber-physical attacks. Finally, protection schemes for output attacks against systems described by purely static models are presented, among others, in [13], [14].

Contributions. The main contributions of this work are as follows. First, for the differential-algebraic model of cyber-physical systems under attacks developed in [1], we design centralized monitors for attack detection and identification. With respect to the existing solutions, in this paper we propose attack detection and identification filters that are effective against both state and output attacks against linear continuous-time differential-algebraic cyber-physical systems. Our monitors are designed by using tools from geometric control theory; they extend the construction of [15] to descriptor systems with direct feedthrough matrix, and they are guaranteed to achieve optimal performance, in the sense that they detect (respectively identify) every detectable (respectively identifiable) attack.

Second, we develop a fully distributed attack detection filter with optimal (centralized) performance. Specifically, we provide a distributed implementation of our centralized attack detection filter based upon iterative local computations by using the Gauss-Jacobi waveform relaxation technique. For the implementation of this method, we rely upon cooperation among geographically deployed control centers, each one responsible for a part of the system. In particular, we require each control center to have access to the measurements of its local subsystem, synchronous communication among neighboring control centers at discrete time instants, and ability to perform numerical integration.

Third, we show that the attack identification problem is inherently computationally hard. Consequently, we design a distributed identification method that achieves identification, at a low computational cost and for a class of attacks, which can be characterized accurately. Our distributed identification methods is based upon a *divide and conquer* procedure, in which first corrupted regions and then corrupted components are identified by means of local identification procedures and cooperation among neighboring regions. Due to cooperation,

our distributed procedure provably improves upon the fully decoupled approach advocated in decentralized control [16].

Fourth, we present several illustrative examples. Besides illustrating our findings concerning centralized and distributed detection and identification, our numerical investigations show that our methods are effective also in the presence of system noise, nonlinearities, and modeling uncertainties.

Finally, as a minor contribution, we build upon the estimation method in [17] to characterize the largest subspace of the state space of a descriptor system that can be reconstructed in the presence of unknown inputs.

Paper organization. Section II contains a mathematical description of the problems under investigation. In Section III we design monitors for attack detection. Specifically, we propose optimal centralized, decentralized, and distributed monitors. In Section IV we show that the attack identification problem is computationally hard. Additionally, we design an optimal centralized and a sub-optimal decentralized attack identification monitor. Finally, Section V and Section VI contain, respectively, our numerical studies, and our conclusion.

II. Problem setup and preliminary concepts

In this section we recall the framework proposed in [1] for cyber-physical systems and attacks. We model a cyber-physical system under attack with the time-invariant descriptor system

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

where $x(t) \in \mathbb{R}^n$, $y(t) \in \mathbb{R}^p$, $E \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$. Here the matrix E is possibly singular, and the input terms $Bu(t)$ and $Du(t)$ are unknown signals describing disturbances affecting the plant. Besides reflecting the genuine failure of systems components, these disturbances model the effect of an attack against the cyber-physical system. For notational convenience and without affecting generality, we assume that each state and output variable can be independently compromised by an attacker. Thus, we let $B = \begin{bmatrix} I, 0 \end{bmatrix}$ and $D = \begin{bmatrix} 0, I \end{bmatrix}$ be partitioned into identity and zero matrices of appropriate dimensions, and, accordingly, $u(t) = \begin{bmatrix} u_x(t)^\mathsf{T}, u_y(t)^\mathsf{T} \end{bmatrix}^\mathsf{T}$. Hence, the unknown input $(Bu(t), Du(t)) = (u_x(t), u_y(t))$ can be classified as *state attack* affecting the system dynamics and as *output attack* corrupting directly the measurements vector.

The attack signal $t \mapsto u(t) \in \mathbb{R}^{n+p}$ depends upon the specific attack strategy. In the presence of $k \in \mathbb{N}_0$, $k \leq n+p$, attackers indexed by the *attack set* $K \subseteq \{1, \dots, n+p\}$ only and all the entries K of $u(t)$ are nonzero over time. To underline this sparsity relation, we sometimes use $u_K(t) \in \mathbb{R}^{|K|}$ to denote the *attack mode*, that is the subvector of $u(t)$ indexed by K. Accordingly, we use the pair (B_K, D_K), where B_K and D_K are the submatrices of B and D with columns in K, to denote the *attack signature*. Hence, $Bu(t) = B_K u_K(t)$, and $Du(t) = D_K u_K(t)$. We make the following assumptions on system (1), a discussion of which can be found in [1]:

(A1) the pair (E, A) is regular, that is, $\det(sE - A)$ does not vanish identically,

(A2) the initial condition $x(0) \in \mathbb{R}^n$ is consistent, that is, $(Ax(0) + Bu(0)) \perp \mathrm{Ker}(E^\mathsf{T}) = 0$; and

(A3) the input signal $u(t)$ is smooth.

The following definitions are inspired by our results in [1]. Let $y(x_0, u, t)$ be the output sequence generated from the initial state x_0 under the attack signal $u(t)$.

Definition 1: ***(Undetectable attack set)*** For the linear descriptor system (1), the attack set K is *undetectable* if there exist initial conditions $x_1, x_2 \in \mathbb{R}^n$, and an attack mode $u_K(t)$ such that, for all $t \in \mathbb{R}_{\geq 0}$, it holds $y(x_1, u_K, t) = y(x_2, 0, t)$.

Definition 2: ***(Unidentifiable attack set)*** For the linear descriptor system (1), the attack set K is *unidentifiable* if there exists an attack set R, with $|R| \leq |K|$ and $R \neq K$, initial conditions $x_K, x_R \in \mathbb{R}^n$, and attack modes $u_K(t)$, $u_R(t)$ such that, for all $t \in \mathbb{R}_{\geq 0}$, it holds $y(x_K, u_K, t) = y(x_R, u_R, t)$.

In our companion paper [1] we characterize undetectable and unidentifiable attacks. In this paper, instead, we design monitors to achieve attack detection and identification.

III. Monitor design for attack detection

A. Centralized attack detection monitor design

In the following we present a centralized attack detection filter based on a modified Luenberger observer.

Theorem 3.1: ***(Centralized attack detection filter)*** Consider the descriptor system (1) and assume that the attack set K is detectable, and that the network initial state $x(0)$ is known. Consider the *centralized attack detection filter*

$$\begin{aligned} E\dot{w}(t) &= (A + GC)w(t) - Gy(t), \\ r(t) &= Cw(t) - y(t), \end{aligned} \tag{2}$$

where $w(0) = x(0)$ and the output injection $G \in \mathbb{R}^{n\times p}$ is such that the pair $(E, A + GC)$ is regular and Hurwitz. Then $r(t) = 0$ at all times $t \in \mathbb{R}_{\geq 0}$ if and only if $u_K(t) = 0$ at all times $t \in \mathbb{R}_{\geq 0}$. Moreover, in the absence of attacks, the filter error $w(t) - x(t)$ is exponentially stable.

Proof: Consider the error $e(t) = w(t) - x(t)$ between the dynamic states of the filter (2) and the descriptor system (1). The error dynamics with output $r(t)$ are given by

$$\begin{aligned} E\dot{e}(t) &= (A + GC)e(t) - (B_K + GD_K)u_K(t), \\ r(t) &= Ce(t) - D_K u_K(t), \end{aligned} \tag{3}$$

where $e(0) = 0$. To prove the theorem we show that the error system (3) has no invariant zeros, that is, $r(t) = 0$ for all $t \in \mathbb{R}_{\geq 0}$ if and only if $u_K(t) = 0$ for all $t \in \mathbb{R}_{\geq 0}$. Since the initial condition $x(0)$ and the input $u_K(t)$ are assumed to be consistent (A2) and non-impulsive (A3), the error system (3) has no invariant zeros if and only if [18, Proposition 3.4] there exists no triple $(s, \bar{w}, g_K) \in \mathbb{C} \times \mathbb{R}^n \times \mathbb{R}^p$ satisfying

$$\begin{bmatrix} sE - (A + GC) & B_K + GD_K \\ C & -D_K \end{bmatrix} \begin{bmatrix} \bar{w} \\ g_K \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{4}$$

The second equation of (4) yields $C\bar{w} = D_K g_K$. Thus, by substituting $C\bar{w}$ by $D_K g_K$ in the first equation of (4), the set of equations (4) can be equivalently written as

$$\begin{bmatrix} sE - A & B_K \\ C & -D_K \end{bmatrix} \begin{bmatrix} \bar{w} \\ g_K \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{5}$$

Finally, note that a solution $(s, -\bar{w}, g_K)$ to above set of equations would yield an invariant zero, zero state, and zero input

for the descriptor system (1). By the detectability assumption,[1] the descriptor model (1) has no zero dynamics and the matrix pencil in (5) necessarily has full rank. It follows that the triple (E, A, C) is observable, so that G can be chosen to make the pair $(E, A+GC)$ Hurwitz [19, Theorem 4.1.1], and the error system (3) is stable and with no zero dynamics. ■

Remark 1: ***(Detection and identification filters for unknown initial condition and noisy dynamics)*** If the network initial state is not available, then, since $(E, A + GC)$ is Hurwitz, an arbitrary initial state $w(0) \in \mathbb{R}^n$ can be chosen. Consequently, the filter converges asymptotically, and some attacks may remain undetected or unidentified. For instance, if the eigenvalues of the detection filter matrix have real part smaller than $c < 0$, with $c \in \mathbb{R}$, then, in the absence of attacks, the residual $r(t)$ exponentially converges to zero with rate less than c. Hence, only inputs $u(t)$ that vanish faster or equal than e^{-ct} may remain undetected by the filter (2). Alternatively, the detection filter can be modified so as to converge in a predefined finite time, see [20], [21]. In this case, every attack signal is detectable after a finite transient.

If the dynamics and the measurements of (1) are affected by modeling uncertainties and noise with known statistics, then the output injection matrix G in (2) should be chosen as to optimize the sensitivity of the residual $r(t)$ to attacks versus the effect of noise. Standard robust filtering or model matching techniques can be adopted for this task [22]. Statistical hypothesis techniques can subsequently be used to analyze the residual $r(t)$ [23]. Finally, as discussed in [1], attacks aligned with the noise statistics turn out to be undetectable. □

Observe that the design of the filter (2) is independent of the particular attack signature (B_K, D_K) and its performance is optimal in the sense that any detectable attack set K can be detected. We remark that for index-one descriptor systems such as power system models, the filter (2) can analogously be designed for the corresponding Kron-reduced model, as defined in [1]. In this case, the resulting attack detection filter is low-dimensional and non-singular but also non-sparse, see [12]. In comparison, the presented filter (2), although inherently centralized, features the *sparse* matrices (E, A, C). This sparsity will be key to develop a distributed attack detection filter.

B. Decentralized attack detection monitor design

Let $G_{\mathrm{t}} = (\mathcal{V}, \mathcal{E})$ be the directed graph associated with the pair (E, A), where the vertex set $V = \{1, \dots, n\}$ corresponds to the system state, and the set of directed edges $\mathcal{E} = \{(x_j, x_i) : e_{ij} \neq 0 \text{ or } a_{ij} \neq 0\}$ is induced by the sparsity pattern of E and A; see also [1, Section IV]. Assume that V has been partitioned into N disjoint subsets as $V = V_1 \cup \cdots \cup V_N$, with $|V_i| = n_i$, and let $G_{\mathrm{t}}^i = (V_i, \mathcal{E}_i)$ be the i-th subgraph of G_{r} with vertices V_i and edges $\mathcal{E}_i = \mathcal{E} \cap (V_i \times V_i)$. 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 $A_D = \mathrm{blkdiag}(A_1, \dots, A_N)$, 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 make the following assumptions:

(A4) the matrices E, C are block-diagonal, that is $E = \mathrm{blkdiag}(E_1, \dots, E_N)$, $C = \mathrm{blkdiag}(C_1, \dots, C_N)$, where $E_i \in \mathbb{R}^{n_i \times n_i}$ and $C_i \in \mathbb{R}^{p_i \times n_i}$,

(A5) each pair (E_i, A_i) is regular, and each triple (E_i, A_i, C_i) is observable.

Given the above structure and in the absence of attacks, the descriptor system (1) can be written as the interconnection of N subsystems of the form

$$\begin{aligned} E_i \dot{x}_i(t) &= A_i x_i(t) + \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} A_{ij} x_j(t), \\ y_i(t) &= C_i x_i(t), \quad i \in \{1, \dots, N\}, \end{aligned} \tag{6}$$

where $x_i(t)$ and $y_i(t)$ are the state and output of the i-th subsystem and $\mathcal{N}_i^{\mathrm{in}} = \{j \in \{1, \dots, N\} \setminus i \ : \ \|A_{ij}\| \neq 0\}$ are the in-neighbors of subsystem i. We also define the set of out-neighbors as $\mathcal{N}_i^{\mathrm{out}} = \{j \in \{1, \dots, N\} \setminus i \ : \ \|A_{ji}\| \neq 0\}$. We assume the presence of a *control center* in each subnetwork G_{t}^i with the following capabilities:

(A6) the i-th control center knows the matrices E_i, A_i, C_i, as well as the neighboring matrices A_{ij}, $j \in \mathcal{N}_i^{\mathrm{in}}$; and

(A7) the i-th control center can transmit an estimate of its state to the j-th control center if $j \in \mathcal{N}_i^{\mathrm{out}}$.

Before deriving a fully-distributed attack detection filter, we explore the question of *decentralized stabilization* of the error dynamics of the filter (2). For each subsystem (6), consider the local residual generator

$$\begin{aligned} E_i \dot{w}_i(t) &= (A_i + G_i C_i) w_i(t) + \sum_{j \in \mathcal{N}_i^{\mathrm{in}}} A_{ij} x_j(t) - G_i y_i(t), \\ r_i(t) &= y_i(t) - C_i w_i(t), \quad i \in \{1, \dots, N\}, \end{aligned} \tag{7}$$

where $w_i(t)$ is the i-th estimate of $x_i(t)$ and $G_i \in \mathbb{R}^{n_i \times p_i}$. In order to derive a compact formulation, let $w(t) = [w_1^{\mathsf{T}}(t) \cdots w_N^{\mathsf{T}}(t)]^{\mathsf{T}}$, $r(t) = [r_1^{\mathsf{T}}(t) \cdots r_N^{\mathsf{T}}(t)]^{\mathsf{T}}$, and $G = \mathrm{blkdiag}(G_1, \dots, G_N)$. Then, the overall filter dynamics (7) are

$$\begin{aligned} E \dot{w}(t) &= (A_D + GC) w(t) + A_C w(t) - G y(t)\,, \\ r(t) &= y(t) - C w(t)\,. \end{aligned} \tag{8}$$

Due to the observability assumption (A5) an output injection matrix G_i can be chosen such that each pair $(E_i, A_i - G_i C_i)$ is Hurwitz [19, Theorem 4.1.1]. Notice that, if each pair $(E_i, A_i + G_i C_i)$ is regular and Hurwitz, then $(E, A_D + GC)$ is also regular and Hurwitz since the matrices E and $A_D + GC$ are block-diagonal. We are now ready to state a condition for the decentralized stabilization of the filter (8).

Lemma 3.2: ***(Decentralized stabilization of the attack detection filter)*** Consider the descriptor system (1), and assume

[1] Due to linearity of the descriptor system (1), the detectability assumption reads as "the attack $(B, D, u(t))$ is detectable if there exist no initial condition $x_0 \in \mathbb{R}^n$, such that $y(x_0, u, t) = 0$ for all $t \in \mathbb{R}_{\geq 0}$."

that the attack set K is detectable and that the network initial state $x(0)$ is known. Consider the attack detection filter (8), where $w(0) = x(0)$ and $G = \text{blkdiag}(G_1, \dots, G_N)$ is such that $(E, A_D + GC)$ is regular and Hurwitz. Assume that

$$\rho\left((j\omega E - A_D - GC)^{-1} A_C\right) < 1 \text{ for all } \omega \in \mathbb{R}\,, \quad (9)$$

where $\rho(\cdot)$ denotes the spectral radius operator. Then $r(t) = 0$ at all times $t \in \mathbb{R}_{\geq 0}$ if and only if $u_K(t) = 0$ at all times $t \in \mathbb{R}_{\geq 0}$. Moreover, in the absence of attacks, the filter error $w(t) - x(t)$ is exponentially stable.

Proof: The error $e(t) = w(t) - x(t)$ obeys the dynamics

$$\begin{aligned} E\dot{e}(t) &= (A_D + A_C + GC)e(t) - (B_K + GD_K)u_K(t), \\ r(t) &= Ce(t) - D_K u_K(t)\,. \end{aligned} \quad (10)$$

A reasoning analogous to that in the proof of Theorem 3.1 shows the absence of zero dynamics. Hence, for $r(t) = 0$ at all times $t \in \mathbb{R}_{\geq 0}$ if and only if $u_K(t) = 0$ at all times $t \in \mathbb{R}_{\geq 0}$.

To show stability of the error dynamics in the absence of attacks, we employ the small-gain approach to large-scale interconnected systems [24] and rewrite the error dynamics (10) as the closed-loop interconnection of the two subsystems

$$\begin{aligned} \Gamma_1 :&\quad E\dot{e}(t) = (A_D + GC)e(t) + v(t)\,, \\ \Gamma_2 :&\quad v(t) = A_C e(t)\,. \end{aligned}$$

Since both subsystems Γ_1 and Γ_2 are causal and internally Hurwitz stable, the overall error dynamics (10) are stable if the loop transfer function $\Gamma_1(j\omega) \cdot \Gamma_2$ satisfies the spectral radius condition $\rho(\Gamma_1(j\omega) \cdot \Gamma_2) < 1$ for all $\omega \in \mathbb{R}$ [22, Theorem 4.11]. The latter condition is equivalent to (9). ∎

Observe that, although control centers can compute the output injection matrix independently of each other, an implementation of the decentralized attack detection filter (8) requires control centers to continuously exchange their local estimation vectors. Thus, this scheme has high communication cost, and it may not be broadly applicable. A solution to this problem is presented in the next section.

C. Distributed attack detection monitor design

In this subsection we exploit the classical waveform relaxation method to develop a fully distributed variation of the decentralized attack detection filter (8). We refer the reader to [25], [26] for a comprehensive discussion of waveform relaxation methods. The Gauss-Jacobi waveform relaxation method applied to the system (8) yields the *waveform relaxation iteration*

$$E\dot{w}^{(k)}(t) = A_D w^{(k)}(t) + A_C w^{(k-1)}(t) - Gy(t)\,, \quad (11)$$

where $k \in \mathbb{N}$ denotes the iteration index, $t \in [0, T]$ is the integration interval for some uniform time horizon $T > 0$, and $w^{(k)} : [0, T] \to \mathbb{R}^n$ is a trajectory with the initial condition $w^{(k)}(0) = w_0$ for each $k \in \mathbb{N}$. Notice that (11) is a descriptor system in the variable $w^{(k)}$ and the vector $A_C w^{(k-1)}$ is a known input, since the value of $w(t)$ at iteration $k-1$ is used. The iteration (11) is said to be (uniformly) *convergent* if

$$\lim_{k\to\infty} \max_{t\in[0,T]} \left\| w^{(k)}(t) - w(t) \right\|_\infty = 0\,,$$

where $w(t)$ is the solution of the non-iterative dynamics (8). In order to obtain a low-complexity distributed detection scheme, we use the waveform relaxation iteration (11) to iteratively approximate the decentralized filter (8).

We start by presenting a convergence condition for the iteration (8). Recall that a function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}^p$ is said to be of *exponential order* β if there exists $\tilde{\beta} \in \mathbb{R}$ such that the exponentially scaled function $\tilde{f} : \mathbb{R}_{\geq 0} \to \mathbb{R}^p$, $f(t) = f(t)e^{-\beta t}$ and all its derivatives exist and are bounded. An elegant analysis of the waveform relaxation iteration (11) can be carried out in the Laplace domain [27], where the operator mapping $w^{(k-1)}(t)$ to $w^{(k)}(t)$ is $(sE - A_D - GC)^{-1}A_C$. Similar to the regular Gauss-Jacobi iteration, convergence conditions of the waveform relaxation iteration (11) rely on the contractivity of the iteration operator.

Lemma 3.3: ***(Convergence of the waveform relaxation [27, Theorem 5.2])*** Consider the waveform relaxation iteration (11). Let the pair $(E, A_D + GC)$ be regular, and the initial condition w_0 be consistent. Let $y(t)$, with $t \in [0, T]$, be of exponential order β. Let α be the least upper bound on the real part of the spectrum of (E, A), and define $\sigma = \max\{\alpha, \beta\}$. The waveform relaxation method (11) is convergent if

$$\rho\left(((\sigma + j\omega)E - A_D - GC)^{-1} A_C\right) < 1 \text{ for all } \omega \in \mathbb{R}. \quad (12)$$

In the reasonable case of bounded (integrable) measurements $y(t)$, $t \in [0, T]$, and stable filter dynamics, we have that $\sigma \leq 0$, and the convergence condition (12) for the waveform relaxation iteration (11) equals the condition (9) for decentralized stabilization of the filer dynamics. We now propose our distributed attack detection filter.

Theorem 3.4: ***(Distributed attack detection filter)*** Consider the descriptor system (1) and assume that the attack set K is detectable, and that the network initial state $x(0)$ is known. Let assumptions (A1) through (A7) be satisfied and consider the *distributed attack detection filter*

$$\begin{aligned} E\dot{w}^{(k)}(t) &= \big(A_D + GC\big)w^{(k)}(t) + A_C w^{(k-1)}(t) - Gy(t)\,, \\ r(t) &= y(t) - Cw^{(k)}(t)\,, \end{aligned} \quad (13)$$

where $k \in \mathbb{N}$, $t \in [0, T]$ for some $T > 0$, $w^{(k)}(0) = x(0)$ for all $k \in \mathbb{N}$, and $G = \text{blkdiag}(G_1, \dots, G_N)$ is such that the pair $(E, A_D + GC)$ is regular, Hurwitz, and

$$\rho\left((j\omega E - A_D - GC)^{-1} A_C\right) < 1 \text{ for all } \omega \in \mathbb{R}\,. \quad (14)$$

Then $\lim_{k\to\infty} r^{(k)}(t) = 0$ at all times $t \in [0, T]$ if and only if $u_K(t) = 0$ at all times $t \in [0, T]$. Moreover, in the absence of attacks, the asymptotic filter error $\lim_{k\to\infty}(w^{(k)}(t) - x(t))$ is exponentially stable for $t \in [0, T]$.

Proof: Since $w^{(k)}(0) = x(0)$, it follows from Lemma 3.3 that the solution $w^{(k)}(t)$ of the iteration (13) converges, as $k \to \infty$, to the solution $w(t)$ of the non-iterative filter dynamics (8) if condition (12) is satisfied with $\sigma = 0$ (due to integrability of $y(t)$, $t \in [0, T]$, and since the pair $(E, A_D + GC)$ is Hurwitz). The latter condition is equivalent to condition (14).

Under condition (14) and due to the Hurwitz assumption, it follows from Lemma 3.2 that the error $e(t) = w(t) - x(t)$ between the state $w(t)$ of the decentralized filter dynamics (8) and the state $x(t)$ of the descriptor model (1) is asymptotically stable in the absence of attacks. Due to the detectability assumption and by reasoning analogous to the proof of Theorem

3.1, it follows that the error dynamics $e(t)$ have no invariant zeros. This concludes the proof of Theorem 3.4. ■

Remark 2: ***(Distributed attack detection)*** The waveform relaxation iteration (11) can be implemented in the following distributed fashion. Assume that each control center i is able to numerically integrate the descriptor system

$$E_i\dot{w}_i^{(k)}(t) = (A_i + G_iC_i)w_i^{(k)}(t) + \sum_{j\in\mathcal{N}_i^{\text{in}}} A_{ij}w_j^{(k-1)}(t) - G_iy_i(t)\,, \tag{15}$$

over a time interval $t \in [0, T]$, with initial condition $w_i^{(k)}(0) = w_{i,0}$, measurements $y_i(t)$, and the neighboring filter states $w_j^{(k-1)}(t)$ as external inputs. Let $w_j^{(0)}(t)$ be an initial guess of the signal $w_j(t)$. Each control center $i \in \{1, \dots, N\}$ performs the following operations assuming $k = 0$ at start:

(1) set $k := k + 1$, and compute the signal $w_i^{(k)}(t)$ by integrating the local filter equation (15),
(2) transmit $w_i^{(k)}(t)$ to the j-th control center if $j \in \mathcal{N}_i^{\text{out}}$
(3) update the input $w_j^{(k)}$ with the signal received from the j-th control center, with $j \in \mathcal{N}_i^{\text{in}}$, and iterate.

If the waveform relaxation is convergent, then, for k sufficiently large, the residuals $r_i^{(k)}(t) = y_i(t) - C_iw_i^{(k)}(t)$ can be used to detect attacks; see Theorem 3.4. In summary, our distributed attack detection scheme requires integration capabilities at each control center, knowledge of the measurements $y_i(t)$, $t \in [0, T]$, as well as synchronous discrete-time communication between neighboring control centers. □

Remark 3: ***(Distributed filter design)*** As discussed in Remark 2, the filter (13) can be implemented in a distributed fashion. In fact, it is also possible to design the filter (13), that is, the output injections G_i, in an entirely distributed way. Since $\rho(A) \leq \|A\|_p$ for any matrix A and any induced p-norm, condition (14) can be relaxed by the small gain criterion to

$$\left\|(j\omega E - A_D - GC)^{-1}A_C\right\|_p < 1 \text{ for all } \omega \in \mathbb{R}\,. \tag{16}$$

With $p = \infty$, in order to satisfy condition (16), it is sufficient for each control center i to verify the following *quasi-block diagonal dominance* condition [28] for each $\omega \in \mathbb{R}$:

$$\left\|(j\omega E_i - A_i - G_iC_i)^{-1}\sum\nolimits_{j=1,j\neq i}^{n} A_{ij}\right\|_\infty < 1. \tag{17}$$

Note that condition (17) can be checked with local information, and it is a conservative relaxation of condition (14). □

D. Illustrative example of decentralized detection

The IEEE 118 bus system shown in Fig. 1 represents a portion of the Midwestern American Electric Power System as of December 1962. This test case system is composed of 118 buses and 54 generators, and its parameters can be found, for example, in [29]. Following [1, Section II.C], a linear continuous-time descriptor model of the network dynamics under attack assumes the form (1).

For estimation and attack detection purposes, we partition the IEEE 118 bus system into 5 disjoint areas, we assign a control center to each area, and we implement our detection procedure via the filter (13); see Fig. 1 for a graphical illustration. Suppose that each control center continuously

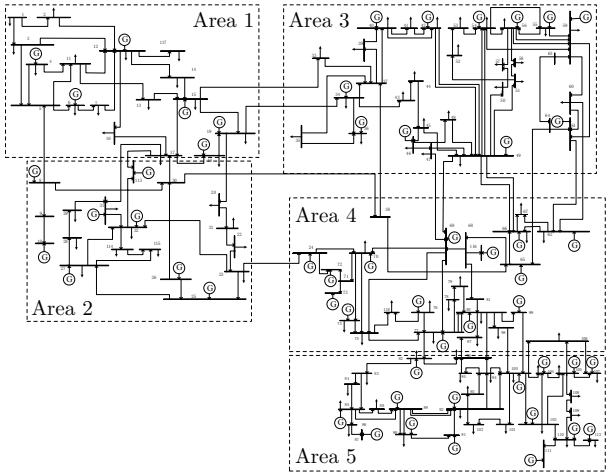

Fig. 1. Partition of IEEE 118 bus system into 5 areas. Each area is monitored and operated by a control center. The control centers cooperate to estimate the state and to assess the functionality of the whole network.

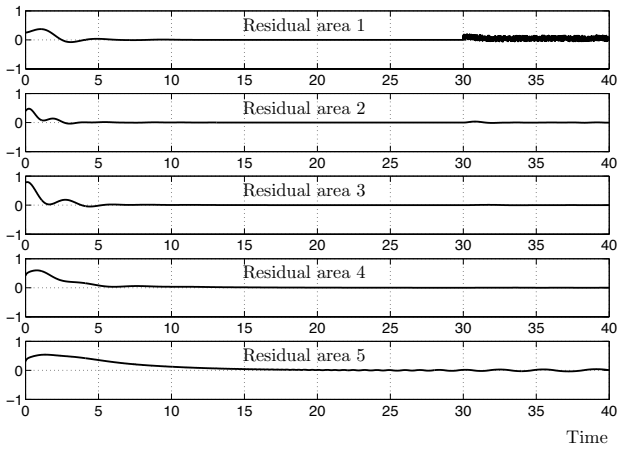

Fig. 2. In this figure we show the residual functions computed through the distributed attack detection filter (13). The attacker compromises the measurements of all the generators in area 1 from time 30 with a signal uniformly distributed in the interval $[0, 0.5]$. The attack is correctly detected, because the residual functions do not decay to zero. For the simulation, we run $k = 100$ iterations of the attack detection method.

measures the angle of the generators in its area, and suppose that an attacker compromises the measurements of all the generators of the first area. In particular, starting at time 30s, the attacker adds a signal $u_K(t)$ to all measurements in area 1. It can be verified that the attack set K is detectable, see [1]. According to assumption (A3), the attack signal $u_K(t)$ needs to be continuous to guarantee a continuous state trajectory (since the power network is a descriptor system of index 1). In order to show the robustness of our detection filter (13), we let $u_K(t)$ be randomly distributed in the interval $[0, 0.5]\,\text{rad}$.

The control centers implement the distributed attack detection procedure described in (13), with $G = AC^{\mathsf{T}}$. It can be verified that the pair $(E, A_D + GC)$ is Hurwitz stable, and that $\rho\left(j\omega E - A_D - GC)^{-1}A_C\right) < 1$ for all $\omega \in \mathbb{R}$. As predicted by Theorem 3.4, our distributed attack detection filter is convergent; see Fig. 2. For completeness, in Fig. 3 we

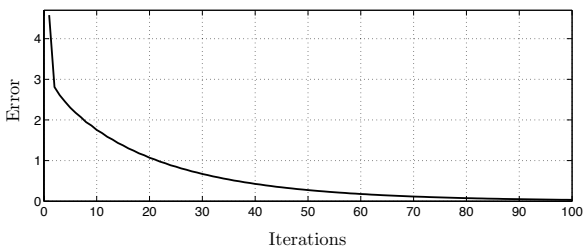

Fig. 3. The plot represents the error of our waveform relaxation based filter (13) with respect to the corresponding decentralized filter (8). Here the error is $\max_{t\in[0,T]} \left\| w^{(k)}(t) - w(t) \right\|_\infty$, that is, the worst-case difference of the outputs of the two filters. As predicted by Theorem 3.4, the error is convergent.

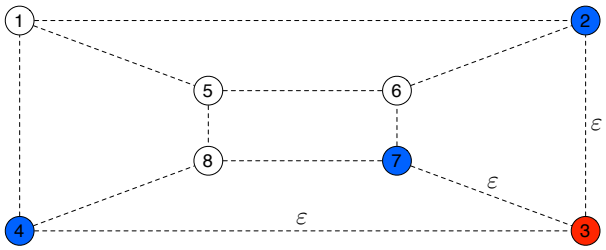

Fig. 4. A regular consensus system (A, B, C), where the state variable 3 is corrupted by the attacker, and the state variables 2, 4, and 7 are directly measured. Due to the sparsity pattern of (A, B, C) any attack of cardinality one is *generically* detectable and identifiable, see [1], [7] for further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

illustrate the convergence of our waveform relaxation-based filter as a function of the number of iterations k. Notice that the number of iterations directly reflects the communication complexity of our detection scheme.

IV. Monitor design for attack identification

A. Complexity of the attack identification problem

In this section we study the problem of attack identification, that is, the problem of identifying from measurements the state and output variables corrupted by the attacker. We start our discussion by showing that this problem is generally *NP-hard*. For a vector $x \in \mathbb{R}^n$, let $\text{supp}(x) = \{i \in \{1, \dots, n\} : x_i \neq 0\}$, let $\|x\|_{\ell_0} = |\text{supp}(x)|$ denote the number of non-zero entries, and for a vector-valued signal $v : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$, let $\|v\|_{\mathcal{L}_0} = |\cup_{t\in\mathbb{R}_{\geq 0}} \text{supp}(v(t))|$. We consider the following cardinality minimization problem: given a descriptor system with dynamic matrices $E, A \in \mathbb{R}^{n\times n}$, 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 arbitrary initial condition $\xi_0 \in \mathbb{R}^n$ that explain the data $y(t)$, that is,

$$\begin{aligned} \min_{v_x,\, v_y,\, \xi_0} \quad & \|v_x\|_{\mathcal{L}_0} + \|v_y\|_{\mathcal{L}_0} \\ \text{subject to} \quad & E\dot{\xi}(t) = A\xi(t) + v_x(t), \\ & y(t) = C\xi(t) + v_y(t), \\ & \xi(0) = \xi_0 \in \mathbb{R}^n \,. \end{aligned} \tag{18}$$

Lemma 4.1: ***(Problem equivalence)*** Consider the system (1) with identifiable attack set K. The optimization problem (18) coincides with the problem of identifying the attack set K given the system matrices E, A, C, and the measurements $y(t)$, where $K = \text{supp}([v_x^\mathsf{T}\ v_y^\mathsf{T}])$.

Proof: Due to the identifiability of K, the attack identification problem consists of finding the smallest attack set capable of injecting an attack $(B_K u_K, D_K u_K)$ that generates the given measurements y for the given dynamics E, A, C, and some initial condition; see Definition 2. The statement follows since $B = [I, 0]$ and $D = [0, I]$ in (1), so that $(B_K u_K, D_K u_K) = (v_x, v_y)$. ∎

As it turns out, the optimization problem (18), or equivalently our identification problem, is generally *NP-hard* [30].

Corollary 4.2: ***(Complexity of the attack identification problem)*** Consider the system (1) with identifiable attack set K. The attack identification problem given the system matrices E, A, C, and the measurements $y(t)$ is NP-hard.

Proof: Consider the NP-hard [31] sparse recovery problem $\min_{\bar{\xi}\in\mathbb{R}^n} \|\bar{y} - \bar{C}\bar{\xi}\|_{\ell_0}$, where $\bar{C} \in \mathbb{R}^{p\times n}$ and $\bar{y} \in \mathbb{R}^p$ are given and constant. In order to prove the claimed statement, we show that every instance of the sparse recovery problem can be cast as an instance of (18). Let $E = I$, $A = 0$, $C = \bar{C}$, and $y(t) = \bar{y}$ at all times. Notice that $v_y(t) = \bar{y} - C\xi(t)$ and $\xi(t) = \xi(0) + \int_0^t v_x(\tau)d\tau$. The problem (18) can be written as

$$\begin{aligned} & \min_{v_x,\, \xi} \ \|v_x\|_{\mathcal{L}_0} + \|\bar{y} - \bar{C}\xi(t)\|_{\mathcal{L}_0} \\ & = \min_{v_x(t),\, \bar{\xi}} \ \|v_x(t)\|_{\mathcal{L}_0} + \|\bar{y} - \bar{C}\bar{\xi} - \bar{C}\textstyle\int_0^t v_x(\tau)d\tau\|_{\mathcal{L}_0}, \end{aligned} \tag{19}$$

where $\bar{\xi} = \xi(0)$. Notice that there exists a minimizer to problem (19) with $v_x(t) = 0$ for all t. Indeed, since $\|\bar{y} - \bar{C}\bar{\xi} - \bar{C}\int_0^t v_x(\tau)d\tau\|_{\mathcal{L}_0} = |\cup_{t\in\mathbb{R}_{\geq 0}} \text{supp}(\bar{y} - \bar{C}\bar{\xi} - \bar{C}\int_0^t v_x(\tau)d\tau)| \geq |\text{supp}(\bar{y} - \bar{C}\bar{\xi} - \bar{C}\int_0^0 v_x(\tau)d\tau)| = \|\bar{y} - \bar{C}\bar{\xi}\|_{\ell_0}$, problem (19) can be equivalently written as $\min_{\bar{\xi}} \|\bar{y} - \bar{C}\bar{\xi}\|_{\ell_0}$. ∎

By Corollary 4.2 the general attack identification problem is combinatorial in nature, and its general solution will require substantial computational effort. In the next sections we propose an optimal algorithm with high computational complexity, and a sub-optimal algorithm with low computational complexity. We conclude this section with an example.

Example 1: ***(Attack identification via ℓ_1 regularization)*** A classical procedure to handle cardinality minimization problems of the form $\min_{v\in\mathbb{R}^n} \|y - Av\|_{\ell_0}$ is to use the ℓ_1 regularization $\min_{v\in\mathbb{R}^n} \|y - Av\|_{\ell_1}$ [31]. This procedure can be adapted to the optimization problem (18) after converting it into an algebraic optimization problem, for instance by taking subsequent derivatives of the output $y(t)$, or by discretizing the continuous-time system (1) and recording several measurements. As shown in [8], 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 performs generally poorly. In what follows, we develop an intuition when and why this approach fails.

Consider a consensus system with underlying network graph (sparsity pattern of A) illustrated in Fig. 4. The dynamics

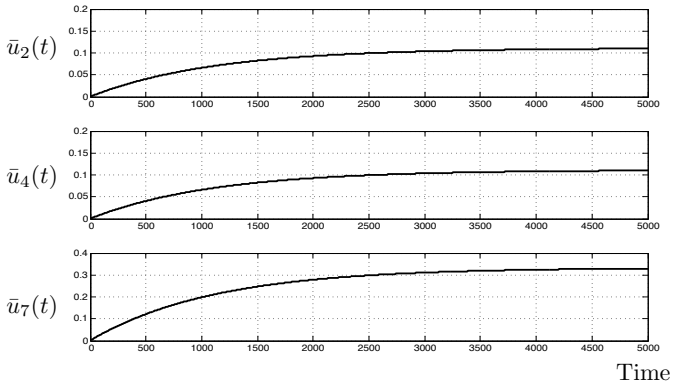

Fig. 5. Plot of the attack mode $\bar{u}(t)$ for the attack set $\bar{K} = \{2, 4, 7\}$ to generate the same output as the attack set $K = \{3\}$ with attack mode $u(t) = 1$. Although $|\bar{K}| > |K|$, we have that $|\bar{u}_i(t)| < |u(t)|/3$ for $i \in \{1, 2, 3\}$.

are described by the nonsingular matrix $E = I$ and the state matrix A depending on the small parameter $0 < \varepsilon \ll 1$ as

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

The measurement matrix C and the attack signature B_K are

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

and we let $G(s) = C(sI - A)^{-1}B_K$. It can be verified that the state attack $K = \{3\}$ is detectable and identifiable.

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

$$B_{\bar{K}}^{\mathsf{T}} = \begin{bmatrix} 0&1&0&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&1&0 \end{bmatrix},$$

and let $\bar{G}(s) = C(sI - A)^{-1}B_{\bar{K}}$. We now adopt the shorthands $u(t) = u_K(t)$ and $\bar{u}(t) = u_{\bar{K}}(t)$, and denote their Laplace transforms by $U(s)$ and $\bar{U}(s)$, respectively. Notice that $\bar{G}(s)$ is right-invertible [32]. Thus, $Y(s) = G(s)U(s) = \bar{G}(s)\left(\bar{G}^{-1}(s)G(s)U(s)\right)$. In other words, the measurements $Y(s)$ generated by the attack signal $U(s)$ can equivalently be generated by the signal $\bar{U}(s) = \bar{G}^{-1}(s)G(s)U(s)$. Obviously, we have that $\|\bar{u}\|_{\mathcal{L}_0} = 3 > \|u\|_{\mathcal{L}_0} = 1$, that is, the attack set K achieves a lower cost than $\bar{K}$ in the optimization problem (18).

Consider now the numerical realization $\varepsilon = 0.0001$, $x(0) = 0$, and $u(t) = 1$ for all $t \in \mathbb{R}_{\geq 0}$. The corresponding attack mode $\bar{u}(t)$ is shown in Fig. 5. Since $|\bar{u}_i(t)| < 1/3$ for $i \in \{1, 2, 3\}$ and $t \in \mathbb{R}_{\geq 0}$, it follows that $\|u(t)\|_{\ell_p} > \|\bar{u}(t)\|_{\ell_p}$ point-wise in time and $\|u(t)\|_{\mathcal{L}_q/\ell_p} > \|\bar{u}(t)\|_{\mathcal{L}_p/\ell_q}$, where $p, q \geq 1$ and $\|u(t)\|_{\mathcal{L}_q/\ell_p} = \left(\int_0^\infty (\sum_{i=1}^{n+p} |u_i(\tau)|^p)^{q/p} d\tau\right)^{1/q}$ is the $\mathcal{L}_q/\ell_p$-norm. Hence, the attack set $\bar{K}$ achieves a lower cost than K for any algebraic version of the optimization problem (18) penalizing a ℓ_p cost point-wise in time or a $\mathcal{L}_q/\ell_p$ cost over a time interval. Since $\|\bar{u}\|_{\mathcal{L}_0} > \|u\|_{\mathcal{L}_0}$, 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 $p, q \geq 1$.

Notice 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. Moreover, large-scale stable systems often exhibit this behavior independently of the system parameters. This can be easily seen in discrete-time systems, where a state attack with attack set K affects the output via the matrix $CA^{r-1}B_K$, where r is the relative degree of (A, B_K, C). Hence, if A is Schur stable and thus $\lim_{k\to\infty} A^k = 0$, then $CA^{r-1}B_K$ converges to the zero matrix for increasing relative degree. In this case, an attack closer to the sensors may achieve a lower $\mathcal{L}_q/\ell_p$ cost than an attack far from sensors independently of the cardinality of the attack set. In short, the ϵ-connections in Fig. 4 can be thought of as the effect of a large relative degree in a stable system. □

B. Centralized attack identification monitor design

As previously shown, unlike the detection case, the identification of the attack set K requires a combinatorial procedure, since, a priori, K is one of the $\binom{n+p}{|K|}$ possible attack sets. The following centralized attack identification procedure consists of designing a residual filter to determine whether a predefined set coincides with the attack set. The design of this residual filter consists of three steps – an input output transformation (see Lemma 4.3), a state transformation to a suitable conditioned-invariant subspace (see Lemma 4.4), and an output injection and definition of a proper residual (see Theorem 4.5).

As a first design step, we show that the identification problem can be carried out for a modified system without corrupted measurements, that is, without the feedthrough matrix D.

Lemma 4.3: ***(Attack identification with safe measurements)*** Consider the descriptor system (1) with attack set K. The attack set K is identifiable for the descriptor system (1) if and only if it is identifiable for the following descriptor system:

$$\begin{aligned} E\dot{x}(t) &= (A - B_K D_K^\dagger C)x(t) + B_K(I - D_K^\dagger D_K)u_K(t), \\ \tilde{y}(t) &= (I - D_K D_K^\dagger)Cx(t). \end{aligned} \tag{20}$$

Proof: Due to the identifiability hypothesis, there exists no attack set R with $|R| \leq |K|$ and $R \neq K$, $s \in \mathbb{C}$, $g_K \in \mathbb{R}^{|K|}$, $g_R \in \mathbb{R}^{|R|}$, and $x \in \mathbb{R}^n \setminus \{0\}$ such that

$$\left[\begin{array}{c|c|c} sE - A & -B_K & -B_R \\ \hline C & D_K & D_R \\ \hline C & D_K & D_R \end{array}\right] \begin{bmatrix} x \\ g_K \\ g_R \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{21}$$

where we added an additional (redundant) output equation [1, Theorem 3.4]. A multiplication of equation (21) from the left by the projectors $\text{blkdiag}\big(I\,,\, D_K D_K^\dagger\,,\, (I - D_K D_K^\dagger)\big)$ yields

$$\left[\begin{array}{c|c|c} sE - A & -B_K & -B_R \\ \hline D_K D_K^\dagger C & D_K & D_K D_K^\dagger D_R \\ \hline (I - D_K D_K^\dagger)C & 0 & (I - D_K D_K^\dagger)D_R \end{array}\right] \begin{bmatrix} x \\ g_K \\ g_R \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

The variable g_K can be eliminated in the first redundant (corrupted) output equation according to

$$g_K = -D_K^\dagger C x - D_K^\dagger D_R g_R + (I - D_K^\dagger D_K)g_K.$$

Thus, $P(s)[x^{\mathsf{T}}\; g_K^{\mathsf{T}}\; g_R^{\mathsf{T}}]^{\mathsf{T}} = 0$ has no solution, where $P(s)$ is

$$\left[\begin{array}{c|c|c} sE - A + B_K D_K^\dagger C & -B_K(I - D_K^\dagger D_K) & -B_R + B_K D_K^\dagger D_R \\ \hline (I - D_K D_K^\dagger)C & 0 & (I - D_K D_K^\dagger)D_R \end{array}\right]$$

The statement follows. ■

The second design step of our attack identification monitor relies on the concept of *conditioned invariant subspace*. We refer to [18], [32], [33] for a comprehensive discussion of conditioned invariant subspaces. Let $\mathcal{S}^*$ be the conditioned invariant subspace associated with the system (E, A, B, C, D), that is, the smallest subspace of the state space satisfying

$$\mathcal{S}^* = \begin{bmatrix} A & B \end{bmatrix} \left(\begin{bmatrix} E^{-1}\mathcal{S}^* \\ \mathbb{R}^m \end{bmatrix} \cap \operatorname{Ker} \begin{bmatrix} C & D \end{bmatrix} \right), \tag{22}$$

and let L be an output injection matrix satisfying

$$\begin{bmatrix} A + LC & B + LD \end{bmatrix} \begin{bmatrix} E^{-1}\mathcal{S}^* \\ \mathbb{R}^m \end{bmatrix} \subseteq \mathcal{S}^*. \tag{23}$$

We transform the descriptor system (20) into a set of canonical coordinates representing $\mathcal{S}^*$ and its orthogonal complement. For a nonsingular system ($E = I$) such an equivalent state representation can be achieved by a nonsingular transformation of the form $Q^{-1}(sI - A)Q$. However, for a singular system different transformations need to be applied in the domain and codomain such as $P^{\mathsf{T}}(sE - A)Q$ for nonsingular P and Q.

Lemma 4.4: ***(Input decoupled system representation)*** For the system (20), let $\mathcal{S}^*$ and L be as in (22) and (23), respectively. Define the unitary matrices $P = \begin{bmatrix} \operatorname{Basis}(\mathcal{S}^*) & \operatorname{Basis}((\mathcal{S}^*)^{\perp}) \end{bmatrix}$ and $Q = \begin{bmatrix} \operatorname{Basis}(E^{-1}\mathcal{S}^*) & \operatorname{Basis}((E^{-1}\mathcal{S}^*)^{\perp}) \end{bmatrix}$. Then

$$P^{\mathsf{T}}EQ = \begin{bmatrix} \tilde{E}_{11} & \tilde{E}_{12} \\ 0 & \tilde{E}_{22} \end{bmatrix}, P^{\mathsf{T}}(A - B_K D_K^{\dagger} C + LC)Q = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix},$$
$$P^{\mathsf{T}}B_K(I - D_K^{\dagger}D_K) = \begin{bmatrix} \tilde{B}_K(t) \\ 0 \end{bmatrix}, (I - D_K D_K^{\dagger})C)Q = \begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 \end{bmatrix}.$$

The attack set K is identifiable for the descriptor system (1) if and only if it is identifiable for the descriptor system

$$\begin{bmatrix} \tilde{E}_{11} & \tilde{E}_{12} \\ 0 & \tilde{E}_{22} \end{bmatrix} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} \tilde{B}_K(t) \\ 0 \end{bmatrix},$$
$$y(t) = \begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}. \tag{24}$$

Proof: Let $\mathcal{L} = E^{-1}\mathcal{S}^*$ and $\mathcal{M} = \mathcal{S}^*$. Notice that $(A + LC)E^{-1}\mathcal{S}^* \subseteq \mathcal{S}^*$ by the invariance property of $\mathcal{S}^*$ [33], [18]. It follows that $\mathcal{L}$ and $\mathcal{M}$ are a pair of *right deflating subspaces* for the matrix pair $(A + LC, E)$ [34], that is, $\mathcal{M} = A\mathcal{L} + E\mathcal{L}$ and $\dim(\mathcal{M}) \leq \dim(\mathcal{L})$. The sparsity pattern in the descriptor and dynamic matrices $\tilde{E}$ and $\tilde{A}$ of (24) arises by construction of the right deflating subspaces P and Q [34, Eq. (2.17)], and the sparsity pattern in the input matrix arises due to the invariance properties of $\mathcal{S}^*$ containing $\operatorname{Im}(B_K)$. The statement follows because the output injection L, the coordinate change $x \mapsto Q^{-1}x$, and the left-multiplication of the dynamics by P^{T} does not affect the existence of zero dynamics. ∎

We call system (24) the *conditioned system* associated with (1). For the ease of notation and without affecting generality, the third and final design step of our attack identification filter is presented for the conditioned system (24).

Theorem 4.5: ***(Attack identification filter for attack set K)*** Consider the *conditioned system* (24) associated with the descriptor system (1). Assume that the attack set is identifiable, the network initial state $x(0)$ is known, and the assumptions (A1) through (A3) are satisfied. Consider the *attack identification filter for the attack signature* (B_K, D_K)

$$\begin{aligned} \tilde{E}_{22}\dot{w}_2(t) &= (\tilde{A}_{22} + \tilde{G}(I - \tilde{C}_1\tilde{C}_1^{\dagger})\tilde{C}_2)w_2(t) - \tilde{G}\bar{y}(t), \\ r_K(t) &= (I - \tilde{C}_1\tilde{C}_1^{\dagger})\tilde{C}_2 w_2(t) - \bar{y}(t), \quad \text{with} \\ \bar{y}(t) &= (I - \tilde{C}_1\tilde{C}_1^{\dagger})\tilde{C}_2 y(t), \end{aligned} \tag{25}$$

where $w_2(0) = x_2(0)$, and $\tilde{G}$ is such that $(\tilde{E}_{22}, \tilde{A}_{22} + \tilde{G}(I - \tilde{C}_1\tilde{C}_1^{\dagger})\tilde{C}_2)$ is Hurwitz. Then $r_K(t) = 0$ for all times $t \in \mathbb{R}_{\geq 0}$ if and only if K coincides with the attack set.

Proof: Let $w(t) = [w_1(t)^{\mathsf{T}}\ w_2(t)^{\mathsf{T}}]^{\mathsf{T}}$, where $w_1(t)$ obeys

$$\tilde{E}_{11}\dot{w}_1(t) + \tilde{E}_{12}\dot{w}_2(t) = \tilde{A}_{11}w_1(t) + \tilde{A}_{12}w_2(t).$$

Consider the filter error $e(t) = w(t) - x(t)$, and notice that

$$\begin{bmatrix} \tilde{E}_{11} & \tilde{E}_{12} \\ 0 & E_{22} \end{bmatrix} \begin{bmatrix} \dot{e}_1(t) \\ \dot{e}_2(t) \end{bmatrix} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \bar{A}_{22} \end{bmatrix} \begin{bmatrix} e_1(t) \\ e_2(t) \end{bmatrix} - \begin{bmatrix} \tilde{B}_K \\ 0 \end{bmatrix} u_K(t),$$
$$r_K(t) = (I - \tilde{C}_1\tilde{C}_1^{\dagger})\tilde{C}_2 e_2(t),$$

where $\bar{A}_{22} = \tilde{A}_{22} + \tilde{G}(I - \tilde{C}_1\tilde{C}_1^{\dagger})\tilde{C}_2)$. Notice that $r_K(t)$ is not affected by the input $u_K(t)$, so that, since $e_2(0) = 0$ due to $w_2(0) = x_2(0)$, the residual $r_K(t)$ is identically zero when K is the attack set. In order to prove the theorem we are left to show that for every set R, with $|R| \leq |K|$ and $R \cap K = \emptyset$, every attack mode $u_R(t)$ results in a nonzero residual $r_K(t)$. From [1, Theorem 3.4] and the identifiability hypothesis, for any $R \neq K$, there exists no solution to

$$\left[\begin{array}{cc|c|c} s\tilde{E}_{11} - \tilde{A}_{11} & s\tilde{E}_{12} - \tilde{A}_{12} & \tilde{B}_K & -B_{R1} \\ 0 & s\tilde{E}_{22} - \bar{A}_{22} & 0 & -B_{R2} \\ \hline \tilde{C}_1 & \tilde{C}_2 & 0 & D_R \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ g_K \\ g_R \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

A projection of the equation $0 = \tilde{C}_1 x_1 + \tilde{C}_2 x_2 + D_R g_R$ onto the image of $\tilde{C}_1$ and its orthogonal complement yields

$$\left[\begin{array}{cc|c|c} s\tilde{E}_{11} - \tilde{A}_{11} & s\tilde{E}_{12} - \tilde{A}_{12} & B_K & -B_{R1} \\ 0 & s\tilde{E}_{22} - \bar{A}_{22} & 0 & -B_{R2} \\ \hline \tilde{C}_1 & \tilde{C}_1\tilde{C}_1^{\dagger}\tilde{C}_2 & 0 & \tilde{C}_1\tilde{C}_1^{\dagger}D_R \\ 0 & (I - \tilde{C}_1\tilde{C}_1^{\dagger})\tilde{C}_2 & 0 & (I - \tilde{C}_1\tilde{C}_1^{\dagger})D_R \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ g_K \\ g_R \end{bmatrix}$$
$$= [0 \quad 0 \quad 0 \quad 0]^{\mathsf{T}}. \tag{26}$$

Due to the identifiability hypothesis the set of equations (26) features no solution $[x_1^{\mathsf{T}}\ x_2^{\mathsf{T}}\ g_K^{\mathsf{T}}\ g_R^{\mathsf{T}}]^{\mathsf{T}}$ with $[x_1^{\mathsf{T}}\ x_2^{\mathsf{T}}]^{\mathsf{T}} = 0$.

Observe that, for every x_2 and g_R, there exists $x_1 \in \operatorname{Ker}(\tilde{C}_1)^{\perp}$ such that the third equation of (26) is satisfied. Furthermore, for every x_2 and g_R, there exist $x_1 \in \operatorname{Ker}(\tilde{C}_1)$ and g_K such that the first equation of (26) is satisfied. Indeed, since $QE^{-1}\mathcal{S}^* = [\operatorname{Im}(I)\ 0]^{\mathsf{T}}$ and $P^{\mathsf{T}}\mathcal{S}^* = [\operatorname{Im}(I)\ 0]^{\mathsf{T}}$, the invariance of $\mathcal{S}^*$ implies that $\mathcal{S}^* = A(E^{-1}\mathcal{S}^* \cap \operatorname{Ker}(C)) + \operatorname{Im}(B_K)$, or equivalently in new coordinates, $\operatorname{Im}(I) = \tilde{A}_{11}\operatorname{Ker}(\tilde{C}_1) + \operatorname{Im}(\tilde{B}_K)$. Finally note that $[(s\tilde{E}_{11} - \tilde{A}_{11})\operatorname{Ker}(\tilde{C}_1)\ \tilde{B}_K]$ is of full row rank due to the controllability of the subspace $\mathcal{S}^*$ [18]. We conclude that there exist no vectors x_2 and g_R such that $(s\tilde{E}_{22} - \bar{A}_{22})x_2 - B_{R2}g_R = 0$ and $(I - \tilde{C}_1\tilde{C}_1^{\dagger})(\tilde{C}_2 x_2 + D_R g_R) = 0$ and the statement follows. ∎

Our identification procedure is summarized in Algorithm 1. Observe that the proposed attack identification filter extends classical results concerning the design of unknown-input fault detection filters. In particular, our filter generalizes the construction of [15] to descriptor systems with direct

Algorithm 1: *Identification Monitor for* (B_K, D_K)

Input : Matrices E, A, B_K, and D_K,;
Require : Identifiability of attack set K;

1 From system (1) define the system (20);
2 Compute $\mathcal{S}^*$ and L for system (20) as in (22) and (23);
3 Apply L, P, and Q as in Lemma 4.4 leading to system (24);
4 For (24), define r_K and apply the output injection $\bar{G}$ as in (25).

feedthrough matrix. Additionally, we guarantee the absence of invariant zeros in the residual dynamics. By doing so, our attack identification filter is sensitive to *every* attack mode. Notice that classical fault detection filters, for instance those presented in [15], are guaranteed to detect and isolate signals that do not excite exclusively zero dynamics. Finally, an attack identification filter for the case of state space or index-one systems is presented in our previous work [12].

Remark 4: ***(Complexity of centralized identification)*** Our centralized identification procedure assumes the knowledge of the cardinality k of the attack set, and it achieves identification of the attack set by constructing a residual generator for $\binom{n+p}{k}$ possible attack sets. Thus, for each finite value of k, our procedure constructs $O(n^k)$ filters. If only an upper bound $\bar{k}$ on the cardinality of the attack set is available, identification can be achieved by constructing $\binom{n+p}{\bar{k}}$ filters, and by intersecting the attack sets generating zero residuals.□

Remark 5: ***(Attack identification filter in the presence of noise)*** Let the dynamics and the measurements of the system (1) be affected, respectively, by the additive white noise signals $\eta(t)$, with $\mathbb{E}[\eta(t)\eta^{\mathsf{T}}(\tau)] = R_\eta \delta(t-\tau)$, and $\zeta(t)$, with $\mathbb{E}[\zeta(t)\zeta^{\mathsf{T}}(\tau)] = R_\zeta \delta(t-\tau)$. Let the state and output noise be independent of each other. Then, simple calculations show that the dynamics and the output of the attack identification filter (25) are affected, respectively, by the noise signals

$$\hat{\eta}(t) = P^{\mathsf{T}}\eta(t) + P^{\mathsf{T}}(L(I - D_K D_K^\dagger) - B_K D_K^\dagger)\zeta(t),$$
$$\hat{\zeta}(t) = -\Big(I - \Big[(I - D_K D_K^\dagger)CQ_1\Big]\Big[(I - D_K D_K^\dagger)CQ_1\Big]^\dagger (I - D_K D_K^\dagger)\Big)\zeta(t),$$

where $Q_1 = \mathrm{Basis}(E^{-1}\mathcal{S}^*)$. Define the covariance matrix

$$R_{\hat{\eta},\hat{\zeta}} = \mathbb{E}\left(\begin{bmatrix}\hat{\eta}(t)\\ \hat{\zeta}(t)\end{bmatrix}\begin{bmatrix}\hat{\eta}^{\mathsf{T}}(t) & \hat{\zeta}^{\mathsf{T}}(t)\end{bmatrix}\right).$$

Notice that the off-diagonal elements of $R_{\hat{\eta},\hat{\zeta}}$ are in general nonzero, that is, the state and output noises of the attack identification filter are not independent of each other. As in the detection case, by using the covariance matrix $R_{\hat{\eta},\hat{\zeta}}$, the output injection matrix $\tilde{G}$ in (25) can be designed to optimize the robustness of the residual $r_K(t)$ against noise. A related example is in Section V. □

We conclude this section by observing that a distributed implementation of our attack identification scheme is not practical. Indeed, even if the filters parameters may be obtained via distributed computation, still $\binom{n+p}{k}$ filters would need to be implemented to identify an attack of cardinality k. Such a distributed implementation results in an enormous communication effort and does not reduce the fundamental combinatorial complexity.

C. Fully decoupled attack identification

In the following sections we develop a distributed attack identification procedure. Consider the decentralized setup presented in Section III-B with assumptions (A4)-(A7). The subsystem assigned to the i-th control center is

$$\begin{aligned} E_i\dot{x}_i(t) &= A_i x_i(t) + \sum_{j\in\mathcal{N}_i^{\mathrm{in}}} A_{ij}x_j(t) + B_{K_i}u_{K_i}(t),\\ y_i(t) &= C_i x_i(t) + D_{K_i}u_{K_i}(t),\ \ i\in\{1,\dots,N\},\end{aligned} \tag{27}$$

where $K_i = (K\cap V_i)\cup K_i^{\mathrm{p}}$ with K being the attack set and K_i^{p} being the set of corrupted measurements in the region G_{t}^i.

As a first distributed identification method we consider the fully decoupled case (no cooperation among control centers). In the spirit of [16], the neighboring states $x_j(t)$ affecting $x_i(t)$ are treated as unknown inputs ($f_i(t)$) to the i-th subsystem:

$$\begin{aligned} E_i\dot{x}_i(t) &= A_i x_i(t) + B_i^{\mathrm{b}} f_i(t) + B_{K_i}u_{K_i}(t),\\ y_i(t) &= C_i x_i(t) + D_{K_i}u_{K_i}(t),\ \ i\in\{1,\dots,N\},\end{aligned} \tag{28}$$

where $B_i^{\mathrm{b}} = [A_{i1}\ \cdots\ A_{i,i-1}\ A_{i,i+1}\ \cdots\ A_{iN}]$. We refer to (28) as to the *i-th decoupled system*, and we let $K_i^{\mathrm{b}}\subseteq V_i$ be the set of *boundary nodes* of (28), that is, the nodes $j\in V_i$ with $A_{jk}\neq 0$ for some $k\in\{1,\dots,n\}\setminus V_i$.

If the attack identification procedure in Section IV-B is designed for the i-th decoupled system (28) subject to unknown inputs $f_i(t)$ and $u_{K_i}(t)$, then a total of only $\sum_{i=1}^N \binom{n_i+p_i}{|K_i|} < \binom{n+p}{|K|}$ need to be designed. Although the combinatorial complexity of the identification problem is tremendously reduced, this decoupled identification procedure has several limitations. The following fundamental limitations follow from [1]:

(L1) if $(E_i, A_i, B_{K_i}, C_i, D_{K_i})$ has invariant zeros, then K_i is not detectable by the i-th control center;
(L2) if there is an attack set R_i, with $|R_i|\leq|K_i|$, such that $(E_i, A_i, [B_{K_i}\ B_{R_i}], C_i, [D_{K_i}\ D_{R_i}])$ has invariant zeros, then K_i is not identifiable by the i-th control center;
(L3) if $K_i\not\subseteq K_i^{\mathrm{b}}$ and $(E_i, A_i, [B_i^{\mathrm{b}}\ B_{K_i}], C_i, D_{K_i})$ has no invariant zeros, then K_i is detectable by the i-th control center; and
(L4) if $K_i\not\subseteq K_i^{\mathrm{b}}$ and there is no attack set R_i, with $|R_i|\leq|K_i|$, such that $(E_i, A_i, [B_i^{\mathrm{b}}\ B_{K_i}\ B_{R_i}], C_i, [D_{K_i}\ D_{R_i}])$ has invariant zeros, then K_i is identifiable by the i-th control center.

Whereas limitations (L1) and (L2) also apply to any centralized attack detection and identification monitor, limitations (L3) and (L4) arise by naively treating the neighboring signals as unknown inputs. Since, in general, 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, we can further state that

(L5) any (boundary) attack set $K_i\subseteq K_i^{\mathrm{b}}$ is not detectable and not identifiable by the i-th control center, and
(L6) any (external) attack set $K\setminus K_i$ is not detectable and not identifiable by the i-th control center.

We remark that, following our graph-theoretic analysis in [1, Section IV], the attack K_i is generically identifiable by the i-th control center if the number of attacks $|K_i|$ on the i-th subsystem is sufficiently small, the internal connectivity of the i-th subsystem (size of linking between unknown inputs/attacks and outputs) is sufficiently high, and the number of unknown signals $|K_i^{\mathrm{b}}|$ from neighboring subsystems is sufficiently small. These criteria can ultimately be used to select an attack-resilient partitioning of a cyber-physical system.

D. Cooperative attack identification

In this section we improve upon the naive fully decoupled method presented in Subsection IV-C and propose an identification method based upon a divide and conquer procedure with cooperation. This method consists of the following steps.
(S1: estimation and communication) 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 $B_i^{\mathrm{b}} f_i(t)$. For this task we build upon existing unknown-input estimation algorithms (see the Appendix for a constructive procedure). Assume that the state $x_i(t)$ is reconstructed modulo some subspace $\mathcal{F}_i$.[2] Let $F_i = \mathrm{Basis}(\mathcal{F}_i)$, and let $x_i(t) = \tilde{x}_i(t) + \hat{x}_i(t)$, where $\hat{x}_i(t)$ is the estimate computed by the i-th control center, and $\tilde{x}_i(t) \in \mathcal{F}_i$. Assume that each control center i transmits the estimate $\hat{x}_i(t)$ and the uncertainty subspace F_i to every neighboring control center.
(S2: residual generation) Observe that each input signal $A_{ij}x_j(t)$ can be written as $A_{ij}x_j(t) = A_{ij}\tilde{x}_j(t) + A_{ij}\hat{x}_j(t)$, where $\tilde{x}_j(t) \in \mathcal{F}_j$. Then, after carrying out step (S1), only the inputs $A_{ij}\tilde{x}_j(t)$ are unknown to the i-th control center, while the inputs $A_{ij}\hat{x}_j(t)$ are known to the i-th center due to communication. Let $B_i^{\mathrm{b}}F_i = [A_{i1}F_1 \cdots A_{i,i-1}F_{i-1}\, A_{i,i+1}F_{i+1} \cdots A_{iN}F_N]$, and rewrite the signal $B_i^{\mathrm{b}}\tilde{x}(t)$ as $B_i^{\mathrm{b}}\tilde{x}(t) = B_i^{\mathrm{b}}F_i f_i(t)$, for some unknown signal $f_i(t)$. Then the dynamics of the i-th subsystem read as

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

Analogously to the filter presented in Theorem 4.5 for the attack signature (B_K, D_K), consider now the following filter (in appropriate coordinates) for (28) for the signature $(B_i^{\mathrm{b}}F_i, 0)$

$$\begin{aligned} E_i\dot{w}_i(t) &= (A_i + L_iC_i)w_i(t) - Ly(t) + B_i^{\mathrm{b}}\bar{x}(t), \\ r_i(t) &= Mw_i(t) - Hy(t), \end{aligned} \tag{29}$$

where L_i is the injection matrix associated with the conditioned invariant subspace generated by $B_i^{\mathrm{b}}F_i$, with $(E_i, A_i + L_iC_i)$ Hurwitz, and $\bar{x}(t)$ is the state transmitted to i by its neighbors. Notice that, in the absence of attacks in the regions $\mathcal{N}_i^{\mathrm{in}}$, we have $B_i^{\mathrm{b}}\bar{x}(t) = B_i^{\mathrm{b}}\hat{x}(t)$. Finally, let the matrices M and H in (29) be chosen so that the input $B_i^{\mathrm{b}}F_i f_i(t)$ does not affect the residual $r_i(t)$.[3] Consider the filter error $e_i(t) = w_i(t) - x_i(t)$, and notice that

$$\begin{aligned} E_i\dot{e}_i(t) &= (A_i + L_iC_i)e_i(t) + B_i^{\mathrm{b}}(\bar{x}(t) - \hat{x}(t)) - B_{K_i}u_{K_i}(t) \\ &\quad - B_i^{\mathrm{b}}F_i f_i(t), \\ r_i(t) &= Me_i(t), \end{aligned} \tag{30}$$

[2]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 $\mathrm{Ker}(C_i)$ [32].

[3]See Section IV-B for a detailed construction of this type of filter.

(S3: cooperative residual analysis) We next state a key result for our distributed identification procedure.

Lemma 4.6: ***(Characterization of nonzero residuals)*** Let each control center implement the distributed identification filter (29) with $w_i(0) = x_i(0)$. Assume that the attack K affects only the i-th subsystem, that is $K = K_i$. Assume that $(E_i, A_i, [B_i^{\mathrm{b}}F_i\, B_{K_i}], C_i)$ and $(E_i, A_i, B_i^{\mathrm{b}}, C_i)$ have no invariant zeros. Then,

(i) $r_i(t) \neq 0$ at some time t, and
(ii) either $r_j(t) = 0$ for all $j \in \mathcal{N}_i^{\mathrm{out}}$ at all times t, or $r_j(t) \neq 0$ for all $j \in \mathcal{N}_i^{\mathrm{out}}$ at some time t.

Proof: Notice that the estimation computed by a control center is correct provided that its area is not under attack. In other words, since $K = K_i$, we have that $B_i^{\mathrm{b}}\hat{x}(t) = B_i^{\mathrm{b}}\bar{x}(t)$ in (30). Since $(E_i, A_i, [B_i^{\mathrm{b}}F_i\, B_{K_i}], C_i)$ has no invariant zeros, statement (i) follows. In order to prove statement (ii), consider the following two cases: the i-th control center provides the correct estimation $\hat{x}_i(t) = \bar{x}_i(t)$ or an incorrect estimation $\hat{x}_i(t) \neq \bar{x}_i(t)$. For instance, if $\mathrm{Im}(B_{K_i}) \subseteq \mathrm{Im}(B_i^{\mathrm{b}})$, that is, the attack set K_i lies on the boundary of the i-th area, then $\hat{x}_i(t) = \bar{x}_i(t)$. Notice that, if $\hat{x}_i(t) = \bar{x}_i(t)$, then each residual $r_j(t)$, $j \neq i$, is identically zero since the associated residual dynamics (30) evolve as an autonomous system without inputs. Suppose now that $\hat{x}_i(t) \neq \bar{x}_i(t)$. Notice that $B_i^{\mathrm{b}}F_i f_i(t) + B_i^{\mathrm{b}}(\hat{x}(t) - \bar{x}(t)) \in \mathrm{Im}(B_i^{\mathrm{b}})$. Then, since $(E_i, A_i, B_i^{\mathrm{b}}, C_i)$ has no invariant zeros, each residual $r_j(t)$ is nonzero for some t. ■

As a consequence of Lemma 4.6 the region under attack can be identified through a distributed procedure. Indeed, the i-th area is safe if either of the following two criteria is satisfied:

(C1) the associated residual $r_i(t)$ is identically zero, or
(C2) the neighboring areas $j \in \mathcal{N}_i^{\mathrm{out}}$ feature both zero and nonzero residuals $r_j(t)$.

Consider now the case of several simultaneously corrupted subsystems. Then, if the graphical distance between any two corrupted areas is at least 2, that is, if there are at least two uncorrupted areas between any two corrupted areas, corrupted areas can be identified via our distributed method and criteria (C1) and (C2). An upper bound on the maximum number of identifiable concurrent corrupted areas can consequently be derived (see the related *set packing* problem in [30]).
(S4: local identification) Once the corrupted regions have been identified, the identification method in Section IV is used to identify the local attack set.

Lemma 4.7: ***(Local identification)*** Consider the decoupled system (28). Assume that the i-th region is under the attack K_i whereas the neighboring regions $\mathcal{N}_i^{\mathrm{out}}$ are uncorrupted. Assume that each control center $j \in \mathcal{N}_i^{\mathrm{in}}$ transmits the estimate $\hat{x}_j(t)$ and the uncertainty subspace F_i to the i-th control center. Then, the attack set K_i is identifiable by the i-th control center if $(E_i, A_i, [B_i^{\mathrm{b}}F_i\ B_{K_i}\ B_{R_i}], C_i, [D_{K_i}\ D_{R_i}])$ has no invariant zeros for any attack set R_i, with $|R_i| \leq |K_i|$.

Proof: Notice that each control center j, with $j \neq i$, can correctly estimate the state $x_j(t)$ modulo $\mathcal{F}_j$. Since this estimation is transmitted to the i-th control center, the statement follows from [1, Theorem 3.4]. ■

The final identification procedure **(S4)** is implemented only on the corrupted regions. Consequently, the combinatorial complexity of our distributed identification procedure is

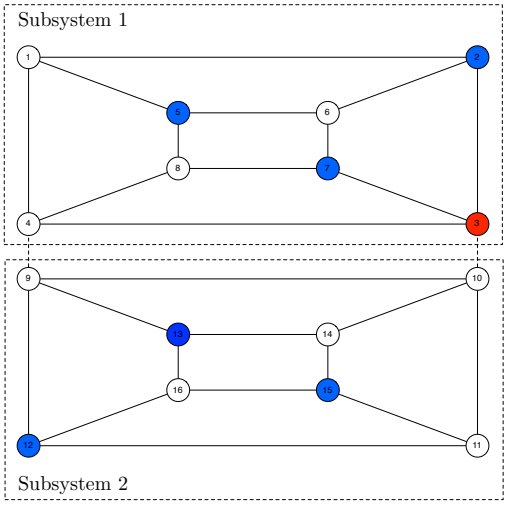

Fig. 6. This figure shows a network composed of two subsystems. A control center is assigned to each subsystem. Each control center knows only the dynamics of its local subsystem. The state of the blue nodes $\{2, 5, 7, 12, 13, 15\}$ is continuously measured by the corresponding control center, and the state of the red node $\{3\}$ is corrupted by an attacker. The decoupled identification procedure presented in Subsection IV-C fails at detecting the attack. Instead, by means of our cooperative identification procedure, the attack can be detected and identified via distributed computation.

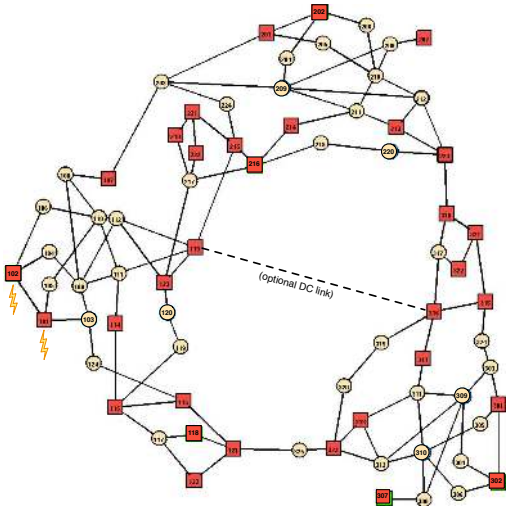

Fig. 7. This figure illustrates the IEEE RTS96 power network [35]. The dynamics of the generators $\{101, 102\}$ are affected by an attacker.

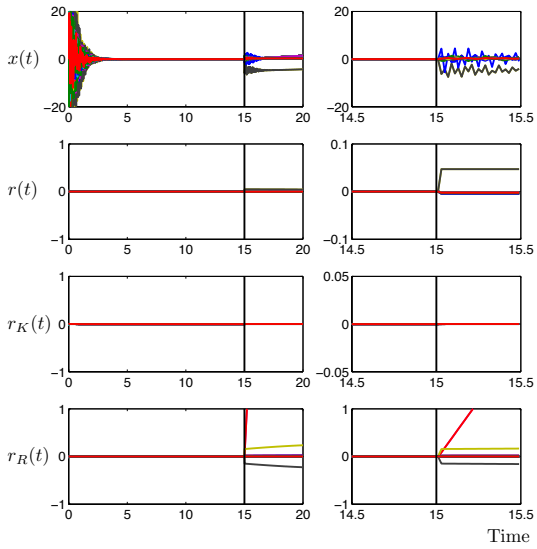

Fig. 8. In this figure we report our simulation results for the case of linear network dynamics without noise and for the proposed detection monitor (2) and identification monitor (25), respectively. The state trajectory $x(t)$ consists of the generators angles and frequencies. The detection residual $r(t)$ becomes nonzero after time 15s, and it reveals the presence of the attack. The identification residual $r_K(t)$ is identically zero even after time 15s, and it reveals that the attack set is $K = \{101, 102\}$. The identification residual $r_R(t)$ is nonzero after time 15s, and it reveals that R is not the attack set.

$\sum_{i=1}^{\ell} \binom{n_i+p_i}{|K_i|}$, where ℓ is the number of corrupted regions. Hence, the distributed identification procedure greatly reduces the combinatorial complexity of the centralized procedure presented in Subsection IV-B, which requires the implementation of $\binom{n+p}{|K|}$ filters. Finally, the assumptions of Lemma 4.6 and Lemma 4.7 clearly improve upon the limitations (L3) and (L4) of the naive decoupled approach presented in Subsection IV-C. We conclude this section with an example showing that, contrary to the limitation (L5) of the naive fully decoupled approach, boundary attacks $K_i \subseteq K_i^{\mathrm{b}}$ can be identified by our cooperative attack identification method.

Example 2: ***(An example of cooperative identification)***

Consider the sensor network in Fig. 6, where the state of the blue nodes $\{2, 5, 7, 12, 13, 15\}$ is measured and the state of the red node $\{3\}$ is corrupted by an attacker. Assume that the network evolves according to nonsingular, linear, time-invariant dynamics. Assume further that the network has been partitioned into the two areas $V_1 = \{1, \ldots, 8\}$ and $V_2 = \{9, \ldots, 16\}$ and at most one area is under attack. Since $\{3, 4\}$ are the boundary nodes for the first area, the attack set $K = 3$ is neither detectable nor identifiable by the two control centers via the fully decoupled procedure in Section IV-C.

Consider now the second subsystem with the boundary nodes $K_2^{\mathrm{b}} = \{9, 10\}$. It can be shown that, generically, the second subsystem with unknown input $B_2^{\mathrm{b}} f_2(t)$ has no invariant zeros; see [1, Section V]. Hence, the state of the second subsystem can be entirely reconstructed. Analogously, since the attack is on the boundary of the first subsystem, the state of the first subsystem can be reconstructed, so that the residual $r_2(t)$ is identically zero; see Lemma 4.6.

Suppose that the state of the second subsystem is continuously transmitted to the control center of the first subsystem. Then, the only unknown input in the first subsystem is due to the attack, which is now generically detectable and identifiable, since the associated system has no invariant zeros; see Lemma 4.7. We conclude that our cooperative identification procedure outperforms the decoupled counterpart in Section IV-C. □

V. A CASE STUDY: THE IEEE RTS96 SYSTEM

In this section we apply our centralized attack detection and identification methods to the IEEE RTS96 power network [35] illustrated in Fig. 7. In particular, we first consider the nominal case, in which the power network dynamics evolve as nominal linear time-invariant descriptor system, as described in [1, Section II.C]. Second, we consider the case of additive state and measurement noise, and we show the robustness of the attack detection and identification monitors. Third, we

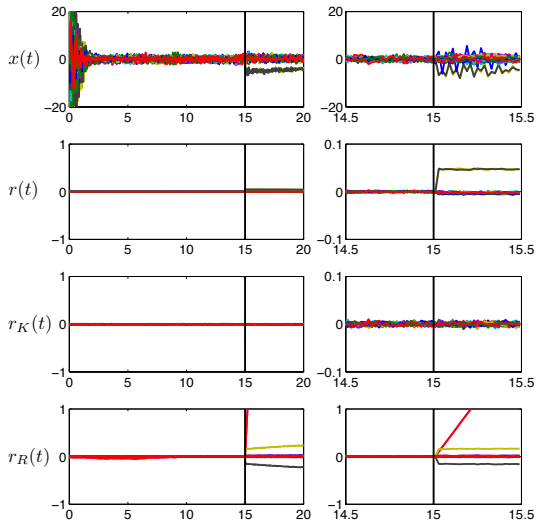

Fig. 9. In this figure we report our simulation results for the case of linear network dynamics driven by state and measurements noise. For this case, we choose the output injection matrices of the detection and identification filters as the corresponding optimal Kalman gain (see Remark 1 and Remark 5). Due to the presence of noise, the residuals deviate from their nominal behavior reported in Fig. 8. Although the attack is clearly still detectable and identifiable, additional statistical tools such as hypothesis testing [23] may be adopted to analyze the residuals $r(t)$, $r_K(t)$, and $r_R(t)$.

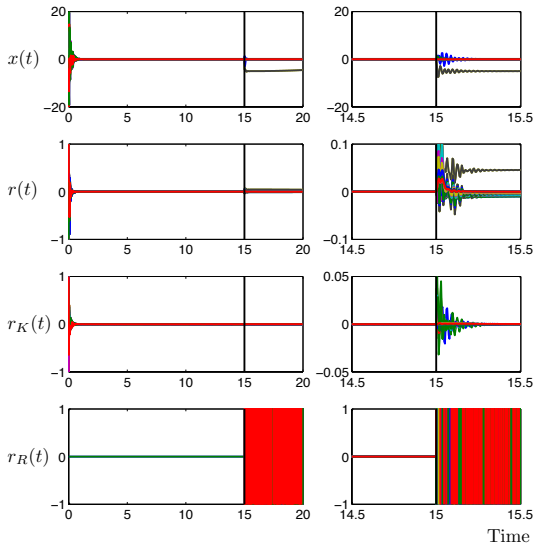

Fig. 10. In this figure we report our simulation results for the case of nonlinear network dynamics without noise. For this case, the detection and identification filters are designed for the *nominal linearized dynamics* with output injection matrices as the corresponding optimal Kalman gain (see Remark 1 and Remark 5). Despite the presence of unmodeled nonlinear dynamics, the residuals reflect their nominal behavior reported in Fig. 8.

consider the case of nonlinear differential-algebraic power network dynamics and show the effectiveness of our methods in the presence of unmodeled nonlinear dynamics.

For our numerical studies, we assume the angles and frequencies of every generator to be measured. Additionally, we let the attacker affect the angles of the generators $\{101, 102\}$ with a random signal starting from time 15s. Since the considered power network dynamics are of index one, the filters are implemented using the nonsingular Kron-reduced system representation [1, Section III.D]. The results of our simulations are in Fig. 8, Fig. 9, and Fig. 10. In conclusion, our centralized detection and identification filters appears robust to state and measurements noise and unmodeled dynamics.

VI. Conclusion

For cyber-physical systems modeled by linear time-invariant descriptor systems, we proposed attack detection and identification monitors. In particular, for the detection problem we developed both centralized and distributed monitors. These monitors are optimal, in the sense that they detect every detectable attack. For the attack identification problem, we developed an optimal centralized monitor and a sub-optimal distributed method. Our centralized attack identification monitor relies upon a combinatorial machinery. Our distributed attack identification monitor, instead, is computationally efficient and achieves guaranteed identification of a class of attacks, which we characterize. Finally, we provided several examples to show the effectiveness and the robustness of our methods against uncertainties and unmodeled dynamics.

Appendix

In this section we present an algebraic technique to reconstruct the state of a descriptor system. Our method builds upon the results presented in [17]. Consider the descriptor model (1) written in the form (see [1, Section IV.C])

$$\begin{aligned} \dot{x}_1(t) &= A_{11}x_1(t) + A_{12}x_2(t) + B_1u(t)\,, \\ 0 &= A_{21}x_1(t) + A_{22}x_2(t) + B_2u(t)\,, \\ y(t) &= C_1x_1(t) + C_2x_2(t) + Du(t)\,. \end{aligned} \tag{A-1}$$

We aim at characterizing the largest subspace of the state space of (A-1) that can be reconstructed through the measurements $y(t)$. Consider the associated nonsingular system

$$\begin{aligned} \dot{\tilde{x}}_1(t) &= A_{11}\tilde{x}_1(t) + B_1\tilde{u}(t) + A_{12}\tilde{x}_2(t), \\ \tilde{y}(t) &= \begin{bmatrix} \tilde{y}_1(t) \\ \tilde{y}_2(t) \end{bmatrix} = \begin{bmatrix} A_{21} \\ C_1 \end{bmatrix} \tilde{x}_1(t) + \begin{bmatrix} A_{22} & B_2 \\ C_2 & D \end{bmatrix} \begin{bmatrix} \tilde{x}_2(t) \\ \tilde{u}(t) \end{bmatrix}. \end{aligned} \tag{A-2}$$

Recall from [32, Section 4] that the state of the system (A-2) can be reconstructed modulo its largest controlled invariant subspace $\mathcal{V}_1^*$ contained in the null space of the output matrix.

Lemma 6.1: ***(Reconstruction of the state*** $x_1(t)$***)*** Let $\mathcal{V}_1^*$ be the largest controlled invariant subspace of the system (A-2). The state $x_1(t)$ of the system (A-1) can be reconstructed only modulo $\mathcal{V}_1^*$ through the measurements $y(t)$.

Proof: We start by showing that for every $x_1(0) \in \mathcal{V}_1^*$ there exist $x_2(t)$ and $u(t)$ such that $y(t)$ is identically zero. Due to the linearity of (A-1), we conclude that the projection of $x_1(t)$ onto $\mathcal{V}_1^*$ cannot be reconstructed. Notice that for every $\tilde{x}_1(0)$, $\tilde{x}_2(t)$, and $\tilde{u}(t)$ yielding $\tilde{y}_1(t) = 0$ at all times, the state trajectory $[\tilde{x}_1(t)\ \tilde{x}_2(t)]$ is a solution to (A-1) with input $u(t) = \tilde{u}(t)$ and output $y(t) = \tilde{y}_2(t)$. Since for every $\tilde{x}_1(0) \in \mathcal{V}_1^*$, there exists $\tilde{x}_2(t)$ and $\tilde{u}(t)$ such that $\tilde{y}(t)$ is identically zero, we conclude that every state $x_1(0) \in \mathcal{V}_1^*$ cannot be reconstructed.

We now show that the state $x_1(t)$ can be reconstructed modulo $\mathcal{V}_1^*$. Let $x_1(0)$ be orthogonal to $\mathcal{V}_1^*$, and let $x_1(t)$, $x_2(t)$, and $y(t)$ be the solution to (A-1) subject to the input $u(t)$. Notice that $\tilde{x}_1(t) = x_1(t)$, $\tilde{y}_1(t) = 0$, and $\tilde{y}_2(t) = y(t)$

is the solution to (A-2) with inputs $\tilde{x}_2(t) = x_2(t)$ and $\tilde{u}(t) = u(t)$. Since $\tilde{x}_1(0)$ is orthogonal to $\mathcal{V}_1^*$, we conclude that $\tilde{x}_1(0) = x_1(0)$, and in fact the subspace $(\mathcal{V}^*)^\perp$, can be reconstructed through the measurements $\tilde{y}_2(t) = y(t)$. ■

In Lemma 6.1 we show that the state $x_1(t)$ of (A-1) can be reconstructed modulo $\mathcal{V}_1^*$. We now show that the state $x_2(t)$ can generally not be completely reconstructed.

Lemma 6.2: **(Reconstruction of the state $x_2(t)$)** Let $\mathcal{V}_1^* = \mathrm{Im}(V_1)$ be the largest controlled invariant subspace of the system (A-2). The state $x_2(t)$ of the system (A-1) can be reconstructed only modulo $\mathcal{V}_2^* = A_{22}^{-1}\,\mathrm{Im}([A_{21}V_1\ B_2])$.

Proof: Let $x_1(t) = \bar{x}_1(t) + \hat{x}_1(t)$, where $\bar{x}_1(t) \in \mathcal{V}_1^*$ and $\hat{x}_1(t)$ is orthogonal to $\mathcal{V}_1^*$. From Lemma 6.1, the signal $\hat{x}_1(t)$ can be entirely reconstructed via $y(t)$. Notice that

$$\begin{aligned} 0 &= A_{21}x_1(t) + A_{22}x_2(t) + B_2u(t), \\ &= A_{21}V_1v_1(t) + A_{21}\hat{x}_1(t) + A_{22}x_2(t) + B_2u(t). \end{aligned}$$

Let W be such that $\mathrm{Ker}(W) = \mathrm{Im}([A_{21}V_1\ \ B_2])$. Then, $0 = WA_{21}\hat{x}_1(t) + WA_{22}x_2(t)$, and hence $x_2(t) = \bar{x}_2(t) + \hat{x}_2(t)$, where $\hat{x}_2(t) = (WA_{22})^\dagger WA_{21}\hat{x}_1(t)$, and $\bar{x}_2(t) \in \mathrm{Ker}(WA_{22}) = A_{22}^{-1}\,\mathrm{Im}([A_{21}V_1\ B_2])$. The statement follows. ■

To conclude the paper, we remark the following points. First, our characterization of $\mathcal{V}_1^*$ and $\mathcal{V}_2^*$ is equivalent to the definition of *weakly unobservable* subspace in [18], and of maximal *output-nulling* subspace in [33]. Hence, we proposed an optimal state estimator for our distributed attack identification procedure, and the matrix V_i in **(S1: estimation and communication)** can be computed as in [18], [33]. Second, a reconstruction of $x_1(t)$ modulo $\mathcal{V}_1^*$ and $x_2(t)$ modulo $\mathcal{V}_2^*$ can be obtained through standard algebraic techniques [32]. Third and finally, Lemma 6.1 and Lemma 6.2 extend the results in [17] by characterizing the subspaces of the state space that can be reconstructed with an algebraic method by processing the measurements $y(t)$ and their derivatives.

REFERENCES

[1] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack Detection and Identification in Cyber-Physical Networks – Part I: Models and Fundamental Limitations," *IEEE Transactions on Automatic Control*, 2012, Submitted.

[2] A. R. Metke and R. L. Ekl, "Security technology for smart grid networks," *IEEE Transactions on Smart Grid*, vol. 1, no. 1, pp. 99–107, 2010.

[3] H. Khurana, "Cybersecurity: A key smart grid priority," *IEEE Smart Grid Newsletter*, Aug. 2011.

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

[5] G. E. Apostolakis and D. M. Lemon, "A screening methodology for the identification and ranking of infrastructure vulnerabilities due to terrorism," *Risk Analysis*, vol. 25, no. 2, pp. 361–376, 2005.

[6] S. Sundaram and C. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Transactions on Automatic Control*, vol. 56, no. 7, pp. 1495–1508, 2011.

[7] 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.

[8] F. Hamza, P. Tabuada, and S. Diggavi, "Secure state-estimation for dynamical systems under active adversaries," in *Allerton Conf. on Communications, Control and Computing*, Sep. 2011.

[9] 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.

[10] E. Scholtz, "Observer-based monitors and distributed wave controllers for electromechanical disturbances in power systems," Ph.D. dissertation, Massachusetts Institute of Technology, 2004.

[11] A. Domınguez-Garcıa and S. Trenn, "Detection of impulsive effects in switched DAEs with applications to power electronics reliability analysis," in *IEEE Conf. on Decision and Control*, Atlanta, GA, USA, Dec. 2010, pp. 5662–5667.

[12] F. Pasqualetti, F. Dörfler, and F. Bullo, "Cyber-physical attacks in power networks: Models, fundamental limitations and monitor design," in *IEEE Conf. on Decision and Control and European Control Conference*, Orlando, FL, USA, Dec. 2011, pp. 2195–2201.

[13] G. Dan and H. Sandberg, "Stealth attacks and protection schemes for state estimators in power systems," in *IEEE Int. Conf. on Smart Grid Communications*, Gaithersburg, MD, USA, Oct. 2010, pp. 214–219.

[14] R. B. Bobba, K. M. Rogers, Q. Wang, H. Khurana, K. Nahrstedt, and T. J. Overbye, "Detecting false data injection attacks on DC state estimation," in *First Workshop on Secure Control Systems*, Stockholm, Sweden, Apr. 2010.

[15] M.-A. Massoumnia, G. C. Verghese, and A. S. Willsky, "Failure detection and identification," *IEEE Transactions on Automatic Control*, vol. 34, no. 3, pp. 316–321, 1989.

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

[17] F. J. Bejarano, T. Floquet, W. Perruquetti, and G. Zheng, "Observability and detectability analysis of singular linear systems with unknown inputs," in *IEEE Conf. on Decision and Control and European Control Conference*, Orlando, FL, USA, Dec. 2011, pp. 4005–4010.

[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] L. Dai, *Singular Control Systems*. Springer, 1989.

[20] A. V. Medvedev and H. T. Toivonen, "Feedforward time-delay structures in state estimation-finite memory smoothing and continuous deadbeat observers," *IEE Proceedings. Control Theory & Applications*, vol. 141, no. 2, pp. 121–129, 1994.

[21] T. Raff and F. Allgöwer, "An observer that converges in finite time due to measurement-based state updates," in *IFAC World Congress*, Seoul, Korea, Jul. 2008, pp. 2693–2695.

[22] S. Skogestad and I. Postlethwaite, *Multivariable Feedback Control Analysis and Design*, 2nd ed. Wiley, 2005.

[23] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, 1993.

[24] M. Vidyasagar, *Input-Output Analysis of Large-Scale Interconnected Systems: Decomposition, Well-Posedness and Stability*. Springer, 1981.

[25] E. Lelarasmee, A. E. Ruehli, and A. L. Sangiovanni-Vincentelli, "The waveform relaxation method for time-domain analysis of large scale integrated circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 1, no. 3, pp. 131–145, 1982.

[26] M. L. Crow and M. D. Ilić, "The waveform relaxation method for systems of differential/algebraic equations," *Mathematical and Computer Modelling*, vol. 19, no. 12, pp. 67–84, 1994.

[27] Z. Z. Bai and X. Yang, "On convergence conditions of waveform relaxation methods for linear differential-algebraic equations," *Journal of Computational and Applied Mathematics*, vol. 235, no. 8, pp. 2790–2804, 2011.

[28] Y. Ohta, D. Šiljak, and T. Matsumoto, "Decentralized control using quasi-block diagonal dominance of transfer function matrices," *IEEE Transactions on Automatic Control*, vol. 31, no. 5, pp. 420–430, 1986.

[29] R. D. Zimmerman, C. E. Murillo-Sánchez, and D. Gan, "MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12–19, 2011.

[30] M. R. Garey and D. S. Johnson, *Computers and Intractability*. Springer, 1979.

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

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

[33] F. L. Lewis, "Geometric design techniques for observers in singular systems," *Automatica*, vol. 26, no. 2, pp. 411–415, 1990.

[34] K. D. Ikramov, "Matrix pencils: Theory, applications, and numerical methods," *Journal of Mathematical Sciences*, vol. 64, no. 2, pp. 783–853, 1993.

[35] C. Grigg, P. Wong, P. Albrecht, R. Allan, M. Bhavaraju, R. Billinton, Q. Chen, C. Fong, S. Haddad, S. Kuruganty, W. Li, R. Mukerji, D. Patton, N. Rau, D. Reppen, A. Schneider, M. Shahidehpour, and C. Singh, "The IEEE Reliability Test System - 1996. A report prepared by the Reliability Test System Task Force of the Application of Probability Methods Subcommittee," *IEEE Transactions on Power Systems*, vol. 14, no. 3, pp. 1010–1020, 1999.