Attack Detection and Identification in Cyber-Physical Systems

Fabio Pasqualetti, Florian Dörfler, and Francesco Bullo

Abstract—Cyber-physical systems are ubiquitous in power systems, transportation networks, industrial control processes, and critical infrastructures. These systems need to operate reliably in the face of unforeseen failures and external malicious attacks. In this paper (i) we propose a mathematical framework for cyber-physical systems, attacks, and monitors; (ii) we characterize fundamental monitoring limitations from system-theoretic and graph-theoretic perspectives; and (iii) we design centralized and distributed attack detection and identification monitors. Finally, we validate our findings through compelling examples.

I. INTRODUCTION

Cyber-physical systems integrate physical processes, computational resources, and communication capabilities. Examples of cyber-physical systems include transportation networks, power generation and distribution networks, water and gas distribution networks, and advanced communication systems. As recently highlighted by the Maroochy water breach [1] in March 2000, multiple recent power blackouts in Brazil [2], the SQL Slammer worm attack on the Davis-Besse nuclear plant in January 2003 [3], the StuxNet computer worm [4] in June 2010, and by various industrial security incidents [5], cyber-physical systems are prone to failures and attacks on their physical infrastructure, and cyber attacks on their data management and communication layer.

Concerns about security of control systems are not new, as the numerous manuscripts on systems fault detection, isolation, and recovery testify [6], [7]. Cyber-physical systems, however, suffer from specific vulnerabilities which do not affect classical control systems, and for which appropriate detection and identification techniques need to be developed. For instance, the reliance on communication networks and standard communication protocols to transmit measurements and control packets increases the possibility of intentional and worst-case attacks against physical plants. On the other hand, information security methods, such as authentication, access control, and message integrity, appear inadequate for a satisfactory protection of cyber-physical systems. Indeed, these security methods do not exploit the compatibility of the measurements with the underlying physical process or the control mechanism, and they are therefore ineffective against insider attacks targeting the physical dynamics [1].

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.

Related work. The analysis of vulnerabilities of cyber-physical systems to external attacks has received increasing attention in the last years. The general approach has been to study the effect of specific attacks against particular systems. For instance, in [8] deception and denial of service attacks against a networked control system are defined, and, for the latter ones, a countermeasure based on semi-definite programming is proposed. Deception attacks refer to the possibility of compromising the integrity of control packets or measurements, and they are cast by altering the behavior of sensors and actuators. Denial of service attacks, instead, compromise the availability of resources by, for instance, jamming the communication channel. In [9] false data injection attacks against static state estimators are introduced. False data injection attacks are specific deception attacks in the context of static estimators. It is shown that undetectable false data injection attacks can be designed even when the attacker has limited resources. In a similar fashion, stealthy deception attacks against the Supervisory Control and Data Acquisition system are studied, among others, in [10]. In [11] the effect of replay attacks on a control system is discussed. Replay attacks are cast by hijacking the sensors, recording the readings for a certain amount of time, and repeating such readings while injecting an exogenous signal into the system. It is shown that these attacks can be detected by injecting a signal unknown to the attacker into the system. In [12] the effect of covert attacks against control systems is investigated. Specifically, a parameterized decoupling structure allows a covert agent to alter the behavior of the physical plant while remaining undetected from the original controller. In [13] a resilient control problem is studied, in which control packets transmitted over a network are corrupted by a human adversary. A receding-horizon Stackelberg control law is proposed to stabilize the control system despite the attack. Recently the problem of estimating the state of a linear system with corrupted measurements has been studied [14]. More precisely, the maximum number of tolerable faulty sensors is characterized, and a decoding algorithm is proposed to detect corrupted measurements. Finally, security issues of specific cyber-physical systems have received considerable attention, such as power networks [15]–[19], linear networks with misbehaving components [20], [21], and water networks [22], [23].

Contributions. The contributions of this paper are as follows. First, we describe a unified modeling framework for cyber-physical systems and attacks (Section II). Motivated by existing cyber-physical systems and existing attack scenarios, we model a cyber-physical system under attack as a descriptor system subject to unknown inputs affecting the state and
the measurements. For our model, we define the notions of detectability and identifiability of an attack by its effect on output measurements. Informed by the classic work on geometric control theory [24], [25], our framework includes the deterministic static detection problem considered in [9], [10], and the prototypical deception and denial of service [8], stealth [16], (dynamic) false-data injection [26], replay attacks [11], and covert attacks [12] as special cases.

Second, we show the fundamental limitations of a class of monitors (Section III-A). This class includes the widely-studied static, dynamic, and active monitors. We prove that (i) a cyber-physical attack is undetectable by our monitors if and only if the attackers’ signal excites uniquely the zero dynamics of the input/output system, and (ii) that undetectable and unidentifiable attacks can be cast without knowing monitoring signals or the system noise.

Third, we provide a graph-theoretic characterization of undetectable attacks (Section III-B). We borrow some tools from the theory of structured systems, and we identify conditions on the system interconnection structure for the existence of undetectable attacks. These conditions are generic, in the sense that they hold for almost all numerical systems with the same structure, and they can be efficiently verified. As a complementary result, we extend a result of [27] on structural left-invertibility to regular descriptor systems. Finally, with respect to our earlier work [20], [21], we consider continuous-time descriptor systems, and we include parameters constraints.

Fourth, we design centralized and distributed monitors (Section IV). Our centralized monitors and our distributed detection monitor are complete, in the sense that they detect and identify every (detectable and identifiable) attack. Our centralized monitors are designed by leveraging on tools from geometric control theory, while our distributed detection monitor relies upon techniques from distributed control and parallel computation. Additionally, we characterize the computational complexity of the attack identification problem.

Fifth and finally, we illustrate the potential impact of our theoretical findings through compelling examples. In particular, (i) we design an undetectable state attack to destabilize the WSSC 3-machine 6-bus power system, (ii) we characterize the resilience to output attacks of the IEEE 14 bus system, (iii) we show the detection performance of our distributed monitor on the IEEE 118 bus system, and (iv) we use the RTS 96 network model to illustrate that our methods are effective also in the presence of system noise, nonlinearities, and modeling uncertainties. Through these examples we show the advantages of dynamic monitors against static ones, and we provide insight on the design of attacks.

II. Problem Setup and Preliminary Results

In this paper we model cyber-physical systems under attack as linear time-invariant descriptor systems subject to unknown inputs. This simplified model neglects system nonlinearities and the presence of noise in the dynamics and the measurements. Nevertheless, such a simplified model has long proven useful in studying stability, faults, and attacks in, for instance, power networks, sensor networks, and water networks. It is our premise that more detailed models are unlikely to change the basic conclusions of this work.

Model of cyber-physical systems under attack. We consider the descriptor system

\[
E \dot{x}(t) = Ax(t) + Bu(t), \quad y(t) = Cx(t) + Du(t),
\]

where \(x(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m, y(t) \in \mathbb{R}^p, E \in \mathbb{R}^{n \times n}, A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}, C \in \mathbb{R}^{p \times n}, \) and \(D \in \mathbb{R}^{p \times m}. \) Here the matrix \(E\) is possibly singular, and the inputs \(Bu\) and \(Du\) are unknown signals describing disturbances affecting the plant. Besides reflecting the genuine failure of systems components, these disturbances model the effect of attacks against the cyber-physical system. Without loss of generality, we assume that each state and output variable can be independently compromised by an attacker, and we let \(B = [I_{n \times n} \ 0_{n \times p}]\) and \(D = [0_{p \times n} \ I_{p \times p}]\).

The attack signal \(t \mapsto u(t) \in \mathbb{R}^{n+p}\) depends upon the specific attack strategy. In particular, if \(K \subseteq \{1, \ldots, n+p\}\) is the attack set, with \(|K| = k\), then all (and 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 sometimes use \(u_K\) to denote the attack signal, that is, the subvector of \(u\) indexed by \(K\). Accordingly, the pair \((B_K, D_K)\), where \(B_K\) and \(D_K\) are the submatrices of \(B\) and \(D\) with columns in \(K\), denote the attack signature. Hence, \(B_u(t) = B_K u_K(t)\), and \(D_u(t) = D_K u_K(t)\). Since the matrix \(E\) may be singular, we make the following assumptions on system (1):

(A1) the pair \((E, A)\) is regular, that is, the determinant \(|sE - A|\) does not vanish identically;

(A2) the initial condition \(x(0) \in \mathbb{R}^n\) is consistent, that is, \((Ax(0) + Bu(0)) \in \text{Im}(E)\); and

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

Assumption (A1) assures the existence of a unique solution \(x(t)\) to (1). Assumptions (A2) and (A3) guarantee smoothness of the state trajectory \(x(t)\) and the measurements \(y(t)\). [28, Lemma 2.5]. If assumptions (A2) and (A3) are dropped, then there are inconsistent initial conditions and impulsive inputs by which a powerful attacker can avoid detection; see Remark 4. Throughout the paper, the cardinality \(k\) of the attack set, or an upper bound, is assumed to be known.

Remark 1: (Examples of cyber-physical systems requiring advanced security mechanisms) Future power grids will combine physical dynamics with a sophisticated coordination infrastructure. The cyber-physical security of the grid has been identified as an issue of primary concern, see [19], [29] and [10], [16]–[18], [30], [31].

Mass transport networks are cyber-physical systems, such as gas transmission and distribution networks [32], large-scale process engineering plants [33], and water networks. Examples of water networks include open channel flows [34] for irrigation purposes and municipal water networks [35], [3].

1The results stated in this paper for continuous-time descriptor systems hold also for discrete-time descriptor systems and nonsingular systems. Moreover, due to linearity of (1), known inputs do not affect our results.
[36]. The vulnerability of open channel networks to cyber-physical attacks has been studied in [12], [22], and municipal water networks are also known to be susceptible to attacks on the hydraulics [1] and biochemical contamination threats [23].

Power networks and mass transport network under attack can be modeled by descriptor systems with unknown inputs. For instance, the small-signal version of the classical structure-preserving power network model reads as [30], [31]

\[
\begin{pmatrix}
I & 0 & 0 \\
0 & M_g & 0 \\
0 & 0 & 0
\end{pmatrix} \begin{pmatrix}
\dot{\delta}(t) \\
\dot{\omega}(t) \\
\dot{\theta}(t)
\end{pmatrix} = \begin{pmatrix}
0 & -I & 0 \\
D_g & L_g & 0 \\
0 & L_{\omega} & 0
\end{pmatrix} \begin{pmatrix}
\delta(t) \\
\omega(t) \\
\theta(t)
\end{pmatrix} + \begin{pmatrix}
0 \\
P_w(t) \\
P_t(t)
\end{pmatrix},
\]

where \( \delta \) and \( \omega \) denote the generator rotor and the voltage angles at the buses, \( \mathcal{L} = \begin{pmatrix} L_g & L_{\omega} & L_{\theta} \end{pmatrix} \) is the network susceptance matrix, \( M_g \) and \( D_g \) are the diagonal matrices of the generator inertial and damping coefficients, and \( P_w \) and \( P_t \) are power injections at the generators and buses. We refer to [35], [36] for the modeling of power networks.

**Model of monitors.** A monitor is a deterministic algorithm \( \Phi : \Lambda \rightarrow \Psi \) with access to continuous-time measurements and knowledge of the system dynamics, that is, \( \Lambda = \{ E, A, C, y(t) \ \forall t \in \mathbb{R}_{\geq 0} \} \). The output of a monitor is \( \Psi = \{ \psi_1, \psi_2 \} \), with \( \psi_1 \in \{ \text{True, False} \} \), and \( \psi_2 \subseteq \{ 1, \ldots, n + p \} \). In particular, the output \( \psi_1 \) reveals the presence of attacks, while \( \psi_2 \) corresponds to the attack set.

Let \( y(x, u, t) \) be the output signal of (1) generated from the initial state \( x \) by the attack input \( u \). Then, the monitoring input \( y(t) = y(x_0, u_K, t) \) at all times, where \( x_0 \) is the system initial state and \( u_K \) is the attack signal of the attack set \( K \). Since we only consider deterministic cyber-physical systems, we assume monitors to be consistent, that is,

(i) \( \psi_1 = \text{True only if the attack set } K \) is nonempty (\( \psi_1 = \text{False}, \text{otherwise} \)),

(ii) \( \psi_2 = \emptyset \) if and only if \( \psi_1 = \text{False} \), and

(iii) \( \psi_2 = K \) only if \( K \) is the (unique) smallest subset \( S \subseteq \{ 1, \ldots, n + p \} \) satisfying \( y(t) = y(x_1, u_K, t) \) for some initial state \( x_1 \) and at all times \( t \in \mathbb{R}_{\geq 0} \) (\( \psi_2 = \{ 1, \ldots, n + p \} \), otherwise).

Observe that, if \( S = \{ 1, \ldots, n + p \} \), then there always exists an attack signal \( u_S \) satisfying \( y(t) = y(x_0, u_K, t) = y(x_1, u_S, t) \). Our consistency assumption ensures that false-alarms are not triggered by our monitors. Examples of monitors can be found in [10], [11], [17].

The objective of a monitor is twofold:

**Definition 1: (Attack detection and identification)** Consider system (1) with nonzero attack \( (B_K u_K, D_K u_K) \). The attack \( (B_K u_K, D_K u_K) \) is detected by a monitor \( \Phi \) if \( \psi_1 = \text{True} \). The attack \( (B_K u_K, D_K u_K) \) is identified by a monitor \( \Phi \) if \( \psi_2 = K \).

An attack is undetectable (respectively unidentifiable) if no monitor detects (respectively identifies) the attack. Of course, an undetectable attack is also unidentifiable, since it cannot be distinguished from the zero attack. An attack set \( K \) is undetectable (respectively unidentifiable) if there exists an undetectable (respectively unidentifiable) attack \( (B_K u_K, D_K u_K) \).

**Model of attacks.** In this work we consider colluding omniscient attackers with the ability of altering the cyber-physical dynamics through exogenous inputs. In particular, we let the attack \( (B u(t), D u(t)) \) in (1) be designed based on knowledge of the system structure and parameters \( E, A, C \), and the full state \( x(t) \) at all times. Additionally, attackers have unlimited computation capabilities, and their objective is to disrupt the physical state or the measurements while avoiding detection or identification. Note that specific attacks may be cast by possibly-weaker attackers.

**Remark 2:** (Existing attack strategies as subcases) The following prototypical attacks can be modeled and analyzed through our theoretical framework:

(i) **stealth attacks** defined in [16] correspond to output attacks compatible with the measurements equation;

(ii) **replay attacks** defined in [11] are state and output attacks which affect the system dynamics and reset the measurements;

(iii) **covert attacks** defined in [12] are closed-loop replay attacks, where the output attack is chosen to cancel out the effect on measurements of the state attack; and

(iv) **(dynamic) false-data injection attacks** defined in [26] are output attacks rendering an unstable mode (if any) of the system unobservable.

A possible implementation of the above attacks in our model is illustrated in Fig. 1.

To conclude this section we remark that our modeling framework captures failures and attacks against power networks and water supply networks. Possible genuine failures include variations in demand and supply of power or water, line outages or pipe leakages, and failures of sensors and actuators. Possible cyber-physical attacks include measurements corruption [9], [10], [22] and attacks on the control architecture or the physical state [1], [15], [18], [19].

**III. FUNDAMENTAL MONITORING LIMITATIONS**

In this section we highlight fundamental monitoring limitations from system-theoretic and graph-theoretic perspectives.
A. System-theoretic monitoring limitations

Following the discussion in Section II, an attack is undetectable if the measurements due to the attack coincide with the measurements due to some nominal operating condition.

**Lemma 3.1: (Undetectable attack)** For the descriptor system (1), the nonzero attack \((B_K u_K, D_K u_K)\) is undetectable if and only if \(y(x_1, u_K, t) = y(x_2, 0, t)\) for some initial states \(x_1, x_2 \in \mathbb{R}^n\) and for all \(t \in \mathbb{R}_{\geq 0}\).

**Proof:** (If) Let \(y(x_1, u_K, t) = y(x_2, 0, t)\). Since monitors are consistent, the monitor inputs \(y_1(t) = y(x_1, u_K, t)\) and \(y_2(t) = y(x_2, 0, t)\) yield the same output \(\{\psi_1, \psi_2\}\). Since monitors are consistent, we have \(\psi_1 = \text{False}\) for the input \(y_2\). Hence, \(\psi_1 = \text{False}\) also for the input \(y_1\), and the attack is undetectable.

(Only if) Suppose that \(y(x_1, u_K, t) \neq y(x_2, 0, t)\) for every initial states \(x_1\) and \(x_2\). Then the attack \((B_K u_K, D_K u_K)\) is distinguishable from nominal operating conditions via the system output. Hence, the attack is detectable. See Section IV for a complete detection monitor.

Analogous to detectability, the identifiability of an attack is the possibility to distinguish from measurements between the action of two distinct attacks. We measure the strength of an attack through the cardinality of the corresponding attack set. Since an attacker can independently compromise any state variable or measurement, every subset of the states and measurements of fixed cardinality is a potential attack set.

**Lemma 3.2: (Identifiable attack)** For the descriptor system (1), the nonzero attack \((B_K u_K, D_K u_K)\) is unidentifiable if and only if \(y(x_1, u_K, t) = y(x_2, u_R, t)\) for some initial states \(x_1, x_2 \in \mathbb{R}^n\), attack \((B_R u_R, D_R u_R)\) with \(|R| \leq |K|\) and \(R \neq K\), and for all \(t \in \mathbb{R}_{\geq 0}\).

A proof of Lemma 3.2 follows the same reasoning as the proof of Lemma 3.1. We now elaborate on the above lemmas to derive fundamental detection and identification limitations. For a vector \(x \in \mathbb{R}^n\), let supp\((x) = \{i \in \{1, \ldots, n\} : x_i \neq 0\}\), and let \(\|x\|_e = |\text{supp}(x)|\) denote the number of non-zero entries.

**Theorem 3.3: (Detectability of cyber-physical attacks)** For the descriptor system (1) and an attack set \(K\), the following statements are equivalent:

(i) the attack set \(K\) is undetectable; and
(ii) there exist \(s \in \mathbb{C}\), \(g \in \mathbb{R}^{|K|}\), and \(x \in \mathbb{R}^n\), with \(x \neq 0\), such that \((sE - A)x - B_K g = 0\) and \(Cx + D_K g = 0\).

Moreover, there exists an undetectable attack set \(K\), with \(|K| = k\), if and only if there exist \(s \in \mathbb{C}\) and \(x \in \mathbb{R}^n\) such that \(||(sE - A)x||_0 + ||Cx||_0 = k\).

**Proof:** By Lemma 3.1 and linearity of system (1), the attack \(u_K\) is undetectable if and only if there exists \(x_0\) such that \(y(x_0, u_K, t) = 0\) for all \(t \in \mathbb{R}_{\geq 0}\), that is, if and only if system (1) features zero dynamics. For a linear descriptor system with smooth input and consistent initial condition, the existence of zero dynamics is equivalent to the existence of invariant zeros as in (ii), see [28, Theorem 3.2 and Proposition 3.4]. The equivalence of statements (i) and (ii) follows. The last statement follows from (ii), and the fact that \(B = [I, 0]\) and \(D = [0, I]\).

Following Theorem 3.3, an attack \((B_K u_K, D_K u_K)\) is undetectable if it excites only zero dynamics for the dynamical system (1). Moreover, the existence of undetectable attacks for the attack set \(K\) is equivalent to the existence of invariant zeros for the system \((E, A, B_K, C, D_K)\). For the notions of zero dynamics and invariant zeros we refer the reader to [25], [28]. The following theorem shows that analogous statements hold for the identifiability of attacks.

**Theorem 3.4: (Identifiability of cyber-physical attacks)** For the descriptor system (1) and an attack set \(K\), the following statements are equivalent:

(i) the attack set \(K\) is unidentifiable; and
(ii) there exists an attack set \(R\), with \(|R| \leq |K|\) and \(R \neq K\), \(s \in \mathbb{C}\), \(g \in \mathbb{R}^{|K|}\), \(g \in \mathbb{R}^{|R|}\), and \(x \in \mathbb{R}^n\), with \(x \neq 0\), such that \((sE - A)x - B_K g - B_R g_R = 0\) and \(Cx + D_K g_K + D_R g_R = 0\).

Moreover, there exists an unidentifiable attack set \(K\), with \(|K| = k \in \mathbb{N}_0\), if and only if there exists an undetectable attack set \(K\), with \(|K| \leq 2k\).

**Proof:** Due to linearity of the system (1), the unidentifiability condition in Lemma 3.2 is equivalent to the condition \(y(x_K - x_R, u_K - u_R, t) = 0\), for some initial conditions \(x_K, x_R\), and attack signals \(u_K, u_R\). The equivalence between statements (i) and (ii) follows analogously to the proof of Theorem 3.3. Finally, the last statement follows from Theorem 3.3, and the fact that \(B = [I, 0]\) and \(D = [0, I]\).

Theorem 3.4 shows that the existence of an undetectable attack set \(K\) of cardinality \(k\) is equivalent to the existence of invariant zeros for the system \((E, A, B_K, C, D_K)\), with \(|K| \leq 2k\).

**Remark 3: (Static and active monitors, and noisy dynamics)** A particular monitor is the so-called static monitor which verifies the consistency of the measurements without knowledge of the system dynamics and without exploiting relations among measurements taken at discrete time instants. For instance, the bad data detector in [9], [37] is a static monitor. Then, an attack \((B_K u_K, D_K u_K)\) is undetectable by a static monitor if and only if, for some state trajectory \(x : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^n\) and for all times \(t \in \mathbb{N}\) it holds \(C x(t) + D_K u_K(t) = 0\). Note that static attacks are undetectable by static monitors [17].

An active monitor injects an auxiliary input \((B v, D v)\) to reveals attacks [11]. Since auxiliary inputs do not alter the invariant zeros of system (1), active monitors share the same fundamental limitations of our monitors.

An analogous reasoning shows that the existence of undetectable attacks for a noise-free system implies the existence of undetectable attacks for the same system driven by noise. The converse does not hold, since attackers may remain undetected by injecting a signal compatible with the noise statistics.

**Remark 4: (Inconsistent initial conditions and impulsive attacks)** If the consistency assumption (A2) is dropped, then discontinuities in the state \(x(t \downarrow 0)\) may affect the measurements \(y(t \downarrow 0)\). For instance, for index-one systems, inconsistent initial conditions lead to initial jumps for the algebraic equations to be satisfied. Consequently, the inconsistent initial value \([0^T \ x_0(0)^T]^T \in \text{Ker}(E)\) cannot be recovered through measurements.
Assumption (A3) requires the attack signal to be sufficiently smooth such that $x(t)$ and $y(t)$ are at least continuous. Suppose that assumption (A3) is dropped and the input $u(t)$ belongs to the class of impulsive smooth distributions $C_{\text{imp}} = C_{\text{smooth}} \cup C_{p-\text{imp}}$, that is, loosely speaking, the class of functions given by the linear combination of a smooth function on $\mathbb{R}_{\geq 0}$ (denoted by $C_{\text{smooth}}$) and Dirac impulses and their derivatives at $t = 0$ (denoted by $C_{p-\text{imp}}$) [28]. In this case, an attacker commanding an impulsive input can reset the initial state and evade detection.

The discussion in the previous two paragraphs can be formalized as follows. Let $V_c$ be the subspace of points $x_0 \in \mathbb{R}^n$ of consistent initial conditions for which there exists an input $u \in C_{\text{smooth}}$ and a state trajectory $x(t)$ in $C_{\text{smooth}}$ to the descriptor system (1) such that $y(t) = 0$ for all $t \in \mathbb{R}_{\geq 0}$. Let $V_d$ (respectively $W$) be the subspace of points $x_0 \in \mathbb{R}^n$ for which there exists an input $u \in C_{p-\text{imp}}^n$ (respectively $u \in C_{p-\text{imp}}^n$) and a state trajectory $x \in C_{p-\text{imp}}^n$ (respectively $x \in C_{p-\text{imp}}^n$) to the descriptor system (1) such that $y(t) = 0$ for all $t \in \mathbb{R}_{\geq 0}$. From [28, Theorem 3.2 and Proposition 3.4] it is known that $V_d = V_c + W + \text{Ker}(E)$.

In this work we focus on the smooth output-nulling sub-space $V_c$, which is exactly space of zero dynamics identified in Theorems 3.3 and 3.4. Hence, for inconsistent initial conditions, the results presented in this section are valid only for strictly positive times. On the other hand, if an attacker injects impulsive signals, then it can avoid detection for initial conditions in $W$.

\section*{B. Graph-theoretic monitoring limitations}

In this section we derive detectability conditions based upon a connectivity property of a graph associated with the dynamical system. For the ease of notation, in this subsection we drop the subscript $K$ from $B_K$, $D_K$, and $u_K$. Let $([E],[A],[B],[C],[D])$ be the tuple of structure matrices associated with the system (1) [27]. We associate a directed input/state/output graph $G_{\text{so}} = (V,E)$ with $([E],[A],[B],[C],[D])$. The vertex set $V = \cup \{X, Y\}$ consists of input, state, and output vertices given by $U = \{u_1, \ldots, u_m\}$, $X = \{x_1, \ldots, x_n\}$, and $Y = \{y_1, \ldots, y_p\}$, respectively. The set of directed edges $E$ is $E_{|E|} \cup E_{|A|} \cup E_{|B|} \cup E_{|C|} \cup E_{|D|}$, where $E_{|E|} = \{(x_j, x_i) : [E]_{ij} \neq 0\}$, $E_{|A|} = \{(x_j, y_i) : [A]_{ij} \neq 0\}$, $E_{|B|} = \{(u_j, x_i) : [B]_{ij} \neq 0\}$, $E_{|C|} = \{(x_j, y_i) : [C]_{ij} \neq 0\}$, and $E_{|D|} = \{(u_j, y_i) : [D]_{ij} \neq 0\}$. In the latter, the expression $[E]_{ij} \neq 0$ means that the $(i,j)$-th entry of $[E]$ is a free parameter. For the graph $G_{\text{so}}$, a set of $l$ mutually disjoint and simple paths between two sets of vertices $S_1, S_2$ is called linking of size $l$ from $S_1$ to $S_2$. Finally, the matrix $s[E] - A$ is structurally non-degenerate if the determinant $|s[E] - A| \neq 0$ for a generic realization of $E$ and $A$, that is, $|s[E] - A| \neq 0$ holds in the whole parameter space of elements of $E$ and $A$ with exception of a low dimensional variety [24, 38].

\textbf{Example 1: (Power network structural analysis)} Consider the power network illustrated in Fig. 2(a), where, being $e_i$ the $i$-th canonical vector, we take $[E] = \text{blkdiag}(1, 1, 1, M_1, M_2, M_3, 0, 0, 0, 0, 0, 0)$, $[B] = [e_8 \ e_9]$, $[C] = [e_1 \ e_4]$. The digraph associated with the structure matrices $([E],[A],[B],[C],[D])$ is in Fig. 2(b). There are four vertices of $\{0\}$ in $G_{\text{so}}$.

Recall from Lemma 3.1 that an attack $u$ is undetectable if $y(x_1, u, t) = y(x_2, 0, t)$ for some initial states $x_1$ and $x_2$. In the following result, we consider the particular case that the system initial state is known. Hence, an attack $u$ is undetectable if $y(x_0, u, t) = y(x_0, 0, t)$ for some initial state $x_0$. Equivalently, the system fails to be left-invertible [25]. We say that the structural system $([E],[A],[B],[C],[D])$ is structurally left-invertible if every admissible realization $(E, A, B, C, D)$ is left-invertible, with exception, possibly, of a low dimensional variety.

\textbf{Theorem 3.5: (Structurally undetectable attack)} Let the parameters space of the structured system

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig2.png}
\caption{Fig. 2(a) shows the WSSC power system with 3 generators and 6 buses. The numerical value of the network parameters can be found in [30]. The digraph associated with the network in Fig. 2(a). The self-loops of the vertices $\{\delta_1, \delta_2, \delta_3\}$, $\{\omega_1, \omega_2, \omega_3\}$, and $\{\theta_1, \ldots, \theta_6\}$ are not drawn. The inputs $u_1$ and $u_2$ affect respectively the bus $b_1$ and the bus $b_5$. The measured variables are the rotor angle and frequency of the first generator.}
\end{figure}
([E],[A],[B],[C],[D]) define a polytope in \( \mathbb{R}^d \) for some \( d \in \mathbb{N}_0 \). Assume that \( s[E] - [A] \) is structurally non-degenerate. The system \( ([E],[A],[B],[C],[D]) \) is structurally left-invertible if and only if there exists a linking of size \([U]\) from \(U\) to \(Y\).

Theorem 3.5 extends the structural left-invertibility results known for nonsingular systems to regular descriptor systems, and its proof relies on classical concepts from structural analysis, algebraic geometry, and graph theory. Additionally, Theorem 3.5 gives a characterization of structurally undetectable attacks. The following result is useful to prove Theorem 3.5.

**Lemma 3.6: (Polytopes and algebraic varieties)** Let \( S \subseteq \mathbb{R}^d \) be a polytope, and let \( T \subseteq \mathbb{R}^d \) be an algebraic variety. Then, either the set \( S \subseteq T \), or the set \( S \setminus (S \cap T) \) is generic in \( S \).

**Proof:** Let \( T \subseteq \mathbb{R}^d \) be the algebraic variety described by the locus of common zeros of the polynomials \( \{\phi_1(x), \ldots, \phi_t(x)\} \), with \( t \in \mathbb{N}, t < \infty \). Then \( S \subseteq T \) if and only if every polynomial \( \phi_i \) vanishes identically on \( S \). Suppose that some polynomials, say \( \phi_i \), do not vanish identically on \( S \). Then, \( S \cap T \neq S \), and \( S \cap T = \{ x \in S : \phi_i(x) = 0 \} \) is nowhere dense in \( S \), since its closure has empty interior \([39]\). Hence, \( S \cap T \) is a meagre subset of \( S \), and its complement \( S \setminus (S \cap T) \) is a generic subset of \( S \)[39].

In Lemma 3.6 interpret the polytope \( S \) as the admissible parameters space of a structured cyber-physical system. Then we have shown that left-invertibility of a cyber-physical system is a generic property even when the admissible parameters space is a polytope of the whole parameters space. Consequently, for a structured cyber-physical system, if the initial state is known, either every admissible realization admits undetectable attacks, or there is no undetectable attack for every realization, except possibly for those lying on a low dimensional variety.

**Proof of Theorem 3.5:** Because of Lemma 3.6, we need to show that, if there are \( |U| \) disjoint paths from \( U \) to \( Y \), then there exists admissible left-invertible realizations. Conversely, if there are at most \( |U| - 1 \) disjoint paths from \( U \) to \( Y \), then every admissible realization is not left-invertible.

(I) Let \((E,A,B,C,D)\), with \([sE-A] \neq 0\), be an admissible realization, and suppose there exists a linking of size \([U]\) from \(U\) to \(Y\). Notice that \( |Y| \geq |U| \), and select \([U]\) outputs on a linking of size \([U]\) from \(U\) to \(Y\) (let \( C \) and \( D \) be the submatrices of \( C \) and \( D \) associated with the smaller set of outputs). Observe that left-invertibility of \((E,A,B,C,D)\) implies left-invertibility of \((E,A,B,C,D)\). For the left-invertibility of \((E,A,B,C,D)\) we need

\[
\begin{bmatrix}
    sE - A & -B \\
    C & D \\
\end{bmatrix} = [sE-A]D + C(sE-A)^{-1}B \neq 0,
\]

and hence we need \([D + C(sE-A)^{-1}B] \neq 0\). Notice that \( D + C(sE-A)^{-1}B \) corresponds to the transfer matrix of the cyber-physical system. Since there are \([U]\) independent paths from \(U\) to \(Y\), the matrix \( D + C(sE-A)^{-1}B \) can be made nonsingular and diagonal by removing some connection lines from the network. In particular, for a given linking of size \([U]\) from \(U\) to \(Y\), a nonsingular and diagonal transfer matrix is obtained by setting to zero the entries of \( E \) and \( A \) corresponding to the edges not in the linking. Then there exist admissible left-invertible realizations, and thus the systems \((E,A,B,C,D)\) and \((E,A,[D],[C],[D])\) are structurally left-invertible.

(II) Take any subset of \([U]\) output vertices, and let the maximum size of a linking from \(U\) to \(Y\) be smaller than \([U]\). Let \([E\) and \([A\)] be such that \(s[E] - [A] = \begin{bmatrix} s[E] - [A] & [B] \\
    [C] & [D] \end{bmatrix}\), where \([C\) and \([D\) are the structured output matrices corresponding to the chosen \([U]\) output vertices. Consider the graph \(G(s[E] - [A])\), that consists of \(N = |X| + |U|\) vertices, and an edge from vertex \( j \) to \( i \) if \( A_{ij} \neq 0 \) or \( E_{ij} \neq 0 \). Notice that a path from \(U\) to \(Y\) in the digraph associated with the structured system corresponds, possibly after relabeling the output variables, to a cycle in involving input/output vertices in \(G(s[E] - [A])\). Observe that there are only \([U] - 1\) such (disjoint) cycles. Hence, there is no cycle family of length \(N\), and the system \((E,A,[B],[C],[D])\) fails to be structurally left-invertible [40, Theorem 1]. Since the same reasoning holds for every set of \([U]\) output vertices, every realization of the pencil \(\begin{bmatrix} s[E] - [A] & [B] \\
    [C] & [D] \end{bmatrix}\) has no invertible minor of size \(N\), and the claimed statement follows.

If the system initial state is unknown, then an undetectable attack \(u\) is characterized by the existence of a pair of initial conditions \(x_1\) and \(x_2\) such that \(y(x_1,0,t) = y(x_2,u,t)\), or, equivalently, by the existence of invariant zeros for the given cyber-physical system. We will now show that, provided that a cyber-physical system is left-invertible, its invariant zeros can be computed by simply looking at an associated nonsingular state space system. Let the state vector \(x\) of the descriptor system \((1)\) be partitioned as \(x = [x_1^T \ x_2^T]^T\), where \(x_1\) corresponds to the dynamic variables. Let the network matrices \(E, A, B, C, D\) be partitioned accordingly, and assume that the descriptor system \((1)\) is given in semi-explicit form, that is, \(E = \text{blkdiag}(E_{11}, 0)\), where \(E_{11}\) is nonsingular.\(^2\) In this case, the descriptor system \((1)\) reads as

\[
E_{11} \dot{x}_1(t) = A_{11} x_1(t) + B_1 u(t) + A_{12} x_2(t) ,
\]

\[
0 = A_{21} x_1(t) + A_{22} x_2(t) + B_2 u(t) ,
\]

\[
y(t) = C_1 x_1(t) + C_2 x_2(t) + D u(t) .
\]

Consider now the associated nonsingular state space system which is obtained by regarding \(x_2(t)\) as an external input to the descriptor system \((3)\) and the algebraic constraint as output:

\[
\dot{x}_1(t) = E_{11}^{-1} A_{11} x_1(t) + E_{11}^{-1} B_1 u(t) + E_{11}^{-1} A_{12} x_2(t) ,
\]

\[
y(t) = [A_{21} C_1] x_1(t) + [A_{22} C_2] x_2(t) + u(t) .
\]

**Theorem 3.7: (Equivalence of invariant zeros)** Consider the structurally left-invertible system \((E,A,B,C,D)\). The invariant zeros of every admissible realization \((3)\) coincide with those of the associated nonsingular system \((4)\), except, possibly, for realizations lying on a low dimensional variety of the parameters space.

\(^2\) Interesting cyber-physical systems, such as power and mass-transport networks \((2)\), are readily given in semi-explicit form.
**Proof:** In the interest of space we omit the proof, which follows from Theorem 3.5, [41, Proposition 8.4] and a manipulation of the system pencil.

Following Theorem 3.7, under the assumption of structural left-invertibility, classical results can be used to investigate the presence of undetectable attacks in structured system with unknown initial state; see [38] for a survey of results on generic properties of linear systems.

IV. **Monitor Design for Attack Detection and Identification**

We now design centralized and distributed filters for attack detection and identification.

A. **Centralized attack detection**

The output of the attack detection filters developed in this subsection will be a residual signal \( r : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^p \). If each monitor is equipped with such an attack detection filter and if the attack is detectable, then the outputs of the monitor and the filter are related as follows: \( \psi_1 = \text{True} \) if and only if \( r(t) = 0 \) for all \( t \in \mathbb{R}_{\geq 0} \). We next present a centralized attack detection filter based on a modified Luenberger observer.

**Theorem 4.1: (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**

\[
E \dot{\bar{w}}(t) = (A + GC)w(t) - Gy(t),
\]

\[
r(t) = Cw(t) - y(t),
\]

where \( w(0) = x(0) \) and the output injection matrix \( G \in \mathbb{R}^{n \times p} \) is such that the pair \((E,A+GC)\) is regular and Hurwitz. \(^3\) 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 \( \bar{w} = w - x \) is exponentially stable.

**Proof:** Consider the error \( e = w - x \) between the filter (5) and system (1). The error system with output \( r(t) \) is

\[
Ee(t) = (A + GC)e(t) - (B_K + GD_K)u_K(t),
\]

\[
r(t) = Ce(t) - D_Ku_K(t),
\]

where \( e(0) = 0 \). To prove the theorem we show that the error system (6) 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 \( u_K \) are assumed to be consistent (A2) and non-impulsive (A3), the error system (6) has no invariant zeros if and only if [28, Proposition 3.4] there exists no triple \((s, \bar{w}, g_K) \in \mathbb{C} \times \mathbb{R}^n \times \mathbb{R}^p \) satisfying

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

The second equation of (7) yields \( C\bar{w} = D_Kg_K \). By substituting \( C\bar{w} \) by \( D_Kg_K \) in the first equation of (7), we obtain

\[
\begin{bmatrix}
    sE - A & B_K \\
    C & -D_K
\end{bmatrix}
\begin{bmatrix}
    \bar{w} \\
    g_K
\end{bmatrix}
= \begin{bmatrix}
    0 \\
    0
\end{bmatrix}.
\]

Note that a solution \((s, \bar{w}, g_K) \) to (8) would yield an invariant zero, zero state, and zero input for the descriptor system (1). By the detectability assumption, the descriptor system (1) has no invariant zeros and the matrix pencil in (8) necessarily has full rank. It follows that the triple \((E,A,C)\) is observable, \( G \) can be chosen to make the pair \((E,A+GC)\) Hurwitz [42], and the error system (6) is stable without zero dynamics.

Notice that, if the initial state \( x(0) \) is not available, then, an arbitrary initial state \( w(0) \in \mathbb{R}^n \) can be chosen. In this case, since \((E,A+GC)\) is Hurwitz, the filter (5) converges asymptotically, and some attacks may remain undetected. Also, if the dynamics and the measurements of (1) are affected by modeling uncertainties and noise with known statistics, then the output injection matrix \( G \) should be chosen to optimize the sensitivity of the residual \( r \) to attacks versus the effect of noise. Statistical testing techniques can [7] subsequently be used to analyze the residual \( r \). Finally, attacks aligned with the noise statistics may remain undetected.

B. **Distributed attack detection**

Notice that a direct implementation of the filter (5) requires continuous communication of measurements to a central processor, which needs to integrate the possibly large-scale system (5). In what follows, we will exploit the sparsity of the filter matrices \((E,A,C)\) to develop a distributed detection filter.

Assume that control centers are geographically deployed in a large scale cyber-physical system to operate the whole plant via distributed computation; see Fig. 3. Let \( G_s = (\mathcal{V}, \mathcal{E}) \) be the directed sparsity graph associated with the pair \((E,A)\), where the vertex set \( \mathcal{V} = \mathcal{X} \) corresponds to the system state, and the set of directed edges \( \mathcal{E} = \{(x_j, x_i) : \epsilon_{ij} \neq 0 \text{ or } a_{ij} \neq 0\} \) is induced by the sparsity pattern of \( E \) and \( A \). Let \( \mathcal{V} \) be 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 \( G_s^i = (\mathcal{V}_i, \mathcal{E}_i) \) be the \( i \)-th subgraph of \( G_s \) 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

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

where $A_i \in \mathbb{R}^{n_i \times n_i}$, $A_{ij} \in \mathbb{R}^{n_i \times n_j}$, and $A_D = \text{blkdiag}(A_1, \ldots, A_N)$. We make the following assumptions:

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

(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

$$
\dot{E}_i \dot{x}_i(t) = A_i x_i(t) + \sum_{j \in \mathcal{N}_i} A_{ij} x_j(t),
$$

$$
y_i(t) = C_i x_i(t), \quad i \in \{1, \ldots, N\},
$$

where $x_i$ and $y_i$ are the state and output of the $i$-th subsystem and $\mathcal{N}_i = \{ j \in \{1, \ldots, N\} \setminus \{i\} : \|A_{ij}\| \neq 0 \}$ are the in-neighbors of subsystem $i$. We also define the set of out-neighbors as $\mathcal{N}_i^\text{out} = \{ j \in \{1, \ldots, N\} \setminus \{i\} : \|A_{ji}\| \neq 0 \}$. We assume the presence of a control center in each subnetwork $\mathcal{G}_i$ 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 $A_{ij}, j \in \mathcal{N}_i^\text{in}$; and

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

Before presenting our distributed attack detection filter, we need the following result on a decentrally stabilized filter.

**Lemma 4.2: (Decentralized stabilization of the 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 attack detection filter

$$
\dot{E} \dot{w}(t) = (A_D + GC)w(t) + A_C w(t) - G y(t),
$$

$$
r(t) = y(t) - C w(t),
$$

where $w(0) = x(0)$ and $G = \text{blkdiag}(G_1, \ldots, G_N)$ is such that $(E, A_D + GC)$ is regular and Hurwitz. Assume that

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

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 $w(t) = w(t) - x(t)$ obeys the dynamics

$$
\dot{E} \dot{w}(t) = (A_D + A_C + GC)w(t) - (B_K + GD_K)u_K(t),
$$

$$
r(t) = C e(t) - D_K u_K(t).
$$

A reasoning analogous to that in the proof of Theorem 4.1 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 systems and rewrite the error dynamics (12) as the closed-loop interconnection of the two subsystems $\Gamma_1: E \dot{e}(t) = (A_D + GC)e(t) + v(t)$ and $\Gamma_2: v(t) = A_C e(t)$. When regarded as input-output systems with respective input-output pairs $(e, v)$ and $(e, v)$, both $\Gamma_1$ and $\Gamma_2$ are causal and internally stable. Hence, by [43, Theorem 4.11], the overall error dynamics (12) 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 $\omega \in \mathbb{R}$. The latter condition is equivalent to (11).

An implementation of the decentrally stabilized filter (10) under assumptions (A1)-(A7) requires the input $A_C w$ and hence continuous communication among control centers. To overcome this continuous communication obstacle we rely on waveform relaxation methods [44, 45] developed for parallel numerical integration. The Gauss-Jacobi waveform relaxation applied to the filter (10) yields the waveform relaxation iteration

$$
\dot{E} \dot{w}^{(k)}(t) = (A_D + GC)w^{(k)}(t) + A_C w^{(k-1)}(t) - G y(t),
$$

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 initial condition $w^{(k)}(0) = w_0$ for each $k \in \mathbb{N}$. Notice that (13) is a descriptor system with state $w^{(k)}$, and known input $A_C w^{(k-1)}$, since the value of $w(t)$ at iteration $k - 1$ is used. The iteration (13) is initialized with an initial profile $w(0): [0, T] \to \mathbb{R}^n$.

The iteration (13) is said to be (uniformly) convergent if

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

where $w$ is the solution of the non-iterative dynamics (10). In order to obtain a detection scheme, we use the waveform relaxation iteration (13) to iteratively approximate the decentralized filter (10).

**Theorem 4.3: (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 the assumptions (A1) through (A7) be satisfied and consider the distributed attack detection filter

$$
\dot{E} \dot{w}^{(k)}(t) = (A_D + GC)w^{(k)}(t) + A_C w^{(k-1)}(t) - G y(t),
$$

$$
r^{(k)}(t) = y(t) - C w^{(k)}(t),
$$

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 $G = \text{blkdiag}(G_1, \ldots, G_N)$ is such that the pair $(E, A_D + GC)$ is regular, Hurwitz, and

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

Then $\lim_{k \to \infty} \|r^{(k)}(t)\|_{\infty} = 0$ 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 [45, Theorem 5.2] that the solution $w^{(k)}$ of the iteration (15) converges, as $k \to \infty$, to the solution $w$ of (10) if

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

where $\sigma = \max\{\alpha, \beta\}$, $\alpha$ is the least upper bound on the real part of the spectrum of $(E, A)$, and $\beta$ is such that the signal
Due to the detectability assumption, it follows from Lemma 4.2 that \( \lim_{t \to \infty} \|u(t)\|_{\infty} = 0 \) at all times \( t \in [0, T] \) if and only if \( u_K(t) = 0 \) at all times \( t \in [0, T] \).

Under condition (16) and due to the Hurwitz assumption, it follows from Lemma 4.2 that the error \( e = w - x \) between the state \( w \) of the decentralized filter dynamics (10) and the state \( x \) of the descriptor model (1) is asymptotically stable in the absence of attacks. This concludes the proof of Theorem 4.3.

The waveform relaxation iteration (13) can be implemented in the following distributed fashion. Assume that control center \( i \) is able to numerically integrate the descriptor system

\[
E_i \dot{w}_i^{(k)}(t) = (A_i + G_i C_i) w_i^{(k)}(t) + \sum_{j \in \mathcal{N}_i} A_{ij} w_j^{(k-1)}(t) - G_i y_i(t),
\]

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

1. set \( k := k + 1 \), and compute the signal \( w_i^{(k)} \) by integrating the local filter equation (18);
2. transmit \( w_i^{(k)} \) to the \( j \)-th control center if \( j \in \mathcal{N}_i \); and
3. update the input \( w_i^{(k)} \) with the signal received from the \( j \)-th control center, with \( j \in \mathcal{N}_i \), and iterate.

Following Theorem 4.3, for \( k \) sufficiently large, the local residuals \( r_i^{(k)}(t) = y_i - C_i w_i^{(k)} \) can be used to detect attacks.

A related large-scale example is given in Section V-C.

Remark 5: (Implementation of distributed attack detection filter) When implementing the distributed attack detection filter (15) in the interval \([0, T]\), control center \( i \) needs to transmit the signal \( w_i^{(k)}(t) \) with \( t \in [0, T] \) at each iteration \( k \). In practice, only an approximation or a finite basis representation \( \hat{w}_i^{(k)}(t) \) can be transmitted. The error due to this approximation can be characterized, and we refer the reader to [46].

\[ \min_{v_x, v_y, \xi_0} \|v_x\|_{L_0} + \|v_y\|_{L_0} \]

subject to

\[
E\dot{\xi}(t) = A\xi(t) + v_x(t),
\]

\[
y(t) = C\xi(t) + v_y(t),
\]

\[
\xi(0) = \xi_0 \in \mathbb{R}^n.
\]

Lemma 4.4: (Problem equivalence) Consider the system (1) with identifiable attack set \( K \). The optimization problem (19) coincides with the problem of identifying the attack set \( K \) given the system matrices \( E, A, C \), and the measurements \( y \), where \( K = \text{supp}(v(x, y)) \).

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 Lemma 3.2. 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 (19), or equivalently our identification problem, is generally NP-hard [47].

Corollary 4.5: (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 \( y \) is NP-hard.

Proof: Consider the NP-hard [48] sparse recovery problem

\[
\min_{\xi, \bar{y}} \| \bar{y} - C\xi \|_{\ell_0}, \text{ where } C \in \mathbb{R}^{P \times n} \text{ and } \bar{y} \in \mathbb{R}^P \text{ are given and constant.}
\]

In order to prove the claimed statement, we show that every instance of the sparse recovery problem can be cast as an instance of (19). Let \( E = I, A = 0, C = \bar{C}, \) and \( y(t) = \bar{y} \) at all times. Notice that \( v_x(t) = \bar{y} - C\xi(t) \) and \( \xi(t) = \xi(0) + \int_0^t v_x(\tau) d\tau \). The problem (19) can be written as

\[
\min_{v_x, \xi} \|v_x\|_{L_0} + \|\bar{y} - C\xi(t)\|_{L_0} \]

\[
\min_{v_x(t), \xi} \|v_x(t)\|_{L_0} + \|\bar{y} - \bar{C}\xi - \bar{C} \int_0^t v_x(\tau) d\tau\|_{L_0},
\]

where \( \bar{C} = \bar{C}(0) \). Notice that there exists a minimizer to problem (20) with \( v_x(t) = 0 \) for all \( t \). Indeed, since \( \|\bar{y} - C\xi - \bar{C} \int_0^t v_x(\tau) d\tau\|_{L_0} \geq \|\bar{y} - \bar{C}\xi - \bar{C} \int_0^t v_x(\tau) d\tau\|_{L_0} = \|\bar{y} - \bar{C}\xi\|_{L_0} \)

By Corollary 4.5 the general attack identification problem is combinatorial in nature, and its general solution will require substantial computational effort. In the next subsection we propose a complete identification algorithm.

D. Centralized attack identification

The identification of the attack set \( K \) requires a combinatorial procedure, since, a priori, \( K \) is one of the \( \binom{|\mathcal{N}|}{|K|} \) 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. Analogously to the attack detection filter developed in Subsections
IV-A and IV-B, the output of the attack equation filter for the attack set \( K \) will be a residual signal \( r_K \). If each monitor is equipped with such an attack identification filter and if the attack \( K \) is identifiable, then the outputs of the monitor and the filter are related as follows: \( \psi_2 = K \) if and only if \( r_K(t) = 0 \) for all \( t \in \mathbb{R}_{\geq 0} \).

The design of this residual filter consists of three steps—an input output transformation, a state transformation, and an output injection and definition of a specific residual. We start by showing that the identification problem can be carried out for a modified system without corrupted measurements.

**Lemma 4.6: (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 without corrupted measurements:

\[
\begin{align*}
E \dot{x}(t) &= (A - B_K D_K^\dagger C)x(t) + B_K(I - D_K^\dagger D_K)u_K(t), \\
\dot{y}(t) &= (I - D_K D_K^\dagger)Cx(t).
\end{align*}
\]  
(21)

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

\[
\begin{bmatrix}
sE - A & -B_K & -B_R \\
C & D_K & D_R \\
C & D_K & D_R \\
\end{bmatrix}
\begin{bmatrix}
x \\
g_K \\
g_r \\
\end{bmatrix} = \begin{bmatrix} 0 \\
0 \\
0 \\
\end{bmatrix},
\]  
(22)

where we added an additional (redundant) output equation (see Theorem 3.4). A multiplication of equation (22) from the left by the projectors \( \text{blkdiag}(I, D_K D_K^\dagger, (I - D_K D_K^\dagger)) \) yields

\[
\begin{bmatrix}
sE - A & -B_K & -B_R \\
D_K D_K^\dagger C & D_K & D_R \\
(I - D_K D_K^\dagger)C & 0 & (I - D_K D_K^\dagger)D_R \\
\end{bmatrix}
\begin{bmatrix}
x \\
g_K \\
g_r \\
\end{bmatrix} = \begin{bmatrix} 0 \\
0 \\
0 \\
\end{bmatrix}.
\]

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

\[
g_K = -D_K^\dagger Cx - D_K^\dagger D_R g_r + (I - D_K^\dagger D_K)g_K.
\]

Thus, \( P(s)[x^T g_K^T g_r^T] = 0 \) has no solution, where \( P(s) \) is

\[
\begin{bmatrix}
sE - A + B_K D_K^\dagger C & -B_K(I - D_K^\dagger D_K) & -B_R + B_K D_K^\dagger D_R \\
(I - D_K D_K^\dagger)C & 0 & (I - D_K D_K^\dagger)D_R \\
\end{bmatrix}
\]

The statement follows.

The second design step of our attack identification monitor relies on the concept of **conditioned invariant subspace**. We refer to [24], [25], [28], [49] for a comprehensive discussion of geometric control theory. Let \( S^* \) be the conditioned invariant subspace associated with the system \((E, A, B, C, D)\), that is, the smallest subspace of the state space satisfying

\[
\begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} E^{-1}S^* \end{bmatrix} \cap \text{Ker} \begin{bmatrix} C & D \end{bmatrix} \subseteq S^*,
\]  
(23)

and let \( L \) be an output injection matrix satisfying

\[
\begin{bmatrix} A + LC & B + LD \end{bmatrix} \begin{bmatrix} E^{-1}S^* \end{bmatrix} \subseteq S^*.
\]  
(24)

Notice that the conditioned invariant \( S^* \) and an output injection \( L \) satisfying (23) and (24) always exist for instance, take \( S^* = \mathbb{R}^n \). We transform the descriptor system (21) into a set of canonical coordinates representing \( S^* \) 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 \( P^T(sE - A)Q \) for nonsingular \( P \) and \( Q \).

**Lemma 4.7: (Input decoupled system representation)** For system (21), let \( S^* \) and \( L \) be as in (23) and (24), respectively. Define the unitary matrices \( P = [\text{Basis}(S^*) \text{Basis}((S^*)^\perp)] \) and \( Q = [\text{Basis}(E^{-1}S^*) \text{Basis}((E^{-1}S^*)^\perp)] \). Then

\[
P^T EQ = \begin{bmatrix} \tilde{E}_{11} & \tilde{E}_{12} \\
0 & \tilde{E}_{22} \end{bmatrix}, \quad P^T B_K(I - D_K^\dagger D_K) = \begin{bmatrix} \tilde{B}_K(t) \\
0 \end{bmatrix},
\]

\[
P^T(A - B_K D_K^\dagger C + LC)Q = [\tilde{A}_{11} \tilde{A}_{12} \tilde{A}_{21} \tilde{A}_{22}],
\]

\( (I - D_K D_K^\dagger)C)Q = [\tilde{C}_1 \tilde{C}_2] \).

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

\[
\begin{bmatrix}
\tilde{E}_{11} & \tilde{E}_{12} \\
0 & \tilde{E}_{22} \\
\end{bmatrix}
\begin{bmatrix}
x_1(t) \\
x_2(t) \\
\end{bmatrix} = [\tilde{A}_{11} \tilde{A}_{12} \tilde{A}_{21} \tilde{A}_{22}]
\begin{bmatrix}
x_1(t) \\
x_2(t) \\
\end{bmatrix} + \begin{bmatrix} \tilde{B}_K(t) \\
0 \\
\end{bmatrix},
\]

\( y(t) = [\tilde{C}_1 \tilde{C}_2] \begin{bmatrix} x_1(t) \\
x_2(t) \end{bmatrix} \).  
(25)

**Proof:** Let \( L = E^{-1}S^* \) and \( M = S^* \). Notice that \((A + LC)E^{-1}S^* \subseteq S^* \) by the invariance property of \( S^* \) [28], [49]. It follows that \( L \) and \( M \) are a pair of right deflating subspaces for the matrix pair \((A + LC, E)\) [50], that is, \( M = AC + EC \) and \( \dim(M) \leq \dim(L) \). The sparsity pattern in the descriptor and dynamic matrices \( E \) and \( A \) of (25) arises by construction of the right deflating subspaces \( P \) and \( Q \) [50, Eq. (2.17)], and the sparsity pattern in the input matrix arises due to the invariance properties of \( S^* \) containing \( \text{Im}(B_K) \). 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.

For the ease of notation and without affecting generality, the third and final design step of our attack identification filter is presented for the pre-conditioned system (25).

**Theorem 4.8: (Attack identification for attack set \( K \)** Consider the pre-conditioned system (25) 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 \((B_K, D_K)\)

\[
\tilde{E}_{22} \tilde{w}_2(t) = \tilde{A}_{22} + \tilde{G} (I - \tilde{C}_1 \tilde{C}_1^T) \tilde{C}_2 w_2(t) - \tilde{G} \tilde{y}(t),
\]

\( r_K(t) = (I - \tilde{C}_1 \tilde{C}_1^T) \tilde{C}_2 w_2(t) - \tilde{y}(t) \), with

\( \tilde{y}(t) = (I - \tilde{C}_1 \tilde{C}_1^T) y(t) \),

where \( w_2(0) = x_2(0) \), and \( \tilde{G} \) is such that \((\tilde{E}_{22}, \tilde{A}_{22} + \tilde{G}(I - \tilde{C}_1 \tilde{C}_1^T) \tilde{C}_2) \) 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 \( w(t) = [w_1(t)^T w_2(t)^T]^T \), where \( w_1(t) \) obeys

\[
\tilde{E}_{11} \tilde{w}_1(t) + \tilde{E}_{12} \tilde{w}_2(t) = \tilde{A}_{11} w_1(t) + \tilde{A}_{12} w_2(t).
\]
Consider the filter error $e(t) = w(t) - x(t)$, and notice that

$$
\begin{bmatrix}
\hat{E}_{11} & \hat{E}_{12} \\
0 & E_{22}
\end{bmatrix}
\begin{bmatrix}
\dot{e}_1(t) \\
\dot{e}_2(t)
\end{bmatrix}
= \begin{bmatrix}
\hat{A}_{11} & \hat{A}_{12} \\
0 & \hat{A}_{22}
\end{bmatrix}
\begin{bmatrix}
\dot{e}_1(t) \\
\dot{e}_2(t)
\end{bmatrix}
- \begin{bmatrix}
\hat{B}_K \\
0
\end{bmatrix}
u_K(t),
$$

$$
r_K(t) = (I - \hat{C}_1\hat{C}_1^\dagger)\hat{C}_2\dot{e}_2(t),
$$

where $\hat{A}_{22} = \hat{A}_{22} + \hat{G}(I - \hat{C}_1\hat{C}_1^\dagger)\hat{C}_2$. Notice that $r_K(t)$ is not affected by the input $u_K(t)$, so that, since $e_2(0) = 0$ due to $u_2(0) = x_2(0)$, 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 $u_R$ results in a nonzero residual $r_K$. From Theorem 3.4 and the identifiability hypothesis, for any $R \neq K$, there exists no solution to

$$
\begin{bmatrix}
\dot{e}_{11} - \hat{A}_{11} & \dot{e}_{12} - \hat{A}_{12} & B_K \\
s\dot{e}_{22} - \hat{A}_{22} & 0 & -B_{R1} \\
0 & C_1 & -B_{R2}
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
g_k
\end{bmatrix}
= \begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}.
$$

A projection of the equation $0 = \tilde{C}_1x_1 + \tilde{C}_2x_2 + D_{RGR}$ onto the image of $\tilde{C}_1$ and its orthogonal complement yields

$$
\begin{bmatrix}
\dot{e}_{11} - \hat{A}_{11} & \dot{e}_{12} - \hat{A}_{12} & B_K \\
s\dot{e}_{22} - \hat{A}_{22} & 0 & -B_{R1} \\
0 & C_1 & -B_{R2}
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
g_k
\end{bmatrix}
= \begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}.
$$

Due to the identifiability hypothesis the set of equations (27) features no solution $[x_1^T \ x_2^T \ g_k^T]^T$ with $[x_1^T \ x_2^T]^T = 0$.

Observe that, for every $x_2$ and $g_R$, there exists $x_1 \in \text{Ker}(\tilde{C}_1)^\perp$ such that the third equation of (27) is satisfied. Furthermore, for every $x_2$ and $g_R$, there exist $x_1 \in \text{Ker}(\tilde{C}_1)$ and $g_K$ such that the first equation of (27) is satisfied. Indeed, since $QE^{-1}S^* = [\text{Im}(I) \ 0]^T$ and $P^TS^* = [\text{Im}(I) \ 0]^T$, the invariance of $S^*$ implies that $S^* = A(E^{-1}S^* \cap \text{Ker}(C)) + \text{Im}(B_K)$, or equivalently in new coordinates, $\text{Im}(I - \hat{C}_1\hat{C}_1^\dagger)\hat{C}_2)$. Finally note that $[(s\dot{E}_{11} - \hat{A}_{11}) \text{Ker}(\tilde{C}_1) \ B_K]$ is of full row rank due to the controllability of the subspace $S^*$ [28]. We conclude that there exist no vectors $x_2$ and $g_R$ such that $<s\dot{E}_{22} - \hat{A}_{22}> x_2 - B_{R2}g_R = 0$ and $(I - \hat{C}_1\hat{C}_1^\dagger)\hat{C}_2x_2 + D_{RGR}) = 0$ and the statement follows.

The design of the attack identification filter (26) is summarized as follows:

1. from system (1) define the system (21);
2. compute $S^*$ and $L$ for system (21) as in (23) and (24), and apply $L$, $P$, and $Q$ as in Lemma 4.7 leading to system (25);
3. for system (25), define $r_K$ and apply the output injection $G$ as in (26).

Remark 6: (Literature comparison) Our identification filter extends classical results concerning the design of unknown-input fault detection filters. In particular, our filter generalizes the construction of [6] to descriptor systems with direct 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 identifiable attack strategy. Notice that classical fault detection filters, for instance those presented in [6], are guaranteed to detect and isolate signals that do not excite exclusively zero dynamics. Finally, an equivalent attack identification filter for nonsingular or index-one systems is presented in our previous work [17].

Remark 7: (Complexity of centralized identification) Our centralized identification procedure assumes the knowledge of the cardinality of the attack set, and it achieves identification by constructing a residual generator for $(n+1)^k$ possible attack sets. Thus, our procedure constructs $O(n^k)$ filters. If only an upper bound $k$ on the cardinality of the attack set is available, identification can be achieved by constructing $(n+1)^k$ filters, and by intersecting the attack sets generating zero residuals.

In Section IV-C we show that this non-polynomial complexity is inherent to the identification problem.

V. ILLUSTRATIVE EXAMPLES

A. An example of state attack against a power network

Consider the power network model analyzed in Example 1 and illustrated in Fig. 2(a). We consider a load altering attack [18] affecting the power demand $P_3$ at the load buses 4 and 5. Assume that the variables $\theta_4$ and $\delta_5$ are affected by the unknown and unmeasurable signals $u_1$ and $u_2$. Suppose that a monitoring unit measures the state variables of the first generator, that is, $y_1 = \delta_1$ and $y_2 = \omega_1$.

Notice from Fig. 4(a) that the maximum size of a linking from the failure to the output vertices is 1, so that, by Theorem
3.5, there exists a structural vulnerability. In other words, for every choice of the network matrices, there exist nonzero $u_1$ and $u_2$ that are not detectable through the measurements.\footnote{When these output-nulling inputs $u_1$, $u_2$ are regarded as additional loads, then they are entirely sustained by the second and third generator.}

We now consider a numerical realization of this system. Let the input matrices be $B = [e_8 e_9]$ and $D = [0 0]^T$, the measurement matrix be $C = [e_1 e_1]^T$, and the system matrix $A$ be as in Remark 1 with $M_g = \text{blkdiag}(.125, .034, .016)$, $D_g = \text{blkdiag}(.125, .068, .048)$, and

$$L = \begin{bmatrix}
0.058 & 0 & 0 & -0.058 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.063 & 0 & 0 & -0.063 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.058 & 0 & 0 & -0.058 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.235 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.085 & -0.092 \\
0 & -0.063 & 0 & 0 & 0 & 0 & -0.296 & 0 & -0.161 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.330 & 0 & -0.170 \\
0 & 0 & 0 & 0 & 0 & 0 & -0.085 & -0.161 & 0 & -0.246 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.072 \\
0 & 0 & 0 & 0 & 0 & 0 & -0.101 & 0 & 0 & -0.173
\end{bmatrix}$$

Let $U(s)$ and $U_2(s)$ be the Laplace transform of the attack signals $u_1$ and $u_2$, and let

$$\begin{bmatrix}
U_1(s) \\
U_2(s)
\end{bmatrix} = \begin{bmatrix}
-1.024s^4 - 5.121s^3 - 10.34s^2 - 9.584s - 3.531 \\
s^5 + 5.8s^4 + 9.865s^3 + 9.173s^2 + 3.531
\end{bmatrix} \tilde{U}(s),$$

for some arbitrary nonzero signal $\tilde{U}(s)$. Then it can be verified that the attack cannot be detected through the measurements $y$. In fact, the transfer matrix mapping $\tilde{U}(s)$ to $U(s)$ coincides with the null space of the input/output transfer matrix. An example is in Fig. 4(b), where the second and the third generator are steered stable by the attack, but the first generator does not deviate from its nominal operating condition.

Let $K$ be the cardinality of the attack set. It can be verified that the attack set $K$ remains undetected by a static detector. In contrary, as shown in [9], a cyber attack may remain undetected by a static detector if it compromises as few as four measurements.

### C. An example of distributed detection

The IEEE 118 bus system shown in Fig. 3 is composed of 118 buses and 54 generators, and its parameters can be found in [51]. Following Section II, a linear continuous-time descriptor model of the system under attack takes the form (1).

For estimation and detection purposes, we partition the IEEE 118 system into 5 disjoint areas, we assign a control center to each area, and we implement our detection procedure via the filter (15); see Fig. 3 for a graphical illustration. Suppose that each control center continuously 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 comprises all measurements in area 1 by adding a signal $u_K$. It can be verified that the attack set $K$ is detectable, see Theorem 3.3. According to assumption (A3), the attack signal $u_K$ needs to be continuous to guarantee a continuous state trajectory (since the associated descriptor model is of index 1). To show the robustness of our detection filter (15), we let $u_K$ be discontinuous and randomly distributed in the interval $[0, 0.5]$ rad.

The control centers implement the distributed attack detection procedure described in (15), with $G = AG^T$. It can be verified that the pair $(E, A_D + GC)$ is Hurwitz stable, and that $\rho(\langle \omega E - A_D - GC \rangle^{-1} A_C) < 1$ for all $\omega \in \mathbb{R}$. As predicted by Theorem 4.3, our distributed attack detection filter is convergent; see Fig. 6(a). For completeness, in Fig. 6(b) we illustrate the convergence rate of our waveform relaxation-based filter as a function of the number of iterations.

![IEEE 14 bus system](image-url)


VI. CONCLUSION

For cyber-physical systems modeled by linear time-invariant descriptor systems, we have analyzed fundamental monitoring limitations. In particular, we have characterized undetectable and unidentifiable attacks from a system-theoretic and a graph-theoretic perspective. Additionally, we have designed centralized and distributed monitors.

Future and ongoing work includes (i) a detailed analysis of the convergence of our distributed monitor, (ii) the design of distributed identification monitors, and (iii) the design of monitors robust to system noise and unmodeled dynamics.

REFERENCES

Fig. 7. In Fig. 7(a) we report our simulation results for the case of linear network dynamics without noise and for the proposed detection monitor (5) and identification monitor (26), 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. In Fig. 7(b) 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. Due to the presence of noise, the residuals deviate from their nominal behavior reported in Fig. 7(a). Although the attack is clearly still detectable and identifiable, additional statistical tools such as hypothesis testing [7] may be adopted to analyze the residuals $r(t), r_K(t), \text{and } r_R(t)$. In Fig. 7(c) 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. Despite the presence of unmodeled nonlinear dynamics, the residuals reflect their nominal behavior reported in Fig. 7(a).


