Control-Theoretic Methods for Cyber-Physical Security

Fabio Pasqualetti, Florian Dörfler, and Francesco Bullo

Cyber-physical systems integrate physical processes, computational resources, and communication capabilities. Cyber-physical systems have permeated modern society becoming prevalent in many domains including energy production, health care, and telecommunications. Examples of cyber-physical systems include sensor networks, industrial automation systems, and critical infrastructures such as transportation networks, power generation and distribution networks, water and gas distribution networks, and advanced communication systems. The integration of cyber technologies with physical processes increases systems efficiency and, at the same time, introduces vulnerabilities that undermine the reliability of critical infrastructures. As recently highlighted by the Maroochy water breach in March 2000 [1], 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 in June 2010 [4], and 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 [6], [7].

Concerns about security of systems are not new, as the numerous manuscripts on systems fault detection, isolation, and recovery testify [8], [9]. Cyber-physical systems, however, suffer from specific vulnerabilities that do not affect classical systems, and for which appropriate detection and identification techniques need to be developed. For instance, the reliance of cyber-physical systems on communication networks and standard communication protocols to transmit measurements and control packets increases the possibility of intentional and unforeseen 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. In fact, these information security methods do not exploit the compatibility of the measurements with the underlying physical process or the control mechanism, and they are ineffective, for instance, against insider attacks and attacks targeting the physical dynamics [1].

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 [10] *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 [11] *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 [12]. In [13] 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 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 random signal unknown to the attacker into the system. In [14] 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 [15] 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 [16]. 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 [17]–[22], linear networks with misbehaving components [23]–[25], and water networks [26]–[28].

This article provides a self-contained presentation of recent control-theoretic approaches to cyber-physical security. We adopt the unified modeling framework for cyber-physical systems and attacks proposed in [29], where cyber-physical systems under attack are modeled as descriptor systems subject to unknown inputs altering the state and the measurements. With respect to [29], we provide a tutorial and self-contained presentation which includes all the necessary background material, detailed modeling sections, and additional examples of attacks against power systems and water networks. The framework presented in this paper is general to include previously described attack scenarios, yet it allows for a rigorous study of the detectability and identifiability of attacks, for a comprehensive analysis of the effects of attacks on the system, and for the design of monitors and attack remedial schemes. We start our presentation with the models of cyber-

physical systems, monitors, and attacks. For these models we define detectability and identifiability of attacks, and we derive fundamental detection and identification limitation from system-theoretic and graph-theoretic perspectives. Finally we discuss the monitor design problem, and we conclude with a case study on coordinated attacks against power networks.

I. Models Of Cyber-physical Systems, Monitors, and Attacks

Cyber-physical systems are ubiquitous in various domains including power networks, water distribution networks, sensor networks, dynamic Leontief models of multi-sector economies, mixed gas-electricity networks, and large-scale industrial control systems. In this section, we model cyber-physical systems under attack as linear time-invariant descriptor systems subject to unknown inputs. This modeling framework is very general and includes most of the existing cyber-physical models, attacks, and faults. In fact, as we show in Sidebar "Power network model" for power networks and in Sidebar "Transport network model" for water distribution networks, important real-world cyber-physical systems contain conserved physical quantities leading to differential-algebraic system descriptions. Additionally, most attack and fault scenarios can be modeled by additive inputs affecting the state and the measurements; see Sidebar "Stealth, replay, covert, and injection attacks".

Model of cyber-physical systems and attacks. We consider the linear time-invariant descriptor system

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

where $x : \mathbb{R} \to \mathbb{R}^n$ and $y : \mathbb{R} \to \mathbb{R}^p$ are the maps describing the evolution of the system state and measurements, respectively, and $E \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$ are constant matrices. In the present paper we allow the matrix E to be singular, and we note that the case of nonsingular systems ($E = I$) is a particular instance of this model. The inputs Bu and Du are unknown signals that describe disturbances affecting the system state and measurements. Besides reflecting the genuine failure of systems components, these disturbances model the effect of attacks against the cyber-physical system (see below for the attack model). Finally, it should be observed that we neglect the presence of known inputs affecting the system (1) because they do not affect the results on the detectability and identifiability of unknown input attacks; see [29] for a complete analysis, and Figure 1 for an illustration of the setup.

For notational convenience and without affecting generality, we assume that each state and output 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, $u = \begin{bmatrix} u_x^\mathsf{T}, u_y^\mathsf{T} \end{bmatrix}^\mathsf{T}$. The *attack* $(Bu, Du) = (u_x, u_y)$ can be classified as *state attack* $(Bu, 0)$ affecting the system dynamics, and as *output attack* $(0, Du)$ corrupting directly the measurements vector.

The attack signal $u : \mathbb{R} \to \mathbb{R}^{n+p}$ depends on the attack strategy. In the presence of $k \in \{1, \dots, n+p\}$ attackers indexed by the *attack set* $K \subseteq \{1, \dots, n+p\}$, all and only the entries K of u are nonzero over time. In order to underline this sparsity relation, we sometimes use u_K to denote the *attack mode*, that is the subvector of u indexed by K. Accordingly, the pair (B_K, D_K) denotes the *attack signature*, where B_K and D_K are the submatrices of B and D with columns in K. Thus, $Bu = B_K u_K$, and $Du = D_K u_K$. Because the matrix E can be singular, we make the following assumptions on system (1):

(A1) the pair (E, A) is regular, that is, the determinant $\det(sE - A)$ does not vanish identically,
(A2) the initial condition $x(0) \in \mathbb{R}^n$ is consistent, that is, the relation $(Ax(0) + Bu(0)) \in \mathrm{Im}(E)$ holds; and
(A3) the input signal u is smooth.

The regularity assumption (A1) ensures the existence of a unique solution x to (1). Assumptions (A2) and (A3) simplify the technical presentation in this article since they guarantee smoothness of the state trajectory x and the measurements y; see [29], [30] for further details.

Model of monitors. A monitor is a device to detect and identify attacks in a cyber-physical system. We consider a general class of monitors with knowledge of the system dynamics and measurements, that is, the monitor knows the system matrices E, A, C, and it has access to the measurements y at all times. We do not impose additional constraints on monitors, and we study their fundamental limitations in detecting and identifying attacks.

An example of monitor is the *bad data detector* [31]. The bad data detector takes as inputs the matrix C and the measurements y, and detects an attack whenever there is no physical state that satisfies the measurement equation $y = Cx$. In other words, the bad data detector detects an attack whenever the residual

$$r = y - CC^\dagger y, \tag{2}$$

is nonzero, where $C^\dagger$ denotes the Moore-Penrose pseudoinverse of the matrix C. Observe that the bad data detector detects only attacks of the form $(0, Du)$ with $Du \notin \mathrm{Im}(C)$. Other examples of monitors can be found in [12], [13], [19], and in Section III.

Model of attackers. 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 (Bu, Du) in (1) be designed based on knowledge of the system matrices E, A, C, and the full state x at all times. Additionally, attackers have unlimited computation capabilities, and their objective is to disrupt the physical state or the measurements while avoiding detection.

For a power network (see Sidebar 1 "Power network model"), attacks and faults modeled by additive inputs include:

(i) a change in the mechanical power input to generator i is described by the attack signature $(B_i, 0)$, and an arbitrary attack mode u_{n+i}. This attack can originate from a genuine loss of generation or load, a malicious attack via the governor control to disrupt the system functionality [17], or an internet-based load altering attack [20];

(ii) a line outage occurring on the line $\{r, s\}$ is modeled by the signature $([B_r\ B_s], [0\ 0])$ and an arbitrary attack mode $[u_r\ u_s]^\mathsf{T}$, see [32]; and

(iii) the failure of sensor i, or the corruption of the i-th measurement by an attacker is captured by the signature $(0, D_{2n+m+i})$ and a non-zero mode u_{2n+m+i}, see [11], [12], [16], [22] for examples of sensor attacks.

Likewise, for a water network (see Sidebar "Transport network model"), faults modeled by additive inputs include leakages, sudden changes of demand, and failures of pumps and sensors. Possible cyber-physical attacks include compromising the flow and pressure measurements to divert flow, and attacks on the hydraulic control architecture (pumps and valves). These attacks are modeled similarly to the power network attacks above.

II. Fundamental Attack Detection and Identification Limitations

In this section, we present system-theoretic and graph-theoretic conditions for the detectability and identifiability of attacks. These conditions are fundamental, in the sense that they hold independently of the monitoring device.

A. System-theoretic conditions

As discussed in Section I, monitors exploit only the system dynamics and measurements to reveal attacks. Consequently, an attack is undetectable if the measurements due to the attack are compatible with the measurements without the attack, that is, they coincide with the measurements due to some nominal operating condition. On the other hand, if the measurements due to the attack are not compatible with the system dynamics and measurements without attacks, then the attack can be detected. The following definitions summarize this discussion, where $y(x_0, u, t)$ denotes the system measurements at time t due to the attack u and initial state x_0.

Definition 1: ***(Undetectable attack)*** For the descriptor system (1) with initial state x_0, the attack $(B_K u_K, D_K u_K)$ is undetectable if $y(x_0, u_K, t) = y(x_1, 0, t)$ for some initial state $x_1 \in \mathbb{R}^n$ and for all $t \in \mathbb{R}_{\geq 0}$.

A more general concern than detectability is identifiability of attacks, that is, the possibility for a monitor to distinguish between two distinct sets of attackers. Recall that attackers can independently compromise any state variable or measurement.

Definition 2: ***(Unidentifiable attack)*** For the descriptor system (1) with initial state x_0, the attack $(B_K u_K, D_K u_K)$ is unidentifiable if $y(x_0, u_K, t) = y(x_1, u_R, t)$ for some initial state $x_1 \in \mathbb{R}^n$, attack $(B_R u_R, D_R u_R)$ with $|R| \leq |K|$ and $R \neq K$, and for all $t \in \mathbb{R}_{\geq 0}$.

Following Definition 1, an attack set is undetectable if it can result in undetectable attacks. Likewise, an attack set is unidentifiable if it can result in unidentifiable attacks.

We now elaborate on the above definitions to derive fundamental detection and identification limitations. Observe that, due to linearity of (1), the detectability condition in Definition 1 can be equivalently rewritten as follows: for the descriptor system (1) with initial state x_0, the attack $(B_K u_K, D_K u_K)$ is undetectable if and only if $y(x_2, u_K, t) = 0$ for some initial state $x_2 \in \mathbb{R}^n$ (namely $x_2 = x_0 - x_1$ for some $x_1 \in \mathbb{R}^n$) and for all $t \in \mathbb{R}_{\geq 0}$. The relation $y(x_2, u_K, t) = 0$ can be satisfied at all times if and only if the attack u_K excites only the *zero dynamics* of the input/output dynamical system; see Sidebar 5 "Invariant zeros and zero dynamics" and [30], [33], [34]. Thanks to this interpretation and the notion of *invariant zeros*, we are led to the following algebraic characterization of undetectable attack sets.

Theorem 2.1: ***(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 $s \in \mathbb{C}$, $g \in \mathbb{C}^{|K|}$, and $x \in \mathbb{C}^n$, with $x \neq 0$, such that

$$\begin{aligned} (sE - A)x - B_K g &= 0, \\ Cx + D_K g &= 0. \end{aligned}$$

In other words, the existence of undetectable attacks for the system (E, A, B_K, C, D_K) is equivalent to the existence of invariant zeros for the same attack/measurements system. On the other hand, undetectable attacks exist only if the cardinality of the attack set is sufficiently large. To see this, let $\|x\|_{\ell_0} = |\text{supp}(x)|$ denote the number of nonzero components of the vector x. Observe that condition (ii) of Theorem 2.1 can be satisfied if and only if the cardinality of the attack set satisfies $|K| \geq \|(sE - A)x\|_0 + \|Cx\|_0$ for some vector x. The choice of the vector x determines the cardinality of the attack set, being therefore a suitable optimization variable for the design of undetectable attacks with smallest cardinality.

Analogously to the detectability condition, the identifiability condition in Definition 2 can be equivalently rewritten as follows: for the descriptor system (1) with initial state x_0, the attack $(B_K u_K, D_K u_K)$ is unidentifiable if and only if $y(x_2, u_K - u_R, t) = 0$ for some initial state $x_2 \in \mathbb{R}^n$, for some attack $(B_R u_R, D_R u_R)$ with $|R| \leq |K|$ and $R \neq K$, and for all $t \in \mathbb{R}_{\geq 0}$. The following result gives an algebraic characterization of identifiability.

Theorem 2.2: ***(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 $R \neq K$, $s \in \mathbb{C}$, $g_K \in \mathbb{C}^{|K|}$, $g_R \in \mathbb{C}^{|R|}$, and $x \in \mathbb{C}^n$, with $x \neq 0$, such that

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

Condition (ii) in Theorem 2.2 can be written by collecting the input matrices as follows:

$$\begin{aligned} (sE - A)x - \begin{bmatrix} B_K & B_R \end{bmatrix} \begin{bmatrix} g_K \\ g_R \end{bmatrix} &= 0, \\ Cx + \begin{bmatrix} D_K & D_R \end{bmatrix} \begin{bmatrix} g_K \\ g_R \end{bmatrix} &= 0. \end{aligned} \tag{3}$$

From equation (3) and Theorem 2.1 we conclude that the existence of unidentifiable attack sets of cardinality k is equivalent to the existence of undetectable attack sets of

cardinality $2k$, that is, to the existence of invariant zeros for the system $(E, A, B_{\bar{K}}, C, D_{\bar{K}})$ with $|\bar{K}| \leq 2k$.

B. Graph-Theoretic conditions

In this section, we describe graph-theoretic conditions for the detectability of attacks. We refer the reader to Sidebar 6 "Graph theory and generic properties" for the notions in graph theory and algebraic geometry used in this section.

For the system (E, A, B, C, D), construct the directed *attack/state/output graph* $\mathcal{G}_{\text{aso}} = (\mathcal{V}_{\text{aso}}, \mathcal{E}_{\text{aso}})$ by defining the vertex set as

$$\mathcal{V}_{\text{aso}} = \mathcal{U}_{\text{aso}} \cup \mathcal{X}_{\text{aso}} \cup \mathcal{Y}_{\text{aso}},$$

where $\mathcal{U}_{\text{aso}} = \{u_1, \ldots, u_m\}$ is the set of attack vertices, $\mathcal{X}_{\text{aso}} = \{x_1, \ldots, x_n\}$ is the set of state vertices, and $\mathcal{Y}_{\text{aso}} = \{y_1, \ldots, y_p\}$ is the set of output vertices, and the edge set as

$$\mathcal{E}_{\text{aso}} = \mathcal{E}_E \cup \mathcal{E}_A \cup \mathcal{E}_B \cup \mathcal{E}_C \cup \mathcal{E}_D,$$

where

$$\begin{aligned}
&\mathcal{E}_E = \{(x_j, x_i) : E_{ij} \neq 0\},\ \mathcal{E}_A = \{(x_j, x_i) : A_{ij} \neq 0\},\\
&\mathcal{E}_B = \{(u_j, x_i) : B_{ij} \neq 0\},\ \mathcal{E}_C = \{(x_j, y_i) : C_{ij} \neq 0\},\\
&\mathcal{E}_D = \{(u_j, y_i) : D_{ij} \neq 0\}.
\end{aligned}$$

Various properties of the dynamical system (E, A, B, C, D) can be expressed as properties of its associated graph $\mathcal{G}_{\text{aso}}$ [35], [36].

The dynamical system $(\bar{E}, \bar{A}, \bar{B}, \bar{C}, \bar{D})$ with attack/state/output graph $\bar{\mathcal{G}}_{\text{aso}} = (\bar{\mathcal{V}}_{\text{aso}}, \bar{\mathcal{E}}_{\text{aso}})$ is *compatible* with (E, A, B, C, D) if $\bar{\mathcal{G}}_{\text{aso}}$ is a subgraph of $\mathcal{G}_{\text{aso}}$ with $\bar{\mathcal{V}}_{\text{aso}} = \mathcal{V}_{\text{aso}}$ and $\bar{\mathcal{E}}_{\text{aso}} \subseteq \mathcal{E}_{\text{aso}}$. In other words, the system $(\bar{E}, \bar{A}, \bar{B}, \bar{C}, \bar{D})$ is compatible with the system (E, A, B, C, D) if the matrices $\bar{E}$, $\bar{A}$, $\bar{B}$, $\bar{C}$, and $\bar{D}$ can be obtained from the matrices E, A, B, C, and D by changing only their nonzero entries. A system property is *generic* if it holds for *almost all* compatible systems. Many system properties turn out to be generic, and hence robust to uncertainties in the system parameters.

Recall from Definition 1 that an attack u is undetectable if $y(x_0, u, t) = y(x_1, 0, t)$ at all times t for some initial states x_0 and x_1. As a particular case, if the system initial state is known, an attack u is undetectable if $y(x_0, u, t) = y(x_0, 0, t)$ for some initial state x_0. This attack undetectability condition is equivalent to the system (E, A, B, C, D) failing to be left-invertible; see Sidebar 5 "Invariant zeros and zero dynamics".

Theorem 2.3: ***(Generically undetectable attack)*** Let $\mathcal{G}_{\text{aso}}$ be the attack/state/output graph associated with the descriptor system (1) and attack set K. Assume that the system initial state is known, and that the determinant $\det(sE - A) \neq 0$ for some values of $s \in \mathbb{C}$. The following statements are equivalent:

(i) the attack set K is generically undetectable; and
(ii) the graph $\mathcal{G}_{\text{aso}}$ contains no linking of size $|K|$ from $\mathcal{U}_{\text{aso}}$ to $\mathcal{Y}_{\text{aso}}$.

In Theorem 2.3 we show that, if the attack/state/output graph is sufficiently connected and the system initial state is known, then there are no undetectable attacks for almost all compatible systems, that is, for almost every choice of the numerical entries of the system matrices. Conversely, if a system admits a generically undetectable attack set, then every compatible system admits an undetectable attack set. See Sidebar 1 "Power network model and attack example" for an illustrative example of this result.

If the system initial state is unknown, then an undetectable attack u is characterized by the existence of a pair of initial conditions x_0 and x_1 such that $y(x_0, u, t) = y(x_1, 0, t)$, 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 $[\xi_1^{\mathsf{T}}\ \xi_2^{\mathsf{T}}]^{\mathsf{T}}$, where ξ_1 corresponds to the dynamic variables. Let the network matrices E, A, B, C, and D be partitioned accordingly, and assume that the descriptor system (1) is given in *semi-explicit* form, that is, $E = \text{blkdiag}(E_{11}, 0)$ and E_{11} is nonsingular. As a matter of fact, many cyber-physical systems, such as power and mass-transport networks, are readily given in semi-explicit form. In this case, the descriptor system (1) reads as

$$\begin{aligned}
E_{11}\dot{\xi}_1 &= A_{11}\xi_1 + A_{12}\xi_2 + B_1 u\,,\\
0 &= A_{21}\xi_1 + A_{22}\xi_2 + B_2 u\,,\\
y &= C_1\xi_1 + C_2\xi_2 + Du\,.
\end{aligned} \tag{4}$$

Consider now the associated nonsingular state space system that is obtained by regarding ξ_2 as an external input and the algebraic constraint as an output:

$$\begin{aligned}
\dot{\xi}_1 &= E_{11}^{-1}A_{11}\xi_1 + E_{11}^{-1}A_{12}\xi_2 + E_{11}^{-1}B_1 u,\\
\tilde{y} &= \begin{bmatrix} A_{21} \\ C_1 \end{bmatrix} \xi_1 + \begin{bmatrix} A_{22} & B_2 \\ C_2 & D \end{bmatrix} \begin{bmatrix} \xi_2 \\ u \end{bmatrix}.
\end{aligned} \tag{5}$$

Under the assumption of left-invertibility of the system (4), the invariant zeros of the systems (4) and (5) coincide. Because the system (5) is nonsingular, graph-theoretic results in control can be used to investigate the presence of generically undetectable attacks in singular cyber-physical systems. For instance, from [35, Theorem 4] we have that system (4) admits generically undetectable attacks if (i) the system initial state is unknown, (ii) the number of attack vertices equals the number of output vertices, (iii) the system is left-invertible, and (iv) in the graph $\mathcal{G}_{\text{aso}}$ the vertices $\mathcal{X}_{\text{aso}}$ are not contained in some linking of size $|K|$ from $\mathcal{U}_{\text{aso}}$ to $\mathcal{Y}_{\text{aso}}$.

III. Design of Attack Detection and Identification Monitors

In the previous sections, we derived fundamental limitations and conditions characterizing attack detectability and identifiability by monitors. In this section, we address the converse problem of designing monitors to detect and identify attacks. Monitors can be designed in different ways, depending on the knowledge of the system dynamics, the available measurements, and the communication constraints. For the considered setup, we design monitors by leveraging and extending fault detection and isolation techniques; see Sidebar 4 "Geometric control theory and its application to fault detection and isolation" and [8]. Within this article, we

focus on the design of centralized monitors with access to all measurements y and with detailed knowledge of the system matrices (E, A, C). We refer to [29], [37], [38] for extensions to distributed monitors with local knowledge of the system dynamics, with access to locally available measurements, and subject to communication constraints.

We first focus on the design of an attack detection monitor, which is designed as a continuous-time residual filter with input the system measurements $y : \mathbb{R}_{\geq 0} \to \mathbb{R}^p$ and output the residual signal $r : \mathbb{R}_{\geq 0} \to \mathbb{R}^p$. Consider the modified Luenberger observer

$$\begin{aligned} E\dot{w} &= (A + GC)w - Gy, \\ r &= Cw - y, \end{aligned} \tag{6}$$

where the output injection matrix $G \in \mathbb{R}^{n\times p}$ is selected so that the pair (E, A) is regular and Hurwitz, that is, its finite spectrum $\sigma(E, A) = \{\lambda \ : \ \lambda \in \mathbb{C}, |\lambda| < \infty, \det(\lambda E - A) = 0\}$ lies in the open left half-plane. If the system initial state $x(0)$ is known and the filter (6) is initialized with $w(0) = x(0)$, then an analysis of the filter error dynamics $w - x$ yields that the residual r is identically zero if and only if the attack $(B_K u_K, D_K u_K)$ is either identically zero (no attack) or undetectable. We conclude that the proposed filter (6) is a *complete* monitor, that is, it detects every detectable attack.

Theorem 3.1: ***(Complete attack detection monitor)*** Consider the descriptor system (1), and assume that the attack set K is detectable and the initial state $x(0) \in \mathbb{R}^n$ is known. Consider the *attack detection filter* (6), where $w(0) = x(0)$ and $G \in \mathbb{R}^{n\times p}$ 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 $t \in \mathbb{R}_{\geq 0}$.

Several comments are in order. First, if the initial state $x(0)$ is not available, then an arbitrary initial state $w(0) \in \mathbb{R}^n$ can be chosen and the filter (6) has an asymptotic performance: the filter error $w-x$ converges asymptotically, and the residual r (in the absence of attacks) becomes zero only in the limit as time goes to infinity. Second, if the filter (6) is implemented only over a finite and nontrivial interval of time, then the residual r being zero in this interval is equivalent to the attack signal u_K being zero for this interval. Third, the filter (6) can be implemented using locally available information and distributed computation; see [38] for details. Fourth, the dynamics and the measurements of (1) may be affected by modeling uncertainties and noise with known statistics. In a practical implementation the output injection matrix G should be chosen to optimize the sensitivity of the residual r to attacks versus the effect of noise, or to optimize the transient behavior of the filter. Statistical hypothesis testing techniques [9] are subsequently used to analyze the residual r for sufficiently large but finite horizons. We remark that attacks hiding in the transient dynamics or aligned with the noise statistics may remain undetected.

In comparison to the attack detection problem, the attack identification problem is inherently combinatorial and computationally hard. If the cardinality of the attack set is known, the identification of the attack set K requires a combinatorial procedure because, a priori, K is one of the $\binom{n+p}{|K|}$ possible attack sets. The following attack identification procedure consists of designing a residual filter for a candidate attack set to determine whether the candidate set coincides with the actual attack set.

For simplicity, we consider the case of nonsingular systems ($E = I$) in the absence of output attacks $D_K = 0$; see [29] for a more general treatment. The identification monitor design is akin to the design of residual generators (S4) in fault detection and isolation, and it relies on the notion of *conditioned invariant* subspaces from geometric control theory; see Sidebar 4 "Geometric control theory and its application to fault detection and isolation". Define the subspace $\mathcal{S}_K$ to be the smallest $(A, \mathrm{Ker}(C))$-conditioned invariant subspace containing $\mathrm{Im}(B_K)$, and let $J_K \in \mathbb{R}^{p\times n}$ be an output injection matrix rendering this subspace invariant, that is,

$$(A + J_K C)\mathcal{S}_K \subseteq \mathcal{S}_K.$$

Consider the orthonormal matrix $T_K = [W_K^{\mathsf{T}} \ P_K^{\mathsf{T}}] \in \mathbb{R}^{n\times n}$, where W_K is a basis of $\mathcal{S}_K$ and P_K is a basis of the quotient space $\mathbb{R}^n \setminus \mathcal{S}_K$. In the coordinates $[\xi_1, \xi_2] = [W_K x, P_K x]$ and with the output injection J_K, system (1) reads as

$$\begin{aligned} \begin{bmatrix} \dot{\xi}_1 \\ \dot{\xi}_2 \end{bmatrix} &= \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ 0 & \hat{A}_{22} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \hat{B}_K \\ 0 \end{bmatrix} u_K \,, \\ y(t) &= \begin{bmatrix} \hat{C}_1 & \hat{C}_2 \end{bmatrix} \begin{bmatrix} \xi_1(t) \\ \xi_2(t) \end{bmatrix} \,, \end{aligned} \tag{7}$$

where $\hat{A} = T_K^{\mathsf{T}}(A + J_K C)T_K$, $\hat{B}_K = T_K^{\mathsf{T}} B_K$, and $\hat{C} = CT_K$. Hence, the effect of the input u_K is contained in the "contaminated" ξ_1-dynamics, and the ξ_2-dynamics are "secure". The measurement equation $y = \hat{C}\xi$ can be projected on the image of $\hat{C}_1$ and its orthogonal complement as

$$\begin{bmatrix} \tilde{y} \\ \bar{y} \end{bmatrix} = \begin{bmatrix} \hat{C}_1 & \hat{C}_1 \hat{C}_1^{\dagger} \hat{C}_2 \\ 0 & (I - \hat{C}_1 \hat{C}_1^{\dagger}) \hat{C}_2 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \,, \tag{8}$$

where $\bar{y} = (I - \hat{C}_1 \hat{C}_1^{\dagger})y$ is the secure component of the output unaffected by ξ_1. Hence, we can design a residual filter for the secure ξ_2-dynamics using the secure output $\bar{y}$.

Theorem 3.2: ***(Complete attack identification monitor for the attack set*** K***)*** Consider the descriptor system (1) with attack set K in the coordinates (7). Assume that the attack set is identifiable and the network initial state $x(0)$ is known. Consider the *attack identification filter for the attack signature* (B_R, D_R), with $|R| = |K|$,

$$\begin{aligned} \dot{w} &= \left(\hat{A}_{22} + G(I - \hat{C}_1 \hat{C}_1^{\dagger}) \hat{C}_2 \right) w - G\bar{y} \,, \\ r_R &= (I - \hat{C}_1 \hat{C}_1^{\dagger}) \hat{C}_2 w - \bar{y} \,, \end{aligned} \tag{9}$$

where $w(0) = \xi_2(0)$, and G is such that $\hat{A}_{22} + G(I - \hat{C}_1 \hat{C}_1^{\dagger}) \hat{C}_2$ is Hurwitz, and $\bar{y}$ is the secure output defined in (8). Then the residual satisfies $r_R(t) = 0$ at all times $t \in \mathbb{R}_{\geq 0}$ if and only if R coincides with the attack set, that is, if and only if $R = K$.

Theorem 3.2 implies that the attack set K can be identified by constructing $\binom{n+p}{|K|}$ residual filters (9), one for each distinct attack set of cardinality $|K|$. In [29] we show that this non-polynomial complexity is inherent to the attack identification problem, which is generally NP-hard. We remark that, for the case of output attacks, an efficient (yet incomplete) approach

is to reformulate the attack identification problem as a convex optimization problem using heuristic convex relaxations [16].

IV. Coordinated Attacks in Power Networks

In this section we consider a network of utility companies that compete in the production of electrical energy. In particular, we consider the case where a group of utility companies form a coalition to compromise the functionality of their business rivals through a coordinated and destabilizing attack. A similar power network scenario is studied in [17]. We start by reviewing the setup and the attack strategy.

Consider a connected power transmission network with n generators $G_{\rm m} = \{g_1, \ldots, g_n\}$, where the generators rotor dynamics are modeled by second-order linear swing equations subject to governor control, and the power flows along lines are modeled by the DC approximation; see Sidebar 1 "Power network model and attack example". Assume that a subset $K = \{k_1, \ldots, k_m\}$ of generators is driven by an additional control action besides the primary frequency control. After elimination of the load bus variables through Kron reduction, the power network dynamics subject to the additional control u at the generators K read as

$$\dot{x} = Ax + B_K u_K, \tag{10}$$

where $x = [\theta^{\mathsf{T}}, \omega^{\mathsf{T}}]^{\mathsf{T}}$ contains the generators rotor angles and frequencies, $A \in \mathbb{R}^{2n \times 2n}$, and $B_K = I_K \in \mathbb{R}^{2n \times m}$, where $I_K = [e_{n+k_1} \cdots e_{n+k_m}]$ and e_i is the i-th canonical vector in $\mathbb{R}^{2n}$. We propose the following attack: the generators K form a coalition, select some sacrificial machines $\bar{K} \subseteq K$, and implement a coordinated control strategy (see below) to destabilize the other generators $G_{\rm m} \setminus K$, while maintaining satisfactory performance within the group $K \setminus \bar{K}$.

The attack strategy relies on the notion of *controlled invariant* subspace from geometric control theory; see Sidebar 4 "Geometric control theory and its application to fault detection and isolation". In particular, the colluding generators inject an attack input that remains undetectable by the generators $K \setminus \bar{K}$, while affecting the generators $G_{\rm m} \setminus K$. The attack input is of the form

$$u_K = Fx + \bar{B}_K^{\dagger} v, \tag{11}$$

where the matrix F and $\bar{B}_K$ satisfy the conditions

$$(A + B_K F)\mathcal{V}^* \subseteq \mathcal{V}^*, \tag{12}$$

and

$$\bar{B}_K = \mathrm{Basis}(\mathcal{V}^* \cap \mathrm{Im}(B_K)). \tag{13}$$

In the above equation (12), $\mathcal{V}^*$ denotes the largest $(A, \mathrm{Im}(B_K))$-controlled invariant subpace contained in $\mathrm{Ker}(C)$, where Cx is the vector of the frequencies of the generators $K \setminus \bar{K}$. Notice that the subspace $\mathrm{Im}(C)$ identifies the generators $K \setminus \bar{K}$, while $\mathrm{Ker}(C)$ identifies the generators $G_{\rm m} \setminus K$ and the sacrificial machines $\bar{K}$.

The attack input (11) consists of two components. The open-loop component $\bar{B}_K^{\dagger} v$ alters the behavior of the sacrificial machines only. In fact, $\mathrm{Im}(\bar{B}_K) \subseteq \mathcal{V}^* \subseteq \mathrm{Ker}(C)$. The input $v : \mathbb{R} \to \mathbb{R}^n$ is an arbitrary signal designed by the attackers to optimize some performance function, such as, the effect of the malicious control on the sacrificial machines, the energy of the malicious control, or the information pattern required to implement the malicious control. The closed-loop component Fx ensures that the generators $K \setminus \bar{K}$ are not affected by networks dynamics evolving in the subspace $\mathcal{V}^*$. In fact, dynamics in the subspace $\mathcal{V}^*$ are invariant due to (12), and they do not affect the generators $K \setminus \bar{K}$ because $\mathcal{V}^* \subseteq \mathrm{Ker}\, C$. Because the open-loop component of the attack excites only dynamics in $\mathcal{V}^*$ due to (13), we conclude that the attack (11) does not affect the generators $K \setminus \bar{K}$, while altering the behavior of the sacrificial machines and, consequently, of the generators $G_{\rm m} \setminus K$. Notice that the attack (11) is undetectable from the measurements taken at the generators $K \setminus \bar{K}$.

Theorem 4.1: ***(Malicious attacks)*** Consider the network-reduced power system model (10) with controlled generators K and sacrificial machines $\bar{K} \subseteq K$. Let $\bar{C} = I_{K \setminus \bar{K}}^{\mathsf{T}}$, let $\mathcal{V}^*$ be the largest $(A, \mathrm{Im}(B_K))$-controlled invariant subspace contained in $\mathrm{Ker}(\bar{C})$, let the state feedback F satisfy $(A + B_K F)\mathcal{V}^* \subseteq \mathcal{V}^*$, let $\bar{B}_K = \mathrm{Basis}(\mathcal{V}^* \cap \mathrm{Im}(B_K))$, and let $\mathcal{S}^*$ be the smallest $(A, \mathrm{Ker}(\bar{C}))$-conditioned invariant subspace containing $\mathrm{Im}(B_K)$. Then, for every input $v : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$, the attack $u = Fx + \bar{B}_K^{\dagger} v$ affects the generators $\bar{K} \cup G_{\rm m} \setminus K$ only.

The attack (11) is very general, and in fact it includes all attacks that can be cast by the generators K without affecting the generators $K \setminus \bar{K}$, including the strategy proposed in [17]. To illustrate the effectiveness of the attack (11), consider an aggregated model of the Western North American power grid as illustrated in Figure 2. This model is often studied in the context of inter-area oscillations [39]. Assume that the generators $\{1, 9\}$ form a coalition, and that generator 9 is the sacrificial machine. Following Theorem 4.1, a malicious attack $u = Fx + \bar{B}_K^{\dagger} v$ is cast by the generators $\{1, 9\}$ such that (i) generator 1 is not affected by the attack, (ii) generator 2 maintains an acceptable working condition even in the presence of the attack, and (iii) large frequency oscillations are induced at all other generators $G_{\rm m} \setminus K$. As a consequence of the attack, the linear model (10) is driven far away from the operating point, and the corresponding original nonlinear model eventually looses stability. In a real-world scenario the generators $G_{\rm m} \setminus K$ would be disconnected to maintain safety.

In the above scenario, assume that each generator monitors its own state variables, and that at most two generators may be colluding to disrupt the network. Notice that detectability of the malicious attacks designed in Theorem 4.1 is guaranteed for each generator affected by the attack. Unfortunately, the colluding generators cannot be identified from the measurements of any single generator. To see this, let B_K be the input matrix associated with any set K of two generators, and let $C_i = e_i^{\mathsf{T}}$ be the output matrix associated with generator i. It can be verified that for every K and i the system (A, B_K, C_i) is right-invertible [33], that is, the output $C_i x$ can be arbitrarily assigned by any coalition of two generators. We conclude that the measurements taken by generator i can be generated by any set of two generators, so that the colluding generators are not identifiable by generator i.

V. Conclusion

Cyber-physical systems are complex systems integrating physical processes with cyber infrastructures. For security assessment, cyber-physical systems can be conveniently modeled by linear time-invariant descriptor systems, where the algebraic constraints capture the presence of conserved physical quantities in the system. For cyber-physical systems modeled by descriptor systems, attacks can be represented by exogenous inputs altering the system dynamics and the measurements. With this representation of attacks it is possible (i) to characterize fundamental attack detection and identification limitations, (ii) to analyze the effect of attacks on the system, and (iii) to design monitors capable of revealing and locating attacks independently of the attack strategy and implementation. This article contains a self-contained discussion of cyber-physical security, including modeling, system-theoretic and graph-theoretic security analysis, monitor design, and illustrative examples.

References

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

[2] J. P. Conti, "The day the samba stopped," *Engineering Technology*, vol. 5, no. 4, pp. 46–47, 06 March - 26 March, 2010.

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

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

[5] G. Richards, "Hackers vs slackers," *Engineering & Technology*, vol. 3, no. 19, pp. 40–43, 2008.

[6] 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.

[7] A. A. Cárdenas, S. Amin, and S. S. Sastry, "Research challenges for the security of control systems," in *Proceedings of the 3rd Conference on Hot Topics in Security*, Berkeley, CA, USA, 2008, pp. 6:1–6:6.

[8] 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.

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

[10] S. Amin, A. Cárdenas, and S. Sastry, "Safe and secure networked control systems under denial-of-service attacks," in *Hybrid Systems: Computation and Control*, vol. 5469, Apr. 2009, pp. 31–45.

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

[12] A. Teixeira, S. Amin, H. Sandberg, K. H. Johansson, and S. Sastry, "Cyber security analysis of state estimators in electric power systems," in *IEEE Conf. on Decision and Control*, Atlanta, GA, USA, Dec. 2010, pp. 5991–5998.

[13] 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.

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

[15] M. Zhu and S. Martínez, "Stackelberg-game analysis of correlated attacks in cyber-physical systems," in *American Control Conference*, San Francisco, CA, USA, Jul. 2011, pp. 4063–4068.

[16] 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.

[17] C. L. DeMarco, J. V. Sariashkar, and F. Alvarado, "The potential for malicious control in a competitive power systems environment," in *IEEE Int. Conf. on Control Applications*, Dearborn, MI, USA, 1996, pp. 462–467.

[18] 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.

[19] 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.

[20] A.-H. Mohsenian-Rad and A. Leon-Garcia, "Distributed internet-based load altering attacks against smart power grids," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 667–674, 2011.

[21] S. Sridhar, A. Hahn, and M. Govindarasu, "Cyber–physical system security for the electric power grid," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 1–15, 2012.

[22] A. Giani, E. Bitar, M. Garcia, M. McQueen, P. Khargonekar, and K. Poolla, "Smart grid data integrity attacks: characterizations and countermeasures π," in *IEEE Int. Conf. on Smart Grid Communications*, Brussels, Belgium, 2011, pp. 232–237.

[23] 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.

[24] 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.

[25] M. Zhu and S. Martínez, "On distributed convex optimization under inequality and equality constraints," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 151–164, 2012.

[26] S. Amin, X. Litrico, S. S. Sastry, and A. M. Bayen, "Stealthy deception attacks on water SCADA systems," in *Hybrid Systems: Computation and Control*, Stockholm, Sweden, Apr. 2010, pp. 161–170.

[27] D. G. Eliades and M. M. Polycarpou, "A fault diagnosis and security framework for water systems," *IEEE Transactions on Control Systems Technology*, vol. 18, no. 6, pp. 1254–1265, 2010.

[28] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," 2012, available at `http://arxiv.org/abs/1212.0226`.

[29] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, Aug. 2012, to appear.

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

[31] A. Abur and A. G. Exposito, *Power System State Estimation: Theory and Implementation*. CRC Press, 2004.

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

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

[34] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 3rd ed. Springer, 1985.

[35] J. M. Dion, C. Commault, and J. van der Woude, "Generic properties and control of linear structured systems: a survey," *Automatica*, vol. 39, no. 7, pp. 1125–1144, 2003.

[36] K. J. Reinschke, *Multivariable Control: A Graph-Theoretic Approach*. Springer, 1988.

[37] F. Pasqualetti, F. Dörfler, and F. Bullo, "Cyber-physical security via geometric control: Distributed monitoring and malicious attacks," in *IEEE Conf. on Decision and Control*, Maui, HI, USA, Dec. 2012, pp. 3418–3425.

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

[39] D. J. Trudnowski, J. R. Smith, T. A. Short, and D. A. Pierre, "An application of Prony methods in PSS design for multimachine systems," *IEEE Transactions on Power Systems*, vol. 6, no. 1, pp. 118–126, 1991.

[40] F. Pasqualetti, A. Bicchi, and F. Bullo, "A graph-theoretical characterization of power network vulnerabilities," in *American Control Conference*, San Francisco, CA, USA, Jun. 2011, pp. 3918–3923.

[41] A. Osiadacz, *Simulation and Analysis of Gas Networks*. Houston, TX, USA: Gulf Publishing Company, 1987.

[42] A. Kumar and P. Daoutidis, *Control of Nonlinear Differential Algebraic Equation Systems*. CRC Press, 1999.

[43] X. Litrico and V. Fromion, *Modeling and Control of Hydrosystems*. Springer, 2009.

[44] J. Burgschweiger, B. Gnädig, and M. C. Steinbach, "Optimization models for operative planning in drinking water networks," *Optimization and Engineering*, vol. 10, no. 1, pp. 43–73, 2009.

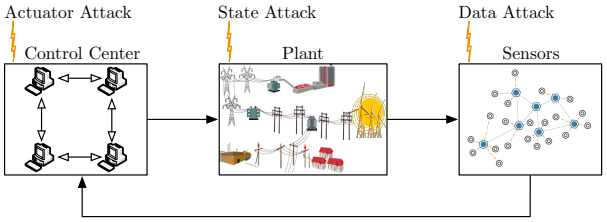

Fig. 1. Cyber-physical systems integrate physical and cyber layers, and are prone to attacks on all components. The dynamics of cyber-physical systems can be represented as $E\dot{x} = Ax$, and attacks as unknown inputs (Bu, Du). State and actuator attacks are modeled by the input Bu that affect directly the system dynamics. Output or data attacks corrupt the system measurements, and are modeled by the input Du. State, actuator, and data attacks can be implemented by physical or cyber tampering with the system components.

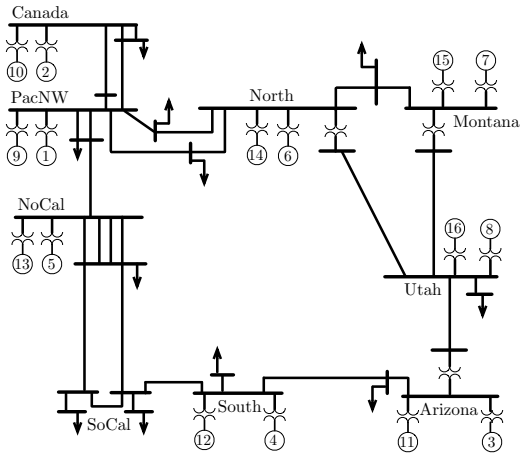

Fig. 2. A schematic diagram of the Western North American power grid.

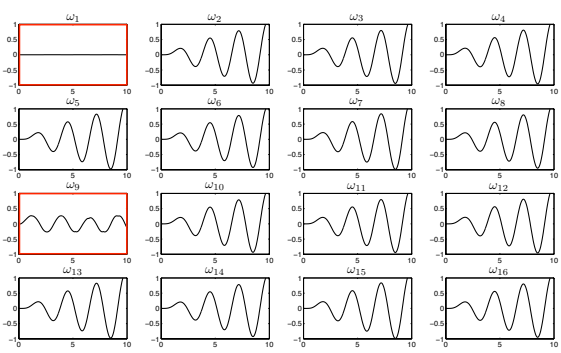

Fig. 3. This figure reports the deviations from steady state of the generators frequencies due to the coordinated attack in Section IV. All deviations have been normalized so that the unit value indicates a safety limit. The attack input is of the form (11), where the input v is chosen such that the infinity norm of ω_9 is minimized, subject to the infinity norm of ω_{16} being no less than 1. Notice that (i) generator 1 is not affected by the attack, (ii) generator 9 maintains satisfactory performance, and (iii) the remaining generators are severely affected by the coordinated attack.

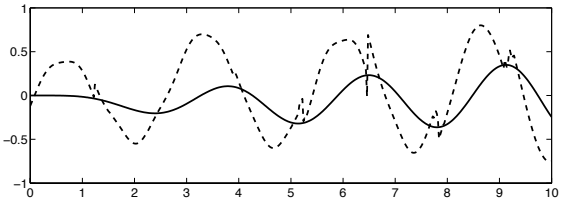

Fig. 4. This figures shows the coordinated attack input discussed in Section IV. The attack is implemented by modifying the governor control input of generator 1 (solid) and generator 9 (dashed). Both plots are in p.u. values and for the linear system (10), that is, measured as deviation from steady state.

[45] P. F. Boulos, K. E. Lansey, and B. W. Karney, *Comprehensive Water Distribution Systems Analysis Handbook for Engineers and Planners*. American Water Works Association, 2006.

[46] L. A. Rossman, "Epanet 2, water distribution system modeling software," US Environmental Protection Agency, Water Supply and Water Resources Division, Tech. Rep., 2000.

[47] Y. Mo and B. Sinopoli, "False data injection attacks in control systems," in *First Workshop on Secure Control Systems*, Stockholm, Sweden, Apr. 2010.

[48] H. L. Trentelman, A. Stoorvogel, and M. Hautus, *Control Theory for Linear Systems*. Springer, 2001.

[49] F. L. Lewis, "A tutorial on the geometric analysis of linear time-invariant implicit systems," *Automatica*, vol. 28, no. 1, 1992.

Sidebar 1: ***Power network model and attack example.*** Future power networks will be equipped with a sophisticated coordination infrastructure to control the volatile physical dynamics due to renewable energy sources and deregulation of energy markets. The cyber-physical security of the future "smart grid" has been identified as an issue of primary concern [6], [21], and it has recently attracted the interest of the control and power systems communities, see [12], [17]–[22], [32], [40].

In order to describe a power network, we adopt the small-signal version of the classic structure-preserving power network model, which we now briefly recall. We refer the interested reader to [32], [40] for a detailed derivation from the full nonlinear structure-preserving power network model. Consider a connected power network consisting of n generators $\{g_1, \dots, g_n\}$ and m load buses $\{b_{n+1}, \dots, b_{n+m}\}$. The interconnection structure of the power network is encoded by a connected susceptance-weighted graph G. The vertices of G are the generators g_i and the buses b_i. The edges of G are the transmission lines $\{b_i, b_j\}$ and the connections $\{g_i, b_i\}$, weighted by their susceptance values. The Laplacian associated with the susceptance-weighted graph is the symmetric susceptance matrix $\mathcal{L} \in \mathbb{R}^{(n+m)\times(n+m)}$ defined by

$$\mathcal{L} = \begin{bmatrix} \mathcal{L}_{\rm gg} & \mathcal{L}_{\rm gl} \\ \mathcal{L}_{\rm lg} & \mathcal{L}_{\rm ll} \end{bmatrix}, \tag{S1}$$

where generators and load buses have been labeled so that the first n rows of $\mathcal{L}$ are associated with the generators and the last m rows of $\mathcal{L}$ correspond to the load buses. The dynamic model of the power network is

$$\begin{bmatrix} I & 0 & 0 \\ 0 & M_{\rm g} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\delta} \\ \dot{\omega} \\ \dot{\theta} \end{bmatrix} = - \begin{bmatrix} 0 & -I & 0 \\ \mathcal{L}_{\rm gg} & D_{\rm g} & \mathcal{L}_{\rm gl} \\ \mathcal{L}_{\rm lg} & 0 & \mathcal{L}_{\rm ll} \end{bmatrix} \begin{bmatrix} \delta \\ \omega \\ \theta \end{bmatrix} + \begin{bmatrix} 0 \\ P_\omega \\ P_\theta \end{bmatrix}, \tag{S2}$$

where $\delta : \mathbb{R} \to \mathbb{R}^n$ and $\omega : \mathbb{R} \to \mathbb{R}^n$ denote the generator rotor angles and frequencies, and $\theta : \mathbb{R} \to \mathbb{R}^m$ are the voltage angles at the buses. The matrices $M_{\rm g}$ and $D_{\rm g}$ are the diagonal matrices of the generator inertial and damping coefficients, and the inputs $P_\omega : \mathbb{R} \to \mathbb{R}^n$ and $P_\theta : \mathbb{R} \to \mathbb{R}^m$ are due to *known* changes in the mechanical input power to the generators or real power demand at the loads.

Consider the power network in Figure S1 subject to a load altering attack [20] at the buses b_4 and b_5. Due to this attack,

the angles θ_4 and θ_5 are altered by the attack signals $u_1 : \mathbb{R} \to \mathbb{R}$ and $u_2 : \mathbb{R} \to \mathbb{R}$, respectively. Suppose that a monitor measures directly the state variables of the first generator as $y_1 = \delta_1$ and $y_2 = \omega_1$. Let the system matrices be as in equations (S1) and (S2) with $M_g = \text{blkdiag}(.125, .034, .016)$, $D_g = \text{blkdiag}(.125, .068, .048)$, and

$$\mathcal{L} = \begin{bmatrix} .058 & 0 & 0 & -.058 & 0 & 0 & 0 & 0 & 0 \\ 0 & .063 & 0 & 0 & -.063 & 0 & 0 & 0 & 0 \\ 0 & 0 & .059 & 0 & 0 & -.059 & 0 & 0 & 0 \\ -.058 & 0 & 0 & .235 & 0 & 0 & -.085 & -.092 & 0 \\ 0 & -.063 & 0 & 0 & .296 & 0 & -.161 & 0 & -.072 \\ 0 & 0 & -.059 & 0 & 0 & .330 & 0 & -.170 & -.101 \\ 0 & 0 & 0 & -.085 & -.161 & 0 & .246 & 0 & 0 \\ 0 & 0 & 0 & -.092 & 0 & -.170 & 0 & .262 & 0 \\ 0 & 0 & 0 & 0 & -.072 & -.101 & 0 & 0 & .173 \end{bmatrix}.$$

Let U_1 and U_2 be the Laplace transform of the attack signals u_1 and u_2, and let

$$\begin{bmatrix} U_1(s) \\ U_2(s) \end{bmatrix} = \underbrace{\begin{bmatrix} \frac{-1.024s^4-5.121s^3-10.34s^2-9.584s-3.531}{s^4+5s^3+9.865s^2+9.173s+3.531} \\ 1 \end{bmatrix}}_{\mathcal{N}(s)} \bar{U}(s),$$

for *some arbitrary* nonzero signal $\bar{U}(s)$. The attack signals u_1 and u_2 are carefully chosen by the attacker to avoid detection by a monitor measuring the variables δ_1 and ω_1. In fact, it can be verified that $\mathcal{N}$ coincides with the null space of the transfer matrix between the attack at the buses b_4 and b_5 and the measurements δ_1, ω_1. From the analysis in Section II, the attack is undetectable by the monitor, because it does not affect the measurements. In Figure S2 we report the frequency of the network generators for a specific choice of $\bar{U}$. Notice that the second and the third generator are driven unstable by the attack input, yet the first generator does not deviate from the nominal operating condition. In other words, if the attack signals u_1 and u_2 are regarded as additional loads, then they are entirely sustained by the second and third generator.

We now apply the graph-theoretic analysis methods to analyze the above load altering attack. The directed graph describing the network Figure S1 is reported in Figure S3. Notice from Figure S3 that the maximum size of a linking from the attack vertices to the output vertices is 1, so that, by Theorem 2.3, the attack configuration admits generically undetectable attacks. In other words, for every choice of the numerical values of the network matrices, there exist nonzero attack modes u_1 and u_2 that are not detectable through the measurements y_1 and y_2.

□

Sidebar 2: ***Water network model and attack example.*** Mass transport networks are cyber-physical systems modeled by differential-algebraic equations. Examples include gas transmission and distribution networks [41], large-scale process engineering plants [42], and water networks. The vulnerability of water networks to cyber-physical attacks has been shown in [14], [26] for the case of open channel networks [43], and in [1], [27] for the case of municipal networks.

We focus on the hydraulics of a municipal water distribution network as described in [44], [45]. Water networks can be modeled as directed graphs with vertex set consisting of reservoirs, junctions, and storage tanks, and with edge set given by pipes, pumps, and valves that are used to convey water from source points to consumers. The key variables are the pressure

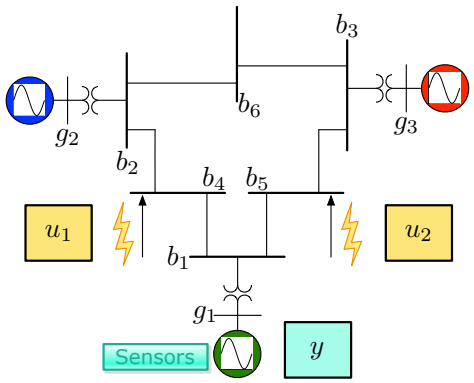

Fig. S1. This figure shows the WSSC power system with 3 generators and 6 buses. The attacker modifies the power injection at the buses b_4 and b_5 via a load altering attack. The monitor measures the rotor angle and frequency of the generator g_1. For some attack inputs u_1 and u_2, the attacker compromises the generators g_2 and g_3 while remaining undetected to the monitor.

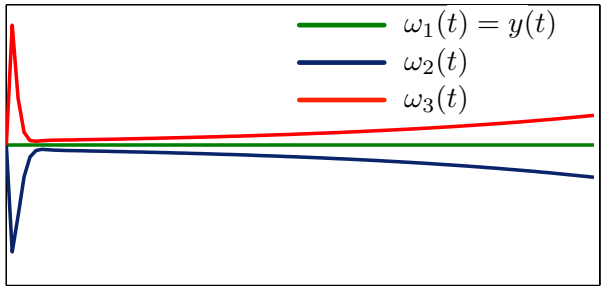

Fig. S2. In this figure we illustrate the effect of the attack discussed in Sidebar 1 "Power network model and attack example" on the generators frequencies. Notice that generators g_2 and g_3 are driven unstable by the attack, while generator g_1 is not affected by the attack.

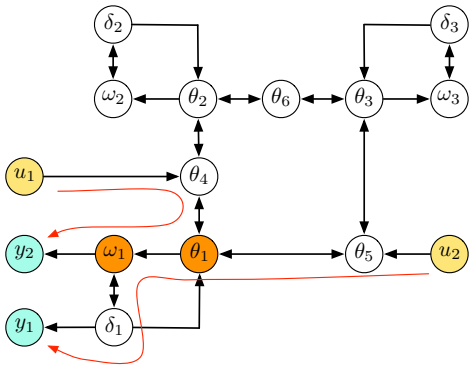

Fig. S3. This figure shows the digraph associated with the network in Figure S1 (the self-loops of the vertices $\{\delta_1, \delta_2, \delta_3\}$, $\{\omega_1, \omega_2, \omega_3\}$, and $\{\theta_1, \ldots, \theta_6\}$ are not drawn). The inputs u_1 and u_2 affect the buses b_4 and b_5, respectively. The measured variables are the rotor angle and frequency of the first generator. Notice that there is no linking of size 2 from the attack vertices to the output vertices. In fact, the vertices θ_1 and ω_1 belong to every path from $\{u_1, u_2\}$ to $\{y_1, y_2\}$. Two sample paths from the attack vertices to the output vertices are depicted in red.

head h_i at each network node i, and the flows Q_{ij} from node i to j. The hydraulic model governing the network dynamics includes constant reservoir heads, flow balance equations at junctions and tanks, and pressure difference equations along all edges:

$$\begin{aligned}
\text{reservoir } i: &\quad h_i = h_i{}^{\text{reservoir}} = \text{constant}\,,\\
\text{junction } i: &\quad d_i = \sum\nolimits_{j\to i} Q_{ji} - \sum\nolimits_{i\to k} Q_{ik}\,,\\
\text{tank } i: &\quad A_i\dot{h}_i = \sum\nolimits_{j\to i} Q_{ji} - \sum\nolimits_{i\to k} Q_{ik}\,,\\
\text{pipe } (i,j): &\quad Q_{ij} = Q_{ij}(h_i - h_j)\,,\\
\text{pump } (i,j): &\quad h_j - h_i = +\Delta h_{ij}{}^{\text{pump}} = \text{constant}\,,\\
\text{valve } (i,j): &\quad h_j - h_i = -\Delta h_{ij}{}^{\text{valve}} = \text{constant}\,.
\end{aligned} \tag{S3}$$

Here, d_i is the demand at junction i, A_i is the (constant) cross-sectional area of storage tank i, and the notation "$j \to i$" denotes the set of nodes j connected to node i. The flow Q_{ij} depends on the pressure drop $h_i - h_j$ along pipe according to the Hazen-Williams equation $Q_{ij}(h_i - h_j) = g_{ij}|h_i - h_j|^{1/1.85-1}\cdot(h_i - h_j)$, where $g_{ij} > 0$ is the pipe conductance.

Consider the water supply network EPANET 3 [46] linearized at steady state with non-zero pressure drops. The topology of the water network and the attack locations are illustrated in Figure S4. For notational convenience, let x_1, x_2, x_3, and x_4 denote, respectively, the pressure at the reservoir R_2, at the reservoir R_1 and at the tanks T_1, T_2 and T_3, at the junction P_2, and at the remaining junctions, respectively. The descriptor model for the EPANET 3 network reads as

$$\begin{bmatrix} \dot{x}_1(t) \\ M\dot{x}_2(t) \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & A_{22} & 0 & A_{24} \\ A_{31} & 0 & A_{33} & A_{34} \\ 0 & A_{42} & A_{43} & A_{44} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix},$$
$$y = \begin{bmatrix} C_1 & C_2 & C_3 & C_4 \end{bmatrix},$$

where the pattern of zeros is due to the network interconnection structure, and $M = \mathrm{diag}(1, A_1, A_2, A_3)$ corresponds to the dynamics of the reservoir R_1 and the tanks T_1, T_2, and T_3.

We consider the following attack where the attacker's intention is to steal water from the reservoir R_2. In order to remain undetected from the sensors measurements, the attacker simultaneously corrupts the measurements at sensor S_1 and modifies the pressure at pump P_2. Formally, the attack matrices read as $B = [B_1\ B_2\ 0]$ and $D = [0\ 0\ D_1]$, with

$$B_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}^{\mathsf{T}},\ B_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}^{\mathsf{T}},\ \text{and}$$
$$D_1 = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^{\mathsf{T}}.$$

Let the attack input be $u = [u_1^{\mathsf{T}}\ u_2^{\mathsf{T}}\ u_3^{\mathsf{T}}]^{\mathsf{T}}$, where $u_1 = -x_1$ (the attacker physically subtracts water from R_2), $u_2 = -A_{31}x_1$, and $u_3 = -x_1$. The dynamics of the EPANET 3 network under attack read as

$$\begin{bmatrix} \dot{x}_1 \\ M\dot{x}_2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & A_{22} & 0 & A_{24} \\ 0 & 0 & A_{33} & A_{34} \\ 0 & A_{42} & A_{43} & A_{44} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix},$$
$$y = \begin{bmatrix} 0 & C_2 & C_3 & C_4 \end{bmatrix} x,$$

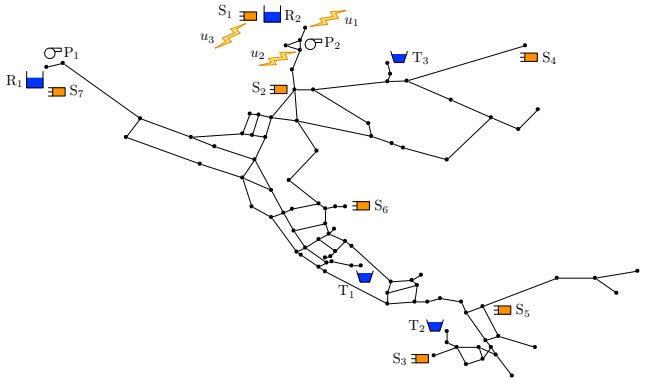

Fig. S4. This figure shows the structure of the EPANET water supply network model # 3, which features 3 tanks (T_1, T_2, T_3), 2 reservoirs (R_1, R_2), 2 pumps (P_1, P_2), 96 junctions, and 119 pipes. Seven pressure sensors ($\mathrm{S}_1, \dots, \mathrm{S}_7$) have been installed to monitor the network functionalities. A cyber-physical attack to steal water from the reservoir R_2 is reported. Notice that the cyber-physical attack features two state attacks (u_1, u_2) and one output attack (u_3).

Observe that the attacker subtracts water from R_2 while remaining undetected from the sensors measurements. In fact, the effect of the attack does not affect the measurements y, because the pressure change at the reservoir R_2 does not appear in the measurements y.

We conclude this section with the following remarks. First, the attack can be implemented with knowledge of the submatrix A_{31} only, without knowing the whole network structure and initial state. Second, the effectiveness of the proposed attack strategy is independent of the sensors measuring the variables x_3 and x_4. On the other hand, if additional sensors are used to measure the flow between the reservoir R_2 and the pump P_2, then the attacker would need to corrupt these measurements as well to remain undetected. Third and finally, due to the reliance on networks to control actuators in cyber-physical systems, the attack u_2 on the pump P_2 could be generated by a cyber attack as for the case of power grids [20].

Sidebar 3: ***Stealth, replay, covert, and injection attacks.*** Several attacks against cyber-physical systems have recently been identified and analyzed. These attacks are particular instances of the general framework introduced in Section I.

Stealth attack [11], [18]. In a stealth attack the attacker modifies some sensors readings by physically tampering with the individual meters or by getting access to some communication channels. Following the notation in Section I, stealth attacks are modeled by the exogenous input $(0, Du)$, with $\mathrm{Im}(D) \subseteq \mathrm{Im}(C)$ and u an arbitrary signal. Notice that stealth attacks modify only the measurements equation, so that the system dynamics become

$$E\dot{x} = Ax,$$
$$y = Cx + Du.$$

See Figure S5 for a block diagram representation of a stealth attack.

Stealth attacks have three important features. First, they can be cast without knowing or tampering with the system dynamics. Second, they are undetectable by bad data detectors (see Section I). To see this, notice that the residual of the bad

data detector $r = y - CC^{\dagger}y$ is identically zero for a stealth attack, because $\mathrm{Im}(D) \subseteq \mathrm{Im}(C)$ (the measurements y are compatible with the measurement matrix C). Third and finally, stealth attacks may be detectable by the monitor (6), because such monitor verifies the compatibility of the measurements with the system dynamics, and not only with the measurements equation. See [19] for a discussion of the detectability of attacks via static and dynamic monitors.

Replay attack [13]. In a replay attack the attacker performs three main actions. First, it records the system output corresponding to a nominal operating condition. Second, it modifies the sensors measurements to replicate the previously recorded measurements corresponding to a nominal operating condition. Third, it injects a control signal to disrupt the system functionality. Replay attacks can be modeled by the input $(Bu, -Cx + C\tilde{x})$, where x and $\tilde{x}$ are the state trajectories of the system under attack and without the attack, respectively. In other words, x and $\tilde{x}$ satisfy the differential equations

$$E\dot{x} = Ax + Bu,$$
$$E\dot{\tilde{x}} = A\tilde{x},$$

and $C\tilde{x}$ are the measurements corresponding to the system without attack. The system dynamics with replay attack read as

$$E\dot{x} = Ax + Bu,$$
$$y = C\tilde{x}.$$

See Figure S6 for a block diagram representation of a replay attack. Notice that replay attacks can be cast without knowing the system dynamics, provided that the attacker has access to all sensors. We refer the reader to [13] for a method to reveal replay attacks.

Covert attacks [14]. Covert attacks are closed-loop replay attacks, where the attacker modifies the system measurements to cancel out only the effect of its attack on the system dynamics. In particular, the covert attack input is $(Bu, -C\tilde{x})$, where $\tilde{x}$ is the state trajectory due to the attack input. The system dynamics with covert attacks read as

$$E\dot{x} = Ax + Bu,$$
$$y = C(x - \tilde{x}),$$

where $\tilde{x}$ satisfies

$$E\dot{\tilde{x}} = A\tilde{x} + Bu, \text{ with } \tilde{x}(0) = 0.$$

See Figure S7 for a block diagram representation of a covert attack. Notice that covert attacks require the attacker to know the exact system dynamics, and to hijack some sensors (only those sensors affected by the state attack). On the other hand, covert attacks are undetectable by static and dynamic monitors [19], and by the active method proposed in [13]. In fact, covert attacks excite only the zero dynamics of the attack/measurements dynamical system, and they are therefore undetectable as discussed in Section II.

Dynamic false-data injection attacks [47]. Dynamic false-data injection attacks can be cast against systems with unstable modes, and they aim at modifying the system measurements

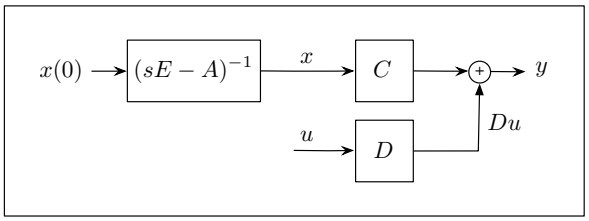

Fig. S5. Block diagram of a stealth attack. The attacker corrupts the measurements y with the signal $Du \in \mathrm{Im}(C)$. Notice that stealth attacks can be cast by tampering with the sensors measurements, and without knowing the system dynamics. In fact, stealth attacks do not alter the system dynamics.

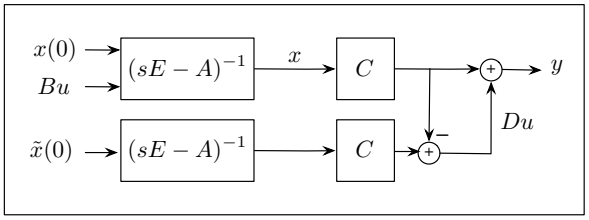

Fig. S6. Block diagram of a replay attack. The attacker corrupts the system dynamics and the measurements. In particular, the attacker resets the measurements to reflect a pre-recorded nominal operating condition $\tilde{x}(0)$, and to hide the effect of the state attack on the system dynamics. Replay attacks can be cast with access to all sensors, and without knowing the system dynamics.

to render unobservable some unstable modes. The attack input for a dynamic false-data injection attack is $(0, -C\tilde{x})$, where

$$E\dot{\tilde{x}} = A\tilde{x},$$

and $\tilde{x}(0)$ is the projection of the system state $x(0)$ along the eigenvector of an unstable mode. Dynamic false data injection attacks are illustrated in Figure S8. As for the case of covert attack, dynamic false-data injection attacks are undetectable as in Definition 1.

□

Sidebar 4: ***Geometric control theory and its application to fault detection and isolation.*** The *geometric approach* to control of dynamical systems aims to develop a set of tools and techniques based on geometric notions and operations, such as subspaces, sets, linear transformation, direct sum, and orthogonalization, to analyze and control dynamical systems. The geometric approach has been developed over the last

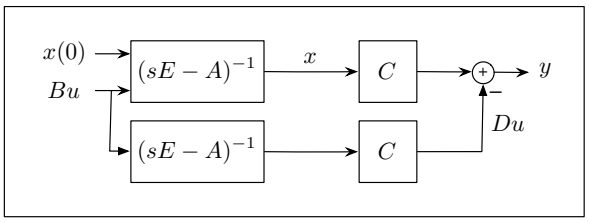

Fig. S7. Block diagram of a covert attack. The attacker corrupts the system dynamics and the measurements. In particular, the attacker modifies the measurements to cancel out the effect of its attack on the system dynamics. Covert attacks are closed-loop replay attacks, and they require access to some sensors and knowledge of the system dynamics to be implemented.

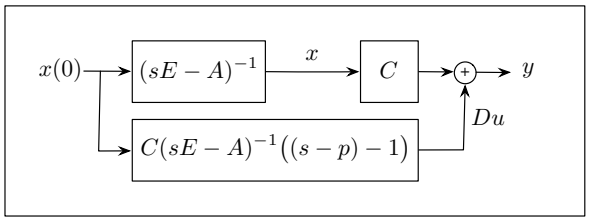

Fig. S8. Block diagram of a dynamic false-data injection attack. The attacker corrupts the system dynamics and measurements to render the unstable mode p unobservable from the measurements. Dynamic false-data injection attacks require access to some sensors and knowledge of the system dynamics to be implemented.

decades, and it has found applicability in many classic control problems. We refer the interested reader to [33], [34], [48] for a comprehensive treatment of the geometric approach to control of linear dynamical systems. We now review some basic concepts and applications of the geometric approach.

We start with the notions of *controlled* and *conditioned* invariant subspaces. For the ease of notation, consider the system

$$\begin{aligned}\dot{x} &= Ax + Bu,\\ y &= Cx,\end{aligned}$$

where A, B, and C are constant matrices of appropriate dimensions. A subspace $\mathcal{V} \subseteq \mathbb{R}^n$ is an $(A, \mathrm{Im}(B))$-controlled invariant subspace [33, Chapter 4] if

$$A\mathcal{V} \subseteq \mathcal{V} + \mathrm{Im}(B),$$

or, equivalently, if there exists a matrix F such that

$$(A + BF)\mathcal{V} \subseteq \mathcal{V}.$$

The notion of controlled invariant subspace refers to the possibility of confining the state trajectory of the system (A, B, C) within a subspace. Specifically, a subspace $\mathcal{V} \subseteq \mathbb{R}^{n\times n}$ is an $(A, \mathrm{Im}(B))$-controlled invariant if, for every initial state $x_0 \in \mathcal{V}$, there exists a control input u such that the state $x \in \mathcal{V}$ at all times $t \in \mathbb{R}_{\geq 0}$. For instance, the controllability subspace $\mathrm{Im}([B\ AB\ \cdots A^{n-1}B])$ is an $(A, \mathrm{Im}(B))$-controlled invariant subspace. The set of controlled invariant subspaces contained in a subspace $\mathcal{E} \subseteq \mathbb{R}^{n\times n}$ admits a supremum $\mathcal{V}^*$, that is, there exists an $(A, \mathrm{Im}(B))$-controlled invariant subspace satisfying $\mathcal{V} \subseteq \mathcal{V}^* \subseteq \mathcal{E}$, for any $(A, \mathrm{Im}(B))$-controlled invariant subspace $\mathcal{V}$. If $\mathcal{E} = \mathrm{Ker}(C)$, then the subspace $\mathcal{V}^*$ contains all the state trajectories driven by the input u and resulting in the output y being identically zero.

Controlled invariant subspaces are dual to conditioned invariant subspaces. A subspace $\mathcal{S} \subseteq \mathbb{R}^n$ is an $(A, \mathrm{Ker}(C))$-conditioned invariant subspace [33, Chapter 4] if

$$A(\mathcal{S} \cap \mathrm{Ker}(C)) \subseteq \mathcal{S},$$

or, equivalently, if there exists a matrix G such that

$$(A + GC)\mathcal{S} \subseteq \mathcal{S}.$$

Condition invariant subspaces arise in the context of state estimation. Specifically, the subspace $\mathcal{S}$ is an $(A, \mathcal{C})$-conditioned invariant if it is possible to estimate the trajectory $x \setminus \mathcal{S}$ by processing the initial condition $x_0 \setminus \mathcal{S}$, the input u and the measurements y through an observer [48, Chapter 5]. For instance, the unobservability subspace $\mathrm{Ker}([C^{\mathsf{T}}\ A^{\mathsf{T}}C^{\mathsf{T}} \cdots (A^{n-1})^{\mathsf{T}}C^{\mathsf{T}}]^{\mathsf{T}})$ is an $(A, \mathrm{Ker}(C))$-conditioned invariant subspace. The set of conditioned invariant subspaces containing $\mathcal{E} \subseteq \mathbb{R}^{n\times n}$ admits an infimum $\mathcal{S}^*$, that is, an $(A, \mathrm{Ker}(C))$-conditioned invariant subspace satisfying $\mathcal{E} \subseteq \mathcal{S}^* \subseteq \mathcal{S}$, for any $(A, \mathrm{Ker}(C))$-conditioned invariant subspace $\mathcal{S}$. If $\mathcal{E} = \mathrm{Im}(B)$, then the subspace $\mathcal{S}^*$ defines the largest subspace of the state space that can be estimated in the presence of an unknown input signal u. Controlled and conditioned invariant subspaces can be extended to systems with direct feedthrough matrix, and to singular systems; see [33], [49].

Several problems, including disturbance decoupling, non interacting control, fault detection and isolation, and state estimation in the presence of unknown inputs, have been addressed and solved in the geometric framework. In the classical Fault Detection and Isolation (FDI) setup, the presence of sensor failures and actuator malfunctions is modeled by adding unknown and unmeasurable inputs $f_i : \mathbb{R}_{\geq 0} \to \mathbb{R}^{m_i}$ to the nominal system. The dynamical system with failures reads as

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Bu \sum_{i=1}^{k} B_i f_i(t),\\ y(t) &= Cx(t),\end{aligned}$$

where $k \in \mathbb{N}$ denotes the number of actuators and sensors failures, and $B_i \in \mathbb{R}^{n\times m_i}$, $i \in \{1, \ldots, k\}$, are known matrices reflecting the failures input directions. The FDI problem is to design a set of *residual generators* of the form

$$\begin{aligned}\dot{w}_i(t) &= F_i w_i(t) + E_i y(t),\\ r_i(t) &= M_i w(t) + H_i y(t),\end{aligned} \tag{S4}$$

to detect and identify failures. The residual generator processes the observables y and the known input u to generate a residual vector r_i that allows to uniquely identify if f_i becomes nonzero, that is, if the failure i occurred in the system. As a result of [8], [33], the i-th failure can be correctly identified if and only if

$$\mathrm{Im}(B_i) \cap (\mathcal{V}^*_{K\setminus\{i\}} + \mathcal{S}^*_{K\setminus\{i\}}) = \emptyset,$$

where $\mathcal{V}^*_{K\setminus\{i\}}$ and $\mathcal{S}^*_{K\setminus\{i\}}$ are the maximal controlled and minimal conditioned invariant subspaces associated with the system $(A, [B_1 \cdots B_{i-1}\ \ B_{i+1} \cdots B_k], C)$. We refer the reader to [8], [33] for a procedure to design residual generators. □

Sidebar 5: ***Invariant zeros and zero dynamics.*** The concept of zero of a dynamical system plays an important role in several control problems, and it refers to the possibility of having a nonzero state trajectory while the system output is identically zero. To be specific, consider the system (E, A, B, C, D), where $E \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$. Assume that

$$\mathrm{Rank}\left(\begin{bmatrix} B \\ D \end{bmatrix}\right) = m,$$

where Rank$(\cdot)$ denotes the rank of a matrix. Define the Rosenbrock matrix associated with the system (E, A, B, C, D) by

$$P(s) = \begin{bmatrix} sE - A & -B \\ C & D \end{bmatrix}.$$

The invariant zeros of (E, A, B, C, D) are the complex values $s \in \mathbb{C}$ satisfying

$$\text{Rank}(P(s)) < n + m.$$

Let z be an invariant zero, and let x_0 and u_0 be such that

$$\begin{aligned} (sE - A)x_0 - Bu_0 &= 0, \\ Cx_0 + Du_0 &= 0. \end{aligned}$$

The vectors x_0 and u_0 are referred to as *state-zero direction* and *input-zero direction*, and they can be used to excite the system (E, A, B, C, D) in a way that the state trajectory is nonzero while the output is identically zero. To see this, let $E = I$, and let the system initial state $x(0)$ and input u be x_0 and $t \to e^{zt}u_0$, respectively. Notice that the state trajectory x is $t \to \exp^{zt} x_0$, because it is the unique solution to the differential equation $\dot{x} = Ax + Bu$. Finally, observe that the system output is identically zero at all times $t \in \mathbb{R}_{\geq 0}$, because

$$y(t) = C \exp^{zt} x_0 + De^{zt}u_0 = e^{zt}\left(Cx_0 + Du_0\right) = 0.$$

The state trajectory x is called *zero dynamics*. Given the relationship between zero dynamics and invariant zeros, it can be shown that a system exhibits zero dynamics if and only if it features invariant zeros. Notice that the number of invariant zeros can be infinite. A system with a finite number of invariant zeros is called *left-invertible*, and it satisfies $y(0, u_1, t) \neq y(0, u_2, t)$ for some times $t \in \mathbb{R}$, and for all inputs u_1 and u_2. Likewise, a system that fails to be left-invertible can be characterized in the Laplace domain by a rank-deficient transfer matrix, as illustrated in the example in Sidebar 1 "Power network model and attack example". □

Sidebar 6: ***Graph theory and generic properties.*** The *graph theoretic* approach to control of dynamical systems aims to express system properties via properties of a properly defined graph [35], [36]. In order to highlight connections between dynamical systems and graphs, we recall necessary notions and definitions in graph theory and algebraic geometry.

A directed graph $G = (\mathcal{V}_G, \mathcal{E}_G)$ consists of a set of vertices $\mathcal{V}_G$ and a set of directed edges $\mathcal{E}_G \subseteq \mathcal{V}_G \times \mathcal{V}_G$. An edge $(v, w) \in \mathcal{E}_G$ is directed from vertex v to vertex w. A subgraph of a graph $G = (\mathcal{V}_G, \mathcal{E}_G)$ is a graph $H = (\mathcal{V}_H, \mathcal{E}_H)$ such that $\mathcal{V}_H \subseteq \mathcal{V}_G$ and $\mathcal{E}_H \subseteq \mathcal{E}_G$. A graph is undirected if $(v, w) \in \mathcal{E}_G$ implies that $(w, v) \in \mathcal{E}_G$, and in this case we write $\{v, w\} \in \mathcal{E}_G$. A path in G is a subgraph $P = (\{v_1, \ldots, v_{k+1}\}, \{e_1, \ldots, e_k\})$ such that $v_i \neq v_j$ for all $i \neq j$, and $e_i = (v_i, v_{i+1})$ for each $i \in \{1, \ldots, k\}$. A set of ℓ mutually disjoint paths between two sets of vertices S_1 and S_2 is called *linking* of size ℓ from S_1 to S_2.

For a system (E, A, B, C, D), let d be the number of its nonzero entries. A system $(\bar{E}, \bar{A}, \bar{B}, \bar{C}, \bar{D})$ is compatible with (E, A, B, C, D) if the two systems have the same pattern of nonzero entries. By collecting the nonzero parameters into a vector, every system compatible with (E, A, B, C, D) can be represented by a point in the Euclidean space $\mathbb{R}^d$. A property which can be asserted on a dynamical system is called *generic* if, informally, it holds for *almost all* compatible systems. To be more precise, a property is generic if and only if the set of compatible systems satisfying such property forms a dense subset of the parameters space. For instance, controllability, observability, and left-invertibility of a dynamical system are generic properties with respect to the parameters space $\mathbb{R}^d$ [35]. We refer the interested reader to [34], [36] for a comprehensive discussion of structured systems and generic properties.

□

PLACE PHOTO HERE

Fabio Pasqualetti (S'07) is an Assistant Professor in the Department of Mechanical Engineering at the University of California, Riverside. He received a Doctor of Philosophy degree in Mechanical Engineering from the University of California, Santa Barbara, in 2012, a Laurea Magistrale degree "summa cum laude" (M.Sc. equivalent) in Automation Engineering from the University of Pisa, Pisa, Italy, in 2007, and a Laurea degree "summa cum laude" (B.Sc. equivalent) in Computer Engineering from the University of Pisa, Pisa, Italy, in 2004.

His main research interest is in secure control systems, with application to multi-agent networks, distributed computing and power networks. Other interests include vehicle routing and combinatorial optimization, with application to distributed patrolling and camera surveillance.

PLACE PHOTO HERE

Florian Dörfler (S'09) is an Assistant Professor in the Department of Electrical Engineering at the University of California Los Angeles, and he is affiliated with the Center for Nonlinear Studies at the Los Alamos National Laboratories. He received a Ph.D. degree in Mechanical Engineering from the University of California at Santa Barbara in 2013, and a Diplom degree in Engineering Cybernetics from the University of Stuttgart in 2008. His primary research interests are centered around distributed control, complex networks, and cyber-physical systems with applications to smart power grids and robotic coordination. He is recipient of the 2009 Regents Special International Fellowship, the 2011 Peter J. Frenkel Foundation Fellowship, the 2010 ACC Student Best Paper Award, and the 2011 O. Hugo Schuck Best Paper Award. As a co-advisor and a co-author, he has been a finalist for the ECC 2013 Best Student Paper Award.

Francesco Bullo (S'95–M'99–SM'03–F'10) is a Professor with the Mechanical Engineering Department at the University of California, Santa Barbara. He received the Laurea degree "summa cum laude" in Electrical Engineering from the University of Padova, Italy, in 1994, and the Ph.D. degree in Control and Dynamical Systems from the California Institute of Technology in 1999. From 1998 to 2004, he was an Assistant Professor with the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign.

His main research interest is multi-agent networks with application to robotic coordination, distributed computing and power networks. Other interests include vehicle routing, geometric control, and motion planning problems. He has published more than 200 papers in international journals, books and refereed conferences. He is the coauthor, with Andrew D. Lewis, of the book "Geometric Control of Mechanical Systems" (Springer, 2004) and, with Jorge Cortés and Sonia Martínez, of the book "Distributed Control of Robotic Networks" (Princeton, 2009). His students' papers were finalists for the Best Student Paper Award at the IEEE Conference on Decision and Control (2002, 2005, 2007), and the American Control Conference (2005, 2006, 2010). He is an IEEE Fellow and has served on the Editorial Boards of the "IEEE Transactions on Automatic Control," the "ESAIM: Control, Optimization, and the Calculus of Variations" and the "SIAM Journal of Control and Optimization."