Subspace Methods for Data Attack on State Estimation: A Data Driven Approach

Jinsub Kim, *Member, IEEE*, Lang Tong, *Fellow, IEEE*, and Robert J. Thomas, *Life Fellow, IEEE*

***Abstract*—Data attacks on state estimation modify part of system measurements such that the tempered measurements cause incorrect system state estimates. Attack techniques proposed in the literature often require detailed knowledge of system parameters. Such information is difficult to acquire in practice. The subspace methods presented in this paper, on the other hand, learn the system operating subspace from measurements and launch attacks accordingly. Conditions for the existence of an unobservable subspace attack are obtained under the full and partial measurement models. Using the estimated system subspace, two attack strategies are presented. The first strategy aims to affect the system state directly by hiding the attack vector in the system subspace. The second strategy misleads the bad data detection mechanism so that data not under attack are removed. Performance of these attacks are evaluated using the IEEE 14-bus network and the IEEE 118-bus network.**

***Index Terms*—Cyber physical system, data framing attack, false data injection, state estimation, subspace method.**

I. Introduction

A cyber physical system (CPS) [1] is a collection of physical devices networked by a cyber infrastructure with integrated sensing, communications, and control. A defining feature of CPS is coordinated operations based on data collected from sensors deployed throughout the system. Major examples of CPS include power grids, intelligent transportation systems, and networked robotics.

An essential signal processing component of many CPSs is real-time state estimation based on sensor measurements [2]. The state estimate provides a CPS with the real-time monitoring and control capability. For instance, the state estimate of a power grid facilitates real-time economic dispatch, contingency analysis, and computation of real-time electricity price [2].

Manuscript received May 08, 2014; revised September 14, 2014; accepted November 29, 2014. Date of publication December 23, 2014; date of current version January 30, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Rui Zhang. This work was supported in part by the National Science Foundation under Grant CNS-1135844. Part of this work was presented at the Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, November 2013.

J. Kim is with the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331 USA (e-mail: jinsub.kim@oregonstate.edu).

L. Tong and R. J. Thomas are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853 USA (e-mail: ltong@ece.cornell.edu; rjt1@cornell.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2014.2385670

The dependency of CPS on data communications makes it vulnerable to cyber attacks where an adversary may break into the network, collect unauthorized information, and intercept and alter sensor data. Because measurements are collected over a wide geographical area by distributed data acquisition systems, sometimes through wireless links, communications networks that support modern CPSs have numerous points of vulnerabilities [3], [4]. For critical infrastructures such as a power grid, a well planned coordinated attack may lead to a cascading failure and a regional blackout [5].

To assess vulnerability of CPS to possible cyber attacks, it is important to study potential attack mechanisms. In this paper, we consider an adversary who can modify certain sensor data such that the corrupted data will mislead the CPS control with a wrong state estimate. We refer to such a data attack on state estimation as a *state attack*. A major challenge of state attack is to avoid being detected and identified by the fusion center.

In the literature, successful state attacks on a CPS, in particular a power grid, have been reported. Liu, Ning, and Reiter [6] presented the first state attack strategy, where an adversary replaces part of "normal" sensor data with "malicious data." They showed that if an adversary can control a sufficiently large number of sensor data, it can perturb the state estimate by an arbitrary degree while avoiding detection at the control center. Subsequent works along this line uncovered numerous attack and protection mechanisms [7]–[14].

Most proposed attack schemes require considerably detailed system information. In particular, the network topology and physical system parameters are often required to construct attacks. Although such information may be obtained by penetrating the control center, security measures can make it difficult in practice to access such information.

A. Summary of Contributions

We consider the problem of data-driven attacks on state estimation, assuming that the adversary is capable of monitoring a subset of system measurements without detailed knowledge of the network topology and system parameters. The key idea in the proposed approach is to exploit the subspace structure of the measurements, in the same spirit of subspace techniques in array processing [15], beamforming [16], and system identification [17].

The main contribution of this paper is the development of subspace techniques for designing a state attack. To this end, we present two techniques with different characteristics. First, we show a construction of an unobservable attack based on the estimated subspace structure of measurements. We show further

that, in constructing the attack, under certain conditions, monitoring only partial measurements may be sufficient. In particular, we present a graph theoretic condition for the existence of an unobservable attack under the partial measurement model.

The second subspace-based attack exploits current bad data detection and removal mechanisms. In particular, the attack purposely triggers the bad data detection mechanism, but it is designed to mislead the fusion center to remove data that have not been tampered by the adversary while retaining some of the falsified data. After such data removal, although the remaining data appear to be consistent with the system model, the resulting state estimate may have an arbitrarily large error. We refer to this type of attack as *data framing attack* in the sense that valid data are "framed" by the adversary and removed incorrectly by the fusion center.

To demonstrate the effectiveness of these attacks, we consider the problem of state estimation in a power system as a practical example of CPS. To this end, we consider the IEEE 14-bus network and the IEEE 118-bus network [18].

An additional complexity of the power system is that the system observation is a nonlinear function of the system state. This raises the issue of whether attacks constructed from a linear model is effective in a nonlinear system. While we do not have theoretical guarantees, simulation results show that the subspace-based data attacks perform well in the presence of the nonlinearities in system equations.

B. Related Work and Organization

This paper extends some of the key results on state attacks that assume that the system parameters and the network topology are known to the attacker. We describe below some of the relevant techniques.

There is a substantial literature on state attacks when the system parameter and the network topology are *known*. As mentioned before, Liu, Ning, and Reiter [6] first introduced an *unobservable attack* on power system state estimation, which can perturb the state estimate without being detected by the bad data detector at the fusion center. Following their seminal work, the link between feasibility of an unobservable attack and power system observability was made in [8], [7], [19]. Consequently, classical power system observability conditions [20] can be modified to check feasibility of unobservable attacks and used to develop countermeasures based on sensor data authentication [7]–[10], [12], [19], [21], [22]. To assess the grid vulnerability against data attacks, the minimum number of adversary-controlled sensors necessary for an unobservable attack was suggested as the *security index* of the grid [8], [23]. The data framing attack, when the system parameters are known, was first proposed in [24] to circumvent the fundamental limit imposed by the security index.

There is limited work on state attacks without system information or with partial system information. The use of independent component analysis in [13] is the most relevant. The authors of [13] proposed to identify a mixing matrix from which to construct an unobservable attack. However, such techniques require that loads are statistically independent and non-Gaussian, and the techniques need full sensor observations. Generating unobservable attacks using partial parameter information was considered in [14]. The authors in [14] showed that an adversary knowing impedance of transmission lines in a cutset of the network topology can construct an unobservable attack. However, how an adversary can learn local parameters is nontrivial. In contrast to the aforementioned approaches, our method requires no system parameter information, and it can be launched with only partial sensor observations. Furthermore, we identify the conditions under which an attacker with partial sensor observations (without other system information) may construct an unobservable attack or a data framing attack. In contrast to the feasibility conditions given in existing works in the literature [7], [8], [19], where an omniscient adversary is assumed, our conditions guarantee a successful attack design for an adversary with limited knowledge and limited access to the system.

Attacks were also studied in the framework of a general dynamic CPS, under the assumption of an omniscient adversary. For instance, an attack on a linear control system equipped with a linear-quadratic-Gaussian controller was studied in [25]. Detectability and identifiability of attacks on general CPS operations was characterized in [26]. The model considered in these papers is more general than the static model studied here. However, their assumption of an adversary with complete system information is stronger than that in the present work.

The rest of this paper is organized as follows. Section II introduces the measurement model, the mathematical model of state estimation and bad data processing, and the attack model. Section III presents the subspace methods of unobservable attack, and Section IV presents the subspace methods of data framing attack. In Section V, the results from simulations with benchmark power grids are presented. Finally, Section VI provides concluding remarks.

II. Mathematical Models

A. Notations

An upper case boldface letter (e.g., $\mathbf{H}$) denotes a matrix, a lower case boldface letter (e.g., $\mathbf{x}$) denotes a vector, and a script letter (e.g., $\mathcal{A}$, $\mathcal{S}$) denotes a set. The entry of $\mathbf{H}$ at the ith row and the jth column is denoted by $\mathbf{H}_{ij}$, and the ith entry of $\mathbf{x}$ is denoted by x_i. In addition, $\mathcal{R}(\mathbf{H})$ and $\mathcal{N}(\mathbf{H})$ denote the column space and the null space of $\mathbf{H}$ respectively. And, $\mathbf{I}$ denotes an identity matrix with an appropriate size.

B. Measurement Model

The *system state* of a CPS is defined as a vector of variables that characterize the current operating condition of the CPS. We assume centralized state estimation at the fusion center. For real-time estimation of the system state $\mathbf{x} \in \mathbb{R}^n$, the fusion center collects measurements from sensors deployed throughout the system. Generally, the sensor measurements are related to the system state $\mathbf{x}$ in a nonlinear fashion, and the relation can be described by the nonlinear measurement model (e.g., the AC model for a power grid [27]):

$$\mathbf{z} = h(\mathbf{x}) + \mathbf{e}, \tag{1}$$

where $\mathbf{z} \in \mathbb{R}^m$ is the measurement vector, $h(\cdot)$ is the measurement function, and $\mathbf{e}$ is the Gaussian measurement noise.

If some sensors malfunction or an adversary injects malicious data, the fusion center observes biased measurements,

$$\bar{\mathbf{z}} = h(\mathbf{x}) + \mathbf{e} + \mathbf{a}, \tag{2}$$

where $\mathbf{a}$ represents a deterministic bias. In such a case, the data are said to be *bad*, and the biased sensor entries are referred to as *bad data entries*. The bad data vector is typically sparse, and its support is unknown to the fusion center. If $\mathbf{a}$ is injected by an adversary, $\mathbf{a}$ is constrained by its support.

In analyzing the attack effect on state estimation, we adopt a linearization of (1) around a nominal state $\mathbf{x}_0$:

$$\mathbf{z} = h(\mathbf{x}_0) + \mathbf{H}(\mathbf{x} - \mathbf{x}_0) + \mathbf{e}, \tag{3}$$

where $\mathbf{H} \in \mathbb{R}^{m \times n}$ is the measurement matrix that relates the system state to the measurement vector, and $\mathbf{e}$ is the Gaussian measurement noise with a covariance matrix $\sigma^2\mathbf{I}$. Without loss of generality, we assume that both $h(\mathbf{x}_0)$ and $\mathbf{x}_0$ are zero vectors[1] and employ the following model:

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}. \tag{4}$$

A system is said to be *observable* if the measurement matrix $\mathbf{H}$ has full column rank (i.e., $\mathbf{x}$ can be uniquely determined from $\mathbf{H}\mathbf{x}$.) System observability is essential for state estimation. In practice, sensors should be placed in the network to satisfy observability. Hence, we assume that the CPS of interest is observable, i.e., $\mathbf{H}$ has full column rank.

In practice, the nonlinear system and the nonlinear iterative state estimation techniques have a certain mitigating effect on attacks designed based on a linear model [28]. It is therefore important to validate performance of an attack strategy based on the nonlinear model (1) using a nonlinear state estimator. Note that, while our attacks are constructed based on (4), our numerical experiments validate their performance using the original nonlinear system (1) with a nonlinear state estimator.

C. State Estimation and Bad Data Processing

This section introduces a popular approach to state estimation and bad data processing [27], [29], which we assume to be employed by the fusion center. The specific approach is a widely used standard implementation in the power grid where the number of states is in the order of 10,000, and the estimates are made every few minutes.

Fig. 1 illustrates an iterative scheme for obtaining an estimate $\hat{\mathbf{x}}$ of the system state, which consists of three functional blocks: state estimation, bad data detection, and bad data identification.

The assumed state estimator is based on the maximum likelihood principle and is implemented in a recursive manner. Iterations begin with the initial measurement vector $\mathbf{z}^{(1)} \triangleq \mathbf{z}$ and the initial measurement function $h^{(1)} \triangleq h$ where the superscript denotes the index for the current iteration.

[1]For general cases, we can simply treat $\mathbf{z}_1 \triangleq \mathbf{z} - h(\mathbf{x}_0)$ and $\mathbf{x}_1 \triangleq \mathbf{x} - \mathbf{x}_0$ as the measurement vector and the state vector and work with $\mathbf{z}_1 = \mathbf{H}\mathbf{x}_1 + \mathbf{e}$.

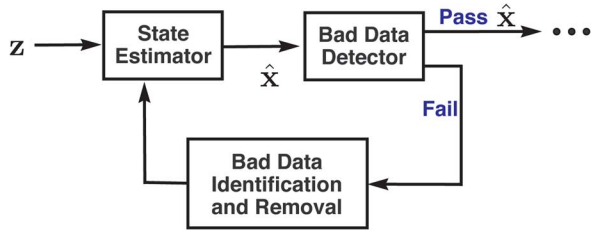


Fig. 1. State estimation and bad data processing.

In the kth iteration, state estimation uses $(\mathbf{z}^{(k)}, h^{(k)})$ as an input and calculates the least squares (LS) estimate of the system state and the corresponding residue vector:

$$\begin{aligned} \hat{\mathbf{x}}^{(k)} &\triangleq \arg\min_{\mathbf{x}} \frac{1}{\sigma^2} \left\| \mathbf{z}^{(k)} - h^{(k)}(\mathbf{x}) \right\|_2^2, \\ \mathbf{r}^{(k)} &\triangleq \mathbf{z}^{(k)} - h^{(k)}\left(\hat{\mathbf{x}}^{(k)}\right), \end{aligned} \tag{5}$$

where $\|\cdot\|_2$ denotes l_2 norm. In practice, the above nonlinear LS estimate can be obtained by iteration of a linearized LS estimation using Newton-Raphson or quasi-Newton methods [27].

Bad data detection employs the $J(\hat{\mathbf{x}})$-test [27], [29]:

$$\begin{cases} \text{bad data} & \text{if } \frac{1}{\sigma^2}\left\|\mathbf{r}^{(k)}\right\|_2^2 > \tau^{(k)}; \\ \text{good data} & \text{if } \frac{1}{\sigma^2}\left\|\mathbf{r}^{(k)}\right\|_2^2 \le \tau^{(k)} \end{cases} \tag{6}$$

where $\tau^{(k)}$ is a predetermined threshold. The $J(\hat{\mathbf{x}})$-test is widely used due to its simplicity and the fact that the test statistic has a χ^2 distribution if the data are good [29]. The latter fact is used to set the threshold $\tau^{(k)}$ for a given false alarm constraint.

If the bad data detector (6) declares that the data are good, the algorithm returns the state estimate $\hat{\mathbf{x}}^{(k)}$ and *terminates*. However, if the bad data detector declares that the data are bad, bad data identification is invoked to identify and remove *one* bad data entry from the measurement vector.

A widely used criterion for identifying a bad data entry is the normalized residue [27], [29]: each $r_i^{(k)}$ is divided by its standard deviation under the hypothesis that $\mathbf{z}^{(k)}$ contains no bad data. Therefore, each normalized residue approximately follows the standard normal distribution if $\mathbf{z}^{(k)}$ contains no bad data. Specifically,

$$\tilde{\mathbf{r}}^{(k)} \triangleq \mathbf{\Omega}^{(k)}\mathbf{r}^{(k)}, \tag{7}$$

where $\mathbf{\Omega}^{(k)}$ is a diagonal matrix with

$$\mathbf{\Omega}_{ii}^{(k)} \triangleq \begin{cases} 0 & \text{if removing } i \text{ makes} \\ & \text{the system unobservable;} \\ \frac{1}{\sqrt{\sigma^2 \mathbf{W}_{ii}^{(k)}}} & \text{otherwise;} \end{cases} \tag{8}$$

[2]and $\mathbf{W}^{(k)}$ is defined as

$$\mathbf{I} - \mathbf{H}^{(k)}\left(\left(\mathbf{H}^{(k)}\right)^T \mathbf{H}^{(k)}\right)^{-1}\left(\mathbf{H}^{(k)}\right)^T \tag{9}$$

[2]If removing the sensor i makes the system unobservable, its residue is always equal to zero [27], and the corresponding diagonal entry of $\mathbf{W}^{(k)}$ is zero. For such a sensor, the normalizing factor is 0 such that its normalized residue is equal to 0.

with $\mathbf{H}^{(k)}$ denoting the Jacobian of $h^{(k)}$ at $\hat{\mathbf{x}}^{(k)}$ (see Appendix of [29] for details.)

Once the normalized residue $\tilde{\mathbf{r}}^{(k)}$ is calculated, the sensor with the largest $|\tilde{r}_i^{(k)}|$ is identified as a bad sensor. The row of $\mathbf{z}^{(k)}$ and the row of $h^{(k)}$ that correspond to the bad sensor are removed, and the updated measurement vector $\mathbf{z}^{(k+1)}$ and measurement function $h^{(k+1)}$ are used as the inputs for the next iteration.

Using the linearized model (4), every step is the same as using the nonlinear model, except that the nonlinear measurement function $h^{(k)}(\mathbf{x})$ is replaced with the linear function $\mathbf{H}^{(k)}\mathbf{x}$ (so, the Jacobian is the same everywhere.) Note that the LS state estimate (5) is replaced with a simple linear LS solution:

$$\hat{\mathbf{x}}^{(k)} = \left(\left(\mathbf{H}^{(k)}\right)^T \mathbf{H}^{(k)}\right)^{-1} \left(\mathbf{H}^{(k)}\right)^T \mathbf{z}^{(k)}, \qquad (10)$$

and thus

$$\mathbf{r}^{(k)} = \mathbf{z}^{(k)} - \mathbf{H}^{(k)}\hat{\mathbf{x}}^{(k)} = \mathbf{W}^{(k)}\mathbf{z}^{(k)}. \qquad (11)$$

D. Adversary Model

An adversary is assumed to be capable of modifying the data from a subset of sensors $\mathcal{S}_{\mathrm{A}}$, referred to as *adversary sensors*. The fusion center observes corrupted measurements $\bar{\mathbf{z}}$ instead of the real measurements $\mathbf{z}$. The adversarial modification is mathematically modeled by:

$$\bar{\mathbf{z}} = \mathbf{z} + \mathbf{a}, \quad \mathbf{a} \in \mathcal{A}, \qquad (12)$$

where $\mathbf{a}$ is an attack vector, and $\mathcal{A}$ is the set of feasible attack vectors defined as

$$\mathcal{A} \triangleq \{\mathbf{a} \in \mathbb{R}^m : \ a_i = 0, \ \forall i \notin \mathcal{S}_{\mathrm{A}}\}. \qquad (13)$$

Liu, Ning, and Reiter [6] presented an *unobservable attack*, which is a powerful attack mechanism capable of perturbing the state estimate without being detected. An unobservable attack can be formally defined as follows.

Definition 2.1: Given a measurement vector $\mathbf{z}$ corresponding to a state $\mathbf{x}$, i.e., $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{e}$, a state attack $\mathbf{a} \in \mathcal{A}$ is *unobservable* if there exists a state $\bar{\mathbf{x}} \neq \mathbf{x}$ such that $\mathbf{z} + \mathbf{a} = \mathbf{H}\bar{\mathbf{x}} + \mathbf{e}$.

The following Lemma shows the algebraic property of the attack; it follows immediately from the definition.

Lemma 2.1: A state attack is unobservable if and only if $\mathbf{a} \neq \mathbf{0}$, and $\mathbf{a} \in \mathcal{R}(\mathbf{H}) \cap \mathcal{A}$. Furthermore, if $\mathbf{a}$ is unobservable, so is $\gamma \cdot \mathbf{a}$ for any nonzero $\gamma \in \mathbb{R}$, and $\|\mathbf{x} - \bar{\mathbf{x}}_\gamma\|_2 \to \infty$ as $\gamma \to \infty$, where $\bar{\mathbf{x}}_\gamma$ denotes the state satisfying $\mathbf{H}\mathbf{x} + \gamma \cdot \mathbf{a} = \mathbf{H}\bar{\mathbf{x}}_\gamma$.

Lemma 2.1 implies that the feasibility of an unobservable attack does not depend on the current operating state $\mathbf{x}$. It only depends on $\mathcal{A}$, which is characterized by the set of adversary sensors, and the subspace $\mathcal{R}(\mathbf{H})$. The feasibility is also closely related to the concept of system observability. In particular, the following connection was found in [8].

Theorem 2.1 (Theorem 1, [8]): An unobservable attack is feasible if and only if removing the adversary sensors makes the grid unobservable (i.e., the measurement matrix does not have full column rank.)

Proof: See Appendix A. ■

III. Subspace Methods for Unobservable Attack

Most existing works on an unobservable attack assumed that an adversary knows the measurement matrix $\mathbf{H}$. In contrast, this section presents a design of an unobservable attack based on the system measurement subspace, without knowledge of $\mathbf{H}$. Employing the linearized measurement model (4), we will present the conditions under which an unobservable attack can be constructed based on the subspace information. We also demonstrate a condition that guarantees the design of an unobservable attack based on partial sensor measurements; for an attack on a power grid, this condition is characterized as a graph condition on the network topology.

A. Feasibility of an Unobservable Attack

Note that designing an unobservable attack is equivalent to finding a nonzero vector in $\mathcal{R}(\mathbf{H})$ satisfying the sparsity pattern defined by $\mathcal{A}$. Therefore, an unobservable attack, if feasible, can be launched by using a basis matrix $\mathbf{U} \in \mathbb{R}^{m \times n}$ of $\mathcal{R}(\mathbf{H})$ without knowing $\mathbf{H}$, as stated in the following theorem. Formally, we refer to $\mathcal{R}(\mathbf{H})$ as the *measurement subspace* because it is the subspace of all possible noiseless measurements.

Theorem 3.1: Let $\mathbf{U}$ be any basis matrix of $\mathcal{R}(\mathbf{H})$ and $\bar{\mathbf{U}}$ a submatrix of $\mathbf{U}$ obtained by removing the rows corresponding to the adversary sensors. Then, the following are true:

1) An unobservable attack is feasible if and only if $\bar{\mathbf{U}}$ does not have full column rank.
2) When feasible, an unobservable attack can be constructed using $\mathbf{U}$: for a nonzero vector $\mathbf{v} \in \mathcal{N}(\bar{\mathbf{U}})$, $\mathbf{a} \triangleq \mathbf{U}\mathbf{v}$ is an unobservable attack vector.

Proof: See Appendix B. ■

Theorem 3.1 states the feasibility condition in Theorem 2.1 (Theorem 1 of [8]; see also Theorem 5 of [21]) as a *subspace* condition. Note that in constructing an unobservable attack vector $\mathbf{U}\mathbf{v}$, the adversary only needs to know a basis matrix $\mathbf{U}$ of $\mathcal{R}(\mathbf{H})$.

B. Unobservable Attack With Partial Measurements

In this section, we show that an unobservable attack can be constructed using the subspace information of *partial* sensor measurements. To formally state the result, we need the notion of a critical set of sensors [27] and partial observability defined as follows.

Definition 3.1: A set of sensors is called a *critical set* if removing the set of sensors from the system renders the system unobservable while removing any strict subset of it does not. Let $\mathcal{S}$ and $\mathcal{X}$ denote a subset of sensors and a subset of state variables respectively. The state variables in $\mathcal{X}$ are said to be *observable with respect to* $\mathcal{S}$ if the state variables in $\mathcal{X}$ can be uniquely determined based on measurements from $\mathcal{S}$.[3] When the state variables in $\mathcal{X}$ are observable with respect to $\mathcal{S}$, a subset $\mathcal{C}$

[3]In other words, every element of $\mathcal{N}(\mathbf{H}_{\mathrm{s}})$ has zero entries for the rows corresponding to the state variables in $\mathcal{X}$, where $\mathbf{H}_{\mathrm{s}} \in \mathbb{R}^{|\mathcal{S}| \times n}$ is the submatrix of $\mathbf{H}$ obtained by retaining only the rows corresponding to the sensors in $\mathcal{S}$.

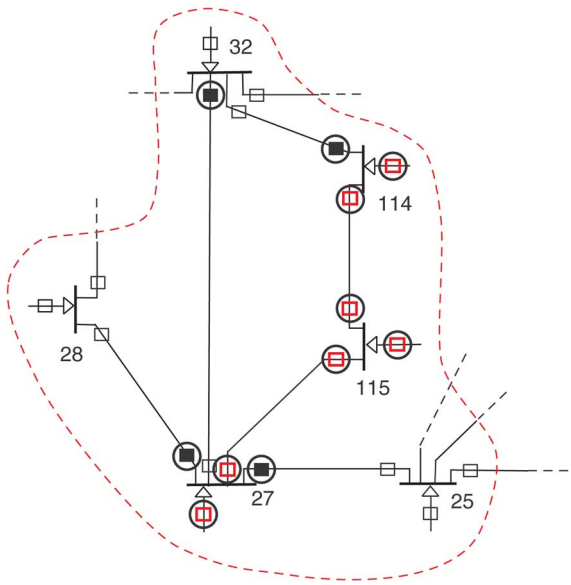


Fig. 2. A part of the IEEE 118-bus network: Rectangles represent the sensor locations. Every bus has an injection sensor, and every line has line flow sensors for both directions.

of $\mathcal{S}$ is a *critical set with respect to* $(\mathcal{S}, \mathcal{X})$ if removing $\mathcal{C}$ from $\mathcal{S}$ makes the state variables in $\mathcal{X}$ no longer observable with respect to $\mathcal{S}$ while removing a strict subset of $\mathcal{C}$ from $\mathcal{S}$ does not.

Consider a subset of sensors $\mathcal{S}_\mathrm{o}$. Let $\mathcal{X}_\mathrm{o}$ denote the set of state variables whose values affect measurements from the sensors in $\mathcal{S}_\mathrm{o}$ (i.e., the $|\mathcal{S}_\mathrm{o}|$ by n submatrix $\mathbf{H}_\mathrm{o}$ of $\mathbf{H}$, consisting of the rows corresponding to the sensors in $\mathcal{S}_\mathrm{o}$, has nonzero columns exactly at the columns corresponding to the state variables in $\mathcal{X}_\mathrm{o}$.)

The following theorem provides the conditions under which an unobservable attack can be constructed based on the subspace information of measurements from $\mathcal{S}_\mathrm{o}$. The conditions roughly mean that (i) based on measurements from $\mathcal{S}_\mathrm{o}$, one can uniquely identify the relevant state variables (i.e., the variables in $\mathcal{X}_\mathrm{o}$,) and (ii) $\mathcal{S}_\mathrm{o}$ contains a set of sensors, which, if controlled by an adversary, is sufficient for launching an unobservable attack and is also critical with respect to $(\mathcal{S}_\mathrm{o}, \mathcal{X}_\mathrm{o})$.

Theorem 3.2: Suppose that

1) the state variables in $\mathcal{X}_\mathrm{o}$ are observable with respect to $\mathcal{S}_\mathrm{o}$,
2) $\mathcal{C} \subset \mathcal{S}_\mathrm{o}$ is a critical set with respect to $(\mathcal{S}_\mathrm{o}, \mathcal{X}_\mathrm{o})$, and
3) removing $\mathcal{C}$ makes the system unobservable.

Let $\mathbf{H}_\mathrm{o} \in \mathbb{R}^{|\mathcal{S}_\mathrm{o}| \times n}$ denote the submatrix of $\mathbf{H}$ obtained by retaining only the rows corresponding to the sensors in $\mathcal{S}_\mathrm{o}$. Then, the following are true:

1) Let $\mathcal{A}_\mathrm{o}$ denote the set of vectors in $\mathcal{R}(\mathbf{H}_\mathrm{o})$ such that $\mathbf{f} \in \mathcal{R}(\mathbf{H}_\mathrm{o})$ is in $\mathcal{A}_\mathrm{o}$ if and only if the rows of $\mathbf{f}$ corresponding to the sensors in $\mathcal{S}_\mathrm{o} \setminus \mathcal{C}$ are equal to zero. Then, the dimension of $\mathcal{A}_\mathrm{o}$ is one.
2) For an arbitrary *nonzero* $\mathbf{a}_\mathrm{o} \in \mathcal{A}_\mathrm{o}$, the attack that modifies the sensor data from $\mathcal{C}$ by adding the corresponding entries in $\mathbf{a}_\mathrm{o}$ to the real data is unobservable.

Proof: See Appendix C. ∎

Note that $\mathcal{A}_\mathrm{o}$ in Theorem 3.2 can be fully characterized based on a basis matrix of $\mathcal{R}(\mathbf{H}_\mathrm{o})$. If the conditions of Theorem 3.2 are met, an attacker knowing a basis matrix of $\mathcal{R}(\mathbf{H}_\mathrm{o})$ can launch an unobservable attack. The following corollary provides the detail of how an attack can be constructed from a basis matrix of $\mathcal{R}(\mathbf{H}_\mathrm{o})$.

Corollary 3.2.1: Suppose that the conditions 1), 2), and 3) of Theorem 3.2 hold. Let $\mathbf{U}_\mathrm{o} \in \mathbb{R}^{|\mathcal{S}_\mathrm{o}| \times |\mathcal{X}_\mathrm{o}|}$ denote a basis matrix of $\mathcal{R}(\mathbf{H}_\mathrm{o})$ and $\bar{\mathbf{U}}_\mathrm{o}$ denote a submatrix of $\mathbf{U}_\mathrm{o}$ obtained by removing the rows corresponding to the sensors in $\mathcal{C}$. Then, the following are true:

1) The dimension of $\mathcal{N}(\bar{\mathbf{U}}_\mathrm{o})$ is one.
2) For any nonzero vector $\mathbf{v} \in \mathcal{N}(\bar{\mathbf{U}}_\mathrm{o})$, the attack that modifies the sensor data from $\mathcal{C}$ by adding the corresponding entries in $\mathbf{U}_\mathrm{o}\mathbf{v}$ to the real data is unobservable.

The three conditions of Theorem 3.2 are all related to system observability or partial observability. In case of a power grid, system observability and partial observability can be checked based on *partial* information about the grid topology and sensor locations. In particular, the graph-theoretical observability criterion in [20] can be employed.

A power grid is a network of buses connected by transmission lines. The *topology* of a grid is naturally defined as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of buses, and $\mathcal{E}$ is the set of connected transmission lines: $\{i, j\}$ is in $\mathcal{E}$ if and only if there exists a connected transmission line between bus i and bus j. We consider two types of legacy sensors: line flow sensors and bus injection sensors. A line flow sensor located on a line $\{i, j\}$ measures the power flowing through the line either from bus i to bus j or from bus j to bus i. A bus injection sensor on bus i measures the total power injected into the network at bus i (see Appendix F for the details of the sensor measurements.)

The following corollary presents the graph conditions that imply the conditions of Theorem 3.2 for an attack on a power grid state estimation. Appendix F provides the details of the graph-theoretical observability criterion in [20], which directly results in the following corollary from Theorem 3.2. To state the corollary, we need to introduce the concept of a *reduced power network*. Given a subset $\mathcal{S}_\mathrm{o}$ of sensors, the reduced network consists of the sensors in $\mathcal{S}_\mathrm{o}$ and the topology $\bar{\mathcal{G}} = (\bar{\mathcal{V}}, \bar{\mathcal{E}})$, where $\{i, j\}$ is in $\bar{\mathcal{E}}$ if and only if a line flow sensor on $\{i, j\}$ is in $\mathcal{S}_\mathrm{o}$, or an injection sensor at bus i or bus j is in $\mathcal{S}_\mathrm{o}$, and $\bar{\mathcal{V}}$ consists of all the endpoints of the lines in $\bar{\mathcal{E}}$. For instance, in the IEEE 118-bus network, Fig. 2 describes a reduced network for $\mathcal{S}_\mathrm{o}$ consisting of the circled sensors. In this example, the vertices and edges inside the dashed boundary form $\bar{\mathcal{G}}$.

Corollary 3.2.2: Let $\mathcal{S}_\mathrm{o}$ be a subset of sensors, $\bar{\mathcal{G}} = (\bar{\mathcal{V}}, \bar{\mathcal{E}})$ the topology of the reduced network for $\mathcal{S}_\mathrm{o}$, and $\mathcal{C}$ a subset of $\mathcal{S}_\mathrm{o}$. Suppose that

1) There exists a cut of the grid topology $\mathcal{G}$ such that $\mathcal{C}$ consists of all line flow sensors on the cutset lines and all injection sensors on the endpoints of the cutset lines.
2) For every sensor s in $\mathcal{C}$, there exists a way to assign each injection sensor in $(\mathcal{S}_\mathrm{o} \setminus \mathcal{C}) \cup \{s\}$ to a line incident to the bus where the sensor is located[4] such that there exists a spanning tree of $\bar{\mathcal{G}}$ with at least one sensor in $(\mathcal{S}_\mathrm{o} \setminus \mathcal{C}) \cup \{s\}$ on every edge of the tree (either a line flow or an assigned injection sensor.)

[4]In other words, for an injection sensor located at bus i, we assign the injection sensor to one of the lines that are incident to bus i. We do this for each injection sensor in $(\mathcal{S}_\mathrm{o} \setminus \mathcal{C}) \cup \{s\}$.

Then, the conditions of Theorem 3.2 hold, and thus the statements in Theorem 3.2 and Corollary 3.2.1 hold.

Note that the conditions of Corollary 3.2.2 are related to the topology and the sensor locations in the reduced network. Therefore, an adversary can exploit partial information about the topology and sensor locations to find an attack setting that enables an unobservable attack with partial sensor observations. For instance, it can be easily checked that the example in Fig. 2 with $\mathcal{C}$ consisting of the circled empty-rectangle sensors satisfies the conditions. In particular, the first condition is satisfied with the cut that isolates bus 115 from the rest of the network.

C. Subspace Attack Algorithm

All the information that is necessary for a subspace attack is the subspace information of $\mathcal{R}(\mathbf{H})$ or $\mathcal{R}(\mathbf{H}_{\mathrm{o}})$. Subspace estimation based on measurement data has been actively studied in the signal processing literature (e.g., [30], [31]), and thus subspace methods naturally lead to a data-driven algorithm for practical attack scenarios. Our focus in this section is to demonstrate how (any) subspace estimator can be used to generate a data-driven attack.

One of the simplest yet effective ways of estimating a basis matrix is to use a sample covariance matrix. Let $\mathbf{z}_1, \ldots, \mathbf{z}_K$ denote measurement vectors at K different sampling instances:

$$\mathbf{z}_i = \mathbf{H}\mathbf{x}_i + \mathbf{e}_i, \quad i = 1, \ldots, K. \tag{14}$$

For simplicity, suppose that the noise vectors $\mathbf{e}_1, \ldots, \mathbf{e}_K$ are independent and identically distributed (i.i.d.), the state vectors $\mathbf{x}_1, \ldots, \mathbf{x}_K$ are i.i.d. with a positive definite covariance matrix $\mathbf{\Sigma}_{\mathbf{x}}$, and the noise vectors and the state vectors are uncorrelated. Then, the covariance matrix of $\mathbf{z}$ is

$$\mathbf{\Sigma}_{\mathbf{z}} \triangleq \mathbb{E}\left[(\mathbf{z}_1 - \mathbb{E}[\mathbf{z}_1])(\mathbf{z}_1 - \mathbb{E}[\mathbf{z}_1])^T\right] = \mathbf{H}\mathbf{\Sigma}_{\mathbf{x}}\mathbf{H}^T + \sigma^2\mathbf{I}. \tag{15}$$

Note that $\mathbf{H}\mathbf{\Sigma}_{\mathbf{x}}\mathbf{H}^T$ has rank n. Therefore, if $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ is a singular value decomposition (SVD) of $\mathbf{\Sigma}_{\mathbf{z}}$, the n columns of $\mathbf{U}$ that correspond to the n largest singular values form a basis of $\mathcal{R}(\mathbf{H}\mathbf{\Sigma}_{\mathbf{x}}\mathbf{H}^T)$. Because $\mathcal{R}(\mathbf{H}\mathbf{\Sigma}_{\mathbf{x}}\mathbf{H}^T)$ is equivalent to $\mathcal{R}(\mathbf{H})$, the same columns form a basis of $\mathcal{R}(\mathbf{H})$.

Therefore, in practice, we can estimate a basis matrix of $\mathcal{R}(\mathbf{H})$ by applying SVD to the sample covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{z}}$:

$$\hat{\mathbf{\Sigma}}_{\mathbf{z}} \triangleq \frac{1}{K-1}\sum_{i=1}^{K}(\mathbf{z}_i - \underline{\mathbf{z}})(\mathbf{z}_i - \underline{\mathbf{z}})^T, \tag{16}$$

where $\underline{\mathbf{z}}$ denotes the sample mean.

Based on the above (or any other) subspace estimator and Theorem 3.1, the data-driven attack with *full* sensor observations operates as follows with the observations $\{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$ and the adversary sensor set $\mathcal{S}_{\mathrm{A}}$ as inputs:

1) **Subspace estimation**: Based on $\{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$, calculate an estimate $\hat{\mathbf{U}} \in \mathbb{R}^{m\times n}$ of a basis matrix of $\mathcal{R}(\mathbf{H})$.
2) **Null space estimation**: Obtain $\hat{\mathbf{U}}_1$ by removing the rows of $\hat{\mathbf{U}}$ that correspond to the sensors in $\mathcal{S}_{\mathrm{A}}$. Find an SVD of $\hat{\mathbf{U}}_1$, $\hat{\mathbf{U}}_1 = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^T$, and let $\mathbf{v}$ denote the column of $\tilde{\mathbf{V}}$ that corresponds to the smallest singular value ($\mathbf{v}$ is an estimate of a nonzero element of $\mathcal{N}(\bar{\mathbf{U}})$ in Theorem 3.1.)
3) **Attack**: Modify the sensor data from $\mathcal{S}_{\mathrm{A}}$ by adding the corresponding entries of $\eta \cdot \hat{\mathbf{U}}\mathbf{v}$ to them, where $\eta \in \mathbb{R}$ is a scaling factor to adjust the degree of perturbation.

The data-driven attack with *partial* sensor observations can be constructed in the same manner based on Corollary 3.2.1. Specifically, the attack receives $(\mathcal{X}_{\mathrm{o}}, \mathcal{S}_{\mathrm{o}}, \mathcal{C})$ and $\{\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_K\}$—the set of measurements from the sensors in $\mathcal{S}_{\mathrm{o}}$ at K different time instances—as inputs and executes the following steps:

1) **Subspace estimation**: Based on $\{\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_K\}$, calculate an estimate $\hat{\mathbf{U}}_{\mathrm{o}} \in \mathbb{R}^{|\mathcal{S}_{\mathrm{o}}|\times|\mathcal{X}_{\mathrm{o}}|}$ of a basis matrix of $\mathcal{R}(\mathbf{H}_{\mathrm{o}})$.
2) **Null space estimation**: Obtain $\hat{\mathbf{U}}_{\mathrm{c}}$ by removing the rows of $\hat{\mathbf{U}}_{\mathrm{o}}$ that correspond to the sensors in $\mathcal{C}$. Find an SVD of $\hat{\mathbf{U}}_{\mathrm{c}}$: $\hat{\mathbf{U}}_{\mathrm{c}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^T$. Let $\mathbf{v}$ denote the column of $\tilde{\mathbf{V}}$ that corresponds to the smallest singular value ($\mathbf{v}$ is an estimate of a nonzero element of $\mathcal{N}(\bar{\mathbf{U}}_{\mathrm{o}})$ in Corollary 3.2.1.)
3) **Attack**: Modify the sensor data from $\mathcal{C}$ by adding the corresponding entries of $\eta \cdot \hat{\mathbf{U}}_{\mathrm{o}}\mathbf{v}$ to them, where $\eta \in \mathbb{R}$ is a scaling factor to adjust the degree of perturbation.

IV. Subspace Methods for Data Framing Attack

The idea of a data framing attack based on full system parameter information was first presented in [24]. In this section, we demonstrate data-driven approaches of data framing attack by exploiting the subspace structure of sensor measurements.

A. Data Framing Attack

A data framing attack aims to enable an adversary to perturb the state estimate by an arbitrary degree even when an unobservable attack with $\mathcal{S}_{\mathrm{A}}$ does not exist. To this end, a data framing attack frames some normally operating meters as sources of bad data such that their data will be removed. A critical parameter of data framing attack is the set of sensors to be framed, denoted by $\mathcal{S}_{\mathrm{F}}$. The framed sensor set $\mathcal{S}_{\mathrm{F}}$ is selected such that $\mathcal{S}_{\mathrm{F}} \cap \mathcal{S}_{\mathrm{A}} = \emptyset$, and if the sensors in $\mathcal{S}_{\mathrm{F}}$ are removed from the system, an unobservable attack with $\mathcal{S}_{\mathrm{A}}$ becomes feasible. Under this selection rule, an adversary may design an attack that becomes unobservable once the sensor data from $\mathcal{S}_{\mathrm{F}}$ are removed by the bad data removal rule.

To successfully remove the data from $\mathcal{S}_{\mathrm{F}}$, one can use an attack vector that maximizes the energy of the normalized residues at $\mathcal{S}_{\mathrm{F}}$ in the *first* iteration of the bad data processing. Such an attack design does not necessarily guarantee that all data from $\mathcal{S}_{\mathrm{F}}$ will be identified as bad. Nevertheless, this is a reasonable heuristic to circumvent the difficulty of analyzing attack effect on normalized residues in all iterations.

To simplify notation, we drop the superscript that denotes the first iteration of bad data processing: all the quantities in this section are from the first iteration unless otherwise specified. The attack *direction* that maximizes the energy of the normalized residues in the first iteration can be constructed by solving the following optimization [24]:

$$\begin{aligned} \max_{\mathbf{a}} \quad & \mathbb{E}\left[\sum_{i\in\mathcal{S}_{\mathrm{F}}}(\tilde{r}_i)^2\right] \\ \text{subj.} \quad & \|\mathbf{a}\|_2^2 = 1, \quad \mathbf{a} \in \mathcal{R}(\mathbf{H}_1) \cap \mathcal{A}, \end{aligned} \tag{17}$$

where $\mathbf{H}_1 \in \mathbb{R}^{m\times n}$ is a matrix obtained from $\mathbf{H}$ by replacing the rows corresponding to the sensors in $\mathcal{S}_\mathrm{F}$ with zero row vectors. The constraint $\mathbf{a} \in \mathcal{R}(\mathbf{H}_1)$ holds if and only if $\mathbf{a}$ is unobservable after the framed sensor data are removed. This constraint guarantees that once the data from $\mathcal{S}_\mathrm{F}$ are removed, the attack can have the same effect as an unobservable attack.

The following theorem states that a solution to (17) can be obtained without knowing $\mathbf{H}$ if we know a basis matrix of $\mathcal{R}(\mathbf{H})$.

Theorem 4.1: An adversary knowing a basis matrix $\mathbf{U} \in \mathbb{R}^{m\times n}$ of $\mathcal{R}(\mathbf{H})$ can find a solution of (17). Specifically, a solution to the following quadratically constrained quadratic programming (QCQP) is also a solution to (17), and vice versa:

$$\begin{aligned} \max_{\mathbf{a}} \quad & \left\|\mathbf{I}_{\mathcal{S}_\mathrm{F}}\tilde{\mathbf{\Omega}}\tilde{\mathbf{W}}\mathbf{a}\right\|_2^2 \\ \text{subj.} \quad & \|\mathbf{a}\|_2^2 = 1, \qquad \mathbf{a} \in \mathcal{R}(\mathbf{U}_1) \cap \mathcal{A}, \end{aligned} \tag{18}$$

where $\mathbf{I}_{\mathcal{S}_\mathrm{F}} \in \mathbb{R}^{|\mathcal{S}_\mathrm{F}|\times m}$ is the row selection operator that retains only the rows corresponding to the sensors in $\mathcal{S}_\mathrm{F}$ out of m rows,

$$\tilde{\mathbf{W}} \triangleq \mathbf{I} - \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T, \tag{19}$$

$\tilde{\mathbf{\Omega}} \in \mathbb{R}^{m\times m}$ is a diagonal matrix with

$$\tilde{\mathbf{\Omega}}_{ii} = \begin{cases} \frac{1}{\sqrt{\tilde{\mathbf{W}}_{ii}}} & \text{if } \tilde{\mathbf{W}}_{ii} > 0; \\ 0 & \text{if } \tilde{\mathbf{W}}_{ii} = 0, \end{cases} \tag{20}$$

and $\mathbf{U}_1 \in \mathbb{R}^{m\times n}$ is a matrix obtained from $\mathbf{U}$ by replacing the rows corresponding to the sensors in $\mathcal{S}_\mathrm{F}$ with zero row vectors.

Proof: See Appendix D. ∎

Note that addition of the attack vector $\mathbf{a}$ changes the mean of the residue vector from $\mathbf{0}$ to $\tilde{\mathbf{W}}\mathbf{a}$. And, $\mathbf{I}_{\mathcal{S}_\mathrm{F}}\tilde{\mathbf{\Omega}}\tilde{\mathbf{W}}\mathbf{a}/\sigma$ is the resulting mean of the normalized residues of the data from $\mathcal{S}_\mathrm{F}$.

B. Sufficiency of Partial Measurements

Similar to sufficiency of partial measurements for an unobservable attack (Theorem 3.2), data framing attack can also be launched based on subspace information of partial measurements, as stated formally in the following theorem. Below, we use the notations defined in Section III-B for the partial measurement case.

Theorem 4.2: Suppose that the conditions 1), 2), and 3) of Theorem 3.2 hold for $\mathcal{S}_\mathrm{o}$, $\mathcal{X}_\mathrm{o}$, and $\mathcal{C}$. Let $\{\mathcal{C}_1, \mathcal{C}_2\}$ denote an arbitrary partition of $\mathcal{C}$. Let $\mathbf{H}_\mathrm{A}$ denote a submatrix of $\mathbf{H}$ consisting of the rows corresponding to the sensors in $\mathcal{S}_\mathrm{o} \setminus \mathcal{C}_2$, $\mathbf{U}_\mathrm{A} \in \mathbb{R}^{|\mathcal{S}_\mathrm{o}\setminus\mathcal{C}_2|\times|\mathcal{X}_\mathrm{o}|}$ denote a basis matrix of $\mathcal{R}(\mathbf{H}_\mathrm{A})$, and $\bar{\mathbf{U}}_\mathrm{A}$ denote a submatrix of $\mathbf{U}_\mathrm{A}$ obtained by removing the rows corresponding to the sensors in $\mathcal{C}_1$. Then, the following are true:

1) The dimension of $\mathcal{N}(\bar{\mathbf{U}}_\mathrm{A})$ is one.
2) For a nonzero vector $\mathbf{v} \in \mathcal{N}(\bar{\mathbf{U}}_\mathrm{A})$, the attack that modifies the sensor data from $\mathcal{C}_1$ by adding the corresponding entries in $\mathbf{U}_\mathrm{A}\mathbf{v}$ to the real data is equivalent to using $\alpha\cdot\mathbf{a}^*$ as an attack vector, where α is a nonzero real number, and $\mathbf{a}^*$ is an optimal solution to (17) with $(\mathcal{S}_\mathrm{A}, \mathcal{S}_\mathrm{F}) = (\mathcal{C}_1, \mathcal{C}_2)$.

Proof: See Appendix E. ∎

Theorem 4.2 implies that under certain conditions, knowledge of a basis matrix of $\mathcal{R}(\mathbf{H}_\mathrm{A})$—the subspace of measurements from $\mathcal{S}_\mathrm{o} \setminus \mathcal{C}_2$—is sufficient for launching a data framing attack with $(\mathcal{S}_\mathrm{A}, \mathcal{S}_\mathrm{F}) = (\mathcal{C}_1, \mathcal{C}_2)$. Note that Theorem 4.2 requires the same conditions as Theorem 3.2. Therefore, for an attack on a power grid, the graph conditions in Corollary 3.2.2 can replace the conditions of Theorem 4.2.

C. Subspace Data Framing Attack Algorithm

Theorem 4.1 and Theorem 4.2 guarantee the sufficiency of subspace information in constructing data framing attacks. Similar to the data-driven algorithms for unobservable attacks, we can incorporate a subspace estimator and SVD to build a data-driven algorithm for data framing attacks.

The data-driven framing attack with *full* sensor observations receives sensor observations $\{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$ at K different time instances and $(\mathcal{S}_\mathrm{A}, \mathcal{S}_\mathrm{F})$ as inputs, and it has two small positive parameters ϵ_1 and ϵ_2 for thresholding rules. Based on the QCQP formulation (18), it works as follows:

1) **Subspace estimation**: Based on $\{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$, calculate an estimate $\hat{\mathbf{U}} \in \mathbb{R}^{m\times n}$ of a basis matrix of $\mathcal{R}(\mathbf{H})$.
2) **Null space estimation**: Obtain $\hat{\mathbf{U}}_1$ by removing the rows of $\hat{\mathbf{U}}$ that correspond to the sensors in $\mathcal{S}_\mathrm{A} \cup \mathcal{S}_\mathrm{F}$. Find an SVD of $\hat{\mathbf{U}}_1$: $\hat{\mathbf{U}}_1 = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^T$. Let $\tilde{\mathbf{V}}$ denote the matrix consisting of the columns of $\tilde{\mathbf{V}}$ whose corresponding singular values are less than ϵ_1. Let $\hat{\mathbf{U}}_\mathrm{A} \in \mathbb{R}^{m\times n}$ be the matrix obtained from $\hat{\mathbf{U}}$ by replacing the rows corresponding to the sensors *not* in $\mathcal{S}_\mathrm{A}$ with zero row vectors. Then, $\hat{\mathbf{U}}_\mathrm{A}\tilde{\mathbf{V}}$ is an estimate of a basis matrix of $\mathcal{R}(\mathbf{U}_1) \cap \mathcal{A}$ in (18).[5]
3) **QCQP parameter estimation**: Calculate

$$\hat{\mathbf{W}} \triangleq \mathbf{I} - \hat{\mathbf{U}}(\hat{\mathbf{U}}^T\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}^T \tag{21}$$

and $\hat{\mathbf{\Omega}} \in \mathbb{R}^{m\times m}$, which is a diagonal matrix with

$$\hat{\mathbf{\Omega}}_{ii} = \begin{cases} \sqrt{\frac{1}{\hat{\mathbf{W}}_{ii}}} & \text{if } \hat{\mathbf{W}}_{ii} > \epsilon_2; \\ 0 & \text{if } \hat{\mathbf{W}}_{ii} < \epsilon_2. \end{cases} \tag{22}$$

4) **QCQP**: Solve maximizing $\|\mathbf{I}_{\mathcal{S}_\mathrm{F}}\hat{\mathbf{\Omega}}\hat{\mathbf{W}}\hat{\mathbf{U}}_\mathrm{A}\hat{\mathbf{V}}\mathbf{y}\|_2^2$ subject to $\|\hat{\mathbf{U}}_\mathrm{A}\hat{\mathbf{V}}\mathbf{y}\|_2^2 = 1$ and $\mathbf{y} \in \mathbb{R}^k$, where k is the number of columns of $\hat{\mathbf{V}}$. Let $\mathbf{y}^*$ denote the solution.
5) **Attack**: Modify the sensor data from $\mathcal{S}_\mathrm{A}$ by adding the corresponding entries of $\eta\cdot\hat{\mathbf{U}}_\mathrm{A}\hat{\mathbf{V}}\mathbf{y}^*$ to them, where $\eta \in \mathbb{R}$ is a scaling factor to adjust the degree of perturbation.

Based on Theorem 4.2, the data-driven framing attack with *partial* sensor observations receives $(\mathcal{X}_\mathrm{o}, \mathcal{S}_\mathrm{o}, \mathcal{C}_1, \mathcal{C}_2)$ and $\{\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_K\}$—the set of measurements from the sensors in $\mathcal{S}_\mathrm{o} \setminus \mathcal{C}_2$ at K different time instances—as inputs and executes the following steps:

1) **Subspace estimation**: Based on $\{\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_K\}$, calculate an estimate $\hat{\mathbf{U}}_\mathrm{A} \in \mathbb{R}^{|\mathcal{S}_\mathrm{O}\setminus\mathcal{C}_2|\times|\mathcal{X}_\mathrm{O}|}$ of a basis matrix of $\mathcal{R}(\mathbf{H}_\mathrm{A})$.
2) **Null space estimation**: Obtain $\hat{\mathbf{U}}_\mathrm{c}$ by removing the rows of $\hat{\mathbf{U}}_\mathrm{A}$ that correspond to the sensors in $\mathcal{C}_1$. Find an SVD of $\hat{\mathbf{U}}_\mathrm{c}$: $\hat{\mathbf{U}}_\mathrm{c} = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^T$. Let $\mathbf{v}$ denote the column of $\tilde{\mathbf{V}}$ that

[5]A basis matrix of $\mathcal{R}(\mathbf{U}_1) \cap \mathcal{A}$ in (18) can be found by noting that $\mathbf{a} \in \mathcal{R}(\mathbf{U}_1) \cap \mathcal{A}$ if and only if $\mathbf{a} = \mathbf{U}_1\mathbf{y}$ for some $\mathbf{y} \in \mathcal{N}(\mathbf{U}_2)$ where $\mathbf{U}_2 \in \mathbb{R}^{(m-|\mathcal{S}_\mathrm{A}\cup\mathcal{S}_\mathrm{F}|)\times n}$ is a submatrix of $\mathbf{U}$ obtained by removing the rows corresponding to the sensors in $\mathcal{S}_\mathrm{A} \cup \mathcal{S}_\mathrm{F}$. In other words, given a basis matrix $\mathbf{B}$ of $\mathcal{N}(\mathbf{U}_2)$, $\mathbf{U}_1\mathbf{B}$ is a basis matrix of $\mathcal{R}(\mathbf{U}_1) \cap \mathcal{A}$.

corresponds to the smallest singular value ($\mathbf{v}$ is an estimate of a nonzero element of $\mathcal{N}(\bar{\mathbf{U}}_{\mathrm{A}})$ in Theorem 4.2.)

3) **Attack**: Modify the sensor data from $\mathcal{C}_1$ by adding the corresponding entries of $\eta \cdot \hat{\mathbf{U}}_{\mathrm{A}}\mathbf{v}$ to them, where $\eta \in \mathbb{R}$ is a scaling factor to adjust the degree of perturbation.

V. Numerical Results

In this section, simulations with benchmark power grids, the IEEE 14-bus network and the IEEE 118-bus network, demonstrate the performance of data-driven attacks. The nonlinear measurement model (1) and the nonlinear state estimator were employed to emulate practical power system state estimation. The power system measurement model is briefly described in Appendix F.

As an attack performance metric, we use the l_2 norm of the state estimation error, i.e., $\|\hat{\mathbf{x}} - \mathbf{x}\|_2$, where $\hat{\mathbf{x}}$ is the state estimate, and $\mathbf{x}$ is the true state.

A. Simulation Methods

In each Monte Carlo run, we used the nonlinear model (1) to generate measurement vectors. State vectors at different time points were assumed to be independent and identically distributed Gaussian random vectors with the mean equal to the operating states given in the IEEE 14-bus and 118-bus data [18]. Both the 14-bus network and the 118-bus network were assumed to be fully measured; i.e., all bus injections and all line flows (in both directions for each line) were measured by sensors.

In each simulation scenario, we compared performance of three attack methods: an attack with full knowledge of $\mathbf{H}$, a data-driven attack with full sensor observations, and a data-driven attack with partial sensor observations. For data-driven attacks, 1,000 samples were used to estimate a basis matrix of the subspace of (either full or partial) measurements unless otherwise specified. Data-driven attacks employed the subspace estimator based on the sample covariance matrix which was described in Section III-C.

Once an attack vector was added to measurements, the control center executed nonlinear state estimation and bad data detection and removal, as described in Section II-C, on the corrupted measurements. The threshold $\tau^{(k)}$ of the bad data detector (i.e., the $J(\hat{\mathbf{x}})$-test) was set to satisfy the false alarm constraint 0.04.

B. Data-Driven Unobservable Attack

1) IEEE 14-Bus Test: In the IEEE 14-bus network, we considered an adversary controlling data from $(\bar{2})$, $(\bar{3})$, $(\bar{4})$, (2, 3), (3, 2), (3, 4), and (4, 3), as illustrated in Fig. 3: $(\bar{i})$ denotes the injection sensor at bus i, and (i, j) denotes the line flow sensor measuring the power flow from i to j. Theorem 2.1 and the spanning tree observability criterion [20] imply that the adversary is capable of launching an unobservable attack (see Appendix F.) In addition, the adversary sensor set is also a critical set, and thus all possible unobservable attack vectors are aligned along the same direction (i.e., the dimension of $\mathcal{A} \cap \mathcal{R}(\mathbf{H})$ is one.)

An adversary with partial sensor observations was assumed to observe data from $(\bar{2})$, $(\bar{3})$, $(\bar{4})$, (2, 3), (3, 2), (3, 4), (4, 3), (2, 1), (2, 4), (4, 5), (4, 7), and (4, 9). In this setting, it can be easily verified that the conditions of Corollary 3.2.2 are satisfied, and thus an adversary with partial observations can construct an unobservable attack under the linearized model assumption.

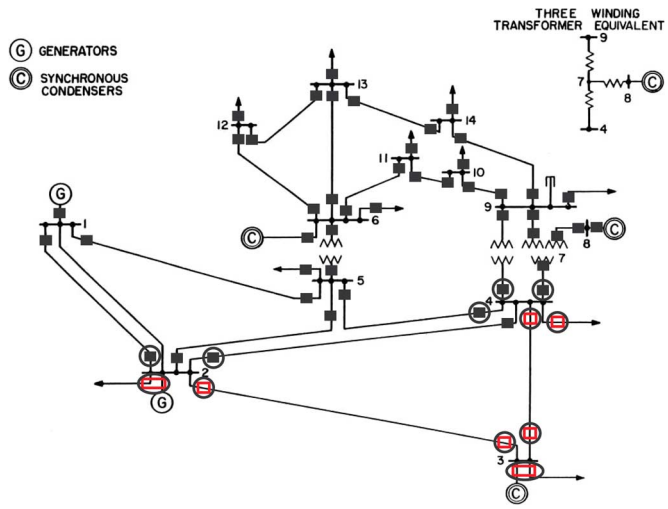


Fig. 3. IEEE 14-bus network: the circled red *empty* rectangles represent the adversary sensors (i.e., the sensors in $\mathcal{S}_{\mathrm{A}}$). The adversary with partial sensor observations can observe all the circled sensors.

Fig. 4 shows the performance of unobservable attacks, especially the plot of the *normalized* state estimation error versus the relative attack magnitude ($\|\mathbf{a}\|_1/\|\mathbf{z}\|_1$). For normalization, state estimation errors are divided by the mean estimation error under the non-attack scenario. In the plot, each marked point (circle, rectangle, or triangle) denotes the mean of a normalized error caused by an attack, and the vertical bar on the marked point is the confidence interval of the normalized error with 90% confidence level (i.e., the normalized error stayed in the interval with probability 0.9 in our simulations.) Data-driven attacks, based on either full or partial sensor observations, perform as well as the attack with full knowledge of $\mathbf{H}$. For all attacks, the resulting state estimate error scales as the attack magnitude scales. The confidence intervals imply that attacks are successful with high probability. The overall results indicate that even in a practical nonlinear power system, the data-driven attacks designed based on the linear model can perform well, and partial sensor observations can provide sufficient information for designing a successful attack.

The proposed subspace approaches approximate the nonlinear measurement model with a linearized model and construct an attack vector based on the subspace structure of the assumed linearized model. The *nonlinearity* of the actual measurement model poses a gap between the theory and practice, introducing the possibility that a data-driven unobservable attack might be detectable in practice. Furthermore, even when the actual measurement model is linear, the subspace estimate will inevitably contain errors due to the *limited sample size*. Therefore, attacks designed based on the estimate cannot be unobservable in a strict sense although they might approximate unobservable attacks.

We examined the probability that an unobservable attack is detected by bad data detection at the control center. Fig. 5 shows

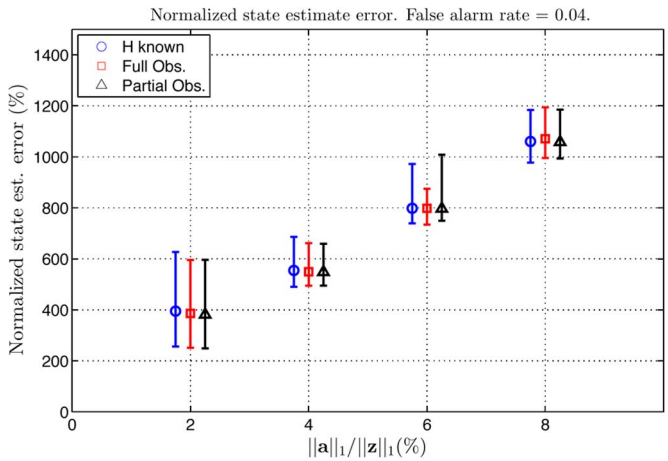


Fig. 4. Unobservable attacks on the 14-bus network: the sensor SNR is 46 dB. Attacks with the relative attack magnitudes 2, 4, 6, and 8% were tested. For each scenario, 1,000 Monte Carlo runs are used.

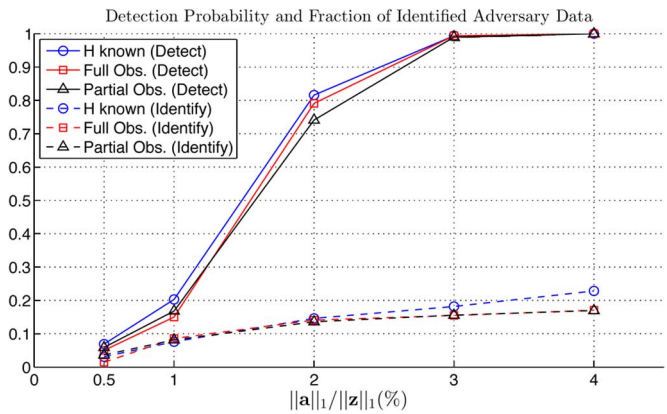


Fig. 5. Detection probability (solid lines) and the average fraction of the adversary data that are identified as bad (dashed lines). Attacks with the relative attack magnitudes 0.5, 1, 2, 3, and 4% were tested. For each scenario, 1,000 Monte Carlo runs are used.

the detection probability for the unobservable attacks, the setting and performance of which were described in Fig. 4. It turns out that when the attack amplitude is large, the attacks are detected by bad data detection with high probability. Considering that the attack with the perfect knowledge of $\mathbf{H}$ and the data-driven attacks resulted in similar detection probabilities, detectability seems to be attributed to nonlinearity of the actual measurement model rather than subspace estimation error. Nevertheless, Fig. 5 also shows the mean fraction of adversary-modified data that are identified as bad by the bad data detection, which turns out to be very small. The results imply that even though the control center notices that some data are not trustworthy, bad data detection filters out only a small fraction of adversary-controlled data thereby leading to successful attack performance as shown in Fig. 4. Note that even when a number of adversary-controlled data are filtered out, the performance of an unobservable attack (i.e., the resulting state estimate error) remains the same as long as some adversary-controlled data remains unremoved.[6] This can be easily seen from the structure of an unobservable attack described in Section II-D.

TABLE I
DETECTION PROBABILITY OF DATA-DRIVEN ATTACKS VERSUS THE SAMPLE SIZE FOR SUBSPACE ESTIMATION: $\|\mathbf{a}\|_1/\|\mathbf{z}\|_1$ WAS SET TO 1%, AND 1,000 MONTE CARLO RUNS WERE USED

	Sample size (K)			
Attack type	$K = 250$	$K = 500$	$K = 750$	$K = 1000$
Full Obs.	0.176	0.180	0.169	0.150
Partial Obs.	0.166	0.146	0.151	0.168

To further examine the effect of subspace estimation error on the attack performance, we tried different sample sizes for subspace estimation (especially, 250, 500, 750, 1000) under the various attack scenarios. As the representative result in Table I demonstrates, the sample size did not affect the detection probability of data-driven attacks. Furthermore, the resulting state estimation errors were barely affected by the sample size for subspace estimation: for instance, when the relative attack magnitude was set to 1%, the mean normalized errors from all data-driven attacks with different sample sizes stayed between 227% and 238%. These results imply that the effect of nonlinearity on attack performance dominates the effect of subspace estimation errors.

However, we believe that the data-driven attacks are asymptotically unobservable if the measurement model is *linear*. First of all, it is well known that the subspace estimate based on the sample covariance converges in probability to a true basis matrix of the subspace as the sample size grows (see Theorem 1 in [32] or Lemma 3.1 in [15].) Second, in a data-driven attack, the mapping from a subspace estimate to an attack vector is continuous. Therefore, the continuous mapping theorem (see Theorem 29.2 in [33]) implies that the data-driven attack vector converges in probability[7] to the attack vector constructed based on the exact subspace information, which is an unobservable attack. Subsequently, Slutsky's theorem (see Theorem 1.11 in [34]) can be used to show that the corrupted measurement vector under a data-driven attack converges in distribution to a corrupted measurement vector under an unobservable attack as the sample size grows. As an experimental evidence, Table II provides the detection probabilities of data-driven attacks when we used the linear model (4) for measurement generation and employed the linear state estimator instead of the nonlinear estimator. The results show that the detection probability converges to the false alarm constraint of the bad data detector as the sample size grows, which means that the attack becomes unobservable as the sample size grows.

2) IEEE 118-Bus Test: In the IEEE 118-bus simulation, we considered unobservable attacks discussed in the example in

[6]An unobservable attack $\mathbf{a}$ is equal to $\mathbf{Hy}$ for some nonzero $\mathbf{y}$. Suppose that bad data detection identifies some adversary-controlled data as bad and removes them. Let $\bar{\mathbf{a}}$ and $\bar{\mathbf{H}}$ denote the attack vector and the measurement matrix respectively after removing all the rows corresponding to the sensors identified as bad by bad data detection. Then, $\bar{\mathbf{a}}$, the remaining attack modification, is equal to $\bar{\mathbf{H}}\mathbf{y}$, and thus $\bar{\mathbf{a}}$ will perturb the state estimate by $\mathbf{y}$.

[7]The continuous mapping theorem implies only the convergence in distribution. However, the convergence in distribution to a *constant* implies the convergence in probability[33].

TABLE II
DETECTION PROBABILITY OF ATTACKS ON THE LINEAR MODEL: $\|\mathbf{a}\|_1/\|\mathbf{z}\|_1$ WAS 4%, AND 4,000 MONTE CARLO RUNS WERE USED

	Sample size (K)				
Attack type	$K = 8$	$K = 12$	$K = 16$	$K = 24$	$K = 40$
Full Obs.	0.998	0.774	0.105	0.044	0.043
Partial Obs.	0.280	0.066	0.053	0.044	0.043

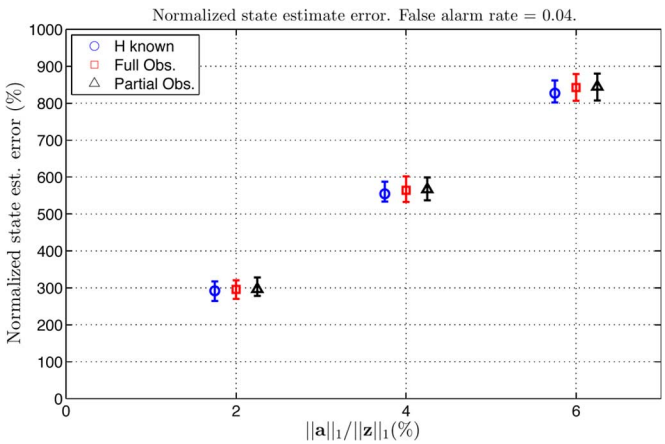


Fig. 6. Unobservable attacks on the 118-bus network: the sensor SNR is 46 dB. Attacks with the relative attack magnitudes 2, 4, and 6% were tested. For each scenario, 200 Monte Carlo runs are used.

Fig. 2 of Section III-B. Fig. 6 shows the plots of the normalized state estimation error versus the relative attack magnitude and the confidence intervals with 90% confidence level. Three methods resulted in almost the same degree of perturbation on the state estimate. In particular, the performance of data-driven attacks with partial sensor observations demonstrates that observing data from a *small fraction* of sensors can be sufficient for designing a successful attack on a large system; only about 2 percent of sensors need to be observed.

C. Data-Driven Framing Attack

1) IEEE 14-Bus Test: For data framing attacks, we considered an adversary who controls (2, 3), (3, 4), and (4, 3), and frames $(\bar{2})$, $(\bar{3})$, $(\bar{4})$, and (3, 2) as sources of bad data. Under this setting, an adversary cannot launch an unobservable attack. In attacks with partial observations, an adversary was assumed to observe data from (2, 3), (3, 4), (4, 3), (2, 1), (2, 4), (4, 5), (4, 7), and (4, 9). This setting satisfies the conditions of Corollary 3.2.2, and thus the conditions of Theorem 4.2 are also satisfied. Hence, an adversary with partial sensor observations is capable of designing a data framing attack under the linearized model assumption.

Fig. 7 shows the plots of the normalized state estimation error versus the relative attack magnitude and the confidence intervals with 90% confidence level. The results show that even when an unobservable attack is not feasible, an adversary may exploit the idea of data framing to perturb the state estimate by an arbitrary degree. Furthermore, the results indicate that data-driven attacks designed based on the linearized model perform well on nonlinear power systems, and partial sensor observations are sufficient for designing a data framing attack. To investigate the effect of subspace estimation error on attack performance, we

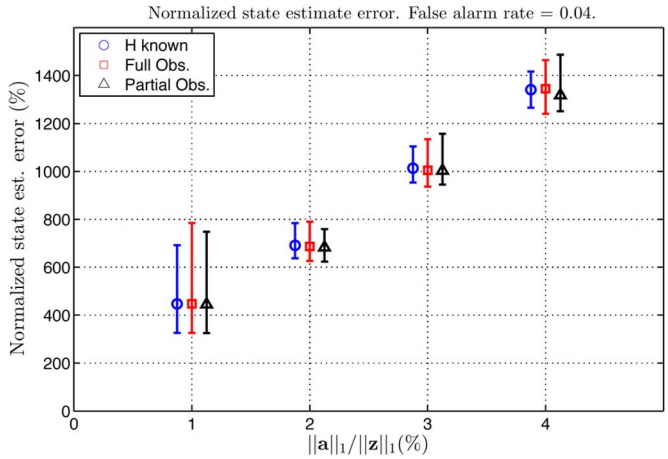


Fig. 7. Data framing attacks on the 14-bus network: the sensor SNR is 46 dB. Attacks with the relative attack magnitudes 1, 2, 3, and 4% were tested. For each scenario, 1,000 Monte Carlo runs are used.

tried different sample sizes (250, 500, 750, and 1,000) for subspace estimation in data framing attacks. Similar to the case of unobservable attacks, the sample size hardly affected the attack performance. This seems to imply that the effect of nonlinearity on attack performance dominates the effect of subspace estimation errors.

2) IEEE 118-Bus Test: We considered an adversary attacking the part of the 118-bus network illustrated in Fig. 2. The adversary was assumed to control (114, 115), (115, 114), and (27, 115), and frame $(1\bar{1}4)$, $(1\bar{1}5)$, $(\bar{2}7)$, and (115, 27) as sources of bad data. An adversary with partial sensor observations was assumed to observe data from the circled sensors in Fig. 2 except $(1\bar{1}4)$, $(1\bar{1}5)$, $(\bar{2}7)$, and (115, 27). The graph conditions of Corollary 3.2.2 are satisfied, and thus an adversary with partial observations is capable of launching a data framing attack under the linearized model assumption.

Fig. 8 shows the plots of the normalized state estimation error versus the relative attack magnitude and the confidence intervals with 90% confidence level. The results demonstrate the sufficiency of partial sensor observations for designing a data framing attack in a large network.

VI. CONCLUSIONS

This paper presents subspace methods of data attacks on state estimators of cyber physical systems. By exploiting the fact that subspace information of measurements is sufficient for designing attacks, we devised data-driven attacks that can be launched based on partial sensor observations. The numerical results demonstrated that the data-driven attacks are as efficient as the attacks based on full system information.

Our results demonstrate that one should not presumably underestimate the ability of an adversary even when system information is secure from the adversary. Even a leak of a small fraction of certain sensor measurements may provide enough data, upon which state attacks can be constructed.

Most countermeasures in the literature focused on protecting certain sensor data from adversarial modification via data authentication, while assuming that system parameters are known

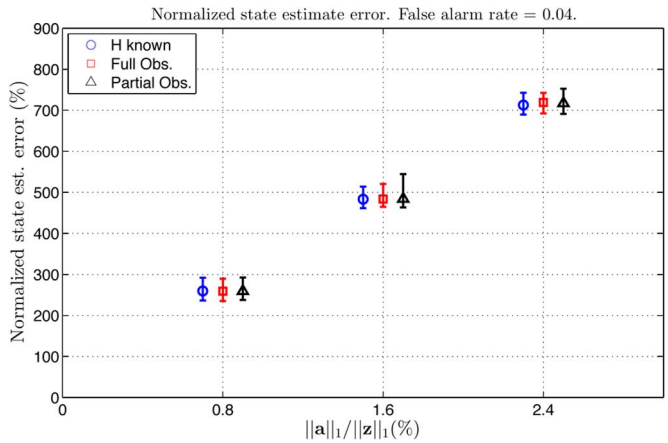


Fig. 8. Data framing attacks on the 118-bus network: the sensor SNR is 46 dB. Attacks with the relative attack magnitudes 0.8, 1.6, and 2.4% were tested. For each scenario, 200 Monte Carlo runs are used.

to adversaries (e.g., [7], [9], [12], [21]). In case that system parameter information is kept secure, our results demonstrate that not only the ability to modify data but also the ability to observe data are critical to an adversary. Therefore, as a countermeasure, on top of a data authentication strategy, one can strategically enhance data encryption and access control protocols to limit the set of data an adversary may eavesdrop.

Lastly, the successful performance of data framing attacks suggests that current bad data detection and removal mechanisms might not be the best in this day and age of cyber security concerns. A bad data processing mechanism based on dynamic state estimation or prior knowledge of sensor qualities (e.g., a Bayesian approach) might be more appropriate for defeating such attacks.

Appendix A
Proof of Theorem 2.1

Let $\bar{\mathbf{H}}$ denote the measurement matrix after the sensors in $\mathcal{S}_{\mathrm{A}}$ are removed; i.e., $\bar{\mathbf{H}}$ is obtained from $\mathbf{H}$ by removing the rows corresponding to the adversary sensors. Then, $\mathbf{Hy}$ is in $\mathcal{A}$ if and only if $\mathbf{y}$ is in $\mathcal{N}(\bar{\mathbf{H}})$—the null space of $\bar{\mathbf{H}}$. This implies that an unobservable attack is feasible if and only if $\bar{\mathbf{H}}$ does not have full column rank (i.e., $\mathcal{N}(\bar{\mathbf{H}})$ has a nonzero dimension.) ■

Appendix B
Proof of Theorem 3.1

The columns of $\bar{\mathbf{U}}$ span $\mathcal{R}(\bar{\mathbf{H}})$. In addition, because $\bar{\mathbf{U}}$ and $\bar{\mathbf{H}}$ have the same number of columns, $\bar{\mathbf{U}}$ does not have full column rank if and only if $\bar{\mathbf{H}}$ does not have full column rank. Therefore, Theorem 2.1 implies that an unobservable attack is feasible if and only if $\bar{\mathbf{U}}$ does not have full column rank.

Suppose that an unobservable attack is feasible. Then, $\bar{\mathbf{U}}$ is rank deficient, and we can find a nonzero vector $\mathbf{v} \in \mathcal{N}(\bar{\mathbf{U}})$. With $\mathbf{a} \triangleq \mathbf{Uv}$, $\mathbf{a}$ is in $\mathcal{A}$ because $\mathbf{Uv}$ has zero entries for the sensors *not* in $\mathcal{S}_{\mathrm{A}}$ (i.e., $\bar{\mathbf{U}}\mathbf{v} = \mathbf{0}$).

In addition, there exists an invertible matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$ such that $\mathbf{H} = \mathbf{UB}$, and $\mathbf{U} = \mathbf{HB}^{-1}$, because $\mathbf{H}$ has full column rank. Therefore, $\mathbf{Uv} = \mathbf{H}(\mathbf{B}^{-1}\mathbf{v})$, and thus $\mathbf{a}$ is an unobservable attack vector. ■

Appendix C
Proof of Theorem 3.2

Let $\bar{\mathbf{H}}$ denote the submatrix of $\mathbf{H}$ obtained by removing the rows corresponding to the sensors in $\mathcal{C}$. Then, $\mathcal{N}(\bar{\mathbf{H}})$ is not null due to the third assumption. Let $\mathbf{y}$ denote a nonzero vector in $\mathcal{N}(\bar{\mathbf{H}})$ and $\mathbf{y}_{\mathrm{o}}$ denote a subvector of $\mathbf{y}$ obtained by retaining only the rows corresponding to the state variables in $\mathcal{X}_{\mathrm{o}}$. In addition, let $\mathbf{H}_{\mathrm{s}}$ denote a submatrix of $\mathbf{H}_{\mathrm{o}}$ obtained by retaining only the columns corresponding to the state variables in $\mathcal{X}_{\mathrm{o}}$ (note that all the other columns of $\mathbf{H}_{\mathrm{o}}$ are zero vectors.) And, $\bar{\mathbf{H}}_{\mathrm{s}}$ denotes a submatrix of $\mathbf{H}_{\mathrm{s}}$ obtained by removing the rows corresponding to the sensors in $\mathcal{C}$.

First, note that $\mathbf{a}_{\mathrm{o}} \in \mathcal{A}_{\mathrm{o}}$ if and only if $\mathbf{a}_{\mathrm{o}} = \mathbf{H}_{\mathrm{s}}\mathbf{p}$ for some $\mathbf{p} \in \mathcal{N}(\bar{\mathbf{H}}_{\mathrm{s}})$. In addition, because $\mathcal{C}$ is a critical set with respect to $(\mathcal{S}_{\mathrm{o}}, \mathcal{X}_{\mathrm{o}})$, $\mathcal{N}(\bar{\mathbf{H}}_{\mathrm{s}})$ has dimension one. Note that $\bar{\mathbf{H}}_{\mathrm{s}}\mathbf{y}_{\mathrm{o}} = \mathbf{0}$ whereas $\mathbf{H}_{\mathrm{s}}\mathbf{y}_{\mathrm{o}} \neq \mathbf{0}$. This implies that $\mathbf{y}_{\mathrm{o}} \neq \mathbf{0}$, and $\{\mathbf{y}_{\mathrm{o}}\}$ is a basis of $\mathcal{N}(\bar{\mathbf{H}}_{\mathrm{s}})$. Therefore, $\{\mathbf{H}_{\mathrm{s}}\mathbf{y}_{\mathrm{o}}\}$ is a basis of $\mathcal{A}_{\mathrm{o}}$.

Therefore, for any nonzero $\mathbf{a}_{\mathrm{o}} \in \mathcal{A}_{\mathrm{o}}$, there exists a nonzero $\alpha \in \mathbb{R}$ such that $\mathbf{a}_{\mathrm{o}} = \alpha \cdot \mathbf{H}_{\mathrm{s}}\mathbf{y}_{\mathrm{o}}$. Furthermore, $\mathbf{H}_{\mathrm{s}}\mathbf{y}_{\mathrm{o}} = \mathbf{H}_{\mathrm{o}}\mathbf{y}$ implies that

$$\mathbf{a}_{\mathrm{o}} = \alpha \cdot \mathbf{H}_{\mathrm{o}}\mathbf{y}. \tag{23}$$

In addition, $\bar{\mathbf{H}}\mathbf{y} = \mathbf{0}$ implies that the attack that modifies the data from $\mathcal{C}$ by adding the corresponding entries of $\mathbf{a}_{\mathrm{o}}$ to the actual data is equivalent to using $\alpha \cdot \mathbf{Hy}$ as an attack vector, which is unobservable. So, the attack is unobservable. ■

Appendix D
Proof of Theorem 4.1

The normalized residues in the first iteration are affected by the attack $\mathbf{a}$ as follows:

$$\tilde{\mathbf{r}} = \mathbf{\Omega W}(\mathbf{z} + \mathbf{a}) = \mathbf{\Omega W e} + \mathbf{\Omega W a}, \tag{24}$$

which can be derived from (7) and (11). Note that $(\mathbf{\Omega We})_i$ follows a standard normal distribution (due to the normalization) if $\{i\}$ is not a critical set; $(\mathbf{\Omega We})_i$ is zero otherwise. Therefore, $\tilde{\mathbf{r}}_i$ follows the normal distribution $\mathcal{N}((\mathbf{\Omega Wa})_i, 1)$ if $\{i\}$ is not a critical set; otherwise, $\tilde{\mathbf{r}}_i$ is equal to $(\mathbf{\Omega Wa})_i$.

Therefore, the expected energy of the normalized residues at $\mathcal{S}_{\mathrm{F}}$ in the presence of the attack $\mathbf{a}$ is

$$\mathbb{E}\left[\sum_{i\in\mathcal{S}_{\mathrm{F}}} (\tilde{r}_i)^2\right] = \sum_{i\in\mathcal{S}_{\mathrm{F}}} (\mathbf{\Omega Wa})_i^2 + C = \|\mathbf{I}_{\mathcal{S}_{\mathrm{F}}}\mathbf{\Omega Wa}\|_2^2 + C, \tag{25}$$

where C is the number of sensors in $\mathcal{S}_{\mathrm{F}}$ that do not form a single element critical set.

Consequently, a solution to (17) is also a solution to the following problem, and vice versa:

$$\begin{aligned} \max_{\mathbf{a}} \quad & \|\mathbf{I}_{\mathcal{S}_{\mathrm{F}}}\mathbf{\Omega Wa}\|_2^2 \\ \text{subj.} \quad & \|\mathbf{a}\|_2^2 = 1, \quad \mathbf{a} \in \mathcal{R}(\mathbf{H}_1) \cap \mathcal{A}, \end{aligned} \tag{26}$$

The theorem statements follow from the following observations: $\mathbf{W}$ is equal to $\tilde{\mathbf{W}}$ as both are orthogonal projections on the same space, and $\mathcal{R}(\mathbf{H}_1)$ is equivalent to $\mathcal{R}(\mathbf{U}_1)$. ■

APPENDIX E
PROOF OF THEOREM 4.2

Let $\bar{\mathbf{H}}$ denote the submatrix of $\mathbf{H}$ obtained by removing the rows corresponding to the sensors in $\mathcal{C}$. First, from the proof procedure of Theorem 3.2, one can derive that the dimension of $\mathcal{N}(\bar{\mathbf{H}})$ is one. This implies that $\mathcal{C}$ contains exactly one critical set. Because, if there were more than one critical sets included in $\mathcal{C}$, $\mathcal{N}(\bar{\mathbf{H}})$ should have a dimension larger than one.

Because $\mathcal{S}_{\mathrm{A}} \cup \mathcal{S}_{\mathrm{F}} = \mathcal{C}$ contains exactly one critical set, the dimension of $\mathcal{R}(\mathbf{H}_1) \cap \mathcal{A}$ in (17) is one. This can be seen as follows. The dimension of $\mathcal{R}(\mathbf{H}_1) \cap \mathcal{A}$ in (17) is equal to the dimension of $\mathcal{N}(\mathbf{H}_2)$ where $\mathbf{H}_2$ is the matrix obtained from $\mathbf{H}$ by removing the rows corresponding to the sensors in $\mathcal{S}_{\mathrm{A}} \cup \mathcal{S}_{\mathrm{F}}$. And, the fact that $\mathcal{S}_{\mathrm{A}} \cup \mathcal{S}_{\mathrm{F}}$ contains exactly one critical set implies that the rank of $\mathbf{H}_2$ is $n-1$, and thus the dimension of $\mathcal{N}(\mathbf{H}_2)$ is 1.

Therefore, (17) has only two feasible points, and they give the same objective function values. In particular, a solution to (17) is the direction given by $\mathbf{H}_1\Delta\mathbf{x}$ where $\Delta\mathbf{x}$ is a nonzero vector in $\mathcal{N}(\mathbf{H}_2)$ (see [24] for more detailed arguments.)

The first and second conditions of Theorem 3.2, which are assumed to hold, imply that the dimension of $\mathcal{N}(\bar{\mathbf{U}}_{\mathrm{A}})$ is one. In addition, it can be seen from Corollary 3.2.1 that the second statement is true for $\mathbf{a}^* = \mathbf{H}_1\Delta\mathbf{x}$ and some nonzero α. ■

APPENDIX F
POWER GRID MEASUREMENT MODEL AND OBSERVABILITY

In this section, we briefly describe the power system measurement model and the spanning-tree observability criterion in [20]. The spanning-tree observability criterion results in Corollary 3.2.2 from Theorem 3.2. For more details about power system models, see [27].

The power system state is defined as the vector of voltage magnitudes and phase angles at all buses except a reference bus, which is an arbitrary bus whose voltage phase angle is set to zero:

$$\mathbf{x} = [V_1\ V_2\ \cdots\ V_n\ \theta_2\ \cdots\ \theta_n]^T \tag{27}$$

where V_i and θ_i denote the voltage magnitude and phase angle at bus i respectively, and bus 1 is set as the reference bus.

We consider two types of legacy sensors: line flow sensors and bus injection sensors.[8] The line flow from bus i to bus j is a complex quantity related to the system state by

$$P_{ij} + \mathrm{j} \cdot Q_{ij} = V_i e^{\mathrm{j}\theta_i} \cdot \left(\frac{V_i e^{\mathrm{j}\theta_i} - V_j e^{\mathrm{j}\theta_j}}{Z_{ij}} \right)^* \tag{28}$$

where $P_{ij} \in \mathbb{R}$ and $Q_{ij} \in \mathbb{R}$ are real and imaginary parts of the line flow respectively, Z_{ij} is the impedance of the line $\{i, j\}$, and X^* denotes the complex conjugate of X. The bus injection at bus i is the sum of all outgoing line flows from bus i.

For computational benefits, the above nonlinear relation is often linearized at the nominal operating point where all bus voltage magnitudes are equal to 1 p.u., and all bus voltage phase angles are equal to zero. This linearization decouples the relation such that the real part of measurements depends only on the voltage phase angles, and the imaginary part depends only on the voltage magnitudes.

The linearized relation between the real part of measurements and the voltage phase angles—the so-called DC model—is often used to analyze power system observability. In the DC model (4), the state $\mathbf{x}$ is defined as the vector of voltage phase angles at all buses except the reference bus:

$$\mathbf{x} = [\theta_2\ \theta_3\ \cdots\ \theta_n]^T. \tag{29}$$

The measurement matrix $\mathbf{H}$ depends on the topology and line impedance.[9]

The power system is observable if and only if $\mathbf{H}$ has full column rank [20]. Verifying this rank condition seems to require knowledge of the line impedance. However, Krumpholz *et al.* [20] showed that system observability can be determined purely based on the topology and sensor locations. In particular, Krumpholz *et al.* [20] showed that a system is observable if and only if there exists a way to assign each injection sensor to any of the lines that are incident to the bus where the sensor is located such that there exists a spanning tree of the topology having at least one sensor (an assigned injection or line flow sensor) on each edge of the tree (see Corollary 2 in [20].)

The spanning tree criterion can also be used to check whether the state variables in $\mathcal{X}_{\mathrm{o}}$ are observable with respect to $\mathcal{S}_{\mathrm{o}}$ (we use the notations in Section III-B.) Without loss of generality, we assume that $\mathcal{S}_{\mathrm{o}}$ contains an injection sensor on the reference bus or a line flow sensor on a line incident to the reference bus.[10] Then, we can simply apply the spanning tree criterion to the reduced network for $\mathcal{S}_{\mathrm{o}}$ (see Section III-B for the definition of a reduced network.) The state variables in $\mathcal{X}_{\mathrm{o}}$ are observable with respect to $\mathcal{S}_{\mathrm{o}}$ if and only if it is possible to assign injection sensors in $\mathcal{S}_{\mathrm{o}}$ to their neighboring lines such that a spanning tree of the reduced network with at least one sensor in $\mathcal{S}_{\mathrm{o}}$ on every edge exists.

[8]Other types of sensors (e.g., phasor measurement units) can also be considered. We impose this restriction merely to facilitate clearer presentation.

[9]To describe the entries of $\mathbf{H}$, we consider a noiseless measurement vector $\mathbf{z} = \mathbf{H}\mathbf{x}$ for simplicity. Suppose that the kth entry of $\mathbf{z}$ is a measurement from a line flow sensor measuring the line flow from bus i to j. Then, if the line is *connected*, $z_k = B_{ij}(\theta_i - \theta_j)$, where B_{ij} is the susceptance of the line; if the line is *not* connected, $z_k = 0$. In case that z_k corresponds to an injection sensor at bus i, z_k is the sum of all the outgoing line flows from bus i.

[10]Note that we can choose the reference bus such that this condition holds.

REFERENCES

[1] E. A. Lee, "Cyber physical systems: Design challenges," EECS Dept., Univ. of Calif., Berkeley, CA, USA, Tech. Rep. UCB/EECS-2008-8, Jan. 2008 [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-8.html

[2] Y.-F. Huang, S. Werner, J. Huang, N. Kashyap, and V. Gupta, "State estimation in electric power grids: Meeting new challenges presented by the requirements of the future grid," *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 33–43, Sep. 2012.

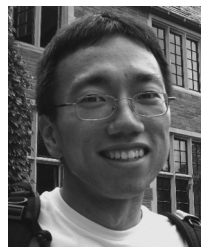
[3] A. Cardenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, and S. S. Sastry, "Challenges for securing cyber physical systems," in *Proc. Workshop Future Direct. Cyber-Phys. Syst. Secur.*, Jul. 2009, DHS.

[4] J. Hull, H. Khurana, T. Markham, and K. Staggs, "Staying in control: Cybersecurity and the modern electric grid," *IEEE Power Energy Mag.*, vol. 10, no. 1, pp. 41–48, 2012.

[5] "Vulnerability analysis of energy delivery control systems," Idaho Nat. Lab., INL/EXT-10-18381, Sep. 2011.

[6] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *Proc. 16th ACM Conf. Comput. Commun. Secur.*, 2009, pp. 21–32.

[7] R. B. Bobba, K. M. Rogers, Q. Wang, H. Khurana, K. Nahrstedt, and T. J. Overbye, "Detecting false data injection attacks on dc state estimation," in *Proc. 1st Workshop on Secure Control Syst. (CPSWEEK 2010)*, Stockholm, Sweden, Apr. 2010.
[8] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Trans. Smart Grid*, vol. 2, no. 4, pp. 645–658, Dec. 2011.
[9] T. Kim and H. Poor, "Strategic protection against data injection attacks on power grids," *IEEE Trans. Smart Grid*, vol. 2, no. 2, pp. 326–333, June 2011.
[10] S. Bi and Y. Zhang, "Defending mechanisms against false-data injection attacks in the power system state estimation," in *Proc. IEEE GLOBECOM Workshops*, Houston, TX, USA, Dec. 2011.
[11] A. Giani, E. Bitar, M. Garcia, M. McQueen, P. Khargonekar, and K. Poolla, "Smart grid data integrity attacks: Characterizations and countermeasures," in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Oct. 2011, pp. 232–237.
[12] J. Kim and L. Tong, "On phasor measurement unit placement against state and topology attacks," in *IEEE Int. Conf. Smart Grid Commun.*, Oct. 2013.
[13] M. Esmalifalak, H. Nguyen, R. Zheng, and Z. Han, "Stealth false data injection using independent component analysis in smart grid," in *Proc. IEEE Int. Conf. Smart Grid Commun.*, Oct. 2011, pp. 244–248.
[14] M. Rahman and H. Mohsenian-Rad, "False data injection attacks with incomplete information against smart power grids," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2012.
[15] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, Cramer-Rao bound," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 37, no. 5, pp. 720–741, 1989.
[16] A. Pezeshki, B. Van Veen, L. Scharf, H. Cox, and M. Nordenvaad, "Eigenvalue beamforming using a multirank mvdr beamformer and subspace selection," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1954–1967, 2008.
[17] M. Viberg, "Subspace-based methods for the identification of linear time-invariant systems," *Automatica*, vol. 31, no. 12, pp. 1835–1851, 1995.
[18] Power Systems Test Case Archive [Online]. Available: http://www.ee.washington.edu/research/pstca/
[19] S. Bi and Y. J. Zhang, "Graphical methods for defense against false-data injection attacks on power system state estimation," *IEEE Trans. Smart Grid*, vol. 5, no. 3, pp. 1216–1227, May 2014.
[20] G. R. Krumpholz, K. A. Clements, and P. W. Davis, "Power system observability: A practical algorithm using network topology," *IEEE Trans. Power Apparat. Syst.*, vol. 99, no. 4, pp. 1534–1542, Jul. 1980.
[21] A. Giani, E. Bitar, M. Garcia, M. McQueen, P. Khargonekar, and K. Poolla, "Smart grid data integrity attacks," *IEEE Trans. Smart Grid*, vol. 4, no. 3, pp. 1244–1253, Sep. 2013.
[22] J. Kim and L. Tong, "On topology attack of a smart grid: Undetectable attacks and countermeasures," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 7, Jul. 2013.
[23] H. Sandberg, A. Teixeira, and K. H. Johansson, "On security indices for state estimators in power networks," in *Proc. 1st Workshop on Secure Control Syst. (CPSWEEK 2010)*, Stockholm, Sweden, Apr. 2010.
[24] J. Kim, L. Tong, and R. J. Thomas, "Data framing attack on state estimation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 7, pp. 1460–1470, Jul. 2014.
[25] Y. Mo and B. Sinopoli, "False data injection attacks in control systems," in *Proc. 1st Workshop Secure Control Syst. (CPS Week)*, 2010.
[26] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 58, no. 11, pp. 2715–2729, Nov. 2013.
[27] A. Abur and A. G. Expósito, *Power System State Estimation: Theory and Implementation*. Boca Raton, FL, USA: CRC, 2000.
[28] L. Jia, J. Kim, R. Thomas, and L. Tong, "Impact of data quality on real-time locational marginal price," *IEEE Trans. Power Syst.*, vol. 29, no. 2, pp. 627–636, Mar. 2014.
[29] E. Handschin, F. C. Schweppe, J. Kohlas, and A. Fiechter, "Bad data analysis for power system state estimation," *IEEE Trans. Power App. Syst.*, vol. PAS-94, no. 2, pp. 329–337, Mar./Apr. 1975.
[30] A. Srivastava, "A Bayesian approach to geometric subspace estimation," *IEEE Trans. Signal Process.*, vol. 48, no. 5, pp. 1390–1400, May 2000.
[31] S. T. Smith, "Covariance, subspace, intrinsic Cramer Rao bounds," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1610–1630, May 2005.
[32] T. W. Anderson, "Asymptotic theory for principal component analysis," *Ann. Math. Statist.*, vol. 34, no. 1, pp. 122–148, 1963.
[33] P. Billingsley, *Probability and Measure*. New York, NY, USA: Wiley, 1995.
[34] J. Shao, *Mathematical Statistics*. New York, NY, USA: Springer-Verlag, 2003.



Jinsub Kim (M'14) received the Ph.D. degree in electrical and computer engineering with minors in applied mathematics and statistics from Cornell University, Ithaca, NY.

He is an Assistant Professor of the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis. His research interest spans statistical inference, learning, and optimization for a smart grid and network security. His graduate study was partially supported by the Samsung Scholarship. Before joining Oregon State University, he was a Postdoctoral Associate at Cornell University.

Lang Tong (S'87–M'91–SM'01–F'05) received the B.E. degree in automation from Tsinghua University, Beijing, China, and the Ph.D. degree in electrical engineering from the University of Notre Dame, Notre Dame, IN.

He is the Irwin and Joan Jacobs Professor in Engineering at Cornell University, Ithaca, NY. He is also the Cornell site director of the Power System Engineering Research Center (PSERC). His current research focuses on inference, optimization, and economic problems in energy and power systems. He is also a coauthor of seven student paper awards.

Dr. Tong received the 1993 Outstanding Young Author Award from the IEEE Circuits and Systems Society, the 2004 Best Paper award from IEEE Signal Processing Society, and the 2004 Leonard G. Abraham Prize Paper Award from the IEEE Communications Society. He received Young Investigator Award from the Office of Naval Research. He was a Distinguished Lecturer of the IEEE Signal Processing Society.

Robert J. Thomas (M'73–SM'83–F'93–LF'08) currently holds the position of Professor Emeritus of Electrical and Computer Engineering at Cornell University, Ithaca, NY. He has had assignments with the U.S. Department of Energy Office of Electric Energy Systems (EES) in Washington, DC and the National Science Foundation as the first Program Director for the Power Systems Program in the Engineering Directorate's Division of Electrical Systems Engineering (ESE). He is the author of more than 100 technical papers and two book chapters. He has published in the areas of transient control and voltage collapse problems as well as technical, economic and institutional impacts of restructuring.

Dr. Thomas is the founding Director of the 13-university member National Science Foundation Center, Power Systems Engineering Research Center (PSerc). He was a member of the USDOE Secretary's Power Outage Study Team (POST) and is a founding member of the Coalition for Electric Reliability Solutions (CERTS). He has received five teaching awards and the IEEE Centennial and Millennium medals. He is a member of Tau Beta Pi, Eta Kappa Nu, Sigma Xi, ASEE, and a Cornell ACSF Fellow.