Data-Driven False Data Injection Attacks against Power Grids: A Random Matrix Approach

Item Type	Article
Authors	Lakshminarayana, Subhash; Kammoun, Abla; Debbah, Merouane; Poor, H. Vincent
Citation	Lakshminarayana, S., Kammoun, A., Debbah, M., & Poor, H. V. (2021). Data-Driven False Data Injection Attacks Against Power Grids: A Random Matrix Approach. IEEE Transactions on Smart Grid, 12(1), 635–646. doi:10.1109/tsg.2020.3011391
Eprint version	Post-print
DOI	10.1109/TSG.2020.3011391
Publisher	Institute of Electrical and Electronics Engineers (IEEE)
Journal	IEEE Transactions on Smart Grid
Rights	(c) 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/ republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.
Download date	28/08/2022 01:37:23
Link to Item	http://hdl.handle.net/10754/661757

Data-Driven False Data Injection Attacks Against Power Grid: A Random Matrix Approach

Subhash Lakshminarayana *Member, IEEE*, Abla Kammoun *Member, IEEE*, Merouane Debbah *Fellow, IEEE* and
H. Vincent Poor *Fellow, IEEE*

Abstract—We address the problem of constructing false data injection (FDI) attacks that can bypass the bad data detector (BDD) of a power grid. The attacker is assumed to have access to only power flow measurement data traces (collected over a limited period of time) and no other prior knowledge about the grid. Existing related algorithms are formulated under the assumption that the attacker has access to measurements collected over a long (asymptotically infinite) time period, which may not be realistic. We show that these approaches do not perform well when the attacker has a limited number of data samples only. We design an enhanced algorithm to construct FDI attack vectors in the face of limited measurements that can nevertheless bypass the BDD with high probability. The algorithm design is guided by results from random matrix theory. Furthermore, we characterize an important trade-off between the attack's BDD-bypass probability and its sparsity, which affects the spatial extent of the attack that must be achieved. Extensive simulations using data traces collected from the MATPOWER simulator and benchmark IEEE bus systems validate our findings.

I. INTRODUCTION

Information and communication technologies (ICTs) play a key role in reducing costs and improving the quality of service in critical infrastructures such as the power grid. However, they also make the infrastructures vulnerable to cyber attacks, which may cause widespread damage as witnessed in a recent attack against the Ukraine power grid [2]. Hence, it is critical to assess the vulnerabilities of ICT-enabled critical infrastructures and devise ways to protect them.

In this work, we study the problem of constructing false data injection (FDI) attacks against state estimation (SE) in a power grid from an attacker's perspective. It has been shown [3] that if the attacker obtains detailed knowledge of the power grid topology and transmission line reactance values – i.e., the system's *measurement matrix* – then he can construct FDI attacks that bypass the grid's bad data detector (BDD). Subsequent research [4], [5], [6], [7] has shown that an attacker can learn the power grid's measurement matrix [4], or learn the structure of its column space by estimating the basis vectors [5], [6], [7] from accessed measurement data (i.e., nodal power injections and line power flows) only. The focus of our work is on constructing these *data-driven* FDI attacks.

S. Lakshminarayana is with the University of Warwick, Coventry, UK (email: subhash.lakshminarayana@warwick.ac.uk). A. Kammoun is with the Electrical Engineering Department, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia (abla.kammoun@gmail.com). M. Debbah is with the Mathematical and Algorithmic Sciences Lab, Huawei Technologies Co. Ltd., France (merouane.debbah@huawei.com). H. Vincent Poor is with the Department of Electrical Engineering, Princeton University, USA (email: poor@princeton.edu).
The work was partially presented at ICASSP-2018 [1].

Prior work on designing data-driven BDD-bypass attacks [5], [6], [7] has studied the setting of a long measurement period encompassing (asymptotically infinitely) many samples. These works are based on *principal component analysis* (PCA), whose basic idea is to use the *sample covariance matrix* to identify the eigen modes along which the data exhibits the greatest variance. It is well known that PCA performs efficiently when the measurement period is significantly large compared to the signal dimension [8]. Thus, FDI attacks constructed using these techniques perform well when the attacker has access to measurements from a large time window.

However, for practical purposes, it is important to understand these attacks under a limited measurement time window. The reasons include (i) active topology control [9] or renewable energy integration [10] that leads to an inherently dynamic operating environment, thereby rendering measurements outdated and irrelevant after some time; and (ii) an attacker's desire or need (e.g., due to limited resources or limited exploitation time windows) to launch the attack quickly. Thus, in a practical scenario, the measurement time period may not be asymptotically large compared to the signal dimension, especially for large power grids (refer to the example presented in Section III-B). It has been demonstrated that under such a regime, the principal component estimated by PCA is inconsistent [11]. Indeed, our experiments show that FDI attacks constructed by the existing PCA-based algorithms [5], [6], [7] do not perform well (in terms of the BDD-bypass probability) when applied in a limited measurement period setting.

To address these shortcomings, in this paper, we analyze the problem of finding BDD-bypassing attack using measurement data collected from a limited time window (comparable to the measurement signal dimension), and identify guiding principles for the solution in this context. Under the limited measurement period setting, a key issue is that only a few eigen modes can be reliably estimated from the sample covariance matrix. This number in turn depends on the length of the measurement period relative to the signal dimension. To bypass the BDD with a high probability, it is important for the attacker to identify these critical eigen modes. Direct application of the PCA method as in [5], [6], [7] does not use this knowledge, and hence, performs poorly. In this work, we propose an enhanced algorithm to construct FDI attacks in the face of limited measurements that can nevertheless bypass the BDD with high probability.

Our algorithm design is based on results from random

matrix theory (RMT). The application is based on an important observation that the power grid's SE utilizes several redundant sensor measurements to filter the effect of measurement noise and obtain an optimal estimate on the system state [12], [13]. In other words, the dimension of the measurement vector is much greater than the size of the system state. Under this setting, the covariance matrix of the sensor measurements has a structure similar to the so-called "spiked models" in RMT [14], [15], which comprises of a low-rank perturbation of a scaled identity matrix. Here, the leading few eigen modes correspond to the subspace spanned by the signal (i.e., system state), where as the bulk of the eigen modes (corresponding to the redundant measurements) correspond to the noise subspace. For data obtained from the spiked model, RMT results can be used to characterize the number of eigen modes that can be estimated accurately as a function of the measurement time window, as well as the corresponding estimation accuracy [14], [15]. Using these results, the attacker can carefully design the attack vector by restricting it to a lower-dimensional subspace that is spanned by the accurately estimated eigen modes only, and bypass the BDD with a high probability. Otherwise, the inaccurately estimated basis vectors may mislead the attack vector to a subspace that is different from the intended one, thereby risking detection by the BDD.

However, restricting the attack vector to a lower-dimensional subspace of the estimated column space limits the attacker's freedom in crafting the FDI attack. In particular, a resource-constrained attacker may wish to minimize the number of the meters that must be compromised, or equivalently find a sparsest attack vector in the execution [16], [17]. Clearly, maximizing the sparsity of the attack vector is best achieved if we have an unconstrained choice of this vector over the full estimated column space of the measurement matrix. Hence, the attacker faces a fundamental tradeoff. On the one hand, as we observed, restricting the attack vector to a lower-dimensional subspace (spanned by the accurately estimated basis vectors) will enhance the BDD-bypass probability under limited measurements; i.e., the restriction makes the attack efficient temporally. On the other hand, this restriction may reduce the sparsity of the optimized attack vector, thus making it less efficient spatially. To understand the tradeoffs between the conflicting objectives, we compute the sparsest attack vector while constraining it to subspaces of varying lower dimensions of the full estimated column space.

To summarize, the contributions of this work are as follows.

- We propose an enhanced algorithm to construct data-driven FDI attacks in the limited measurement period setting that can bypass the BDD with high probability using results from RMT.
- We characterize an important trade-off between the FDI attack's BDD-bypass probability and the number of power meters in the grid that the attacker has to compromise in achieving the attack.
- We illustrate the fundamental trade-off by performing extensive simulations using benchmark IEEE bus systems. The results show that the attacker can significantly enhance the BDD-bypass probability using our proposed approach.

Our results provide important understanding about the design of FDI attacks by a temporal and/or spatial resource-limited attacker against power systems. We note that although this work analyzes the problem from an attacker's point of view, our results have practical relevance in the design of defense strategies as well. For instance, moving target defense (MTD) has been recently proposed as an effective strategy to thwart FDI attacks by invalidating the attacker's existing knowledge of the grid [18]. The frequency of MTD perturbations depends on the attacker's ability to learn the system parameters [18], since the system must be perturbed before the attacker can gather sufficient information to learn the MTD perturbations. In this context, our analysis presents a fundamental understanding of the attacker's capabilities in designing FDI attacks by accessing the system measurements.

To the best of our knowledge, this work is the first to apply RMT results in the context of smart grids. While RMT results have found wide application in other domains such as wireless communications, finance, physics etc. (we refer the reader to reference [19], Chapter 1 for a comprehensive review of RMT applications), its application to smart grids has been scarce. In particular, the application of the RMT spiked model results to FDI attack construction is novel and has not been considered previously, and this is a key contribution of our work.

The rest of the paper is organized as follows. We describe the system model in Section II. We review existing subspace method based algorithms to construct data-driven FDI attacks and point our their drawbacks in Section III. In Section IV, we present data-driven FDI attacks with limited number of measurements using RMT results and analyze its performance. The trade-offs in data-driven FDI attacks are discussed in Section V. The simulation results are presented in Section VI. Finally, conclusions are drawn in Section VII. The technical proofs are omitted due to the lack of space and can be found in the supplementary material.

II. SYSTEM MODEL

We consider a power grid that is characterized by a set of buses $\mathcal{N} = \{0, 1, 2, \ldots, N\}$ and transmission lines $\mathcal{L} = \{1, 2, \ldots, L\}$. The node with index 0 is used to represent the reference node. The grid is assumed to operate in a time slotted manner indexed by $t = 1, 2, \ldots, T$. To model power flows within the grid, we adopt the direct current (dc) power flow model [13]. Under this model, the system state corresponds to the nodal voltage phase angles, which we denote by $\boldsymbol{\theta}[t] = [\theta_1[t], \ldots, \theta_N[t]]^T$; i.e., $\theta_i[t], i \in \mathcal{N}$ is the voltage phase angle at bus i during the time slot t. For the reference bus, $\theta_0[t] = 0, \forall t$. We assume that the system state fluctuates around a mean value, e.g., due to the temporal variations of the load. Thus, $\boldsymbol{\theta}[t] = \bar{\boldsymbol{\theta}} + \boldsymbol{\epsilon}[t]$, where $\boldsymbol{\epsilon}[t]$ is assumed to be an independent and identically distributed (i.i.d.) random vector (across time) whose fourth order moment is bounded and its covariance matrix is given by $\sigma_\theta^2 \mathbf{I}$, (where $\mathbf{I}$ denotes an identity matrix). Here in, $\bar{\boldsymbol{\theta}}$ represents the bus voltage phase angles due at a base load (e.g., obtained by solving the optimal power flow considering a base load). The temporal independence assumption of the system state

fluctuations can be met by taking measurements with sufficient load/angle variations over time.

State Estimation & Bad Data Detection: The system state $\boldsymbol{\theta}[t]$ is monitored using sensors deployed at the buses and transmission lines. These sensors measure respectively the nodal power injections and the forward/reverse line power flows. Under the DC power flow model, these measurements, which we denote by $\boldsymbol{z}[t] \in \mathbb{R}^M$ (where M denotes the number of measurements), are related to the system state $\boldsymbol{\theta}[t] \in \mathbb{R}^N$ as

$$\boldsymbol{z}[t] = \mathbf{H}\boldsymbol{\theta}[t] + \mathbf{n}[t], \quad t = 1, 2, \cdots, T, \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{M \times N}$ is the measurement matrix and $\mathbf{n}[t]$ is the sensor measurement noise. The noise is assumed to be zero-mean Gaussian with covariance matrix $\sigma_n^2 \mathbf{I}$, and independent of the system state $\boldsymbol{\theta}[t]$. It is also assumed to be i.i.d. across the time slots. Without the loss of generality we set $\sigma_n^2 = 1$ throughout the paper, and scale the covariance of the $\boldsymbol{\theta}[t]$ appropriately (i.e., we set $(\sigma_\theta')^2 = (\sigma_\theta^2/\sigma_n^2)$ in our analysis). The measurement matrix $\mathbf{H}$ depends on the system topology (i.e., the bus connectivity) and the branch reactances [13]. We assume that within the considered time interval T, $\mathbf{H}$ does not change. The estimate of the system state, denoted by $\widehat{\boldsymbol{\theta}}[t]$, is recovered from the measurement vector $\boldsymbol{z}[t]$ using a maximum-likelihood (ML) technique [12]: $\widehat{\boldsymbol{\theta}}[t] = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\boldsymbol{z}[t]$.

After state estimation, the residual vector is given by $\mathbf{r}[t] = \boldsymbol{z}[t] - \mathbf{H}\widehat{\boldsymbol{\theta}}[t]$. The BDD checks for possible measurement inconsistencies in $\boldsymbol{z}[t]$ works by comparing the norm of the residual vector $r = ||\mathbf{r}[t]||^2$ against a pre-defined threshold τ. It raises an alarm if $r \geq \tau$. The threshold τ is selected to ensure a certain false-positive (FP) rate.

Attacker Model: We consider an attacker who can eavesdrop on the measurement data communicated between the field devices and the control center by exploiting vulnerabilities in the communication system. However, the attacker is assumed to be unaware of the semantics of the accessed data. Furthermore, the attacker has no other information about the grid (e.g., its topology or bus system).

The attacker's objective is to craft FDI attacks against the state estimation. Denote the attack vector by $\mathbf{a}[t] \in \mathbb{R}^M$, the sensor measurements under attack by $\mathbf{z}_a[t]$, where $\mathbf{z}_a[t] = \boldsymbol{z}[t] + \mathbf{a}[t]$, and the BDD residual under attack by $r_a[t] = ||\mathbf{z}_a[t] - \mathbf{H}\widehat{\boldsymbol{\theta}}_a[t]||^2$. It has been shown [3] that for an attack of the form $\mathbf{a}[t] = \mathbf{H}\mathbf{c}[t]$, the residual value remains unchanged under the attack, i.e., $r_a[t] = r[t]$. Hence, the BDD's detection probability for such attacks is no greater than the FP rate. We will henceforth refer to these attacks as *undetectable* attacks. Note that constructing an undetectable FDI attack requires the knowledge of $\mathbf{H}$. In data-driven FDI attack, the attacker strives to construct an undetectable FDI attack by learning the system parameters using the accessed measurement data.

III. SUBSPACE METHOD BASED ALGORITHM AND THE DRAWBACKS

In this section, we review existing subspace based approach for constructing undetectable data-driven FDI attacks [5], [6], [7], and point out its drawbacks under a practical regime of limited observation time window.

A. Algorithm Description

Note that designing an undetectable attack is equivalent to finding a non-zero vector in $Col(\mathbf{H})$, or equivalently, a linear combination of the basis vectors that span $Col(\mathbf{H})$. The attacker must estimate the basis vectors using the noisy measurement data $\mathbf{z}[t], \ t = 1, \ldots, T$. This problem is well studied in the signal processing literature [20], and has been used to guide the construction of data-driven FDI attacks.

The key idea is to use the covariance matrix of the measurements $\boldsymbol{\Sigma}_\mathbf{z} = \mathbb{E}[(\mathbf{z}[t] - \mathbb{E}[\mathbf{z}[t]])(\mathbf{z}[t] - \mathbb{E}[\mathbf{z}[t]])^T]$. From (1), it follows that

$$\boldsymbol{\Sigma}_\mathbf{z} = \sigma_\theta^2 \mathbf{H}\mathbf{H}^T + \mathbf{I}. \tag{2}$$

Let $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ be the SVD of $\boldsymbol{\Sigma}_\mathbf{z}$, where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_M]$, is a matrix consisting of the eigen vectors of $\boldsymbol{\Sigma}_\mathbf{z}$, and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_M)$ is a matrix consisting of the eigen values of $\boldsymbol{\Sigma}_\mathbf{z}$. Note that the rank of the matrix $\sigma_\theta^2 \mathbf{H}\mathbf{H}^T$ is N. Thus, the first N columns of $\mathbf{U}$ corresponding to the N largest singular values must form the basis vectors of $Col(\sigma_\theta^2 \mathbf{H}\mathbf{H}^T)$. Since, $Col(\sigma_\theta^2 \mathbf{H}\mathbf{H}^T)$ is equivalent to $Col(\mathbf{H})$, they also form the basis vectors of $Col(\mathbf{H})$ [20]. Thus, the attacker must estimate the eigen vectors of $\boldsymbol{\Sigma}_\mathbf{z}$ in order to construct an undetectable FDI attack vector.

We note that the attacker cannot directly execute the procedure stated above since the $\boldsymbol{\Sigma}_\mathbf{z}$ is unknown. However, it can be estimated using the measurement data $\{\mathbf{z}[t]\}_{t=1}^T$. Based on this observation, the procedure to construct data-driven FDI attacks is summarized in Alg.1. (We use the superscript $\widehat{}$ to denote estimates of the corresponding quantities. The notation $\mathbf{A}_s$ denotes a matrix consisting of the first s columns of the matrix $\mathbf{A}$, i.e., $\mathbf{A}_s = [\mathbf{a}_1, \ldots, \mathbf{a}_s]$, for any integer value s.).

ALGORITHM 1: Construction of Data-driven FDI attack

1. Using measurements $\{\mathbf{z}[1], \ldots, \mathbf{z}[T]\}$, compute the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$ as

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \frac{1}{T-1} \sum_{t=1}^{T} \left(\boldsymbol{z}[t] - \bar{\mathbf{z}}\right)\left(\boldsymbol{z}[t] - \bar{\mathbf{z}}\right)^T,$$

where $\bar{\mathbf{z}}$ denotes the sample mean given by $\bar{\mathbf{z}} = \frac{1}{T-1}\sum_{t=1}^{T}\boldsymbol{z}[t]$.

2. Perform singular value decomposition (SVD) of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$ as $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \widehat{\mathbf{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{U}}^T$, where $\widehat{\mathbf{U}} = [\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_M]$ and $\widehat{\boldsymbol{\Lambda}} = \text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_M)$.

3. Construct an undetectable FDI attack vector as $\mathbf{a}[t] = \widehat{\mathbf{U}}_N \mathbf{c}[t]$, where $\mathbf{c}[t] \in \mathbb{R}^N$.

B. Drawbacks of Existing Techniques

The aforementioned subspace estimation algorithm performs well in a classical setting when the number of temporal measurements are far greater than the system dimension (i.e., $T \gg M, M/T \to 0$). However, under a practical setting, it is unreasonable to expect the availability of an "infinite time window", especially for large bus systems. For instance, consider the IEEE-118 bus system which has $M = 490$ measurements per time slot (assuming a fully measured system). Under an
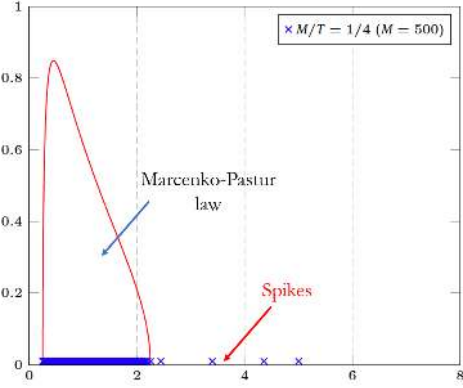
Fig. 1: Limting eigen value distribution of sample covariance matrix. Figure due to [19].

optimistic assumption of obtaining a temporally decorrelated measurement every minute, the attacker would require a measurement time window of 4900 minutes, or approximately, 80 hours, for the ratio of $M/T = 10$. However, the system topology may have changed well before this duration.

Thus, we focus on a practically relevant scenario, where the number of measurements M and the measurement time window T are reasonably large, where as their ratio of $M/T = p > 0$ is a finite constant. This scenario is especially relevant for large power grids. Under this regime, the principal components estimated by Algorithm 1 are known to be inconsistent [11]. Thus, in rest of the paper, we present an enhanced algorithm for strengthening the attack's BDD-bypass probability when the attacker has access to measurements from a limited time window. Furthermore, we characterize an important trade-off between the attack's BDD-bypass probability and the number of compromised measurements in executing the attack.

IV. DATA-DRIVEN FDI ATTACKS WITH LIMITED NUMBER OF MEASUREMENTS

In this section, we present an enhanced algorithm for data-driven FDI attacks based on RMT results [19]. From the discussion in Section III, note that the problem at hand is equivalent to estimating the principal eigen values/vectors of $\mathbf{\Sigma_z}$ from the corresponding sample covariance matrix $\widehat{\mathbf{\Sigma}}_\mathbf{z}$. In this section, we characterize analytical results that address this problem.

The covariance matrix of the measurements $\mathbf{\Sigma}_z$ can be expressed through eigen decomposition as

$$\mathbf{\Sigma}_z = \mathbf{I} + \sum_{i=1}^{N} \mu_i \mathbf{u}_i \mathbf{u}_i^H, \qquad (3)$$

where $\{\mu_i\}_{i=1}^N$ denote the eigen values of $\sigma_\theta^2 \mathbf{H}\mathbf{H}^H$ and $\{\mathbf{u}_i\}_{i=1}^N$ the corresponding eigen vectors. We will henceforth refer to $\mu_1, \ldots, \mu_N$ as the "spike" eigen values. The attacker must estimate the spikes and the corresponding eigen vectors from $\widehat{\mathbf{\Sigma}}_\mathbf{z}$ (which we will henceforth refer to as the sample covariance matrix and its eigen values are given by $\widehat{\lambda}_i$). This can be done as follows.

First, consider the case with no spikes, i.e., measurements generated as $\mathbf{z}[t] \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_z})$ where $\mathbf{\Sigma_z} = \mathbf{I}$. The eigen values of $\mathbf{\Sigma_z}$ are given by $[1, \ldots, 1] \in \mathbb{R}^M$. For this model, the limiting eigen value distribution[1] of $\widehat{\mathbf{\Sigma}}_\mathbf{z}$ is known to converge to the Marcenko-Pastur (MP) law [21]. We denote the limiting distribution by F, which is shown in Fig. 1. Now consider measurements whose covariance matrix given is by (3). The eigen values of $\mathbf{\Sigma_z}$ are given by

$$[\underbrace{\mu_1 + 1, \ldots, \mu_N + 1}_{N \text{ terms}}, \underbrace{1, \ldots, 1}_{M-N \text{ terms}}]. \qquad (4)$$

We focus on the limiting eigen value distribution of $\widehat{\mathbf{\Sigma}}_\mathbf{z}$ for this model (i.e., measurements from the covariance model (3)). Following (4), one would expect the "leading N eigen values" of $\widehat{\mathbf{\Sigma}}_\mathbf{z}$ to be found outside the distribution of MP law, F (see Fig. 1). Surprisingly, the number of eigen values that can be found outside F depends critically on the ratio $p = M/T$. This result was formalized in [14] and stated in Theorem 2, Appendix A. Herein, we only present the main idea. Consider

$$\mu_1 > \mu_2 > \cdots > \mu_s > \sqrt{p}, \qquad 1 \le s \le N, \qquad (5)$$

where $s \le N$ is the number of spike eigen values that are less than $\sqrt{p}$. Then, the result [14] states that for all $\mu_i > \sqrt{p}$, when $M, T \to \infty, M/T = p > 0$, there exists a deterministic and one-to-one mapping between eigen value of the sample covariance matrix ($\widehat{\mathbf{\Sigma}}_\mathbf{z}$), i.e., between $\widehat{\lambda}_i$ and μ_i. In other words, all μ_i which satisfy $\mu_i > \sqrt{p}$ can be recovered from the eigen values of sample covariance matrix. A similar result also holds for estimating the corresponding eigen vectors [15], i.e., the corresponding eigen vectors (for which $\mu_i > \sqrt{p}$) can be reliably recovered from the eigen vectors of the sample covariance matrix (see Theorem 2). It is important to note that for eigen modes corresponding to $\mu_i < \sqrt{p}$, these relationships do not hold, and the corresponding eigen value/vectors cannot be recovered. Thus, the quantity $\sqrt{p}$ represents a fundamental "phase transition" point in estimating the spike eigen values/vectors from the sample covariance matrix.

The result is important in this context, since it precisely characterizes the information about $Col(\mathbf{H})$ that the attacker can recover from the measurements as a function of the observation time window T (specifically, the ratio $p = M/T$). To construct a data-driven FDI that can bypass the BDD with a high probability, the attacker must first estimate the number of eigen values/vectors, s, that can be reliably recovered from the measurements $\{\mathbf{z}[t]\}_{t=1}^T$. Note that the attacker cannot directly use (5) to determine s, since he does not have the knowledge of μ_i. Using the result of Theorem 2 (Appendix A), it follows that for $\mu_i > \sqrt{p}$, with probability 1, we have $\widehat{\lambda}_i > (1+\sqrt{p})^2$. Thus the attacker can determine s by counting the number of eigen values of the sample covariance matrix that exceed $(1 + \sqrt{p})^2$, i.e.,

$$s = \{\#i, \widehat{\lambda}_i > (1 + \sqrt{p})^2\}. \qquad (6)$$

Note that direct application of subspace estimation algorithm as proposed in [5], [6], [7] (Algorithm 1) uses all N estimated

---

[1]Limiting eigen value distribution is the distribution of eigen values of $\widehat{\mathbf{\Sigma}}_\mathbf{z}$ when $M, T \to \infty, M/T = p > 0$.

eigen modes for the construction of the FDI attack. However, following the application of RMT spiked model results, it is clear that eigen modes for which $\mu_i < \sqrt{p}$ cannot be recovered from the sample covariance matrix, and hence, must not be used in the construction of FDI attack.

After determining s, the attacker can construct a data-driven FDI attack as $\mathbf{a} = \widehat{\mathbf{U}}_s \mathbf{c}_s$, where $\mathbf{c}_s \in \mathbb{R}^s$ denotes an $s-$dimensional vector. In particular, the vector $\mathbf{c}_s$ can be tuned by the attacker to achieve his objectives, such as minimizing the attack's detection probability or causing a desired attack impact. In this rest of this section, we describe how the attacker can achieve these objectives using results from RMT.

We first focus on the attack detection probability. To this end, we characterize the BDD residual with data driven FDI attack.

Lemma 1. *For a data-driven FDI attack $\mathbf{a} = \widehat{\mathbf{U}}_s \mathbf{c}_s$, the BDD residual with attack, r_a, follows a non-central χ^2 distribution with $M-N$ degrees of freedom and a non-centrality parameter ν given by*

$$\nu = \mathbf{c}_s^T \mathbf{c}_s - \mathbf{c}_s^T \widehat{\mathbf{U}}_s^T \mathbf{U}_N \mathbf{U}_N^T \widehat{\mathbf{U}}_s \mathbf{c}_s. \tag{7}$$

For $M, T \to \infty, M/T = p$ the second term of the right hand side of (7) *converges to*

$$\mathbf{c}_s^T \widehat{\mathbf{U}}_s^T \mathbf{U}_N \mathbf{U}_N^T \widehat{\mathbf{U}}_s \mathbf{c}_s - \mathbf{c}_s^T \mathbf{\Omega}_s \mathbf{c}_s \xrightarrow{\text{a.s.}} 0, \tag{8}$$

where $\mathbf{\Omega}_s = diag(\omega_1, \ldots, \omega_s)$ and

$$\omega_i = \frac{1 - p/\mu_i^2}{1 + p/\mu_i}, i = 1, \ldots, s. \tag{9}$$

Further, the attacker can obtain a consistent estimator $\widehat{\omega}_i$ of ω_i as $|\omega_i - \widehat{\omega}_i| \xrightarrow{\text{a.s.}} 0, i = 1, \ldots, s$ where,

$$\widehat{\omega}_i = \frac{1 - p/\widehat{\mu}_i^2}{1 + p/\widehat{\mu}_i}, i = 1, \ldots, s, \tag{10}$$

$$and \quad \widehat{\mu}_i = \frac{\widehat{\lambda}_i + 1 - p + \sqrt{(\widehat{\lambda}_i + 1 - p)^2 - 4\widehat{\lambda}_i}}{2} - 1.$$

Thus, it follows that

$$\nu - \mathbf{c}_s^T (\mathbf{I} - \widehat{\mathbf{\Omega}}_s) \mathbf{c}_s \xrightarrow{\text{a.s.}} 0, \tag{11}$$

where $\widehat{\mathbf{\Omega}}_s = diag(\widehat{\omega}_1, \ldots, \widehat{\omega}_s)$.

Proof. The proof is omitted due to the lack of space and can be found in Appendix A, Part I of the supplementary material. □

The result of (7) along with the asymptotic approximations (8)-(10) provides a tractable expression for the attacker to compute the detection probability for a given data-driven FDI attack $\mathbf{a} = \widehat{\mathbf{U}}_s \mathbf{c}_s$. Observe that all the quantities required to compute the asymptotic approximation of ν depend on the estimated parameters only (i.e., $\widehat{\lambda}_i$). Using these expressions, the attacker can tune $\mathbf{c}_s$ to minimize the attack's detection probability. We analyze the results further.

Note that the entries of the matrix $\widehat{\mathbf{U}}_s^T \mathbf{U}_N \mathbf{U}_N^T \widehat{\mathbf{U}}_s$ represent the projection of the eigen vectors of the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathbf{z}}$ onto the eigen vectors of the population covariance matrix $\mathbf{\Sigma}_{\mathbf{z}}$. In particular, the result of Lemma 1 states that

asymptotically, the estimated eigen vectors $\widehat{\mathbf{u}}_i$ are orthogonal to $\mathbf{u}_j, j \neq i$, since $\mathbf{\Omega}_s$ is diagonal. Specifically, $|\widehat{\mathbf{u}}_i^T \mathbf{u}_i|^2 \xrightarrow{\text{a.s.}} \omega_i$ and $|\widehat{\mathbf{u}}_i^T \mathbf{u}_j|^2 \xrightarrow{\text{a.s.}} 0, i \neq j$. The following lemma illustrates the relationship between the projections.

Lemma 2. *The diagonal elements $\{\omega_i\}_{i=1}^s$ and $\{\widehat{\omega}_i\}_{i=1}^s$ follow $1 > \omega_1 \geq \omega_2 \geq, \ldots, \geq \omega_s > 0$ and $1 > \widehat{\omega}_1 \geq \widehat{\omega}_2 \geq, \ldots, \geq \widehat{\omega}_s > 0$ respectively.*

Proof. The proof can be found in Appendix B of the supplementary material. □

From Lemma 1 and Lemma 2, it follows that the projection of $\widehat{\mathbf{u}}_i$ onto $\mathbf{u}_i$ is in the decreasing order of the eigen mode index. Note that minimizing the detection probability is equivalent to minimizing the non-centrality parameter ν of the χ^2 distribution. From (11), it follows the attacker can compute $\mathbf{c}_s$ which minimizes $\mathbf{c}_s^T (\mathbf{I} - \widehat{\mathbf{\Omega}}_s) \mathbf{c}_s$. However, directly minimizing this expression would result in a trivial solution $\mathbf{c}_s = \mathbf{0}$, (i.e., a zero attack). Thus, we must constrain the attack impact in order to obtain a meaningful attack.

We quantify the attack impact in terms of the second norm of the error in state estimate (for the system operator) due to the FDI attack. Specifically, we let $\widehat{\boldsymbol{\theta}}^a$ denote the estimate of the system state from measurements with FDI attack, $\mathbf{z}^a$. Then, $\Delta\boldsymbol{\theta} \triangleq \widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}^a$ is the error in the state estimate due to the FDI attack. Using this, the data-driven FDI attack can be formulated as the following optimization problem:

$$\min_{\mathbf{c}_s} \quad \mathbf{c}_s^T (\mathbf{I} - \widehat{\mathbf{\Omega}}_s) \mathbf{c}_s \tag{12}$$
$$\text{s.t.} \quad ||\Delta\boldsymbol{\theta}||_2^2 \geq \tau$$

In the optimization problem (12), the attacker designs $\mathbf{c}_s$ to minimize the probability of detection among all attacks that satisfy $||\Delta\boldsymbol{\theta}||_2^2 \geq \tau$. However, (12) cannot be solved by the attacker directly, as $||\Delta\boldsymbol{\theta}||_2^2$ depends on the measurement matrix $\mathbf{H}$, that is unknown to the attacker. To address this issue, we present a consistent estimate of $||\Delta\boldsymbol{\theta}||_2^2$ in the large system regime that depends only on the attacker's estimated parameters in the following lemma:

Lemma 3. *For $M, T \to \infty$, the quantity $||\Delta\boldsymbol{\theta}||_2^2$ converges to*

$$||\Delta\boldsymbol{\theta}||_2^2 - \sigma_\theta^2 \mathbf{c}_s^T \mathbf{M}^{-1} \mathbf{\Omega} \mathbf{c}_s \xrightarrow{\text{a.s.}} 0. \tag{13}$$

Further, we have

$$\sigma_\theta^2 \mathbf{c}_s^T \mathbf{M}^{-1} \mathbf{\Omega} \mathbf{c}_s - \widehat{\sigma}_\theta^2 \mathbf{c}_s^T \widehat{\mathbf{M}}^{-1} \widehat{\mathbf{\Omega}} \mathbf{c}_s \xrightarrow{\text{a.s.}} 0, \tag{14}$$

where $\mathbf{M} = diag(\mu_1, \ldots, \mu_N)$ and $\widehat{\mathbf{M}} = diag(\widehat{\mu}_1, \ldots, \widehat{\mu}_N)$.

Proof. The proof is presented in Appendix A, Part II of the supplementary material. □

Note from Lemma 3 that the asymptotic approximation of $||\Delta\boldsymbol{\theta}||_2^2$ depends on the estimate of the variance of the system state $\widehat{\sigma}_\theta^2$. The attacker can estimate this by monitoring historical fluctuations of the system load (note that this is a second-order statistic and hence, need not be estimated in real time).

Based on the result of Lemma 3, optimization problem (12) can be reformulated as follows:

$$\min_{\mathbf{c}_s} \quad \mathbf{c}_s^T(\mathbf{I} - \widehat{\boldsymbol{\Omega}}_s)\mathbf{c}_s \tag{15}$$
$$\text{s.t.} \quad \widehat{\sigma}_\theta^2 \mathbf{c}_s^T \widehat{\mathbf{M}}^{-1}\widehat{\boldsymbol{\Omega}}_s \mathbf{c}_s \geq \tau$$

The solution to (15) can be characterized in closed form and its result leads to the following theorem:

Theorem 1. *For $M, T \to \infty, M/T = p$, the optimal data-driven FDI attack that solves (15) is given by*

$$\mathbf{a} = \sqrt{\frac{\tau}{\widehat{\sigma}_\theta^2(\widehat{\omega}_1/\widehat{\mu}_1)}}\widehat{\mathbf{u}}_1. \tag{16}$$

Proof. The proof of Theorem 1 follows by noting that the solution to optimization problem (15) is given by $c_1 = \sqrt{\frac{\tau}{\widehat{\sigma}_\theta^2(\widehat{\omega}_1/\widehat{\mu}_1)}}$ and $c_2 = c_3 = \cdots = c_s = 0$. The details are presented in Appendix C of the supplementary material. □

Theorem 1 implies that the attacker can minimize the detection probability by aligning the attack vector along $\widehat{\mathbf{u}}_1$ while achieving the desired attack impact. Coincidentally, from Lemma 2, $\widehat{\mathbf{u}}_1$ is also the most accurately estimated eigen mode.

Theorem 1 also implies that the optimal attack must be restricted to a $1-$dimensional subspace of the estimated space $\widehat{\mathbf{U}}_N$. A natural question is whether there a cost to pay for this restriction? We will address this question in the next section where we consider the attack's sparsity in addition to the factors considered in this section.

A. Discussion

We note from Theorem 2 (Appendix A) that the application of spiked model requires N to be fixed and independent of M and T. Strictly speaking, the power grid model does not satisfy this condition, since N (dimension of the state vector) also grows for large grids. However, despite this limitation, we will show by simulations in Section VI that RMT spiked model results are accurate for various power grid bus configurations as long as the number of sensor measurements M is large compared to N. In other words, there are a significant number of redundant measurements, which is reasonable for the state estimation problem [12], [13]. Thus, the RMT spiked model can be used for analysing the data-driven FDI attacks.

V. TRADE-OFFS IN DATA-DRIVEN FDI ATTACKS

The analysis considered thus far in this paper only focusses on attacker's learning of $Col(\mathbf{H})$. The learning phase only requires the attacker to obtain *read* access to the sensor measurements. However, executing the FDI attack requires the attacker to modify the sensor measurements values, which in turn requires *write* access. A graphical illustration is presented in Fig. 2. Note that from an attacker's point of view, *read* access to sensor measurements is easier to obtain compared to *write* access, since it only involves passive sniffing of the network data, where as write access requires modification of the network packets. Thus, a resource constrained may wish to minimize the number of sensors he must compromise to
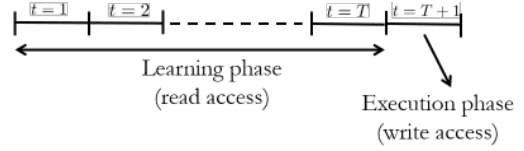

Fig. 2: Learning and execution phase of data-driven FDI attack.

execute the FDI attack, or equivalently, maximize the attack's sparsity.

Recall that from our analysis in the previous section, the optimal data-driven FDI attack is one that is restricted to a $1-$ dimensional subspace of the estimated space. Restricting the attack to a lower-dimensional subspace makes it hard to enforce sparsity. Naturally, maximum sparsity of the attack vector can be achieved if we have an unconstrained choice of the attack vector over the full estimated column space. On the other hand, using the innacurately estimated basis vectors will increase the attack's detection probability (recall that the first basis is estimated most accurately, followed by the second, etc. refer to Lemma 2). Thus, the attacker faces a fundamental trade-off between the attack's BDD-bypass probability and the attack's sparsity. In particular, the number of estimated eigen modes for constructing the FDI attack must be chosen to balance between the two factors.

To formalize this trade-off, we cast the FDI attack construction as a sparse optimization problem for all $m = 1, \ldots, s$ as follows:

$$K_m^* = \min_{\mathbf{c}_m} \quad \|\widehat{\mathbf{U}}_m \mathbf{c}_m\|_0, \tag{17}$$
$$\text{s.t.} \quad \mathbf{c}_m^T \widehat{\mathbf{M}}^{-1}\widehat{\boldsymbol{\Omega}}\mathbf{c}_m \geq \tau.$$

The objective function of (17) gives the number of non-zero elements in the FDI attack vector while restricting the attack vector to a $m-$dimensional subspace of the estimated column space, where $m \leq s$ ((6)). The optimization problem (17) can be solved using an l_1-relaxation based approach. We omit the details here and refer the reader to [17]. We illustrate the trade-off by simulations in Section VI.

VI. SIMULATION RESULTS

In this section, we present the simulation results. All the simulations are conducted using the MATPOWER simulator [22]. In these simulations, the eigen modes are estimated following the steps 1 and 2 of Algorithm 1. The measurement data is generated according to (1), where $\boldsymbol{\theta}[t] = \bar{\boldsymbol{\theta}} + \boldsymbol{\epsilon}[t]$. Here in, $\bar{\boldsymbol{\theta}}$ is obtained by solving the optimal power flow formulation considering base load values provided in the MATPOWER case file. The fluctuations $\boldsymbol{\epsilon}[t]$ are assumed to be i.i.d. Gaussian random vectors with variance σ_θ^2. The FDI attacks are constructed using the estimated eigen modes and their detection probability is computed by averaging the BDD's detection results over 1000 independent trials. The BDD threshold is adjusted such that the FP rate is set to 0.02. The results are presented next.

First, we examine the estimation accuracy of different eigen modes by evaluating the projection metric $|\widehat{\mathbf{u}}_i^T \mathbf{u}_i|^2$. We also verify the accuracy of the RMT approximation in Theorem 2.
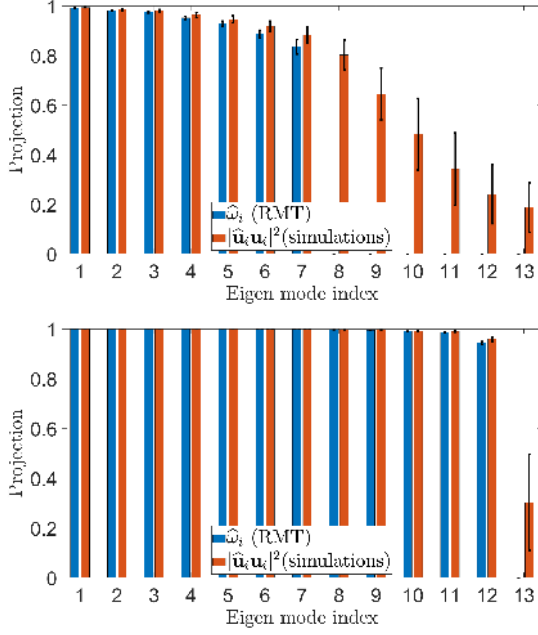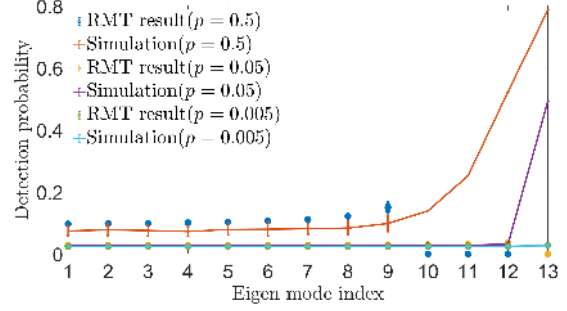
Fig. 4: Detection probability for FDI attacks constructed using different eigen modes. $s = 9$ for $p = 0.5$, $s = 12$ for $p = 0.05$, and $s = 13$ for $p = 0.005$ respectively.
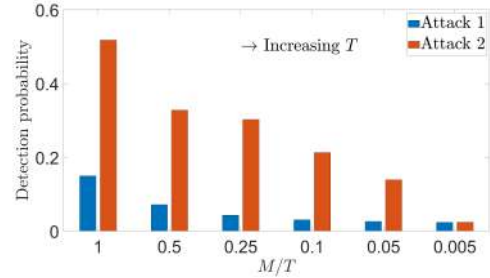

Fig. 5: Detection probability for FDI attacks as a function of the measurement time window. Attack 1 : optimal data-driven FDI attack (Theorem 1), Attack 2: FDI attack constructed using the entire estimated subspace.

Fig. 3: Eigen mode estimation accuracy using simulations and RMT approximation. Top: non-asymptotic regime, $p = 0.5$ ($s = 7$), Bottom: asymptotic regime, $p = 0.005$ ($s = 12$).

To this end, we compare $|\widehat{\mathbf{u}}_i^T \mathbf{u}_i|^2$ obtained from simulations with $\widehat{\omega}_i$ computed according to (10). For the measurement time window T, we consider two regimes, (i) a non-asymptotic regime with $T = 0.5M$ and (ii) an asymptotic regime with $T = 100M$. The results are plotted in Fig. 3 by averaging across 1000 trails. The bars represent mean values over the trials and the vertical lines (on top of the bars) represent the flutuation around this mean value. We make the following observations.

Firstly, in the non-asymptotic regime, the estimation accuracies of the different eigen modes vary. In particular, they are arranged in the decreasing order of the eigen mode index. This is consistent with our observation in Lemma 2. Secondly, it can be observed that the RMT approximations $\widehat{\omega}_i$ (red bars) are reasonably accurate, though there is a non-zero but negligible gap between the simulations and RMT results. The gap exists due to the fact that the number of spikes in the power grid model are large and equal to the dimension of the state vector (see the discussion in Section IV-A). However, despite this limitation, the gap is small and the RMT results are a good approximation. Thirdly, recall that RMT approximations only exist for the eigen modes $i \leq s$ where s is computed according to (8). The value of s for each of the simulation cases is indicated in the figure description. It can be observed that the estimation accuracy for eigen modes beyond this value of s is poor. Hence, they must not be utilized for FDI attack construction as prescribed by our analysis based on the RMT spiked model. Finally, we observe that in the asymptotic regime however, i.e. $T = 100M$, all the eigen modes can be estimated with a high accuracy and Algorithm 1 can be used directly for the design of FDI attack.

Next, we examine the detection probability of FDI attacks constructed using different estimated eigen modes. Specifi-

cally, for each estimated eigen mode i, the FDI attack is constructed as $\mathbf{a} = c_i \widehat{\mathbf{u}}_i$, where c_i is set to $c_i = \sqrt{\frac{\tau}{\widehat{\omega}_i / \widehat{\mu}_i}}$, such that is satisfies the constraint of (15). Recall that asymptotically this ensures that $\|\Delta \boldsymbol{\theta}\|_2^2 \geq \tau$. We conduct 1000 simulation trials and plot the results in Fig. 4. The RMT approximations of the detection probability are also plotted in Fig. 4. They are computed by evaluating the $\mathbb{P}(X \geq \tau)$, where X is a χ^2 distributed random variable with $M - N$ degrees of freedom and a non-centrality parameter $\nu = \mathbf{c}_s^T \widehat{\boldsymbol{\Omega}}_s \mathbf{c}_s$ (following the result of Lemma 1). Once again, we make the following observations. Firstly, the detection probability increases with the eigen mode index and the attack $\mathbf{a} = c_1 \widehat{\mathbf{u}}_1$ has the lowest detection probability, confirming the result of Theorem 1. Secondly, it can be observed for $i \geq s$, the detection probability becomes very high, thus confirming the phase transition phenomenon of the RMT spiked model. Finally, the detection probability becomes lower as we increase the training time T.

We also compare our results to data-driven FDI attacks proposed in prior work [5], [6] in Figure 5. Here in, attack 1 is constructed according to Theorem 1. Attack 2 is constructed using the entire estimated subspace, i.e., $\mathbf{a}_2 = \widehat{\mathbf{U}}_N \mathbf{c}_N$, where the elements of $\mathbf{c}_N$ are set to $c_i = \sqrt{\frac{\tau}{N \widehat{\omega}_i / \widehat{\mu}_i}}, i = 1, \ldots, N$. Note that $\mathbf{c}_N$ is adjusted to satisfy $\|\Delta \boldsymbol{\theta}\|_2^2 \geq \tau$. As expected, the detection probability of attack 1 is significantly lower compared to attack 2.

Next, we illustrate the trade-off between attack's sparsity and the detection probability in Fig. 6. The points on the trade-off curve are obtained by varying m in (17). Specifically,
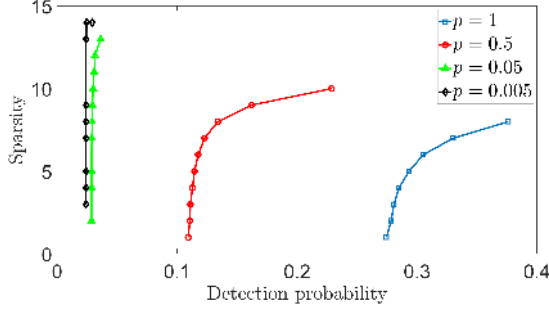
Fig. 6: Trade-off between attack sparsity and the attack detection probability.

we compute an attack vector for each value of m, and then compute the corresponding detection probability and the attack's sparsity. Note that sparsity of the attack vector is equal to $M - K_m^*$. We repeat the simulations for different training times T (i.e. varying p). It can be observed that in the non-asymptotic regime (i.e, small T), the attack's sparsity can be enhanced if the attacker can tolerate an increase in the attack detection probability (refer to the red and blue curves). For large T however, the attacker can simply utilize the entire estimated subspace without having to compromise the attack's detection probability (green and black curves). In practice, the attacker can make use of such trade-off curves to select suitable parameters for the construction of the FDI attack, e.g., based on the resources available to him.

VII. CONCLUSIONS

We have studied the construction of data-driven FDI attacks when the attacker has access to only a limited number of measurements. We showed that in this regime, the attacker can enhance the BDD-bypass probability by constraining the attack vector to a lower-dimensional subspace spanned by the accurately estimated basis vectors. We used results from RMT spiked model to analyze the algorithm performance. We also characterized an important trade-off between the attacker's ability to bypass the BDD and the sparsity the attack vector. Our framework gives practical guidance to a resource-constrained attacker in designing stealthy FDI attacks. In the future, we will explore how the results of this work can be used to address the defense problem against these attackers (e.g., MTD).

APPENDIX A: RELEVANT RESULTS FROM RANDOM MATRIX THEORY

Theorem 2. *Consider $\mathbf{y}[t] = \mathcal{N}(0, \boldsymbol{\Sigma}_z) \in \mathbb{R}^{M \times 1}$, where $\boldsymbol{\Sigma}_z = \mathbf{I} + \sum_{i=1}^{N} \mu_i \mathbf{u}_i \mathbf{u}_i^H$. Let $\widehat{\mathbf{U}} \widehat{\boldsymbol{\Lambda}} \widehat{\mathbf{U}}^T$ denote the eigen value decomposition of $\widehat{\boldsymbol{\Sigma}}_z = \frac{1}{T} \sum_{t=1}^{T} \mathbf{y}[t] \mathbf{y}[t]^T$, where $\widehat{\mathbf{U}} = [\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_M]$, and $\widehat{\boldsymbol{\Lambda}} = diag(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_M)$. Assume N is fixed, and independent of M and T. Then, when $M, T \to \infty, M/T = c$, for all $\mu_i > \sqrt{p}$, with probability one, $\widehat{\lambda}_i \geq (1 + \sqrt{p_N})^2$ and*

$$\left| \widehat{\lambda}_i - 1 - \mu_i - \frac{p(1+\mu_i)}{\mu_i} \right| \xrightarrow{\text{a.s.}} 0. \quad (18)$$

Moreover, $|\mu_i - \widehat{\mu}_i| \xrightarrow{\text{a.s.}} 0$, where $\widehat{\mu}_i$ can be obtained from $\widehat{\lambda}_i$ as $\widehat{\mu}_i = \frac{\widehat{\lambda}_i + 1 - p + \sqrt{(\widehat{\lambda}_i + 1 - p)^2 - 4\widehat{\lambda}_i}}{2} - 1$. Further, for all $\mu_i > \sqrt{p}$, we also have

$$\left| \widehat{\mathbf{u}}_i^T \mathbf{u}_j \mathbf{u}_j^T \widehat{\mathbf{u}}_i - \frac{1 - p/\mu_i^2}{1 + p/\mu_i} \delta_{i=j} \right| \xrightarrow{\text{a.s.}} 0, i, j = 1, \ldots, s. \quad (19)$$

APPENDIX B: UNDETECTABLE ATTACKS

Part I: Proof of Lemma 1

First recall that the residual vector is given by

$$\mathbf{r} = \mathbf{z} - \mathbf{H}\widehat{\boldsymbol{\theta}} = (\mathbf{I} - \mathbf{K})\mathbf{z}, \quad (20)$$

where in (20), we have substituted $\widehat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{z}$ (note $\mathbf{W} = \mathbf{I}_M$) and denoted $\mathbf{K} = \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T$. Further substituting $\mathbf{z} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n}$, we obtain,

$$\mathbf{r} = (\mathbf{I} - \mathbf{K})(\mathbf{H}\boldsymbol{\theta} + \mathbf{n}) = (\mathbf{I} - \mathbf{K})\mathbf{n}, \quad (21)$$

where (21) follows since $(\mathbf{I} - \mathbf{K})\mathbf{H} = \mathbf{0}$.

Now consider the residual for measurements with FDI attack $\mathbf{z}_a = \mathbf{z} + \mathbf{a}$. The residual denoted by $\mathbf{r}_a$ is given by:

$$
\begin{aligned}
\mathbf{r}_a &= (\mathbf{I} - \mathbf{K})\mathbf{z}_a \\
&\overset{(a)}{=} (\mathbf{I} - \mathbf{K})(\mathbf{n} + \mathbf{a}) \\
&\overset{(b)}{=} (\mathbf{I} - \mathbf{U}_N \mathbf{U}_N^T)(\mathbf{n} + \mathbf{a}),
\end{aligned}
\quad (22)
$$

where in (a), we have once again use $(\mathbf{I} - \mathbf{K})\mathbf{H} = \mathbf{0}$, and in (b), we have used the fact that $\mathbf{K}$ can be decomposed as $\mathbf{K} = \mathbf{U}_N \mathbf{U}_N^T$. Since the noise is Gaussian, $\|\mathbf{r}_a\|_2^2$ follows a non central chi-square distribution with $M - N$ degrees of freedom and noncentrality parameter ν given by

$$\mathbb{E}\left[\|\mathbf{r}_a\|_2^2 \right] = \nu = \mathbf{a}^T \mathbf{a} - \mathbf{a}^T \mathbf{U}_N \mathbf{U}_N^T \widehat{\mathbf{U}}_s \mathbf{a}_s. \quad (23)$$

In particular, for a data-driven FDI attack of the form $\mathbf{a} = \widehat{\mathbf{U}}_s \mathbf{c}_s$, we have

$$\nu = \mathbf{c}_s^T \mathbf{c}_s - \mathbf{c}_s^T \widehat{\mathbf{U}}_s^T \mathbf{U}_N \mathbf{U}_N^T \widehat{\mathbf{U}}_s \mathbf{c}_s. \quad (24)$$

For $M, T \to \infty, M/T = c$, using the result of Theorem 2 (main paper), the matrix $\widehat{\mathbf{U}}_s^T \mathbf{U}_N \mathbf{U}_N^T \widehat{\mathbf{U}}_s$ converges to a diagonal matrix whose diagonal elements are given by ω_i defined in (9). Since N is assumed to be fixed (and finite) and independent of M and T (see Theorem 2, main paper), we obtain,

$$\mathbf{c}_s^T \widehat{\mathbf{U}}_s^T \mathbf{U}_N \mathbf{U}_N^T \widehat{\mathbf{U}}_s \mathbf{c}_s - \mathbf{c}_s^T \boldsymbol{\Omega}_s \mathbf{c}_s \xrightarrow{\text{a.s.}} 0.$$

Further, using the result $|\mu_i - \widehat{\mu}_i| \xrightarrow{\text{a.s.}} 0, i = 1, \ldots, s$ (Theorem 2, main paper), from continuous mapping theorem [23], it follows that

$$|\omega_i - \widehat{\omega}_i| \xrightarrow{\text{a.s.}} 0, i = 1, \ldots, s.$$

Note that ω_i is a continuous function μ_i (see (9)).

A. Part II: Proof of Lemma 3

Since $\widehat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{z}$, it follows that $\Delta\widehat{\boldsymbol{\theta}}$ is given by

$$\Delta\widehat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T(\mathbf{z}^a - \mathbf{z})$$
$$\Delta\widehat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\widehat{\mathbf{U}}_s\mathbf{c}_s. \tag{25}$$

From (25), we obtain,

$$\begin{aligned}
||\Delta\widehat{\boldsymbol{\theta}}||_2 &= \mathbf{c}_s^T\widehat{\mathbf{U}}_s^T\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-2}\mathbf{H}^T\widehat{\mathbf{U}}_s\mathbf{c}_s \\
&\stackrel{(a)}{=} \mathbf{c}_s^T\widehat{\mathbf{U}}_s^T\mathbf{U}_N\mathbf{D}^{-1}\mathbf{U}_N^T\widehat{\mathbf{U}}_s\mathbf{c}_s, \\
&\stackrel{(b)}{=} \sigma_\theta^2\mathbf{c}_s^T\widehat{\mathbf{U}}_s^T\mathbf{U}_N\mathbf{M}^{-1}\mathbf{U}_N^T\widehat{\mathbf{U}}_s\mathbf{c}_s.
\end{aligned}$$

where (a) follows since $\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-2}\mathbf{H}^T = \mathbf{U}_N^T\mathbf{D}^{-1}\mathbf{U}_N^T$. Here in, $\mathbf{D} = \text{diag}(d_1,\ldots,d_N)$, where $\{d_i\}_{i=1}^N$ are the first N eigen values of $\mathbf{H}\mathbf{H}^T$ (in decreasing order). In (b), recall that $\mathbf{M} = \text{diag}(\mu_1,\ldots,\mu_N)$, where μ_i are the eigen values of $\sigma_\theta^2\mathbf{H}\mathbf{H}^T$. Similar to the proof of Lemma 1 (Part I), it can be shown that

$$\mathbf{c}_s^T\widehat{\mathbf{U}}_s^T\mathbf{U}_N\mathbf{M}^{-1}\mathbf{U}_s^T\widehat{\mathbf{U}}_N\mathbf{c}_s - \widehat{\sigma}_\theta^2\mathbf{c}_s^T\widehat{\mathbf{M}}^{-1}\widehat{\boldsymbol{\Omega}}\mathbf{c}_s \xrightarrow{\text{a.s.}} 0.$$

APPENDIX C: PROOF OF LEMMA 2

First, we consider the proof of $1 > \omega_1 \geq \omega_2 \geq,\ldots,\geq \omega_s > 0$.

We first show that $0 \leq \omega_i \leq 1, \forall i$. By definition $\mu_i > \sqrt{p}$ for $i = 1,\ldots,s$. For $\mu_i > \sqrt{p}$, we have $1 - p/\mu_i^2 > 0$. Thus, $\omega_i > 0, \forall i$.

Also, it is straighforward to note that $1 - p/\mu_i^2 < 1$ and $1 + p/\mu_i > 1$. Thus, $\omega_i = \frac{1-p/\mu_i^2}{1+p/\mu_i} < 1, \forall i$.

Finally, note that the derivative of ω_i with respect to μ_i is given by

$$\frac{d\omega_i}{d\mu_i} = \frac{\mu_i^2 p + 2\mu_i p + p^2}{\mu_i(\mu_i + p)} > 0. \tag{26}$$

where in (26), the inequality follows since all terms in the derivative are positive. Thus, we conclude that ω_i is an increasing function of μ_i. Since by definition, $\mu_1 \geq \mu_2 \geq,\ldots,\geq \mu_s$, it follows that $1 > \omega_1 \geq \omega_2 \geq,\ldots,\geq \omega_s$.

Next, we consider the proof of $1 > \widehat{\omega}_1 \geq \widehat{\omega}_2 \geq,\ldots,\geq \widehat{\omega}_s > 0$.

Once again, by definition, we have $\widehat{\lambda}_i > (1+\sqrt{p})^2$ for $i = 1,\ldots,s$. It can be verified from Theorem 2 (Appendix A) that for $\widehat{\lambda}_i > (1+\sqrt{p})^2$, $\widehat{\mu}_i > \sqrt{p}$. Thus, $1 > \widehat{\omega}_1 \geq \widehat{\omega}_2 \geq,\ldots,\geq \widehat{\omega}_s > 0$ can be proved by arguments identical to the previous case (i.e., the proof of $1 > \omega_1 \geq \omega_2 \geq,\ldots,\geq \omega_s > 0$).

APPENDIX D: PROOF OF THEOREM 1

Note that optimization problem (15) can be rewritten as

$$\min_{\mathbf{c}} \quad \sum_{i=1}^s (1-\widehat{\omega}_i)c_i^2 \tag{27}$$
$$\text{s.t.} \quad \sum_{i=1}^s \widehat{\sigma}_\theta^2\left(\frac{\widehat{\omega}_i}{\widehat{\mu}_i}\right)c_i^2 \geq \tau$$

By a simple replacement of the variable $y_i = c_i^2$, (27) becomes

$$\min_{\mathbf{y}} \quad \sum_{i=1}^s (1-\widehat{\omega}_i)y_i \tag{28}$$
$$\text{s.t.} \quad \sum_{i=1}^s \widehat{\sigma}_\theta^2\left(\frac{\widehat{\omega}_i}{\widehat{\mu}_i}\right)y_i \geq \tau$$

Note that (28) is a linear programming (LP) problem. Since the coefficients of the objective function as well as the constraints are positive (see Appendix C), the optimal solution of (28) must satisfy the constraint with equality, i.e., $\sum_{i=1}^s \left(\frac{\widehat{\omega}_i}{\widehat{\mu}_i}\right)y_i = \tau$. Thus, we can replace the inequality constraint of (28) with equality. We perform one more change of variable as

$$y_i = \eta_i\left(\frac{\tau}{\widehat{\sigma}_\theta^2(\widehat{\omega}_i/\widehat{\mu}_i)}\right), i = 1,\ldots,s.$$

The LP (28) along with replacing the constraint with equality now becomes

$$\min_{\boldsymbol{\eta}} \quad \sum_{i=1}^s \left(\frac{1-\widehat{\omega}_i}{\widehat{\sigma}_\theta^2(\widehat{\omega}_i/\widehat{\mu}_i)}\right)\eta_i \tag{29}$$
$$\text{s.t.} \quad \sum_{i=1}^s \eta_i = 1.$$

It can be verified that the coefficients of the objective function

$$\frac{1-\widehat{\omega}_i}{\widehat{\sigma}_\theta^2(\widehat{\omega}_i/\widehat{\mu}_i)}$$

is a decreasing function of $\widehat{\mu}_i$. (This can be verified by differentiating the coefficient terms with respect to μ_i and noting that the derivative is negative.) Since $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_s$, we have,

$$\frac{1-\widehat{\omega}_1}{\widehat{\sigma}_\theta^2(\widehat{\omega}_1/\widehat{\mu}_1)} \geq \frac{1-\widehat{\omega}_2}{\widehat{\sigma}_\theta^2(\widehat{\omega}_2/\widehat{\mu}_2)} \geq \cdots \geq \frac{1-\widehat{\omega}_s}{\widehat{\sigma}_\theta^2(\widehat{\omega}_s/\widehat{\mu}_s)}.$$

The solution to (29) is thus given by $\eta_1 = 1, \eta_2 = \eta_3 = \cdots = \eta_s = 0$. Hence, the solution to (27) becomes

$$c_1 = \sqrt{\frac{\tau}{\widehat{\sigma}_\theta^2(\widehat{\omega}_i/\widehat{\mu}_i)}}$$

and $c_2 = c_3 = \cdots = c_s = 0$.

REFERENCES

[1] S. Lakshminarayana, F. Wen, and D. K. Y. Yau, "Trade-offs in data-driven false data injection attacks against the power grid," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 2022–2026.

[2] "Confirmation of a coordinated attack on the Ukrainian power grid," http://bit.ly/1OmxfnG.

[3] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security*, vol. 14, no. 1, pp. 1–33, May 2011.

[4] X. Li, H. V. Poor, and A. Scaglione, "Blind topology identification for power systems," in *Proc. IEEE International Conference on Smart Grid Communications*, Oct. 2013.

[5] J. Kim, L. Tong, and R. J. Thomas, "Subspace methods for data attack on state estimation: A data driven approach," *IEEE Transactions on Signal Processing*, vol. 63, no. 5, pp. 1102–1114, Mar. 2015.

[6] Z. Yu and W. Chin, "Blind false data injection attack using PCA approximation method in smart grid," *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1219–1226, May 2015.

[7] W. Chin, C. Lee, and T. Jiang, "Blind false data attacks against AC state estimation based on geometric approach in smart grid communications," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6298–6306, Nov. 2018.

[8] T. W. Anderson, "Asymptotic theory for principal component analysis," *Ann. Math. Statist.*, vol. 34, no. 1, pp. 122–148, Mar. 1963.

[9] D. Divan and H. Johal, "Distributed FACTS; A new concept for realizing grid power flow control," *IEEE Trans. Power Syst.*, vol. 22, no. 6, pp. 2253–2260, Nov 2007.

[10] R. Verzijlbergh, L. De Vries, G. Dijkema, and P. Herder, "Institutional challenges caused by the integration of renewable energy sources in the european electricity sector," *Renewable and Sustainable Energy Reviews*, vol. 75, pp. 660–667, 2017.

[11] I. M. Johnstone and A. Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 682–693, 2009.

[12] A. Abur and A. G. Exposito, *Power system state estimation: theory and implementation*. CRC press, 2004.

[13] A. Wood and B. Wollenberg, *Power Generation, Operation, and Control*. A Wiley-Interscience, 1996.

[14] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *Journal of Multivariate Analysis*, vol. 97, no. 6, pp. 1382–1408, July 2006.

[15] D. Paul, "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," *Statistica Sinica*, vol. 17, no. 4, pp. 1617–1642, 2007.

[16] M. Ozay, I. Esnaola, F. T. Y. Vural, S. R. Kulkarni, and H. V. Poor, "Sparse attack construction and state estimation in the smart grid: Centralized and distributed models," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 7, pp. 1306–1318, July 2013.

[17] T. T. Kim and H. V. Poor, "Strategic protection against data injection attacks on power grids," *IEEE Transactions on Smart Grid*, vol. 2, no. 2, pp. 326–333, Jun. 2011.

[18] S. Lakshminarayana and D. K. Y. Yau, "Cost-Benefit analysis of moving-target defense in power grids," in *Proc. IEEE/IFIP Dependable Systems and Networks (DSN)*, June 2018, pp. 139–150.

[19] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. New York, NY, USA: Cambridge University Press, 2011.

[20] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[21] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, pp. 457–483, Apr. 1967.

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

[23] P. Billingsley, *Probability and Measure*, 2nd ed. John Wiley and Sons, 1986.