Detection and Estimation of the Invisible Units Using Utility Data Based on Random Matrix Theory

Xing He, Robert C. Qiu, *Fellow, IEEE*, Lei Chu, Qian Ai, Zenan Ling, Jian Zhang

***Abstract*—Invisible units refer mainly to small-scale units that are not monitored, and thus are invisible to utilities and system operators, e.g., small-scale distributed units like unauthorized roof-top photovoltaics (PVs), and plug-and-play units like electric vehicles (EVs). Massive integration of invisible units into power systems could significantly affect the way in which the distribution grid is planned and operated. This paper, based on random matrix theory (RMT), proposes a data-driven approach for the detection, identification, and estimation of the existing invisible units only using easily accessible utility data. The concatenated matrices and linear eigenvalue statistic (LES) indicators are suggested as the main ingredients of this solution. Furthermore, the hypothesis testing is formulated for anomaly detection according to the statistical characteristic of LES indicators. The proposed approach is promising for anomaly detection in a complex grid—it is able to detect invisible power usage, fraud behavior and even to locate the suspect's location. The case studies, using both simulated data and actual data, validate the proposed method.**

***Index Terms*—Invisible unit, utility data, fraud behavior, anomaly detection, random matrix theory, data-driven, concatenated matrix, linear eigenvalue statistic.**

I. Introduction

FUTURE grids are fundamentally different from current ones [1]. Technology development, environment pressure, and market reform have greatly spurred the deployment and penetration of the distributed, the renewable, and even the plug-and-play units, on both the power generation side and the power consumption side. The worldwide small-scale roof-top photovoltaics (PVs) installation reached 23 GW at the end of 2013, and the growth is predicted to be 20 GW per year until 2018 [2]. The up-take of electric vehicles (EVs) also continues to increase. At least 665,000 electric-driven light-duty vehicles, 46,000 electric buses, and 235 million electric two-wheelers were in the worldwide market in early 2015 [3].

These distributed units are mostly invisible to utilities, i.e., they are not monitored by, and thus not visible to, power system operators. 1) Accessing distributed units operation data into utility systems requires an enormous amount of cost paid for data acquisition, communication, storage, calculation, and security [4]. 2) It is hard to describe these units using a fixed model or in a united way; they are small-scale and mostly with high uncertainty or individuality. 3) Some anomaly behaviors are essentially invisible. In 2009, over 20% of total electricity generated is lost from theft in India alone [5]. In 2014, the system in Hawaii, with the highest penetration of PVs in the U.S., recognized a large number of unauthorized PV installations [2].

Lack of visibility may result in incorrect planning and operation of power systems, and even worse, damaging system equipment such as transformers, voltage regulators, and customer appliances. For a highly distributed energy resource penetration environment, utilities are facing technical problems related to overvoltage, frequency control, back feeding flow, and other issues such as a rapid decrease in revenue. The prosumers are also bringing many unknowns and risks that need to be identified and managed [3].

To solve the above problems, many distribution utilities have begun deploying high-precision distribution phasor measurement units (PMUs) for monitoring, diagnostic, and control purposes [6]. High resolution voltage and current phasor measurements can be used in a plethora of applications concerning real-time system operation and long-term planning, such as state estimation, model validation, load characterization, and event detection and localization [7].

Many researchers have studied the impacts and risks of invisible units, especially PVs, on distribution systems [8]; little attention, however, has been paid to the detection and estimation of the invisible units, especially in a complex distributed grid. Some related research is found in the special issue of "Big Data Analytics for Grid Modernization" [9]. Reference [10] proposes a change-point detection algorithm for a time series. The change-point concept is relevant to our paper in spirit. The proposed algorithm, however, is effective only if the characteristics of all other units before and after the change point are similar. In addition, the spatial information of the utility data (data distributed across nodes) are not used. Reference [2] takes the uncertainty in PV sites into account, and estimates the power generation of invisible solar photovoltaic sites using the data generated by a small set of selected representative sites. Reference [11] proposes an approach of big data characterization for smart grids and a two-layer dynamic optimal synchrophasor measurement devices selection algorithm for fault detection, identification, and causal impact analysis. Our previous work [1, 12–15], based on random matrix theory (RMT), also outlines a data-driven methodology to conduct big data analytics for power systems. Our approach utilizes the temporal-spatial statistics.

A. Contribution

This paper proposes an approach aimed at detection and estimation of the invisible units in a complex distribution grid; the analysis of these results will give insight into distribution network characteristics and consumer behaviors. Based on RMT, the proposed approach handles raw data in an unsupervised way and obtains Linear Eigenvalue Statistic (LES) indicators, which are in high-dimensional vector space and thus robust when considering data errors (e.g., data loss, data

out-of-synchronization) [13]. Furthermore, using the statistical characteristics of LES indicators, hypothesis testing can be formulated for anomaly detection. The data analytics only rely on easily accessible utility data such as node voltage magnitudes and node power injections. Finally, the proposed method is validated using both simulated data of a complex grid and field data of a certain distribution system in China. The heart of the method is presented in Sec. III-B2.

II. Problem Formulation

This paper attempts to conduct situation awareness in a non-omniscient distribution network. More precisely, we try to obtain the load/generator ingredients and their weights, and the power usage behaviors at the node level.

For any node, its customers are divided into two categories—typical load pattern units (TLPs) and uncertain load pattern ones (ULPs).

1) The TLPs operate according to a well-defined profile, and are denoted as vectors $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n$. For instance, street-lamps are turned on at 18:00 and turned off at 6:00; their load pattern is modeled as

$$\mathbf{p}_{\text{Lamp}}(t) = \begin{cases} 1 & t \in [00:00, 06:00] \cup [18:00, 24:00] \\ 0 & t \in [06:00, 18:00] \end{cases}.$$

If the sampling interval is 6 hours, $\mathbf{p}_{\text{Lamp}} = [1, 0, 0, 1]$.

2) The ULPs are denoted as vectors $\mathbf{p}_{\text{u}1}, \mathbf{p}_{\text{u}2}, \cdots, \mathbf{p}_{\text{u}m}$, and might be further divided into three categories—completely random behavior, invisible behavior, and fraudulent behavior. We have already successfully distinguished completely random behavior from the others in our previous work [1, 13] by using random matrix tools. Next, we will focus on the detection and identification of invisible and fraudulent behavior. The former often causes a chain reaction and has an impact on other parameters. For instance, unauthorized residential PV installation and plug-in EV charging changes the power flow. The latter often causes parameter deviation in isolation. For instance, some metering error or cyber attack might merely reduce data value of power consumption P without affecting voltage U.

Motivated by the above observations, we propose to study a general model for each node:

$$\mathbf{p}_\Sigma = a_1\mathbf{p}_1 + a_2\mathbf{p}_2 + \cdots + a_n\mathbf{p}_n + b_1\mathbf{p}_{\text{u}1} + b_2\mathbf{p}_{\text{u}2} + \cdots + b_m\mathbf{p}_{\text{u}m}, \tag{1}$$

where vectors $\mathbf{p}_i, i = 1, ..., n$ and $\mathbf{p}_{\text{u}j}, j = 1, ..., m$ are the daily patterns of TLPs and ULPs, with coefficients $a_i, b_j, i = 1, ..., n, j = 1, ..., m$, respectively. Thus, for $i = 1, ..., n, j = 1, ..., m$, vector $a_i\mathbf{p}_i$ is the daily power usage for the i-th TLP, and similarly vector $b_j\mathbf{p}_{\text{u}j}$ is the daily power usage for the j-th ULP.

If all the units patten and behaviors are known in advance, i.e., no $\mathbf{p}_{\text{u}j}$ exists, or if ULPs are able to be modeled as $\mathbf{p}_{i+j}$ instead of uncertain $\mathbf{p}_{\text{u}j}$, then Eq. (1) can be rewritten as

$$\mathbf{p}_\Sigma = a_1\mathbf{p}_1 + a_2\mathbf{p}_2 + \cdots + a_{n+m}\mathbf{p}_{n+m}, \tag{2}$$

Our first step is to formulate the problem in terms of a classical optimization

$$\arg\min_{a_i} \left(\mathbf{p}_{\text{N}} - \mathbf{p}_{\text{L}} - \mathbf{p}_\Sigma \begin{pmatrix} a_1 & a_2 & \cdots & a_{m+n} \end{pmatrix}\right), \tag{3}$$

where vectors $\mathbf{p}_{\text{N}}$ and $\mathbf{p}_{\text{L}}$ are the power injections of nodes and power losses of nodes, respectively, which are measurable and calculable. In addition, it is worth mentioning that the analysis for the reactive power Q may be conducted similarly.

For the modern distribution network, as described in Sec I, ULPs play an important role: $b_j P_{\text{u}j}, j = 1, ..., m$ are present and their influences need to be considered. They violate the prerequisites of most algorithms (e.g., least square method) and have significant effects on the final values of coefficients $a_i, i = 1, ..., n$ in Eq. (1). In most cases, it is reasonable to model $\mathbf{p}_{\text{u}j}, j = 1, ..., m$ as a step signal. This is the case when the plug-in EVs charge and/or unauthorized PVs generate during t_a to t_b. *Determining the start point and the end point of the step signal* is at the heart of the problem. Based on random matrix theory (RMT) and linear eigenvalue statistics (LES), a statistical, data-driven solution, rather than its deterministic, empirical or model-based counterpart, is proposed to solve the problem.

III. Mathematical Foundation

A. Random Matrix Theory

1) Statistics based on Random Matrix Theory:

Random matrices have been an important issue in multivariate statistical analysis since the landmark work of Wishart on fixed size Gaussian matrices. The asymptotic theory on the limiting spectrum of large random matrices was initially proposed in several works [16] by Wigner in the 1950s, motivated by problems in quantum physics. Since then, research on the finite spectral analysis of high dimensional random matrices has come under heated discussion by scholars in numerous disciplines. The RMT, as a statistical tool with profound theoretical basis, is adapted to multivariate analysis. It can help model many intractable practical systems, especially those with numerous variables.

2) Laws for Spectral Analysis:

RMT mainly concerns two ensemble random matrices—Gaussian unitary ensemble (GUE) and Laguerre unitary ensemble (LUE).

$$\mathbf{A} = \begin{cases} \frac{1}{2}\left(\mathbf{Y} + \mathbf{Y}^H\right) & , \mathbf{Y} \in \mathbb{C}^{N\times N} \text{ ,GUE;} \\ \frac{1}{N}\mathbf{Y}\mathbf{Y}^H & , \mathbf{Y} \in \mathbb{C}^{N\times T} \text{ ,LUE.} \end{cases}, \tag{4}$$

where $\mathbf{Y}$ is the standard Gaussian Random Matrix.

Let $p_{\mathbf{A}}(x)$ be the empirical density of $\mathbf{A}$, and define its empirical spectral distribution (ESD) $F_{\mathbf{A}}(x)$:

$$F_{\mathbf{A}}(x) = \frac{1}{N}\sum_{i=1}^{N} I_{\{\lambda_i \le x\}}, \tag{5}$$

where $\mathbf{A}$ is GUE or LUE matrix, $I(\cdot)$ represents the event indicator function. We investigate the rate of convergence of the expected ESD $\mathbb{E}\{F_{\mathbf{A}}(x)\}$ to Wigner's Semicircle Law or Wishart's M-P Law.

Let $g_{\mathbf{A}}(x)$ and $G_{\mathbf{A}}(x)$ denote the empirical eigenvalue density and ESD of $\mathbf{A}$, and the Wigner's Semicircle Law [16] and Wishart's Marchenko-Pastur (M-P) Law [17] say:

$$g_{\mathbf{A}}(x)=\begin{cases}\frac{1}{2\pi}\sqrt{4-x^2} & ,x\in[-2,2]\ \text{,GUE;}\\ \frac{1}{2\pi cx}\sqrt{(x-a)(b-x)} & ,x\in[a,b]\ \text{,LUE;}\end{cases}, \tag{6}$$

where $a=(1-\sqrt{c})^2, b=(1+\sqrt{c})^2$.

$$G_{\mathbf{A}}(x)=\int_{-\infty}^{x} g_{\mathbf{A}}(u)\,du. \tag{7}$$

Then, we denote the Kolmogorov distance between $\mathbb{E}\{F_{\mathbf{A}}(x)\}$ and $G_{\mathbf{A}}(x)$ as Δ:

$$\Delta=\sup_x\left|\mathbb{E}\{F_{\mathbf{A}}(x)\}-G_{\mathbf{A}}(x)\right|. \tag{8}$$

Gotze and Tikhomirov, in their work [18], prove an optimal bound for Δ of order $O\left(N^{-1}\right)$.

B. Linear Eigenvalue Statistics and its Central Limit Theorem

The LES τ of an arbitrary matrix $\mathbf{\Gamma}\in\mathbb{C}^{N\times N}$ is defined in [19, 20] via the continuous test function $\varphi:\mathbb{C}\to\mathbb{C}$,

$$\tau(\varphi,\mathbf{\Gamma})=\mathcal{N}_N[\varphi]=\textstyle\sum_{i=1}^{N}\varphi(\lambda_i)=\mathrm{Tr}\varphi(\mathbf{\Gamma}), \tag{9}$$

where the trace of the function of a random matrix is involved.

1) Law of Large Numbers:

The Law of Large Numbers tells us that $N^{-1}\mathcal{N}_N[\varphi]$ converges in probability to the limit

$$\lim_{N\to\infty}\tfrac{1}{N}\mathcal{N}_N[\varphi]=\int\varphi(\lambda)\rho(\lambda)\,d\lambda, \tag{10}$$

where $\rho(\lambda)$ is the probability density function of λ.

2) Central Limit Theorem:

The CLT [20] as the natural second step, aims to study the LES fluctuations [21]. Consider covariance matrix $\mathbf{M}=\frac{1}{N}\mathbf{X}\mathbf{X}^H$. The CLT for $\mathbf{M}$ is given as follows [20]:

Theorem III.1 (M. Sheherbina, 2009)**.** *Let the real valued test function φ satisfy condition $\|\varphi\|_{3/2+\varepsilon}<\infty\,(\varepsilon>0)$. Then $\mathcal{N}_N{}^{\circ}[\varphi]$ defined in* (10)*, in the limit $N,T\to\infty, c=N/T\le 1$, converges in the distribution to the Gaussian random variable with zero mean and the variance:*

$$V_{SC}[\varphi]=\frac{2}{c\pi^2}\iint\limits_{-\frac{\pi}{2}<\theta_1,\theta_2<\frac{\pi}{2}}\psi^2(\theta_1,\theta_2)(1-\sin\theta_1\sin\theta_2)\,d\theta_1 d\theta_2 + \frac{\kappa_4}{\pi^2}\left(\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\varphi(\zeta(\theta))\sin\theta d\theta\right)^2, \tag{11}$$

where $\psi(\theta_1,\theta_2)=\frac{[\varphi(\zeta(\theta))]|_{\theta=\theta_2}^{\theta=\theta_1}}{[\zeta(\theta)]|_{\theta=\theta_2}^{\theta=\theta_1}}$, $[\zeta(\theta)]|_{\theta=\theta_2}^{\theta=\theta_1}=\zeta(\theta_1)-\zeta(\theta_2)$, and $\zeta(\theta)=1+1/c+2/\sqrt{c}\sin\theta$; $\kappa_4=\mathbb{E}\left(X^4\right)-3$ is the 4-th cumulant of entries of $\mathbf{X}$.

Eq. (8) has been used in a power grid in our previous work [14]. This paper takes a fundamentally different approach from (8). To study the convergence as a function of N, we study the LES instead of the probability distribution of eigenvalues in (8). For an arbitrary test function with enough smoothness, the LES Y is a (positive) scalar random variable defined in (9). As $N\to\infty$, the asymptotic limit of its expectation, $\mathbb{E}[Y]$, is given in (10). As $N\to\infty$, the asymptotic limit of its variance, $\mathrm{Var}[Y]$, is given in (11). These two equations are sufficient to study the scalar random variable Y. This approach can be viewed as a dimensionality reduction. The random data matrix of size $N\times T$ is reduced to a (positive) scalar random variable Y! This dimension reduction is mathematically rigorous only when $N\to\infty, T\to\infty$ but $\frac{N}{T}\to c$. Experiences demonstrate, however, that moderate values of N and T are accurate enough for our practical purposes.

3) Change Point Detection using LES:

Change-point detection began with Page's (1954, 1955) classical formulation, which was further developed by Shiryaev (1963) and Lorden (1971) [22]. Change-point detection is such a problem: Suppose $X_1, X_2, \cdots, X_m$, are independent observations. For $j\le M$ they have the distribution F_0, while for $j>M$ they have the distribution F_1. The distributions F_1 may be completely specified or may depend on unknown parameters. In the case of a fixed number m of observations, we would like to test the null hypothesis of no change, that $F_0=F_1$, and perhaps to estimate M.

This paper formulates the hypothesis test in terms of the statistical characteristics of LES indicators. Theorem III.1 says that the LES indicator τ_φ, in the limit $N,T\to\infty, c=N/T\le 1$, converges in the distribution to a Gaussian random variable with mean $\mathbb{E}(\tau_\varphi)$ and variance $\sigma(\tau_\varphi)$. Due to the Gaussian property, following a standard procedure, the detection is modeled as a binary hypothesis test: normal hypothesis $\mathcal{H}_0$ (no anomaly present) and abnormal one $\mathcal{H}_1$, denoted by:

$$\begin{cases}\mathcal{H}_0: \left|\dfrac{\tau_\varphi-\mathbb{E}(\tau_\varphi)}{\sigma(\tau_\varphi)}\right|<\epsilon,\\ \mathcal{H}_1: \left|\dfrac{\tau_\varphi-\mathbb{E}(\tau_\varphi)}{\sigma(\tau_\varphi)}\right|\ge\epsilon,\end{cases} \tag{12}$$

where ϵ is the threshold value, that needs to be preset based on experiences.

C. Concatenation Operation

Numerous *causing factors* affect the *system state* in different ways; sensitivity analysis is a valuable and hot topic. Assuming that there are N state variables and M factors, their sampling data are multiple time-series. In a fixed period of interest t_i $(i=1,2,\ldots,T)$, the sampling data of N state variables consist of a matrix $\mathbf{B}\in\mathbb{C}^{N\times T}$ (i.e. *state matrix*), and the factors consist of $\mathbf{c}_j\in\mathbb{C}^{1\times T}$ $(j=1,2,\ldots,M)$ (i.e. *factor vector*). Two matrices with the same length can be put together and a concatenated matrix is formed; in such a way, we obtain a new matrix $\mathbf{A}$ using the state matrix $\mathbf{B}$ and the factor matrix $\mathbf{c}_j$.

In order to balance the proportion (to increase the statistic correlation), a factor matrix is formed for each factor vector. First, for the factor $\mathbf{c}_j$, we duplicate it for K times[1] to construct a matrix $\mathbf{D}_j$, written as

$$\mathbf{D}_j=\begin{bmatrix}\mathbf{c}_j{}^T & \mathbf{c}_j{}^T & \cdots & \mathbf{c}_j{}^T\end{bmatrix}^T\in\mathbb{C}^{K\times T}.$$

[1] K is appropriated to $0.3\times N$

Then, white noise is introduced into $\mathbf{D}_j$ to avoid extremely strong cross-correlations. Thus, the factor matrix $\mathbf{C}_j$ for the factor vector $\mathbf{c}_j$ is expressed as

$$\mathbf{C}_j = \mathbf{D}_j + \eta_j \mathbf{R} \quad (j = 1, 2, \ldots, m), \tag{13}$$

where η_j is related to the signal-to-noise ratio (SNR), and the entries $R_{i,j}$ of the matrix $\mathbf{R}$ are Gaussian random variables.

Through the trace function $\mathrm{Tr}(\cdot)$, the SNR of the factor matrix $\mathbf{C}_j$ is defined as

$$\rho_j = \frac{\mathrm{Tr}(\mathbf{D}_j \mathbf{D}_j^H)}{\mathrm{Tr}(\mathbf{R}\mathbf{R}^H)\eta_j^2} \quad (j = 1, 2, \ldots, m). \tag{14}$$

In parallel, we can construct the concatenated matrix with each factor $\mathbf{c}_j$, expressed as

$$\mathbf{A}_j = \begin{bmatrix} \mathbf{B} \\ \mathbf{C}_j \end{bmatrix} \quad (j = 1, 2, \ldots, m). \tag{15}$$

The relationships between causing factors $\mathbf{c}_j$ and system state $\mathbf{B}$ can be revealed by the concatenated matrix $\mathbf{A}_j$. The concatenated model is compatible with different units and different measurements for each variable data (in the form of rows of $\mathbf{A}_j$), due to the normalisation during the data preprocessing. Besides, it is worth to mention that some simple mathematical methods, e.g., interpolation, may be applied to handle data source with different sampling rates.

D. Experiment Design Using Variable Data of Power Systems

The operating states of power systems can be estimated by various kinds of state variables, such as frequencies, voltages, currents, and power flows. In this paper, the state matrix $\mathbf{U} \in \mathbb{C}^{N \times T}$ is made up of $U_{i,j} (i = 1, 2, \cdots, N, t = 1, 2, \cdots, T)$, and the k-th factor matrix $\mathbf{P}_{\Sigma k} \in \mathbb{C}^{K \times T}$ is made up of $P_{\Sigma k,j} (j = 1, 2, \cdots, T)$ according to (13). Similar to (15), we obtain

$$\mathbf{F}_k = \begin{bmatrix} \mathbf{U} \\ \mathbf{P}_{\Sigma k} \end{bmatrix} \in \mathbb{C}^{(N+K) \times T} \quad (k = 1, 2, \ldots, N). \tag{16}$$

IV. Simulation Cases

A. Background

Simulations are based on the IEEE-33 bus system for a distribution network, shown as Fig. 1. For node k, its gross power usage $\mathbf{p}_{\Sigma,k}$ and voltage magnitude $\mathbf{u}_k$ are sampled at a high rate, for example, 9600 points per day (0.11 Hz). Then we introduce the white noise to the power injections as

$$\tilde{y}_{nt} = y_{nt} (1 + \gamma_1 Z_1) + \gamma_2 Z_2, \tag{17}$$

where Z_1 and Z_2 are two standard Gaussian random variables, i.e. $\mathcal{N}(0, 1)$; $\gamma_1 = 0.005, \gamma_2 = 0.02$. In this way, the related power flow is obtained via the software package Matpower.

As mentioned in Sec. II, we mainly focus on fraudulent behavior and invisible power usage. Determining the start point and the end point of the $\mathbf{p}_{\mathrm{u}i}$ is the focus of this paper. For the longstanding anomaly without any step signals in the observed data segment, long-term indicators, such as monthly line loss rate, might be sensitive. This is another topic that will be explored elsewhere.

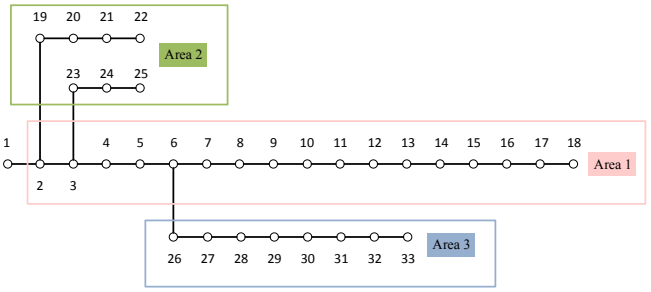
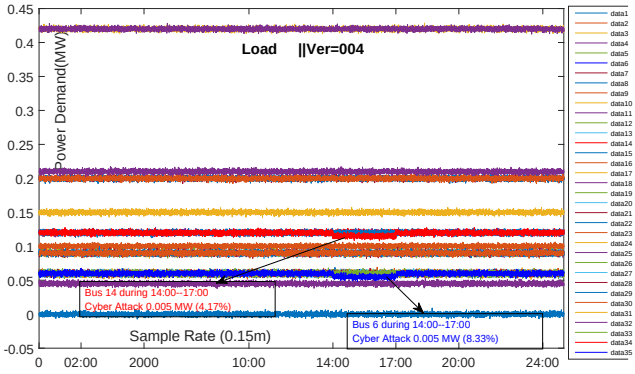


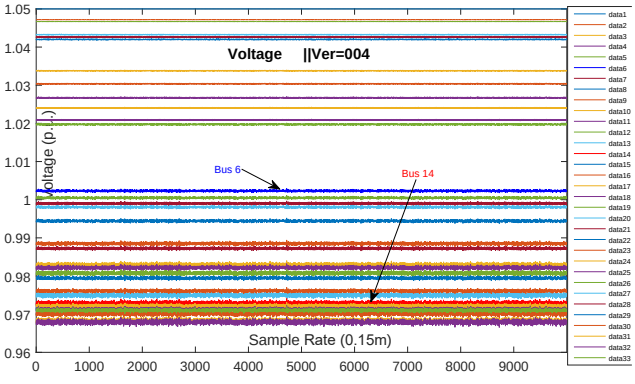
Fig. 1: Topology of the IEEE 33-bus distribution network.

B. Fraud events in a Simple Scenario

Fraud events often cause parameter deviation. Suppose that the active power values P for each node are at their initial points with fluctuations defined in (17). From $14:00$ to $17:00$, some fraud events on node-6 and node-14 cause a reduction of 0.005 MW (8.33% of the total P_6, and 4.17% of P_{14}). The sampling data, power consumptions and voltage magnitudes of each node, are shown as Fig. 2. The lines with legends data 1 to data 33 are for actual power consumption of node 1 to node 33, and lines with data 34 and data 35 are for measured power consumption of node 14 and node 6, respectively. According to the actual power consumption, i.e., data 1 to data 33, the voltage magnitudes are obtained in Fig. 2b. Note that due to the fraud events, the data 14 and data 6 of Fig. 2a are unreachable.



(a) P of each node



(b) V of each node

Fig. 2: Power demand and voltage magnitudes of each node

The matrix concatenation operation and the split window method are used to handle the sampling data. Using (9) for $N=33, T=100, \Delta T=1$, we choose Chebyshev polynomials T_2: $\varphi(x)=2x^2-1$ as the test function. The LES indicators τ_{T_2} of state matrix $\mathbf{B}$ and concatenated matrix $\mathbf{A}_i$ $(i=1,\cdots,33,$ referring Eq. (16)) are obtained as Fig. 3.

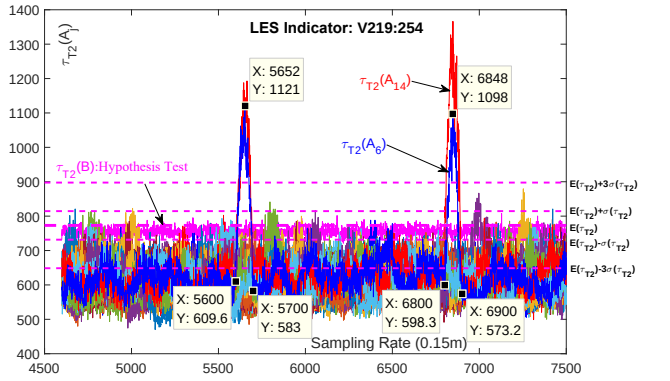


Fig. 3: LES indicator in the simple scenario

In Fig. 3, the LES indicator τ_{T_2} of state matrix $\mathbf{B}$, namely, $\tau_{\mathrm{T}_2}(\mathbf{B})$ is almost constant. From a statistic view, the theoretical expectation $\mathbb{E}(\tau_{\mathrm{T}_2})$ and the standard deviation $\sigma(\tau_{\mathrm{T}_2})$ are accessible via random matrix theory, or rather, via Eq. (6), (9), and (11). It is found that the experimental indicator $\tau_{\mathrm{T}_2}(\mathbf{B})$ is exactly bounded between $\mathbb{E}(\tau_{\mathrm{T}_2})-\sigma(\tau_{\mathrm{T}_2})$ and $\mathbb{E}(\tau_{\mathrm{T}_2})+\sigma(\tau_{\mathrm{T}_2})$. According to Eq. (12), we should accept the hypothesis $\mathcal{H}_0$—there is no factor actually affecting the system state during the observation period. On the other hand, τ_{T_2} of state matrix $\mathbf{A}_i$, namely, $\tau_{\mathrm{T}_2}(\mathbf{A}_i)$ has four spikes: two spikes for $\tau_{\mathrm{T}_2}(\mathbf{A}_6)$ and two spikes for $\tau_{\mathrm{T}_2}(\mathbf{A}_{14})$. Our previous work [1] tells us that the anomaly should last T time points (i.e. $T\times 0.15$ m) and have an extreme point at $T/2$. This phenomenon is observed on the $\tau_{\mathrm{T}_2}(\mathbf{A}_6)-t$ curve and $\tau_{\mathrm{T}_2}(\mathbf{A}_{14})-t$ curve:

$$5700-5600=100=T,\quad 5652-5600\approx 50=T/2.$$

C. Invisible Power Usage and Fraud Events in a Complex Scenario

This subsection proposes a data-driven solution for the problem given in Sec. II—determining the start point and the end point to model the invisible power usage $\mathbf{p}_{\mathrm{u}i}$ as a step signal. Firstly, we assume a complex scenario:

1) The power usage of each bus (e.g., bus i) generally consists of four TLPs and one ULP, denoted as

$$P_{i\Sigma}=a_{i1}P_1+a_{i2}P_2+a_{i3}P_3+a_{i4}P_4+b_{i1}P_{\mathrm{u}1}. \quad (18)$$

The daily load profiles of TLPs are set as Tab. I and shown as Fig. 4. Note that the blue-filled rectangle means that the load profiles have a dramatic change at this time point. According to work [10], these special time points are denoted as change points (CPs). The coefficients a_i, b_j are assumed as Tab II.

2) We assume that there exists invisible power usage events on node 20 and 31: the periods are 1:00–5:00 and 14:00–20:00, and the percentages are 30% and 50%, respectively.

3) We assume that there exist fraud events on node 6, 14 and 27, the periods are 20:00–22:00, 14:00–17:00 and 18:00–19:00, and the percentages are 7%, 8% and 12%, respectively.

TABLE II: Coefficients of TLPs and ULP of each node.

	a_1	a_2	a_3	a_4	b_1		a_1	a_2	a_3	a_4	b_1
1	0.25	0.25	0.25	0.25	0	2	0	0.7	0.1	0.2	0
3	0	0.1	0.8	0.1	0	4	0.05	0.75	0.1	0.1	0
5	0	0.1	0.8	0.1	0	6	0.1	0.2	0.5	0.2	0
7	0.8	0.05	0.1	0.05	0	8	0.85	0.05	0	0.1	0
9	0.1	0.15	0.6	0.15	0	10	0	0.15	0.8	0.05	0
11	0	0.2	0.75	0.05	0	12	0.05	0.1	0.75	0.1	0
13	0.05	0.05	0.85	0.05	0	14	0.7	0.05	0.2	0.05	0
15	0	0.05	0.9	0.05	0	16	0	0	0.95	0.05	0
17	0	0.1	0.8	0.1	0	18	0	0.7	0.1	0.2	0
19	0	0.5	0.1	0.4	0	20	0	0.2	0.2	0.3	0.3
21	0	0.8	0.1	0.1	0	22	0.1	0.75	0	0.15	0
23	0.2	0.6	0	0.2	0	24	0.85	0	0.05	0.1	0
25	0.75	0.1	0.1	0.05	0	26	0.2	0	0.7	0.1	0
27	0.1	0	0.75	0.15	0	28	0.25	0.1	0.6	0.05	0
29	0.8	0.05	0.1	0.05	0	30	0.9	0	0.05	0.05	0
31	0.1	0.1	0.05	0.25	0.5	32	0.9	0	0	0.1	0
33	0.95	0	0	0.05	0						

Using (16), we obtain the active power $P_{\Sigma k,j}(j\in 1,2,\cdots,T)$ and then calculate the voltages $U_{i,j}(i\in 1,2,\cdots,N, t\in 1,2,\cdots,T)$ for the assumed complex scenario above; the results are shown as Fig. 5a, 5c and 5b.

With a similar procedure to that of Sec IV-B, the $\tau_{\mathrm{T}_2}-t$ curve is obtained in Fig. 5d. Based on the curve of Fig. 5d, we make the following observations:

- The brown line at the bottom is the indicator $\tau_{\mathrm{T}_2}(\mathbf{B})$; it is relatively smooth.
- The results shown in Fig. 5d match the settings of the daily load pattern in Tab. I. Taking TLP P_1as an example, Fig 5d shows that the indicators of nodes 25, 24, 32, 30, etc, have bright spikes at 3:00; in fact, 3:00 is a CP of TLP P_1 in Tab. I. The coefficients in Table II tell us that these listed nodes are the exact ones of which the TLP P_1 takes a dominant part.
- For the fraud events, the limit points are located at $t=$ 5553, 6856, 7655, etc. According to Sec IV-B, the key time points are $t=$ 14:00 (5600), 17:00 (6800), 19:00 (7600), etc., respectively.
- For the invisible power usage, we can locate them using the special time points $t=200, 700$ and node 31, 20. For time points $t=200$, 700, the change point is $t=400$.[2] With similar procedure, the CPs are found as $t=2000, 5600, 8000$, and these CPs are at $t=$ 1:00, 5:00, 14:00 and 20:00. These results agree with the daily load pattern of Table I and the coefficients of Table II. The step signal for $P_{\mathrm{u}1}$ is modeled based on this analysis.

V. Real-World Case Studies

A. Data

We use a power grid with 5 substations in China (Fig. 6a). For each substation, its three-phase voltage data V and current data I are recorded using a three-minute sampling-rate. We take a two-day time period as the data set, depicted as Fig. 6c, 6d, 6e, and 6f.

[2] 400=[(200+50-50)+(700-50-50)]/2

	P_1	P_2	P_3	P_4	P_{u1}		P_1	P_2	P_3	P_4	P_{u1}
0	88	20	25	100	0	12	94	77	35	0	0
1	87	20	23	100	100	13	86	80	30	0	0
2	88	20	22	100	100	14	86	86	33	0	100
3	100	21	22	100	100	15	88	86	44	0	100
4	96	20	27	100	100	16	85	87	50	100	100
5	100	20	31	100	0	17	87	35	56	100	100
6	98	20	29	0	0	18	88	25	85	100	100
7	97	30	28	0	0	19	85	25	80	100	100
8	88	40	31	0	0	20	84	20	70	100	0
9	82	85	37	0	0	21	83	20	76	100	0
10	82	85	42	0	0	22	86	20	43	100	0
11	95	82	42	0	0	23	88	15	30	100	0

Note: blue-filled rectangle means CP.

TABLE I: Typical Loads and their 24-hour power demand.

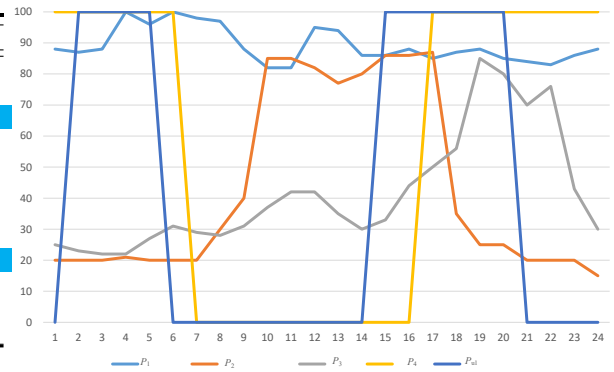


Fig. 4: Daily power demands for typical loads

B. Results: Ring Law and LES Indicator

If we choose $\mathbf{X}_0$ (in Fig. 6c), i.e., the voltage data during 0 a.m. to 2 a.m., the ring distribution is obtained according to our previous work [1], shown as Fig. 6b. Most eigenvalues are distributed between the inner circle and the outer circle. This implies that the real-world data does follow the Ring Law. With a similar process, and setting the test function as Chebyshev Polynomials T_2: $\varphi(x) = 2x^2 - 1$ and the Likelihood Ratio Function LR : $\varphi_{\mathrm{LR}}(x) = x - \ln(x) - 1$, respectively, the LES $t - \tau$ curves are obtain as Fig. 7a, 7b, 7c, and 7d. The grid is relatively smooth during 0 a.m. to 8 a.m. and has dramatic changes at around 8:30 a.m., 11:30 a.m., etc. This observation agrees with our common sense. For the field data, the test function will influence the result in some complicated ways, although the indicators have a similar trend at most CPs.

VI. Conclusion

This paper extends our framework of using large random matrices to model a power grid in several ways. First, *a model-free, data-driven statistical approach* is proposed for the detection and estimation of the *invisible units*, a stressing problem in industry. Behind this approach, we exploit the statistical property of massive datasets in a high-dimensional vector space. The temporal variations (T sampling instants) are simultaneously observed together with spatial variations (N grid nodes). Based on mathematically rigorously random matrix theory, time and space must be unified through their ratio $c = T/N$. What matters is the ratio c, rather than N and T! This observation is valid when N and T are large and comparable in size, which is often true in practice.

Second, we explore numerous practical aspects. Hypothesis tests, change point detection, and concatenation operations are investigated. The statistical features of Linear Eigenvalue Statistics (LES), i.e. $\mathbb{E}(\tau)$ and $\sigma(\tau)$, are studied. Based on these features, the hypothesis test is designed for the detection of fraud behavior and anomaly behavior.

Third, real-world data are tested using our algorithms. We find that the experimental LES indicators agree with the theoretical predictions: the Ring Law is valid. Both the simulated cases and real-world cases validate the proposed approach as a powerful and effective way to gain insight into the distribution network characteristics and consumer behaviors.

We pave the way for future work with this paper. First, in the context of cyber attacks in distribution networks, our approach can locate these attacks. Second, the power of our algorithms depends on the selection of the test function; more test functions need to be studied and optimized using metrics.

References

[1] X. He, Q. Ai, R. C. Qiu, W. Huang, L. Piao, and H. Liu, "A big data architecture design for smart grids based on random matrix theory," *IEEE Transactions on Smart Grid*, vol. 8, no. 2, pp. 674–686, 2017. [Online]. Available: http://arxiv.org/pdf/1501.07329.pdf

[2] H. Shaker, H. Zareipour, and D. Wood, "Estimating power generation of invisible solar sites using publicly available data," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2456–2465, Sept 2016.

[3] Y. Parag and B. K. Sovacool, "Electricity market design for the prosumer era," *Nature Energy*, vol. 1, p. 16032, 2016.

[4] J. Hu and A. V. Vasilakos, "Energy big data analytics and security: Challenges and opportunities," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2423–2436, Sept 2016.

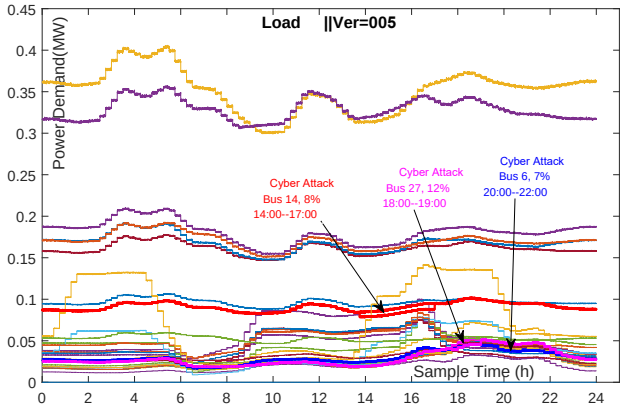
[5] V. Gaur and E. Gupta, "The determinants of electricity theft: An empirical analysis of indian states," *Energy Policy*, vol. 93, pp. 127–136, 2016.

[6] A. Von Meier, D. Culler, A. McEachern, and R. Arghandeh, "Micro-synchrophasors for distribution systems," in *Innovative Smart Grid Technologies Conference (ISGT), 2014 IEEE PES*. IEEE, 2014, pp. 1–5.

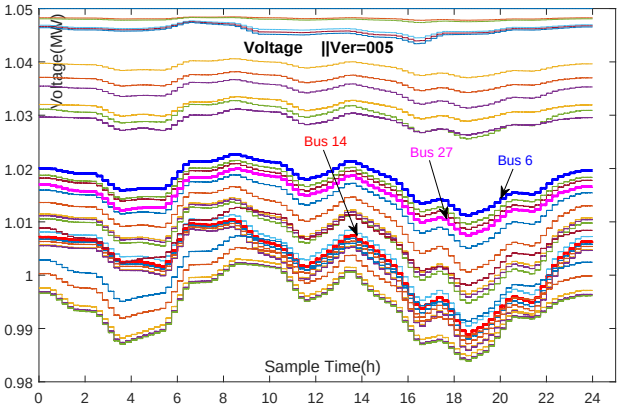
[7] O. Ardakanian, Y. Yuan, R. Dobbe, A. von Meier, S. Low, and C. Tomlin, "Event detection and localization in distribution grids with phasor measurement units," *arXiv preprint arXiv:1611.04653*, 2016.

[8] A. Samadi, L. Söder, E. Shayesteh, and R. Eriksson, "Static equivalent of distribution grids with high penetration of pv systems," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1763–1774, 2015.

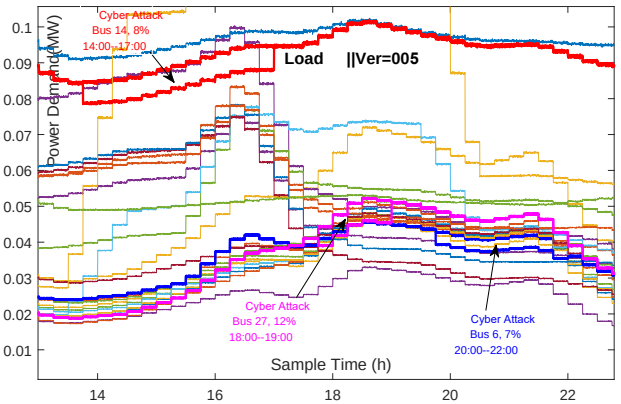
[9] T. Hong, C. Chen, J. Huang, N. Lu, L. Xie, and H. Zareipour, "Guest editorial big data analytics for grid modernization," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2395–2396, Sept 2016.



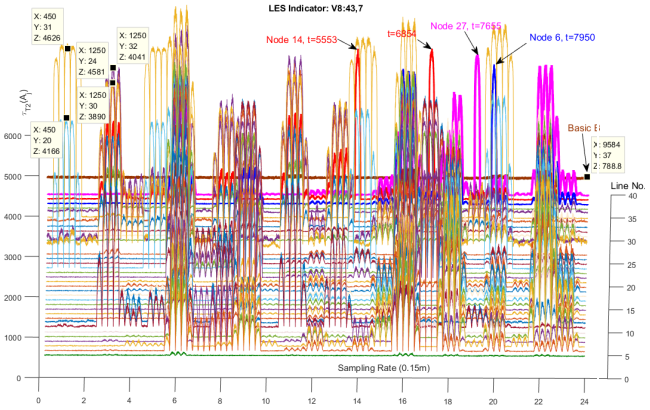
(a) Active Power $P_{\Sigma k,j}$ of each node



(b) Voltage $U_{i,j}$ of each node

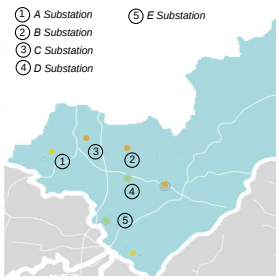


(c) Zoom in of Power $P_{\Sigma k,j}$

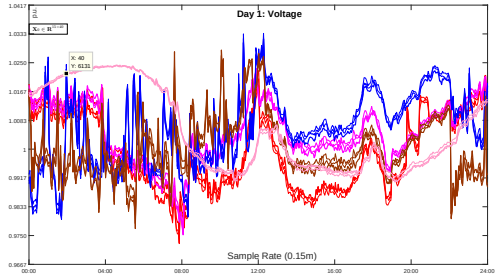


(d) $t - \tau_{T_2}$ Curve

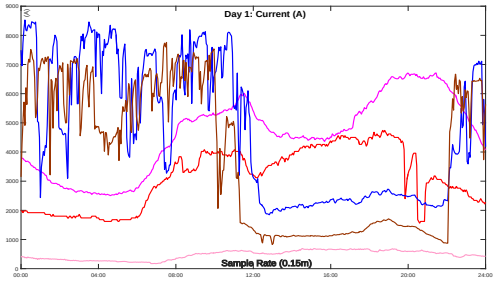
Fig. 5: Illusion of the data and analysis of a complicated scenario for behavior analysis.



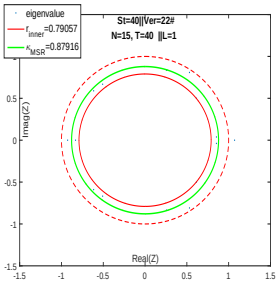
(a) The Grid Network



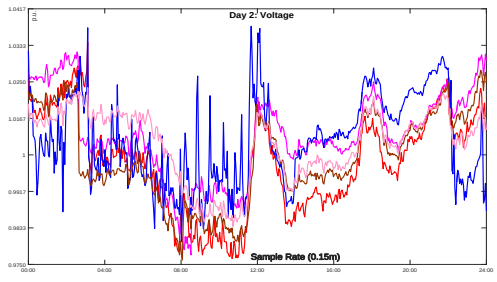
(c) Day 1: Voltage V



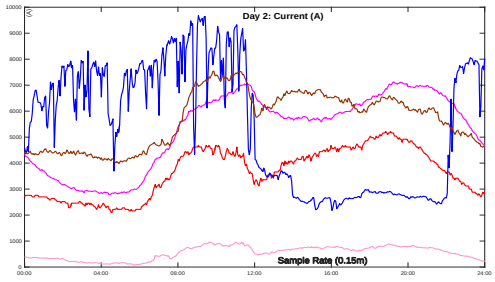
(e) Day 1: Current I



(b) Ring Law of $\mathbf{X}_0$



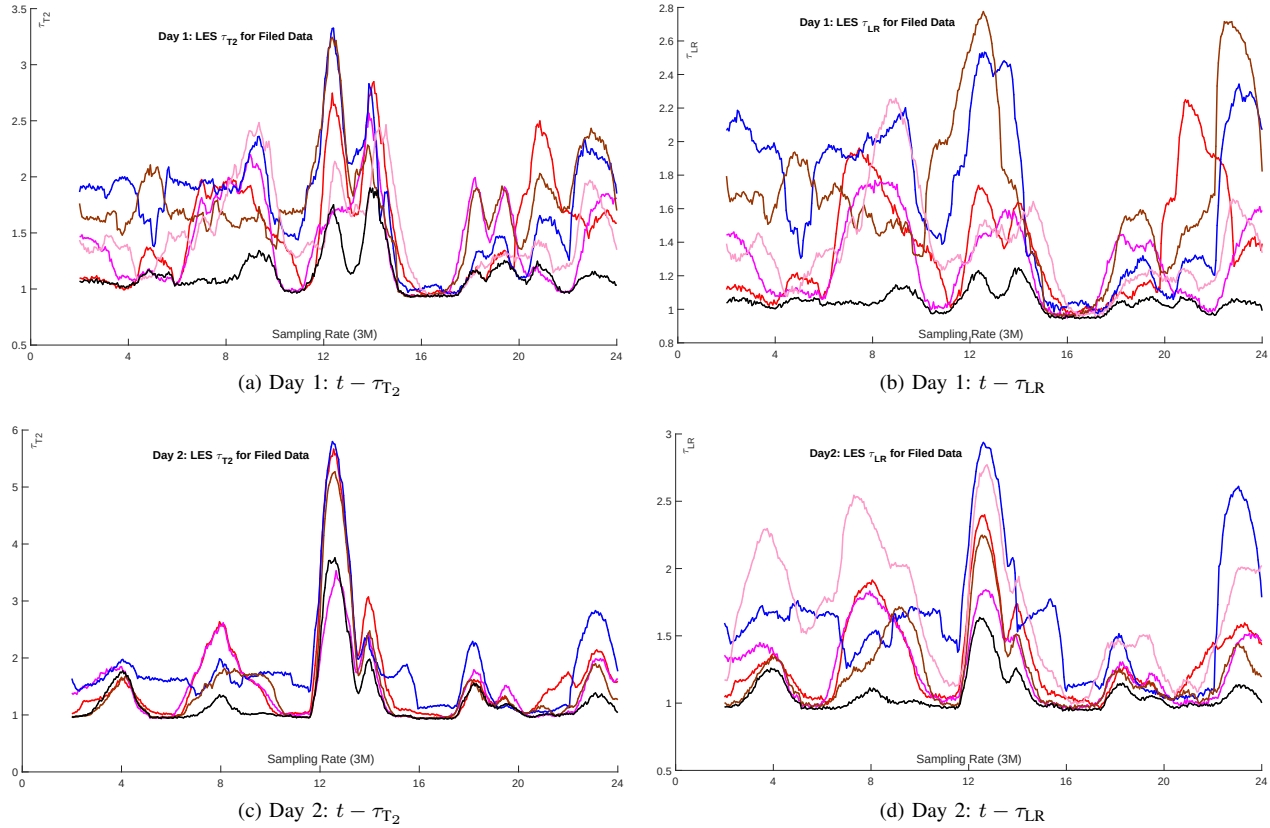
(d) Day 2: Voltage V



(f) Day 2: Current I

Fig. 6: Grid Network and Raw Data of Real Case

Note: For each substation, the 3-phase data are quite similar and only B-phase data are chosen.



(a) Day 1: $t - \tau_{\mathrm{T}_2}$ (b) Day 1: $t - \tau_{\mathrm{LR}}$

(c) Day 2: $t - \tau_{\mathrm{T}_2}$ (d) Day 2: $t - \tau_{\mathrm{LR}}$

Fig. 7: Illusion of the LES indicators of field data.

[10] X. Zhang and S. Grijalva, "A data-driven approach for detection and estimation of residential pv installations," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2477–2485, Sept 2016.

[11] H. Jiang, X. Dai, D. W. Gao, J. J. Zhang, Y. Zhang, and E. Muljadi, "Spatial-temporal synchrophasor data characterization and analytics in smart grid fault detection, identification, and impact causal analysis," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2525–2536, Sept 2016.

[12] X. Xu, X. He, Q. Ai, and C. Qiu, "A correlation analysis method for power systems based on random matrix theory," *IEEE Transactions on Smart Grid*, vol. PP, no. 99, pp. 1–10, 2016. [Online]. Available: http://arxiv.org/pdf/1506.04854.pdf

[13] X. He, R. C. Qiu, Q. Ai, L. Chu, X. Xu, and Z. Ling, "Designing for situation awareness of future power grids: An indicator system based on linear eigenvalue statistics of large random matrices," *IEEE Access*, vol. 4, pp. 3557–3568, 2016. [Online]. Available: http://arxiv.org/pdf/1512.07082.pdf

[14] X. He, L. Chu, Q. Ai, R. C. Qiu, and Z. Ling, "A Data-driven Situation Awareness Method Based on Random Matrix for Future Grids," *ArXiv e-prints*, Oct. 2016. [Online]. Available: https://arxiv.org/pdf/1610.05076.pdf

[15] L. Chu, R. C. Qiu, X. He, Z. Ling, and Y. Liu, "Massive streaming pmu data modeling and analytics in smart grid state evaluation based on multiple high-dimensional covariance tests," *IEEE Transactions on Big Data*, vol. PP, no. 99, 2017.

[16] E. P. Wigner, "On the distribution of the roots of certain symmetric matrices," *Annals of Mathematics*, vol. 67, no. 2, pp. 325–327, Mar. 1958. [Online]. Available: http://www.jstor.org/stable/197008

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

[18] F. Götze and A. Tikhomirov, "The rate of convergence for spectra of gue and lue matrix ensembles," *Open Mathematics*, vol. 3, no. 4, pp. 666–704, 2005.

[19] A. Lytova, L. Pastur *et al.*, "Central limit theorem for linear eigenvalue statistics of random matrices with independent entries," *The Annals of Probability*, vol. 37, no. 5, pp. 1778–1840, 2009.

[20] M. Shcherbina, "Central limit theorem for linear eigenvalue statistics of the wigner and sample covariance random matrices," *ArXiv e-prints*, Jan. 2011. [Online]. Available: http://arxiv.org/pdf/1101.3249.pdf

[21] R. Qiu and P. Antonik, *Smart Grid and Big Data*. John Wiley and Sons, 2015.

[22] D. Siegmund, "Change-points: From sequential detection to biology and back," *Sequential Analysis*, vol. 32, no. 1, pp. 2–14, 2013.