A Novel Data-Driven Situation Awareness Approach for Future Grids—Using Large Random Matrices for Big Data Modeling

Xing He, Lei Chu, Robert C. Qiu, Fellow, IEEE, Qian Ai, Senior Member, IEEE, Zenan Ling

Abstract—Data-driven approaches, when tasked with situation awareness, are suitable for complex grids with massive datasets. It is a challenge, however, to efficiently turn these massive datasets into useful big data analytics. To address such a challenge, this paper, based on random matrix theory (RMT), proposes a data-driven approach. The approach models massive datasets as large random matrices; it is model-free and requiring no knowledge about physical model parameters. In particular, the large data dimension $N$ and the large time span $T$, from the spatial aspect and the temporal aspect respectively, lead to favorable results. The beautiful thing lies in that these linear eigenvalue statistics (LESs) built from data matrices follow Gaussian distributions for very general conditions, due to the latest breakthroughs in probability on the central limit theorems of those LESs. Numerous case studies, with both simulated data and field data, are given to validate the proposed new algorithms.

Index Terms—Big data analytics, situation awareness, random matrix theory, linear eigenvalue statistics, statistical indicator

I. INTRODUCTION

SITUATION awareness (SA) is of great significance for power system operation, and a reconsideration of SA is essential for future grids [1]. These future grids are always huge in size and complex in topology. Operating under a novel regulation, their management mode is much different [2]. Data are more and more easily accessible, on the other hand, and data-driven approaches become natural for future grids. Towards this vision, we are facing the following challenges:

• There are massive data in power grids. The so-called curse of dimensionality [3] occurs inevitably.
• The resource cost (time, hardware, human, etc.) for extracting big data analytics should be tolerable.
• For a massive data source, there often exist realistic “bad” data, e.g. the incomplete, the inaccurate, the asynchronous, and the unavailable. For system operations, decisions such as relay actions, should be highly reliable.

This paper is built upon our previous work in the last several years. See Section II-B for details. Motivated for data mining, our line of research is based on the high-dimensional statistics. By high-dimensionality, we mean that the datasets are represented in terms of large random matrices. These data matrices can be viewed as data points in high-dimensional vector space—each vector is very long.

Data-driven approaches and data utilization for smart grids are current stressing topics, as evidenced in the special issue of “Big Data Analytics for Grid Modernization” [1]. This special issue is most relevant to our paper in spirit. Several SA topics are discussed. We highlight the anomaly detection and classification [4–5], the estimation of active ingredients such as PV installations [6,7], and finally the online transient stability evaluation using real-time data [8].

In addition, we point out research about the improvement in wide-area monitoring, protection and control (WAMPAC) and the utilization of PMU data [9,11], together with the fault detection and location [12,13]. Xie et al. based on principal component analysis (PCA), propose an online application for early event detection by introducing a reduced dimensionality [14]. The work has a special connection to our paper. Lim et al. study the quasi-steady-state operational problem relevant to the voltage instability phenomenon based on SVD using PMU data [15].

A. Contributions of Our Paper

Randomness is critical to future grids since rapid fluctuations in voltages and currents are ubiquitous. Often, these fluctuations exhibit Gaussian statistical properties [15]. Our central interest in this paper is to model these rapid fluctuations using the framework of random matrix theory (RMT). Our new algorithms are made possible due to the latest breakthroughs in probability on the central limit theorems of the linear eigenvalue statistics (LESs) [16, Chapter 7]. See [17] for a recent review.

1) Starting from fundamental formulas of power systems, a theoretical justification is given for the validity of modeling complex grids as large random matrices. This data modeling framework ties together the RMT and the power system analysis. This part is basic in nature.
2) We study numerous basic problems including the technical route and applied framework, data-processing and relevant procedures, evaluation system and indicator sets, and the advantages over classical methodologies.
3) We make a comparison between RMT-based approach and PCA-based one.
4) On the basis of big data analytics, we study some power system applications: anomaly detection and location, empirical spectral density test, sensitivity analysis, statistical...
indicator system and its visualization, and, finally, robustness against asynchronous data.

B. Relationship to Our Previous Work

Our work [2] is the first attempt to introduce the mathematical tool of RMT into power systems. Later, numerous papers demonstrate the power of this tool. Ring Law and Marchenko-Pastur (M-P) Law are regarded as the statistical foundation, and Mean Spectral Radius (MSR) is proposed as the high-dimensional indicator. Then we move forward to the second stage—paper [18] studies the correlation analysis under the above framework. The concatenated matrix \( A_i \) is the object of interest. It consists of the basic matrix \( B \) and a factor matrix \( C_i \), i.e., \( A_i = [B : C_i] \). In order to seek the sensitive factors, we compute the advanced indicators that are based on the LESs of these concatenated matrices \( A_i \). This study contributes to fault detection and location, line-loss reduction, and power-stealing prevention. We also conduct analysis for power transmission equipment based on the same theoretical foundation [19]. Paper [20] is the third step in which the LES set is studied. Based on the LES set, a statistical and data-driven indicator system, rather than its deterministic and model-based counterpart, is built to describe the system from a high-dimensional perspective. The robustness against spatial data error, precisely, data losses in the core area, is emphasized.

C. Advantages of RMT-based Approach

The data-driven RMT-based approach conducts analysis requiring no prior knowledge of the system topology, the unit operation/control mechanism, the causal relationship, etc. Comparing with classical data-driven methodologies such as PCA-based method, the RMT-based counterpart has some unique advantages:

1) The massive dataset of power systems are in a high-dimensional vector space; the temporal variations (\( T \) sampling instants) are simultaneously observed together with spatial variations (\( N \) grid nodes). The extraction of information from the above temporal-spatial variations is a challenge that does not meet the prerequisites of most classical statistical algorithms. Unifying time and space through their ratio \( c = T/N \), RMT deal with such kind of data mathematically rigorously.

2) The statistical indicator is generated from all the data in the form of matrix entries. This is not true to principal components—we really do not know the rank of the covariance matrix. The large size of the data enhances the robustness of the final decision against the bad data (inaccuracy, losses, or asynchronization), as well as those challenges in classical data-driven methods, such as error accumulations and spurious correlations [18].

3) For the statistical indicator, a theoretical or empirical value can be obtained in advance. The statistical indicator such as LES follows a Gaussian distribution, and its variance is bounded [21] and decays very fast in the order of \( O(N^{-2}) \) for a given data dimension \( N \), say \( N = 118 \).

4) We can flexibly handle heterogenous data to realize data fusion via matrix operations, such as the blocking [2], the sum [22], the product [22], and the concatenation [18] of the matrices. Data fusion is guided by the latest mathematical research [16], Chapter 7.

5) Only eigenvalues are used for further analyses, while the eigenvectors are omitted. This leads to a much faster data-processing speed and less required memory space. Although some information is lost, there is still rich information contained in the eigenvalues [23], especially those outliers [24, 25].

6) Particularly, for a certain RMM, various forms of LES, in the form of \( \tau = \sum_{i=1}^{N} \varphi (\lambda_{N_i}) \), can be constructed by designing test functions \( \varphi (\cdot) \) without introducing any system error. Each LES, similar to a filter, provides a unique view-angle. As a result, the system is understood piece by piece. With a proper LES, we can trace some specific signal.

Section II gives the mathematical background and theoretical foundation. Spectrum test is introduced as a novel tool. Section III studies the details about the RMT-based method. Section IV and Section V, using the simulated data and field data respectively, study the function designing based on the proposed method. Section VI concludes this paper.

II. MATHEMATICAL BACKGROUND AND THEORETICAL FOUNDATION

A. Random Matrix Modeling

Operating in a balance situation, power grids obey

\[
\begin{align*}
\Delta P_i &= P_{is} - P_i(V, \theta) \\
\Delta Q_i &= Q_{is} - Q_i(V, \theta)
\end{align*}
\]

where \( P_{is} \) and \( Q_{is} \) are the power injections of node \( i \), and \( P_i(V, \theta) \) and \( Q_i(V, \theta) \) are the power injections of the network, satisfying

\[
\begin{align*}
P_i &= V_i \sum_{j=1}^{n} V_j (G_{ij} \cos \theta_{ij} + B_{ij} \sin \theta_{ij}) \\
Q_i &= V_i \sum_{j=1}^{n} V_j (G_{ij} \sin \theta_{ij} - B_{ij} \cos \theta_{ij})
\end{align*}
\]

Combining (1) and (2), we obtain

\[
w_0 = f(x_0, y_0),
\]

where \( w_0 \) is the vector of nodes’ power injections depending on \( P_{is}, Q_{is}, x_0 \) is the system status variables depending on \( V_i, \theta_i \), and \( y_0 \) is the network topology parameters depending on \( B_{ij}, G_{ij} \).

Then, the system fluctuations, thus randomness in datasets, are formulated as

\[
w_0 + \Delta w = f(x_0 + \Delta x, y_0 + \Delta y).
\]

With a Taylor expansion, (4) is rewritten as

\[
w_0 + \Delta w = f(x_0, y_0) + f_x(x_0, y_0) \Delta x + f_y(x_0, y_0) \Delta y
\]

\[
+ \frac{1}{2} f_{xx}(x_0, y_0)(\Delta x)^2 + \frac{1}{2} f_{yy}(x_0, y_0)(\Delta y)^2
\]

\[
+ f_{xy}(x_0, y_0) \Delta x \Delta y + \cdots.
\]


The value of system status variables \( \mathbf{x} \) are relatively stable, which means we can ignore the second-order term \( (\Delta \mathbf{x})^2 \) and higher-order terms. Besides, \( 3 \) shows that \( f''_{xy}(x, y) = 0 \). As a result, \( 3 \) is turned into

\[
\Delta \mathbf{w} = f'_x(x_0, y_0) \Delta \mathbf{x} + f'_y(x_0, y_0) \Delta \mathbf{y} + f''_{xy}(x_0, y_0) \Delta \mathbf{x} \Delta \mathbf{y}.
\]

Suppose the network topology is unchanged, i.e., \( \Delta \mathbf{y} = 0 \). From \( 3 \), we deduce that

\[
\Delta \mathbf{x} = (f'_{x}(x_0, y_0))^{-1} (\Delta \mathbf{w}) = S_0 \Delta \mathbf{w}.
\]

On the other hand, suppose the power demands is unchanged, i.e., \( \Delta \mathbf{w} = 0 \). From \( 3 \), we deduce that

\[
\Delta \mathbf{x} = S_0 \Delta \mathbf{w}_y,
\]

where \( \mathbf{w}_y = [I + f''_{xy}(x_0, y_0) \Delta \mathbf{s}_0]^{-1} [f'_y(x_0, y_0)] \).

Note that \( S_0 = (f'_{x}(x_0, y_0))^{-1} \), i.e., the inversion of the Jacobian matrix \( J_0 \).

Thus, we describe the power system operation using a random vector. If there exists an unexpected active power change or short circuit, the corresponding change of system status variables \( x_0 \), i.e., \( V_1, \theta_1 \), will obey \( 3 \) or \( 4 \) respectively.

For a practical system without dramatic changes, rich statistical empirical evidence indicates that the Jacobian matrix \( J \) keeps nearly constant, so does \( S_0 \). Considering \( T \) random vectors observed at time instants \( i = 1, \ldots, T \), we build a relationship in the form of \( \Delta \mathbf{X}_v = S_0 \Delta \mathbf{W} \) with a similar procedure as \( 3 \) to \( 5 \), where \( \Delta \mathbf{X}_v \) denotes the variation of state \( [\Delta x_1, \ldots, \Delta x_T] \), and \( \Delta \mathbf{W} \) denotes the variation of power injections or topology parameters accordingly.

Taking the case in \( 20 \) as an example, for an equilibrium operation system (the topology is unchanged, the reactive power is almost constant or changes much more slowly than the active one), the relationship model between voltage magnitude and active power is just like the Multiple Input Multiple Output (MIMO) model in wireless communication \( 16, 22 \). We write \( \mathcal{V} = \mathbb{E} \mathcal{P} \). Note that most variables of vector \( \mathcal{V} \) are random due to the ubiquitous noises, e.g., small random fluctuations in \( \mathcal{P} \). Furthermore, with the normalization, we can build the standard random matrix model (RMM) in the form of \( \tilde{\mathcal{V}} = \mathbb{E} \mathcal{R} \), where \( \mathcal{R} \) is a standard Gaussian random matrix.

### B. Anomaly Detection Based on Asymptotic Empirical Spectral Distribution

Often, these rapid fluctuations exhibit Gaussian statistical properties \( 15 \), as pointed out above. In practice, Gaussian unitary ensemble (GUE) and Laguerre unitary ensemble (LUE) are used in our models:

\[
\mathbf{A} = \begin{cases} \frac{1}{2} (\mathbf{X} + \mathbf{X}^H), & \mathbf{X} \in \mathbb{C}^{N \times N}, \text{GUE}; \\ \frac{1}{N} \mathbf{X} \mathbf{X}^H, & \mathbf{X} \in \mathbb{C}^{N \times T}, \text{LUE}. \end{cases}
\]

where \( \mathbf{X} \) is the standard Gaussian random matrix whose entries are independent identically distributed (i.i.d.) complex Gaussian random variables.

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

\[
F_{\mathbf{A}}(x) = \frac{1}{N} \sum_{i=1}^{N} I_{\{\lambda_i \leq x\}},
\]

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

Let \( g_{\mathbf{A}}(x) \) and \( G_{\mathbf{A}}(x) \) denote the true eigenvalue density and the true spectral distribution of \( \mathbf{A} \), and the Wigner's Semicircle Law and Wishart's M-P Law say:

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

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

\[
G_{\mathbf{A}}(x) = \int_{-\infty}^{x} g_{\mathbf{A}}(u) du.
\]

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

\[
\Delta = \sup_{x} \left| \mathbb{E}\{F_{\mathbf{A}}(x)\} - G_{\mathbf{A}}(x) \right|.
\]

Gotze and Tikhomirov, in their work \( 26 \), prove an optimal bound for \( \Delta \) of order \( O(N^{-1}) \).

**Lemma II.1.** There exists a positive constant \( C \) such that, for any \( N \geq 1 \),

\[
\Delta \leq CN^{-1}.
\]

They also prove that the convergence of the density of standard Semicircle Law and M-P Law to the expected spectral density \( f_{\mathbf{A}}(x) \) satisfies following lemmas.

**Lemma II.2.** For GUE matrix, there exists a positive constant \( \varepsilon \) and \( C \) such that, for any \( x \in [-2 + N^{-\frac{4}{5}} \varepsilon, -2 - N^{-\frac{4}{5}} \varepsilon] \),

\[
|f_{\mathbf{A}}(x) - g_{\mathbf{A}}(x)| \leq \frac{C}{N(4 - x^2)}.
\]

**Lemma II.3.** For LUE matrix, let \( \beta = N/T \), there exists some positive constant \( \beta_1 \) and \( \beta_2 \) such that \( 0 < \beta_1 \leq \beta \leq \beta_2 < 1 \), for all \( N \geq 1 \). Then there exists a positive constant \( C \) and \( \varepsilon \) depending on \( \beta_1 \) and \( \beta_2 \) and for any \( N \geq 1 \) and \( x \in [a + N^{-\frac{4}{5}} \varepsilon, b - N^{-\frac{4}{5}} \varepsilon] \),

\[
|f_{\mathbf{A}}(x) - h_{\mathbf{A}}(x)| \leq \frac{C}{N(x-a)(b-x)}.
\]

Lemma II.2 and II.3 also describe how fast the population distribution functions converge to the asymptotic empirical spectral distribution limit. This ESD-based test is interesting for anomaly detection about a complex grid; the effectiveness is validated in Section IV. We exploit the mathematical knowledge that the ESD converges to its limit with an optimal convergence rate of \( N^{-1} \).
III. THE METHOD OF SITUATION AWARENESS

A. Technical Route and Practical Procedures

The proposed RMT-based method consists of three procedures as illustrated in Fig. 1: 1) big data model—to model the system using experimental data for the RMM; 2) big data analysis—to conduct big data analytics for the indicator system; 3) engineering interpretation—to visualize and interpret the statistical results to operators for the decision-making.

![Fig. 1: RMT-based Method for SA](image)

This method is universal. We have already made numerous successful attempts in the field of anomaly detection and diagnosis for both the grid network and the transmission equipment. In addition, Zhang et al., based on RMT, study the steady stability and transient stability in [27] and [28] respectively.

B. Paradigms and Method

We would like to refer to Fig. 2 in book [29] as a clue. We are now entering the age of 4th-paradigm—data-intensive scientific discovery. Besides, the summaries for the classical decision-making approaches and for our proposed ones, obtained initially in [2], are improved as Fig. 3.

![Fig. 2: Science paradigms](image)

The second and third paradigms are typically model-based—they use equations, formulas, and simulations to describe the system. The blue line in Fig. 3 depicts the general procedure and corresponding tools. These tools cannot deal with massive data due to the essence of mechanism models—the models are in low dimensions, leading to deterministic results which are fully dependent upon only a few parameters. It will raise some problems, causing inefficient or even incorrect big data analytics. For instance, only under ideal conditions, is the wind power proportional to the cube of wind speed. Moreover, some physical parameters, e.g., admittance matrices, will introduce system error due to the ubiquitous randomness and uncertainty.

Under classical statistical framework, only two typical data matrices in the form of \( X \in \mathbb{R}^{N \times T} \) are at our disposal: 1) \( N, T \) are small, and 2) \( N \) is small, \( T \) is very large (compare with \( N \)). This prerequisite greatly restricts the utilization of the massive data; we should enable more data to speak for themselves. In other words, model-based framework is not able to turn massive data into useful big data analytics. Although these massive data can contribute to model improvement and parameters correction, we can hardly conduct analysis more precisely with extremely large data volumes. Even worse, more data mean more errors; if we take those bad data into the fixed model, poor results are obtained almost surely. Besides, the bias, caused by challenges such as error accumulations and spurious correlations, will not be alleviated via a low-dimensional procedure—the dimensions of the procedure are limited by the dimensions of the model. The belief that data-driven mode is adapted to the future grid’s analysis agrees with the core viewpoint of the 4th-paradigm. The classical data utilization methodology needs be revisited.

C. Classical Dimensionality Reduction Algorithm—PCA

Data-driven methodology is an alternative; it is model-free and able to process massive data in a holistic way. Principal component analysis (PCA) is one of the classical data processing algorithms which are sensitive to relative scaling original variables. It uses an orthogonal transformation to convert a set of possibly correlated raw variables into a set of linearly uncorrelated variables called principal components. The number of principal components is often much less than the number of original variables. In [14], PCA is used for dimensionality reduction from 14 PMU datasets to extract the event indicators. For PCA, the procedure consists of three parts: 1) Singular Value Decomposition (SVD); 2) Projection, and 3) Indicators.

This procedure is applied to conduct early event detection; details can be found in [14]. We will make a comparison between the PCA-based approach and the RMT-based approach, and the advantages of the later is summarized in C.

D. Data-Driven Approach Based on Random Matrix Theory

The procedure based on RMT is outlined below.

1) Ring Law and MSR: Ring Law Analysis conducts SA as follows:

\[ y = ax^2 + bx + c \]

E.g., \( y = ax^2 + bx + c \) is a 3-dimensional model—the relationship between \( x \) and \( y \) fully depends on \( a, b, \) and \( c \).
Fig. 3: Data utilization method for power systems. The above, middle, and below parts indicate the data processing procedures and the work modes for G1, G2, and G3, respectively.

Steps of Ring Law Analysis

1) Select arbitrary raw data (or all available data) as data source $\Omega$.
2) At a certain time $t_i$, form $\mathbf{X}$ as random matrix.
3) Obtain $\mathbf{Z}$ by matrix transformations ($\mathbf{X} \rightarrow \mathbf{X} \rightarrow \mathbf{X}_t \rightarrow \mathbf{Z} \rightarrow \mathbf{Z}^\top$).
4) Calculate eigenvalues $\lambda_2$ and plot the Ring on the complex plane.
5) Conduct high-dimensional analysis.
6a) Observe the experimental ring and compare it with the reference.
6b) Calculate $\hat{\theta}_{\text{MSR}} = \sum_{i=1}^{N} |\mathbf{X}_t| / N$ as the statistical indicators.
6c) Compare $\hat{\theta}_{\text{MSR}}$ with the theoretical value $\mathcal{E}(\theta_{\text{MSR}})$.
7) Repeat 2)-6) at the next time point ($t_i = t_{i+1}$).
8) Visualize $\theta$ on the time series, i.e. draw $\theta$-t curve.
8) Make engineering explanations.

In Steps 2–7, with a high-dimensional procedure, one conducts SA without any prior knowledge, assumption, or simplification. In step 2, arbitrary raw data, even those from distributed nodes or intermittent time periods, are at our disposal. The size of $\mathbf{X}$ is controllable, and as a result the dimensionality curse is relieved in some ways.

2) M-P Law and LES: For the M-P Law Analysis, the steps are very similar, except for the following differences:

Partial Steps of M-P Law Analysis

3) Obtain $\mathbf{M}$ by matrix transformations ($\mathbf{M} = \mathbf{X}\mathbf{X}^\top$).
4) Calculate eigenvalues $\lambda_M$.
5a) Calculate $\tau_{\text{MSR}} = \sum_{i=1}^{N} \phi(\lambda_{M,i})$ as the statistical indicators.
5b) Compare $\tau$ with the theoretical value $\mathcal{E}(\tau)$.

Notice that Ring Law maps the information from datasets to the complex plane ($\mathbb{C}^{N \times T} \mapsto \mathbb{C}$), while M-P law does this to the right half real-axis ($\mathbb{C}^{N \times T} \mapsto \mathbb{R}^+$). This fundamental difference plays a critical role in data visualization.

IV. CASE STUDIES USING SIMULATED DATA

A. Background and Assumption of the Case

We adopt a standard IEEE 118-node system as Fig. 16 and assume the events as Table II. Thus, the power demand on each node is obtained as the system injections (Fig. 2a), the voltage is also obtained (Fig. 15). Suppose that the power demand data is unknown or unqualified for SA due to the low sampling frequency or the bad quality. For further analysis, we just start with data source $\Omega$:

- For $\mathbf{v} : \mathbf{v}_{i,j} \in \mathbb{R}^{118 \times 2500}$ and assign the analysis matrix as $\mathbf{X} \in \mathbb{R}^{118 \times 2500}$ (4 minutes’ time span). Firstly, we conduct category for the system operation status; the results are given in Fig. 4c. In general, according to the
data feature (events on time-series) and the matrix length (time span, i.e., \(T\)), we divide the operation status into 8 stages. Note that \(S_4\), \(S_5\), and \(S_6\) are transition stages, and their time span is right equal to the analysis matrix length minus ones, i.e., \(T-1=239\). These stages are described as follows:

- For \(S_0\), \(S_1\), \(S_2\), white noises play a dominant part. \(P_{Node,52}\) is rising in turn.
- For \(S_3\), \(P_{Node,52}\) keeps a sustained and stable growth.
- \(S_4\), transition stage. Ramping signal exists.
- \(S_5\), \(S_6\), transition stages. Step signal exists.
- For \(S_7\), voltage collapse.

Two typical data sections, at stage \(S_0\) and \(S_6\) respectively, are selected: \(X_0 \in \mathbb{R}^{118 \times 240}\), covering period \(t=[61:300]\) and at sampling time \(t_{end}=300\), and 2) \(X_6 \in \mathbb{R}^{118 \times 240}\), covering period \(t=[662:901]\) and at sampling time \(t_{end}=901\).

**B. Anomaly Detection**

1) **Based on Ring Law and M-P Law:**

According to our previous work [2], RMM \(\hat{V}\) is build from the raw voltage data. Then, \(\tau_{MSR}\) is employed as a statistical indicator to conduct anomaly detection. For the selected data section \(X_0\) and \(X_6\), their M-P Law and Ring Law Analysis are shown as Fig 5a, 5b, 5c and 5d. With sliding-window, the \(\tau_{MSR-t}\) curve is obtained as Fig. 5e.

Fig. 5 shows that when there is no signal in the system, the experimental RMM well matches Ring Law and M-P Law, and the experimental value of LES is approximately equal to the theoretical value. This validates the theoretical justification for modeling rapid fluctuation of each node using white Gaussian noises, as shown in Section II-A. On the other hand, Ring Law and M-P Law are violated at the very beginning \((t_{end}=901)\) of the step signal. Besides, the proposed high-dimensional indicator \(\tau_{MSR}\) is extremely sensitive to the anomaly—at \(t_{end}=901\), the \(\tau_{MSR}\) starts the dramatic change (Fig. 5e, \(\tau_{MSR-t}\) curve), while the raw voltage magnitudes are still in the normal range (Fig. 4c). Moreover, following the previous work [20], we design numerous kinds of LES \(\tau\) and define \(\mu_0 = \tau/E(\tau)\). The results are shown in Fig. 6 proving that different indicators have different effectiveness; this suggests another topic to explore in the future.

2) **Based on Spectrum Test:**
We still set the sampling time at \( t_{\text{end}} = 300 \) and \( t_{\text{end}} = 901 \). Following Lemma II.2 and Lemma II.3, \( \mathbf{Y}_0, \mathbf{Y}_\theta \in \mathbb{R}^{118 \times 240} \) (span \( t = [61:300] \) and \( t = [662:901] \)), and \( \mathbf{Z}_0, \mathbf{Z}_\theta \in \mathbb{R}^{118 \times 118} \) (span \( t = [183 : 300] \) and \( t = [784 : 901] \)) are selected. The results are shown in Fig. 7 and Fig. 8. These results validate that empirical spectral density test is competent to conduct anomaly detection—when the power grid is under a normal condition, the empirical spectral density \( f_\lambda(x) \) and the ESD function \( F_\lambda(x) \) are almost strictly bounded between the upper bound and the lower bound of their asymptotic limits. On the other hand, these results also validate that GUE and LUE are proper mathematical tools to model the power grid operation.

\[ \text{C. Steady Stability Evaluation} \]

The \( V - P \) curve (also called nose curve) and the smallest eigenvalue of the Jacobian Matrix \( \mathbf{J} \) are two clues for steady stability evaluation. In this case, we focus on \( E4 \) stage during which \( P_{\text{Node},E2} \) keeps increasing until the system exceeds its steady stability limit. The \( V - P \) curve and \( \lambda - P \) curve, respectively, are given in Fig. 9a and Fig. 9b. Furthermore, we choose some data section, \( T_1 : [1601 : 1840] \); \( T_2 : [1901 : 2140] \); \( T_3 : [2101 : 2340] \), shown as Fig. 9a. The RMT-based results are shown as Fig. 10. The outliers become more evident as the stability degree decreases. The statistics of the outliers, in some sense, are similar to the smallest eigenvalue of Jacobian Matrix, Lyapunov Exponent or the entropy.

For further analysis, we take the signal and stage division into account. In general, sorted by the stability degree, the stages are ordered as \( S_0 > S_1 > S_2 \gg \max(S_3, S_4, S_5) > \min(S_3, S_4, S_5) \approx S_6 \approx S_7 \). According to Fig. 6, we make the Table I. The high-dimensional indicators \( \gamma_{K\mathbf{R}} \) has the same trend as the stability degree order. These statistics have the potential for data-driven stability evaluation. Moreover, based on the Gaussian property of LES indicators, hypothesis tests are designed for the anomaly detection; see [32] for details.
TABLE I: Indicator of Various LESs at Each Stage.

| Node | Pre-Fault | Post-Fault | A | B | C | D | E | F |
|------|-----------|------------|---|---|---|---|---|---|
| S0   | 0.165     | 0.180      | 0.174 | 0.182 | 0.168 | 0.180 | 0.168 | 0.184 |
| S1   | 0.165     | 0.180      | 0.174 | 0.182 | 0.168 | 0.180 | 0.168 | 0.184 |
| S2   | 0.165     | 0.180      | 0.174 | 0.182 | 0.168 | 0.180 | 0.168 | 0.184 |
| S3   | 0.165     | 0.180      | 0.174 | 0.182 | 0.168 | 0.180 | 0.168 | 0.184 |
| S4   | 0.165     | 0.180      | 0.174 | 0.182 | 0.168 | 0.180 | 0.168 | 0.184 |
| S5   | 0.165     | 0.180      | 0.174 | 0.182 | 0.168 | 0.180 | 0.168 | 0.184 |
| S6   | 0.165     | 0.180      | 0.174 | 0.182 | 0.168 | 0.180 | 0.168 | 0.184 |
| S7   | 0.165     | 0.180      | 0.174 | 0.182 | 0.168 | 0.180 | 0.168 | 0.184 |

*τX_R = τX/|E(τ)|.

D. Correlation Analysis

The key for correlation analysis is the concatenated matrix \( A_i \), which consist of two part—the basic matrix \( B \) and a certain factor matrix \( C_i \), i.e., \( A_i = [B; C_i] \). For details, see our previous work [18]. The LES of each \( A_i \) is computed in parallel, and Fig. [11] shows the results.

In Fig. [11] the blue dot line (marked with None) shows the LES of basic matrix \( B \), and the orange line (marked with Random) shows the LES of the concatenated matrix \( [B; R] \) (\( R \) is the standard Gaussian Random Matrix). Fig. [11] demonstrates that: 1) node 52 is the causing factor of the anomaly; 2) sensitive nodes are 51, 53, and 58; and 3) nodes 11, 45, 46, etc., are not affected by the anomaly. Based on this algorithm, we can conduct behavior analysis, e.g., detection and estimation of residential PV installations [6]. It is another hot topic, and we expand it in [32].

V. CASE STUDIES USING FIELD DATA

Some power grid of China are selected, with 34 PMUs collecting power flow data. The raw data are shown as Fig. [13]. It is quite obvious that the fault begins at sampling time \( t_0 = 3271 \). The ring distribution and M-P law pre-fault (3101 – 3100), during fault (3173 – 3272), and post-fault (7201 – 7300) are given as Fig [14]. This implies that the real-world data do follow the Ring Law and M-P Law under normal condition, and they violate these laws when the fault is occurring. Moreover, the LES \( t - \tau \) curves of basic matrix \( B \) and concatenated matrix \( C_i \) are obtain as Fig. [15]. It shows that Node 8, 9, 26, 27, 28, 10, 11, 12 are most relevant to this fault; while Node 1 – 7 are not so sensitive.
data modeling and analysis. Our approach exploits the massive spatial-temporal datasets of power systems. Random matrix theory (RMT) appears very natural for the problem at hand; in a random matrix of $CN \times T$, we use $N$ nodes to represent the spatial nodes and $T$ data samples to represent the temporal samples. When the number of nodes $N$ is large, very unique mathematical phenomenon occurs, such as concentration of measure \[16\]. Phase transition as a function of data size $N$ is a result of this deep mathematical phenomenon. This is the very reason why the proposed algorithms are so powerful in practice.

Explicitly expressed in forms of linear eigenvalue statistics (LESs) \[17\], the proposed RMT-based algorithms have numerous unique advantages. They are especially suitable for complex systems. In the form of a large random matrix, they handle massive data that are in high dimensions and within a wide time span all at once. The trick is to treat these data as a whole at the disposal of RMT. In this way, highly reliable decisions are still attainable with some imperfect data, e.g., the asynchronous data. Moreover, with the statistical processing such as test function setting, the proposed data-driven approach has the potential to balance the perspectives of the speed, the sensitivity, and the reliability in practice.

The stability evaluation and behavior analysis are two big topics along this direction. Besides, the statistical indicators are good starting points for artificial intelligence and machine learning. For example, we can extract the linear eigenvalue statistics as features; those extracted features are used for further data processing in the pipeline using algorithms such as random forest, decision trees, and support vector machine. Our whole framework starts with the use of sample covariance matrix to replace the true covariance matrix. It is well known that this replacement is far from optimal. The almost optimal estimation of large covariance matrices using tools from RMT \[33\] can be used, instead.

**APPENDIX A**

This paper has made significant progress on the basis of our previous work in the context of big data analytics for future grids. Randomness and uncertainty are at the heart of this...
APPENDIX B

TABLE II: Series of Events

| Stage     | E1       | E2       | E3       | E4       |
|-----------|----------|----------|----------|----------|
| Time (s)  | 1–500   | 501–900  | 901–1300 | 1301–2500|
| \( P_{\text{node-2}} \) (MW) | 0       | 30       | 120      | t/4 – 205 |

\( P_{\text{node-2}} \) is the power demand of node 2.

The power demand of other nodes are assigned as

\[
y_{\text{load-nt}} = y_{\text{load-nt}} \times (1 + \gamma_{\text{Mul}} \times r_1) + \gamma_{\text{Acc}} \times r_2,
\]

where \( r_1 \) and \( r_2 \) are the element of standard Gaussian random matrix; \( \gamma_{\text{Acc}}=0.1, \gamma_{\text{Mul}}=0.001 \).

REFERENCES

[1] T. Hong, C. Chen, J. Huang et al., “Guest editorial big data analytics for grid modernization,” IEEE Transactions on Smart Grid, vol. 7, no. 5, pp. 2395–2396, Sept. 2016.

[2] X. He, Q. Ai, R. C. Qiu et al., “A big data architecture design for smart grids based on random matrix theory,” Jan. 2015, accepted by IEEE Trans on Smart Grid.

[3] L. Moulin, A. Alves da Silva, M. El-Sharkawi et al., “Support vector machines for transient stability analysis of large-scale power systems,” Power Systems, IEEE Transactions on, vol. 19, no. 2, pp. 818–825, 2004.

[4] M. Rafferty, X. Liu, D. M. Laverty, and S. McLoone, “Real-time multiple event detection and classification using moving window pca,” IEEE Transactions on Smart Grid, vol. 7, no. 5, pp. 2537–2548, Sept. 2016.

[5] H. Jiang, X. Dai, D. W. Gao et al., “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.

[6] 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.

[7] H. Shaker, H. Zareipour, and D. Wood, “A data-driven approach for estimating the power generation of invisible solar sites,” IEEE Transactions on Smart Grid, vol. 7, no. 5, pp. 2466–2476, Sept. 2016.

[8] B. Wang, B. Fang, Y. Wang, H. Liu, and Y. Liu, “Power system transient stability assessment from phasor measurement unit data,” IEEE Trans on Smart Grid, vol. 7, no. 5, pp. 2561–2570, Sept. 2016.

[9] A. Phadke and R. M. de Moraes, “The wide world of wide-area measurement,” Power and Energy Magazine, IEEE, vol. 6, no. 5, pp. 52–65, 2008.

[10] V. Terzija, G. Valverde, D. Cai, P. Regulski, V. Madani, J. Fitch, S. Skok, M. M. Begovic, and A. Phadke, “Wide-area monitoring, protection, and control of future electric power networks,” Proceedings of the IEEE, vol. 99, no. 1, pp. 80–93, 2011.

[11] L. Xie, Y. Chen, and H. Liao, “Distributed online monitoring of quasi-static voltage collapse in multi-area power systems,” Power Systems, IEEE Transactions on, vol. 27, no. 4, pp. 2271–2279, 2012.

[12] Q. Jiang, X. Li, B. Wang, and H. Wang, “Pmu-based fault location using voltage measurements in large transmission systems,” Power Delivery, IEEE Transactions on, vol. 27, no. 3, pp. 1644–1652, 2012.

[13] A. H. Al-Mohammed and M. Abido, “A fully adaptive pmu-based fault location algorithm for series-compensated lines,” Power Systems, IEEE Transactions on, vol. 29, no. 5, pp. 2129–2137, 2014.

[14] L. Xie, Y. Chen, and P. Kumar, “Dimensionality reduction of synchrophasor data for early event detection: Linearized analysis,” Power Systems, IEEE Transactions on, vol. 29, no. 6, pp. 2784–2794, 2014.

[15] J. M. Lim and C. L. DeMarco, “Svd-based voltage stability assessment from phasor measurement unit data,” IEEE Transactions on Power Systems, vol. PP, no. 99, pp. 1–9, 2015.

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

[17] R. C. Qiu, Big Data for Complex Network. CRC, 2016, ch. Large Random Matrices and Big Data Analytics.

[18] X. Xu, X. He, Q. Ai, and C. Qiu, “A correlation analysis method for power systems based on random matrix theory,” Jun. 2015, accepted by IEEE Trans on Smart Grid.

[19] Y. Yan, G. Sheng, H. Wang, Y. Liu, Y. Chen, X. Jiang, and Z. Guo, “The key state assessment method of power transmission equipment using big data analyzing model based on large dimensional random matrix,” Proceedings of the CSEE, vol. 36, no. 2, pp. 435–445, Jan. 2016.

[20] X. He, R. C. Qiu, Q. Ai, L. Chu, and X. Xu, “Linear eigenvalue statistics: An indicator ensemble design for situation awareness of power systems,” Dec. 2015.

[21] 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

[22] C. Zhang and R. C. Qiu, “Massive mimo as a big data system: Random matrix models and testbed,” IEEE Access, vol. 3, pp. 837–851, 2015.

[23] J. R. Ipsen and M. Kieburg, “Weak commutation relations and eigenvalue statistics for products of rectangular random matrices,” Physical Review E, vol. 89, no. 3, 2014, Art. ID 032106.

[24] F. Benaych-Georges and J. Rochet, “Outliers in the single ring theorem,” Probability Theory and Related Fields, pp. 1–51, May 2015. [Online]. Available: http://dx.doi.org/10.1007/s00440-015-0632-x

[25] T. Tao, “Outliers in the spectrum of iid matrices with bounded rank perturbations,” Probability Theory and Related Fields, vol. 155, no. 1–2, pp. 231–263, 2013.

[26] F. Götzé 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.

[27] Q. Wu, D. Zhang, D. Liu, W. Liu, and C. Deng, “A method for power system steady stability situation assessment based on random matrix theory,” 2016.

[28] W. Liu, D. Zhang, X. Wang, D. Liu, and Q. Wu, “Power system transient stability analysis based on random matrix theory,” 2016.

[29] A. J. Hey, S. Tansley, K. M. Tolle et al., The fourth paradigm: data-intensive scientific discovery. Microsoft Research Redmond, WA, 2009, vol. 1.

[30] J. Gray, “Jim gray on esience: A transformed scientific method,” The fourth paradigm: Data-intensive scientific discovery, pp. xvii–xxxi, 2009.

[31] R. Kitchin, “Big data and human geography opportunities, challenges and risks,” Dialogues in Human Geography, vol. 3, no. 3, pp. 262–267, 2013.

[32] X. He, R. Qiu, L. Chu, Q. Ai, Z. Ling, and J. Zhan, “Detection and estimation of the invisible units using utility data based on random matrix theory,” ArXiv e-prints, 2017. [Online]. Available: [http://arxiv.org/pdf/1710.10745.pdf]

[33] J. Bun, J.-P. Bouchaud, and M. Potters, “Cleaning large correlation matrices: tools from random matrix theory,” arXiv preprint arXiv:1610.08104, 2016.