A Big Data Architecture Design for Smart Grids Based on Random Matrix Theory

Xing He, Qian Ai, *Member, IEEE*, Robert Caiming Qiu, *Fellow, IEEE*, Wentao Huang, Longjian Piao, and Haichun Liu

***Abstract*—Model-based analysis tools, built on assumptions and simplifications, are difficult to handle smart grids with data characterized by volume, velocity, variety, and veracity (i.e., 4Vs data). This paper, using random matrix theory (RMT), motivates data-driven tools to perceive the complex grids in high-dimension; meanwhile, an architecture with detailed procedures is proposed. In algorithm perspective, the architecture performs a high-dimensional analysis and compares the findings with RMT predictions to conduct anomaly detections. Mean spectral radius (MSR), as a statistical indicator, is defined to reflect the correlations of system data in different dimensions. In management mode perspective, a group-work mode is discussed for smart grids operation. This mode breaks through regional limitations for energy flows and data flows, and makes advanced big data analyses possible. For a specific large-scale zone-dividing system with multiple connected utilities, each site, operating under the group-work mode, is able to work out the regional MSR only with its own measured/simulated data. The large-scale interconnected system, in this way, is naturally decoupled from statistical parameters perspective, rather than from engineering models perspective. Furthermore, a comparative analysis of these distributed MSRs, even with imperceptible different raw data, will produce a contour line to detect the event and locate the source. It demonstrates that the architecture is compatible with the block calculation only using the regional small database; beyond that, this architecture, as a data-driven solution, is sensitive to system situation awareness, and practical for real large-scale interconnected systems. Five case studies and their visualizations validate the designed architecture in various fields of power systems. To our best knowledge, this paper is the first attempt to apply big data technology into smart grids.**

***Index Terms*—Architecture, big data, group-work mode, high-dimension, large-scale distributed system, mean spectral radius (MSR), random matrix, smart grid.**

I. INTRODUCTION

BIG DATA technology is a new scientific trend [1], [2]. Driven by data analysis in high-dimension, big data technology works out data correlations (indicated by statistical parameters) to gain insight to the inherent mechanisms. Data-driven results only rely on an unrestrained selection of system raw data (the space can be whole system or only a region, the time can be long or short, and the size can be large or small) and a general statistical procedure (for data processing). On the other side, procedures for traditional model-based analysis, especially decoupling a practical interconnected system, are always based on assumptions and simplifications. Model-based results rely on identified causalities, specific parameters, sample selections, and training processes; imprecise or incomplete formulas/expressions, biased sample selections, and improper training processes will all lead to bad results. The results are often barely satisfied or even unsatisfied as the system size grows and complexity increases. Generally speaking, data-driven analysis tools, rather than model-based ones, are more suitable to complex large-scale interconnected systems with readily accessible data.

Data in power systems have increased dramatically, leaving gaps and challenges; data processing is a major concern and its urgency increases with data growth. The 4Vs data (data with features of volume, variety, velocity, and veracity) [3] in smart grids, which can hardly be handled within a tolerable elapsed time or hardware resources by traditional model-based tools, have encouraged the development of an emerging paradigm—big data technology for power systems [4]–[6]. Big data technology does not conflict with classical analyses or pretreatments. Actually, big data technology has already been successfully applied as a powerful data-driven tool for numerous phenomena, such as quantum systems [7], financial systems [8], [9], biological systems [10], as well as wireless communication networks [11]–[13]. Major tasks of the architecture for these applications seem similar: 1) big data modeling; and 2) big data analysis. It is believed that big data technology will also have a wide applied scope in power systems, and the results will be fruitful.

A. Contribution

This paper, aiming to apply big data technology into smart grids, proposes a feasible architecture with detailed procedures. Firstly, we introduce random matrix theory (RMT) as our mathematical foundations. According to RMT, a standard random matrix is systematically formed to map the system. Then, we conduct the high-dimensional analysis and compare the findings [i.e., empirical spectrum density (ESD) and

Manuscript received January 30, 2015; revised April 15, 2015; accepted June 8, 2015. This work was supported by the National Key Technology Research and Development Program of Science and Technology under Grant 2013BAA01B04. Paper no. TSG-00106-2015.

X. He, Q. Ai, W. Huang, L. Piao, and H. Liu are with the Department of Electrical Engineering, Research Center for Big Data Engineering Technology, State Energy Smart Grid Research and Development Center, Shanghai Jiaotong University, Shanghai 200240, China (e-mail: hexing_hx@126.com).

R. C. Qiu is with the Department of Electrical and Computer Engineering, Tennessee Technological University, Cookeville, TN 38505 USA (e-mail: rqiu@tntech.edu).

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

Digital Object Identifier 10.1109/TSG.2015.2445828

kernel density estimation (KDE)] with the RMT theoretical predictions [i.e., Marchenko–Pastur (M–P) law and Ring law] to tell signals from white noises. Within the above mathematical procedure, a high-dimensional statistic—mean spectral radius (MSR)—is proposed to indicate the data correlations. More than that, MSR also clarifies the parameter interchanged among the utilities under the group-work mode for distributed calculation. In addition, power grids in three different periods, involving their features, management modes, and data flows and energy flows, are summarized as general backgrounds and foundations for applying big data in power systems.

Based on the architecture, we also conduct five case studies and summarize the most interesting results.

1) The comparisons between the experimental findings and the RMT predictions, as well as the proposed indicator MSR, are sensitive to event detections. In addition, data in different dimensions are correlative under high-dimensional perspective.
2) The MSR is somehow qualitatively correlated with the quantitative parameters of system performance.
3) The architecture, besides for event detections, can also be used as a new method to find the critical active power point at any bus node, taking account of probable grid fluctuations.
4) The architecture is compatible with the block calculation only using the regional small database, and practical for real large-scale distributed systems. In addition, the high-dimensional comparative analysis is sensitive to situation awareness for grids operation, even with imperceptible different measured data.
5) The architecture is suitable not only for the power flow analysis, but also for the fault detection. To our best knowledge, this paper represents the first such attempt in the literature on power systems.

B. Related Work

It is well-established that data, as a vital resource, should be utilized much more efficiently in power systems. References [14]–[16] showed the improvement in wide-area monitoring, protection and control with utilizing phasor measurement units data. Kanao *et al.* [17] proposed a practical data utilization method based on harmonic state-estimation for power system harmonic analysis. It was a data processing method in a specific field and only available when the engineering model is accurate. Alahakoon and Yu [18] proposed advanced analytic refer to a number of techniques in many specific fields: data mining tools, knowledge discovery tools, machine learning technologies, and so on. Recently, Xu and Yong [19] initiated power disturbance data analytics to explore useful aspects of power quality monitoring data, and showed a wide applied scope in the future. The mathematical foundations and system frameworks were missing yet. For data utilization methods in power systems, although many researches, especially those methods based on specific physical models, were done in various fields, little attention has been paid to the design of a universal architecture, which is based on solid mathematical foundations and statistical procedures.

TABLE I
SOME FREQUENTLY USED NOTATIONS IN THE THEORY

Notations	*Means*
$\mathbf{X}$,$\mathbf{x}$,x, $x_{i,j}$	a matrix, a vector, a single value, an entry of a matrix
$\hat{\mathbf{X}}$,$\hat{\mathbf{x}}$,$\hat{x}$	hat: raw data
$\tilde{\mathbf{X}}$,$\tilde{\mathbf{x}}$,$\tilde{x}$,$\tilde{\mathbf{Z}}$	tilde: transformation data, formed by normalization
$\overline{\hat{\mathbf{x}}_i}$,$\overline{\tilde{\mathbf{x}}_i}$	overline: average
N, T, c	the numbers of rows and columns, $c = N/T$
$\mathbb{C}^{\mid N \times T}$	$N \times T$ dimensional complex space
$\mathbf{X}_u$	the singular value equivalent of the matrix $\tilde{\mathbf{X}}$
$\mathbf{S}$	Covariance matrix of $\mathbf{X}$: $\mathbf{S} = \frac{1}{N}\mathbf{X}\mathbf{X}^H \in \mathbb{C}^{\mid N \times N}$
$\mathbf{Z}$,L	L independent matrices product: $\mathbf{Z} = \prod_{i=1}^{L} \mathbf{X}_{u,i}$
$\lambda_{\mathbf{S}}$, $\lambda_{\mathbf{Z}}$	the eigenvalue of matrix $\mathbf{S}$, $\mathbf{Z}$
$\lambda_{\mathbf{S},i}$	the i-th eigenvalue of matrix $\mathbf{S}$
r	the circle radius on the complex plane of eigenvalues
κ_{MSR}	mean value of radius for all the eigenvalues of $\tilde{\mathbf{Z}}$: $\overline{\mathbf{r}_{\lambda_{\tilde{\mathbf{Z}}}}}$
$\mu\,(x)$,$\sigma^2(x)$	mean, variance for x

II. BIG DATA, RANDOM MATRIX THEORY, AND DATA PROCESSING PROCEDURE

The nomenclature is given as Table I.

Big data is an emerging paradigm applied to datasets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Various technologies are being discussed to support the handling of big data such as massively parallel processing databases, scalable storage systems, cloud computing platforms, and MapReduce. In the context of this paper, massively parallel processing databases is relevant to address the real-time operation within tolerable elapsed time.

A. Big Data Definition and its Features

Big data is a data-driven cognitive approach; it perceives the world through data—it works out the statistical correlations indicated by high-dimensional parameters using a nonparameter model. Currently, there exists no standardized definition for big data. This paper gives a mathematical definition below as our past work [11]–[13], [20].

1) For each sampling time, data of N-dimension are modeled as vectors, say $\mathbf{x}_i \in \mathbb{R}^{\mid N}$.
2) The number of data samples, say T, is large.
3) A function, $f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$, is able to be defined.

One of the big data fundamental characteristics is huge volume of data represented by heterogeneous and diverse dimensions. $\mathbf{X} \in \mathbb{C}^{\mid N \times T}$ is a natural model, formed by the data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, to describe a large-scale system or subsystem [21]; $\mathbf{X}$ is a nonparameter model formed almost based on a minimum hypothesis. While the size (rows N for dimensions number; columns T for samples number) increases, so do the complexity and the relationship underneath the data. Whereas, when the size are sufficiently large, some unique phenomena, such as concentration of spectral measure [4], will occur.

B. Random Matrix Theory

RMT has emerged as a particularly useful framework for many theoretical questions, especially for those concerning multivariate data. There are two frameworks

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

for RMT: 1) assumes the asymptotic convergence, and the dimensions are infinite and 2) the nonasymptotic solution, assuming finite matrix size. For the asymptotic results, in theory, we require the infinite size of the matrix, which is infeasible in practical world. However, the results are remarkably accurate, even for relatively moderate matrix sizes such as tens. This is the very reason why this random matrix model is penetrating so many areas from financial engineering to wireless network. Our initial motivation for this model was from large-scale wireless network. The new trends for RMT are: 1) finite matrix and 2) non-Gaussian matrix entries.

1) Marchenko–Pastur Law (M–P Law): M–P law describes the asymptotic behavior of singular values of large rectangular random matrices. Let $\mathbf{X} = \{x_{i,j}\}$ be a $N \times T$ random matrix whose entries, with the mean $\mu(x) = 0$ and the variance $\sigma^2(x) < \infty$, are independent identically distributed [independent identically distributed (i.i.d.)]. As $N, T \to \infty$ with the ratio $c = N/T \in (0, 1]$, the ESD of the corresponding sample covariance matrix $\mathbf{S} = (1/N)\mathbf{X}\mathbf{X}^H \in \mathbb{C}^{|N \times N}$ converges to the distribution of M–P law [20], [22] with density function

$$f_{\mathrm{ESD}}(\lambda_{\mathbf{S}}) = \begin{cases} \dfrac{1}{2\pi\lambda c\sigma^2}\sqrt{(b-\lambda)(\lambda-a)}, & a \le \lambda \le b \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

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

2) Kernel Density Estimation: A nonparametric estimate [23] of the ESD of the sample covariance matrix $\mathbf{S} \in \mathbb{C}^{|N \times N}$ is used

$$f_{\mathrm{ESD}}(\lambda_{\mathbf{S}}) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - \lambda_{\mathbf{S},i}}{h}\right) \quad (2)$$

where $\lambda_{\mathbf{S},i}$ $(i = 1, 2, \ldots, N)$ are the eigenvalues of $\mathbf{S}$, and $K(\cdot)$ is the kernel function for bandwidth parameter h.

3) Ring Law: Ring law for (large) non-Hermitian matrices is one of the most remarkable developments in the modern probability [24], [25]. Except for our previous work [4], [11]–[13], little research has been done to leverage this new tool in the context of modeling massive datasets. This mathematical structure is general enough to model many unprecedented problems.

Consider the product of L non-Hermitian random matrices $\mathbf{Z} = \prod_{i=1}^{L}\mathbf{X}_{u,i}$, where $\mathbf{X}_u \in \mathbb{C}^{|N \times N}$ is the singular value equivalent [26] of the rectangular non-Hermitian random matrix $\tilde{\mathbf{X}} \in \mathbb{C}^{|N \times T}$, whose entries are i.i.d. variables with the mean $\mu(\tilde{x}) = 0$ and the variance $\sigma^2(\tilde{x}) = 1$. This product $\mathbf{Z}$ allows us to study the streaming datasets generated as a function of both space and time. The basic target of this paper is the prospective architecture and its effectiveness. For simplification, we set L to one in this paper and need not to discuss more on L. Beyond that, the product $\mathbf{Z}$, by a transform which make the variance to $1/N$, can be converted to $\tilde{\mathbf{Z}}$ [i.e., $\sigma^2(\tilde{z}) = 1/N$]. Thus, the ESD of $\tilde{\mathbf{Z}}$ converges almost surely to the limit given by

$$f_{\mathrm{ESD}}(\lambda_{\tilde{\mathbf{Z}}}) = \begin{cases} \dfrac{1}{\pi c L}|\lambda|^{(2/L-2)}, & (1-c)^{L/2} \le |\lambda| \le 1 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

as $N, T \to \infty$ with the ratio $N/T = c \in (0, 1]$.

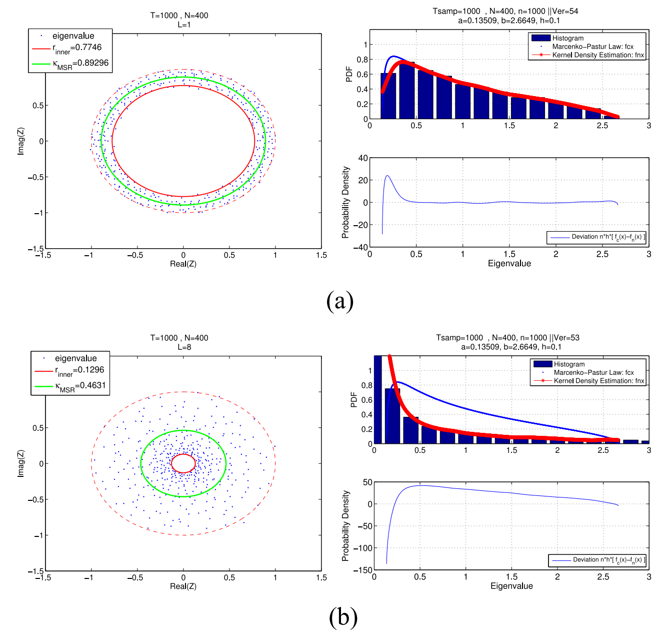


Fig. 1. Ring law and M–P law for $\mathbf{Z} = \prod_{i=1}^{L}\mathbf{X}_{u,i}$. (a) $L = 1$. (b) $L = 8$.

On the complex plane of eigenvalues, the inner circle radius is $(1-c)^{L/2}$ and the outer circle radius is unity. Especially, $\mathbf{S} = \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^H = (1/N)\mathbf{Y}\mathbf{Y}^H$ ($\mathbf{Y} = \sqrt{N}\tilde{\mathbf{Z}} \in \mathbb{C}^{|N \times N}$, $\sigma^2(\tilde{z}) = 1/N$, $\sigma^2(y) = \sigma^2(\sqrt{N}\tilde{z}) = 1$) is able to be acquired, and $\mathbf{S}$ conforms to M–P law. Ring law and M–P law with $L = 1$ and $L = 8$ are shown as Fig. 1. Furthermore, we propose κ_{MSR} (defined at the end of this section) to indicate the eigenvalues distribution of $\tilde{\mathbf{Z}}$ and the data correlation as a statistical parameter. In Fig. 1, κ_{MSR} is depicted in green line.

C. Big Data Analysis

Data processing in power systems is a typical big data challenge. For a system, lots of measured data or simulated ones $\hat{x}$ are readily accessible; for a certain time t_i, they are arranged as a vector $\hat{\mathbf{x}}_{t_i}$. As time goes by, vectors are acquired one by one and a sheet is naturally formed as dataset to map the system. Any section in the dataset, according to our intend, is available as a raw data source $\Omega\hat{\mathbf{x}}$ for further analyses. Thus, $\Omega\hat{\mathbf{x}}$ consists of sample vectors on a series of times, which are denoted as $\hat{\mathbf{x}}_{t_1}, \hat{\mathbf{x}}_{t_2}, \ldots, \hat{\mathbf{x}}_{t_i}, \ldots$; at any time t_i, the vector $\hat{\mathbf{x}}_{t_i}$ consists of sample data in various dimensions, which are denoted as $\hat{x}_{t_i,n_1}, \hat{x}_{t_i,n_2}, \ldots, \hat{x}_{t_i,n_j}, \ldots$. The upper limit of dimensions n (i.e., max j) is subject to the variety of the data at a single sampling time, and the upper limit of time length t (i.e., max i) is subject to the volume of the database and generally big enough.

For the raw data source $\Omega\hat{\mathbf{x}}$, we can focus on any data area (e.g., N rows, T columns) as a split-window; a raw matrix $\hat{\mathbf{X}} \in \mathbb{C}^{|N \times T}$ is naturally formed. Then, $\hat{\mathbf{X}}$ is converted to a normalized non-Hermitian matrix $\tilde{\mathbf{X}} \in \mathbb{C}^{|N \times T}$ row-by-row with the following algorithm:

$$\tilde{x}_{i,j} = \left(\hat{x}_{i,j} - \overline{\hat{\mathbf{x}}_i}\right) \times \left(\sigma\left(\tilde{\mathbf{x}}_i\right)/\sigma\left(\hat{\mathbf{x}}_i\right)\right) + \overline{\tilde{\mathbf{x}}_i}, \quad 1 \le i \le N;\ 1 \le j \le T \quad (4)$$

where $\hat{\mathbf{x}}_i = (\hat{x}_{i1}, \hat{x}_{i2}, \ldots, \hat{x}_{iT})$ and $\overline{\tilde{\mathbf{x}}_i} = 0$, $\sigma^2(\tilde{\mathbf{x}}_i) = 1$.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
NOTATIONS FOR ARCHITECTURE AND CASE STUDIES

Notations	*Means*
$\mathbf{A}_{\text{Time}}$	time area to focus on the data split-window
$\mathbf{A}_{\text{Node}}$	node area to focus on the data split-window
t	time: t_0 for current time, t_s for sampling time
$P_{\text{Bus-}n}$	power demand of load at bus-n
$P_{\text{max\_Bus-}n}$	critical point on P_{ower}–V_{oltage} curve for bus-n
γ_{Acc}	addition factor for load fluctuations of the grid
γ_{Mul}	multiplication factor for load fluctuations of the grid

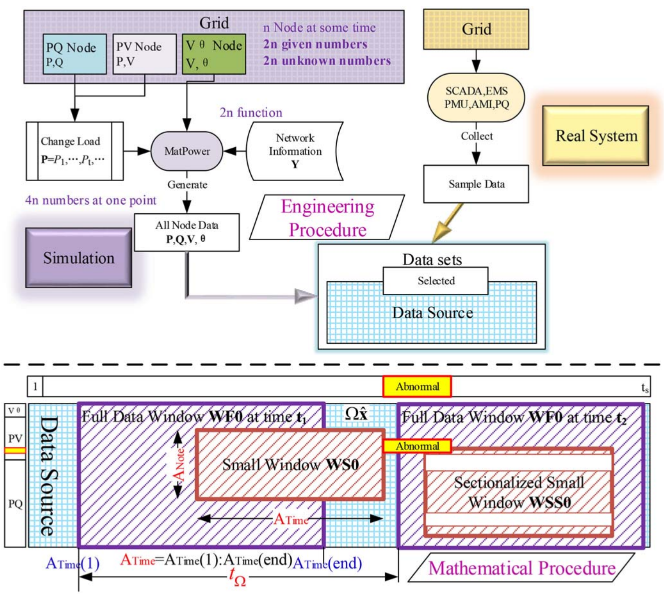


Fig. 2. Designed big data architecture for smart grids. The part above the dot line illustrates an engineering procedure for big data modeling, during which the raw data source $\Omega\hat{\mathbf{x}}$ are formed to map the physical system. The other part below the dot line illustrates a mathematical procedure for big data analysis. It is fully independent of engineering parameters, and during which the analyses are extracted from data source $\Omega\hat{\mathbf{x}}$.

The matrix $\mathbf{X}_u \in \mathbb{C}^{|N\times N}$ is introduced as the singular value equivalent [4] of $\tilde{\mathbf{X}} \in \mathbb{C}^{|N\times T}$ by

$$\mathbf{X}_u = \sqrt{\tilde{\mathbf{X}}\tilde{\mathbf{X}}^H}\mathbf{U} \tag{5}$$

where $\mathbf{U} \in \mathbb{C}^{|N\times N}$ is a Haar unitary matrix, $\mathbf{X}_u\mathbf{X}_u{}^H \equiv \tilde{\mathbf{X}}\tilde{\mathbf{X}}^H$.

For L arbitrarily assigned independent non-Hermitian matrices $\hat{\mathbf{X}}_i$ $(i = 1, \ldots, L)$ in the raw data source $\Omega\hat{\mathbf{x}}$, the matrices product $\mathbf{Z} = \prod_{i=1}^{L}\mathbf{X}_{u,i} \in \mathbb{C}^{|N\times N}$ is able to be acquired. Then, $\tilde{\mathbf{Z}}$ is calculated row-by-row with the following formula:

$$\tilde{\mathbf{z}}_i = \mathbf{z}_i / \left(\sqrt{N}\sigma(\mathbf{z}_i)\right),\ 1 \leq i \leq N \tag{6}$$

where $\mathbf{z}_i = (z_{i,1}, z_{i,2}, \ldots, z_{i,N})$, $\tilde{\mathbf{z}}_i = (\tilde{z}_{i,1}, \tilde{z}_{i,2}, \ldots, \tilde{z}_{i,N})$.

Furthermore, for the radius of all eigenvalue of $\tilde{\mathbf{Z}}$ on the complex plane, we calculate its mean value and denote it as κ_{MSR} (i.e., $\kappa_{\text{MSR}} = \overline{\mathbf{r}_{\lambda_{\tilde{\mathbf{Z}}}}}$). In general, we conduct high-dimensional analysis, only with its dataset, to reveal the properties of a power system. $\tilde{\mathbf{Z}}$ and its ESD are analyzed based on the newly developed Ring law, and the high-dimensional statistic κ_{MSR} is calculated and visualized as an indicator. In addition, the corresponding sample covariance matrix $\mathbf{S} = \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^H$ is calculated for the comparison among the results of its histogram, KDE, and M–P law.

III. BIG DATA ARCHITECTURE FOR SMART GRIDS AND ITS ADVANTAGES

A. Big Data Architecture for Smart Grids

The frequently used notations are shown in Table II.

The designed architecture is illustrated as Fig. 2.

It consists of two independent procedures to connect smart grids and big data—big data modeling as an engineering procedure, following by the big data analysis as a mathematical procedure. During the engineering procedure, the raw data source $\Omega\hat{\mathbf{x}}$ is acquired as described in Section II; during the mathematical procedure, the following steps are conducted.

Steps of Mathematical Procedure
1) Set the initial parameters
1a) Set $\mathbf{A}_{\text{Time0}}$ and $\mathbf{A}_{\text{Node0}}$ to focus on the first data window
1b) Set t_Ω and $k = 0$ to slide the moving split-window (MSW)
2) Focus on correspond window to form $\hat{\mathbf{X}}$ ($\mathbf{A}_{\text{Time}} = \mathbf{A}_{\text{Time0}} + k$)
3) Calculate $\tilde{\mathbf{X}}$, $\mathbf{X}_u$, $\mathbf{Z}$, $\tilde{\mathbf{Z}}$, $\mathbf{S}$
4) Calculate the eigenvalues $\lambda_{\tilde{\mathbf{Z}}}$,$\lambda_{\mathbf{S}}$
5) Conduct ESD analysis and compare the result according to RMT
6) Calculate κ_{MSR} with $\lambda_{\tilde{\mathbf{Z}}}$
7) Visualize the results
8) Judge as times goes by:
8a) $k < t_\Omega \Rightarrow k{+}{+}$; back to *step 2)*
8b) $k \geqslant t_\Omega \Rightarrow$ END)

Especially, in *step 2) focus on the data window*, we are able to conduct: 1) real-time analysis: focusing on the real-time data window whose last edge of the sampling time area is current time [i.e., $\mathbf{A}_{\text{Time}}(\text{end}) = t_0$]; and 2) block-calculation for decoupling interconnected system: focusing on a smaller window consisting of data only in designated dimensions, but not in all dimensions. Besides, as the split-window, in a fixed size, slides across the data source $\Omega\hat{\mathbf{x}}$ with t_Ω and k set in *1b)*, a series of κ_{MSR} is got for further research and visualization.

B. Advantages in Data Processing

1) Algorithm: This architecture analyzes data in high-dimension as illustrated by the solid purple lines in Fig. 4. It is a universal procedure with four steps as follows.

Steps of Data Management for G3
1) Form standard random matrices $\tilde{\mathbf{X}}$
2) Acquire $\tilde{\mathbf{Z}}$, $\mathbf{S}$ by variables transforming ($\tilde{\mathbf{X}} \to \mathbf{X}_u \to \mathbf{Z} \to \tilde{\mathbf{Z}} \to \mathbf{S}$)
3) Conduct high-dimensional analysis based on RMT
4) Conduct engineering interpretations

On the other hand, the procedure of traditional data processing algorithms, in most cases, relies highly on specific simplifications and assumptions to build models. Taking genetic algorithm for an example, two steps are essential to achieve the result. One is to transcode the engineering variables to gene as the input of the gene model. It is a subjective selection procedure for the specific roles in engineering system, and only a few variables can be taken into account in final model. The other one is to perform the genetic algorithm through operations of selection, crossover, and mutation. Many problems, such as improper settings of the population size, of the crossover probabilities, or of the mutation probabilities, will inevitably make the result worse.

Compared to traditional algorithms, big data analysis, driven by data, enables us to analyze the interrelation and interaction among all the elements in system, seen as correlations indicated by high-dimensional parameters. Using a pure mathematical procedure without physical models and hypotheses, big data analysis is easier in logic and faster in speed. Moreover, except step *4) conduct Engineering Interpretation*, the whole procedure is objective without introducing or accumulating the systematic errors; moreover, the accidental errors can be eliminated either with the matrix size growing, or by repletion test and parallel computing due to the independence of the algorithm.

2) Distributed Calculation for Interconnected Grids: Smart grids operation are featured with autonomous sources and decentralized controls. Due to the potential transmission cost and privacy concerns, aggregating distributed data sources to a centralized site for mining is systematically prohibitive. On the other hand, although we are able to carry out model-based mining at each distributed site, the decoupling procedure for connected sources is highly related to simplifications and assumptions. Accordingly the result are often barely satisfied and leads to biased views and decisions.

Large random matrices provide a natural and universal data-driven solution. For a specific zone-dividing interconnected system, each site is able to form a small matrix only with its own data. In this way, the integrated matrix can be divided into blocks for distributed calculation and vice versa. For the overall system, we can conduct high-dimensional analysis by integrating the regional matrices, or even by processing a few regional high-dimensional parameters. The mathematical foundation is kept invariant as RMT; the scalability, however, depends on our intention. This architecture, decoupling the systems in the form of statistical matrices or high-dimensional parameters instead of models, is practical for real large-scale interconnected grids. The distributed calculation is a deep research, and in this paper, we just provide a relative simple applied example—*case 4* in the next section *five case studies*.

C. Advantages in Management Mode

Some brief but novel introductions and analyses about the development of power grids, mainly under the perspective of data managing mode and information communication technology (ICT), are given as the related background for applying big data into smart grids. Meanwhile, new situations and challenges, and advanced management mode for future grids are further discussed from the perspective. Especially, it is group-work mode, in our opinion, that breaks through the regional limitation for energy flows and data flows. As a result, some data-driven functions, e.g., comparative analysis, and distributed calculation, are able to be carried out.

Generally, the power grids evolutions are summarized as three generations—G1–G3 [27]. Their own network structures

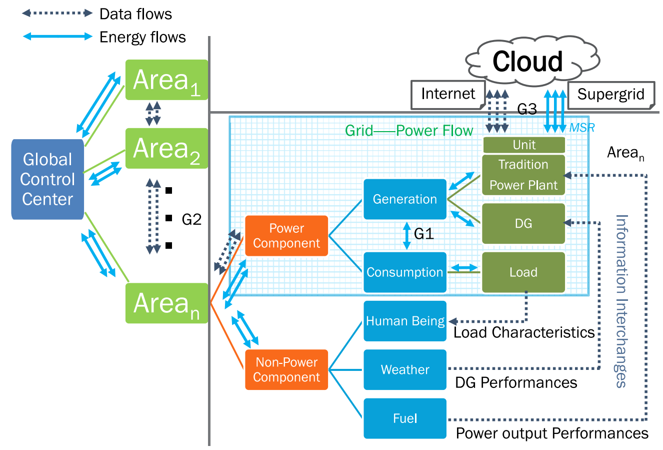


Fig. 3. Data flows and energy flows for three generations of power systems. The single lines, double lines, and triple lines indicate the flows of G1–G3, respectively.

are depicted in Fig. 12 [28]. Meanwhile, their data flows and energy flows, as well as corresponding data management systems and work modes, are quite different [29], which are shown in Figs. 3 and 4, respectively. In the following discussions, we will come to a conclusion that the group-work mode is the precondition for data-driven analysis, and the trend for smart grids.

1) G1 (Small-Scale Isolated Grids): G1 was developed from around 1900 to 1950, featured by small-scale isolated grids. For G1, components interchange energy and data within the isolated grid to keep stable. The components are fully controlled by decentralized control system and operating under individual-work mode. It means that each apparatus collects designated data, and makes corresponding decisions only with its own application, just as shown at the above part of Fig. 12(a). The individual-work mode works with an easy logic and little information communication. Whereas, it means few advanced functions and inefficient utilization for resources. It is only suitable for small grids.

2) G2 (Large-Scale Interconnected Grids): G2 was developed from about 1960 to 2000, featured by zone-dividing large-scale interconnected grids. For G2, utilities interchange energy and data within the adjacent ones. The components are dispatched by control center and operating under team-work mode. The regional team leaders, likes local dispatching centers, substations, or microgrid control centers, aggregate their own team-members (i.e., components in the region) into a standard black-box model. These standard models will be further aggregated by the global control center for the control or prediction purposes. The two aggregations above are achieved by four steps, which are data monitoring, data preprocessing, data storage, and data processing, respectively. However, lots of engineering technologies or sciences are essential as the foundations—cognitive radio wireless network [30]–[32], specific communication service mapping as ICT [33], cloud storage, parallel computing as computer science, and modelling building, parameter identification as mathematical modeling [34], [35]. The description above are illustrated by dotted blue lines in Fig. 4. In general, the team-work mode

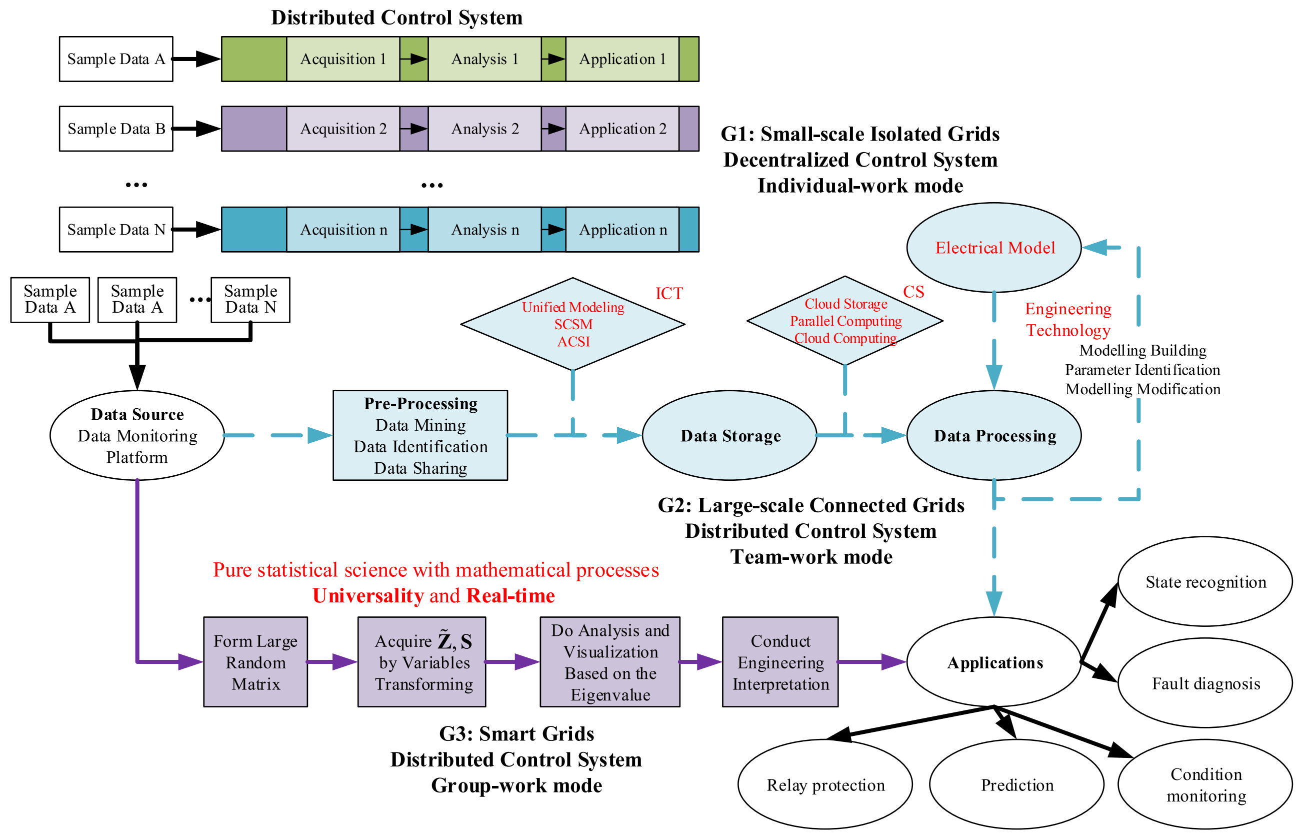


Fig. 4. Data management systems and work modes for three generations of power systems. The above, middle, and below parts indicate the data management systems and the work modes of G1–G3, respectively. For G1, each grid works independently. For G2, global and local control centers are operating under the team-work mode. For G3, the group-work mode breaks through the regional limitation for energy and data flows and has a better performance.

conducts model-based analysis, and mainly concerns with the system stability rather than the individual benefit; it will not work well for smart grids with 4Vs data as described in Section I.

3) G3 (Smart Grids): The development of G3 was launched at the beginning of the 21st century; and for China, it is expected to be completed around 2050 [27]. Fig. 12(c) shows that the clear-cut partitioning is no longer suitable for G3, as well as the team-work mode which is based on the regional leader. For G3, the control force of the regional center (if still exist) is greatly released by individual units. The high-performance and self-control individuals results in much more flexible flows, for both energy exchange and data communication, to improve utilization by sharing resources among the whole grid [36]. Accordingly, the group-work mode is proposed. Under this mode, the individuals play a dominant part in the system under the authority of the global control centers [29]. Virtual power plants [37] and multi-microgrids [38], for instance, are typically G3 utilities. These group-work mode utilities provide a relaxed environment to benefit both the individuals and the grids: the former (i.e., individuals), driven by their own interests and characteristics, are able to create or join a relatively free group to benefit mutually from sharing their respective superior resources; meanwhile, these utilities are generally big and controllable enough to be good customers or managers to the latter (i.e., the smart grids).

IV. Five Case Studies

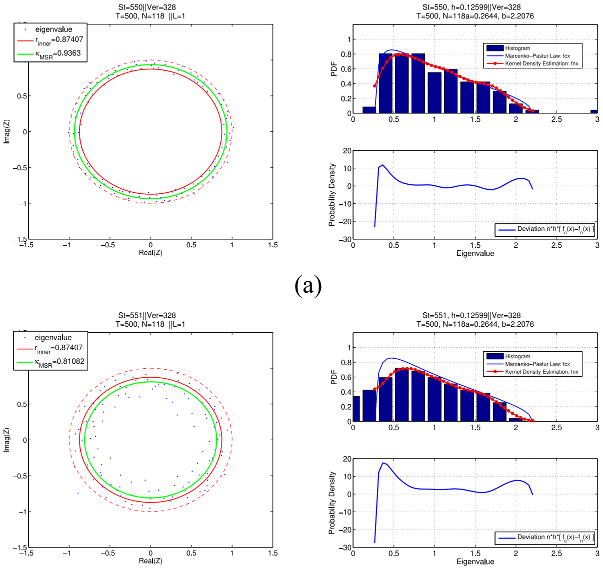
For the following designed experiments, all the data are obtained in two scenarios: 1) only white noises (i.e., benchmark, whose statistical results agree with M–P law and Ring law); and 2) signals plus noises. We treat small random loads fluctuations and sample errors as white noises, and sudden changes and faults as signals.

Cases 1–4, based on Matpower, belongs to the field of power system stability and control. The grid is a standard IEEE 118-bus system with six partitions displayed as Fig. 14 [39]. Detailed information about the test bed is referred to the *case118.m* in *Matpower package* and *Matpower 4.1 users manual* [40]. Case 5, based on power systems computer aided design/electromagnetic transients for direct current, is about fault detection.

A. Case 1: Observation From the Split-Window With Full Network and 500 Sample Points—$N = 118$, $T = 500$

Case 1 is a simple study to validate that the architecture is able to quickly detect the signals from the noises with the real-time data flow. Let $N = 118$, $T = 500$, and $c = N/T = 0.236$. Thus each split-window has $n = NT = 59\,000$ data. Table III shows the series of assumed events, and the accordingly RMT analysis results visualization at critical time and MSRs on the time series are depicted as Figs. 5 and 6, respectively.

1) Sampling Time $t_s = 550$ s, Time Area $\mathbf{A}_{Time} = 51{:}550$ s: There are only small load fluctuations in this split-window,



(a)

(b)

Fig. 5. Ring law, and the comparation among histogram, KDE, and M–P law at critical time for case 1. (a) $t_s = 550$ s. (b) $t_s = 551$ s.

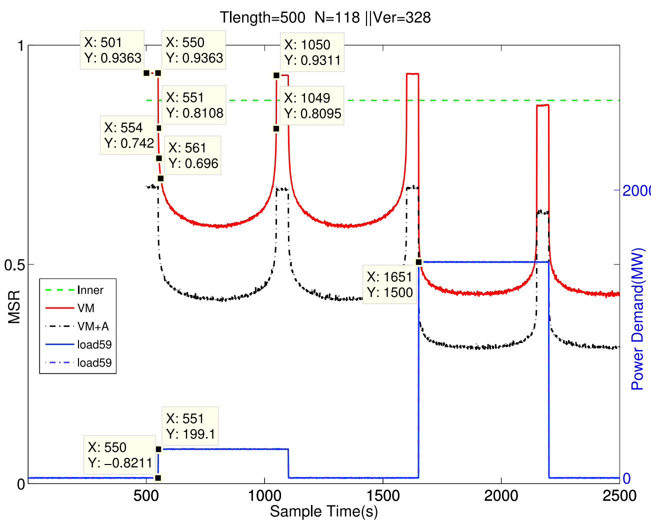


Fig. 6. MSRs on the time series for case 1.

which means that white-noises play a dominant part. Then, we compare the histogram, KDE and M–P law. Inspection of Fig. 5(a) indicates that, in a white-noises dominated system, the KDE (in red line) matches the histogram (in blue bar) very well. Moreover, the histogram curve and the KDE curve agree with M–P law (in blue line).

2) Sampling Time $t_s = 551$ s, Time Area $\mathbf{A}_{Time} = 51{:}550$ s: For this split-window, Fig. 5(b) shows that the eigenvalues of Ring law collapse to the circle center, and both histogram and KDE deviate from M–P law. It means that there are some signals, any deviation of the benchmark (white noises only), in the system. Indeed, just at time $t = 551$ s, the $P_{\text{Bus-59}}$ suddenly changes somehow from 0 to 200 MW.

Fig. 6 depicts that the κ_{MSR} change dramatically in a short time since $t = 550$ s. As the length of time area $T = 500$, the step change as the signal occurring at time $t = 551$ s, is included during all the sampling times from

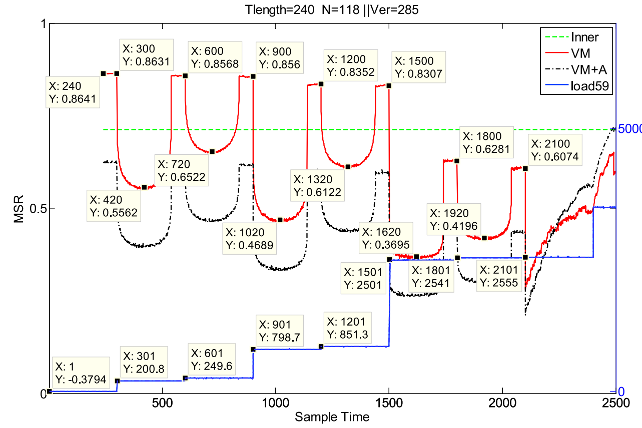


Fig. 7. MSRs on the time series for case 2.

$t_s = 551$ s to $t_s = 1049$ s. It results in the deviation $(0.9363, 0.8108, 0.742, \ldots)$. However, at the sampling time $t_s = 1050$ s, when the time area $\mathbf{A}_{\text{Time}} = 551 : 1050$ s, the step signal is no longer exist, as well as the deviation of the histogram and KDE from M–P law, and κ_{MSR} is back to 0.9311.

In addition, it is found that κ_{MSR} for the data of V (red line) which has definite physical meaning, and of $V + i\theta$ (black line) without any physical meaning, have the same trend. It indicates that MSR is a high-dimensional statistic, which is independent and robust to physical model in some way. The green line indicates the inner radius of the ring for the analyzing matrix, whose value only depends on the matrix size as formula (3) in Section II.

This case indicates that the presented statistic MSR is sensitive to signal. Meanwhile, there are some inherent relations for the data in different dimensions. It means that real-time analysis for event detection can be carried out with less kinds of data.

B. Case 2: Observation From the Smaller Split-Window With Full Network and 240 Sample Points—N = 118, T = 240

Case 2 is similar to the previous one except that T decreases to 240 s, and the events are rearranged. In this case, we try to find the qualitative relationships between the MSR and the quantitative parameters of system performances. The series of assumed events and the κ_{MSR}–t curve are depicted as Table IV and Fig. 7, respectively.

1) During once step-change, there is a negative correlation between the min value of MSR (i.e., $\kappa_{\text{min\_MSR}}$) and the step-change value of $P_{\text{Bus-59}}$ (i.e., $\Delta\ P_{\text{Bus-59}}$).

$\Delta P_{\text{Bus-59}}$	200	50	550	1650	⋯
$P_{\text{Bus-59}}$	$0 \to 200$	$200 \to 250$	$250 \to 800$	$850 \to 2500$	⋯
$\kappa_{\text{min\_MSR}}$	0.5562	0.6522	0.4689	0.3695	⋯

2) When $P_{\text{Bus-59}}$ is steady at a higher level, κ_{MSR} is steady at a lower level.

$P_{\text{Bus-59}}$	0	200	250	800	⋯	2540
κ_{MSR}	0.8631	0.8568	0.8560	0.8352	⋯	0.6074

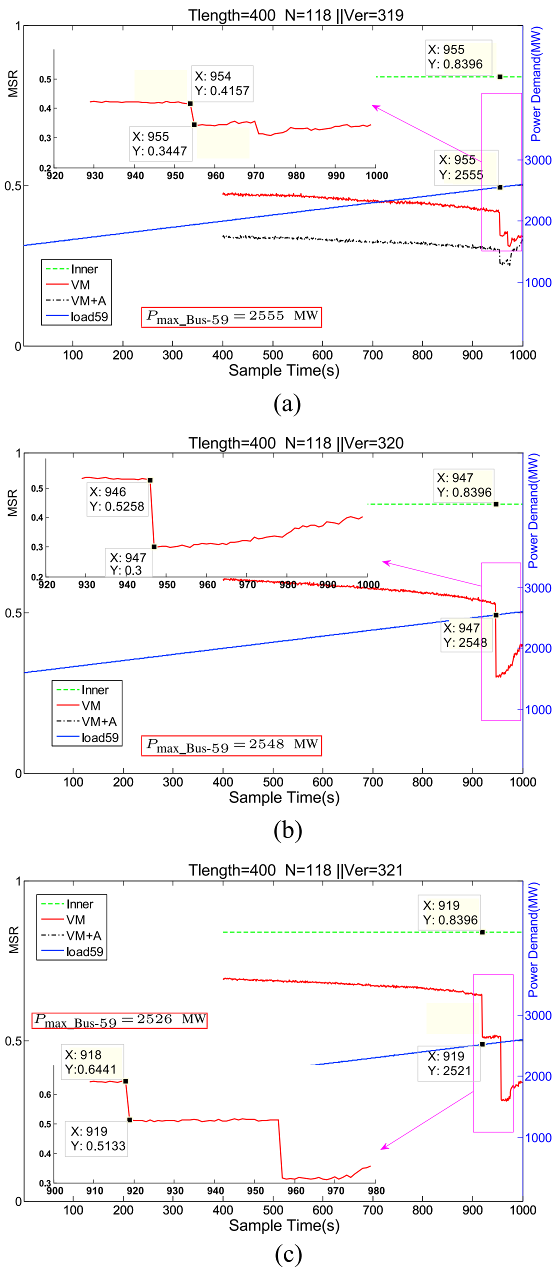


Fig. 8. Critical power point estimation taking account of grid fluctuations. (a) $\gamma_{\text{Acc}} = 0$, $\gamma_{\text{Mul}} = 0$. (b) $\gamma_{\text{Acc}} = 1$, $\gamma_{\text{Mul}} = 0.02$. (c) $\gamma_{\text{Acc}} = 5$, $\gamma_{\text{Mul}} = 0.1$.

3) When the $P_{\text{Bus-59}}$ approaches to the critical active power point $P_{\max\_\text{Bus-59}}$, a little step change of $P_{\text{Bus-59}}$ will lead to a small value of $\kappa_{\min\_\text{MSR}}$. The feature b) and c) is available to conduct vulnerable node identification [41] as a new method. And when $P_{\text{Bus-59}}$ is beyond $P_{\max\_\text{Bus-59}}$ (i.e., $P_{\text{Bus-59}} > 2555$ MW), κ_{MSR} is no longer steady.

$\Delta P_{\text{Bus-59}}$	50	50	40	15
$P_{\text{Bus-59}}$	200→250	800→850	2500→2540	2540→2555
$\kappa_{\min\_\text{MSR}}$	0.6522	0.6122	0.4196	*unsteady*
$\Delta\kappa_{\text{MSR}}$	0.2046	0.2230	0.2085	*unsteady*

C. Case 3: Critical Power Point Estimation

This case, based on the feature c) of the previous one, is designed as a new method to find the critical point $P_{\max\_\text{Bus-}n}$,

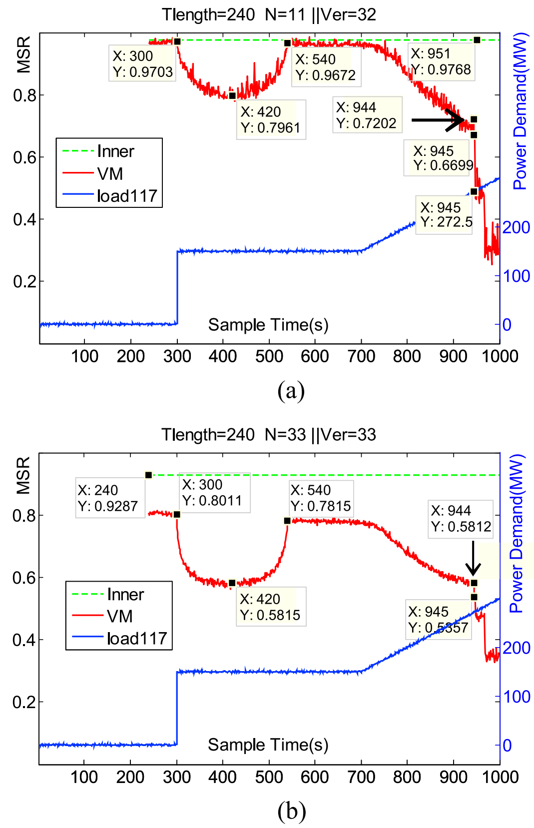


Fig. 9. MSR series of single region. (a) $\mathbf{A}_{\text{Node}} = \text{A1}$. (b) $\mathbf{A}_{\text{Node}} = \text{A2}$.

especially that the grid fluctuations is taking into consideration. The grid fluctuations are set by γ_{Acc} and γ_{Mul}

$$\tilde{y}_{\text{load}\_nt} = y_{\text{load}\_nt} \times (1 + \gamma_{\text{Mul}} \times x_1) + \gamma_{\text{Acc}} \times x_2 \tag{7}$$

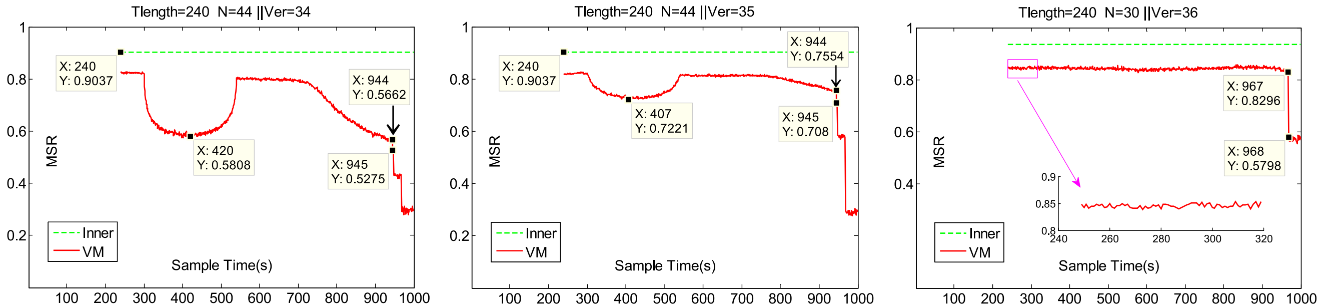
where x_1 and x_2 are random numbers from a standard Gaussian Distribution. The result are shown in Fig. 8.

In this model, the increase of grid fluctuations means the decrease of signal-noise ratio, which will cause a raise of $\kappa_{\min\_\text{MSR}}$. Meanwhile, it will also cause the decrease of $P_{\max\_\text{Bus-}n}$ for a certain node (2555, 2548, and 2521 MW), which meets our common knowledge and experience.

D. Case 4: Group-Work Mode

For the above cases, the anomaly can be detected by traditional model-based tools. These cases validate that the designed data-driven architecture is also a solution to anomaly detection. In case 4, however, we will discuss a scenario in which the traditional tools fail to detect the anomalies.

This case is based on a zone-dividing system with six regions (A1 to A6) depicted in Fig. 14. A PQ node (node with constant active power and reactive power) far from slack bus, i.e., bus-117 in A1, is chosen as signal source. It is much more vulnerable than PQ nodes (nodes with constant active power and voltage magnitude) such as bus-59. With the same procedures of the former case studies, $P_{\max\_\text{Bus-117}}$ and the κ_{MSR}–t curve for the overall system are depicted by Table V, Figs. 15 and 16, respectively. Under group-work mode, each regional center, just like the global one, calculates MSR with its own data, such as Fig. 9(a) and (b) for A1 and A2, respectively.



(a) (b) (c)

Fig. 10. MSR series of union regions. (a) $\mathbf{A}_{\text{Node}} = [\text{A1, A2}]$. (b) $\mathbf{A}_{\text{Node}} = [\text{A3, A5}]$. (c) $\mathbf{A}_{\text{Node}} = [\text{A4, A6}]$.

When the data split-window is not big, just as A1 which only has 11 nodes (i.e., $N = 11$), the signal is still able to detected, but the curve is not smooth. Thus, some regions are combined to smooth the κ_{MSR}–t curve—A1 and A2, A3 and A5, and A4 and A6 in Fig. 10(a)–(c), respectively. In the Appendixes, the raw data of $\hat{\mathbf{V}}$ and their low-dimensional visualization for all PQ buses in A3 and A5 around time $t_s = 301$ s, when the step-change of $P_{\text{Bus-117}}$ happened, are shown as Fig. 17(a) and (b).

Fig. 16 shows that $P_{\text{max\_Bus-117}} = 272.5$ MW at $t_s = 945$ s. This point is observed by all the regional centers as shown in Figs. 9 and 10. Meanwhile, the signal occurring just at time $t = 301$ s is also able to be detected in the system. According to κ_{MSR} of Fig. 10(a)–(c), it is found that to response to the signal, the distribution of $\Delta\kappa_{\text{MSR}}$ in distributed regions is just like the contour line and the A1 and A2 is the mountaintop. As a result, we conjecture that the signal is generated at A1 and A2 rather than A3 and A5 or A4 and A6. The events set in Table V validates the conjecture.

Interchanging MSRs among distributed utilities, some useful analyses are able to be extracted. It is sensitive to anomaly detection, even with imperceptible different raw data just as shown in Fig. 17—the raw voltage amplitude dataset $\Omega\hat{\mathbf{V}}$ of A3 and A5 changes too little to be utilized in low-dimension (e.g., Mean—1-D statistic and Variance—2-D statistic). This case study also validates the architecture is compatible with block calculation only using regional small database; in this way, the architecture decouples the interconnect systems from statistical parameters perspective naturally, and is practical for real large-scale distributed systems.

E. Case 5: Fault Detection for Active Distribution Network

This case shows the application of the designed architecture in another field of power systems. Due to the integration and variation of renewable generators and energy storage units, faults and disturbances detection has become increasingly complicated in active distribution networks [42]. Table VI sets the events series, and Fig. 13 illustrates the fault mode.

Fig. 11 indicates that some signals are detected by the κ_{MSR}–t curve. Especially, it is conjectured that the most influent events are happening around $t = 3000$ ms and $t = 13\,000$ ms, which means that the three-phase and line-to-line short circuit have more influence than the single-phase ones. This case shows that the architecture is also compatible with other fields in power systems.

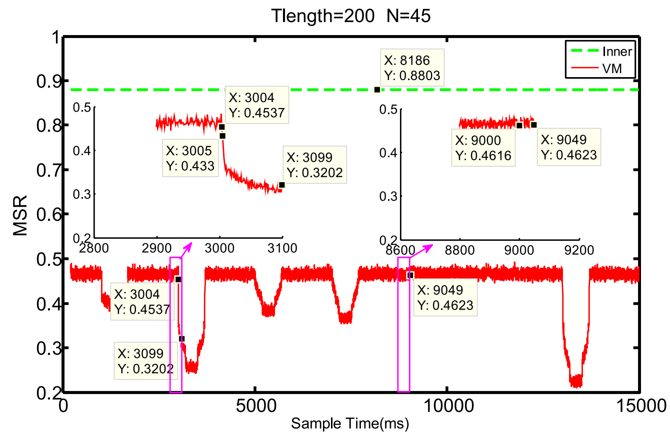
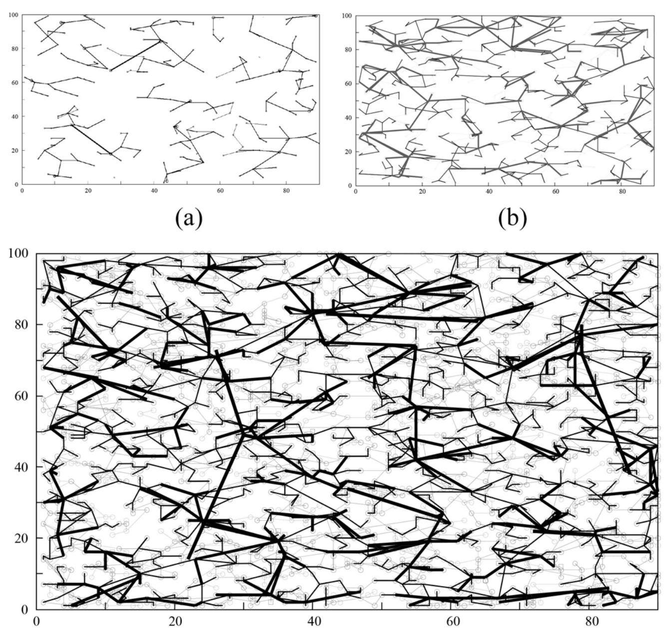


Fig. 11. MSR series for failure events.



(a) (b) (c)

Fig. 12. Simulation of network structures. The above two are G1—small-scale isolated grids, and G2—large-scale interconnected power grids; and the below one is G3—smart grids, which have complex network structures without clear-cut partitioning. (a) G1. (b) G2. (c) G3.

TABLE III
SERIES OF EVENTS FOR CASE 1

t	[001:550]	[551:1100]	[1101:1650]
$P_{\text{Bus-59}}$	0	200	0
t	[1651:2200]	[2201:2500]	
$P_{\text{Bus-59}}$	1500	0	

*t: (s), $P_{\text{Bus-59}}$: (MW)

TABLE IV
SERIES OF EVENTS FOR CASE 2

t	[001:300]	[301:600]	[601:900]
$P_{\text{Bus-59}}$	0	200	250
t	[901:1200]	[1201:1500]	[1501:1800]
$P_{\text{Bus-59}}$	800	850	2000
t	[1801:2100]	[2101:2400]	[2401:2500]
$P_{\text{Bus-59}}$	2540	2555	3500

*t: (s), $P_{\text{Bus-59}}$: (MW)

TABLE V
SERIES OF EVENTS FOR CASE 4

t	[001:300]	[301:700]	[701:1000]
$P_{\text{Bus-117}}$	0	150	$t/2-200$

*t: (s), $P_{\text{Bus-117}}$: (MW)

TABLE VI
SERIES OF EVENTS FOR CASE 5

t	[1000:1500]	[3000:3500]	[5000:5500]
EVENT	$\text{A} \rightarrow \text{G}$	$\text{AB} \rightarrow \text{G}$	$\text{B} \rightarrow \text{G}$
t	[7000:7500]	[9000:11000]	[13000:13500]
EVENT	$\text{C} \rightarrow \text{G}$	BCS OPEN	$\text{ABC} \rightarrow \text{G}$

*t: (ms)

V. CONCLUSION

This paper aims to apply big data technology into smart grids. It proposes an architecture with two independent procedures, as a data-driven solution, to conduct anomaly detections. In addition, moving split-window technology was introduced for real-time analysis, and a new statistic MSR was proposed to indicate the data correlations, as well as to clarify the parameter interchanged among the utilities.

The algorithm of this architecture is based on RMT; it is a fixed objective procedure, which is easy in logic and fast in speed. The nonasymptotic framework of RMT enables us to conduct high-dimensional analysis for real systems, even with relatively moderate datasets. It also provides us a natural way to decouple the interconnected systems from data perspective. The group-work model of utilities in the systems, meanwhile, makes some data-driven functions possible, such as distributed calculation and comparative analysis.

Big data technology does not conflict with classical analysis or pretreatment. Instead, combination between block calculation and traditional zone-dividing structure realizes comparative analysis—it is sensitive way to detect the event and locate the source in the grid network, even with imperceptible different measured/simulated data. Besides, the sparse representation of a random vector is carried out with the random matrix theory [43], [44]. An incomplete matrix data representation of data may be considered: 1) random band matrices [45]; and 2) low rank matrix recovery [11]. All these

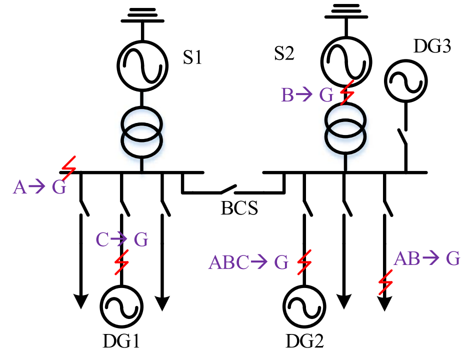


Fig. 13. Fault model for case 5.

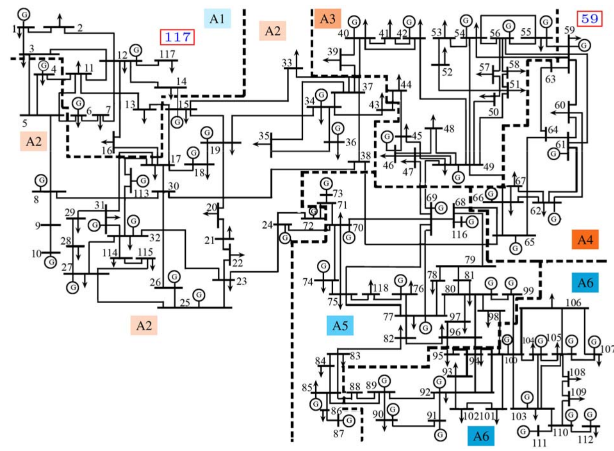


Fig. 14. Partitioning network for IEEE 118-bus system. There are six partitions, i.e., A1–A6.

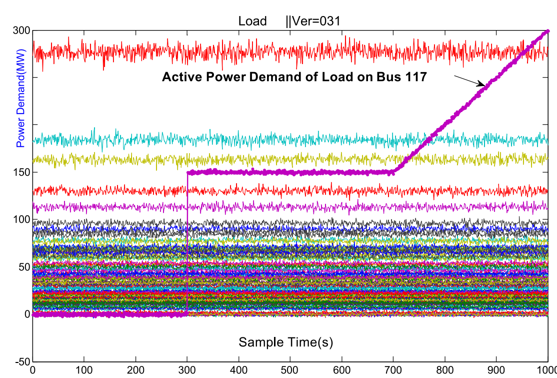


Fig. 15. Grid load change for case 4: $\gamma_{\text{Acc}} = 1$, $\gamma_{\text{Mul}} = 0.02$.

topics are considered through the unified framework of random matrix theory [11].

It is a long way and big topic to apply big data in power systems, especially for a specific field in real systems. Some designed methods are rough in this paper, e.g., *to estimate critical power point* in *case 3*, *to detect fault in power systems* in *case 5*, as well as some descriptions, e.g., *group-work mode* mentioned in *Section III*, and *new method for vulnerable node identification* mentioned in *case 2*. But these special methods/applications all have one thing in common: they are all based on the proposed architecture, and driven only by data of voltage amplitude, which are the most basic in the grids. With work ongoing, including:

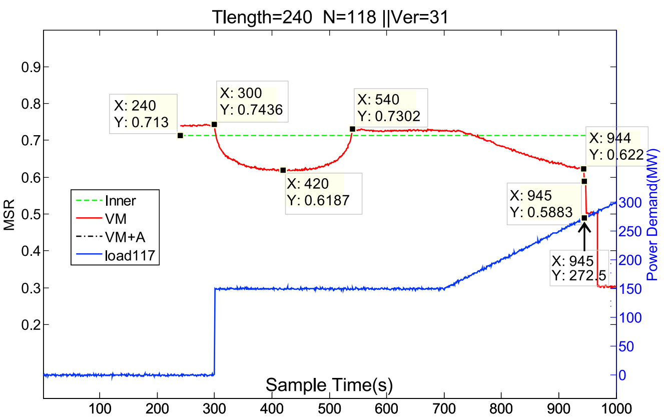
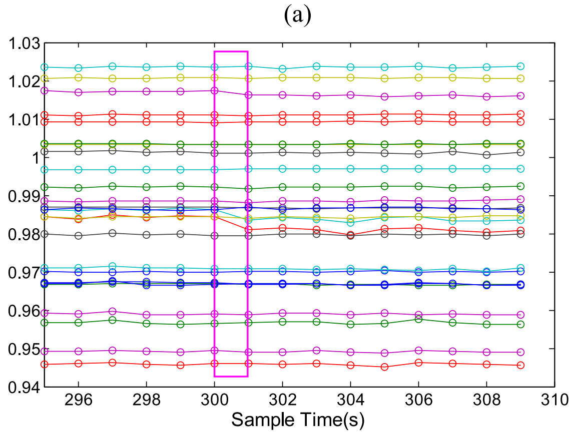


Fig. 16. Global MSR series ($P_{\mathrm{max\_Bus\text{-}117}} = 272.5$ MW).

Bus t_{ime}	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309
39	0.970005	0.969983	0.969848	0.970096	0.969903	0.969995	0.970145	0.970012	0.969999	0.970005	0.970245	0.969933	0.970099	0.96996	0.970206
41	0.966754	0.966685	0.966918	0.967009	0.966976	0.966935	0.966676	0.966833	0.966471	0.96676	0.9666	0.966547	0.966563	0.966763	0.966713
44	0.984399	0.98382	0.984807	0.98424	0.984683	0.984323	0.980976	0.981477	0.980984	0.979607	0.981137	0.981585	0.980765	0.980326	0.980837
45	0.986159	0.985919	0.986475	0.986035	0.9867	0.986232	0.983618	0.984244	0.983821	0.982742	0.984142	0.984361	0.983372	0.983277	0.983466
47	1.017287	1.017029	1.0172	1.017073	1.017246	1.017314	1.016328	1.016143	1.016083	1.01625	1.015837	1.016058	1.016139	1.01585	1.016079
48	1.020617	1.020854	1.020498	1.020576	1.020711	1.020752	1.020535	1.020788	1.020864	1.020696	1.020589	1.020722	1.020721	1.020486	1.020468
50	1.001555	1.001473	1.001716	1.001359	1.00154	1.001021	1.001008	1.001145	1.000924	1.001311	1.001024	1.000831	1.001528	1.000546	1.001234
51	0.966875	0.967021	0.967666	0.966543	0.966493	0.966801	0.966852	0.966935	0.967005	0.96647	0.966767	0.967124	0.966838	0.966583	0.966414
52	0.956708	0.956717	0.957301	0.956364	0.956208	0.956561	0.956812	0.956951	0.956984	0.956343	0.956366	0.957502	0.956705	0.956349	0.956143
53	0.945825	0.94593	0.946299	0.945711	0.945679	0.945927	0.946001	0.945844	0.945985	0.945577	0.945086	0.946315	0.946051	0.945759	0.945598
57	0.9711	0.971013	0.971416	0.970994	0.971002	0.970781	0.970845	0.970836	0.970461	0.970899	0.970459	0.970249	0.970715	0.97007	0.971018
58	0.959209	0.95909	0.959548	0.958857	0.958812	0.959062	0.95881	0.959101	0.959202	0.958819	0.959251	0.958807	0.959037	0.95876	0.958663
68	1.00324	1.00323	1.003206	1.003232	1.003228	1.003243	1.00331	1.003313	1.00331	1.003309	1.003314	1.003313	1.00331	1.003312	1.003313
71	0.986899	0.986883	0.986889	0.986923	0.986839	0.986878	0.98675	0.986715	0.986537	0.986658	0.986604	0.986713	0.986726	0.986554	0.986701
75	0.967256	0.967232	0.967471	0.967343	0.967142	0.967213	0.966722	0.966731	0.96682	0.966516	0.966431	0.966914	0.966975	0.966566	0.966639
78	1.003414	1.003523	1.003464	1.003526	1.003334	1.003337	1.003373	1.003359	1.003399	1.003556	1.003355	1.003497	1.003392	1.00341	1.003502
79	1.009228	1.009288	1.009245	1.009251	1.00911	1.009066	1.009132	1.009142	1.009278	1.009384	1.00912	1.009328	1.009163	1.0092	1.009263
81	0.99679	0.996753	0.996715	0.996781	0.996739	0.996784	0.996959	0.996914	0.996917	0.996911	0.996939	0.996942	0.996941	0.996898	0.996945
82	0.988579	0.988269	0.988548	0.988544	0.988559	0.98849	0.988129	0.988476	0.98843	0.98825	0.988646	0.9886	0.988418	0.988725	0.988934
83	0.984489	0.984081	0.984519	0.984369	0.984482	0.984381	0.983894	0.984484	0.984253	0.984027	0.984471	0.984465	0.98424	0.98456	0.984689
84	0.979856	0.979474	0.980017	0.979697	0.979877	0.979412	0.979491	0.979964	0.979838	0.979406	0.97987	0.979662	0.979954	0.979511	0.979906
86	0.986346	0.986656	0.986421	0.986177	0.986133	0.986261	0.986869	0.986215	0.986793	0.986812	0.986981	0.986868	0.986432	0.986516	0.986139
96	0.992125	0.992003	0.992298	0.992142	0.992337	0.992212	0.991713	0.99215	0.99216	0.99215	0.992362	0.992366	0.991981	0.992263	0.992469
97	1.011039	1.010851	1.011261	1.010953	1.011123	1.011065	1.010781	1.011048	1.01108	1.010941	1.011274	1.011171	1.01102	1.011111	1.011266
98	1.023595	1.023298	1.023731	1.023517	1.023812	1.02356	1.023673	1.023153	1.023715	1.023551	1.02344	1.023858	1.023239	1.0235	1.023679
118	0.949278	0.949199	0.949534	0.949275	0.949086	0.949337	0.948943	0.948904	0.949399	0.948898	0.948805	0.949365	0.949212	0.949062	0.948935

(a)



(b)

Fig. 17. Raw data $\hat{\mathbf{V}}$ and the visualization around $t_s = 300$ s. (a) Raw data $\hat{\mathbf{V}}$ (orange for A3, and blue for A5). (b) Visualization of the above $\hat{\mathbf{V}}$ (in low-dimension).

1) designing 3-D power-map by combining high-dimensional analysis and visualization [46]; and 2) conducting event/fault detection for real systems [47], [48], we can say with confidence that the designed architecture is practical for real world; and the keys are the RMT with nonasymptotic framework, high-dimensional algorithm, block calculation, and augmented matrix. This paper, as an initial one, just presents the universal architecture; for special fields, it aims to raise many open questions than actually answers ones. One wonders if this new direction will be far-reaching in years to come toward the age of big data.

Appendix A
Grid Network Structures of G1, G2, and G3

See Fig. 12.

Appendix B
Event Series for Five Case Studies

See Tables III–VI and Fig. 13.

Appendix C

See Figs. 14–17.

References

[1] Nature. (Sep. 2008). *Big Data (Specials)*. [Online]. Available: http://www.nature.com/news/specials/bigdata/index.html

[2] Science. (Feb. 2011). *Special Online Collection: Dealing With Data*. [Online]. Available: http://www.sciencemag.org/site/special/data/

[3] IBM. (Jul. 1, 2015). *The Four V's of Big Data*. [Online]. Available: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[4] R. Qiu and P. Antonik, *Smart Grid and Big Data*. Hoboken, NJ, USA: Wiley, 2014.

[5] *Managing Big Data for Smart Grids and Smart Meters*, IBM, Armonk, NY, USA, May 2009.

[6] M. Kezunovic, L. Xie, and S. Grijalva, "The role of big data in improving power system operation and protection," in *Proc. Symp. Bulk Power Syst. Dyn. Control-IX Optim. Security Control Emerg. Power Grid (IREP)*, Rethymno, Greece, 2013, pp. 1–9.

[7] T. A. Brody *et al.*, "Random-matrix physics: Spectrum and strength fluctuations," *Rev. Modern Phys.*, vol. 53, no. 3, pp. 396–480, 1981.

[8] L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud, "Random matrix theory and financial correlations," *Int. J. Theor. Appl. Finance*, vol. 3, no. 3, pp. 391–397, 2000.

[9] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Quart.*, vol. 36, no. 4, pp. 1165–1188, 2012.

[10] D. Howe *et al.*, "Big data: The future of biocuration," *Nature*, vol. 455, no. 7209, pp. 47–50, 2008.

[11] R. Qiu and M. Wicks, *Cognitive Networked Sensing and Big Data*. New York, NY, USA: Springer, 2013.

[12] C. Zhang and R. C. Qiu, "Data modeling with large random matrices in a cognitive radio network testbed: Initial experimental demonstrations with 70 nodes," *ArXiv e-prints*, Apr. 2014. [Online]. Available: http://arxiv.org/pdf/1404.3788.pdf

[13] X. Li, F. Lin, and R. C. Qiu, "Modeling massive amount of experimental data with large random matrices in a real-time UWB-MIMO system," *ArXiv e-prints*, Apr. 2014. [Online]. Available: http://arxiv.org/pdf/1404.4078.pdf

[14] A. Phadke and R. M. de Moraes, "The wide world of wide-area measurement," *IEEE Power Energy Mag.*, vol. 6, no. 5, pp. 52–65, Sep./Oct. 2008.

[15] V. Terzija *et al.*, "Wide-area monitoring, protection, and control of future electric power networks," *Proc. IEEE*, vol. 99, no. 1, pp. 80–93, Jan. 2011.

[16] L. Xie, Y. Chen, and H. Liao, "Distributed online monitoring of quasi-static voltage collapse in multi-area power systems," *IEEE Trans. Power Syst.*, vol. 27, no. 4, pp. 2271–2279, Nov. 2012.

[17] N. Kanao *et al.*, "Power system harmonic analysis using state-estimation method for Japanese field data," *IEEE Trans. Power Del.*, vol. 20, no. 2, pp. 970–977, Apr. 2005.

[18] D. Alahakoon and X. Yu, "Advanced analytics for harnessing the power of smart meter big data," in *Proc. IEEE Int. Workshop Intell. Energy Syst. (IWIES)*, Vienna, Austria, 2013, pp. 40–45.

[19] W. Xu and J. Yong, "Power disturbance data analytics–New application of power quality monitoring data," *Proc. CSEE*, vol. 33, no. 19, pp. 93–101, 2013.

[20] R. C. Qiu, Z. Hu, H. Li, and M. C. Wicks, *Cognitive Radio Communication and Networking: Principles and Practice*. Hoboken, NJ, USA: Wiley, 2012.

[21] R. C. Qiu, "The foundation of big data: Experiments, formulation, and applications," *ArXiv e-prints*, Dec. 2014. [Online]. Available: http://arxiv.org/pdf/1412.6570.pdf

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

[23] G. Pan, Q. Shao, and W. Zhou, "Universality of sample covariance matrices: Clt of the smoothed empirical spectral distribution," *ArXiv e-prints*, Nov. 2011. [Online]. Available: http://arxiv.org/pdf/1111.5420.pdf

[24] A. Guionnet, M. Krishnapur, and O. Zeitouni, "The single ring theorem," *ArXiv e-prints*, Sep. 2009. [Online]. Available: http://arxiv.org/pdf/0909.2214.pdf

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

[26] J. R. Ipsen and M. Kieburg, "Weak commutation relations and eigenvalue statistics for products of rectangular random matrices," *Phys. Rev. E*, vol. 89, no. 3, 2014, Art. ID 032106.

[27] X. Zhou, S. Chen, and Z. Lu, "Review and prospect for power system development and related technologies: A concept of three-generation power systems," *Proc. CSEE*, vol. 33, no. 22, pp. 1–11, 2013.

[28] S. Mei, Y. Gong, and F. Liu, "The evolution model of three generation power systems and characteristic analysis," *Proc. CSEE*, vol. 34, no. 7, pp. 1003–1012, 2014.

[29] X. He, Q. Ai, Z. Yu, Y. Xu, and J. Zhang, "Power system evolution and aggregation theory under the view of power ecosystem," *Power Syst. Protect. Control*, vol. 42, no. 22, pp. 100–107, 2014.

[30] R. C. Qiu *et al.*, "Cognitive radio network for the smart grid: Experimental system architecture, control algorithms, security, and microgrid testbed," *IEEE Trans. Smart Grid*, vol. 2, no. 4, pp. 724–740, Dec. 2011.

[31] H. Li *et al.*, "Efficient and secure wireless communications for advanced metering infrastructure in smart grids," *IEEE Trans. Smart Grid*, vol. 3, no. 3, pp. 1540–1551, Sep. 2012.

[32] H. Li, L. Lai, and R. C. Qiu, "Scheduling of wireless metering for power market pricing in smart grid," *IEEE Trans. Smart Grid*, vol. 3, no. 4, pp. 1611–1620, Dec. 2012.

[33] D. Baigent, M. Adamiak, and R. Mackiewicz, "IEC 61850 communication networks and systems in substations: An overview for users," *Protect. Control J.*, vol. 8, pp. 61–68, Jul. 2009. [Online]. Available: http://www.gedigitalenergy.com/multilin/journals/issues/PCJ_Spring2009.pdf

[34] R. Yuan, Q. Ai, and X. He, "Research on dynamic load modelling based on power quality monitoring system," *IET Gener. Transm. Distrib.*, vol. 7, no. 1, pp. 46–51, Jan. 2013.

[35] Q. Ai, X. Wang, and X. He, "The impact of large-scale distributed generation on power grid and microgrids," *Renew. Energy*, vol. 62, pp. 417–423, Feb. 2014.

[36] Z. Hong, D. Zhao, C. Gu, F. Li, and B. Wang, "Economic optimization of smart distribution networks considering real-time pricing," *J. Modern Power Syst. Clean Energy*, vol. 2, no. 4, pp. 350–356, 2014.

[37] Y. Ji, "Multi-agent system based control of virtual power plant and its application in smart grid," M.S. thesis, School Electron. Inf. Elect. Eng., Shanghai Jiaotong Univ., Shanghai, China, 2011.

[38] X. He, Q. Ai, P. Yuan, and X. Wang, "The research on coordinated operation and cluster management for multi-microgrids," in *Proc. IET Int. Conf. Sustain. Power Gener. Supply (SUPERGEN)*, Hangzhou, China, 2012, pp. 1–3.

[39] X. Ni, Q. Ruan, S. Mei, and G. He, "A new network partitioning algorithm based on complex network theory and its application in Shanghai power grid," *Power Syst. Technol.*, vol. 31, no. 9, pp. 6–12, May 2007.

[40] R. Zimmerman, C. Murillo-Sánchez, and D. Gan, *Matpower User's Manual, Version 4.1*, Power Syst. Eng. Res. Center, Tempe, AZ, USA, 2011.

[41] Y. Zhao, Y. An, and Q. Ai, "Research on size and location of distributed generation with vulnerable node identification in the active distribution network," *IET Gener. Transm. Distrib.*, vol. 8, no. 11, pp. 1801–1809, Nov. 2014.

[42] W. Huang *et al.*, "An impedance protection scheme for feeders of active distribution networks," *IEEE Trans. Power Del.*, vol. 29, no. 4, pp. 1591–1602, Aug. 2014.

[43] J. Xu, G. Yang, Y. Yin, H. Man, and H. He, "Sparse-representation-based classification with structure-preserving dimension reduction," *Cogn. Comput.*, vol. 6, no. 3, pp. 608–621, 2014.

[44] G. E. Pfander, H. Rauhut, and J. Tanner, "Identification of matrices having a sparse representation," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5376–5388, Nov. 2008.

[45] I. Jana, K. Saha, and A. Soshnikov, "Fluctuations of linear eigenvalue statistics of random band matrices," *ArXiv e-prints*, Dec. 2014. [Online]. Available: http://arxiv.org/pdf/1412.2445.pdf

[46] X. He *et al.*, "3d power-map for smart grids—An integration of high-dimensional analysis and visualization," *ArXiv e-prints*, Mar. 2015. [Online]. Available: http://arxiv.org/pdf/1503.00463.pdf

[47] X. He, Q. Ai, R. C. Qiu, W. Huang, and J. Long. "An unsupervised learning method for early event detection in smart grid with big data," *ArXiv e-prints*, Jan. 2015. [Online]. Available: http://arxiv.org/pdf/1502.00060.pdf

[48] Y. Cao *et al.*, "A random matrix theoretical approach to early event detection using experimental data," *ArXiv e-prints*, Mar. 2015. [Online]. Available: http://arxiv.org/pdf/1503.08445.pdf

Xing He received the Bachelor's degree from Southeast University, Nanjing, China, and the Master's degree from Shanghai Jiao Tong University, Shanghai, China, in 2008 and 2012, respectively, both in electrical engineering. He is currently pursuing the Ph.D. degree with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University.

His current research interests include microgrid, intelligent algorithms, and big data.

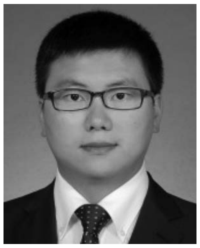
Qian Ai (M'03) received the Bachelor's degree from Shanghai Jiao Tong University, Shanghai, China; the Master's degree from Wuhan University, Wuhan, China; and the Ph.D. degree from Tsinghua University, Beijing, China, in 1991, 1994, and 1999, respectively, all in electrical engineering.

He was with Nanyang Technological University, Singapore, for one year; the University of Bath, Bath, U.K., for two years; and then with Shanghai Jiao Tong University, where he is currently a Professor with the School of Electronic Information and Electrical Engineering. His current research interests include power quality, load modeling, smart grids, microgrid, and intelligent algorithms.

Robert Caiming Qiu (S'93–M'96–SM'01–F'14) received the Ph.D. degree in electrical engineering from New York University, New York, NY, USA, in 1996.

He is currently a Full Professor with the Department of Electrical and Computer Engineering, Center for Manufacturing Research, Tennessee Technological University, Cookeville, TN, USA, where he was an Associate Professor in 2003 and a Full Professor in 2008. He was with General Telephone & Electric Laboratories Inc. (now Verizon), Waltham, MA, USA, and Bell Laboratories, Lucent, Whippany, NJ, USA. In 1998, he developed the first three courses on 3G for Bell Laboratories researchers. He was an Adjunct Professor with Polytechnic University, Brooklyn, NY, USA. He was the Founder-CEO and the President of Wiscom Technologies Inc., Clark, NJ. Wiscom was sold to Intel in 2003. His current research interests include wireless communication and networking, machine learning, and the smart grid technologies. He has researched on wireless communications and network, machine learning, smart grid, digital signal processing, electromagnetic scattering, composite absorbing materials, radio frequency microelectronics, ultrawideband (UWB), underwater acoustics, and fiber optics. He holds over six patents, and has authored over 70 journal papers/book chapters and 90 conference papers. He has 15 contributions to the 3rd Generation Partnership Project and IEEE standards bodies. He has co-authored the books *Cognitive Radio Communication and Networking: Principles and Practice* (Wiley, 2012) and *Cognitive Networked Sensing: A Big Data Way* (Springer, 2013), and authored the book *Introduction to Smart Grid* (Wiley, 2014).

Dr. Qiu serves as an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and other international journals. He is a Guest Book Editor for *Ultra-Wideband Wireless Communications* (Wiley, 2005), and three special issues on UWB, including the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE TRANSACTION ON SMART GRID. He serves as a Member of Technical Program Committee for the Global Telecommunications Conference, International Conference on Communications, Wireless Communications and Networking Conference, Military Communications Conference, and International Conference on Ubiquitous Wireless Broadband. He served on the advisory board of the New Jersey Center for Wireless Telecommunications. He is included in Marquis Who's Who in America.

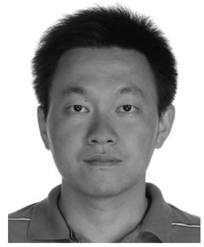


Wentao Huang was born in Anhui, China, in 1988. He received the B.Sc. degree from Shanghai Jiao Tong University, Shanghai, China, in 2010, where he is currently pursuing the Ph.D. degree, both in electrical engineering.

His current research interests include the protection and control of active distribution system, microgrids, smart grid, and renewable energy.

Longjian Piao is currently pursuing the M.Sc. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China.

His current research interests include application of multiagent system in smart grid and intelligent transport system, user interface design and programming of supervisory control and data acquisition systems, and coordinated charging strategies for electric vehicles.



Haichun Liu received the Bachelor's degree from Shanghai University, Shanghai, China, and the Master's degree from Shanghai Jiao Tong University, Shanghai, in 2004 and 2008, respectively, both in electrical engineering. He is currently pursuing the Ph.D. degree with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University.

His current research interests include machine learning, mobile robot control, and data mining.