SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

June 2023, Vol. 66 162301:1–162301:15
https://doi.org/10.1007/s11432-021-3522-0

Mining KPI correlations for non-parametric anomaly diagnosis in wireless networks

Tengfei SUI, Xiaofeng TAO*, Huici WU, Xuefei ZHANG, Jin XU & Guoshun NAN

National Engineering Laboratory for Mobile Network Technologies, Beijing University of Posts and Telecommunications, Beijing 100876, *China*

Received 29 October 2021/Revised 10 March 2022/Accepted 24 June 2022/Published online 10 May 2023

Abstract The increase in mobile data traffic has imposed unprecedented pressure on wireless network management. KPI-based anomaly diagnosis can alleviate such pressure by automatically identifying the cause of abnormalities in the traffic and providing end-to-end monitoring and optimization. Previous approaches mainly focus on finding a subset of anomaly-inducing KPIs on the basis of supervised learning procedures. These studies have two possible limitations: (1) the inherent correlations between KPIs that are proven to be effective for the anomaly diagnosis, are still largely underexplored; (2) machine learning models heavily rely on human annotations, which are expensive and labor-intensive. Therefore, we propose random matrix theory-based KPI identification (RKI), a novel method that automatically mines rich interactions between KPIs for anomaly diagnosis without using any learnable parameters or human annotations. Specifically, RKI diagnoses the abnormal KPIs in two steps. First, we build a matrix for anomaly KPI detection to mine the spectrum of its covariances. Second, another new matrix is reconstructed to calculate the correlation difference. By doing so, the anomaly KPIs that have larger correlation difference scores can be efficiently identified in the wireless traffic without any trainable parameters. In extensive experiments on a public dataset, RKI yields a 6.5% higher true diagnostic rate and 11.36% lower false alarming rate than the statistical model, demonstrating its effectiveness. A 100× larger scale synthetic dataset also demonstrates the capabilities of RKI to explore massive data traffic under real-word scenarios. Finally, we discuss RKI's potential applications of our method in future 6G wireless networks.

Keywords random matrix theory, spectral distribution, anomaly diagnosis, data analysis, wireless networks

Citation Sui T F, Tao X F, Wu H C, et al. Mining KPI correlations for non-parametric anomaly diagnosis in wireless networks. Sci China Inf Sci, 2023, 66(6): 162301, https://doi.org/10.1007/s11432-021-3522-0

1 Introduction

Previous studies reported that because of the increasing popularity of mobile applications and the growing complexity of cellular networks, mobile data traffic would reach 77.5 exabytes per month by 2022, annually growing by 46% from 2019 [1]. This exponential expansion of data traffic will impose unprecedented management pressure on delivering satisfactory wireless services with expected failure-to-response time, leading to frequent anomalies and causing atypical changes in wireless network traffic. In addition, the expansion will also significantly increase both operational and capital expenditures of operators [2]. The issue will be further aggravated in the upcoming launch of 6G networks with much more data traffic and higher service requirements [3, 4]. In this context, diagnosing the cause of network failures can help facilitate the end-to-end optimizations and assist operators in making the wireless network more reliable.

Early studies on anomaly diagnosis relied on excessive human efforts and additional expert knowledge, which are not practical and flexible under complex network conditions with massive data traffic [5]. Network can be well-monitored by various key performance indicators (KPIs), such as the number of dropped and blocked calls, failed handovers, and network traffic volume. The recently proposed KPI identification-based anomaly diagnosis [6] is a promising direction for a more reliable network because

* Corresponding author (email: taoxf@bupt.edu.cn)

KPIs help operators understand the network conditions in their daily tasks. Key performance indicators are sensitive to network status changes, which have been widely used for anomaly KPI detection [7–9]. Furthermore, KPIs can be customized based on the basis of service requirements.

As highlighted in [10,11], the identification of anomaly KPIs from large datasets is essential for anomaly detection and diagnosis and the design of self-management and self-evolution in 6G networks. Despite some underlying connections between the anomaly detection and diagnosis, their differences are obvious. The anomaly detection corresponds to detect patterns of observable network traffic or feature that do not conform to the expected notion of normal behaviors [12]. On the other hand, anomaly diagnosis aims to discover and characterize critical anomalies affecting the networks. Whereas anomaly diagnosis is a multi-stage process with anomaly detection identifying the abnormal network behaviors and categorizing the root cause of the anomaly [13]. This article studies the anomaly diagnosis by identifying the anomaly KPIs that cause anomalous network behaviors.

Previous studies on the KPI identification-based anomaly diagnosis can be categorized into two groups, i.e., statistical and neural methods. The former uses the classic statistical methods to learn parameters, such as clustering [14], principle component analysis (PCA) [15], KMeans [16], and one-class support vector machines (SVMs) [17]. The latter leverages long short term memory (LSTM) [18], convolutional neural networks (CNN) [19] to learn and diagnose. These methods automatically learn useful knowledge from data during the training process for anomaly diagnosis.

Despite their effectiveness, previous KPI identification based anomaly diagnosis approaches may have two limitations. (1) The inherent correlations between KPIs, which have proven to be effective for the anomaly diagnosis [8, 20], are still largely underexplored. This article will empirically show that mining correlations for anomaly diagnosis can greatly improve the overall performance. (2) The aforementioned machine learning models heavily rely on human annotations, which are expensive and labor-intensive. With a small scale of training data, a model trained for one case may not be well-generalized for another case.

To this end, we propose random matrix theory (RMT)-based KPI identification (RKI), a novel method that can automatically mine rich interactions between KPIs for anomaly diagnosis, without relying on any learnable parameters and human annotations. Particularly, we diagnose the abnormal KPIs in two steps. First, we build a matrix for anomaly KPI detection to mine the spectrum of its covariances. Second, we generate another matrix for correlation difference calculation. Consequently, the anomaly KPIs, which have larger correlation difference scores, can be efficiently identified without any trainable parameters. The rationale behind is that RMT [21, 22] used in our method can reveal the inter-relationship between entries in the matrix that follows various probability distributions, where the entries are referred to as random matrix ensembles without training. Experiments on two datasets showed the effectiveness of our approach. We also introduce a 100× larger synthetic dataset to validate the capabilities of RKI in exploring massive data traffic under real-world scenarios.

- We propose RKI, a novel method that aims to mine the rich correlations between KPIs for better anomaly diagnosis. The proposed RKI is a non-parametric approach without relying on any trainable parameters.
- We show some very interesting findings during the diagnosis procedure. For example, (1) the Pearson correlation results of each KPI suggest that anomaly KPIs are more correlated than normal KPIs; (2) the anomaly KPI can be identified by the larger correlation difference scores.
- We conduct extensive experiments on both real-world network data and synthetic data under various settings to show the effectiveness of our RKI method. We observe that computation cost of our approach is much lower than the one of the neural network based method, while the performance is comparable.
- We also discuss the potential application of our method in the next generation wireless network (6G), including the energy efficiency and latency control.

The rest of the study is organized as follows. Existing studies about anomaly diagnosis are summarized in Section 2. Section 3 formulates the problem by addressing the correlation analysis in anomaly networks. Section 4 presents the new RKI method for anomaly diagnosis. The RMT spectral analysis for anomaly KPI detection and the cleaned correlation matrix construction are addressed, wherein an anomaly KPI identification method is also introduced. In Section 5, both a real-world dataset and synthetic dataset with anomalies are employed to validate the effectiveness and advantages of our proposed RKI method. The supportive experiments and potential applications to future 6G networks are also discussed. Section 6 concludes the study.

2 Related work

Wireless network data traffic not only can be mapped onto specific time-domain patterns related to geographical locations, but also are becoming increasingly correlated in time and space [23]. An important consequence of the high correlation in data lies in network anomalies, which may jeopardize the network performance.

2.1 KPI identification in anomaly diagnosis

The approaches for KPI identification in anomaly diagnosis can be roughly classified into two categories, statistical model and neural methods. The statistical model based approaches focus on locating different characteristics of KPI data. For example, clustering based methods [14] cluster different data samples to find anomalies without a full consideration of the data correlation or structure. The authors in [15] performed a PCA-based approach for a selection of KPIs by means of dimension reduction. KMeans clustering algorithm [16] is also utilized for anomaly detection and diagnosis, which trains data to classify new data as normal or anomalous. By extracting an optimum feature subset, a one-class SVMs [17] approach is applied to anomaly detection in temporal correlated data. A supervised filtering technique based overlapping (OV) approach [9] is put forward for KPI selection in automatic diagnosis, which examines the dissimilarities of the statistical behaviors under different network states. Yang et al. [24] put forth a correlation-based traffic classification method, which has taken into consideration an average correlation between datasets, but not the inter-correlation of individual entries. Notably, the performance of traditional statistical techniques referenced above crucially depends on the labeled samples selected from the network experts' experience, and assumptions are made to fit the models for simplification's sake rather than capturing the data correlations properly.

Neural methods have emerged as a productive endeavor to detect and diagnose hidden anomalies in temporal correlated data. In [18], LSTM networks are utilized for anomaly detection and diagnosis in multivariate time series data. CNN [19] is also applied to network anomaly detection, which employs a convolutional encoder to encode the temporal correlations. A deep learning based Transformer approach [25] for multivariate time series anomaly detection in IoT is employed, which utilizes temporal dependency to describe the anomaly information flow between network nodes. Although these model based methods have demonstrated their advantages in various detection applications to multivariate highly correlated network data traffic, it is an inevitable cost with respect to the huge labor and computation resource.

2.2 Application of RMT in anomaly diagnosis

Originated from physics, mathematical statistics and numerical analysis at the beginning of the 20th century, RMT has become one of the statistical foundations for big data analytics [26], and has been widely applied to a variety of fields, such as quantum systems [27, 28], financial systems [29, 30], political strategies [31], biological systems [32], smart grid systems [21, 22], and wireless communication networks [33–37]. A previous study [7] proposed a data decomposition aided RMT method for anomaly detection in wireless networks, which can decompose data into stochastic and regular components. The RMT-based method is applied to non-Gaussian environments in power systems [38], treating the power system data as a time series. A data-driven approach is developed for early anomaly detection and localization in distribution network on the basis of RMT [39]. By analyzing the structure information of the data in wireless networks. The authors in [40] proposed a spatio-temporal correlation analysis approach for anomaly detection and location. The studies suggest the effectiveness of RMT in analyzing highly-correlated datasets for anomaly detection in wireless networks. On the basis of successful anomaly KPI detection using RMT, we intend to build a real-time non-parametric anomaly diagnosis method.

3 Problem formulation

Since the data structure tends to be damaged in univariate statistical results and model-based methods, and the data correlation information is often overlooked, leading to partial diagnoses, we take into account the data correlations of diverse KPIs in anomaly KPI identification. As the correlations of KPIs are

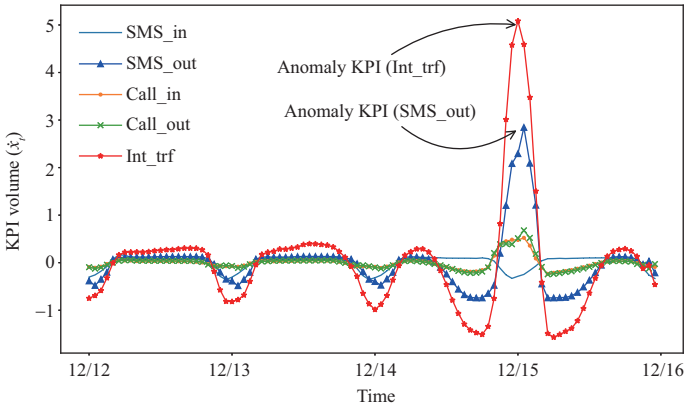

Figure 1 (Color online) Moving average of raw data statistical results.

connected with network status, the network anomalies can be diagnosed from the identified anomaly KPIs. In this section, we focus on the KPI correlation problem.

3.1 Correlation analysis for anomaly KPIs

We identify the anomaly KPIs by investigating the correlation differences of each KPI between the raw and the cleaned correlation matrices, so it is crucial to exploring the data correlation first.

A real-world network traffic dataset with anomalies is utilized for demonstration of the data correlation. The dataset logs the traffic volumes of each base station, including SMS in (SMS_in) activity, SMS out (SMS_out) activity, Call in (Call_in) activity, Call out (Call_out) activity, and Internet traffic (Int_trf) activity, with a 10-min time interval over two months, from November to December, 2013[1)]. For any KPI in the dataset, the network traffic KPI volume at time t is defined as x_t. To roughly estimate the trend of the normal and anomalous KPIs, a moving average (MA) filter is applied with the entry x_t rewritten as $\dot{x}_t$. The MA filter is given as $\dot{x}_t = \frac{1}{d}\sum_{i=0}^{d-1} x_{t-i}, d \leqslant t$, where d denotes the size of the MA window, and t stands for the time order of the data.

The statistical results of the volume of the normal and anomalous KPIs are plotted in Figure 1, where the curves indicate the raw data averaged by a 3-point moving average filter ($d = 3$). It shows that the anomaly KPIs accumulate a much more network traffic volume than normal KPIs. The similar fluctuations of anomalous and normal KPIs suggest that they are both time-correlated. Thereby the anomaly network status can be deduced from the anomaly KPI identification.

3.2 Correlation coefficients

Our goal is to expose the numerical property differences between anomalous and normal KPIs in wireless networks. In particular, we endeavor to show the intrinsic data correlation differences in the corresponding matrices. Theoretically, the correlations between different KPIs can be derived from the standard Pearson correlation coefficient (SPCC) [41,42]. Let $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ be two zero-mean real-valued vectors, and then the SPCC between them can be defined as

$$\rho(\boldsymbol{e}_i, \boldsymbol{e}_j) = \frac{E[\boldsymbol{e}_i \boldsymbol{e}_j]}{\sigma_{\boldsymbol{e}_i} \sigma_{\boldsymbol{e}_j}}, \tag{1}$$

where $E[\boldsymbol{e}_i \boldsymbol{e}_j]$ is the cross-correlation between $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$, $\sigma^2_{\boldsymbol{e}_i} = E[\boldsymbol{e}_i^2]$ and $\sigma^2_{\boldsymbol{e}_j} = E[\boldsymbol{e}_j^2]$ are the variance of $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$, respectively.

One of the most important properties of SPCC is $0 \leqslant \rho(\boldsymbol{e}_i, \boldsymbol{e}_j) \leqslant 1$. SPCC serves as an indicator of the strength of the relationship between the two KPIs $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$. If $\rho(\boldsymbol{e}_i, \boldsymbol{e}_j) = 0$, then $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ can be considered uncorrelated. The closer the value of $\rho(\boldsymbol{e}_i, \boldsymbol{e}_j)$ is to 1, the stronger the correlation between the KPIs is.

Figure 2 displays the correlation matrices of the dataset with and without anomaly KPIs. The diagonals represent the auto-correlation of the KPIs, and the off-diagonals represent the cross-correlation. The value of off-diagonals is relatively smaller than that of diagonals in both Figures 2(a) and (b), which suggests

1) The milano grid spatial description. 2014. https://dandelion.eu/datagems/SpazioDati/milano-grid/description/.

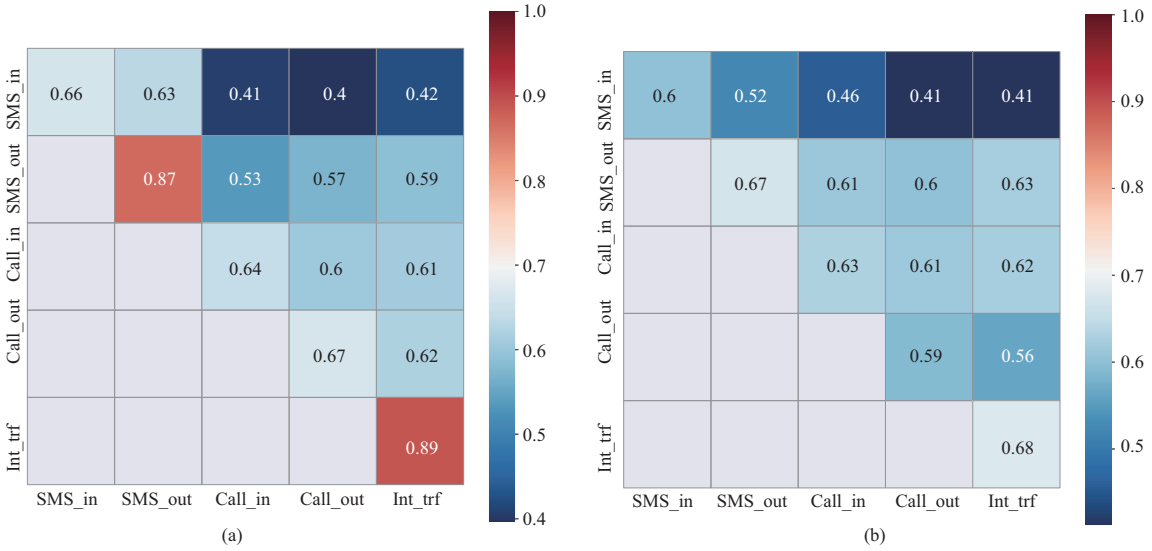

Figure 2 (Color online) Pearson correlation coefficient of real data. (a) SPCC of anomaly data; (b) SPCC of anomaly-free data.

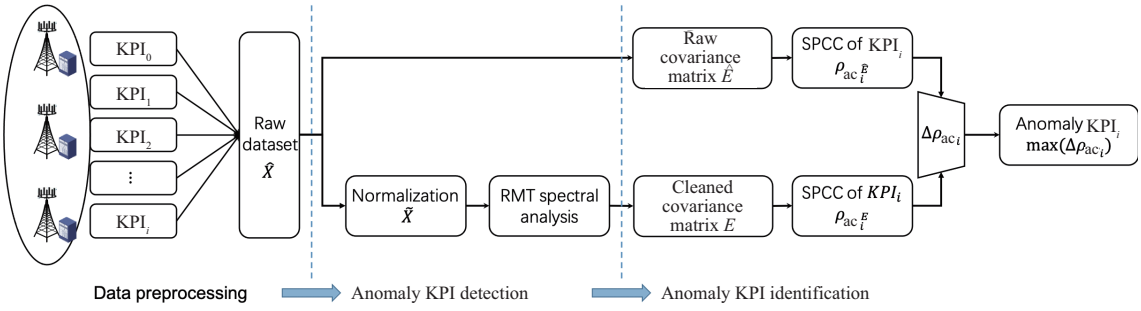

Figure 3 (Color online) Illustration of the RKI framework for anomaly diagnosis in wireless networks. The data collected from real networks are preprocessed before the detection of anomaly KPIs. A comparison between raw and cleaned covariance matrices is conducted for anomaly KPI identification.

that the cross-correlations of different KPIs are not as sensitive to anomalies as the auto-correlation. It is also illustrated in Figure 2(a) that SMS out and Internet traffic are the most correlated KPIs with their auto-correlations at 0.87 and 0.89, whereas all the other KPIs show much smaller correlations. Crucially, the characteristic higher correlation of the anomaly KPIs in Figure 2(a) is consistent with our observations in Figure 1, as the SMS out and Internet traffic show much higher auto-correlations than the other KPIs. Therefore, the anomaly KPIs with larger correlation difference scores can be efficiently identified in the wireless traffic with two consecutive steps. (1) Firstly, a raw matrix for anomaly KPI detection is built to mine the spectrum of its covariances. (2) Then another new matrix with anomaly KPI removed (cleaned matrix) is reconstructed for correlation difference calculation.

4 RMT based KPI identification

Since anomaly KPIs are characterized by higher auto-correlations than normal ones, anomaly KPIs can be identified by investigating the auto-correlational difference of sampled matrix with and without anomalies. The RKI method that we propose for network anomaly diagnosis is based on this consideration. First of all, we detect the anomalies with the aid of RMT spectral analyses. The application of RMT to anomaly detection and anomaly KPI identification is illustrated in Figure 3, which shows the RKI framework for anomaly diagnosis in wireless networks.

Firstly, we build a matrix for anomaly KPI detection to mine the spectrum of its covariances. Then another new matrix is reconstructed for correlation difference calculation. By doing so, the anomaly KPIs, which have larger correlation difference scores, can be efficiently identified in the wireless traffic without any trainable parameters.

4.1 Anomaly KPI detection of the raw matrix using RMT spectral analysis

Based on the analysis of the empirical spectrum distribution (e.s.d.) of the covariance matrix of both the normal and anomalous network data traffic, we apply RMT to the anomaly detection of highly correlated network data traffic, and use the linear statistics of the eigenvalues as an indicator of anomalies.

Assuming the number of KPIs is N and the sampling time is T, without loss of generality, for individual KPI i and its sampling time j, we model the traffic volume of the KPI as $x_{i,j}$, and treat all the sampled data as a vector $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,T}) \in \mathbb{C}^{1\times T}$. By combining different sampling times and KPIs, the raw network traffic KPI matrix can be defined as $\hat{\boldsymbol{X}} = (\boldsymbol{x}_1^{\mathrm{T}}, \ldots, \boldsymbol{x}_N^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{C}^{N\times T}$, where T denotes the matrix transpose. We use the sample correlation matrix $\hat{\boldsymbol{E}}$ shown in (2) to estimate the raw matrix [22],

$$\hat{\boldsymbol{E}} = \frac{1}{T}(\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^{\mathrm{T}}) \triangleq (\boldsymbol{e}_1^{\mathrm{T}}, \boldsymbol{e}_2^{\mathrm{T}}, \ldots, \boldsymbol{e}_N^{\mathrm{T}})^{\mathrm{T}}, \tag{2}$$

where $\boldsymbol{e}_i = (e_{i,1}, \ldots, e_{i,N})^{\mathrm{T}} \in \mathbb{C}^{1\times N}$. Thereby the raw matrix for correlation computation can be obtained.

The detection of anomaly KPIs by the deviated spiked eigenvalues can be derived from a random matrix e.s.d., whose definition is provided in (3). It is an important metric to describe the eigenvalue distribution of a matrix.

Definition 1 (Empirical spectral distribution [43])**.** Considering an $N \times N$ Hermitian matrix $\boldsymbol{T}_N$, the e.s.d. $F^{\boldsymbol{T}_N}$ of the matrix $\boldsymbol{T}_N$ is defined as

$$F^{\boldsymbol{T}_N}(\lambda) = \frac{1}{N}\sum_{j=1}^{N} 1_{\{\lambda,\lambda_j\leqslant\lambda\}}(\lambda), \tag{3}$$

where $1_S(\cdot)$ is an indicator function over the set S, and $\{\lambda_1, \ldots, \lambda_N\}$ denotes the eigenvalues of $\boldsymbol{T}_N$.

With the definition of e.s.d. $F^{\boldsymbol{T}_N}(\lambda)$, the average eigenvalues that are smaller than a particular variable λ constitute a cumulative density function. The probability density function (PDF) $\rho(\lambda)$ of a standard correlation matrix is given in (4). The Marcenko-Pastur (M-P) law is a closed-form e.s.d. of one particular type of random matrix [43], which is defined in Theorem 1.

Theorem 1. Consider a matrix $\boldsymbol{X}_{N,T}$ with i.i.d. entries $(\frac{1}{\sqrt{T}}\boldsymbol{X}_{i,j})$, such that $\boldsymbol{X}_{1,1}$ has zero mean and unit variance. As $N, T \to \infty$ with $\frac{N}{T} \to c \in (0, \infty)$, the e.s.d. of $\boldsymbol{R}_{N,N} = \boldsymbol{X}_{N,T}\boldsymbol{X}_{N,T}^{\mathrm{T}}$ converges weakly and almost surely to a non-random distribution function F_c with density ρ_c:

$$\rho_c(\lambda) = (1-\lambda^{-1})1_{\{\lambda=0\}}(\lambda) + \frac{1}{2\pi c\lambda}\sqrt{(\lambda-\lambda_-)^+(\lambda_+-\lambda)^+}, \tag{4}$$

where $\lambda_- = (1-\sqrt{c})^2$, $\lambda_+ = (1+\sqrt{c})^2$, $1_{\{x=0\}}(x)$ is the indicator function, and $(\cdot)^+$ denotes $\max(0, x)$. The distribution of λ is restricted to the interval $[\lambda_-, \lambda_+]$.

Theorem 1 states that when the entries of a random matrix are i.i.d., the eigenvalue distribution of the correlation matrix follows from (4); otherwise the theorem cannot be satisfied. Note the sharp edges of the spectrum, which are only valid when $N, T \to \infty$. For finite N, T, there is a small probability to find eigenvalues out of the interval $[\lambda_-, \lambda_+]$. The analysis of the e.s.d. of the raw KPI matrix $\hat{\boldsymbol{X}}$ requires normalization to zero mean and unit variance. Thus the raw data matrix $\hat{\boldsymbol{X}} \in \mathbb{C}^{N\times T}$ is normalized to a non-Hermitian matrix $\tilde{\boldsymbol{X}} \in \mathbb{C}^{N\times T}$ with entries defined as $\tilde{x}_{i,j} = (x_{i,j} - \bar{x}_j)/\sigma(x_j)$, where $1 \leqslant i \leqslant N, 1 \leqslant j \leqslant T$, and $\bar{x}_j$ and $\sigma(x_j)$ are the mean and the standard deviation of the jth column of the matrix $\hat{\boldsymbol{X}}$, respectively. The eigenvalues of the correlation matrix $\tilde{\boldsymbol{E}} = \frac{1}{T}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^{\mathrm{T}}$ are calculated before the comparison of the histogram of the eigenvalues with the M-P law.

Figure 4 depicts the e.s.d. of a random matrix $\tilde{\boldsymbol{E}}$. The green histograms indicate the e.s.d. of the raw data with or without anomalies, and the red solid line is the M-P law derived from (4). Figure 4 shows that the bulk of the eigenvalues fits the M-P law in the anomaly-existing scenarios, but the spike suggests that the eigenvalues are populated by anomaly signals.

The eigenvalues outside the region of the M-P law are called spikes, which functions as a benchmark for the hypothesis of the i.i.d. matrix. The high eigenvalues correspond to the meaningful information in anomaly KPI detection, and the largest eigenvalue deviation signifies the different distribution of anomaly KPIs from the normal. Therefore, the anomaly KPIs can be diagnosed by the existence of spikes.

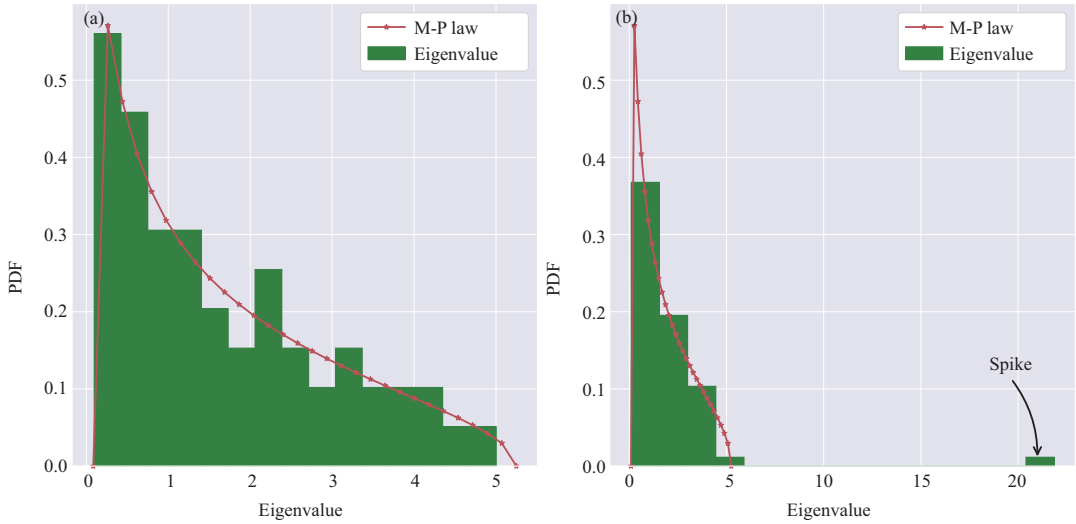

Figure 4 (Color online) The e.s.d. of $\tilde{\boldsymbol{E}}$ and its comparison with the M-P law in correspondence to the normal and anomaly network status. (a) $\tilde{\boldsymbol{X}}$ is a 60×100 random Gaussian matrix. (b) $\tilde{\boldsymbol{X}}$ is a 60×100 random Gaussian matrix with one row of anomaly signals.

On the basis of the M-P law, we can draw the conclusion that the e.s.d. of random matrices are different with and without anomalies, and we utilize it for anomaly KPI detection [7]. The linear eigenvalue statistics (LES) serves as an indicator for the existence of anomaly KPIs, which can be defined as

$$\mathcal{N}_\phi = \sum_{i=1}^{n} \phi(\lambda_i), \tag{5}$$

where λ_i $(i = 1, 2, \ldots, n)$ are the eigenvalues in descending order, and $\phi(\cdot)$ is a test function that makes a linear mapping for the eigenvalues λ_i. Many test functions have been widely investigated. In this paper we only utilize the information entropy function, which is defined as $\phi(\lambda_i) = -\lambda_i \ln \lambda_i$. More details about the test functions can be found in [39]. By observing the change of $\mathcal{N}_\phi$, we can confirm the existence of anomaly KPIs.

4.2 Cleaned correlation matrix construction

The data correlations of anomalous and normal KPIs constitute the vital measurement for our proposed anomaly KPI identification method. Since we have obtained the raw matrix, we need to construct a cleaned correlation matrix for correlation comparison. We adopt RMT to reveal the matrix characteristics with respect to the data structure and statistics for a cleaned correlation matrix. In particular, we use the M-P law to analyze the spectral distribution of the raw correlation matrix. The M-P law is typically employed to analyze the spectral distributions of a random matrix. Two steps are essentially involved in the construction of the cleaned correlation matrix: (i) the detection of the anomaly KPIs by the deviated eigenvalues obtained from the M-P law, which can be considered spiked eigenvalues, and (ii) the identification of anomaly KPIs for anomaly diagnosis.

The comparison of raw data with RMT not only facilitates the detection of spiked eigenvalues in anomaly correlation matrices but also provides a means to reduce them. Considering that the e.s.d. of the anomaly KPI matrix is more like a spiked model, we can reduce the spike and analyze the majority eigenvalues in the anomaly KPI dataset. One of the reduction methods in RMT is eigenvalue cleansing, for less than 5% of the eigenvalues carry most of the information in the spiked correlation matrix [43]. So our first step consists of analyzing the correlation matrix with the spectral decomposition, for instance, the singular value decomposition [44] in (6),

$$\tilde{\boldsymbol{E}} = \sum_{i=1}^{N} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^{\mathrm{T}}, \tag{6}$$

where $\lambda_N \geqslant \cdots \geqslant \lambda_1$ are the eigenvalues of the matrix $\tilde{\boldsymbol{E}}$ in a descending order, and $\{\boldsymbol{v}_k\}$ denotes the set of the corresponding eigenvectors.

By the M-P law, only the eigenvalues outside the M-P law contain useful information. Hence, supposing there are k eigenvalues outside the M-P law, the cleaned correlation matrix can be constructed as

$$\boldsymbol{E} = \sum_{i=0}^{k-1} \lambda_{N-i} \boldsymbol{v}_{N-i} \boldsymbol{v}_{N-i}^{\mathrm{T}} + \alpha \boldsymbol{I}_N, \tag{7}$$

where $\boldsymbol{I}_N$ is an $N \times N$ identity matrix, and α is a parameter chosen to preserve the trace of the matrix $\tilde{\boldsymbol{E}}$ shown as $\alpha = \frac{1}{N} \sum_{i=1}^{k} \lambda_i$.

4.3 KPI identification for anomaly diagnosis

Based on the reconstructed correlation matrix obtained from (7), we can further conduct the anomaly KPI identification by comparing the correlational difference between the anomaly and the normal KPI matrices.

Concerning the anomaly KPI identification, we specifically consider the anomaly-existing scenario. In other words, there is an anomaly KPI in the T consecutive sampling period. The auto-correlation of each KPI between the raw and the cleaned correlation matrices can be defined as $\rho_{\mathrm{ac}_i}^{\hat{\boldsymbol{E}}}$ and $\rho_{\mathrm{ac}_i}^{\boldsymbol{E}}$, respectively. The diagnostic scheme can be conducted by computing the difference of auto-correlation for each KPI between the raw correlation matrix $\hat{\boldsymbol{E}}$ obtained from (2) and the cleaned correlation matrix $\boldsymbol{E}$ obtained from (7) via

$$\mathrm{KPI}_i = \begin{cases} \text{Anomaly}, & \text{if } \Delta\rho_{\mathrm{ac}\,i} \geqslant \overline{\Delta\rho_{\mathrm{ac}}}, \\ \text{Normal}, & \text{others}, \end{cases} \tag{8}$$

where $\Delta\rho_{\mathrm{ac}\,i} = |\rho_{\mathrm{ac}_i}^{\hat{\boldsymbol{E}}} - \rho_{\mathrm{ac}_i}^{\boldsymbol{E}}|$, and $\overline{\Delta\rho_{\mathrm{ac}}} = \frac{1}{N} \sum_{i=1}^{N} \Delta\rho_{\mathrm{ac}\,i}$, which is the mean of $\Delta\rho_{\mathrm{ac}}$. In the anomaly-existing scenario, the change of $\Delta\rho_{\mathrm{ac}\,i}$ can help identify anomaly KPIs, as a KPI's $\Delta\rho_{\mathrm{ac}\,i}$ is positively correlated with the network anomalies, the larger the $\Delta\rho_{\mathrm{ac}\,i}$ of a KPI is, the more closely it is correlated with the network anomalies, and vice versa. Therefore the $\Delta\rho_{\mathrm{ac}\,i}$ bigger than the mean of $\Delta\rho_{\mathrm{ac}}$ corresponds to the identified anomaly KPI.

Couched within the framework presented above, we design the algorithm of anomaly KPI identification for anomaly diagnosis with highly-correlated network data traffic. The fundamental steps are given in Algorithm 1. Steps 1–3 construct the raw correlation matrix, and step 4 computes the auto-correlation of the raw correlation matrix. Steps 5–9 build the cleaned correlation matrix, and step 10 computes the auto-correlation of the cleaned correlation matrix. The derivation of the auto-correlational difference for each KPI in Steps 11 and 12 enables the identification of the anomaly KPI, which exhibits the largest $\Delta\rho_{\mathrm{ac}\,i}$.

Algorithm 1 Steps of RKI approach to anomaly diagnosis in wireless networks

1: Model the raw dataset into a raw correlation matrix $\hat{\boldsymbol{X}}$;
2: Normalize matrix $\hat{\boldsymbol{X}}$ into $\tilde{\boldsymbol{X}}$;
3: Construct the sample correlation matrix $\hat{\boldsymbol{E}}$ with (2);
4: Anomaly KPI detection
5: Calculating the e.s.d. with (3);
6: Detect the anomaly KPIs by observing the $\mathcal{N}_\phi$ with (5);
7: Calculate the auto-correlation of each KPI in $\hat{\boldsymbol{E}}$;
8: Obtain the $\rho_{\mathrm{ac}_i}^{\hat{\boldsymbol{E}}}$ with (1);
9: Calculate the auto-correlation of each KPI in $\boldsymbol{E}$;
10: Decompose the normalized matrix $\tilde{\boldsymbol{X}}$ with (6);
11: Reconstruct matrix $\tilde{\boldsymbol{E}}$ into $\boldsymbol{E}$ with (7);
12: Obtain the $\rho_{\mathrm{ac}_i}^{\boldsymbol{E}}$ with (1);
13: Calculate the auto-correlation difference $\Delta\rho_{\mathrm{ac}}$ of each KPI;
14: Identify the corresponding anomaly KPI_i with $\Delta\rho_{\mathrm{ac}\,i}$ by (8).

5 Experiments

This section presents experiments conducted using an open public dataset obtained from the city of Milan as well as a synthetic dataset to demonstrate the effectiveness of our proposed RKI approach. We give descriptions of the data preprocessing, experimental settings, and performance metrics before

Table 1 Dataset description

Date	Time	SMS_in activity	SMS_out activity	Call_in activity	Call_out activity	Int_trf activity
2013/12/10	00:00:00	5.91618767	9.56081341	0.45037942	0.69351206	201.178145
	01:00:00	0.2854841	0.30205163	0.46022041	0.07162857	158.693158
	⋮	⋮	⋮	⋮	⋮	⋮
	23:00:00	1.93045836	2.71307298	0.50709557	0.38724501	120.690649
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2013/12/19	00:00:00	2.74104861	5.54177034	0.2231084	2.69619718	142.865953
	01:00:00	1.23858358	0.69721883	0.14460397	0.38352118	121.255268
	⋮	⋮	⋮	⋮	⋮	⋮
	23:00:00	5.51559624	3.65261306	2.71335039	2.77525932	243.212695

providing evaluation of the RKI performance and comparison with different schemes. The simulations were conducted on a computer with 3.40 GHz Intel(R) Core(TM) i7-6800K CPU, 16.00 GB RAM and GeForce GTX 1070 GPU. We use Python 3.6 to implement our proposed method.

5.1 Experiments on real-network data traffic

Dataset description. The real-world spatio-temporal network traffic dataset is composed of computation over the call detail records (CDRs), which consists of network data traffic collected from a real LTE network of Telecom Italia at Milan, a major city located in the north of Italy, as described in Subsection 3.1. The dataset is made public for the Big Data Challenge 2014 competition [45]. To facilitate the data analysis, the Milan region is divided into 100×100 grids (named as Milano Grid), with each grid covering a square of 0.055 km^2, and all these base stations can be mapped into individual grids. When there are several BSs in one grid, all the traffic loads are aggregated into one traffic load. Grid 5848 with a Convention Center is selected for analysis as it contains verified network anomalies [7, 45]. For expository purpose, we choose a specific region in nine days' sampling time (from 2013/12/10 00:00:00 to 2013/12/19 23:00:00). The dataset is described in Table 1.

Main results. The main experiment results on real-network data traffic based on Algorithm 1 are as follows. A consecutive $\mathcal{N}_\phi$-t curve is shown in Figure 5(a). The surge of $\mathcal{N}_\phi$-t at 22:00 on December 14th can be considered an indication of the abnormal network status. A step further, Figure 5(b) provides the e.s.d. of the sampled raw network traffic KPI matrix. The green histograms indicate the e.s.d. of the raw data with anomalies, and the red solid line is the M-P law derived from (4). The sampling time T in this practical scenario is 24, and the number of involved KPI N is 5. The sampling time is not much larger than the number of the KPIs, where $c = N/T \approx 0.2083$, $\lambda_- = 0.2952$, and $\lambda_+ = 2.1210$. A spike can be observed outside of the M-P law, which corresponds to the $\mathcal{N}_\phi$-t curve. With one spike captured as displayed in Figure 5(a), the existence of anomaly KPIs can be determined. We set k to 1 in (7) for computational convenience. Figure 5(c) illustrates the correlational comparison between the cleaned and the raw correlation matrices obtained from Algorithm 1, where the solid red line denotes the cleaned correlation matrix derived from (7), and the dash blue line denotes the raw correlation matrix derived from (2).

The comparison shows that after the raw matrix is cleaned, the anomaly KPI in the dataset identified by (8) is the Int_trf activity ($\Delta\rho_{\text{Int\_trf}} = 0.19$), and the SMS_out activity ($\Delta\rho_{\text{SMS\_out}} = 0.18$) with $\overline{\Delta\rho_{\text{ac}}} = 0.086$. The anomaly identification results coincide with the statistical results in Figure 2(a).

We introduce two performance metrics, i.e., true diagnostic rate (TDR) and false alarm rate (FAR), to compare our RKI to other approaches. The definitions of the two metrics are given as follows:

$$\text{TDR} = \frac{N_{\text{cr}}}{N_{\text{gt}}}, \tag{9}$$

$$\text{FAR} = \frac{N_{\text{al}} - N_{\text{cr}}}{N_{\text{al}}}, \tag{10}$$

where N_{cr} denotes the instances of correctly identified anomaly KPIs in the total instances of Monte Carlo simulations, N_{gt} denotes the sum of ground-truth anomaly KPIs in the total simulation instances, and N_{al} is the instances of all identified anomaly KPIs. Given that the proposed RKI method can only

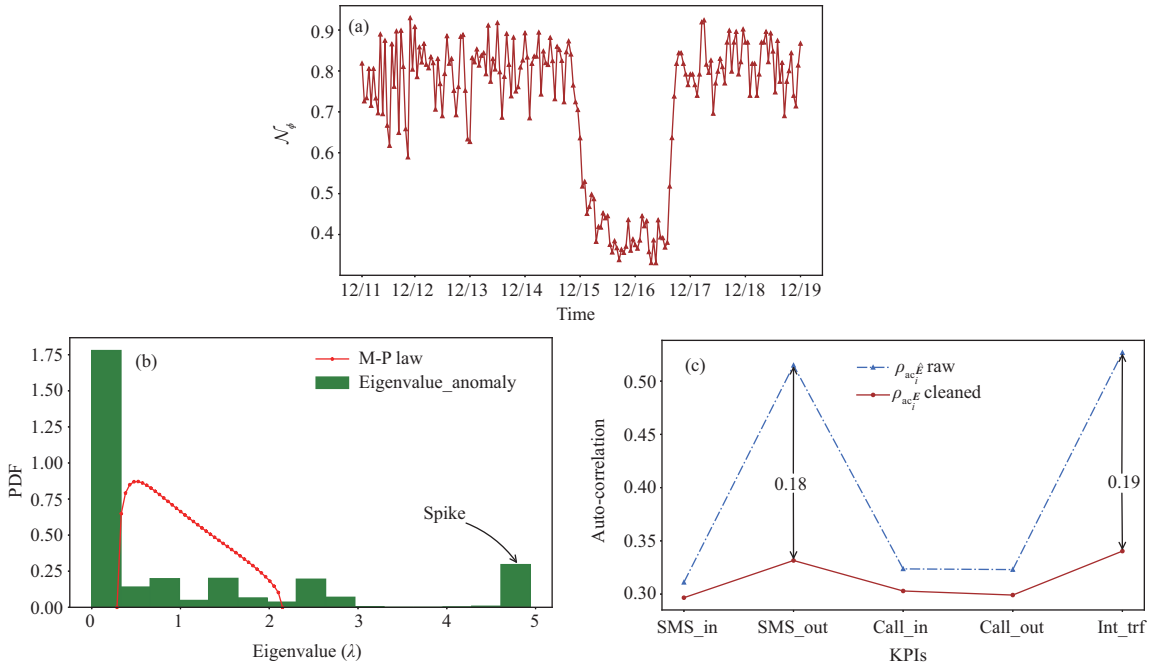

Figure 5 (Color online) Main simulation results of real data. (a) Anomaly KPI detection result of RKI with real data with $\mathcal{N}_\phi$-t curve exhibiting an abrupt change at 22:00, indicating possible occurrence of anomaly; (b) M-P law of anomaly real data with anomalies, and a spike is shown outside of the M-P law; (c) main correlational comparison between the cleaned and the raw correlation matrices of real data with anomaly KPI (SMS_out and Internet traffic) identified.

Table 2 Parameter settings involved in KPI identification approaches

Approach	Parameter setting
SVMs	The upper bound on the fraction of training errors v : 0.1; the kernel function: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\frac{1}{2N})\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$
KMeans+PCA	The number of clusters and components: 10, 5; the outliers fraction: 1.0%
OV	The overlapping function: $\text{OV}L(i,j,p) = \int \min(f_p(x\|i), f_p(x\|j))\text{d}x$; proportion for selection set, training set, test set: 0.4, 0.4, 0.2
LSTM	The model depth: 1 layer, 2 layers; the activation function: sigmoid; the optimizer: Adam
CNN	The kernel size: $2 \times 2, 3 \times 3$; the activation function: ReLu; the optimizer: Adam
Transformer	The encoder and decoder layers: 4, 6; the optimizer: Adam

detect the existence of anomaly KPIs, but cannot work out the exact instances of the anomaly and normal entries in the sampled raw and cleaned matrices, we thereby solely employ the times of identified anomaly KPIs in Monte Carlo simulations for TDR and FAR performance evaluation. The higher the TDR and the smaller the FAR, the better performance can the anomaly KPI identification algorithm achieve. Meanwhile, the average computation time (ACT) for each sampling time is utilized for efficiency comparison. For the neural methods, we consider the training time in ACT for teach testing sample.

In order to provide a more comprehensive description of the superior of our proposed RKI method, we compare the statistical model (SVMs, KMeans+PCA, OV) with the neural methods (LSTM, CNN, Transformer). Furthermore, we present experimental results from two datasets of different sizes, which are the small size (5×216) and the large size (5×1296). The settings of the simulation parameters for different KPI identification approaches are listed in Table 2, and the simulation results are shown in Table 3.

From Table 3 we can see that the RKI method can achieve a higher TDR and smaller FAR than methods based on statistical models. In the small data size (5×216) scenario, it can reach a 9.25% higher TDR and 11.30% smaller FAR than the SVMs method. Meanwhile, the RKI method achieves

Table 3 Main comparative results of different KPI identification approaches

Category	Approach	TDR (%) (small)	FAR (%) (small)	ACT (s) (small)	TDR (%) (large)	FAR (%) (large)	ACT (s) (large)
Statistical model	SVMs	76.25	27.92	0.0062	82.25	28.22	0.0090
	KMeans+PCA	82.75	22.56	0.00870	85.50	21.62	0.0102
	OV	84.00	21.38	0.0102	87.75	18.20	0.0142
Neural methods	$\text{LSTM}_{(1\ \text{layer})}$	83.50	20.00	0.2008	88.50	16.61	0.4016
	$\text{LSTM}_{(2\ \text{layers})}$	85.50	18.75	0.4297	88.75	14.71	1.0257
	$\text{CNN}_{(2\times2)}$	82.50	18.57	0.4032	89.50	12.32	0.6101
	$\text{CNN}_{(3\times3)}$	86.00	15.26	0.5620	91.25	13.10	1.3124
	$\text{Transformer}_{(4)}$	84.00	17.78	0.5122	88.50	14.72	0.7070
	$\text{Transformer}_{(6)}$	88.00	16.25	0.9107	89.75	12.20	1.4160
Non-parametric	RKI	**85.50**	**16.62**	**0.0230**	**89.50**	**13.69**	**0.0910**

Table 4 The synthetic data with anomaly signal in large dimensional scenario

Row number	Sampling time	Traffic load
1–48	t_s = 1–10000	Unchanged
49	t_s = 1–5000	$0.1\times\sin(\frac{\pi}{10}t_s+10)$
	t_s = 5001–10000	$1.0\times\sin(\frac{\pi}{10}t_s+10)$
50	t_s = 1–5000	$0.5\times\sin(\frac{\pi}{20}t_s+20)$
	t_s = 5001–10000	$1.0\times\sin(\frac{\pi}{20}t_s+20)$

a 2.75% higher TDR, 5.94% smaller FAR than the KMeans+PCA method, and a 1.50% higher TDR, 4.76% smaller FAR than the OV method. The TDR comparison suggests that the RMT based method can provide improved diagnostic performance compared with the statistically based SVMs, OV, and KMeans+PCA based methods.

Yet our proposed RKI method has the largest ACT compared to methods based on statistical models, due to the e.s.d. calculation and the cleaned correlation matrix construction. Considering that the real network data used in our experiment are sampled every 10 min, the proposed RKI approach is applicable to anomaly diagnosis in real networks. The simulation results of the neural method show a higher TDR and FAR than the proposed RKI method, which means the neural method can achieve a much higher diagnosis accuracy than the RKI method. Discussion is given in Subsection 5.3.

5.2 Experiments on synthetic data

Dataset description. Further validations of the RKI method are severely constrained by the scale of real data. Therefore, synthetic data of a large scale are specifically designed to provide proof for its practical use in the era of big data. The data consist of a 50×10000 matrix (100 times larger than the real data) $\boldsymbol{C}$ with the traffic volume of rows 1–48 remaining unchanged in the 10000 sampling times. In particular, an anomaly signal was set by altering the traffic volume at rows 49 and 50. It stays $0.1\times\sin(\frac{\pi}{10}t_s+10)$ (row 49) or $0.5\times\sin(\frac{\pi}{20}t_s+20)$ (row 50) during sampling time 1–5000, but changes to $1.0\times\sin(\frac{\pi}{10}t_s+10)$ (row 49) or $1.0\times\sin(\frac{\pi}{20}t_s+20)$ (row 50) during the sampling time of 5000–10000. Details about the large scale dataset can be found in Table 4.

For the sake of randomness, white noise $\boldsymbol{E}\sim N(0,1)$ is introduced in for randomness, which can be defined as $\boldsymbol{C}=\boldsymbol{C}+\gamma\boldsymbol{E}$, where $\gamma=\sqrt{\frac{\text{Tr}(\boldsymbol{CC})^{\text{H}}}{\text{Tr}(\boldsymbol{EE})^{\text{H}}\times\tau_{\text{SNR}}}}$ is the magnitude of the added white noise and τ_{SNR} is the signal-to-noise-ratio which is set to 200 in the current experiment.

Main results. Figure 6(a) depicts the $\mathcal{N}_\phi$-t curve obtained by consecutive split-windows. The detection and location processes are interpreted as follows. During t_s = 1–5000, the $\mathcal{N}_\phi$-t curve remains almost constant, which means the network is generating normal data. From $t_s = 5001$, the curve exhibits an abrupt decline, which suggests the occurrence of anomaly in the network. The MP law of $t_s = 5001$ is shown in Figure 6(b) with one spike spotted outside of the M-P law. The sampling time T in this synthetic scenario is 1000, and the number of involved rows N is 50. The sampling time is much larger than the number of rows, where $c = N/T = 0.050$, $\lambda_- = 12.0560$, and $\lambda_+ = 29.9392$. The anomaly KPIs are identified in rows 49 and 50 as displayed in Figure 6(c).

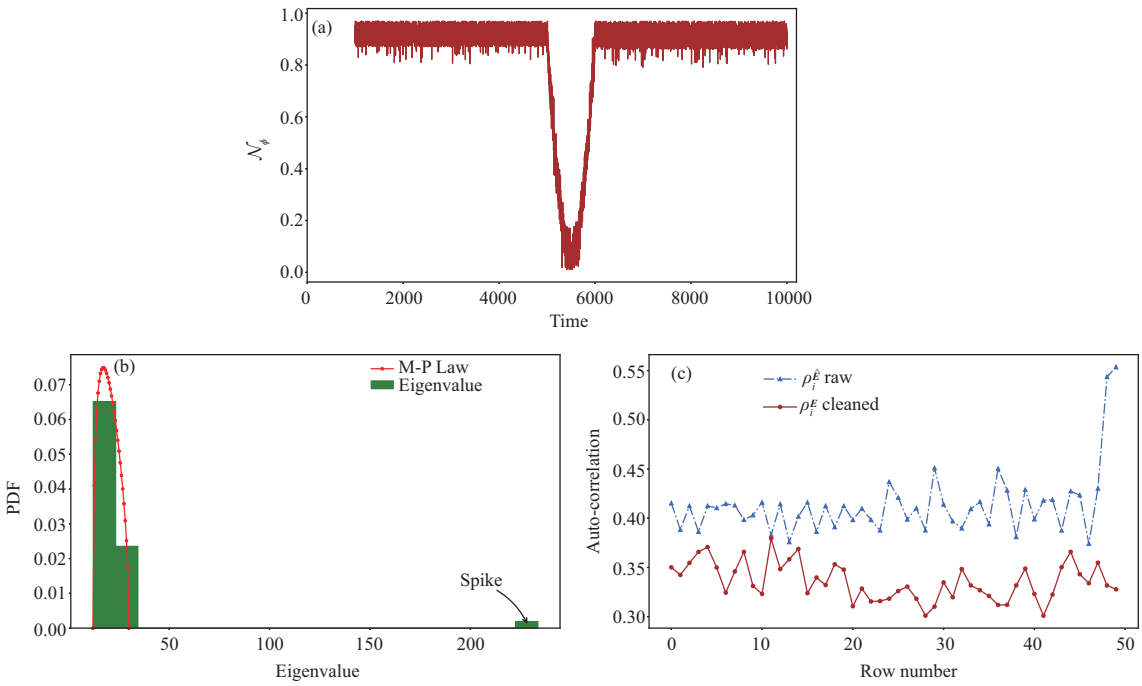

Figure 6 (Color online) Main simulation results of RKI with synthetic data. (a) Anomaly KPI detection with synthetic data. The $\mathcal{N}_\phi$-t curve shows an abrupt decline at $t_s = 5001$. (b) M-P law of anomaly synthetic data. A spike is spotted outside the M-P law. (c) Main correlational comparison between the cleaned and the raw correlation matrices of synthetic data with anomaly KPI (rows 49 and 50) identified.

5.3 Discussion

Computation efficiency. From Table 3, we observe that with the increase of the data size, the TDRs of LSTM, CNN and Transformer grow higher than the proposed RKI method. However, at the same time their ACTs are also much bigger than that of RKI, that is due to the much higher computing consumption in training and labeling. RKI can achieve at least 8.73 times lower ACT than the neural methods, thanks to its non-parametric characteristic. Datasets of different sizes, 1 (5×216), 2 (5×432), 3 (5×648), 4 (5×1296) in ascending order, are utilized for computational efficiency analysis. Specifically, $\text{LSTM}_{(2\ \text{layers})}$, $\text{CNN}_{(3\times3)}$ and $\text{Transformer}_{(6)}$ are selected for the evaluation of computational efficiency. From Figure 7(a) we can draw the conclusion that with the same dataset our proposed RMT-based RKI method can achieve more desirable TDR while reserving a great advantage with ACT.

Real time anomaly diagnosis. Our proposed RKI method is applicable to various streaming data containing verified network anomalies for a dynamic KPI selection. It can reveal the anomaly correlated KPIs in the anomaly network behaviors with real time characteristics. Figure 7(b) presents the auto-correlation of different KPIs versus time (24 h). The RKI method is able to identify the pertaining anomaly KPIs from the abnormal network behaviors by observing the change of $\Delta\rho_{\text{ac}}$. Any KPI_i with $\Delta\rho_{\text{ac}i} \geqslant \overline{\Delta\rho_{\text{ac}}}$ can be considered an anomaly KPI. In this practical scenario, $\overline{\Delta\rho_{\text{ac}}}$ ranges from 0.05 to 0.25. From 9:00 a.m. to 12:00 p.m. (noon), and 18:00 to 21:00, SMS out and Internet traffic are the identified anomaly KPIs that caused the network status fluctuation, thus the network anomalies can be diagnosed. The results demonstrate that the auto-correlation of each KPI is sensitive to time, and different KPIs may have varying impact on the network status.

The results are congruent with the empirical observation, as the data are collected from a convention center where before and after the meeting the network status fluctuates intensively in a short period of time when massive users access the network with frequent service handover, which constitutes anomaly behaviors. Thereby the proposed RKI approach offers a criterion to dynamically identify the anomaly KPIs.

5.4 Potential applications in future 6G wireless networks

The upcoming 6G will demand much higher data rates (up to 1 Tb/s), and extremely low end-to-end latency (< 1 ms), extremely high end-to-end reliability (99.99999%) [46]. The proposed non-parametric

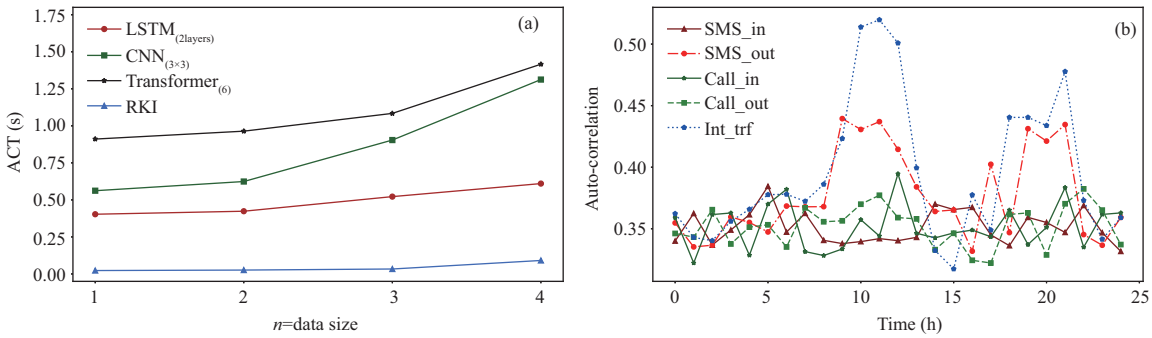

Figure 7 (Color online) Discussions of computation efficiency and real time anomaly diagnosis. (a) ACT comparison; (b) dynamic KPI selection.

foundation of the RKI method makes it robust to multiple forms of highly-correlated data, while being sensitive to correlation variations with desirable diagnosis accuracy and better latency control. The RKI method contributes to the development of non-parametric real-time anomaly detection and diagnosis for the 6G networks.

Energy efficiency of anomaly diagnosis in 6G. As a result of the increasing deployment of artificial intelligence (AI), Internet-of-Things (IoT) architectures and new intelligence scenarios in 6G, large volumes of multidimensional and highly correlated data are accumulating massively [4]. And the transmitting and mining requirements are becoming even more challenging. So it is an urgent appeal for energy efficient anomaly detection and diagnostic methods [47]. Our proposed RKI method is such a non-parametric anomaly diagnostic approach that is computationally much less costly than the popular neural methods. Moreover, a large-scale synthetic data has proved its competence to uncover anomalies in large dimensional data.

Latency control of anomaly diagnosis in 6G. 6G is expected to go beyond mobile Internet and support ubiquitous AI services from the core to the end of the network. AI-enabled intelligent architecture of 6G networks aims at realizing knowledge discovery, smart resource management, automatic network adjustment and intelligent service provision [48, 49]. The emerging new scenarios in 6G have led to a much higher latency control requirement, where highly correlated random and anomalous data traffic is commonly hard to diagnose and results to higher time consumption. Therefore, real-time anomaly detection and diagnosis methodologies are required when mining highly correlated information to ensure the superior low-latency requirement in 6G network performances [50]. Results of the ACT comparison have proved the superior time efficiency performance of the RKI method than neural methods with close TDR.

6 Conclusion

In this paper, we have presented a novel RKI method for anomaly diagnosis, which applies random matrix theory to identify the anomaly network key factors that affect the network status. It estimates the time-varying trend of anomalous and normal network data traffic with a moving average filter, and adopts the M-P law in RMT to uncover the structural properties of normal and anomaly KPI matrices. By examining the correlations of the structural characteristics of the data, the anomaly KPIs can be successfully diagnosed. Furthermore, we have advanced a true diagnostic rate as the evaluation metric to assess the performance of our proposed RKI method. Case studies on both real wireless networks and synthetic data show that the RKI method can efficiently identify the anomaly KPIs with a substantially higher true diagnostic rate and smaller false alarming rate than the traditional methods. Our study contributes to the endeavors of improving anomaly KPI identification with RMT in anomaly diagnosis for the design of next generation wireless networks.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61941105, U21A20449), Beijing Natural Science Foundation (Grant No. L212003), and the 111 Project of China (Grant No. B16006).

References

1 Index C V N. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022 White Paper. Cisco: San Jose, 2019

2 Aliu O G, Imran A, Imran M A, et al. A survey of self organisation in future cellular networks. IEEE Commun Surv Tutorials, 2013, 15: 336–361
3 Strinati E C, Barbarossa S, Gonzalez-Jimenez J L, et al. 6G: the next frontier: from holographic messaging to artificial intelligence using subterahertz and visible light communication. IEEE Veh Technol Mag, 2019, 14: 42–50
4 You X H, Wang C-X, Huang J, et al. Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts. Sci China Inf Sci, 2021, 64: 110301
5 Fernandes J G, Rodrigues J J P C, Carvalho L F, et al. A comprehensive survey on network anomaly detection. Telecommun Syst, 2019, 70: 447–489
6 Callado A, Kamienski C, Szabo G, et al. A survey on Internet traffic identification. IEEE Commun Surv Tut, 2009, 11: 37–52
7 Sui T, Tao X, Xia S, et al. A real-time hidden anomaly detection of correlated data in wireless networks. IEEE Access, 2020, 8: 60990–60999
8 Munoz P, Barco R, Serrano I, et al. Correlation-based time-series analysis for cell degradation detection in SON. IEEE Commun Lett, 2016, 20: 396–399
9 Palacios D, de-la-Bandera I, Gomez-Andrades A, et al. Automatic feature selection technique for next generation self-organizing networks. IEEE Commun Lett, 2018, 22: 1272–1275
10 Marnerides A K, Schaeffer-Filho A, Mauthe A. Traffic anomaly diagnosis in Internet backbone networks: a survey. Comput Networks, 2014, 73: 224–243
11 Jiang D, Yuan Z, Zhang P, et al. A traffic anomaly detection approach in communication networks for applications of multimedia medical devices. Multimed Tools Appl, 2016, 75: 14281–14305
12 Ahmed M, Mahmood A N, Hu J. A survey of network anomaly detection techniques. J Network Comput Appl, 2016, 60: 19–31
13 Jurdak R, Wang X R, Obst O, et al. Wireless sensor network anomalies: diagnosis and detection strategies. In: Intelligence-Based Systems Engineering. Berlin: Springer, 2011. 309–325
14 He Z, Xu X, Deng S. Discovering cluster-based local outliers. Pattern Recogn Lett, 2003, 24: 1641–1650
15 Brint A, Genovese A, Piccolo C, et al. Reducing data requirements when selecting key performance indicators for supply chain management: the case of a multinational automotive component manufacturer. Int J Production Economics, 2021, 233: 107967
16 Münz G, Li S, Carle G. Traffic anomaly detection using k-means clustering. In: Proceedings of GI/ITG Workshop MMBnet, 2007. 13–14
17 Khreich W, Khosravifar B, Hamou-Lhadj A, et al. An anomaly detection system based on variable N-gram features and one-class SVM. Inf Software Tech, 2017, 91: 186–197
18 Zhang C, Song D, Chen Y, et al. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2019. 33: 1409–1416
19 Kwon D, Natarajan K, Suh S C, et al. An empirical study on network anomaly detection using convolutional neural networks. In: Proceedings of IEEE 38th International Conference on Distributed Computing Systems (ICDCS), 2018. 1595–1598
20 Kibria M G, Nguyen K, Villardi G P, et al. Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks. IEEE Access, 2018, 6: 32328–32338
21 He X, Ai Q, Qiu R C, et al. A big data architecture design for smart grids based on random matrix theory. IEEE Trans Smart Grid, 2015. doi: 10.1109/TSG.2015.2445828
22 Qiu R C, Antonik P. Smart Grid Using Big Data Analytics: A Random Matrix Theory Approach. Hoboken: John Wiley & Sons, 2017
23 Xu F, Li Y, Wang H, et al. Understanding mobile traffic patterns of large scale cellular towers in urban environment. IEEE ACM Trans Networking, 2016, 25: 1147–1161
24 Yang J, Ma Z, Dong C, et al. An empirical investigation into CDMA network traffic classification based on feature selection. In: Proceedings of the 15th International Symposium on Wireless Personal Multimedia Communications, 2012. 448–452
25 Chen Z, Chen D, Zhang X, et al. Learning graph structures with transformer for multivariate time-series anomaly detection in IoT. IEEE Internet Things J, 2022, 9: 9179–9189
26 Tao T. Topics in Random Matrix Theory. Providence: American Mathematical Society, 2012
27 Brody T A, Flores J, French J B, et al. Random-matrix physics: spectrum and strength fluctuations. Rev Mod Phys, 1981, 53: 385–479
28 Guhr T, Müller-Groeling A, Weidenmüller H A. Random-matrix theories in quantum physics: common concepts. Phys Rep, 1998, 299: 189–425
29 Laloux L, Cizeau P, Potters M, et al. Random matrix theory and financial correlations. Int J Theor Appl Finan, 2000, 03: 391–397
30 Yeo J, Papanicolaou G. Random matrix approach to estimation of high-dimensional factor models. 2016. ArXiv:1611.05571
31 Chen H, Tao X F, Li N, et al. A data analysis of political polarization using random matrix theory. Sci China Inf Sci, 2020, 63: 129303
32 Luo F, Zhong J, Yang Y, et al. Application of random matrix theory to biological networks. Phys Lett A, 2006, 357: 420–423
33 Chen H, Tao X, Li N, et al. Physical layer data analysis for abnormal user detecting: a random matrix theory perspective. IEEE Access, 2019, 7: 169508
34 Couillet R, Debbah M. Random Matrix Methods for Wireless Communications. Cambridge: Cambridge University Press, 2011
35 Qiu R C, Hu Z, Li H, et al. Cognitive Radio Communication and Networking: Principles and PRACTICE. Hoboken: John Wiley & Sons, 2012
36 He Y, Yu F R, Zhao N, et al. Big data analytics in mobile cellular networks. IEEE Access, 2016, 4: 1985–1996
37 Tulino A M, Verdú S. Random matrix theory and wireless communications. FNT Commun Inf Theor, 2004, 1: 1–182
38 Han B, Luo L, Sheng G, et al. Framework of random matrix theory for power system data mining in a non-Gaussian environment. IEEE Access, 2016, 4: 9969–9977
39 Shi X, Qiu R, He X, et al. Early anomaly detection and localisation in distribution network: a data-driven approach. IET Gener Transm Distrib, 2020, 14: 3814–3825
40 Shi X, Qiu R, Ling Z, et al. Spatio-temporal correlation analysis of online monitoring data for anomaly detection and location in distribution networks. IEEE Trans Smart Grid, 2019, 11: 995–1006
41 Bun J, Bouchaud J P, Potters M. Cleaning large correlation matrices: tools from random matrix theory. Phys Rep, 2017, 666: 1–109

42 Benesty J, Chen J, Huang Y. On the importance of the Pearson correlation coefficient in noise reduction. IEEE Trans Audio Speech Lang Process, 2008, 16: 757–765
43 Bai Z, Silverstein J W. Spectral Analysis of Large Dimensional Random Matrices. 2nd ed. New York: Springer, 2010
44 Edelman A, Rao N R. Random matrix theory. Acta Numerica, 2005, 14: 233–297
45 Parwez M S, Rawat D B, Garuba M. Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network. IEEE Trans Ind Inf, 2017, 13: 2058–2065
46 Alwis C D, Kalla A, Pham Q V, et al. Survey on 6G frontiers: trends, applications, requirements, technologies and future research. IEEE Open J Commun Soc, 2021, 2: 836–886
47 Letaief K B, Chen W, Shi Y, et al. The roadmap to 6G: AI empowered wireless networks. IEEE Commun Mag, 2019, 57: 84–90
48 Yang H, Alphones A, Xiong Z, et al. Artificial-intelligence-enabled intelligent 6G networks. IEEE Network, 2020, 34: 272–280
49 Xiao Y, Shi G, Li Y, et al. Toward self-learning edge intelligence in 6G. IEEE Commun Mag, 2020, 58: 34–40
50 Han G, Tu J, Liu L, et al. Anomaly detection based on multidimensional data processing for protecting vital devices in 6G-enabled massive IIoT. IEEE Internet Things J, 2021, 8: 5219–5229