Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3644592, 13 pages
https://doi.org/10.1155/2022/3644592

Research Article

Random Matrix Theory-Based ROI Identification for Wireless Networks

Tengfei Sui, Xiaofeng Tao, and Jin Xu

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

Correspondence should be addressed to Xiaofeng Tao; taoxf@bupt.edu.cn

Received 19 January 2022; Accepted 19 May 2022; Published 21 June 2022

Academic Editor: Bithas Petros

The identification of region of interests (ROIs) in wireless networks holds the potential to resolve the challenging problems of resource allocation and network traffic prediction for large scale traffic data generated by mobile applications. The rationale is that ROIs are capable of gathering single regions that share similar network characteristics, which promotes better network traffic prediction performance. Previous studies show that spatiotemporal information in network traffic data, such as user behaviors and network status, is nontrivial to ROI identification. However, the modeling between these clues regarding spatiotemporal information is not yet fully explored. To this end, we propose a random matrix theory-based ROI identification (RRI) approach. By observing the intensification or diminution of network characteristic differences, i.e., divergence, between adjacent single regions, the ROIs can be identified. Firstly, we leverage the spatiotemporal information of area network traffic data with a spike model which can be described as a zero mean random matrix with a deterministic perturbation matrix. Then, we put forward an average divergence capacity model for ROI identification by estimating the divergent degree of adjacent regions. Case studies on three real-world network traffic datasets demonstrate the effectiveness of our proposed RRI method. The ROI identification greatly improves the network traffic prediction performance, yielding a decrease of root mean square error and mean absolute error by 36.87% and 52.26%, respectively.

1. Introduction

In the upcoming 2030s+, wireless network services and scenarios will become more diversified, and user needs will be more personalized than ever [1]. Meanwhile, data generated by the use of extremely heterogeneous networks, diverse communication scenarios, and large numbers of devices have undergone an exponential expansion to an unprecedented scale [2]. In particular, due to the increasingly diversified and complex networks, we have ushered in an era of big data with 77.5 exabytes of wireless network traffic data produced per month by 2022 [3]. 6G networks are expected to enable on-demand services for better user satisfactions [4].

A more accurate network traffic prediction of diverse region of interests (ROIs) with similar network traffic characteristics can help network operators understand the diversified network status, optimize the resource allocation, improve users' quality of experience (QoE), and reduce the capital expenditure (CAPEX) and operating expenditure (OPEX) [5–7]. Yet the pervasive and exponentially increasing multidimensional and highly correlated data impose imminent challenges on area network traffic characteristic modeling and prediction in diverse regions [8, 9].

The area network traffic data have become increasingly correlated in time and space [10, 11]. Big data modeling and analysis of the multidimensional and highly correlated wireless network data plays a pivotal role in predicting the network traffic and understanding the network characteristics of ROIs [12–14]. Data-driven network traffic understanding and prediction have attracted great attention and produced fruitful results [15–17]. For example, a data-driven framework for network behavior analysis in cellular networks for Industry 4.0 is proposed in [18]. Human

mobility patterns using spatiotemporal correlated urban big data are provided for vehicular social networks in [15, 19].

Great progress in network traffic prediction has been achieved by neural network-based methods. For example, Long Short-Term Memory (LSTM) [17], Gated Recurrent Units (GRU) [20], and Stacked Autoencoders (SAEs) [9] have reported better performance in predicting time series data than statistically based methods. While these methods study traffic time series for each individual location, recent studies further utilize spatial information. An attention-based neural network is proposed in [6] for traffic prediction, and a deep learning method for wireless network traffic prediction is put forward in [13] with temporal and spatial characteristics of wireless network traffic data modeled for prediction.

However, these neural network-based researches mainly focus on prediction in isolated single regions, which overlooks the spatiotemporal information of adjacent regions, which thus may lead to inaccurate prediction results.

Intuitively, with more data obtained from adjacent regions with similar network traffic characteristics, a higher prediction accuracy can be achieved. The main challenges to this hypothesis are as follows: (1) how to shape the network traffic characteristics with a comprehensive data model, (2) how to evaluate the network traffic characteristic differences of adjacent regions, and (3) how to aggregate the adjacent regions with similar network traffic characteristics as an identified ROI.

Network traffic data can be considered time series for prediction [21, 22]. AutoRegressive Integrated Moving Average (ARIMA) [23] and Support Vector Regression (SVR) [24] are the representative approaches to time series modeling. The ARIMA model tends to focus on the mean value of the past data regardless of the nonlinear variations underlying the traffic flow [25]. The limitation of SVR lies in the difficulty to determine the key parameters [25]. Notably, the excessive dependence on historical data with spatial information ignored, in particular that of adjacent regions, may lead to unsatisfying prediction performance [26].

To this end, we propose a spike model to describe the spatiotemporal information of adjacent regions with random matrix theory (RMT) spectral verifications. By revealing the differences of data structure among multidimensional datasets with the spectral analysis, RMT is able to analyze the divergent degree of different datasets [27–29]. This paper is an extension of our previous work which utilizes RMT for anomaly detection in wireless networks [30]. In [30], we apply RMT to distinguish anomalous data from normal data by observing the eigenvalue distribution, but a deeper investigation of the spectral distribution is lacking. In this paper, we propose a data model and derive its spectral distribution for area network traffic characterization and a new capacity model for divergence degree evaluation in ROI identification. The correctly identified ROI can promote better network traffic prediction performance and higher resource allocation efficiency in the upcoming 6G networks. To summarize, the main contributions of our work are as follows:

(i) We propose a novel method of RMT-based ROI identification (RRI), to identify the ROIs by evaluating the network traffic differences of adjacent regions modeled by a spike model

(ii) The spike model is a zero mean random matrix with a deterministic perturbation matrix utilized for modeling the network traffic characteristics. The RMT spectral analysis is employed to theoretically verify the model, showing that the empirical spectral distribution of the spike model confirms the raw eigenvalue distribution

(iii) An average divergence capacity model is proposed to identify the ROIs by evaluating the divergent degree of adjacent single regions modeled by the spike model. We aggregate the adjacent single regions with shrinking divergence as an identified ROI

(iv) Numerical results show that the proposed RRI approach can identify ROIs with ground truth verifications. Moreover, with the aid of RRI, the performance of prediction in ROIs can be improved with a decrease of 36.87% root mean square error and 52.26% mean absolute error

The rest of the paper is organized as follows. Section 2 presents the data description and some preliminary data analysis. The background knowledge about the RMT spectral analysis and RMT-based theoretical verification for the spike modeling method is laid out in Section 3. In addition, a real-world area network traffic dataset is employed to validate the effectiveness of the proposed model. In Section 4, an average divergence capacity model for evaluating the divergent degree of adjacent regions is presented for the RRI method. Case studies of ROI identification and network traffic predictions are carried out in Section 5. Section 6 concludes the paper.

2. Data Description and Preliminary Data Analysis

As the spatiotemporal correlated data accumulate to an enormous scale, the network traffic differences of diverse regions are no longer static, and thus, the network traffic prediction for isolated regions is not applicable to the fulfillment of on-demand network in the era of big data [26].

ROI identification in wireless network can contribute to a more accurate network traffic prediction. An appropriate data model that can describe the network characteristics with spatiotemporal information preserved is a good start to begin with. In this aspect, a universal data model for network traffic characteristic difference evaluation can greatly facilitate the ROI identification and further improve the prediction performance. Table 1 summarizes the notations used in this paper.

2.1. Dataset Description. In order to model the wireless area network traffic for the RMT analysis, we first present a description of the dataset. It is a real-world spatiotemporal correlated network traffic dataset that is comprised of computation over the Call Detail Records (CDRs) consisting of

Table 1: Summary of notations.

Notations	Meaning
$Y, y, y_{i,j}$	A matrix, a vector, an entry of a matrix
N, T, c	The numbers of rows and columns, $c = N/T$
S	Covariance matrix of Y
$\mathbb{C}^{N\times T}$	Complex space
$\lambda_1, \cdots, \lambda_N$	Eigenvalues of S
$[\lambda_1^-, \lambda_1^+], [\lambda_2^-, \lambda_2^+]$	Support of the bulk and spike
σ	Variance
r, r_1, r_2	Different regions
I^{r_1,r_2}	Average divergence capacity of r^1 and r^2
$vi, \widehat{vi}$	Observed and predicted traffic volumes
F^{S_N}	Empirical spectral distribution
$m_f(\cdot)$	Stieltjes transform
$G(\cdot)$	Inversion of Stieltjes transform

the network traffic data collected from a real LTE network of Telecom Italia at Milan, Italy. This public dataset was officially provided to the Big Data Challenge 2014 competition [31]. It was collected from 3,450 base stations (BSs), which logged the network traffic of each base station over two months, from November to December, 2013. The dataset includes SMS activity, call activity, and Internet traffic activity, which can be considered key performance indicators (KPIs) of the region characteristics. To facilitate the data analysis, the Milan region is divided into 100×100 grids named as Milan Grids, with all the BSs mapped into individual grids, or single regions. When there are several BSs in a single region, all the traffic loads are aggregated into one traffic load [32]. Although these data were recorded nearly a decade ago, due to the fact that they truly included the characteristics of spatiotemporal information in real geographic scenarios, they have been widely utilized for network traffic analyses in recent years [17, 30, 33, 34].

For expository purposes, we select an area with 16 regions (grids) as depicted in Figure 1(a). The area includes three typical social function regions, which are the Convention Center (Grid 5848), Shopping Center (Grid 5849), and Central Park (Grids 5748 and 5749). Figure 1(b) depicts the statistic results of the accumulated spatially correlated network traffic data of the 16 single regions within 24 hours. The network traffic volume of adjacent regions shows great similarities, but notably, the network traffic volume of the regions adjacent to the Shopping Center and Convention Center is much higher than that of the regions adjacent to Central Park. The observation is consistent with the ground truth that Central Park consumes much less network resource than the Convention Center and the Shopping Center [33, 34].

2.2. Preliminary Data Analysis. For a preliminary analysis of the characteristics of the three different regions, the statistical results are presented with each network traffic dataset grouped into a matrix, whose rows represent the individual traffic of specific regions and the columns indicate the sampling time. Assume the number of KPIs is N and the total sampling time is T. Without loss of generality, for different KPI i at the sampling time j, we model the raw KPI volume as $y_{i,j}$. All the sampled KPI i can be treated as a vector $y_i = (y_{i,1}, \cdots, y_{i,T}) \in \mathbb{C}^{1\times T}$.

Figure 2 describes the network traffic data of three adjacent different single regions within a duration of two days, which exhibits strong diverse time series characteristics. Whereas the data consumed by Central Park peak at around 10:00 a.m. and the Convention Center reaches its maximal data consumption at around 20:00 p.m., the Shopping Center displays a plateau of data consumption during the daytime. With respect to the different network traffic characteristics of diverse regions, we can draw the conclusion that the data are also spatially correlated. If we can aggregate the adjacent regions with similar network traffic characteristics, the ROIs can be identified accordingly. Therefore, before we utilize the network traffic characteristic differences for ROI identification, it is necessary to model the spatiotemporal information of individual regions.

3. Network Traffic Data Modeling

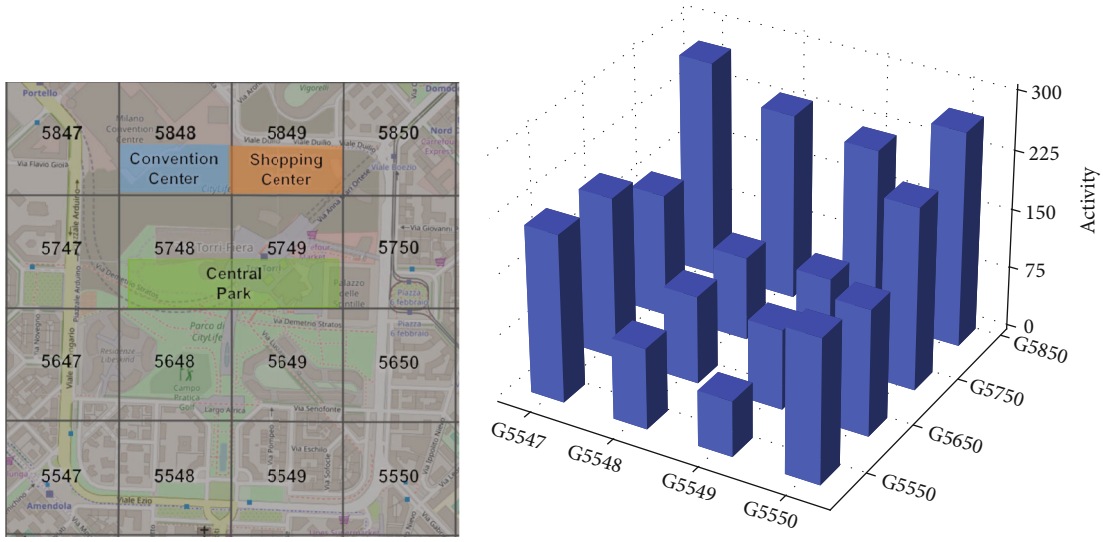
RMT has been widely applied to the analysis of highly correlated big wireless network data that contain a number of random variables [27, 33, 35]. Most researches pertaining to RMT utilize it as a benchmark for anomaly detection by simply observing the eigenvalue distribution, yet lack a mathematical intrinsic modeling investigation [33, 36, 37]. In [38], RMT is employed to analyze the time series data for anomaly detection, which extends the RMT applications to a non-Gaussian distribution scenario. In terms of a thorough analysis of the network traffic data differences, pioneering works in [6, 30, 33] have proposed to apply the RMT spectral analysis to anomaly detection. In this section, we extend the application of RMT to the modeling of network traffic characterization.

3.1. Data Modeling. Wireless network traffic data can be decomposed into regular components and residual components [21, 33]. But we present a more intuitive data model hypothesis of the network traffic volume with the raw data $y_{i,j}$ decomposed into two parts as shown in

$$y_{i,j} = x_{i,j} + \sigma e_{i,j}, \tag{1}$$

where $x_{i,j}$ represents the deterministic network traffic pattern in one region, $e_{i,j}$ is an independent identically distributed (i.i.d.) random variable with zero mean and unit variance, and σ is the variance. Thereby, the raw data $y_{i,j}$ can be considered a random variable with nonzero mean by the probability, and the sampling matrix of the network traffic dataset from a specific region r can be considered a random matrix as formulated in

$$Y_{N,T}^r = X_{N,T}^r + \sigma^r E_{N,T}, \tag{2}$$



(a) The selected 16 regions (Grids)

(b) Statistical results of the selected regions

Figure 1: Dataset description.

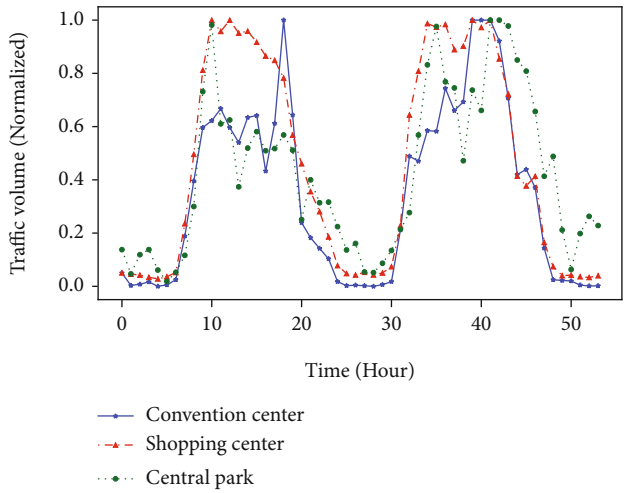


Figure 2: Periodic network traffic volume of three different regions.

where N stands for the number of KPIs, T denotes the total sampling times, the matrix $E = \{e_{i,j}\}_{N\times T}$ is a non-Hermitian random matrix with i.i.d. zero-mean Gaussian distribution entries $e_{i,j}$, and $X^r_{N,T}$ is the deterministic matrix of a specific ROI with all single valued entries.

3.2. Theoretical Verification for the Data Model. Since the area traffic dataset has been constructed as (2), which is a multidimensional and highly correlated random matrix, RMT can be applied as a mathematical tool to theoretically verify the model with spectral analysis. The RMT spectral analysis can reveal the intrinsic data structure information from the perspective of eigenvalue distribution. Therefore, we focus on investigating the eigenvalue properties of the data model in this section.

In the light of the random matrix $Y^r_{N,T} \in \mathbb{C}^{N\times T}$ in (2), its covariance matrix can be derived as

$$S^r_N = \frac{1}{T} Y^r_{N,T} Y^{r\mathrm{T}}_{N,T}, \tag{3}$$

where T stands for the matrix transpose. The matrix $Y^r_{N,T}$ of a specific ROI can be formulated with the data model proposed in (2). Then, the covariance matrix of the raw data matrix in (2) can be denoted as

$$S^r_N = \frac{1}{T}\left(X^r_{N,T} + \sigma^r E_{N,T}\right)\left(X^r_{N,T} + \sigma^r E_{N,T}\right)^{\mathrm{T}}. \tag{4}$$

Having obtained the covariance matrix S^r_N, the asymptotic spectrum of the data model can be derived with the empirical spectral distribution (ESD) given in Definition 1 for mathematical verification, which is an important metric to describe the eigenvalue distribution of a matrix.

Definition 1 (empirical spectral distribution [39]). Consider an$N \times N$ Hermitian matrix S_N, the ESD F^{S_N} of the matrix S_N is defined as

$$F^{S_N}(\lambda) = \frac{1}{N}\sum_{j=1}^{N} 1_{\{\lambda,\lambda_j\le\lambda\}}(\lambda), \tag{5}$$

where $1_S(\cdot)$ is an indicator function over a set S and $\{\lambda_1, \cdots, \lambda_N\}$ denotes the eigenvalues of S_N.

By the definition of ESD $F^{S_N}(\lambda)$, the average eigenvalues that are smaller than a particular variable λ constitute a cumulative density function, based on which the eigenvalue distribution of S^r_N can be derived. As illustrated in Figure 3, the ESD of S^r_N shows two components, the bulk and the

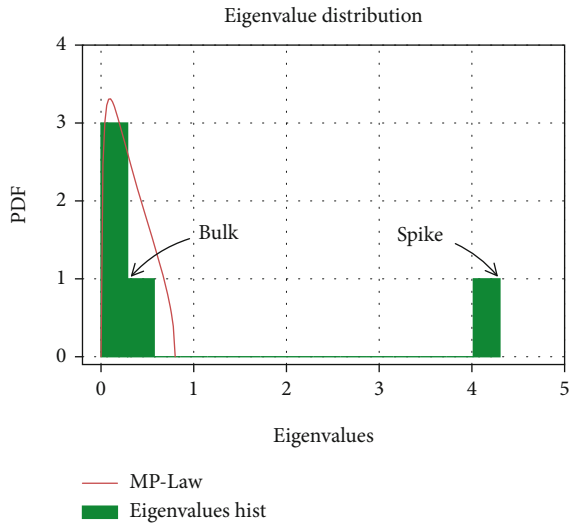


FIGURE 3: Eigenvalue distribution of S_N^r with the bulk denoting the noise and the spike denoting the signal.

spike. The bulk mainly arises from the random noise or fluctuations of the stochastic part $E_{N,T}$ in (4), and the spike represents the unusual network traffic volumes or anomalies in the deterministic part $X_{N,T}$ in (4). This kind of data model can be analogically and mathematically considered a spike model in RMT [40].

Generally speaking, the ESD of a random matrix is difficult to be deduced, especially after basic elementary mathematical calculations. So a diversion is necessary before the derivation of ESD. Stieltjes transform is an elementary but indispensable transformation in RMT given in Definition 2.

Definition 2 (Stieltjes transform [39]). Consider F as a spectral distribution of a given matrix; its Stieltjes transform is defined as

$$m_f(z) \triangleq \int_{-\infty}^{\infty} \frac{1}{\lambda - z} dF(\lambda), \lambda \in \mathbb{C}, z \in \mathbb{C}^+, \tag{6}$$

where $z \in \mathbb{C} \equiv \{z \in \mathbb{C} : \Im[z] > 0\}$ and $\Im$ denotes the imaginary part.

A correspondence exists between the spectral distribution and the Stieltjes transform, which can be described as the convergence characteristics of finite measures [40, 41]. For any distribution function G, the inversion of Stieltjes transform can be defined by

$$G(\lambda) = \lim_{w \longrightarrow 0^+} \frac{1}{\pi} \Im\left[m_f(\lambda + jw)\right], \tag{7}$$

where $j = \sqrt{-1}$ is the imaginary unit.

Although Stieltjes transform is a way to deduce the ESD of a given matrix, in practical scenarios, only some simple structured random matrices can be derived with such an explicit expression. For example, the classic Marchenko-Pastur Law (M-P Law) is a close-form ESD of one particular type of random matrix [40]. The M-P Law offers a deeper insight into the correspondence between ESD and its Stieltjes transform, which has become the foundation to derive the ESD of complex matrices, as illustrated by the red line in Figure 3. The M-P Law has been commonly applied as a benchmark for anomaly detection in wireless networks [28, 33]. The asymptotic theoretical spectrum of the spike model can be obtained with Theorem 3.

Theorem 3 (spike model [40]). *Given a matrix S_N^r defined as in (4) and a non-Hermitian random matrix $E_{N,T}$ in (2) with i.i.d. zero-mean and unit variance Gaussian distribution entries, such that the ESD of $(1/T)X_{N,T}X_{N,T}^T$ converges to the function X with $sup_N\|(1/T)X_{N,T}X_{N,T}^T\| < \infty$ and the Stieltjes transform $m_X(z)$. Denote $c_N = N/T$ and assume $c_N \longrightarrow c$, positive and finite; then, the ESD of S_N^r converges almost surely to a limit distribution G with the Stieltjes transformation $m_G(z)$ derivable from*

$$\frac{m}{1 + \sigma^{r2} cm} = m_X\left(z\left(1 + \sigma^{r2} cm\right)^2 - \sigma^{r2}(1 - c)\left(1 + \sigma^{r2} cm\right)\right). \tag{8}$$

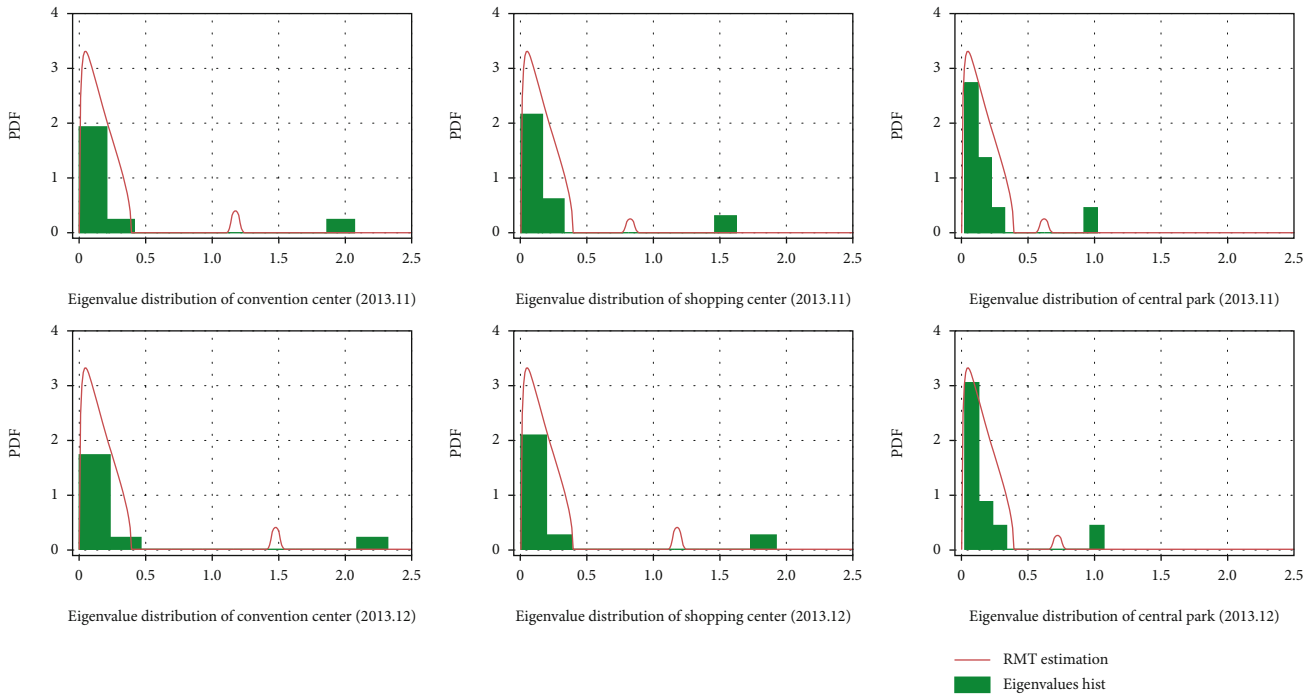
The solution of m satisfies the conditions of the Stieltjes transformation is $m_G(z)$. The theorem presents the Stieltjes transformation of S_N^r given in (4) with an implicit equation. Notably, the deterministic matrix of X in Theorem 3 can be generalized to any given matrix, and the rank of the matrix remains uncertain, which means the spike model can be generalized to a variety of data analysis scenarios.

In a practical scenario such as that shown in Figure 3, the bulk of the eigenvalue distribution mainly arises from the random noise or fluctuations of the stochastic component $E_{N,T}$ in (4), and the spike is usually originated from the deterministic component of $X_{N,T}^r$ in (4). With only one spike spotted in Figure 3, we can deduce that there is only one nonzero eigenvalue in the deterministic matrix $X_{N,T}^r$, and its rank is 1. It is due to the fact that the network traffic exhibits identical network behavior characteristics in a same ROI. Once we can obtain the Stieltjes transform of the deterministic matrix from (2) and the variance σ of the random component, the ESD of the covariance matrix S_N^r can be derived.

Firstly, let us compute the Stieltjes transform of the deterministic matrix in (2). Since the rank of $X_{N,T}^r$ is 1, the only nonzero eigenvalue of the covariance matrix of $(1/T)$ $X_{N,T}^r X_{N,T}^{r\mathrm{T}}$ can be denoted as $\widehat{\lambda}_X = \mathrm{Tr}((1/T)X_{N,T}^r X_{N,T}^{r\mathrm{T}})$ with the probability $p = (1/N)$, while the other eigenvalues $\lambda = 0$ with the probability $1 - p = (N - 1)/N$. Thereby, the Stieltjes transform $m_X(z)$ of $(1/T)X_{N,T}^r X_{N,T}^{r\mathrm{T}}$ can be derived as

$$m_X(z) = \int \frac{1}{\lambda - z} dF_X(\lambda) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\lambda_i - z} = \frac{1}{N}\left[\frac{N-1}{-z} + \frac{1}{\lambda_X - z}\right]. \tag{9}$$

By means of the numerical operation of substituting (9) into (8), the solution of m satisfies the conditions of the



(a) Eigenvalues of Convention Center (b) Eigenvalues of Shopping Center (c) Eigenvalues of Central Park

Figure 4: RMT estimations and eigenvalue distributions of three adjacent ROIs (2013.11-12).

Stieltjes transformation constitutes $m_G(z)$. In turn, the ESD of the spike model covariance matrix can be derived by substituting the obtained $m_G(z)$ into the inversion formula for the Stieltjes transform given in (7).

3.3. Numerical Verification for the Data Model. In this subsection, we present validations for the data model by comparing the theoretical ESD with the practical eigenvalue distributions in three adjacent ROIs as depicted in Figure 1(a). The Stieltjes transform calculations are repeated for 50 times with averaged results.

Figure 4 shows the validations of the RMT estimation of three adjacent regions, which are Convention Center, Shopping Center, and Central Park, respectively. The solid red line illustrates the theoretical RMT estimation of the ESD of the spike model covariance matrix, and the green histograms indicate the practical eigenvalue distributions of the raw data source. The theoretical ESD of the RMT estimation of the spike model can also be separated into two components, the bulk and the spike, which theoretically and practically converge to the proposed model.

Figure 4(a) is the eigenvalue distribution of the Convention Center (Grid 5848), with the bulk more centralized. The theoretical RMT estimation corresponds to the empirical network traffic volume of the Convention Center with more perturbations. Figure 4(b) demonstrates the eigenvalue distribution of the Shopping Center (Grid 5849) with a more regular network status routine. Similarly, the deviations between the bulk and the spike grow larger from November to December for both the Convention Center and the Shopping Center, which suggests their ESD difference enlarges with time advancement.

Figure 4(c) demonstrates the eigenvalue distribution of the Central Park (Grids 5748 and 5749) with the most regular network status routine, as the number of people that go to the park remains almost constant. The deviations of the Central Park between the bulk and the spike almost stay static in November and December implying that the ESD difference of the Central Park hardly changes in the two months.

The verifications have proved the convergence of the spike model. The gaps between the theoretical RMT estimation and the empirical eigenvalue distribution are primarily caused by the estimation of $X^r_{N,T}$ and the limitation of the data size.

3.4. Support of the ESD. A step further, we investigate the ESD separation phenomenon of the spike model. Generally speaking, the raw dataset can be influenced by various factors, which results in the separation of ESD to different components. As $T \longrightarrow \infty$, the ESD deviation of the bulk and the spike can be deduced by deriving their support [42], which is denoted as $[\lambda_1^-, \lambda_1^+]$ and $[\lambda_2^-, \lambda_2^+]$ given in (11) and (12), respectively. The ESD separation of the covariance matrix is given in Lemma 4.

Lemma 4 (spike model support [42]). *Considering an eigenvalue $\lambda_X > \sigma^2\sqrt{c}$, then equation (10) holds with probability 1:*

$$\widehat{\lambda}_N \overset{T\longrightarrow\infty}{\longrightarrow} \phi(\lambda_X, c), \tag{10}$$

and $\widehat{\lambda}_N \longrightarrow^{T\longrightarrow\infty} \lambda^2(1+\sqrt{c})^2$, where $\widehat{\lambda}_N$ is the largest eigenvalue of the model if $\lambda \le \sigma^2\sqrt{c}$. Yet if $\lambda_X > \sigma^2\sqrt{c}$, the

corresponding support of the bulk and the spike can be derived as

$$\left[\lambda_1^-, \lambda_1^+\right] = \left[\sigma^2\left(1-\sqrt{c}\right)^2 + \mathcal{O}\left(\frac{1}{T}\right), \sigma^2\left(1+\sqrt{c}\right)^2 + \mathcal{O}\left(\frac{1}{T}\right)\right], \tag{11}$$

$$\left[\lambda_2^-, \lambda_2^+\right] = \left[\phi\left(\lambda_X, c\right) - \mathcal{O}\left(\frac{1}{\sqrt{T}}\right), \phi(\lambda_X, c) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)\right], \tag{12}$$

where $\mathcal{O}(\cdot)$ *represents the approximate value and* $\phi(\lambda_X, c)$ *is defined as*

$$\phi(\lambda_X, c) = \frac{\left(\lambda_X + \sigma^2 c\right)\left(\lambda_X + \sigma^2\right)}{\lambda_X}. \tag{13}$$

From Lemma 4, we can deduce that the support intervals of the bulk given in (11) and (12) are largely dominated by the parameters of σ and c in $\sigma^r E_{N,T}$ in (2). The spectral distribution of the matrix S_N^r in (4) will give rise to a spike with a large enough λ_X. On the other hand, if λ_X is much smaller, the ESD will be a bulk. The deviation between the bulk and the spike is closely correlated to the deterministic matrix X in (2), which can be utilized to evaluate the difference between different datasets.

4. Divergent Region Difference Evaluation for RRI

Since the theoretical ESD of the spike model matches empirical eigenvalue distribution of the raw network traffic matrix, we will identify ROIs by utilizing the spike model to reconstruct the network traffic data. An average divergence capacity model is proposed to mathematically quantify the divergent degree of adjacent regions for RRI.

4.1. Average Divergence Capacity Model for RRI. In order to numerically quantify the divergent degree of adjacent regions with different datasets, the network traffic volume difference of adjacent regions is defined as (14) according to the spike model we proposed in Section 3.

$$Y_{N,T}^{r1,r2} = Y_{N,T}^{r1} - Y_{N,T}^{r2} = \left(X_{N,T}^{r1} + \sigma^{r1} E_{N,T}\right) - \left(X_{N,T}^{r2} + \sigma^{r2} E_{N,T}\right), \tag{14}$$

where $r1$ and $r2$ represent different adjacent regions. Moreover, (14) can be further expressed as

$$Y_{N,T}^{r1,r2} = X_{N,T}^{r1,r2} + \left(\sigma^{r1,r2}\right) E_{N,T}, \tag{15}$$

where $\left(\sigma^{r1,r2}\right)^2 = \left(\sigma^{r1}\right)^2 + \left(\sigma^{r2}\right)^2$, σ^{r1} and σ^{r2} denote variances of adjacent regions and $E_{N,T}$ stands for the randomness that follows the Gaussian distribution with zero mean and unit variance entries. $X_{N,T}^{r1,r2}$ is defined as $X_{N,T}^{r1} - X_{N,T}^{r2}$, which is the deterministic matrix that indicates the network characteristic difference of two adjacent regions.

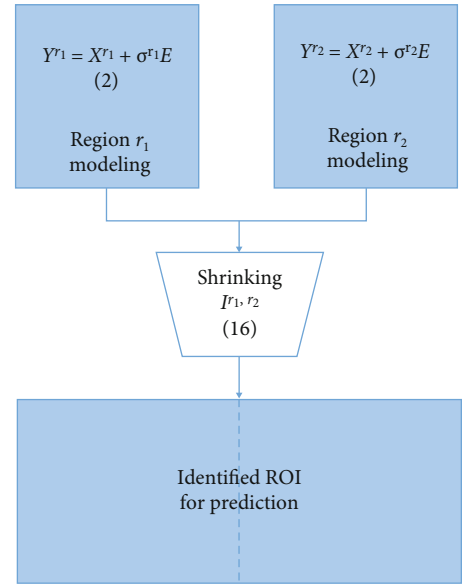


Figure 5: Architecture of the proposed RRI approach.

We present an average divergence capacity model to numerically quantify the different divergent degrees of adjacent regions for ROI identification. Inspired by the definition of the channel capacity, we consider $X_{N,T}^{r1,r2}$ defined in (15) as a signal running through an additive white Gaussian noise channel.Thus, the average divergence capacity model can be analogically defined as

$$I^{r1,r2} = \frac{1}{N}\log_2 \det\left(I_N + \frac{1}{T}\frac{X_{N,T}^{r1,r2} X_{N,T}^{r1,r2^{\mathrm{T}}}}{\left(\sigma^{r1,r2}\right)^2}\right), \tag{16}$$

where I_N is an $N \times N$ identity matrix and T stands for the matrix transpose. The model can quantify the uncertainty of the data with a unit of bits from the perspective of information theory [43], thereby providing a numerical quantification measurement for the ROI identification problem. The evaluation model is a mapping of multidimensional raw support $\mathbb{D}$ to evaluation results $\mathbb{R}^+$, which can be expressed as $F : \mathbb{D} \longrightarrow \mathbb{R}^+$.

4.2. Parameter Estimation. Before we can employ the proposed average divergence capacity model to analyze different datasets, we need to compute the unknown parameter σ in (16). We apply a large dimensional approach (LDA) to σ calculation. The classical LDA assumes that the samples are numerous, i.e., $T \longrightarrow \infty$, so it can accommodate much more diversities in the total samplings. As $T \longrightarrow \infty$, the ratio of the matrix dimensions $c = (N/T) \longrightarrow 0$, and the distribution of the largest eigenvalue of $S \triangleq (1/T) Y_{N,T}^{r1,r2} Y_{N,T}^{r1,r2^{\mathrm{T}}}$ converges almost surely to $\sigma^2 + M$, where $M = \mathrm{Tr}(X_{N,T}^{r1,r2} X_{N,T}^{r1,r2^{\mathrm{T}}})$, and σ is the covariance. Whereas the distribution of the rest eigenvalues of S converges almost surely to the parameter σ^2.

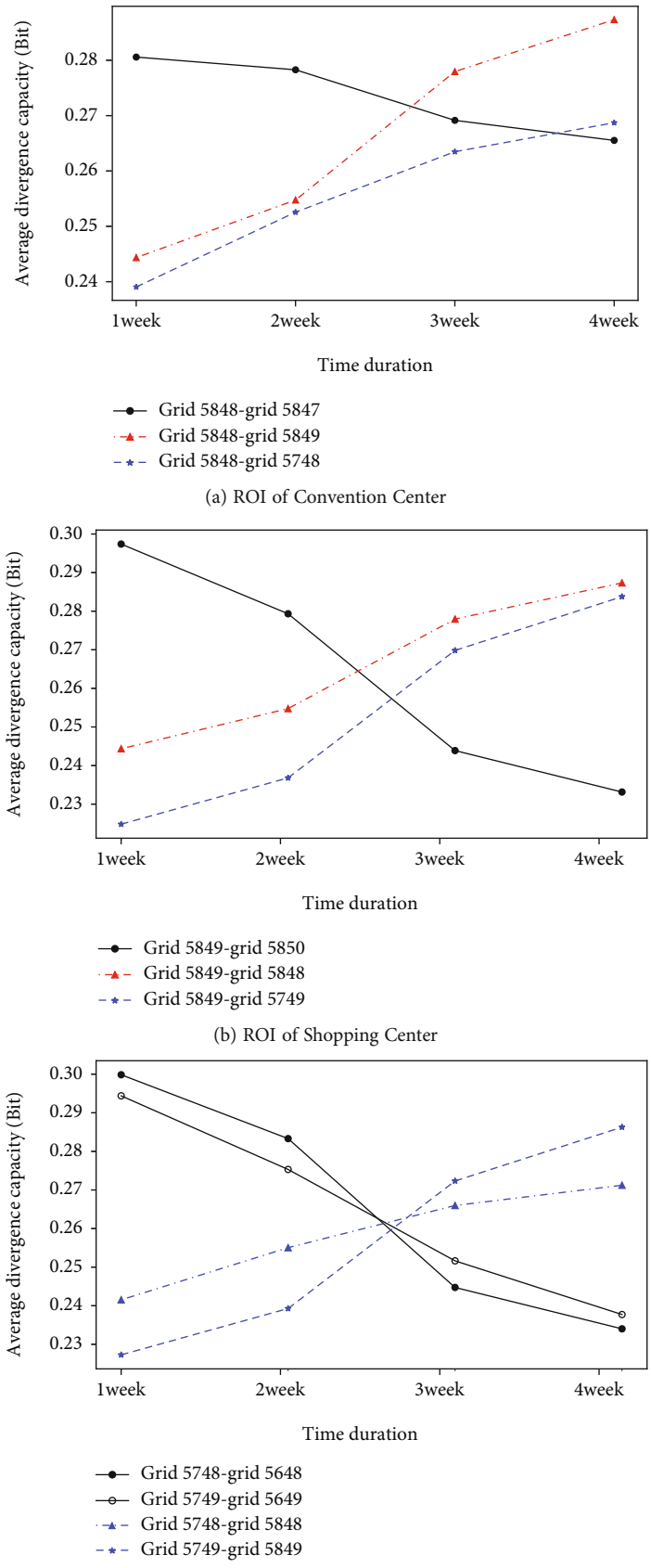


Figure 6: Simulation results of the average divergence capacity of three different ROIs from RRI.

However, in practice, the parameter σ^2 is most likely unknown, so we use the smaller $N-1$ eigenvalues to estimate σ^2. Thereby, the estimation of M can be derived as

$$\widehat{M}_1 = \lambda_N - \widehat{\sigma^2}, \tag{17}$$

where $\widehat{\sigma^2} = (1/(N-1))\sum_{i=1}^{N-1}\lambda_i$ and $\{\lambda_1 \le \lambda_2 \le \cdots \le \lambda_N\}$ are the eigenvalues of S in an ascending order.

The same results can also be obtained from (11) and (12). As $c = (N/T) \longrightarrow 0$, the eigenvalue set of supp $(\lambda_1^-, \lambda_1^+)$ and supp $(\lambda_2^-, \lambda_2^+)$ can be simplified to (18) and (19), respectively,

$$\left[\lambda_1^-, \lambda_1^+\right] = \left[\sigma^2 - \mathcal{O}\left(\frac{1}{T}\right), \sigma^2 + \mathcal{O}\left(\frac{1}{T}\right)\right], \tag{18}$$

$$\left[\lambda_2^-, \lambda_2^+\right] = \left[\left(\lambda_X + \sigma^2\right) - \mathcal{O}\left(\frac{1}{\sqrt{T}}\right), \left(\lambda_X + \sigma^2\right) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)\right]. \tag{19}$$

As $T \longrightarrow \infty$, $\mathcal{O}(1/T)$, and $\mathcal{O}(1/\sqrt{T})$ converge to 0, we hence obtain another result $\widehat{M}_2 = \lambda_N - \widehat{\sigma^2}$, which is the same σ^2 estimation as (17).

5. Experiments on RRI and Network Traffic Prediction

In this section, we conduct experiments on ROI identification using RRI and network traffic prediction with real-world datasets described in Section 2.1. Figure 5 displays the architecture of our proposed ROI identification method.

We present three comprehensive case studies of different ROIs by evaluating the average divergence capacity $I^{r1,r2}$ of adjacent single regions r_1 and r_2 derived from (16). The experiments on network traffic predictions of the three identified ROIs are given in Section 5.1. The three ROIs are Convention Center, Shopping Center, and Central Park.

5.1. Case Studies of the RRI. In order to prove the effectiveness of the proposed ROI identification method, experiments on adjacent single regions are conducted.

5.1.1. Identification for the ROI of Convention Center. The ROI identification starts with the Convention Center in region Grid 5848 as depicted in Figure 1(a), which has been verified with Google Map [33, 34]. By evaluating the average divergence capacity of Grid 5848 with adjacent single regions, the ROI of the Convention Center can be identified. Figure 6(a) illustrates the average divergence capacity of region Grid 5848 with adjacent single regions.

The black solid line indicates the average divergence capacity between the Convention Center (Grid 5848) and the adjacent single region (Grid 5847), which decreases gradually with time advancement. It suggests that the divergent nature of the regional boundary between Grid 5848 and 5847 is growing blurry; in other words, the area network traffic characteristics between the two adjacent single grids are becoming more similar. The red dashed line and the blue dotted line stand for the average divergence capacities of the Convention Center and the adjacent regions of Grids 5849 and 5748, which is intensified as time advances. The intensification of the regional differences (between Grids 5848 and 5849, 5748) indicates that the area network traffic characteristics of the three adjacent regions are becoming more diversified. Therefore, the ROI of the Convention Center can be modified to a bigger area, with Grids 5848 and 5847 aggregated.

Figure 7: Aggregated adjacent ROIs obtained from RRI. The blue shade represents the ROI of Convention Center, the orange shade represents the ROI of Shopping Center, and the green shade represents the ROI of Central Park.

Similar operations are performed on the other two regions of the Shopping Center and Central Park, and their ROIs can be identified as well.

5.1.2. Identification for the ROI of Shopping Center. The ROI identification begins with the Shopping Center in the region of Grid 5849 as denoted by the orange shade in Figure 1(a), which has also been verified with Google Map. We conduct similar computations of the average divergence capacity of Grid 5849 with adjacent single regions (Grids 5850 and 5749) to identify the ROI of the Shopping Center. The results are illustrated in Figure 6(b), from which we can observe that the average divergence capacity of Grids 5849 and 5850 is fast declining; thus, the two adjacent single regions can be aggregated into one ROI. Meanwhile, we note that the average divergence capacity of Grid 5849 with 5848 and 5749 is enhanced with time in contrast, which provides the evidence that the latter two Grids do not belong to the ROI of the Shopping Center.

5.1.3. Identification for the ROI of Central Park. The ROI identification initiates with the Central Park in the regions of Grids 5748 and 5749 as indicated by the green shade in Figure 1(a), which has been verified by Google Map as well. Similarly, we performed computation of the average divergence capacity of Grids 5748 and 5749 with adjacent single regions (Grids 5648 and 5649) to identify the ROI of Central Park. The results are displayed in Figure 6(c). Again, two contrastive tendencies are clearly observable. The average divergence capacities of Grids 5748 and 5749 with Grids

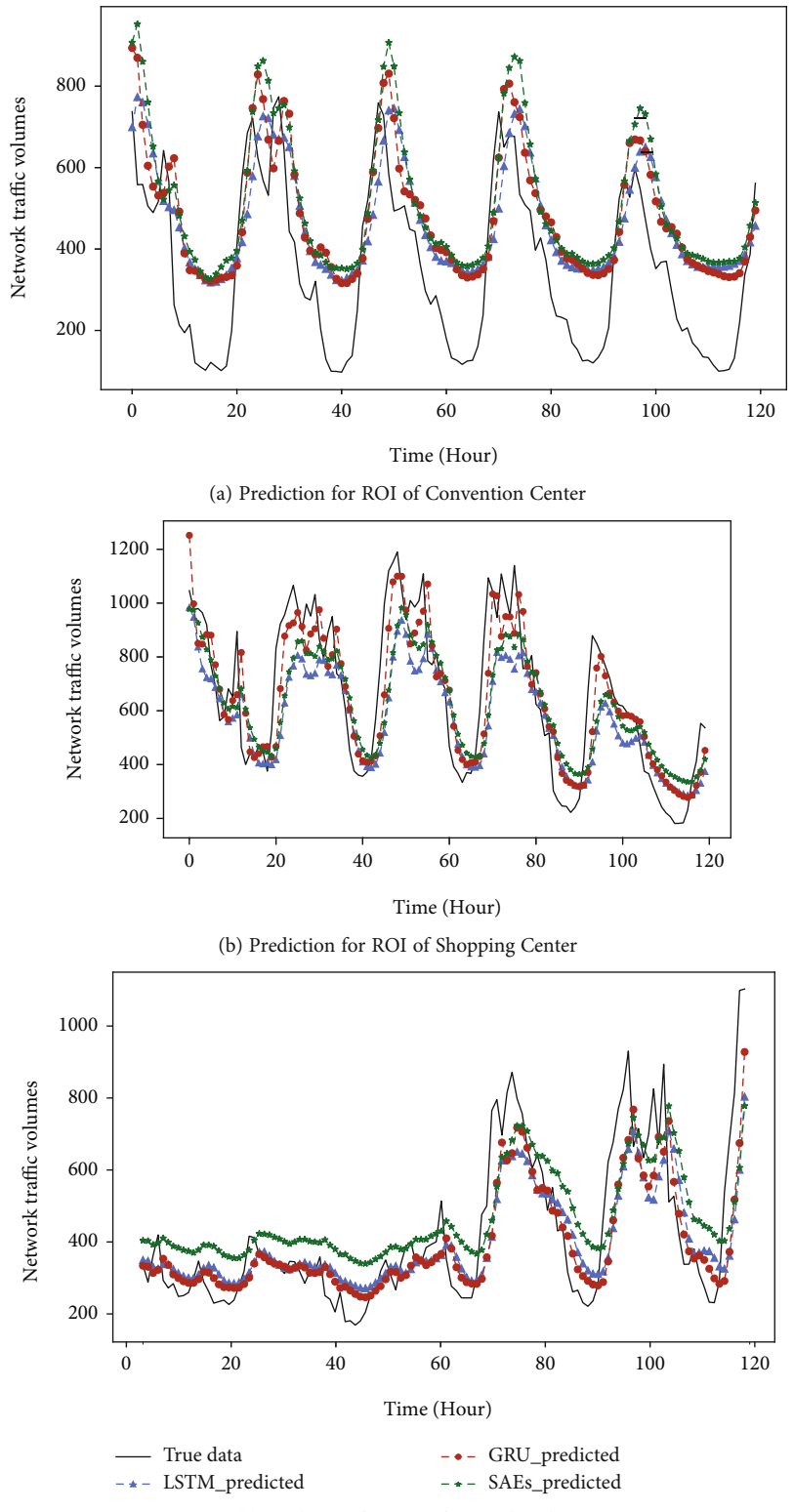


Figure 8: Simulation results of the network traffic volume prediction of three identified ROIs.

5648 and 5649 decline substantially with time advancement in a similar pattern, which indicates that the four adjacent single regions can be aggregated into one ROI. On the other hand, the average divergence capacities of Grids 5748 and 5749 with Grids 5848 and 5849 are intensified over time despite at slightly different speeds, which suggests that the former two Grids (5748 and 5749) do not belong to the same ROI as the latter two.

Table 2: Comparative results of RMSE and MAE prediction schemes on three identified ROIs.

ROIs	Schemes	RMSE		MAE	
		Single	Aggregated	Single	Aggregated
Convention Center	LTSM	577.63	415.33	380.82	222.54
	GRU	437.42	276.16	333.65	185.92
	SAEs	615.23	401.33	506.67	241.86
Shopping Center	LTSM	194.30	178.89	145.30	137.15
	GRU	144.47	125.20	105.10	96.16
	SAEs	250.73	161.00	192.10	125.35
Central Park	LTSM	175.10	148.48	127.11	109.84
	GRU	167.01	132.62	118.16	94.53
	SAEs	159.34	156.93	143.01	130.66

The aggregated adjacent ROIs obtained from RRI are presented in Figure 7, with the blue shade denoting the ROI of the Convention Center, the orange that of the Shopping Center, and the green that of Central Park.

5.2. Case Studies on Network Traffic Prediction in ROIs Identified by RRI. Accurate and timely network traffic prediction plays a pivotal role in intelligent resource allocation [17]. When the divergence between adjacent regions substantially decreases, the aggregation of adjacent regions to one dataset contributes to the improvement of the network traffic prediction performance in the identified ROI.

In order to demonstrate the strength of the aggregated ROI, we apply three neural network-based schemes to predict the network traffic volumes hour by hour, including LSTM, GRU [20], and SAEs [9]. The three prediction methods share the same parameter settings with 3 hidden layers and $\sigma(x) = 1/(1 + e^x)$ as the sigmoid activation function in performance evaluation.

We apply two metrics to evaluate the effectiveness of the prediction performance of the three schemes on aggregated and single ROIs. The first one is root mean square error (RMSE), which measures the difference between the predicted network traffic volumes and the ground truth volumes as defined in

$$\text{RMSE} = \left[\frac{1}{T}\sum_{t=1}^{T}\left(|v_i - \widehat{v}_i|\right)^2\right]^{1/2}, \tag{20}$$

where T is the total time, v_i is the observed network traffic volume, and $\widehat{v}_i$ is the predicted network traffic volume.

The second evaluation index is mean absolute error (MAE), which measures the average of absolute differences between the predicted volumes and the ground truth volumes as defined in

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|v_i - \widehat{v}_i|. \tag{21}$$

5.2.1. ROI of Convention Center. The identified ROI of the Convention Center is the aggregation of Grids 5847 and 5848 as denoted by the blue shade in Figure 5. The data used for the prediction of the Convention Center ROI come from the CDR dataset of the two Grids (5847 and 5848) [31]. Specifically, the training data consist of the aggregated CDR dataset of Grids 5847 and 5848, while the testing data comprise the CDR dataset of Grid 5848 from December 14th to 20th. The prediction results of the three methods on the identified ROI of the Convention Center with aggregated single regions are presented in Figure 8(a).

5.2.2. ROI of Shopping Center. The identified ROIs of the Shopping Center with adjacent single regions are Grids 5849 and 5850, as indicated by the orange shade in Figure 5. The aggregated ROI data of Grids 5849 and 5850 are utilized as the training set, while the data of Grid 5850 from December 14th to 20th are selected as the testing set. The prediction results of the three prediction methods on the identified ROI of the Shopping Center with aggregated single regions are shown in Figure 8(b).

5.2.3. ROI of Central Park. The identified Central Park ROIs with adjacent single regions are Grids 5748, 5749, 5648, and 5649, as denoted by the green shade in Figure 5. The data used for the prediction training are comprised of the aggregated CDR dataset of Grids 5748, 5749, 5648, and 5649 [31], while the data of Grids 5748 and 5749 from December 14th to 20th constitute the testing set. The prediction results of the three prediction methods on the identified ROI of Central Park with aggregated single regions are illustrated in Figure 8(c).

5.3. Discussion on the Prediction Performance of Identified ROIs. We evaluate the performance of the three prediction methods on the identified ROIs by means of the RMSE and MAE metrics in order to demonstrate the strengths of ROI identification with aggregated adjacent single regions.

The comparative results by the RMSE and MAE metrics on the single versus aggregated region prediction performance of the three different schemes on the identified ROIs are displayed in Table 2. A general pattern that emerges from Table 2 is that for all three prediction schemes both of the RMSE and MAE results on identified ROIs with aggregated regions are much smaller than those with single regions. In particular, the largest RMSE difference is found

with the GRU prediction scheme on the identified Convention Center ROI, which is a decrease of 36.87 percent from 437.42 to 276.16. And the largest MAE difference is observed with the SAE prediction scheme on the identified Convention Center ROI, which is a decline of 52.26 percent from 506.67 to 241.86. Based on the above evidence we can conclude that the identified ROIs can substantially improve the performance of network traffic prediction.

6. Conclusion

In this paper, we propose a novel method of RRI which utilizes RMT to analyze the dynamic network traffic characteristics between adjacent regions for ROI identification. By means of RMT, we are able to derive the empirical spectral distribution of the covariance matrix to prove the validity of the spike model. In order to evaluate the divergent degree of identified ROIs, we employ an average divergence capacity model to illustrate the ideological differences with respect to time and region, from which we conclude that the diversity of network traffic in different regions varies with time advancement, and an aggregated ROI can be identified with diminishing diversity between adjacent regions. With our proposed RRI method, we are able to provide more accurate predictions of the network traffic in identified ROIs, which will contribute to the improvement of the system performance, in particular pertaining to energy efficiency and resource allocation.

Data Availability

Data used to support the findings of this study are available at https://doi.org/10.7910/DVN/QJWLFU.

Disclosure

A preprint has previously been published in [30].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Key R&D Program of China (2020YFB1806602).

References

[1] E. Calvanese Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez et al., "6G: the next frontier: from holographic messaging to artificial intelligence using subterahertz and visible light communication," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 42–50, 2019.

[2] X. You, C. X. Wang, J. Huang et al., "Towards 6g wireless communication networks: vision, enabling technologies, and new paradigm shifts," *SCIENCE CHINA Information Sciences*, vol. 64, no. 1, pp. 1–74, 2021.

[3] C. V. N. Index, *Cisco visual networking index: global mobile data traffic forecast update, 2017–2022 white paper*, CA, Cisco: San Jose, USA, 2019.

[4] G. Liu, Y. Huang, N. Li et al., "Vision, requirements and network architecture of 6g mobile network beyond 2030," *China Communications*, vol. 17, no. 9, pp. 92–104, 2020.

[5] P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour, "Design considerations for a 5g network architecture," *IEEE Communications Magazine*, vol. 52, no. 11, pp. 65–75, 2014.

[6] X. Shi, R. Qiu, Z. Ling, F. Yang, H. Yang, and X. He, "Spatio-temporal correlation analysis of online monitoring data for anomaly detection and location in distribution networks," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 995–1006, 2020.

[7] X. Cao, L. Liu, Y. Cheng, and X. S. Shen, "Towards energy-efficient wireless networking in the big data era: a survey," *IEEE Communications Surveys Tutorials*, vol. 20, no. 1, pp. 303–332, 2018.

[8] Y. Shu, M. Yu, O. Yang, J. Liu, and H. Feng, "Wireless traffic modeling and prediction using seasonal Arima models," *IEICE Transactions on Communications*, vol. 88, no. 10, pp. 3992–3999, 2005.

[9] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2014.

[10] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 1147–1161, 2016.

[11] P. Munoz, R. Barco, I. Serrano, and A. Gomez-Andrades, "Correlation-based time-series analysis for cell degradation detection in son," *IEEE Communications Letters*, vol. 20, no. 2, pp. 396–399, 2016.

[12] C. Zhang, H. Zhang, J. Qiao, D. Yuan, and M. Zhang, "Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1389–1401, 2019.

[13] M. Li, Y. Wang, Z. Wang, and H. Zheng, "A deep learning method based on an attention mechanism for wireless network traffic prediction," *Ad Hoc Networks*, vol. 107, p. 102258, 2020.

[14] C. Qiu, Y. Zhang, Z. Feng, P. Zhang, and S. Cui, "Spatio-temporal wireless traffic prediction with recurrent neural network," *IEEE Wireless Communications Letters*, vol. 7, no. 4, pp. 554–557, 2018.

[15] X. Kong, F. Xia, Z. Ning et al., "Mobility dataset generation for vehicular social networks based on floating car data," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 3874–3886, 2018.

[16] F. Xia, J. Wang, X. Kong, Z. Wang, J. Li, and C. Liu, "Exploring human mobility patterns in urban scenarios: a trajectory data perspective," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 142–149, 2018.

[17] W. Wang, C. Zhou, H. He, W. Wu, W. Zhuang, and X. S. Shen, "Cellular traffic load prediction with lstm and gaussian process regression," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pp. 1–6, Dublin, Ireland, June 2020.

[18] D. Jiang, Y. Wang, Z. Lv, S. Qi, and S. Singh, "Big data analysis based network behavior insight of cellular networks for

industry 4.0 applications," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1310–1320, 2019.

[19] X. Kong, K. Wang, M. Hou, F. Xia, G. Karmakar, and J. Li, "Exploring human mobility for multi-pattern passenger prediction: a graph learning framework," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2022.

[20] R. Fu, Z. Zhang, and L. Li, "Using lstm and gru neural network methods for traffic flow prediction," in *31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 324–328, Wuhan, China, Nov. 2016.

[21] F. Xu, Y. Lin, J. Huang et al., "Big data driven mobile traffic understanding and forecasting: a time series approach," *IEEE Transactions on Services Computing*, vol. 9, no. 5, pp. 796–805, 2016.

[22] B. Yu, M. Li, J. Zhang, and Z. Zhu, "3d graph convolutional networks with temporal graphs: a spatial information free framework for traffic forecasting," 2019, http://arxiv.org/abs/1903.00919.

[23] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[24] Y.-S. Jeong, Y.-J. Byon, M. M. Castro-Neto, and S. M. Easa, "Supervised weighting-online learning algorithm for short-term traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 4, pp. 1700–1707, 2013.

[25] W.-C. Hong, "Application of seasonal svr with chaotic immune algorithm in traffic flow forecasting," *Neural Computing and Applications*, vol. 21, no. 3, pp. 583–593, 2012.

[26] J. Wang, J. Tang, Z. Xu et al., "Spatiotemporal modeling and prediction in cellular networks: a big data enabled deep learning approach," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pp. 1–9, Atlanta, GA, USA, May 2017.

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

[28] H. Chen, X. Tao, N. Li, S. Xia, and T. Sui, "Physical layer data analysis for abnormal user detecting: a random matrix theory perspective," *IEEE Access*, vol. 7, no. 169, pp. 169508–169517, 2019.

[29] H. Chen, X. Tao, N. Li, and Z. Han, "A data analysis of political polarization using random matrix theory," *SCIENCE CHINA Information Sciences*, vol. 63, no. 2, p. 129303, 2020.

[30] T. Sui, X. Tao, H. Wu, X. Zhang, and J. Xu, "Dimension increased random matrix method for anomaly detection in wireless networks," in *2021 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, pp. 60–65, Xiamen, China, July 2021.

[31] "First edition of the big data challenge," 2014, http://theodi.fbk.eu/openbigdata/.

[32] T. Italia, "Milano Grid," *Harvard Dataverse*, 2015.

[33] T. Sui, X. Tao, S. Xia et al., "A realtime hidden anomaly detection of correlated data in wireless networks," *IEEE Access*, vol. 8, pp. 60990–60999, 2020.

[34] M. S. Parwez, D. B. Rawat, and M. Garuba, "Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 2058–2065, 2017.

[35] L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud, "Random matrix theory and financial correlations," *International Journal of Theoretical and Applied Finance*, vol. 3, no. 3, pp. 391–397, 2000.

[36] X. Xu, X. He, Q. Ai, and R. C. Qiu, "A correlation analysis method for power systems based on random matrix theory," *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1811–1820, 2017.

[37] Y. Cao, L. Cai, C. Qiu et al., "A random matrix theoretical approach to early event detection using experimental data," 2015, http://arxiv.org/abs/1503.08445.

[38] B. Han, L. Luo, G. Sheng, G. Li, and X. Jiang, "Framework of random matrix theory for power system data mining in a non-gaussian environment," *IEEE Access*, vol. 4, pp. 9969–9977, 2016.

[39] Z. Bai and J. W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, Springer, New York, 2nd ed. edition, 2010.

[40] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*, Cambridge University Press, 2011.

[41] O. Ryan and M. Debbah, "Free deconvolution for signal processing applications," *IEEE International Symposium on Information Theory*, vol. 2007, pp. 1846–1850, 2007.

[42] P. Vallet, P. Loubaton, and X. Mestre, "Improved subspace estimation for multivariate observations of high dimension: the deterministic signals case," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 1043–1068, 2012.

[43] T. M. Cover, *Elements of Information Theory*, John Wiley & Sons, 1999.