Spatio-Temporal Analysis and Prediction of Cellular Traffic in Metropolis

Xu Wang, *Student Member, IEEE*, Zimu Zhou, *Member, IEEE*, Fu Xiao, Kai Xing, *Member, IEEE*, Zheng Yang, *Senior Member, IEEE*, Yunhao Liu, *Fellow, IEEE*, and Chunyi Peng, *Senior Member, IEEE*

Abstract—Understanding and predicting cellular traffic at large-scale and fine-granularity is beneficial and valuable to mobile users, wireless carriers, and city authorities. Predicting cellular traffic in modern metropolis is particularly challenging because of the tremendous temporal and spatial dynamics introduced by diverse user Internet behaviors and frequent user mobility citywide. In this paper, we characterize and investigate the root causes of such dynamics in cellular traffic through a big cellular usage dataset covering 1.5 million users and 5,929 cell towers in a major city of China. We reveal intensive spatio-temporal dependency even among distant cell towers, which is largely overlooked in previous works. To explicitly characterize and effectively model the spatio-temporal dependency of urban cellular traffic, we propose a novel decomposition of in-cell and inter-cell data traffic, and apply a graph-based deep learning approach to accurate cellular traffic prediction. Experimental results demonstrate that our method consistently outperforms the state-of-the-art time-series based approaches and we also show through an example study how the decomposition of cellular traffic can be used for event inference.

Index Terms—Machine learning, prediction methods, predictive models, mobile computing, communication systems, mobile communication

1 Introduction

With billions of mobile devices accessing the Internet via 3G/4G/5G networks, cellular traffic has skyrocketed in the past few years. It is predicted that over 50 percent of the global devices and connections will be mobile (e.g., smartphones, tablets) and the monthly global mobile data traffic will surpass 30.6 exabytes (10^{18}) by 2020 [1]. This trend will continue in the foreseeable future.

Despite the volume of mobile cellular traffic data, we have limited knowledge on its spatio-temporal patterns at urban scale. Tremendous records of cellular traffic collected at cell towers have been widely adopted for daily network management and diagnosis. However, we argue that the big traffic data from cell towers are also beneficial to *understand* and *predict* urban cellular traffic, which can be extremely valuable for individual cell towers, cellular carriers, and city authorities. In the coming 5G era, cell towers can adapt themselves to the dynamics in traffic load based on Software-Defined Networking / Network Function Virtualization (SDN / NFV) technology. Accurate prediction of cellular traffic will facilitate carriers to schedule resources to ensure the overall quality of service and network performance and reduce unnecessary operation cost by allocating energy and bandwidth tightly based on the future traffic demand. Cellular data traffic prediction also assists city authorities to discover certain social events in time for urban governance. For instance, city managers can take prevention measures or preemptive actions for spontaneous gatherings of people to avoid injury caused by crowd stampede.

However, it is extremely challenging to predict mobile cellular traffic at both large-scale and fine granularity. The reasons are three-fold. (i) Due to the diverse network demand of Internet-based applications (e.g., mobile videos, location-based games, VoIP) and user behaviors (e.g., at work, in transit, during sleep), the cellular traffic at an individual cell tower can have a wide dynamic range. According to our dataset in a major city of China, the traffic volume can easily peak at around 1 GB per hour during rush hours (e.g., 16:00), which is 100,000 times greater than the traffic during the least active times (e.g., 04:00) in the same cell tower. (ii) User mobility introduces spatial dependencies into the cellular traffic among spatially distributed cell towers. Our dataset reveals that data traffic from mobile users (i.e., entering or leaving a cell within a time interval) can account for up to 90 percent of the entire data traffic at cell towers in transportation hubs. We also observe that such spatial dependencies can occur even between distant cell towers, as efficient urban transportation easily enables mobile users to travel across cities within half an hour. (iii)

- *X. Wang, Z. Yang, and Y. Liu are with the School of Software and TNLIST, Tsinghua University, Beijing 100084, China. E-mail: xu-wang11@mails.thu.edu.cn, {yangzheng, yunhao}@tsinghua.edu.cn.*
- *Z. Zhou is with the Computer Engineering and Networks Laboratory, ETH Zurich, Zürich 8092, Switzerland. E-mail: zzhou@tik.ee.ethz.ch.*
- *F. Xiao is with the Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing University of Posts and Telecommunications, Nanjing 210028, China. E-mail: xiaof@njupt.edu.cn.*
- *K. Xing is with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China. E-mail: kxing@ustc.edu.cn.*
- *C. Peng is with the Department of Computer Science, Purdue University, West Lafayette, IN 47907. E-mail: chunyi@purdue.edu.*

Manuscript received 30 Oct. 2017; revised 21 Aug. 2018; accepted 3 Sept. 2018. Date of publication 17 Sept. 2018; date of current version 7 Aug. 2019. (Corresponding author: Xu Wang.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TMC.2018.2870135

Geographical distribution of cellular traffic at the urban scale is also influenced by many other factors, including land use, population, holidays, and various social activities. These influential factors further complicate the spatio-temporal dependencies among cell tower traffic citywide.

Network traffic analysis and prediction have been well studied, covering from wired Internet to cellular networks. In the past, mainstream research has modeled traffic patterns using statistics or probabilistic distribution in the time domain [2], space domain [3], or both [4]. Although these works provide a comprehensive understanding of the Internet traffic, the outcomes cannot be utilized to predict traffic load for individual cell tower continuously. Traffic prediction is more challenging than characterization. Existing solutions either totally ignore the spatial influence of cell towers at different locations [5], or use an approximate model (e.g., spatial aggregation [6], [7] or statistical covariance [8]). These solutions fail to capture the intensive and often long-distance spatial dependency of individual cell towers induced by citywide user mobility, let alone the interaction between temporal and spatial factors.

In this work, we carefully investigate the characteristics of urban cellular traffic with a large-scale cellular data usage dataset covering 1.5 million users and 5,929 cell towers in a major city of China. We demonstrate that there is strong, long-distance, and pervasive spatial dependency among cell tower cellular traffic, which was overlooked in previous research. To explicitly account for the spatial dependency, we propose to decompose the total data traffic volume into in-tower traffic and inter-tower traffic, and verify the importance of such decomposition in cellular traffic prediction. We leverage a graphical representation to comprehensively capture the spatio-temporal correlations of cellular traffic, and exploit graph neural network to learn an efficient spatio-temporal model from our massive dataset. We also demonstrate that the combination of in-tower and inter-tower traffic patterns can be applied for network or social event inference, which is practically helpful for mobile carriers to adjust network configuration, or for city authorities to take preventive measures. Evaluations on our massive dataset validate the effectiveness of our graphical spatio-temporal model.

The contributions of this work are summarized as follows.

- Conceptually, we decompose cellular traffic into in-tower and inter-tower traffic to characterize the spatial dependency among cell towers, which was rarely taken into account previously. We show the necessity and benefits to account for spatial dependency in cellular traffic prediction through measurements on a massive dataset.
- We jointly consider temporal and spatial dependency among cell towers and exploit a deep neural network based model for cellular traffic prediction. The model effectively parameterizes in-tower and inter-tower traffic, and is able to learn long-distance spatial correlations for accurate traffic prediction. This is the first work that applies deep learning for individual cell tower traffic prediction at urban scale with massive real-world datasets.
- We achieve large-scale (metropolis) and fine-grained (individual cell tower, half an hour) traffic prediction and outperform the state-of-the-art by 13.2 percent in Mean Absolute Error (MAE) and 17.5 percent in Mean Absolute Relative Error (MARE). To the best of our knowledge, our work is the first that is capable of predicting in such scale and granularity and has been evaluated on a real-world big dataset.

In the rest of the paper, we review related works in Section 2 and describe our dataset and motivating observations in Sections 3 and 4, respectively. On this basis, we propose our graph-based prediction model in Section 5 and evaluate the performance in Section 6. Section 7 finally concludes this study.

2 Related Work

Network traffic analysis and prediction has a long history, from the emergence of data network several decades ago to today, spanning from Internet [9], [10] to cellular networks (e.g., voice call [11], [12] and SMS [13]).

In recent years, the explosive growth of mobile Internet applications has drawn numerous research attempts to characterize cellular traffic from both academy and industry. Network traffic prediction could benefit plenty of applications. According to the sources of data traffic records, existing works can be generally divided into two categories: (i) data recorded by mobile devices; (ii) traces collected by cellular operators. For the first category, phone usage information, including locations, applications, network performance, is automatically monitored and logged by mobile APPs [14], [15], some of which work in a crowdsensing manner [16], [17], [18]. The major limitation of this approach lies in its scale; that is, the user coverage of data traces is restricted within a group of people who install some specific APPs on their mobile devices. As a result, the statistics obtained cannot truly reflect the whole set of users associated with a cellular tower, not to mention the global characteristics of a large scale cellular network. On the other hand, in the dataset collected by cellular operators, users are passively monitored without awareness. In fact, mobile carriers are able to record a wealth of information in many aspects [19], [20], [21], depending on where and in which layer the monitors are placed in cellular networks. Datasets of the second category are often employed for analyzing network-wide statistics with a large user population. Our work, relying on the dataset collected from citywide cell towers, falls into the second category.

Most previous works on cellular traffic focus on how to characterize and understand their statistics and patterns under the circumstance of various temporal and spatial factors, device types, application categories, user groups, etc. From the time domain, the traffic dynamics of IP traffic can be well captured by Markov models [2]. An interesting fact from [22] is that only five basic temporal patterns of traffic exist among cell towers in a city and each of the pattern maps to one type of geographical locations related to urban ecology. This branch of researches focuses mostly on temporal variations without spatial relationship of cells at different locations. From the space domain, Gotzner et al. [3] breach the homogeneous assumption in regular spectrum frequency analysis and propose to model the spatial inhomogeneity of real cellular traffic with log-normal distributions. [5]

demonstrates that the spatial distribution of the traffic density can be approximated by the log-normal or Weibull distribution. Furthermore, [4] finds that the mobile traffic loads follow a trimodal distribution, which is the combination of compound-exponential, power-law and exponential distributions, in terms of both spatial and temporal dimension. [23] analyzes the application-level traffic and demonstrates that the traffic at a service or application has α-stable modeled property in the temporal domain and the sparsity in the spatial domain. These approaches provide comprehensive understanding of cellular traffic, e.g., temporal dynamics and spatial inhomogeneity. However, it still has a long away to realize traffic prediction.

For the more challenging problem of prediction, by treating data traffic of cell towers as a time series, the mainstream solutions employ time domain data analysis and modeling, such as AutoRegressive Integrated Moving Average (ARIMA), Markov Decision Process (MDP) [24] and Holt-Winters method[25], stochastic process models, Kalman filtering, etc.

Applying statistical covariance functions, a recent work [8] considers temporal and spatial factors simultaneously, which is the most similar to our work. However, the traffic transfer among cell towers caused by human mobility is not taken into account. Another attempt [6] models the spatial correlation by clustering nearby cell towers with the same traffic pattern into groups and predicts aggregated traffic for each group. Such relaxation in space domain neither captures spatial correlation of faraway cell towers, nor provides useful results for individual cell tower. Recently, more advanced machine learning techniques are utilized for cellular traffic prediction [7], [26], [27] where spatial dependencies are usually learned by Convolutional Neural Networks (CNN). However, these methods only work with grid-based region partition, which is unpractical for cellular network management. We adopt an upgraded version of [6] as the state of the art comparison in our evaluation, via replacing Elman neural network used in [6] with an up-to-date Long short-term memory (LSTM) neural network. Different from these two similar works, our proposed solution explicitly models spatio-temporal dependency from inter-tower and in-tower traffic and predicts cellular traffic at a large scale and fine granularity, through a neural network that can learn from a graph structure.

The locations of people are estimated by meticulously designed schemas in both indoor context [28], [29] and outdoor context [30], which can be applied to understanding human mobility. The footprint of mobile devices recorded by cellular network has been widely used to understand human mobility [31], [32], [33] and predict human's location [34], [35], [36]. Different from these works, our work focus on people transferring between regions and analyze network traffic consumed by them. We believe that more fine-grained human mobility model will benefit network traffic prediction and we leave this work for future to further explore.

3 Background and Data Set

This section presents the architecture of a typical cellular network monitoring system and gives an overview of our dataset.

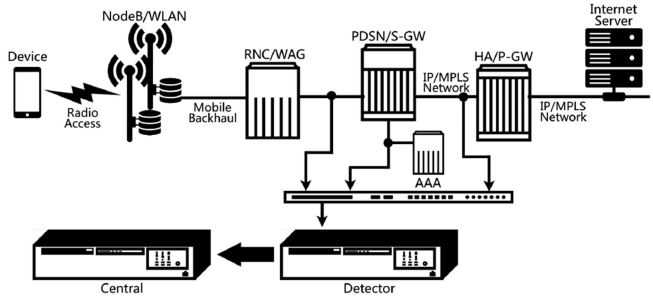


Fig. 1. An illustration of cellular network architecture and data monitoring system.

3.1 Data Monitoring in Cellular Networks

Fig. 1 shows the architecture of a typical cellular network monitoring system. To facilitate network management, a monitoring system is deployed for operators to analyze data traffic, monitor network performance and detect anomalies. The monitoring system consists of a *detector* and a *central*. The detector is deployed in the packet core and the central is deployed in the Network Operations Center (NOC). Every packet sent by a mobile device will be monitored on the interface between the Packet Data Serving Node (PDSN) / Serving Gateway (S-GW) and Home Agent (HA) / Packet Data Network Gateway (P-GW) in CDMA and LTE networks. The monitoring system records important information for network diagnosis and forensics. Each record contains bidirectional flow information with the following key fields: source / destination pair by IP and / or International Mobile Station Equipment Identification (IMEI) / International Mobile Subscriber Identification (IMSI) or Network Access Identifier (NAI) / Electronic Serial Number (ESN) / Mobile Station Identifier (MSID), application(s) and wireless network resources consumed (e.g., traffic volume, airtime, connection setup counts).

3.2 Dataset Description

Our dataset was collected by a major cellular carrier in a big city of China as a large structured flow-level data table. Each entry in the table records the user ID, the flow create time, the flow connected cell tower ID, App ID, device type ID, uplink traffic and downlink traffic. To protect the privacy, (i) The user ID is anonymized by hashing the IMEI / IMSI which is unique for each device. (ii) All of the researchers have been authorized by the cellular network operator to utilize the dataset for research purposes, and are bounded by strict non disclosure agreements. (iii) We store all the data in a secure off-line server and only the aggregated traffics for cell towers are accessible for our researchers. Application types are classified by analyzing the HTTP header, the host name and HTTP referrer for HTTP packages and the host name for HTTPS packages. In our study, we exploit downlink traffic for analysis. Table 1 summarizes the basic statistics for the dataset. Fig. 2 visualizes all available cell towers in our dataset.

Since each record corresponds to each flow between the devices and the cellular towers whenever active or passive network activities, the dataset is much fine-grained and covers people in the whole city. To the best of our knowledge, the dataset is the one of the largest urban-scale cellular traffic dataset in terms of the number of mobile users and cell towers. The wide coverage in mobile users and cell towers

TABLE 1
Dataset Description

Statistics	Value
Flow Records	1.7×10^{10}
Cell Towers	5.9×10^{3}
Covered Users	1.5×10^{6}
Covered Apps	7.0×10^{2}
Covered Area	1.0×10^{4} km^2
Date	June 5th-18th, 2016

promises to capture comprehensive varieties and spatial dynamics in cellular traffic patterns.

3.3 Preprocessing

For better capturing the citywide traffic pattern, we preprocess our dataset in three steps: (i) Some records in our dataset are not complete (i.e., missing user ID or cell tower ID), we remove these records from the dataset. (ii) Note that devices outside the city may also be recorded in the dataset due to the ISPs mobile roaming policy. This portion of mobile records is also removed by checking the coordinates of the cell towers. (iii) As we aim to predict cellular traffic for each individual cell tower at a granularity of half an hour, the preliminary aggregation of the downlink traffic is done according to each cell tower ID and user ID for every half an hour, as a result, we could aggregate the traffic much faster according to our applications.

4 Preliminary Observation and Motivation

4.1 Spatio-Temporal Distribution of Cellular Traffic

Fig. 3 shows the spatial distribution of cellular traffic (bytes per half-hour per km^2) at different times (04:00, 10:00, 16:00, 00:00) on June 15th, 2016, a weekday. Specifically, cell towers in the city center and some industry areas exhibit heavy cellular traffic throughout the day. Overall the cellular traffic before dawn (04:00) is low as most residents in the city are sleeping. At 10:00 the traffic at most areas starts to increase as most people start working. We observe intensive cellular traffic widespread through the city at 16:00, indicating people are involved in diverse activities and are highly mobile. Surprisingly, we find high volume of cellular traffic even at midnight. The reason might lie in the fact that it is a metropolis and there can be rich night life during summer.

Further, we examine the correlations among cell towers using the Moran's I [37], which is a well-known measure of spatial autocorrelation. Formally, let $x_t(c_i)$ denote the traffic of a cell tower c_i at time t, then the definition of Moran's I is

$$I(t) = \frac{n}{\sum_{i,j} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_t(c_i) - \overline{x_t})(x_t(c_j) - \overline{x_t})}{\sum_{i=1}^{n}(x_t(c_i) - \overline{x_t})^2}, \quad (1)$$

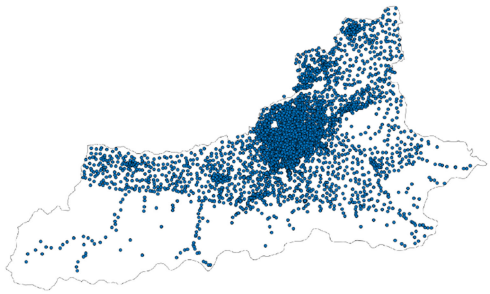


Fig. 2. An illustration of cell towers in our cellular dataset.

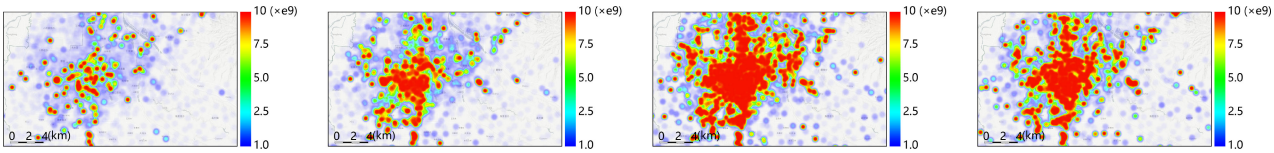
where $w_{ij} = \frac{1}{d_i}$ when the distance of c_i and c_j is less than 2 kilometres, d_i is the degree of c_i. For a normalized weight matrix (i.e., $\sum_j w_{ij} = 1$), values significantly below $\frac{-1}{N-1}$ indicate negative spatial autocorrelation and values significantly above $\frac{-1}{N-1}$ indicate positive spatial autocorrelation. Fig. 4 shows the values of Moran's I of cell traffic for six days. As shown, almost all the Moran's I values are greater than 0.2, indicating positive spatial autocorrelation.

4.2 Characterizing Cell Tower Traffic Using in-Tower and Inter-Tower Traffic

Existing solutions rarely consider data traffic mobility induced by human mobility, which plays a critical role in prediction according to our measurement. To understand the spatio-temporal varieties of cell tower traffic, we propose to distinguish *in-tower* and *inter-tower* traffic. Specifically, instead of considering the traffic of a cell tower c_i at time t (denoted as $x_t(c_i)$) as a whole, we propose to decompose $x_t(c_i)$ into *evisein-tower traffic* $x_t^A(c_i)$ and *inter-tower traffic* $x_t^B(c_i)$, where A is the set of mobile devices residing within the coverage of cell tower c_i, and B stands for the set of mobile devices just entering the coverage of c_i from another cell tower. Formally, let $P_t(c_i)$ be the set of mobile devices at cell tower c_i at time t. Then $x_t(c_i) = x_t^A(c_i) + x_t^B(c_i)$, where $A = P_t(c_i) \cap P_{t-1}(c_i)$ and $B = P_t(c_i) \setminus P_{t-1}(c_i)$.

In particular, if a mobile device traverses several cells in a unit time interval (i.e., half an hour in our case), it will be associated with c_k and put into the set $P(c_k)$, where c_k is the last cell along its trajectory. Accordingly, its traffic during the time interval will be put on the last cell entirely. This approximation is feasible since the most devices (i.e., 82.8 percent) only visit one or two cell in a unit time interval as show in Fig. 6, and the induced error is ignorable compared with the traffic of a cell tower.

Fig. 5 shows the normalized hourly traffic characteristics of three representative cell towers. As shown, the total traffic of all the three cell towers exhibits dramatic temporal dynamics within the day, and the three cell towers demonstrate distinctive in-tower and inter-tower traffic patterns. In the first type of cell towers (Fig. 5a), in-tower traffic



(a) 04:00 (b) 10:00 (c) 16:00 (d) 00:00

Fig. 3. Spatial distribution of urban cellular traffic at different times of a day.

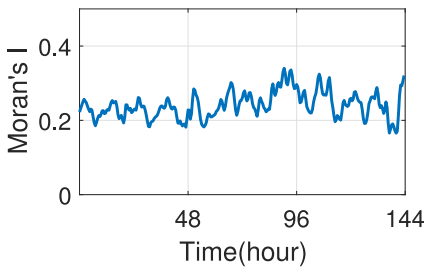


Fig. 4. Moran's I of cell traffic for six days.

overwhelms the entire cellular traffic, indicating the total traffic volume is dominated by the Internet access behaviours of a fixed group of users. In the second type of cell towers (Fig. 5b), inter-tower traffic takes up a notable portion of the total traffic in the daytime. These type of cell towers are likely to be located in places with continuous and intensive mobility in the daytime, e.g., shopping malls. In the third type of cell towers (Fig. 5c), inter-tower traffic surges during the morning and evening rush hour and lunch time. A cell tower at public transportation hubs within the city may demonstrate such a traffic pattern.

Compared with the first type of cell towers, the data traffic of the last two types of cell towers is largely determined by inter-tower traffic, which is induced by mobility. Mobility from cell to cell introduces *spatial* correlations among cell tower traffics. As a consequence, when predicting the traffic of one individual cell tower, it is important to take into account the spatial distributions and dependencies of cellular traffic of cell towers at different locations. In fact, as we will show in the evaluation, the state-of-the-art time-series based cellular traffic prediction approaches [38], [39] perform poorly for the latter two types of cell towers, as they ignore the spatial correlations among cell towers induced by mobility.

5 Spatio-Temporal Modelling for Traffic Prediction

Motivated by the observations from the above in-tower and inter-tower traffic pattern analysis, in this section, we propose a deep graph model for fine-grained (individual cell tower, half-hour granularity) cellular traffic prediction. Furthermore, we decompose cellular traffic into explainable components and design an inference scheme to discover network or social events.

5.1 Problem Statement

Formally, our cellular traffic prediction problem can be formulated as follows. Let $C = \{c_i | i = 1, \ldots, N_c\}$ denote the set

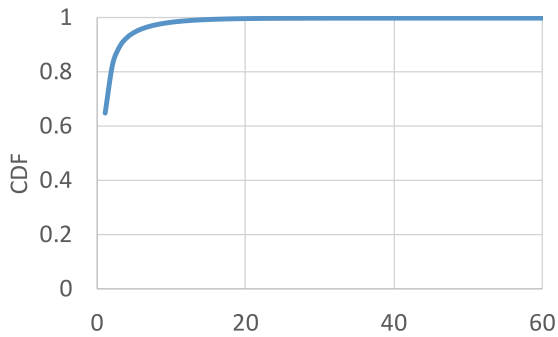


Fig. 6. CDF of the number of cell towers mobile devices traverse in a unit time interval.

of cell towers, where c_i is the ith cell tower, and N_c is the number of cell towers. Let $x_t(c_i)$ denote the traffic of a cell tower c_i at time t. Given a set of cellular traffic $\{x_1, \ldots, x_{t-1}\}$, our goal is to predict $x_t(c)$ for all $c \in C$. As demonstrated in Section 4.2, the majority of the total data traffic can be inter-tower traffic induced by mobility. Thus it is necessary and beneficial to account for spatial dependencies for accurate cellular traffic prediction. That is, we assume $x_t(c_i)$ is dependent on both $\{x_k(c_i)\}$ (temporal factor) and $\{x_k(c_j), j \neq i\}$ due to user mobility (spatial factor), where $k = 1, \ldots, t-1$.

5.2 Graph Representation of Spatio-Temporal Dependencies

We model the spatio-temporal dependencies of cellular traffic using a graph representation. Specifically, given a directed graph $G = (V, E)$, we denote each vertex $c \in V$ as a cell tower, and each edge $e = (c, d) \in E$, where $d \in V$, as the spatial dependency of a cell tower c on d. The neighbours of c, which have incoming edges or outgoing edges to c, are represented by $NBR(c)$, and $CO(c)$ stands for the edges (both incoming and outgoing) connected to c. Given an edge $e = (d, c) \in CO(c)$, its weight $w(e)$ is defined as the total data traffic vector of mobile devices that move from d to c at each time t, i.e., $[x_1(d \to c) \ldots x_t(d \to c) \ldots]$.

Due to the large scale of our dataset (5,929 cell towers), a complete directed graph can contain huge amounts of edges (approximately 2 million), thus hindering efficient learning of model parameters. Besides, in some edges, $w(e)$ only has little number of nonzero items, which should be regarded as noise. Therefore we prune edges with small weights to speed up the learning process and empirically set a threshold of nonzero count in w_e to balance prediction accuracy and computing efficiency.

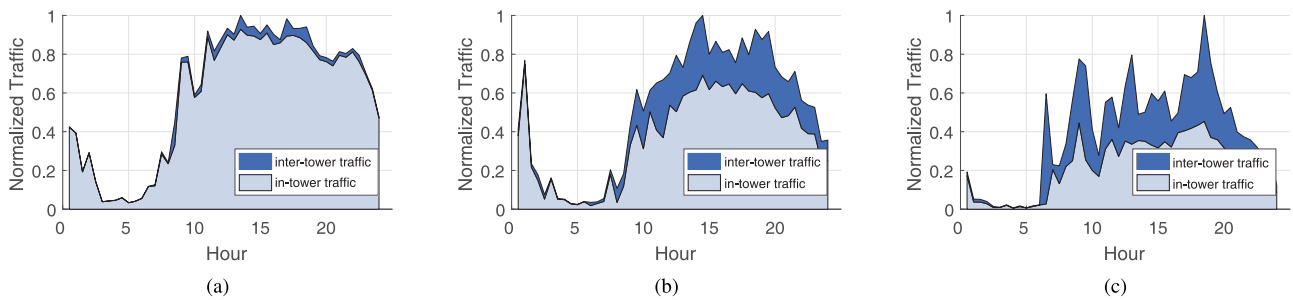
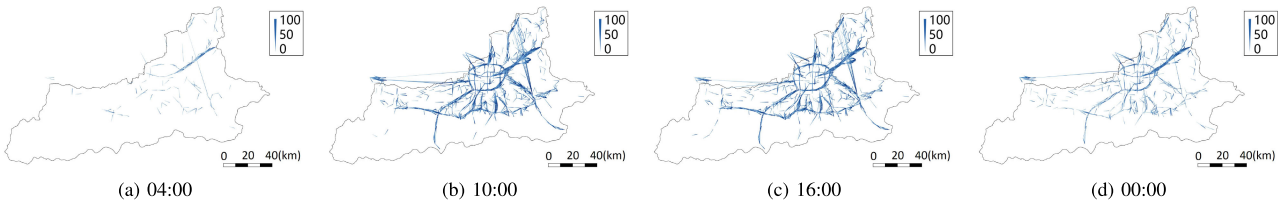
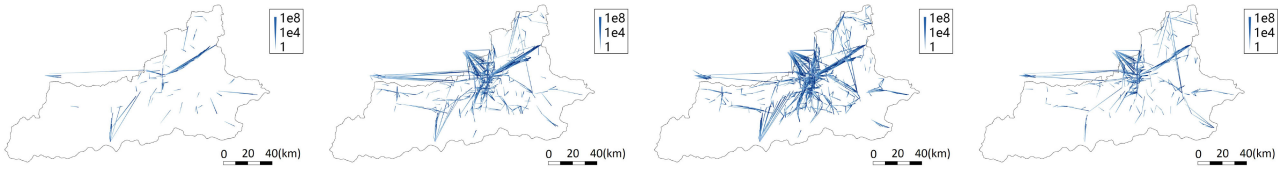


Fig. 5. An illustration of three typical in-tower and inter-tower cell tower data traffic characteristics. (a) In-tower traffic dominant, collected from a cell tower in a residential area; (b) inter-tower traffic consistently notable during the daytime, collected from a cell tower in a shopping mall; and (c) inter-tower traffic surges at certain times, collected from a cell tower in a transit station.



(a) 04:00 (b) 10:00 (c) 16:00 (d) 00:00

Fig. 7. Distribution of user mobility at different times of a day. An edge with gradient color from white to dark blue represents the direction of user mobility between a pair of cell towers. The line width of the edge shows the intensity of user mobility.



(a) 04:00 (b) 10:00 (c) 16:00 (d) 00:00

Fig. 8. Distribution of inter-tower traffic volume at different times of the same day. The line width of each edge represents the volume of inter-tower traffic between a pair of cell towers, aggregated by half an hour.

One important observation in our study is that the inter-tower traffic is highly correlated with user mobility in the city. Fig. 8 shows the spatial distribution of the corresponding inter-tower traffic (per half-hour) at 04:00, 10:00, 16:00 and 00:00. The line width of each edge shows the volume of the corresponding inter-tower traffic. We omit edges of distances shorter than 3 km to investigate whether there is extensive mobility between distant cell towers.

Fig. 7 plots the spatial distribution of user mobility within the city at 04:00, 10:00, 16:00 and 00:00. User mobility is quantified by summing the total number of mobile devices travelling between pairs of cell towers within each half-hour unit, and is represented by a directional edge with its line width showing the intensity of user mobility. We omit edges of distances shorter than 3 km to investigate whether there is extensive mobility between distant cell towers.

Fig. 8 shows the spatial distribution of the corresponding inter-tower traffic (per half-hour) at the same times of the day. The line width of each edge shows the volume of the corresponding inter-tower traffic.

Comparing the user mobility and the corresponding inter-tower traffic, we find that (1) both the intensity and geographical distribution of user mobility are tightly correlated to inter-tower traffic; and (2) inter-tower traffic between two distant cell towers can be dependent because there may be intense user mobility.

The first finding justifies our decomposition of cellular traffic into in-tower traffic and inter-tower traffic to characterize spatial dependencies induced by user mobility. The second finding indicates that the spatial dependencies among cell tower traffic may not necessarily be local (e.g., within the same residential district). The rapid development of urban transportation has made it easy to transit among urban transportation hubs (e.g., airport, train stations), and popular locations (e.g., hot tourist attractions) within half an hour. Such fast urban-scale mobility introduces enormous long-distance spatial dependency to cell towers within the whole city. Thus the neighborhood defined in our graph representation is based on device mobility, rather than geographical distances among cell towers.

5.3 Learning Spatio-Temporal Dependencies via Graph Neural Networks

To efficiently learn the spatio-temporal dependency in cell tower traffic, we adopt a Graph Neural Network (GNN) [40] model. A GNN is a general neural network architecture defined on a graph structure $G = (V, E)$. Nodes $v \in V$ take unique values from $1, \ldots, |V|$ representing each cell tower in our model, and edges are pairs $e = (v, v') \in V \times V$. Since inter-tower traffic is directional, we denote (v, v') as a directed edge $v \to v'$. In a GNN, each node v is assigned with a hidden state named node representation, which is denoted by $h_v \in \mathbb{R}^r$, where r is the dimension of node representation. Node representation h_v models the spatio-temporal features for each node v and is to be trained, which will be explained later.

To separately encode in-tower traffic and inter-tower traffic into a GNN, we assign in-tower traffic as the labels for each node, and inter-tower traffic as the labels of each edge. Specifically, each node v is associated with a node label sequence $x^A(v) \in \mathbb{R}^l$, where l is the history data length for prediction and x^A is the identifier for the node labels. Similarly, each edge e is associated with an edge label sequence $x^B(e) \in \mathbb{R}^l$, where x^B is the identifier for the edge labels. Afterwards, it is natural to use the in-tower traffic series to represent the node sequence, i.e., $x_t^A(v) = (x_{t-l}^A(v), x_{t-l+1}^A(v), \ldots, x_{t-1}^A(v))^T$, and the inter-tower traffic series to represent the edge sequence, i.e., $x_t^B(e) = (x_{t-l}^B(e), x_{t-l+1}^B(e), \ldots, x_{t-1}^B(e))^T$. In the above modelling, the GNN captures both temporal dependency of individual cell towers and spatio-temporal dependency among cell towers network-wide of an order of l, which is an important parameter for tuning.

For ease of presentation, we overload notations and let $x^A(S) = \{x^A(s)|s \in S\}, h(S) = \{h(s)|s \in S\}, x^B(S) = \{x^B(s)|s \in S\}$, and the function $In(v) = \{v'|(v', v) \in E\}$ returns the set of predecessor nodes v' with $v' \to v$. Analogously, $OUT(v) = \{v'|(v, v') \in E\}$ is the set of successor nodes v' with edges $v \to v'$. The set of all nodes neighboring v is $NBR(v) = IN(v) \cup OUT(v)$, and the set of all edges incoming to or outgoing from v is $CO(v) = \{(v', v'') \in E|v = v'||v = v''\}$.

A GNN model maps the graphs to the outputs (prediction of cellular traffic) in two steps: (1) a propagation step

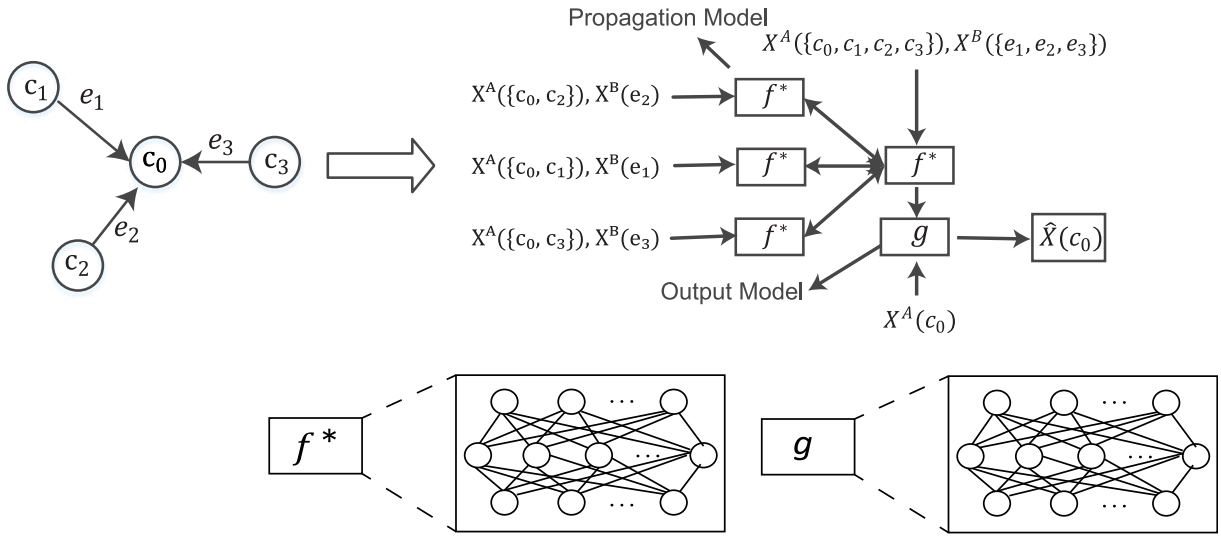


Fig. 9. An illustration of the cellular traffic prediction process for cell tower c_0 using GNN. The left shows the graph representation, where c_0 has inter-tower traffic $x^B(e_i)$ from $\{c_1, c_2, c_3\}$. The right describes the GNN model. For each cell c_i, the in-tower traffic X^A and the inter-tower traffic X^B of all the nodes and edges connected to c_i are input into f^* (propagation step) to learn spatio-temporal features. Then, the feature learned and the in-tower traffic of c_0 are fed into g (output model) to get the prediction $O(c_0)$, i.e., the predicted traffic for c_0.

that computes the node representations $h(v)$ for each node v, and (2) an output model $o(v) = g(h(v), x^A(v))$ that maps the node representations (spatio-temporal features) and the corresponding node label sequence (cellular traffic history of the target cell tower) to an output $o(v)$ for each $v \in V$.

Propagation Step. The propagation step is an iterative procedure that propagates node representations. Initial node representations $h^{(1)}(v)$ are set to arbitrary values. Then each node representation is updated following the recurrence below until convergence, where i denotes the time-step, f^* denotes a function implemented by a multi-layer neural network

$$h^{(i)}(v) = f^*\Big(x^A(v), x^A(NBR(v)), x^B(CO(v)), h^{(i-1)}(NBR(v))\Big). \tag{2}$$

Output Model. An output model is defined per node and is a differentiable function $o(v) = g(h(v), x^A(v))$. In our work, we adopt graph-level regression for the output model instead of mapping output g for each node to emphasize the spatial dependency among cell towers. To train a GNN, all parameters (i.e., node representations h_v and output model g) are learned jointly using gradient-based optimization such as Almeida-Pineda algorithm[41], [42], which runs the propagation step to convergence and computes gradients based on the converged solution.

In this work, we train a GNN model for each cell tower. f^* is implemented by recurrent neural networks and g is implemented by full-connected neural networks. The model parameters will be carefully tuned for performance which will be explained in Section 6. To predict $x(v)$, we feed input $x^A(v)$, $x^A(NBR(v))$, $x^B(CO(v))$ into the GNN model. Since the traffic prediction model for each cell c only depends on a small set of nodes and edges, the training process for all cell towers can be easily distributed and paralleled.

Fig. 9 shows an illustrative example of the encoding network of our model to predict the cellular traffic for cell tower c_0. The figure on the left shows the graph representation of cell tower c_0 and its neighbours $\{c_1, c_2, c_3\}$. The in-tower traffic of each cell tower will be encoded into c_i and the inter-tower traffic will be encoded into the edges $\{e_1, e_2, e_3\}$. The figure on the right describes the GNN model. f^* and g are implemented by two-layer full-connected neural networks in this example. Specifically, node representations $\{h(c_i)\}$, i.e., the spatio-temporal features, will be learned for each cell c_i in the propagation step using f^* based on the in-tower traffic X^A and inter-tower traffic X^B of all the nodes and edges connected to c_i. Then the node representation $h(c_0)$, together with $X^A(c_0)$, will be fed into the output model g to get the prediction result $O(c_0)$, i.e., the predicted cellular traffic for c_0.

5.4 Event Inference

In addition to data traffic prediction, the decomposition of in-tower (x^A) and inter-tower (x^B) traffic also benefits mining network (or social) events.

We decompose the cellular traffic x into 3 explainable components: a seasonal component s capturing the periodic pattern; a trend component u capturing the offset from the periodic pattern in a period, and a residual component r capturing the instantaneous changes. That is, given a prediction x for cellular traffic using the GNN model, we rewrite $x = s + u + r$ with the ratio-to-moving-average method [38], and we conduct the decomposition for both the in-tower traffic and the inter-tower traffic, i.e.,

$$x^A = s^A + u^A + r^A$$

$$x^B = s^B + u^B + r^B.$$

In particular, r^A reflects the instantaneous traffic variation caused by mobility-independent online events, such as an online live show or an Internet failure, while r^B corresponds to the instantaneous traffic changes caused by mobility-dependent social events, e.g., a sports game, pop-music concerts, presidential campaign speech, large rallies, traffic jams, etc. Detecting such events will be beneficial for

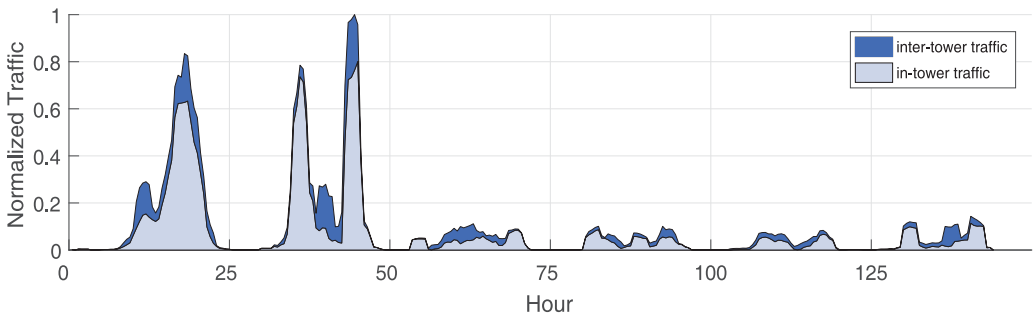
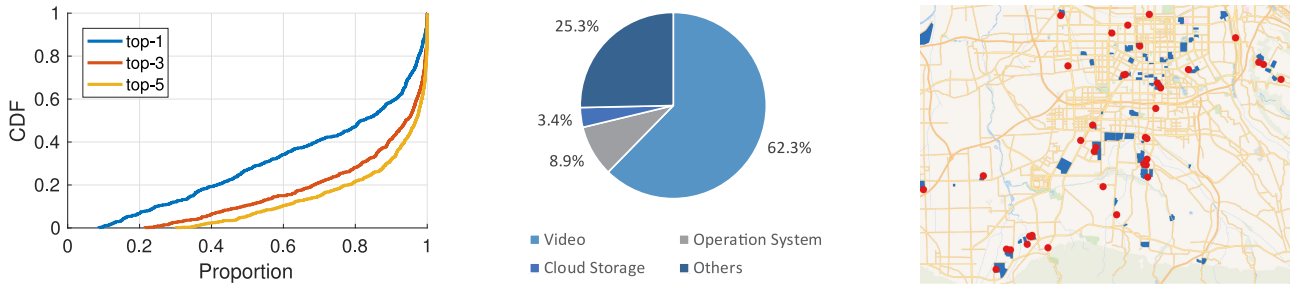


Fig. 10. An example study of cellular traffic based event inference. We detect the start of the online registration of a large activity at 12 p.m. and a people gathering event due to a concert at 4 p.m. on June 5th.



(a) (b) (c)

Fig. 11. (a) CDF of p_i. (b) Apps used by top-1 users. (c) The distribution of events and the locations of schools.

network troubleshooting and optimization as well as for urban governance and taking necessary prevention actions.

For mobility-independent online events, we take it as an event case when r^A is larger than a threshold (e.g., 1 GB per hour). Then we calculate the proportion of traffic p_i consumed by top-k users who use more traffic. The result in Fig. 11a shows that for almost 55 percent cases, only one user consumes more than 80 percent traffic and the number goes to 75 percent when considering top-5 users, illustrating that for most mobility-independent online events, the minority contributes the most of traffic. Further, we visualize the application used by the minority. Fig. 11b displays that 62.3 percent of mobility-independent online events dominated by the minority result from watching videos, 8.9 percent from operation system (i.e., system update) and 3.4 percent from Cloud Storage.

Mobility-independent online events caused by the minority is complex to analyze and predict because of complex individual behaviors. For this reason, we explore mobility-independent online events caused by group users. After filtering out the events caused by the minority, We collect POIs of the locations of these events. POI is a specific point location of a certain function such as restaurant and school. The distribution of an areas POI reflects its function. The POI data are available from the API of Baidu Map, which is one of the main map service providers. The POIs collection are categorized into, five main types of POI, including resident (16.6 percent), transport(6.4 percent), education (48.0 percent), business(15.3 percent) and entertainment(13.7 percent). The POIs distribution implies that it is much more possible for areas like schools to generate mobility-independent online events. Further, we visualize the locations of mobility-independent online events caused by group users on map as red dots and the school areas as blue blocks in Fig. 11c. As we can see, the locations of online events are highly related to the school regions spatially, which enlightens the carriers that more attention should be paid in school regions to avoid the network failure caused by mobility-independent online events.

For mobility-dependent social events, we show an example of event inference based on traffic patterns in Fig. 10. From the information of r^A and r^B, we can observe that the traffic patterns of both r^A and r^B are notably different in the first two days (0-48h). On June 5th (0-24h), 2016, r^B surged twice during morning and evening and the data volume reached 525 percent compared with normal days (the latter four days). Soon following r^B's changes, r^A also experienced large increases twice on June 5th, 2016, with a data traffic escalation of 3,930 percent. It can be observed that both r^A and r^B changed rapidly and it took less than half an hour from the starting to the ending of the dramatic traffic shift. This means the sampling interval of half an hour fails to capture the details of changes. This is one of the most important reasons why mobile traffic prediction is extremely challenging.

According to our previous analysis, r^B's changes can be attributed to user mobility. Thus, we infer that there were people gathering at 10 a.m. and 5 p.m. and people dispersing at 1 p.m. and 9 p.m. on June 5th, 2016. On the other hand, the volume of r^A shows that people accessed the Internet heavily between 10 a.m. to 1 p.m., and between 3 p.m. and 9 p.m., respectively. This observation indicates that there was either a hot online issue that many people followed through the Internet or a local event that people were willing to share on the Internet. Combining the information of both r^A and r^B, we believe that the second possibility is more likely.

Through thoroughly search on the Internet, we find that there was a "shake run" activity with participants signing in at 12 p.m. and a concert of two pop music stars starting at 4 p.m. on June 5th. This verifies our event inference.

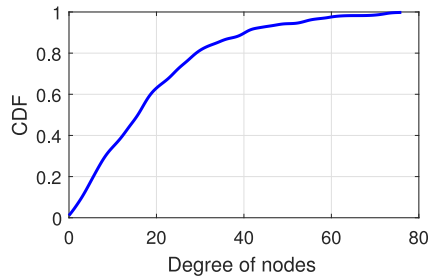


Fig. 12. The degree distribution of extracted graph from our cellular dataset.

We believe that event inference based on traffic patterns is valuable but challenging. The difficulty lies in the lack of ground-truth information of various social events at city-wide scale. The city where our dataset is collected is both an ancient and modern metropolis with a large population of more than 10 million. Numerous activities take place each day, which obstructs our search of events corresponding to traffic variation. It is common that we succeed in detecting abnormal traffic patterns but fail to verify our inference, because such activities may not be posted online, or its information hides in the ocean of Internet data and we are looking for a needle in a haystack. As a result, we are unable to conduct extensive evaluation for event inference at the current stage. However, the findings so far motivate us a lot and we will continue our exploration in this direction and leave it as a future work.

6 Evaluation

In this section, we evaluate the performance of our spatio-temporal prediction model and compare it with the state-of-the-art time-series based prediction approaches.

6.1 Experimental Settings

6.1.1 Dataset

We evaluate the performance of our cellular traffic prediction model on the dataset discussed in Section 3.2. It covers comprehensive cellular data usage traces of 1.5 million users monitored at 5,929 cell towers from June 5th to 18th, 2016. As we aim at a temporal resolution of half an hour, we aggregate the traffic at each cell tower every half an hour. We extract the dependency relationship G as Section 5 describes. Fig. 12 illustrates the degree distribution of G. From the figure, we can see that most cell towers (about 90 percent) have less than 40 neighbors. We choose data from the last two days as the testing data, and all data before that as training data.

6.1.2 Baselines

We compare the performance of our proposed *GNN* with **D**ecomposed Cellular Traffic model (*GNN-D*) with the following baselines.

NAIVE: The NAIVE method simply predicts the traffic at a certain time based on the traffic at the same time of the last day. For instance, its prediction for 15:00 - 15:30, June 17th, 2016 is the traffic volume for 15:00 - 15:30, June 16th, 2016.

ARIMA: Auto-Regressive Integrated Moving Average (ARIMA) model [38] is commonly used for modelling time series behaviours and has been widely adopted in time series prediction [43].

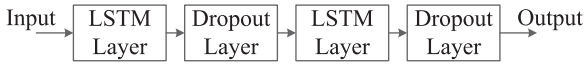


Fig. 13. Architecture of LSTM model used for evaluation.

LSTM: Long-Short Term Memory (LSTM) [39] is a Recurrent Neural Network (RNN) architecture. Unlike traditional RNNs, LSTM uses "gates" instead of activation functions, as a result, LSTM could maintain state as well as output. Since state can be more persistent than output, LSTMs are able to capture longer term memory in comparison with traditional RNNs.

HW: Holt-Winters (HW) is used for exponential smoothing to make short-term forecasts by using additive or multiplicative models with increasing or decreasing trend and seasonality.

GNN-A: We apply the *GNN* model with **A**ggregated cellular traffic (*GNN-A*) as a basic GNN model. In GNN-A model, we assign the entire cellular traffic without decomposition as node labels and don't absort edge labels into the model.

6.1.3 Metrics

We evaluate the performance of cellular traffic prediction based on two metrics: Mean Absolute Error (MAE) and Mean Absolute Relative Error (MARE)

$$MAE = \frac{1}{|T| \cdot |C|} \sum_{i,j} |x_i(c_j) - \hat{x}_i(c_j)| \tag{3}$$

$$MARE = \frac{1}{|T| \cdot |C|} \sum_{i,j} \frac{|x_i(c_j) - \hat{x}_i(c_j)|}{x_i(c_j)}, \tag{4}$$

where $|T|$ is the size of the testing set, $i \in [1, |T|]$ is the index of each testing sample, $j \in [1, |C|]$ is the ID of each cell, and $\hat{x}_i(c_j)$ denotes the prediction for $x_i(c_j)$.

We choose these two metrics to jointly evaluate the prediction performance of traffic volumes considering the wide dynamic range (from 10^4 bytes to 10^9 bytes) of our dataset.

6.1.4 Model Training

We implement the ARIMA model using the "forecast" R package [44]. The package automatically selects the best model parameters based on the given order constraints. We implement the LSTM model using the "Keras" python library [45]. Fig. 13 shows the structure of the LSTM model used in our evaluation. We carefully tune the model parameters and choose 5 neurons for each LSTM layer with a 0.2 dropout rate and linear activation function, and we use the history of the last 3 time units for the best performance. We implement HW model using the "stats" R package. We implement our GNN-A and GNN-D model using the GNN toolbox [46]. We choose two propagation layers and two output layers. Before training, features and targets will be normalized by max normalizer. We set the GNN parameters as $r = 2, l = 3$, and use 5 hidden neurons for each layer of the propagation model with *linear* activation function and 6 neurons for each layer of the output model with *tanh* activation function.

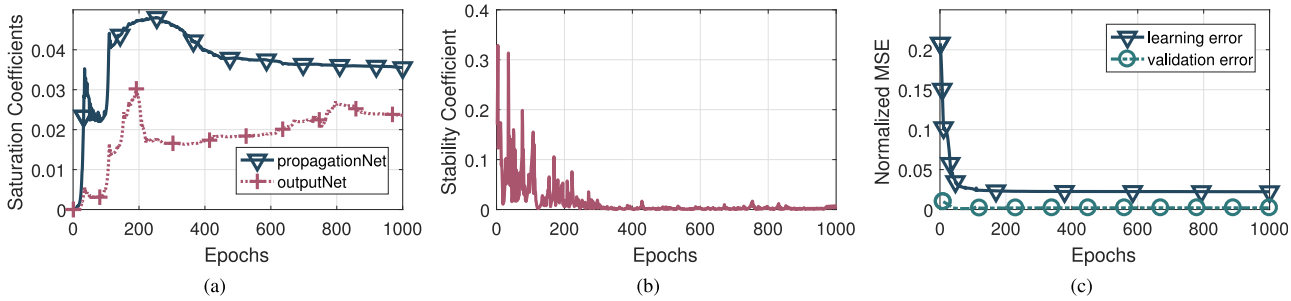


Fig. 14. Quantitative monitoring of the training process in 1,000 epochs. (a) Saturation coefficients of the models in both propagation and output steps. (b) Stability coefficient. (c) Normalized MSE showing the convergence of training.

To evaluate whether our GNN model has been trained properly, we monitor the training process (1,000 epochs) with qualitative metrics as shown in Fig. 14.

Fig. 14a plots the saturation coefficients during our training. Saturation coefficient is defined as the mean square error of the hidden layer's output in our experiment. For tanh layers, a hidden unit is saturated if the saturation coefficient is close to 1 or -1. For linear models, certain degree of saturation is essential, while for non-linear models, saturated layer will decrease both the information capacity and the learning ability of a neural network [47]. We monitor the saturation coefficients for both the propagation step (denoted as "propagationNet", linear) and for the output model (denoted as "outputNet", non-linear). As shown, the saturation coefficients for the propagationNet and the outputNet are both in the range $(0.01, 0.05)$, indicating the training process is nonmalignant.

Fig. 14b shows the stability coefficients during the training process. Stability coefficient measures the difference of outputs between successive epochs, i.e., $\frac{\sum |(o^{(t)} - o^{(t-1)}|}{\sum |o^{(t)}|}$, and indicates whether the training converges. As shown, the stability coefficient gradually drops to nearly zero after 400 epochs, indicating that the training process will converge.

Fig. 14c shows the normalized mean squared errors (Normalized MSE) on traffic prediction using the validation set. As shown, after about 80 epochs, both the learning error and the validation error converge, demonstrating the effectiveness of the training process.

6.2 Prediction Performance

Table 2 summarizes the traffic prediction performance of our GNN-D methods and the baselines. GNN-D consistently and significantly outperforms all the baselines in both metrics. Specifically, GNN-D achieves 62.2, 19.7, 16.3, 13.2, 18.5 percent smaller MAE than NAIVE, ARIMA, LSTM, HW and GNN-A, respectively, and demonstrates 43.7, 62.6, 17.5, 39.2, 14.3 improvement in MARE than NAIVE, ARIMA, LSTM, HW and GNN-A, respectively. Fig. 15 shows a comparison between prediction results and the ground truth of a sample region. We can see that the prediction results well match the trend of the actual values, which indicates the high performance of our method.

TABLE 2
Overall Prediction Performance

Method	MAE ($\times 10^7$ bytes)	MARE
NAIVE	2.784	1.402
ARIMA	1.309	2.114
LSTM	1.255	0.958
HW	1.210	1.301
GNN-A	1.289	0.922
GNN-D	1.051	0.790

For more insights on the applicability of our method, we investigate the impact of different factors on the cellular traffic prediction performance as follows.

Impact of Cellular Traffic Volume. Fig. 16 plots the cumulative distribution function (CDF) of per half-hour traffic volume in our dataset, which can be fitted by a lognormal mixtures distribution with three components. The Kolmogorov-Smirnov test (K-S test) is used to check for goodness-of-fit of empirical data to test distributions, which indicates that the distribution of log-normal mixtures is well accepted with p-value $= 0.58$. As shown, both light traffic ($< 10^6$ bytes) and heavy traffic ($> 10^9$ bytes) take up only a small percentage, but can be more difficult to predict accurately. To evaluate the prediction performance for different traffic volume, we divide the testing set into four subsets based on traffic volume levels: $[0, 10^6), [10^6, 10^7), [10^7, 10^8), [10^8, +\infty]$.

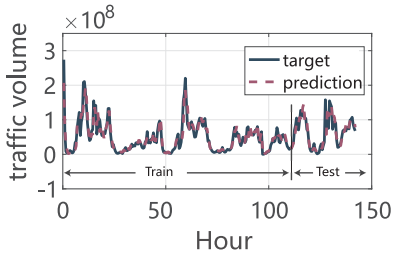


Fig. 15. Prediction vs. the ground truth for a sample cell tower.

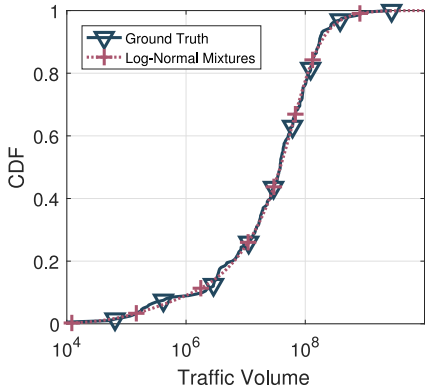


Fig. 16. PDF of half-hour data traffic in our dataset.

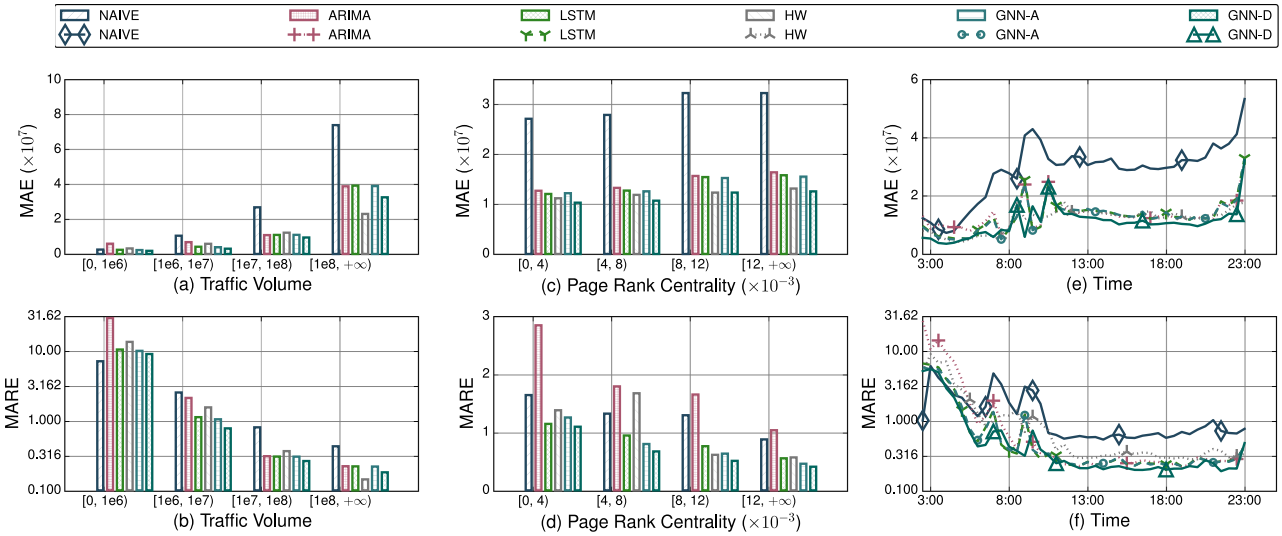


Fig. 17. Different perspectives on predicting performance.

Figs. 17a and 17b present the prediction performance for different traffic volume levels. We make the following observations from the results. (i) For all prediction methods, MAE increases with the traffic volume while MAPE decreases with the traffic volume. (ii) GNN-A performs similarly to LSTM. It could be explained that the propagation model in GNN-A is equivalent to a basic recurrent neural network. Without traffic decomposition, the neural network has similar encoding capacity for spatial time series to LSTM. (iii) GNN-D method outperforms the baselines in all levels of traffic volume. The results indicate that even with an imbalanced training dataset, our method still outperforms others in predicting extremely light or heavy traffic and it is applicable to traffic volumes spanning a wide dynamic range.

Impact of Spatial Dependency. This experiment demonstrates the prediction performance of our method for cell tower traffic with different levels of spatial dependency. We quantify the spatial dependency of cell towers using centrality, which is a graph term to measure the importance of vertices. There are various ways to define centrality such as closeness centrality, betweenness centrality, Eigenvector centrality and PageRank centrality [48]. PageRank centrality measures centrality by considering three distinct factors: (i) the number of links it receives, (ii) the link propensity of the linkers, and (iii) the centrality of the linkers. Thus PageRank centrality is the most suitable to measure the levels of spatial dependency of each cell tower. Fig. 18 shows the distribution of PageRank centrality in the main district of the city where our measurements were collected. Each point represents a cell tower and its color stands for the value of its PageRank centrality. We select three cell towers with high PageRank centrality, denoted by 'A', 'B' and 'C' in Fig. 18. After checking the city maps, we find that cell tower A is located at a railway station, cell tower B is within a major shopping mall, and cell tower C is in a university. All of them are busy and crowded locations with high user mobility. In all of the three cell towers, inter-tower traffic takes up a large portion of the total traffic volume and we expect high spatial dependency of their cellular traffic on the neighbouring cell towers.

After showing that PageRank centrality can act as an indicator for spatial dependency of cellular traffic, we plot the prediction performance of our method for cell towers within different PageRank centrality levels in Figs. 17c and 17d. As shown, with the increase of PageRank centrality (and thus spatial dependency), MAE increases slightly while MAPE decreases significantly for all the prediction schemes. Considering that cells with high spatial dependency usually have high traffic volume, the slightly increased MAE in fact leads us to conclude that our method is more accurate for highly spatial-dependent cells. And GNN-D consistently outperforms the baselines. For instance, for cell towers with high centrality ($\geq 1.2 \times 10^{-2}$), GNN-D is 23.3, 20.5, 12.6, 19.0 percent better than ARIMA, LSTM, HW and GNN-A in MAE, and 60.7, 26.2, 28.6, 11.2 percent better than ARIMA, LSTM, HW and GNN-A in MARE.

Temporal View of Prediction Errors. This experiment shows the temporal trend of the prediction errors of all the methods. Since both the traffic volume and its changing rate usually vary with time, the experiment demonstrates how the prediction accuracy is affected by traffic volume and traffic changing rate. Figs. 17e and 17f plot the MAE and MARE metrics during one day. In accord with our evaluations of the impact of traffic volume on the prediction performance, during busy hours in the morning (7:00, 9:00, 11:00), MAE increases while MARE decreases, as the traffic peaks at these times. One

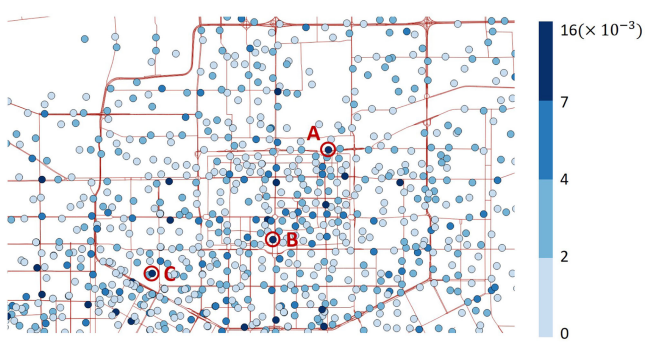


Fig. 18. Spatial distribution of PageRank centrality.

exception is that at 10:00, there is a valley in cellular traffic, and it is expected that for all the methods, MAE will decrease and MARE will increase. However, both ARIMA and LSTM have large MAE, indicating that they fail to predict even modest traffic volume. A closer look at the traffic pattern around 10:00 shows that the cellular traffic changes rapidly around 10:00, suggesting intensive mobility, and thus strong mobility. As ARIMA, LSTM, HW and GNN-A fail to account for spatial dependency of cellular traffic, they naturally perform poorly when there is intensive mobility. In contrast, GNN-D encodes the spatial dependency into the graph model, so that it can still predict the fast-changing cellular traffic accurately and achieve small MAE.

7 Conclusion

Motivated by the decomposition of in-tower and inter-tower data traffic, we model the spatio-temporal features of traffic patterns in a metropolis by a directed graph and propose a powerful deep learning approach that can learn from a graph structure. We achieve large-scale and fine-grained prediction based on a big data set of cellular network records collected by a mobile carrier.

Experiment results show that the spatial dependency and the interaction of spatial and temporal factors play an important role in accurate and robust prediction. Our findings also include how their interaction can be used for event inference through example study. In the future, along with the explosive growth of mobile Internet and continuously evolving traffic patterns, known models are outdated and new opportunities are sprouting.

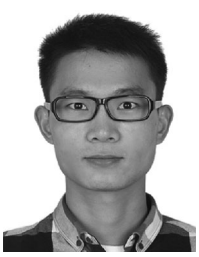
Acknowledgments

This work is supported in part by the National Key Research Plan under grant No. 2016YFC0700100, NSFC under grant 61832010, 61332004, 61572366, 61472057, and The Nature Science Foundation of Jiangsu for Distinguished Young Scientist under Grant BK20170039. A preliminary version of this article appeared in the IEEE International Conference on Network Protocols (IEEE ICNP 2017).

References

[1] C. V. N. I. Cisco, "Global mobile data traffic forecast update, 2015–2020 white paper," 2016.
[2] M. Z. Shafiq and L. E. Ji, "Characterizing and modeling internet traffic dynamics of cellular devices," in *Proc. ACM SIGMETRICS Joint Int. Conf. Meas. Model. Comput. Syst.*, 2011, pp. 305–316.
[3] U. Gotzner and R. Rathgeber, "Spatial traffic distribution in cellular networks," in *Proc. Veh. Technol. Conf.*, 1998, pp. 1994–1998.
[4] H. Wang and J. E. Ding, "Characterizing the spatio-temporal inhomogeneity of mobile traffic in large-scale cellular data networks," in *Proc. Int. Workshop Hot Topics Planet-Scale Mobile Comput. Online Social Netw.*, 2015, pp. 19–24.
[5] D. E. Lee, "Spatial modeling of the traffic density in cellular networks," *IEEE Wireless Commun.*, vol. 21, no. 1, pp. 80–88, Feb. 2014.
[6] Y. Zang, F. Ni, Z. Feng, S. Cui, and Z. Ding, "Wavelet transform processing for cellular traffic prediction in machine learning networks," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, 2015, pp. 458–462.
[7] J. Wang, J. Tang, Z. Xu, Y. Wang, G. Xue, X. Zhang, and D. Yang, "Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach," in *Proc. IEEE INFOCOM*, May 2017, pp. 1–9.
[8] X. Chen, Y. Jin, S. Qiang, W. Hu, and K. Jiang, "Analyzing and modeling spatio-temporal dependence of cellular traffic at city scale," in *Proc. IEEE Int. Conf. Commun.*, 2015, pp. 3585–3591.
[9] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube traffic characterization: A view from the edge," in *Proc. Conf. Internet Meas. Conf.*, 2007, pp. 15–28.
[10] J. Cao and W. S. E. Cleveland, "Stochastic models for generating synthetic http source traffic," in *Proc. IEEE INFOCOM*, 2004, pp. 1546–1557.
[11] C. Sarraute, P. Blanc, and J. Burroni, "A study of age and gender seen through mobile phone usage patterns in mexico," in *Proc. Int. Conf. Advances Social Netw. Anal. Mining*, 2014, pp. 836–843.
[12] D. Willkomm, et al., "Primary users in cellular networks: A large-scale measurement study," in *Proc. Int. Symp. Dyn. Spectr. Access Netw.*, 2008, pp. 1–11.
[13] P. Zerfos, X. Meng, S. H. Wong, V. Samanta, and S. Lu, "A study of the short message service of a nationwide cellular network," in *Proc. Conf. Internet Meas. Conf.*, 2006, pp. 263–268.
[14] A. Noulas, et al., "Exploiting foursquare and cellular data to infer user activity in urban environments," in *Proc. Int. Conf. Mobile Data Manage.*, 2013, pp. 167–176.
[15] Y. Li and C. E. Peng, "Mobileinsight: Extracting and analyzing cellular network information on smartphones," in *Proc. Annu. ACM Int. Conf. Mobile Comput. Netw.*, 2016, pp. 202–215.
[16] B. Guo, Q. Han, H. Chen, L. Shangguan, Z. Zhou, and Z. Yu, "The emergence of visual crowdsensing: Challenges and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2526–2543, Oct.–Dec. 2017.
[17] X. Zhang, Z. Yang, W. Sun, Y. Liu, S. Tang, K. Xing, and X. Mao, "Incentives for mobile crowd sensing: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 54–67, Jan.–Mar. 2016.
[18] X. Zhang, Z. Yang, Z. Zhou, H. Cai, L. Chen, and X. Li, "Free market of crowdsourcing: Incentive mechanism design for mobile sensing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 12, pp. 3190–3200, Dec. 2014.
[19] Y. Jin and N. E. Duffield, "Characterizing data usage patterns in a large cellular network," in *Proc. ACM SIGCOMM Workshop Cellular Netw.: Operations Challenges Future Des.*, 2012, pp. 7–12.
[20] Y. Zhang and A. Årvidsson, "Understanding the characteristics of cellular data traffic," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 4, pp. 461–466, 2012.
[21] F. Xu, P. Zhang, and Y. Li, "Context-aware real-time population estimation for metropolis," in *Proc. Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 1064–1075.
[22] H. Wang, F. Xu, Y. Li, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in Urban environment," in *Proc. Conf. Internet Meas. Conf.*, 2015, pp. 225–238.
[23] R. Li, Z. Zhao, J. Zheng, C. Mei, Y. Cai, and H. Zhang, "The learning and prediction of application-level traffic data in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3899–3912, Jun. 2017.
[24] S. Chinchali, P. Hu, T. Chu, M. Sharma, M. Bansal, R. Misra, M. Pavone, and K. Sachin, "Cellular network traffic scheduling with deep reinforcement learning," *AAAI Conf. Artif. Intell.*, 2018, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16638
[25] V. Sciancalepore and K. E. Samdanis, "Mobile traffic forecasting for maximizing 5G network slicing resource utilization," in *Proc. IEEE INFOCOM*, 2017, pp. 2583–2591.
[26] C. Zhang and P. Patras, "Long-term mobile traffic forecasting using deep spatio-temporal neural networks," in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2018, pp. 231–240.
[27] C. Zhang, X. Ouyang, and P. Patras, "ZipNet-GAN: Inferring fine-grained mobile traffic patterns via a generative adversarial neural network," in *Proc. 13th Int. Conf. Emerging Netw. Exp. Technol.*, 2017, pp. 363–375.
[28] C. Wu, Z. Yang, and C. Xiao, "Automatic radio map adaptation for indoor localization using smartphones," *IEEE Trans. Mobile Comput.*, vol. 17, no. 3, pp. 517–528, Mar. 2018.
[29] L. Shangguan, Z. Yang, A. X. Liu, Z. Zhou, and Y. Liu, "STPP: Spatial-temporal phase profiling-based method for relative RFID tag localization," *IEEE/ACM Trans. Netw.*, vol. 25, no. 1, pp. 596–609, Feb. 2017.
[30] Q. Xu, A. Gerber, Z. M. Mao, and J. Pang, "AccuLoc: Practical localization of performance measurements in 3G networks," in *Proc. 9th Int. Conf. Mobile Syst. Appl. Serv.*, 2011, pp. 183–196.
[31] A. Sridharan and J. Bolot, "Location patterns of mobile users: A large-scale study," in *Proc. IEEE INFOCOM*, 2013, pp. 1007–1015.

[32] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, 2008, Art. no. 779.
[33] A. K. Das, P. H. Pathak, C.-N. Chuah, and P. Mohapatra, "Contextual localization through network traffic analysis," in *Proc. IEEE INFOCOM*, 2014, pp. 925–933.
[34] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, "Exploring human mobility with multi-source data at extremely large metropolitan scales," in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw.*, 2014, pp. 201–212.
[35] Y. Wang, N. J. Yuan, D. Lian, L. Xu, X. Xie, E. Chen, and Y. Rui, "Regularity and conformity: Location prediction using heterogeneous mobility data," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1275–1284.
[36] X. Ouyang, C. Zhang, P. Zhou, and H. Jiang, "DeepSpace: An online deep learning framework for mobile big data to understand human mobility patterns," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1610.07009
[37] H. Li, C. A. Calder, and N. Cressie, "Beyond Moran's I: Testing for spatial dependence based on the spatial autoregressive model," *Geographical Anal.*, vol. 39, no. 4, pp. 357–375, 2007.
[38] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley, 2015.
[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[40] F. Scarselli, et al., "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.
[41] L. B. Almeida, "A learning rule for asynchronous perceptrons with feedback in a combinatorial environment," in *Artificial Neural Networks*. Piscataway, NJ, USA: IEEE Press, 1990, pp. 102–111.
[42] F. J. Pineda, "Generalization of back-propagation to recurrent neural networks," *Phys. Rev. Lett.*, vol. 59, no. 19, 1987, Art. no. 2229.
[43] J. D. Hamilton, *Time Series Analysis*, vol. 2. Princeton, NJ, USA: Princeton Univ. Press, 1994.
[44] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: The forecast package for R," *J. Statistical Softw.*, vol. 26, no. 3, pp. 1–22, 2008.
[45] F. Chollet, "Keras," 2015. [Online]. Available: https://github.com/fchollet/keras
[46] W. Uwents, G. Monfardini, H. Blockeel, M. Gori, and F. Scarselli, "Neural networks for relational learning: An experimental comparison," *Mach. Learn.*, vol. 82, no. 3, pp. 315–349, 2011.
[47] A. Rakitianskaia and A. Engelbrecht, "Measuring saturation in neural networks," in *Proc. IEEE Symp. Series Comput. Intell.*, 2015, pp. 1423–1430.
[48] P. Bonacich, "Power and centrality: A family of measures," *Amer. J. Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

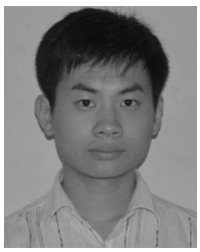


Xu Wang received the BE degree from the School of Software, Tsinghua University, in 2015. He is currently working toward the PhD degree in the School of Software, Tsinghua University. He is a member of the Tsinghua National Lab for Information Science and Technology. His research interests include mobile computing and machine learning. He is a student member of the IEEE.

Zimu Zhou received the PhD degree from the Hong Kong University of Science and Technology, in 2015. He is currently a post-doc researcher with ETH Zurich. His research interests include sensor, mobile, and ubiquitous computing systems. He is a member of the ACM and the IEEE.

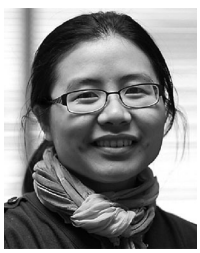
Fu Xiao received the PhD degree in computer science and technology from the Nanjing University of Science and Technology, Nanjing, China, in 2007. He is currently a professor and a PhD supervisor with the School of Computers, Nanjing University of Posts and Telecommunications. His main research interests include wireless sensor networks and mobile computing. He is a member of the IEEE Computer Society and the Association for Computing Machinery.

Kai Xing received the MS and PhD degrees in computer science from George Washington University, Washington, DC, in 2006 and 2009, respectively. He is currently an associate professor with the Department of Computer Science and Technology, University of Science and Technology of China, Hefei, China. His current research interests include cyber physical networking systems, wireless networks, mobile computing, mesh, ad hoc and sensor networking, in-network information processing, and network security. He is a member of the IEEE and ACM.



Zheng Yang received the BE degree in computer science from Tsinghua University, in 2006, and the PhD degree in computer science from the Hong Kong University of Science and Technology, in 2010. He is currently a professor with Tsinghua University. His main research interests include wireless ad-hoc/sensor networks and mobile computing. He is a senior member of the IEEE.

Yunhao Liu received the BS degree from the Automation Department, Tsinghua University, the MA degree from Beijing Foreign Studies University, China, and the MS and PhD degrees from Computer Science and Engineering Department, Michigan State University. He is now the Chang-Jiang professor with Tsinghua University. His research interests include sensor network and IoT, localization, RFID, distributed systems, and cloud computing. He is a fellow of the ACM and the IEEE.



Chunyi Peng received the PhD degree in computer science from the University of California, Los Angeles, in 2013. She was an associate researcher with Microsoft Research Asia. She is currently an assistant professor with the Department of Computer Science and Engineering, Purdue University. Her research interests focus on mobile networks, mobile sensing systems, wireless networking, and network security. She is a senior member of the IEEE.