

Mobile Cellular Big Data: Linking Cyberspace and Physical World with Social Ecology

Fengli Xu, Yong Li, *Member, IEEE*, Min Chen, *Senior Member, IEEE*, and Sheng Chen, *Fellow, IEEE*

Abstract

Understanding mobile big data inherent within large-scale cellular towers in urban environment is extremely valuable for service providers, mobile users, and government managers of modern metropolis. By extracting and modeling the mobile cellular data associated with over 9,600 cellular towers deployed in a metropolitan city of China, this article aims to link the cyberspace and physical world with the social ecology via such big data. We first extract human mobility and cellular traffic consumption trace from the dataset, and then investigate human behaviour in the cyberspace and physical world. Our analysis reveals that human mobility and the consumed mobile traffic have strong correlations and both have distinct periodical patterns in time domain. In addition, both human mobility and mobile traffic consumption are linked with the social ecology, which in return helps us to better understand human behaviour. We believe that the proposed big data processing and modeling methodology, combined with the empirical analysis on the mobile traffic, human mobility and social ecology, paves the way toward a deep understanding of the human behaviors in large-scale metropolis.

Index Terms

Mobile big data, mobile cellular data, mobile networks, cyberspace, social ecology

F. Xu is with School of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan 430074, China (E-mail: fenglixu@mails.hust.edu.cn).

Y. Li is with Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (E-mail: liyong07@tsinghua.edu.cn).

M. Chen is with School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China (E-mail: minchen2012@hust.edu.cn).

S. Chen is with School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK (E-mail: sqc@ecs.soton.ac.uk), and also with King Abdulaziz University, Jeddah 21589, Saudi Arabia.

I. INTRODUCTION

The past few years have seen a dramatic growth in mobile traffic, contributed by billions of mobile devices as the first-class citizens of the Internet. The global cellular network traffic from mobile devices is expected to surpass 24 exabytes (an exabyte is approximately equal to 10^{18} bytes) per month by 2019 [1], which is 9 times larger than the traffic served by the existing cellular network in 2014. Such a huge volume of mobile traffic forms a large-scale mobile big data recording human's activities in the physical world, behaviours in the cyberspace, and interactions with the urban social ecology. Here, social ecology refers to the complex relationship between human behaviors and urban environments. More specifically, in this article, the study of social ecology is carried out through investigating urban functional regions, such as transport hub, business district, shopping mall, residential area, etc. Therefore, while we are embracing a world with ambient cellular connectivity, there is also a critical and challenging problem – how to understand the patterns of data traffic in cyberspace and human mobility in physical world profoundly [2]–[5], especially their inherent relationship.

On a more practical note, understanding the hidden patterns of human's activities and behaviors in the cyberspace, physical world and social ecology in large-scale urban environment is extremely valuable for service providers, mobile users, and government managers of modern cities [6], [7]. If the traffic patterns of cellular network can be *identified and modeled*, the service provider can exploit the modeled traffic patterns and customize a strategy for its individual cellular tower for providing services, instead of using a same uniform strategy, such as using the same load balancing and data pricing algorithms on each tower. Mobile users also benefit from the traffic modeling because they can then choose towers with predicted lower traffic and enjoy better services. More profoundly, management departments of government will benefit from such mobile big data analysis as well because they may infer the social ecology and human economy activities by interpreting these data recorded by the mobile networks [8].

On the other hand, understanding human's behaviors in cyberspace and physical world as well as their interaction with social ecology by analyzing the mobile big data is challenging for three reasons. First, the recorded data experienced by thousands of cellular towers deployed in large-scale modern cities is highly complicated and hard to analyze. For example, our measurement includes over 9,600 cellular towers and 150,000 subscribers, where lots of redundant and

conflicting logs are observed. To identify patterns and behaviors embedded in the data associated with thousands of cellular towers, designing a system that is able to clean and handle large-scale big data is needed. Second, we do not have a priori human behaviour patterns in cyberspace and physical world. Without these profiles of human behaviour patterns, it is challenging to group huge amount of the data experienced by thousands of cellular towers into a small number of meaningful patterns, which are vital for further understanding human behaviours. Third, the traffic of a cellular tower is affected by many factors, such as time and location, etc. These factors often correlate with each other and further complicate analysis task. For example, significant traffic variation is observed at both fine-grained (hours) and coarse-grained (days) time scale, and across towers deployed in different locations [9]. By addressing these challenges, in this article, we investigate how to extract and model the user behaviors and patterns embedded in thousands of cellular towers in a large-scale urban environment via a credible dataset collected by one of the largest commercial mobile operators.

Our main contribution comprises three parts. First, we reveal people's behaviour patterns in cyberspace and physical world, in terms of traffic consumption and human mobility pattern, respectively. Specifically, we find out that cyberspace traffic consumption and physical world human mobility have temporal patterns and are tightly correlated with each other. Second, we link the cyberspace and physical world with the social ecology by first detecting the key mobility patterns embedded in the dataset and then investigating their links with urban functional regions. Third, with the established link between the cyberspace and physical world and the social ecology, we find that the average traffic-consuming rate and human migration pattern are correlated with the social ecology. More importantly, we further analyze the characteristics of human behavior in different urban functional regions, which deepens our understanding of human behaviours in large-scale urban environment.

The rest of this article is organized as follows. The introduction of the mobile big data investigated and the required preprocessing techniques employed are first presented, followed by an overall visualization of the temporal features of the mobile big data. Then, how people behave in the cyberspace and physical world is investigated, and we further our understanding of people's behaviours in the cyberspace and physical world by linking them with the social ecology. Finally, the last section summarizes our study and discusses the future works.

II. DATASET, PREPROCESSING AND OVERALL VISUALIZATION

This section provides the detailed information of the mobile big data investigated and the preprocessing needed. In addition, we visualize the temporal distribution of cellular traffic and subscribers, which benefits analysis.

A. Dataset Description

Our utilized mobile big data is an anonymized cellular trace collected by one of the largest mobile service providers in Shanghai, during the whole month of August, 2014. The trace contains the detailed mobile data usage record of 150,000 users, and each entry in the trace includes the identify ID of device (anonymized), start-end time of data connection, base station (BS) ID, address of BS, and the amount of 3G or LTE data consumed in each connection. The trace logs 1.96 billion tuples of the described information, contributed by over 9,600 BSs, which contains the traffic logs of 2.8 petabytes (petabyte = 10^{15} bytes) cellular data traffic, 92 terabytes (terabyte = 10^{12} bytes) per day and 7 gigabytes (gigabyte = 10^9 bytes) per BS on average. This large-scale and fine-grained dataset ensures that our human behavior analysis and modeling is credible.

B. Preprocessing

The trace collected by the service provider needs to be preprocessed because of the existence of redundant and conflicting traffic logs as well as the incomplete information of BSs' locations. The preprocessing includes three steps. First, the redundant and conflicting logs is eliminated, such as the identical traffic logs, caused by technical issues. Second, to solve the problem of incomplete information, we convert the addresses of BSs to their geographical longitudes and latitudes through APIs provided by online map service. This conversion gives us the precise geographical location of each BS, which is important for analyzing the ground truth of urban functional regions. The last step of preprocessing is segmenting the 31-day traffic trace of a tower into thousands of chunks each of which contains 10-minutes traffic log. The 10-minutes segmentation is chosen because it is the smallest time interval that a cellular tower can experience non-zero traffic.

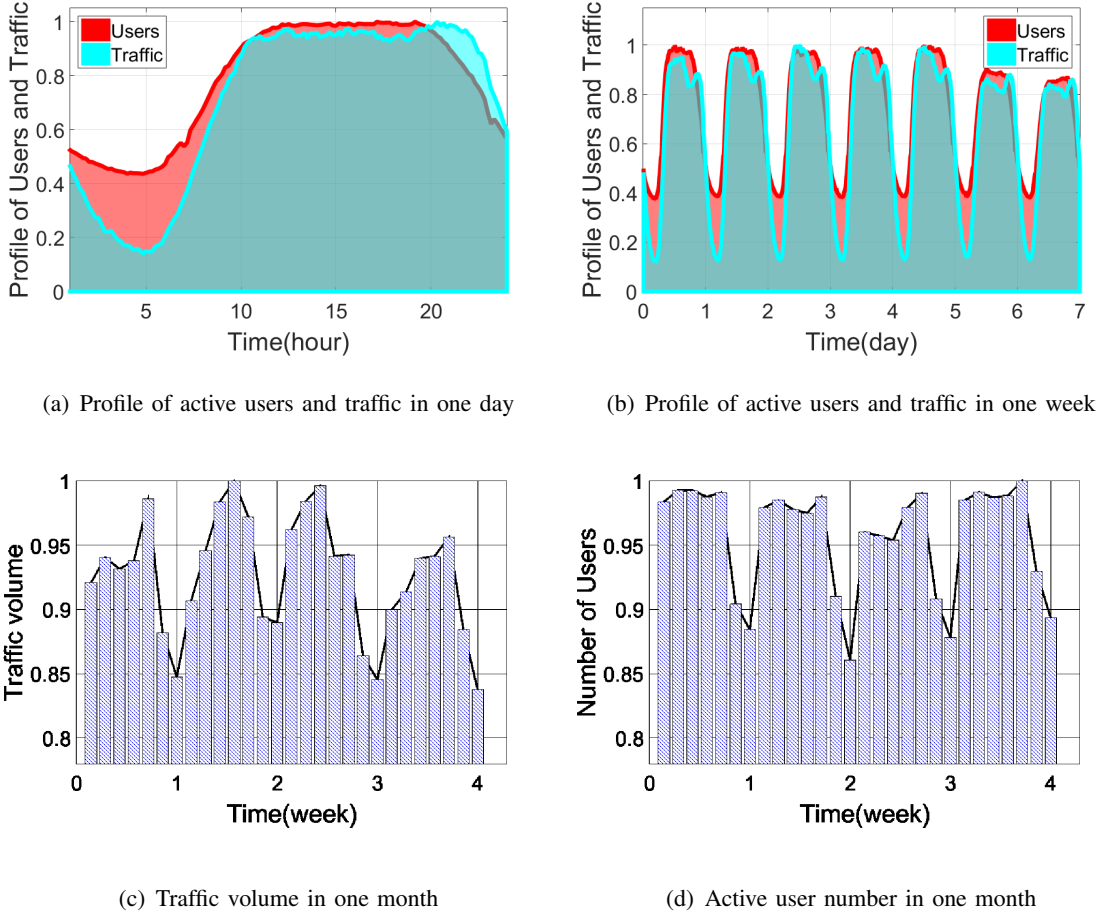


Fig. 1. Variation of the normalized traffic and the number of active users at different time scales.

C. Data Visualization

Before diving into a deep analysis of mobile data traffic, the visualization is first displayed for the distributions of the temporal traffic and number of active users provided by the 9,600 BSs, from which two interesting observations can be concluded.

First, the data embeds the fundamental temporal patterns of mobile data traffic. Fig. 1 shows the aggregated and normalized traffic and the number of active users at different time scales. More specifically, Fig. 1 (a) depicts the profile of normalized traffic and number of active users in one day (August 7, 2015, Thursday), where the aggregated network traffic looks similar with the profile of active users, and both are tightly coupled with the sleep patterns of humans, namely, high cellular traffic and large number of active users are observed during the day and low volumes of them are experienced during midnight. Fig. 1 (b) shows the profile of normalized

traffic and number of active users over one week (August 3-9, 2015). In addition to the repeated daily patterns of Fig. 1 (a), we observe from Fig. 1 (b) that the peak traffic and the number of active users at weekend are lower than those on a weekday. This suggests that mobile users are less active during the weekend and consume less cellular data traffic, which has also been found in [3]. Fig. 1 (c) illustrates the traffic distribution over the month (August 3-31, 2015), which shows that the traffic exhibits a periodical pattern on the order of a week, and weekend traffic is less than weekday traffic. Fig. 1 (d) depicts the temporal patterns of the number of active mobile users over the month. On one hand, the profile of number of active users given in Fig. 1 (d) exhibits similar patterns with the cellular traffic profile shown in Fig. 1 (c). On the other hand, the number of users is more stable during weekday than the cellular traffic, which indicates that the number of active users in Shanghai does not vary significantly during different weekdays.

III. HUMAN BEHAVIOURS IN CYBERSPACE AND PHYSICAL WORLD

In this section, human mobility and cellular traffic consumption patterns are investigated, respectively, in order to understand human behaviours in the cyberspace and physical world. In addition, the relationships between human mobility in physical world and traffic consumption patterns in cyberspace are further analyzed to provide insights of the link between the cyberspace and physical world.

A. Human Mobility in the Physical World

Human mobility is an important topic, which has been extensively studied in the past decade [10], [11]. With our mobile big data, mobile users can be located by checking the locations of the BSs that they are connected to. Therefore, it provides fine-grained location of large-scale mobile users, which is ideal for studying human mobility. In particular, human mobility can be studied through investigating dynamic distribution of mobile user population, which offers a new angle aiding our understanding of how human move in urban environment at a macro scale.

The bottom-half two plots of Fig. 2 (a) and (b) show the spatial distributions of number of active mobile users at two different times of a day (August 7, 2015, Thursday). From these results, it can be seen that the city center possesses the highest density of mobile users. Also we can see that the spatial distribution of mobile users varies with time during a day. In particular, the spatial distribution at 4 AM is different with that in 4 PM. This suggests that human mobility

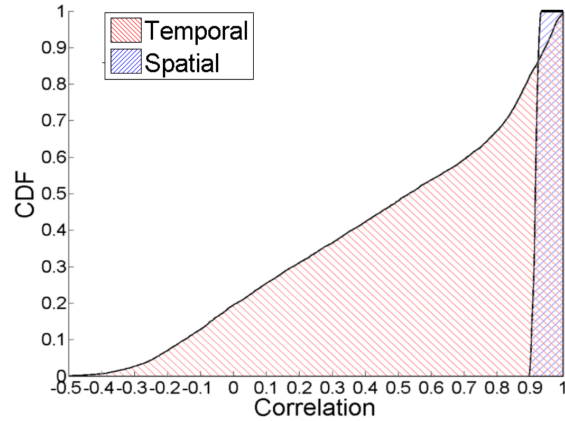
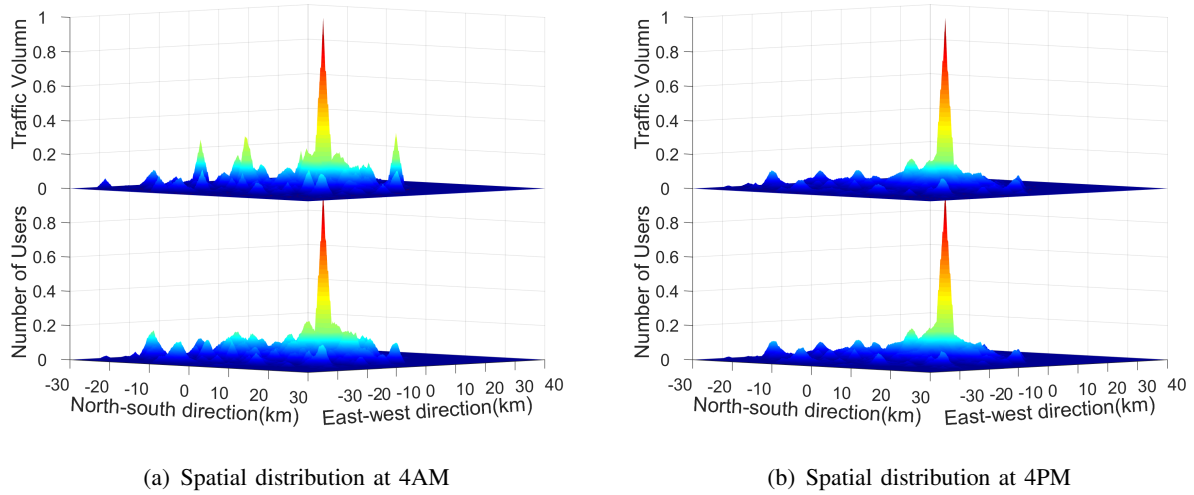


Fig. 2. The spatial distribution of normalized traffic consumption and number of active users, and the CDF of their correlation.

patterns vary during a day in urban environment, which is probably governed by activity patterns of humans.

B. Traffic Consumption Patterns in the Cyberspace

Understanding the traffic consumption patterns in urban environment is of great importance for cellular network load balancing, green operating and smart pricing. With the help of the traffic logs recorded in our big dataset, we are able to analyze such patterns in urban environment up to the period of one month. Let us specifically study the spatial distribution of cellular traffic during a chosen day (August 7, 2015, Thursday).

The top-half two plots of Fig. 2(a) and (b) present the spatial distributions of normalized

cellular traffic at two different times. From these two figures, we observe that the highest traffic consumption rate always occurs in the city center for different times, which is probably associated with the highest density of mobile users at the city center. Furthermore, from Fig. 2 (b), we can see that the spatial distribution of cellular traffic is similar to the spatial distribution of mobile users at 4 PM. However, observe from Fig. 2 (a) that the cellular traffic's spatial distribution is different from the distribution of mobile users at 4 AM. This indicates that the traffic consumption is not only correlated with the number of users, but also affected by other factors, such as traffic demand.

C. Relationship Analysis

From Fig. 2 (a) and (b), it can be seen that the traffic consumption is correlated with the number of users. Understanding the correlations between them will help us better understanding human's behaviour in the physical world and cyberspace. Therefore, we analyse and quantify the correlations between human mobility and cellular traffic patterns.

To understand the relationship between traffic consumption and number of users, the cumulative distribution functions (CDFs) of the spatial and temporal correlations between them are analyzed, respectively, which are presented in Fig. 2 (c). The spatial correlation is derived by computing the spearman correlation coefficient at each time slot, while the temporal correlation is computed on each BS. Observing the results of Fig. 2 (c), the number of users and the traffic consumption rate has a strong correlation in spatial domain, with most time slots having a correlation coefficient larger than 0.9. This suggests that at every time slot an area with more number of users is very likely to have higher traffic consumption. By contrast, the number of users and the traffic consumption exhibits a surprisingly low correlation in temporal domain, with about 20% of the BSs having negative correlation and about 40% of the BSs having a correlation coefficient lower than 0.4. This implies that in 50% of the BSs the number of users has a weak correlation with the traffic consumption rate.

IV. LINKING WITH SOCIAL ECOLOGY

Based on the above analysis, we have some basic understanding of human behaviour in the cyberspace and physical world in urban environment. A natural question to ask is does human behaviour relate to the social ecology? To answer this question, the links between the physical

world and social ecology is established by detecting the key patterns of human mobility. Then, with the knowledge of the social ecology, we deepen our understanding of human behaviour in the cyberspace and physical world.

A. *Discovering the Links with Social Ecology*

Discovering the links between the physical world, cyberspace and social ecology is nontrivial, because we have little knowledge of the relationships among them. However, inspired by a key observation that the human mobility patterns of the same geographical context tend to be similar, we implement and evaluate a system to discover the links by detecting the key mobility patterns.

1) *Detecting Key Mobility Patterns:* Our system is composed by three key elements: data cleaner, pattern identifier and metric tuner.

Data Cleaner: Data cleaner is a distributed traffic analysis system implemented in Hadoop, which is able to tackle large-scale unstructured mobile big data. The key of designing the data cleaner is a parallel transformer, which takes the time-domain logs of thousands of cellular towers as its input and converts each cellular tower's logs into a vector. A vector is constructed in two phases — aggregation and folding. In the first phase, each BS's number of users is aggregated in each 10-minutes time slot to generate a vector representing its user number pattern. Then, cellular towers' user number patterns of a month are converted into the patterns of week (7 days) by averaging. The purpose of averaging is smoothing burst events experienced by cellular towers, such as parade.

Pattern Identifier: Pattern identifier takes the vectorized data from the cleaner and runs a unsupervised machine learning algorithm to identify the key patterns of human mobility. The pattern identifier addresses one key challenge of the mining process — unknown patterns, by exploiting hierarchical clustering [12]. The basic idea of hierarchical clustering is iteratively merging the nearest two clusters. It first considers each input point as a cluster and then bottom-up iteratively merges the nearest two clusters until the stop condition is met. In our application, correlation distance is utilized as distance metric and the distance between clusters is defined as average-linkage distance. In addition, a threshold value is set as the stop condition, which stops clustering when the distance between every pair of clusters is above the threshold value. To be more specific, the pattern identifier operates in the following three steps. Firstly, it receives the predefined threshold value, takes the vectorized data as input and considers each cell tower's

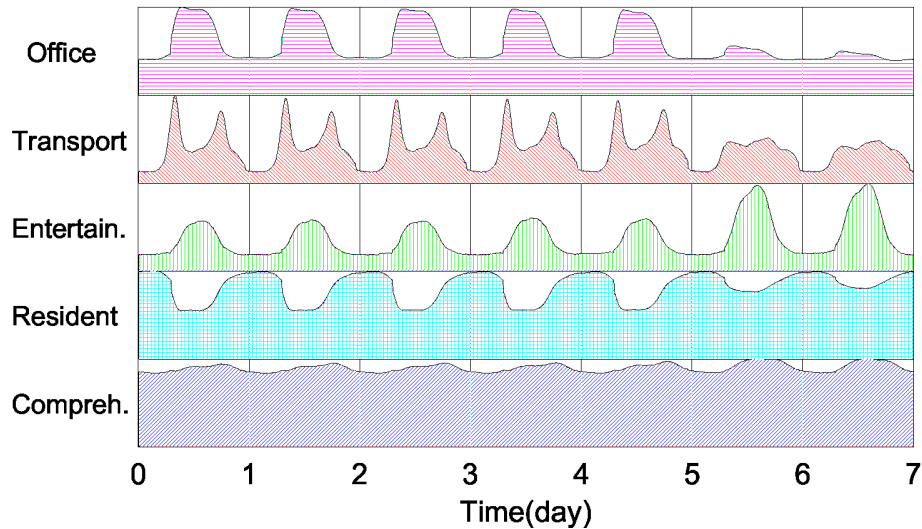


Fig. 3. Patterns of number of active users for the five identified clusters.

data as a cluster. Secondly, it calculates the distances for all pairs of clusters. Thirdly, it finds the minimum distance from the set of all the distances and compares it with the threshold value. If the minimum distance is above the threshold, the clustering is stopped, and the number of clusters gives the number of patterns identified while the average pattern of every cluster is outputted as the identified pattern for each cluster. Otherwise, it merges the nearest two clusters and return to the second step.

Metric Tuner: As the patterns of user variation is unknown, a key question is when should the identifier stop its clustering. In our system, Davies-Bouldin index is utilized [13] to explicitly inform the identifier that the optimum number of patterns has been identified. Davies-Bouldin index is utilized because it measures both the separation of clusters and cohesion within clusters, which mathematically guarantees good clustering result. When a minimum Davies-Bouldin index is obtained, the optimum number of patterns is identified.

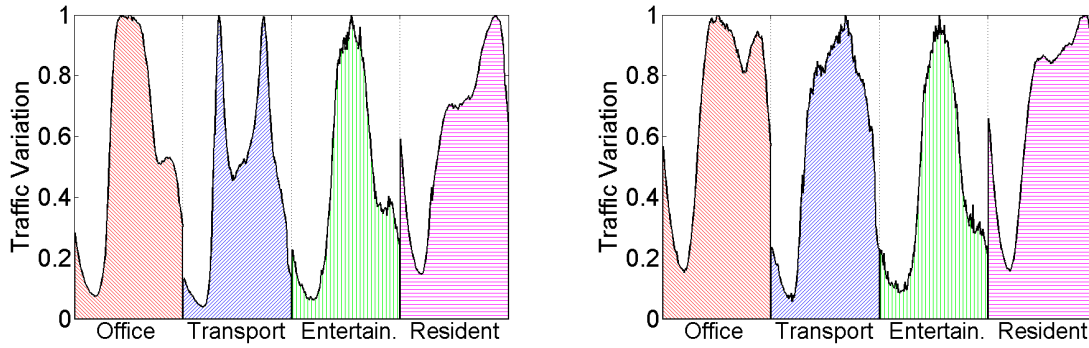
Fig. 3 shows the five time-domain patterns identified by our system from over 9,600 cellular towers. The five clusters differ in terms of the time where peak of user number appears as well as the number of users experienced during weekdays and weekend. The percentage of each cluster's cellular towers is shown in the left-hand part of Table I, which indicates that the first cluster has most cellular towers and the second cluster has the least.

TABLE I
 PERCENTAGE AND AVERAGED NORMALIZED POINTS OF INTEREST OF CELLULAR TOWERS CLASSIFIED IN EACH
 CLUSTER.

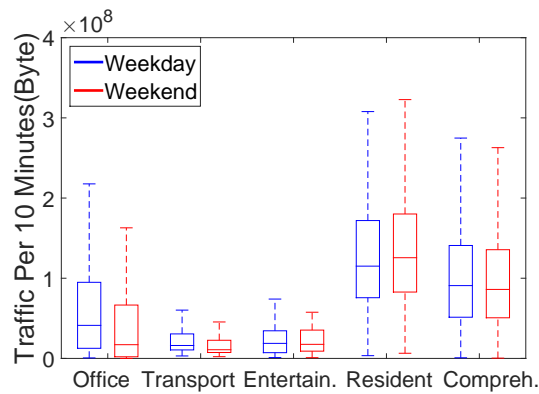
Functional Regions	Cluster Index	Percentage	Points of Interest			
			Office	Transport	Entertain	Resident
Office	1	45.72%	0.1034	0.0813	0.0515	0.0439
Transport	2	2.58%	0.1012	0.2000	0.1020	0.0473
Entertainment	3	9.35%	0.0976	0.1201	0.1674	0.0474
Resident	4	17.55%	0.0232	0.0285	0.0269	0.0528
Comprehensive	5	24.81%	0.0453	0.0373	0.04030	0.0508

2) *Linking Mobility Patterns with Social Ecology*: After obtaining the clustered BSs, the next question to ask is how to link these clusters to the social ecology of the city, i.e., the urban functional regions? We build the linker via investigating the ground truth of urban functions in different regions.

To start with, the distribution of point of interests (POIs) is investigated in each cluster to establish the links between human mobility patterns and urban functional regions. POI is a specific point location of a certain function such as restaurant and shopping mall. An area's POI distribution reflects its urban function and can be considered as ground truth [14]. Therefore, studying the POI distribution of an area can help us to accurately find out the urban function of that area. To calculate the POI distribution, we measure the numbers of four types of POIs, which are resident, transport, office and entertainment, within 200m of each cellular tower. Then, different regions' POI distributions are summarized in the right-hand part of Table I. The maximum value of each column is marked with color, which shows the dominant urban function in the corresponding row, i.e., cluster. According to Table I, cluster 1 corresponds to office area, cluster 2 corresponds to transport area, cluster 3 corresponds to entertainment area, cluster 4 corresponds to residential area, and cluster 5 corresponds to comprehensive area. Therefore, regions are classified into their dominant urban functions. If a region does not have an obvious dominant urban function, it is classified as a comprehensive type. With the help of POI data, we manage to establish the links between the human mobility of physical world with the urban social ecology.



(a) Normalized traffic variations of different regions in weekday (b) Normalized traffic variations of different regions in weekend



(c) Average traffic consumption rates in different regions

Fig. 4. Characteristics of cellular traffic patterns in different functional regions.

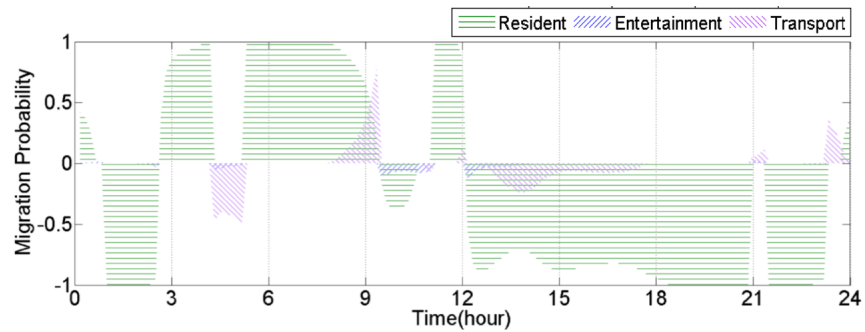
B. Understanding Human Behaviours with Social Ecology

After discovering the links between the cyberspace, physical world and social ecology, we are able to further our analysis to better understand human behaviour in both the cyberspace and physical world by focusing on human behaviours in different functional regions of social ecology. To characterize the features of cellular traffic patterns in different urban functional regions, the normalized traffic patterns both in weekday and weekend are presented in Fig. 4.

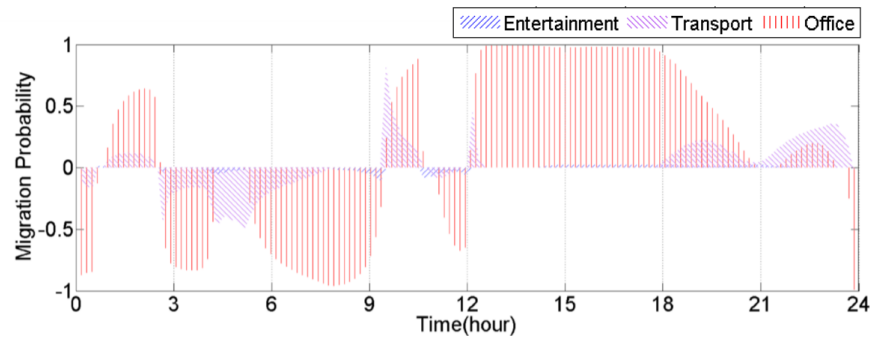
Comparing Fig. 4 (a) with Fig. 4 (b), the traffic patterns in different urban functional regions have distinct features in weekday and in weekend. In weekday, the traffic patterns of office area and entertainment area reach their peaks around noon, while the traffic pattern of transport area has two peaks in morning and afternoon and the traffic pattern of resident area experiences high

value during the night. By contrast, in weekend, the traffic in transport area has only one peak, and the traffic patterns in resident area and office area are different from those experienced in weekday, while entertainment area's pattern does not vary too much from weekday.

To characterize the deep patterns of traffic consumption, the average traffic consumptions for different urban functional regions are presented in Fig. 4(c). Observing from Fig. 4(c), the resident area possesses the highest traffic consumption, while the transport area has the least traffic consumption. In addition, in the office and transport areas, the traffic consumption is lower in weekend than in weekday, while in the resident and entertainment areas, the traffic consumption is higher in weekend, which is consistent with human activity patterns.



(a) To office area



(b) To residential area

Fig. 5. Migration probabilities to office area and residential area.

The information of social ecology can also benefit our understanding of human behaviours in the physical world. Human's migration between different urban functional regions is an important aspect for understanding human's mobility in urban environment. Therefore, the migration probabilities (from other areas) to office and residential areas are presented in Fig. 5 (a) and (b), respectively. The migration probability from one region A to another region B is calculated by

dividing the number of users migrating from A to B with the total number of users moving out of region A. In addition, a positive value represents people actually migrate from A to B, while a negative value represents people actually migrate from B to A. Therefore, the migration probability to each area sums up to 1 or -1 in each time slot, with negative value indicates people migrate out of this area and positive value suggests people migrate into this area. Observing from Fig. 5 (b), most people migrate into office area from resident area from 5 AM to 9 AM, and most people migrate out of office area to resident area from 12 PM to 9 PM. In addition, people begin to migrate from transport area to office area from 8 AM to 10 AM, while people begin to migrate from office area to transport area from 1 PM to 5 PM, as can be seen from Fig. 5 (a). Furthermore, the migration probability to resident area is the opposite of that to office area, as can be clearly seen by comparing Fig. 5 (a) with Fig. 5 (b). This simply confirms that there is a strong connection between office area and resident area, with most migration happens between these two areas. The above results clearly suggest that going to work is the main purpose of human migration between urban functional regions.

V. PROSPECTS AND DISCUSSION

With the rapid growth of mobile devices and ubiquitous cellular access, the cellular mobile network has become a gigantic sensing platform, which captures human behaviours in the physical world and cyberspace. For example, cellular network records human access of all kind applications as well as human mobility and locations in the physical world. These detailed information enables us not only to analyze human mobility and traffic consumption patterns but also to study the links between human behaviours in the physical world, cyberspace and social ecology. In our future works, we plan to further investigate the links from the following aspects.

Spatial domain: In our current study, we find out that the number of active mobile users has a strong correlation with cellular data traffic in spatial domain. With the penetration rate of mobile devices reaches up to 96% over the world, mobile devices have become the best agent to monitor the traces of human mobility. Therefore, based on cellular big data, future works can be carried out to model the dynamic distribution of population, which not only is an important topic in human mobility, but also plays an important role in disease controlling, transportation scheduling and other urban planning applications.

Temporal domain: In our current study, we have shown that human behaviours have strong temporal periodicity in the physical world as well as in the cyberspace. Moreover, the patterns of human behaviours differ significantly in different urban functional regions. Therefore, based on the links with the social ecology, we can better characterize and model human behaviour in the physical world and cyberspace. Future works can be carried out to model the temporal patterns of human behaviours with the social ecology in mind and to develop applications based on it, such as cellular network's dynamic load balancing schemes.

Events oriented: Detecting anomaly events in the physical world is an interesting topic, which is of great importance in public safety. In our current study, we have observed that human behaviours in the physical world is tightly coupled with those in the cyberspace. For example, a parade may cause spikes in cellular data traffic in particular regions. Therefore, by investigating the links between human behaviours in the physical world and cyberspace, our future works plan to develop an effective system to detect the anomaly events in urban environment.

VI. CONCLUSION AND DISCUSSION

In this article, we carry out, to the best of our knowledge, the first study of human behaviours in the cyberspace and physical world embedded in large-scale 3G and LTE cellular towers deployed in an urban environment. Through investigating human mobility patterns and traffic consumption patterns, we characterize the features of human behaviours in the physical world, cyberspace and social ecology. Our analysis reveals that human mobility and traffic consumption have strong correlation and both have distinct periodical patterns in time domain. Moreover, they are both linked with social ecology, which helps us better understand human behaviours. We believe that our analysis provides a systematic and comprehensive understanding of human behaviour in social-physic-cyber space, and opens a set of new research directions.

REFERENCES

- [1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014-2019 White Paper, Feb. 3, 2015.
- [2] A. K. Das, P. H. Pathak, C.-N. Chuah, and P. Mohapatra, "Contextual localization through network traffic analysis," in *Proc. INFOCOM 2014* (Toronto, ON), Apr. 27-May 2, 2014, pp. 925–933.
- [3] B. Cici, M. Gjoka, A. Markopoulou, and C. T. Butts, "On the decomposition of cell phone activity patterns and their connection with urban ecology," in *Proc. ACM MobiHoc 2015* (Hangzhou, Chin), Jun. 22-25, 2015, pp. 317–326.
- [4] Z. Su, Q. Xu, and Q. Qi, "Big data in mobile social networks: a QoE-oriented framework," *IEEE Network*, vol. 30, no. 1, pp. 52–57, Jan.-Feb. 2016.

- [5] Z. Su, P. Ren, and Y. Chen, "Consistency control to manage dynamic contents over vehicular communication networks," in *Proc. GLOBECOM 2011* (Houston, TX), Dec. 5-9, 2011, pp. 1–5.
- [6] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla, "Inferring user demographics and social strategies in mobile social networks," in *Proc. ACM SIGKDD 2014* (New York, USA), Aug. 24-27, 2014, pp. 15–24.
- [7] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3G cellular data network," in *Proc. INFOCOM 2012* (Orlando, FL), Mar. 25-30, 2012, pp. 1341–1349.
- [8] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proc. National Academy of Sciences of USA*, vol. 110, no. 15, pp. 5802–5805, 2013.
- [9] H. Wang, J. Ding, Y. Li, P. Hui, J. Yuan, and D. Jin, "Characterizing the spatio-temporal inhomogeneity of mobile traffic in large-scale cellular data networks," in *Proc. 7th ACM Int. Workshop HotPost* (Hangzhou, China) June 22, 2015, pp. 19–24.
- [10] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018-1021, 2010.
- [11] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.
- [12] F. Corpet, "Multiple sequence alignment with hierarchical clustering," *Nucleic Acids Research*, vol. 16, no. 22, pp. 10881–10890, 1988.
- [13] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002.
- [14] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proc. ACM SIGKDD 2012* (Beijing, China), Aug. 12-16, 2012, pp. 186–194.