2016 17th IEEE International Conference on Mobile Data Management

Characterizing User Mobility from the View of 4G Cellular Network

Weijian Sun, Dandan Miao
Dept. of Electronic Engineering & Information Science
University of Sicence and Technology of China
Hefei, China
{*sunwj,ddmiao*}*@mail.ustc.edu.cn*

Xiaowei Qin, Guo Wei
Dept. of Electronic Engineering & Information Science
University of Sicence and Technology of China
Hefei, China
{*qinxw, wei*}*@ustc.edu.cn*

***Abstract*—The mobility models obtained from mobile data are expected to affect numbers of fields, including urban planning, road traffic engineering, human sociology, epidemiology of infectious diseases, or telecommunication networks. Current user mobility models are mainly extracted from Call Detail Records (CDR) data or WiFi traces. However, CDR data only captures user movements during telephone calls or short message service (SMS) and WiFi traces can not provide intuitive understanding to mobility behavior of cellular network users in large scale. In this paper, we take the first step to investigate if the characteristics of mobility derived from 4G cellular data network is different from the previous findings utilizing other data sources, especially the widely used, CDR based approach. Utilizing our Hadoop based mobile big data processing platform together with a systematic analysis framework of user mobility, we present a comprehensive characterization of the mobility models from the 6 Terabyte(TB) 4G data traffic. Through the comparison with CDR models and 3G models from the view of user occurrence patterns, user movement patterns and dominant locations of users, we find that the 4G data traffic can provide finer granularity of mobility and location information.**

***Keywords*-Cellular data traffic, Mobility Modeling, Important Location, Trajectory Reconstruction**

I. Introduction

Cellular data network has become ubiquitous, fueled by smart mobile devices (cellphones, netbooks and tablet devices) as well as the plethora of mobile applications. Thus, mobile data traffic is an excellent source to reveal the individuals and communities mobility behavior. From the view of cellular network providers, user mobility modeling has broad usage in infrastructure planning, resource optimization, real-time network service provisioning, and new mechanism evaluations. From the view of service providers, the location and mobility information also drives a vibrant ecosystem of location-based services, e.g., personalized proximity advertising, itinerary recommendation, destination based contextualized reminders, and pre-caching of data.

Many attempts have been made to depict and characterize people's real mobility behavior in the past few years. GPS [1] and WiFi traces are the first to be utilized to study human mobility. However, the GPS and WiFI traces data are limited to a small group of participating users and a small geographic area. Call Detail Records (CDRs)[2], [3] are by far the most common type of mobile traffic data employed by researches of user mobility modeling. Despite the promising results demonstrated by the CDRs based approach, there are a few studies start to investigate the accuracy of CDR data for mobility inference. Among which, the most common used data source is 3G traffic data [4], [5].

In this paper, we seek to answer the following question, *comparing with mobility models extracted from CDRs or 3G data records, will those models extracted from 4G data network records exhibit differences?* More specifically, in this paper, we utilize our Hadoop [6] based mobile big data processing platform together with a systematic mobility analysis framework to describe user mobility behavior from our 6TB 4G data records. Each user's trace is constructed from the sequence of eNodeBs which provide 4G data service to users. The extracted mobility models are compared with their corresponding mobility models derived from other sources, e.g., the CDRs or 3G data records. Despite that many recent researches [7] [8] start to investigate cellular data traffic, we make the following contributions in this work.

- **Systematically framework of mobility analysis for 4G data traffic:** To the best of our knowledge, we are the first to study user mobility from the data usage of 4G data network. Especially, we propose a systematically framework for mobility analysis for this type of data.
- **Confidence based method to reconstruct traces:** We propose a confidence based method to reduce such false positives. It should be note that our mobility models may not be applicable to other regions with significantly different spatial structure, population and density, but we believe that our trajectory extraction method can be applied in general.

The remainder of the paper is organized as follows. In Section II, we describe our mobility analysis framework. In Section III, we compare our models with other models. Finally, we conclude this paper in Section IV.

2375-0324/16 $31.00 © 2016 IEEE
DOI 10.1109/MDM.2016.19

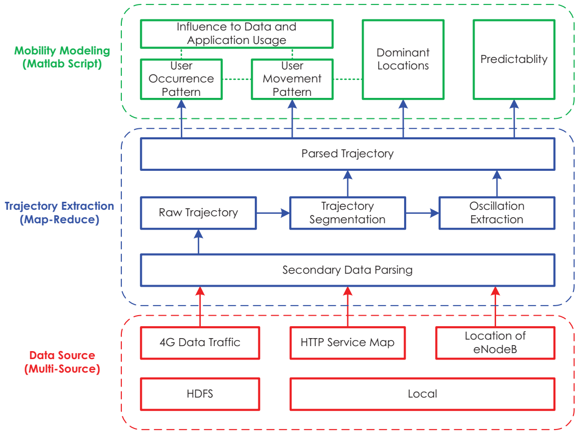


Figure 1: Systematical framework to analyze user mobility from 4G data traffic

II. Background

A. Framework of Mobility Analysis

We propose a systematical framework to analyze user mobility from the view of 4G data traffic in Fig.1. Though trying our best, we admit that some results may still be specific to our dataset and may not be generalized. However, our analysis framework should applicable to any 4G data consistent with our data format.

B. Data Set

We use one anonymous data set from a tier-1 4G cellular network carrier in China. All user identifiers are anonymized to protect privacy without affecting the usefulness of our analysis. In general, our data set contains about 400,000 4G users and their billions of HTTP records during 3 days of April in 2015 in one northwest city of China. Furthermore, 707 kinds of HTTP applications are extracted by our OTTCAP [9] through parsing HTTP tuple, namely, <*HOST*, *URI*, *User-Agent*> offline. We classify them into 18 classes based on their function.

C. Trajectory Extraction

As mentioned above, ping-pang effect of overlapped eNodeBs will introduce false positives in detecting true movements of 4G users. In this section, we propose a confidence based method to reduce such false positives. There are three parts of our method and the parsing procedure is shown in the middle of Fig.1. We will discuss each part in detail in the following subsection.

1) Raw Trajectory: We organize the consecutive HTTP records into a sequence of L three-tuple records first. The i^{th} HTTP record for each user u is

$$X_i^u = < tim_i^u, loc_i^u, App_i^u >, i = 1...L \tag{1}$$

where tim_i^u is the beginning timestamp, loc_i^u means the location of serving cell and App_i^u represents the application info of each HTTP record. Next, we group the consecutive HTTP records of u until a new serving cell, namely, loc_{k+1}^u, according to the rule below.

$$\forall i^{'}, loc_{i^{'}}^u = loc_i^u \ \& \ loc_{k+1}^u \neq loc_i^u \ where \ i \leq i^{'} \leq k \tag{2}$$

we also combine the App_i^u of continuous X_i^u which belongs to a same serving cell, for example, *traffic(KB)* or *duration(s)* of each application are cumulated for every presented application. After the combination step, we get the raw trajectory of user u

$$T_{i\to j}^u = \{X_i^u, X_{i+1}^u, ..., X_j^u\} \tag{3}$$

We name the new combined X_i^u trajectory tuple.

2) Trajectory Segmentation: We further parse the raw trajectory $T_{i\to j}^u$. A trajectory segment is a ***maximal*** ordered subset of a raw trace in which the time between trajectory tuples does not exceed a specified timeout value δ. For specific, we get trajectory segment *m* as

$$S_m^u = \{X_{m(1)}^u, X_{m(2)}^u, ..., X_{m(M)}^u\} \tag{4}$$

according to the rule below

$$\forall k, \ tim_{m(k+1)}^u - tim_{m(k)}^u < \delta, \ where \ 1 \leq k \leq M-1 \tag{5}$$

Overall, δ is a value that distinguish continuous mobility segment. Here, the location set of S_m^u is represented by $S_m^u.Loc$ and $S_m^u.Loc = \{X_{m(1)}^u.loc, X_{m(2)}^u.loc, ..., X_{m(M)}^u.loc\}$.

3) Overlap Extraction: For each cell, namely, loc_c, we first collect its candidate overlapped cell, namely, loc_d by computing the confidence

$$Con(loc_c \to loc_d) = \frac{\sigma(loc_c \bigcup loc_d)}{\sigma(loc_c)} \tag{6}$$

from all users' (U) trajectory segment $S^U.Loc$ and $\sigma(loc_x)$ is defined by

$$\sigma(loc_x) = |\{S_n^U.Loc | loc_x \in S_n^U.Loc \ , S_n^U.Loc \in S^U.Loc\}| \tag{7}$$

where $1 \leq n \leq N$ and N represents the total number of trajectory segments of our 4G data. loc_d is considered to be the overlapped cell of loc_c if

$$P(loc_d | loc_c) > Pr_\delta \ \& \ Dis(loc_d, loc_c) < Dis_\delta \tag{8}$$

where Pr_δ is a probability threshold based on the distribution of conditional probability while $Dist_\delta$ is the distance threshold to be overlapped pairs. We use $Dist_\delta = 70km$, because the maximum radius of a cell defined by 3GPP is 35 km. After these two steps, we obtain a candidate set for each cell.

4) Trajectory Reconstruction: In the Trajectory Reconstruction part, based on each cell's overlapped candidates, ping-pang phenomenon will be removed from each trajectory segment S_n^U. We first sort the occurrence times of each cell in trajectory segment S_n^U from big to small. Then, for the sorted cell set of S_n^U, we iteratively (from big to small) replace each cell's candidate overlapped cells if they are also in S_n^U. Third, we resume the order of replaced or non-replaced cells. For example, if $S.Loc = \{loc_1, loc_2, loc_1, loc_3, loc_4\}$ and the candidate overlapped cells of loc_1 are loc_2 and loc_3, then we get $S.Loc_{new} = \{loc_1, loc_1, loc_1, loc_1, loc_4\}$. Finally, we get the reconstructed trajectory $T_{i \to j}^{u'}$.

$$T_{i \to j}^{u'} = \{X_i^{u'}, X_{i+1}^{u'}, ..., X_j^{u'}\} \tag{9}$$

Our further mobility models are all extracted from the parsed user trajectories.

D. Dominant Locations of Users

Most previous models focus on two most important places of users, namely, home and work place. They extract these two places either from data of census or using the time-of-day heuristics, for example, considering a location as a user's home if he spends most of his time between 10PM and 6PM at this location [10]. However, from the view of data network, it is not clear that if home or work place will exhibit the same features primarily because of the availability of Wi-Fi connection. Therefore, more dimensions of features should be defined to extract dominant locations of users. In this paper, we investigate this problem from two dimensions of significance, namely, the consumption of applications(duration and traffic) and user occurrence. We speculate that dominant locations are possibly score high in both dimensions.

III. User Mobility Analysis

A. Basic characterization

In this section, we aim at the unique mobility properties provided by the 4G cellular data network. Since we do not have access to CDR or 3G data, we perform comparison with existing published researches. Our results are presented from three aspects, the occurrence patterns, the mobility patterns and the important locations.

1) User occurrence: We first quantify how often users use 4G network service. We group continuous trajectory tuples of one user together into one usage instance by granularity of 1 minute, and then compute the time interval between any two consecutive usages of a same user. Fig.2(a) shows the interval in minutes (x-axis) and its corresponding probability distribution (y-axis) in three days. Users tend to user 4G network in short bursts. As shown in Fig.2(a), the dominant fractions (90% of intervals) are within 10 minutes (600 seconds). It also has a long-tail distribution, suggesting long periods with no network usage. However, comparing

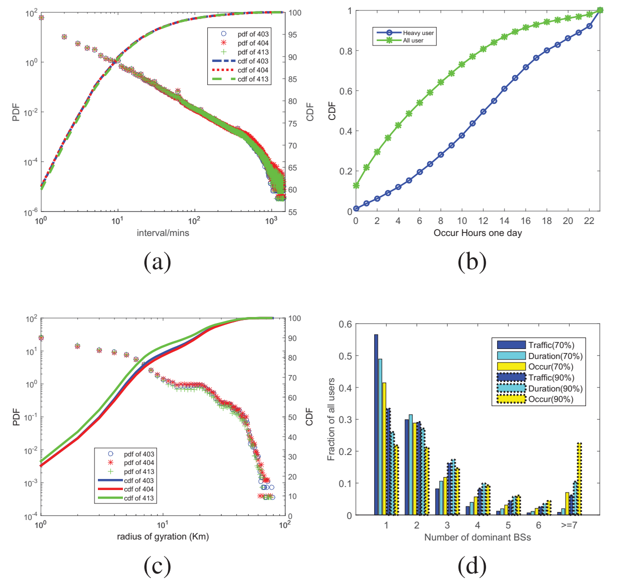


Figure 2: (a) PDF and CDF of 4G usage interval in three days. (b) CDF of hours of user occur in one day. (c) CDF and PDF of radius of gyration in three days. (d) Property of dominant locations

with CDR based [11] and 3G data based approach [5], the usage gap here is much smaller.

The overall shorter gap mentioned above means that we can locate users more continuously when they are using 4G network. However, although users tend to use 4G network more frequently in short period, during those long gaps in the tail of Fig.2(a), we still do not know about the users location. Thus, we also measure the fractions of usage instance when user's location is known per hour and plot its CDF in Fig.2(b). The distribution are more spread out compared to the unknown fractions in CDR and 3G data [5]. For example, for 70% of all users, more than 10% of their locations are known in one day. However, the variance across users is large, suggesting one could draw more conclusion from heavy users. For specific, as the blue line shown in Fig.2(b), for nearly 50% of heavy users, we can locate them more than 12 times from the granularity of hour in one day. These two figures both demonstrate that 4G data provides more locations daily for analysis than CDR and 3G data, as expected.

Next, we analyze the temporal property of user occurrence through how many times each user uses the network per hour. The maximal number of times per hour is 60 since we group continuous trajectory tuples by the granularity of 1 minute. We perform clustering on all users using Kmeans [12] due to its simplicity and use DBI [13] to determine the best clustering choice. Fig.3 shows the distribution of average relative occurrence for 10 clusters. We sort these

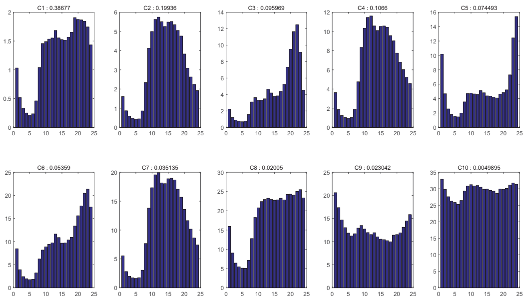


Figure 3: 10 user occurrence patterns extracted by Kmeans and DBI

cluster centers by the range of occurrence times per hour (y-axis). Different cluster describes different user occurrence patterns. Specifically, the largest cluster $C1$ contains nearly 40% of all users and its pattern is in accordance to human routine. Comparing with $C1$, users in $C6$ and $C10$ tend to access 4G network more frequently, especially for $C10$, the smallest occurrence times happens in the mid-night is 10 times larger than the biggest occurrence times in $C1$. The high occurrence in $C10$ provide us probability to locate users more frequently. $C2$, $C4$ and $C7$ are with almost the same patterns but with different level of hourly occurrence. They both have a occurrence burst before lunch time and drop down after work (around 6 oćlock). These 3 clusters could be the users who access network through WI-FI at home. $C3$ and $C6$ both have burst occurrence rate during evening time but low frequency during work time. This pattern makes for the typical routine of "work hard play hard". $C5$ and $C9$ are almost with the same patterns, but their occurrence times in mid-night is also higher comparing with $C3$ and $C6$ which means that their bursts of occurrence are more durable.

In summary, these findings of temporal occurrence behavior between users are very different from CDR. In CDR (Figure 4 of [5]), two clusters are detected: one group has more afternoon calls and the other makes more calls in the evening. This is a key finding of the 4G data approach, as besides the expected time-of-day dynamics, 4G data approach contains more regular access from the users, thus it allows us to capture more movements than CDR.

2) User mobility: One common metric to quantify how far a user moves over a time period is the radius of gyration [11]. It is defined as

$$r_g^u(t) = \sqrt{\frac{1}{n_c^u(t)} \sum_{i=1}^{n_c^u(t)} (\overrightarrow{r_i^u} - \overrightarrow{r_{cm}^u})^2} \tag{10}$$

where $\overrightarrow{r_i^u}$ represents the $i = 1, .., n_c^u(t)$ location recorded for user u and $\overrightarrow{r_{cm}^u} = 1/n_c^u(t) \sum_{i=1}^{n_c^u(t)} \overrightarrow{r_i^u}$. Fig.2(c) shows the CDF and PDF of gyration radius in three days. The CDF of day *20150404* is a little lower than that of *20150403* and *20150404* which means that people tends to have high mobility in holiday, since *20150404* is an official holiday - Chinese dragon boat festival.

We find that 80% of users' daily movements are confined to an area of 10 km, while a few users (10%) travel more than 30 km. Comparing with CDRs results [11], it obeys the same truncated power-law distribution model, but with smaller radius. It could be a consequence of more frequent usage on the applications on the smartphones in 4G network. [14] finds that 34% of all users display a radius of gyration of less than 10 km, while only 14% have a radius of gyration larger than 500 km. This comparison means that although different data source discovers similar model (truncated power-law distribution), the model's parameters are different. It mainly depends on the temporal and spatial granularity that the data source can capture about the human movement behavior. From this point of view, 4G data is a better one.

Similar to the method of describing user occurrence, for each user, we also compute the vectors of his unique locations from the parsed trajectories per period of time. The time period is set to 0 $o'clock \rightarrow$ 5 $o'clock$, 6 $o'clock \rightarrow$ 8 $o'clock$, 9 $o'clock \rightarrow$ 12 $o'clock$, 13 $o'clock \rightarrow$ 14 $o'clock$, 15 $o'clock \rightarrow$ 18 $o'clock$, 19 $o'clock \rightarrow$ 21 $o'clock$ and 22 $o'clock \rightarrow$ 23 $o'clock$, which are shown in the y-axis of Fig.4. Kmeans together with DBI determines 8 clusters as the best choice. The clustering results are shown respectively in Fig.4. Overall, the percentage of each cluster is relatively uniformity comparing with the clusters of user occurrence. The largest cluster is $C4$ which contains nearly 20% of all users. $C3$ and $C4$ can be classified as one class since each of them has a relatively random and flat pattern with less human schedule behavior. $C1$, $C2$ and $C7$ could be classified as one class since each of them has a mobility burst, namely, in the afternoon, noon and forenoon respectively. Furthermore, $C5$, $C6$ and $C8$ could be grouped because every of them have two burst period, specifically, users in $C5$ are with high mobility in the afternoon and late evening while users belong to $C6$ have high mobility in the morning and afternoon, cluster $C8$ is in the morning and early evening. We speculate that $C5$ tends to be the users who only work in the afternoon while $C6$ and $C8$ are composed of general workers with different after working time.

Comparing with the 3G mobility patterns clustered in Figure 6 of [5], we get more patterns. We believe that these new added human mobility patterns generated from 4G data could also exist in 3G users, we suspect that their disappearances are mainly caused by the relatively low usage frequency of 3G network. In summary, together with user occurrence patterns, 4G network has changed user behavior of network usage and thus provide more information of user mobility. In our later analysis, we leverage the mobility

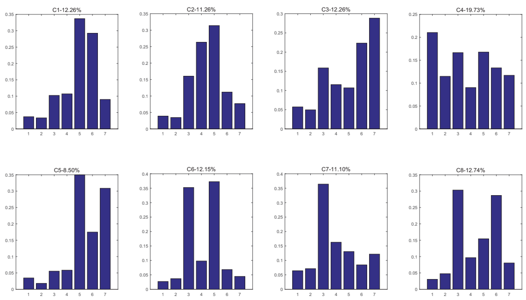
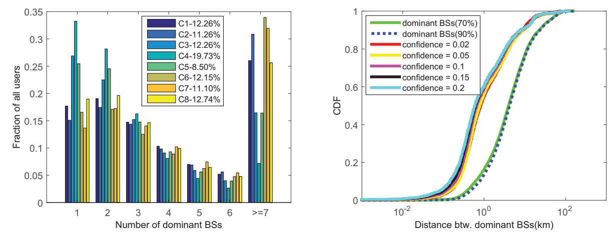


Figure 4: 8 user mobility patterns extracted by Kmeans and DBI



(a) (b)

Figure 5: (a) Mobility patterns VS. Dominant locations. (b) Distance distribution of ping-pong pairs and dominant locations.

clustering results to perform further comparisons with their corresponding dominant locations models.

3) Dominant Locations: The dominant locations are defined as those locations where a user spends top p% of his all network consumption (traffic and duration) or his occurrences. We show $p = 70\%$ and $p = 90\%$ in Fig.2(d) first to explore different dominant levels. For specific, for 27% of all users, the top two locations account for 95% of their network consumption (traffic and duration) and it is similar when measuring on user occurrence. Comparing with the user occurrence results of 3G data [5], we find 22% users whose 95% of occurrences are generated in their top 1 location, much larger than 5% mention in [5]. We speculate that it is mainly caused by the behavior of location-based network usage for 4G users, they tend to access network in specific locations more frequently than 3G users. This phenomenon is also depicted through the consumption of data from the view of traffic and duration in Fig.2(d), since nearly 35% of all users generate 95% of their traffic in their top 1 location. We also compare results with CDR-based models, although 90% of the 4G network consumption and occurrences are covered by the top 6 locations for most users, we can see that the number of important locations are much larger and more spread out since [3] finds that most mobility are accounted for just the top 2 locations with which a user is associated.

Furthermore, we use the extracted mobility patterns in last subsection to perform comparison with their corresponding dominant locations. We study if these 8 groups of users show different spatial properties. We only show $p = 90\%$ of user occurrence in Fig.5(a) due to space limitation. Overall, different mobility pattern indeed has disparate models of dominant locations. Interestingly, we find that $C4$ has the smallest number of dominant locations, for specific, nearly 80% of users in $C4$ have one or two dominant locations. The same phenomenon happens to $C3$ and $C5$, with relatively smaller percent. As we mentioned in last subsection, $C3$ and $C4$ both have relatively flat mobility patterns while $C5$ have high mobility in afternoon and late evening. We speculate that users who tend to move randomly or late in a day have smaller dominant locations and we can predict their locations of appearance with high probability. However, $C6$ has the biggest number of dominant locations, nearly 35% of users in it have more than 7 dominant locations. Comparing with $C6's$ pattern, we speculate that the normal workers tend to have more dominant locations since they could pass by lots of locations while going to work or off duty with network usage, simultaneously.

Finally, we look into how far the dominant locations are apart. For each user, we get a list of his dominant locations, and then compute the distance between any pair of them. Take occurrence based dominant locations as example, the CDF lines of $p = 70\%$ and $p = 90\%$ are shown in Fig.5(b). The closeness of the two lines means that the definition of dominant locations are not sensitive to the choice of p. Since 80% of the dominant pairs are within 10 km and we speculate that it is the distance between home and work place.

As mentioned above, we use the confidence based method to smooth out the false positives introduced by overlapped cells. However, we do not have ground truth to validate our approach. Thus, we examine the distances between each ping-pang pair and compare them with the distance between dominant cells in Fig.5(b). It is shown that most of our detected ping-pang pairs are much closer, i.e., for $confidence = 0.02$, 55% of distances are within 1km and 90% of them are within 10km. We also compare the distance between ping-pang pairs among different confidence values in Fig.5(b) and the differences are not obvious. Overall, results show that our trajectory reconstruction method is not sensitive to the choice of Pr_δ and $Dist_\delta$.

IV. Conclusion

In this paper, we utilize our Hadoop based mobile big data processing platform together with a systematic mobility analysis framework to describe user mobility behavior from our 6TB 4G data records. We propose a confidence based method to reduce ping-pang effects when reconstruct user trajectories, and the extraction method is validated by the comparison of distance between dominant locations and ping-pang pairs, which show convincing results. Based on the parsed user trajectory tuples, some unique mobility properties are extracted from 4G data. Our results are presented from three aspects, the occurrence patterns, the mobility patterns and the important locations. We compare our mobility results with those generated from CDR or 3G data carefully and present their differences in detail. We conclude that the 4G data traffic can provide finer granularity of mobility and location information.

ACKNOWLEDGEMENT

This work is supported by the National Basic Research Program of China (973 Program: 2013CB329004), National High Technology Research and Development Program of China (863 Program: 2015AA01A706), National nature science fund of China(61271272).

REFERENCES

[1] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM transactions on networking (TON)*, vol. 19, no. 3, pp. 630–643, 2011.

[2] E. Halepovic and C. Williamson, "Characterizing and modeling user mobility in a cellular data network," in *Proceedings of the 2nd ACM international workshop on Performance evaluation of wireless ad hoc, sensor, and ubiquitous networks*. ACM, 2005, pp. 71–78.

[3] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[4] G. Ranjan, H. Zang, Z.-L. Zhang, and J. Bolot, "Are call detail records biased for sampling human mobility?" *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 16, no. 3, pp. 33–44, 2012.

[5] Y. Zhang, "User mobility from the view of cellular data networks," in *INFOCOM, 2014 Proceedings IEEE*. IEEE, 2014, pp. 1348–1356.

[6] Apache, "Hadoop," http://hadoop.apache.org/.

[7] W. Sun, X. Qin, and G. Wei, "Modeling and mining the temporal patterns of service in cellular network," *Communications, China*, vol. 12, no. 9, pp. 11–21, 2015.

[8] R. Li, Z. Zhao, C. Qi, X. Zhou, Y. Zhou, and H. Zhang, "Understanding the traffic nature of mobile instantaneous messaging in cellular networks: A revisiting to $alpha$-stable models," *Access, IEEE*, vol. 3, pp. 1416–1422, 2015.

[9] W. Sun and X. Qin, "End-to-end delay analysis of wechat video call service in live dc-hspa+ network," in *Wireless Communications and Signal Processing (WCSP), 2014 Sixth International Conference on*. IEEE, 2014, pp. 1–5.

[10] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Measuring serendipity: connecting people, locations and interests in a mobile 3g network," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. ACM, 2009, pp. 267–279.

[11] 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.

[12] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Applied statistics*, pp. 100–108, 1979.

[13] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 2, pp. 224–227, 1979.

[14] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services." *ICWSM*, vol. 2011, pp. 81–88, 2011.