CellPred: A Behavior-aware Scheme for Cellular Data Usage Prediction

ZHOU QIN, Rutgers University, USA
FANG CAO, Southeast University, China
YU YANG, Rutgers University, USA
SHUAI WANG, Southeast University, China
YUNHUAI LIU, Peking University, China
CHANG TAN, iFlytek, China
DESHENG ZHANG, Rutgers University, USA

Cellular data usage consumption prediction is an important topic in cellular networks related researches. Accurately predicting future data usage can benefit both the cellular operators and the users, which can further enable a wide range of applications. Different from previous work focusing on statistical approaches, in this paper, we propose a scheme called *CellPred* to predict cellular data usage from an individual user perspective considering user behavior patterns. Specifically, we utilize explicit user behavioral tags collected from subscription data to function as an external aid to enhance the user's mobility and usage prediction. Then we aggregate individual user data usage to cell tower level to obtain the final prediction results. To our knowledge, this is the first work studying cellular data usage prediction from an individual user behavior-aware perspective based on large-scale cellular signaling and behavior tags from the subscription data. The results show that our method improves the data usage prediction accuracy compared to the state-of-the-art methods; we also comprehensively demonstrate the impacts of contextual factors on *CellPred* performance. Our work can shed light on broad cellular networks researches related to human mobility and data usage. Finally, we discuss issues such as limitations, applications of our approach, and insights from our work.

CCS Concepts: • **Human-centered computing → Mobile devices**; • **Information systems → Mobile information processing systems**.

Additional Key Words and Phrases: Urban Computing, Cellular networks, Deep Learning, Behaviors, Prediction

ACM Reference Format:
Zhou Qin, Fang Cao, Yu Yang, Shuai Wang, Yunhuai Liu, Chang Tan, and Desheng Zhang. 2020. *CellPred*: A Behavior-aware Scheme for Cellular Data Usage Prediction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 40 (March 2020), 24 pages. https://doi.org/10.1145/3380982

Authors' addresses: Zhou Qin, Rutgers University, 96 Frelinghuysen Rd, Piscataway, NJ, 08854, USA, zq58@cs.rutgers.edu; Fang Cao, Southeast University, Nangong Rd, Nanjing, Jiangsu, China, i.caofang@gmail.com; Yu Yang, Rutgers University, USA, yy388@cs.rutgers.edu; Shuai Wang, Southeast University, China, shuaiwang@seu.edu.cn; Yunhuai Liu, Peking University, 5 Yiheyuan Rd, Beijing, Beijing Shi, China, yunhuai.liu@pku.edu.cn; Chang Tan, iFlytek, Hefei, Anhui, China, changtan2@iflytek.com; Desheng Zhang, Rutgers University, USA, desheng.zhang@cs.rutgers.edu.

2474-9567/2020/3-ART40 $15.00
https://doi.org/10.1145/3380982

1 INTRODUCTION

Cellular networks play an essential role in our daily life. Billions of devices, including smartphones, tablets, and IoT devices, are connected to cellular networks, and more are expected in the future [6, 16, 32]. The performance of cellular networks draws attention from both operators and users due to the ubiquity of cellular networks. Cellular data usage estimation and prediction are needed for broad applications from wearable devices, cellphones to the smart Internet of Things (IoT) devices, and autonomous driving [11, 41, 51]. There are lots of existing researches focusing on the data usage consumption prediction. *Xu et al.* demonstrated the effectiveness of a time series analysis approach, which decomposed the traffic patterns into regular and random components [50]. *Zhang et al.* proposed a deep spatial-temporal neural network scheme, which improved long-term data usage prediction accuracy compared to commonly used prediction algorithms [53]. *Wang et al.* approached the data usage prediction from a new spatiotemporal dependency aspect and demonstrated the superiority of this perspective compared to state-of-the-art methods [42]. However, all these latest researches on data usage prediction focused on aggregated cell tower level, which did not address the fundamental cause of data usage dynamics: cellular users.

There are indeed some works studying cellular networks under various scenarios from a user perspective by collecting user-end data, for example, *Tu et al.* focused on how phone calls may affect the data service performance in LTE networks [18]. The performance and reliability of VoLTE (Voice over Long-Term Evolution) are validated by *Jia et al.* [21]. But few researches investigate data usage prediction based on individual cellular users. More importantly, spatial and temporal dynamics of cellular data usage, essentially, are brought by individual users: diverse users' usage distributions, various online behaviors, and different mobility patterns, etc. Mobility patterns and user behaviors have a direct influence upon data usage across the cellular networks, e.g., car driver users may use navigation apps during morning and evening rush hours, some users may follow a regular commuting route during workdays, etc. Thus we can investigate users' cellular usage patterns and potentially make use of them for a better usage understanding and prediction performance.

In short, there is little work, if any, using cellular signaling data concentrating on individual user level behaviors and carrying further behavioral analysis due to the lack of detailed large-scale user-level data. In this paper, we analyze the data usage consumption from the individual user perspective to see whether user behaviors and mobility patterns can help better predict data usage consumption. We argue that by enhancing the understanding of characteristics of subscribed cellular users, prediction of cellular data consumption would be more effective. In detail, by working with one of the largest cellular operators in China, we analyze cellular data usage of users based on city-scale signaling data and behavior-related tags generated by exclusive data package subscription data. The cellular data are signaling data covering more than 3 million users, providing information including signaling types, timing, and attached cell towers. The subscribed cellular users also come with behavioral tag data from the operator, including six major categories of tags related to the users' behaviors. Then, based on the data, we propose the framework *CellPred* to comprehensively utilize features provided by individual users and tags, to boost user data usage prediction and mobility pattern prediction. As a result, we achieve a better performance compared to traditionally cellular data prediction on the cell tower level.

Contributions of this paper can be summarized below:

- To our knowledge, this is the first paper to study cellular data prediction from an individual user perspective with behaviors considered. Mobility patterns and data usage patterns are revealed by detailed signaling data together with behavioral tags. Our work is based on real-world signaling data generated by one of the major cellular operators from the Chinese city Hefei, covering more than 3 million users. We will release anonymized sample data for the reproducibility of our work in a broader sense. Details will be given in Section 6.
- We propose a framework called *CellPred* for cellular data usage prediction by considering individual user's mobility patterns and data usage patterns. In particular, we design this framework by taking historical

mobility and usage sequences, mobility and usage features as inputs with extra aiding from behavior tags manipulated by a soft attention module. The prediction framework consists of two parts, i.e., an encoder embedding inputs, features, and tags, and a decoder for mobility and usage prediction. Then we aggregate individual results to connected cell towers to obtain cell tower level cellular data usage.

- We evaluate our framework *CellPred* over the three weeks long signaling data with a daily amount of 267 million records and a daily size of 20GB from more than 3 million users, together with the tag data set including 3,931 behavior related tags. Through the evaluation of the large-scale data sets, we summarize lessons learned and provide a set of discussions regarding the limitations and potential applications of this work.

2 MOTIVATION

Data usage prediction plays an essential role in cellular networks; the amount of data consumed draws attention from both subscribed users and operators. For the operators, accurate data prediction can provide guidance for resource allocation of cellular networks, especially during big events [45]; for the users, better load management can provide them with a better experience. However, currently researches only focus on aggregate level analysis and prediction. It is not clear how different users utilize cellular data and whether patterns extracted from historical data usage of individual users can be used to help predict data usage, e.g., what kind of users tends to consume more data usage, when and where they are likely to heavily use cellular data, etc. With the smart city initiatives, we are lucky to have offline access to a city-scale signaling data set together with tag data generated by exclusive data package subscription, e.g., what kind of apps users buy data packages for. Considering all of these, we plan to make full use of users' historical data usage records by considering hidden patterns of data usage and mobility from user behaviors to aid data usage prediction. To verify the feasibility of doing so, we demonstrate the patterns and phenomena extracted from cellular users regarding their mobility and data usage.

2.1 Challenges of Data Usage Prediction

Despite lots of researches done regarding cellular traffic prediction, the task of predicting data usage for a large-scale cellular network with fine granularity remains tricky. Particularly, we summarize the challenges as three types via data investigation on our city-scale signaling data set: 1) temporal dynamics: in our data set, the data consumption during different time can be quite different, e.g., we calculate the data usage for every one minutes, and the highest data usage for one minute can be 96 times greater than the lowest data usage of one minute at the overall cellular network level; 2) spatial dynamics: different cell towers bear different data usage load in the real-world. There are even cell towers with no data usage but used for other purposes, such as voice services. The data usage can be high as 127 GB and as low as 0 per day for all the cell towers in this cellular network; 3) individual dynamics: different cellular users in the networks also show usage dynamics, e.g., daily data usage of a user can be high as 2.87 GB and can be low as 0. In sum, high dynamics exist in the usage data set from different perspectives. Directly predicting the data usage of cell towers would be prohibitively challenging considering the drastic dynamics.

2.2 Opportunity

However, approaching the data usage prediction from individual users with consideration of people's behaviors may provide us with new opportunities to identify the hidden patterns of data usage, since the data usage is contributed by different individual users, who may reveal regular usage and mobility patterns.

2.2.1 Data Usage Dynamics: Network Level vs. Individual Level. To begin with, we show the data usage dynamics of aggregate cell tower level and individual user level by entropy to illustrate why the user level perspective is more effective. Specifically, we discretize and normalize the daily amount of data usage of cell towers and

individual users into ten categories, calculated as below:

$$Category_i = 10 \times \frac{usage_i - MIN_{usage}}{MAX_{usage} - MIN_{usage}} \quad (1)$$

MAX_{usage} and MIN_{usage} refer to the maximum and minimum daily usage of cell towers or cellular users. The $usage_i$ here refers to the daily usage of a specific cell tower or cellular user. After this calculation, we then show the results of the entropy of the discretized data usage of cell towers and cellular users across different days in Fig. 1, and average daily usage entropy is shown in Fig. 2. We can observe that the data usage entropy of cell towers is generally larger than that of individual users, no matter it is across days or daily based.

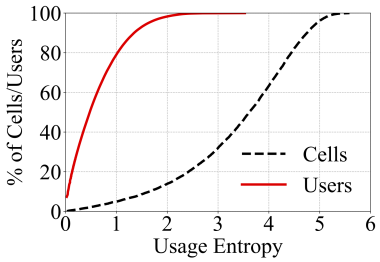

Fig. 1. Usage Dynamics Across Days

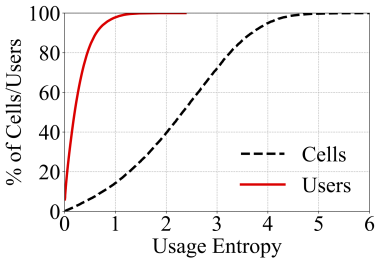

Fig. 2. Daily Usage Dynamics

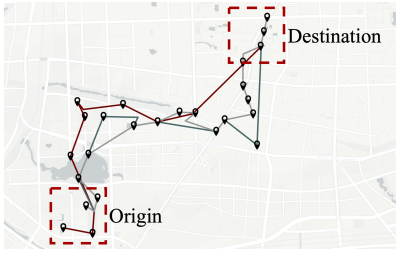

Fig. 3. Repeated Traces

2.2.2 Usage Regularity: Temporal & Spatial Pattern. The signaling data are logged due to various events that happened in the cellular networks, e.g., users making phone calls, visiting the Internet, etc. For each record, we have the timing and corresponding attached cell tower, essentially we have users' spatiotemporal traces at cell tower level, with a high uploading frequency (average 2.85 minutes for each user).

As suggested by Fig. 1 and Fig. 2, we can see that users tend to have a smaller data usage dynamic across multiple days and on a daily basis, showing better predictability of the individual user level from the temporal perspective. We also use the example of users' traces to demonstrate that users have regular patterns from a spatial perspective, e.g., users connect to the same cell tower (or the proximity of the cell tower) during some regular time. For example, origin and destination spots can be inferred from a regular user, where we regard users with n routes with starting cell towers and ending cell towers in the proximity of circle with 1km radius [1]. Here we implement the similarity calculation between two traces as described in [52], we then find users with high trace similarity values across different days. In Fig. 3, we show one user's three traces from three different days. We can see that the origins and destinations are geographically close in three different traces. Besides, their paths are similar except connected cell towers are slightly different. Patterns similar to this can then be considered in predicting a user's future location, which enhances the performance of mobility prediction.

2.2.3 Contextual: User Behavior. Except for the signaling data set, we also have access to the tag data generated by the data package subscriptions, which can give us more information about users. We argue that different tags may indicate different mobility and usage characteristics of users; a group of users with the same tag may show similar statistical features (indicated by Fig. 4). To be general, we show the similarity of data usage among users sharing the same tag and users not sharing any tag in Fig. 5. Specifically, we calculate the similarity of users' temporal usage sequences, i.e., the sequences of data usage for every 1 minute across 24 hours, among users sharing the same tag and users sharing different tags. We can observe that users sharing the same tag tend to have a higher temporal usage resemblance. Regarding mobility, we measure the similarity based on the number of distinct cell towers connected every 1 minute to quantify the mobility dynamics. The results are shown in Fig. 6, which suggests that users sharing the same tag also have a higher mobility similarity than users with different tags.

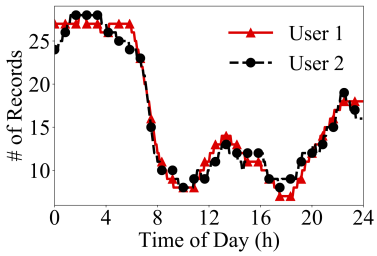

Fig. 4. Sample Users Similarity

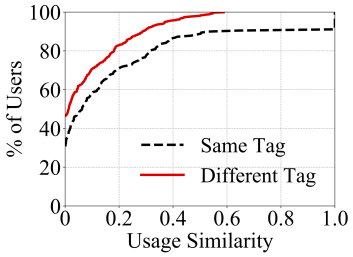

Fig. 5. Users Usage Similarity

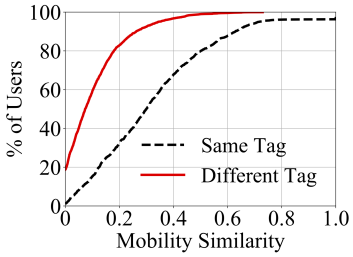

Fig. 6. Users Mobility Similarity

3 DATA SETS

In this section, we first introduce the cellular signaling data and then show some overall statistical features. We also specifically explain the data we will use for this paper. At last, we introduce the tag data from the subscription data, covering all the users in the same cellular network, together with how the tag data can help us.

3.1 Cellular Signaling Data

In this work, the data processing and model implementation are considered on the cellular signaling data from one of the three major cellular operators, in the Chinese city Hefei. The signaling data are generated from the cellular networks triggered by several necessary interactive protocols among User Equipment (UE), cell towers, and Mobility Management Entity (MME). The data were collected by cellular operators to understand, test, and improve these protocols. Signaling data mainly contain information including an encrypted device ID, time/date, an encrypted cell tower ID (i.e., ID for a base station), and a particular type of procedure denoting why this record is generated. An example of a signaling record and some daily based statistics are shown in Table 1.

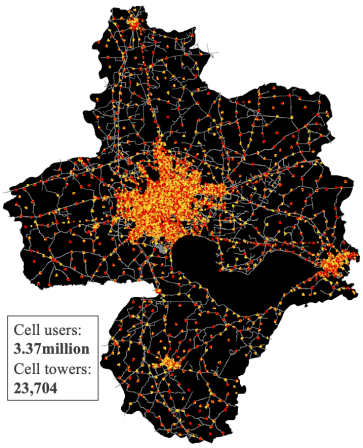

Fig. 7. Cell Towers Distribution

The details of some basic data fields are as follows:

- Timestamp: the time when this record was generated.
- Tower ID: a unique encrypted ID of the currently connected cell tower.
- User ID: a unique encrypted ID for a UE.
- Signalling Type: the related signaling protocol this record belongs to.

There are different signaling types in our data set, but we only focus on one type, i.e., *sRequest*, in this work since we work on the data usage. Notably, we find out that 53% of the raw signaling data are *sRequest* data. Also, a *sRequest* record suggests that the UE requests services from the networks, which can work as an indicator of how active the UE is in the cellular networks. It also indirectly reflects the cellular data usage consumption of this UE [3, 26]. User's mobility status can also be revealed by the signaling data by handovers [12], because the connected tower of the users is revealed in the data, correspondingly we can know the proximate location of the user. In sum, the cellular data usage and mobility status can be inferred via the signaling data.

All the above signaling data are associated with the 23,704 cell towers in the studied city. A signaling-density-based visualization of all cell towers is shown in Fig. 7, where a brighter yellow dot indicates a cell tower with more signaling data associated with it, and a darker red dot indicates a cell tower with fewer signaling data. We observe that signaling data are densely concentrated in the center of the city, together with three smaller counties located in the north, the south, and the east.

We also show some statistics related to our work from this data set in Fig. 8 and Fig. 9. In Fig. 8, we show the number of signaling records generated by all users during a day, and we can see that 80% of users generated

Table 1. Signaling Data Example

Field	Value
Timestamp	2017/06/12 00:23:55.357
Tower ID	110613042
User ID	3B8xZpNZJpTjfOnwYbcdyA==
Web Type	4G
Signaling Type	*sRequest*
Number of Daily Records: 267 millions	
Number of Daily Users: 3.37 millions	
Number of Cell Towers: 23,704	

Table 2. Tag Data Summary

Categories	Tag/Tags(%)	Users/Users(%)
Phone usage	38.64	87.25
Car related	35.87	83.68
Demographic	17.91	46.74
Services	3.95	12.19
Financial	2.14	6.56
Shopping	1.49	2.98
Average number of tags per user: 4		
Number of users with tags: 3.37 millions		

around 50 records per day. By comparing the CDF of all signaling records and *sRequest* records, we find that users tend to generate more *sRequest* records given a certain percentage of users (note that the same percentage of users are different users for two CDF curves). We also show the temporal patterns of the number of records and the number of users during the different time of the day. We observe that the number of records presents a clear diurnal pattern with two rush-hour peaks near 8:00 and 16:00, respectively. However, the second rush hour peak for the number of users is not obvious compared to the records number distribution.

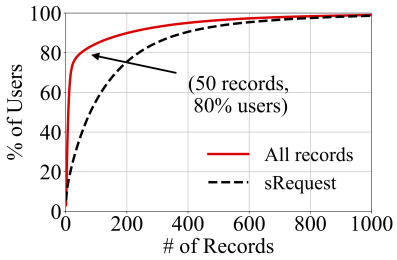

Fig. 8. Dist. of Users' Records

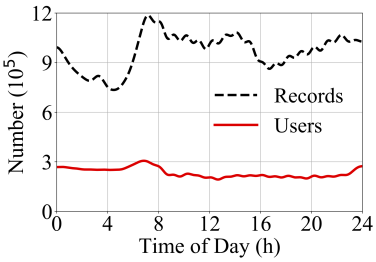

Fig. 9. Users & Records

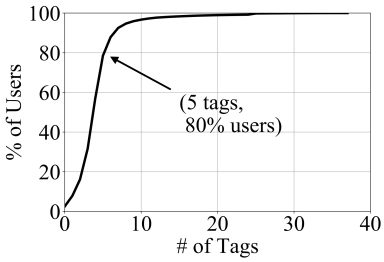

Fig. 10. Tag Distribution

3.2 User Tag Data

We also have access to the user tag data collected from the subscription data and other cellular operator metadata. The subscription data come from a new commercial pattern emerging in China, where IT companies cooperate with cellular operators to dedicate cellular data plan for specific apps or services with a lower cellular data price, e.g., providing a 5GB exclusive data package which can only be used for certain video apps. Then these tags can indirectly provide us with more behavioral information about users. For example, "car" tags can indicate this user may often drive a car, which further indicates that this user may use a navigation app or listen to streaming music, resulting in more cellular data consumption. In total, all subscribed cellular users in our signaling data have tags. For details, we show the distribution of the number of tags per user in Fig. 10. Around 80% of users have fewer than 5 tags, the maximum number of tags for a user is 37, and the average number of tags for all users is 4.

In summary, the six categories and statistics are shown in Table 2. The first column shows the categories, the second column denotes the percentage of corresponding tags of this category out of all the tags, and the last column presents the percentage of users with tags of this category out of all the users. For clarification and brevity, here we show some representative examples of tags for categories. For phone usages, typical tags include video, music, etc.; for car-related category, tags contain car-related properties, such as the user has car(s); demographic tags include gender, age range (e.g., the twenties), etc.; services refer to tags related to services in life, such as Internet services; financial tags include financial services such as stocks; shopping category includes tags like

shopping preferences, etc. More comprehensive analyses from the perspective of the tag data are presented in Section 4.

4 FRAMEWORK

4.1 Overview

The overall goal of this work is to predict the data usage consumption of cellular networks accurately. Specifically, we approach it at the individual user level by considering users' behaviors. The architecture of *CellPred* is depicted in Fig. 11, including preprocessing, mobility and usage prediction, and potential applications based upon it. We reveal the details of each part in the following subsections.

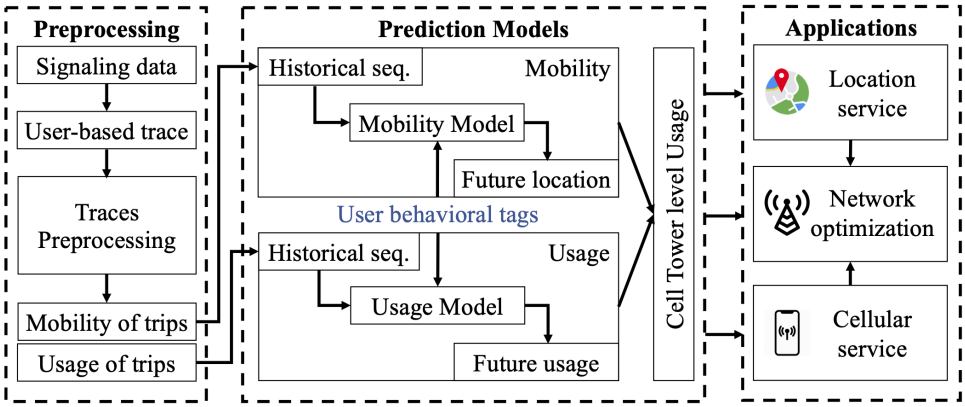

Fig. 11. Overall Framework of *CellPred*

4.2 Preprocessing

The preprocessing part includes raw data processing for preparing input data of mobility and usage prediction. For the cellular data usage inference, we directly use the number of *sRequest* records as a metric to measure the cellular data usage, since *sRequest* is generated when the users are actively using cellphones, as indicated in [53]. For individual user mobility inference, the raw signaling data of users can hardly be treated accurately in terms of mobility analysis due to several reasons, including Ping-Pong Effect [19], outliers, etc. Instead of directly making use of raw signaling data generated from interactions between users and cellular networks, we need to preprocess for the data for further investigation and application. Here we show a sample trace of one user and related statistics in Fig. 12 and Fig. 13 to demonstrate the necessity of the preprocessing step. Several things can be observed here: i) there are certain outliers in this trace, such as points *A* and *B* in Fig. 12 or points (marked by "star") with values as high as a few kilometers (almost 16km) in Fig. 13. These outliers are much further away than other locations in the user's cellular trace. Possible reasons for this is the roaming mechanism of 4G/LTE networks [24]. ii) Ping-Pong Effect exists in real-world signaling records [19], e.g., cell tower connection switches frequently among different cell towers within a short interval. We can observe this from the thick black "line" in Fig. 13, meaning the distance between cell towers from two consecutive records is fixed around 1 km (918 m), which further indicates that the connection is transferring back and forth between two cell towers. iii) trace discontinuity exits, e.g., there are relatively long temporal gaps between different trips for a user during the different time of day, such as the annotated "7.4 hr" in Fig. 13. To achieve better and reasonable results based on the raw signaling data, we implement the next several preprocessing steps.

4.2.1 Trip Segmentation. Due to the long temporal gaps existing in users' traces within one day, we partition users' signaling records into several different trips by setting a 30-minute threshold similar to the implementation

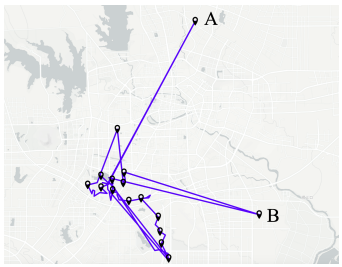

Fig. 12. Sample Trace

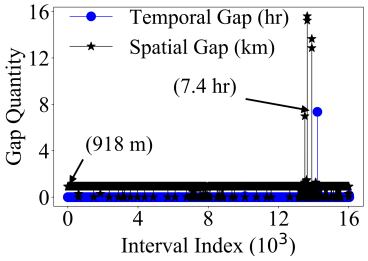

Fig. 13. Trace Statistics

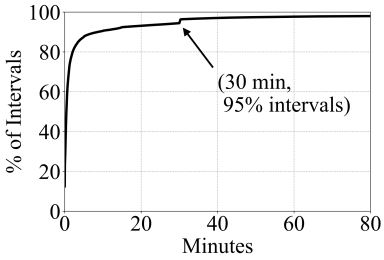

Fig. 14. Interval Distribution

in [22]. That also means if the temporal gap between consecutive records is longer than 30 minutes, we will divide them into two separate trips for further consideration, which is widely used in other cellular data based mobility investigation works [22, 49]. The trip segmentation will provide further convenience and efficiency in our later prediction model, since it is more reasonable to feed the model with trips, i.e., when the users are active in the cellular networks, instead of treating the whole day data as one trip, which would bring lots of empty values due to cellular inactiveness. To justify the choice of the temporal threshold, we show the distribution of time interval of two consecutive records across all users in Fig. 14, where we can see that 95% of the records intervals are shorter than 30 minutes, where the distribution curve "jitters", indicating the distribution difference before and after 30 minutes. Moreover, the choice of the temporal threshold actually poses little impact on the performance of our model since truncation will be implemented for formatting neural networks inputs, which works the same as trip segmentation.

4.2.2 Outlier Filtering. We also need to remove outliers existing in the trips to achieve a satisfying usage and mobility prediction. Here we consider two types of outliers, i.e., spatial outliers and temporal outliers. The first case is that there are spatially outlying records in terms of mobility modeling, where connected cell towers are far away from previously connected cell towers (e.g., 5 km) within a short period (e.g., 10 seconds). We implement the method in [22] to filter such cases. The second case is that if multiple records are showing the connection to the same cell tower within a short time interval, then we regard this as temporally outlying records and only keep one record for further processing. In detail, suppose we have a user's trace expressed as $\{< t_1, CT_1 >, < t_2, CT_2 >, ..., < t_n, CT_n >, \}$, where $< t_i, CT_i >$ indicates the signaling record time t_i and connected cell tower CT_i. The first case can be generalized as:

$$\begin{aligned} Dist(CT_{i-1}, CT_i) > D_{THR_{Max}} \ \& \ Dist(CT_i, CT_{i+1}) > D_{THR_{Max}} \\ t_i - t_{i-1} < T_{THR} \ \& \ t_{i+1} - t_i < T_{THR} \end{aligned} \tag{2}$$

Here the $D_{THR_{Max}}$ is the maximum distance threshold for the distance of two cell towers given a temporal threshold: T_{THR}, which is the time threshold for interval between two signaling records. If the above conditions are satisfied, then the record $< t_i, CT_i >$ is discarded. And the second case can be similarly generalized:

$$CT_{i-1} = CT_i, \ t_i - t_{i-1} < T_{THR} \tag{3}$$

We also discard record $< t_i, CT_i >$ if such conditions are met.

4.2.3 Ping-Pong Effect. Another case we should pay attention to is the Ping-Pong Effect, which will dramatically disturb the performance of user mobility analysis and inference if we directly use the raw signaling trace. Here we utilize the method in [22] at a high level to eliminate the side effects brought by Ping-Pong Effect. We summarize the conditions of Ping-Pong Effect here:

$$\begin{aligned} Dist(CT_{i-1}, CT_{i+1}) < D_{THR_{Min}} \\ t_{i+1} - t_{i-1} < T_{THR} \end{aligned} \tag{4}$$

It describes the scenario that in a temporal window with three consecutive records, if the distance between the cell towers in the first and last records is smaller than a threshold or equal to zero (which is the typically Ping-Pong Effect case) and the temporal interval between these two records is shorter than T_{THR}, then we discard the record $< t_i, CT_i >$. In this way, repetitive signaling data existing between two cell towers can be removed, and the processed data are smoother to reflect a more realistic physical trace of the user.

4.2.4 Temporal and Spatial Granularity. To infer future data usage at the individual user level, we need to quantify the data usage consumption from both temporal and spatial perspectives. That is to say, we need explicit temporal and spatial granularity to quantify data usage consumption status. For the temporal perspective, we first use fine-grained one minute (even shorter than the average interval: 2.85 minutes) as the time slot unit, investigation on impact brought by the time slot length will be presented in Section 5. For the spatial perspective, we need to format the cell tower location data properly. Directly predicting the cell tower ID or associated coordinates is not feasible since the cell tower ID is encoded by rules unknown to us (as the data example shown in Table 1) and exact coordinates are too specific to predict. Considering that the coverage radius of a normal cell tower is 300 meters in the cellular networks of our work, here we partition the city into grids with the size $300m \times 300m$ [20]. Then we cluster the cell towers into the grids according to the locations of the cell towers. After that, we use the gird as a geographical unit to represent the location of the users, providing us with convenience for prediction, also achieves more explainable prediction results. Besides, cell towers deployment presents a clear clustering pattern geographically, as shown in Fig. 15, where we visualize some cell towers (black dots) in the suburban region in Hefei city. Thus it makes more sense to cluster cell towers in grid-based units as location information for mobility prediction. More details about the cell tower formatting will be shown in Section 5.

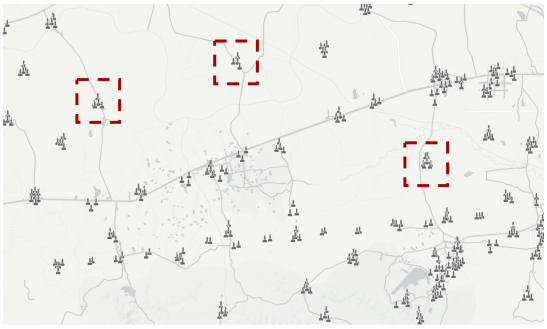

Fig. 15. Cell Tower Clusters

4.3 Behavior-aware Features

To obtain a better prediction performance, temporal sequences of users' locations and cellular data usage are not enough. We need more information to aid the prediction [12]. In this subsection, we investigate potential features that can help the prediction models work better. Specifically, we consider features related to user behaviors in terms of user mobility and data usage and analyze how the tags affect these features. To begin with, we emphasize that all the features extraction are based on trips after data preprocessing, i.e., we analyze and infer mobility and usage based on each trip of users, and there could be multiple trips for one user within a day.

4.3.1 Mobility Status Features. The mobility can be addressed by a broad range of features [2, 57], e.g., the traveling distance of one trip, duration of one trip, locations visited in one trip, etc. We summarize the features we considered here:

- *Radius of gyration*: a metric which can indicate the traveling distance of a user during one trip. The calculation is revealed as the below Equation [2].

$$r_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{r}_i - \boldsymbol{r}_0)^2} \tag{5}$$

Where the $\boldsymbol{r}_0$ denotes the center of mass of the user's trip, and $\boldsymbol{r}_i$ is the location vector of *ith* record, N is the total number of records in one trip. The calculated historical radius of gyration can help us understand the traveling range of one trip, i.e., one aspect of mobility dynamics.

- *Entropy of one trip*: here we also consider the entropy of one trip for a user based on the list of visited grids. The entropy is an indicator of possible visiting locations, i.e., the higher the entropy, the more locations the associated trip covers. Specifically, the entropy is formalized as Equation 6 according to [40].

$$Entropy = -\sum_{i=1}^{n} p_i log(p_i),\ p_i = \frac{e_i}{e} \tag{6}$$

where i = indicator of a certain associated grid
n = the number of grids visited by the user
p_i = the probability of the user visiting grid i
e_i = the number of visits of the user in grid i
e = the total number of visits of the user during one trip

- *Number of trips per day and during night*: we also consider the number of trips during one day and during the night time (18:00-6:00) for a user, where the number of trips is inferred according to the trip segmentation method. The larger number of trips during a day may indicate that the user's visited locations may be more scattered.
- *Average number of trips per day*: we observe from the data that different users can have a different number of trips, and they can also appear on different dates. Thus we also include the average number of trips per day since it also affects user's mobility status.

To see the impact brought by tags, here we visualize the distribution of the features among different users to show how the users with and without certain tags display different feature distribution. For demonstration, we show a few tags affecting the patterns of users regarding different features. The radius of gyration distribution of users with and without the "car" tag is shown in Fig. 16, where we can see that users with the "car" tag tend to have a larger radius of gyration during their trips, denoting they travel longer distances. The corresponding distribution regarding visiting spots entropy of the same groups of users is shown in Fig. 17, where a similar trend can be observed that users with the "car" tag are more likely to visit more locations than users without the "car" tag. Moreover, we show another example of the "navigation" tag, which might also present different patterns of different groups of users. The trip entropy of users with and without the "navigation" tag is shown in Fig. 18, where we can see that users with the "navigation" tag are inclined to have a higher trip entropy, i.e., visiting more places. Specifically, 80% of users with the "navigation" tag have a trip entropy smaller than 2.42, while the corresponding value of the users without the "navigation" tag is 1.04. In general, the users with tags semantically indicating higher mobility dynamics also tend to present higher entropy or radius of gyration.

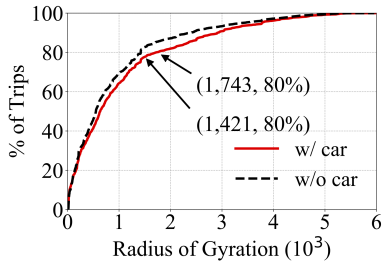

Fig. 16. Radius of Gyration

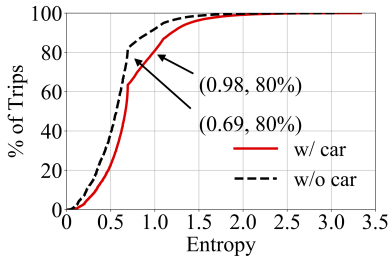

Fig. 17. Trip Entropy: w/wo "car"

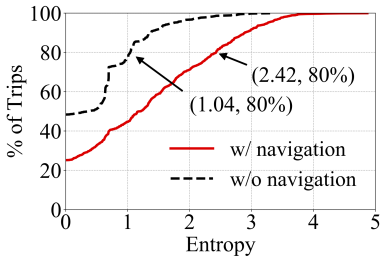

Fig. 18. Trip Entropy: w/wo "navigation"

4.3.2 Usage Data Features. For the data usage traffic, there are also corresponding features that are related to the cellular data usage consumption, such as average cellular data usage per trip, the variance of cellular data usage during a day, etc. In our setting, we use the number of *sRequest* to indicate the data usage consumed by the user [3, 53]. Further, we investigate different usage features to analyze the distribution of two groups of users' data usage consumption, i.e., the amount of data usage consumed per day for each user. We summarize the features considered for the cellular data usage prediction:

- *Average cellular data usage per trip*: The trip here is determined by trip segmentation in data preprocessing. Then we calculate the average cellular data usage during a certain trip.
- *Daily average cellular data usage*: After we have the average cellular data usage for each trip, we calculate the overall data usage per day by summing up the usages normalized by the trip duration. Finally, we can calculate the total daily average of cellular data usage across different days.
- *Variance of cellular data usage*: We also take account of the variance of cellular data usage of a trip, i.e., the variance of usage list for every time slot, indicating the dynamics of data usage during the trip.
- *Daily variance of cellular data usage*: Similarly, we calculate the overall variance of different trips and average them across one day. Then we calculate the overall average across different days for one user.
- *Trip Duration*: The trip duration may also affect the amount of cellular data consumed, and presumably, a longer trip duration will lead to more cellular data consumption.

Similar to the mobility features, we investigate the impact of tags upon usage patterns by quantified features. We first show the distribution of the trip duration in cellular networks in Fig. 19, where users with and without the "video" tag are compared. We can see that users with the "video" tag are more likely to have a longer trip, i.e., a longer period with continuously using cellular networks. The reason behind this may be that users with the "video" tag are users with a data plan subscription to video apps, who tend to watch more videos during the day, i.e., using cellular networks for a relatively long time. Data usage distribution of users with and without the "video" tag is shown in Fig. 20, where users with the "video" tag tend to generate more data usage compared to users without the "video" tag, e.g., 80% of users with the "video" tag generate 287MB data per day while the corresponding value for users without the "video" tag is around 4MB. A similar pattern can be observed from the comparison of users with and without the "music" tag, as shown in Fig. 21. All three figures demonstrate that users with tags semantically indicating heavy cellular usage have higher data consumption compared to users without those tags.

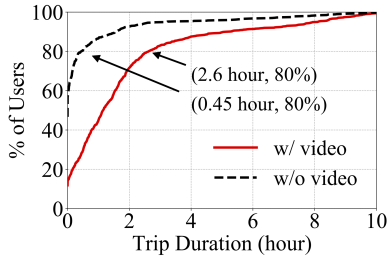

Fig. 19. Trip Duration: w/wo "video"

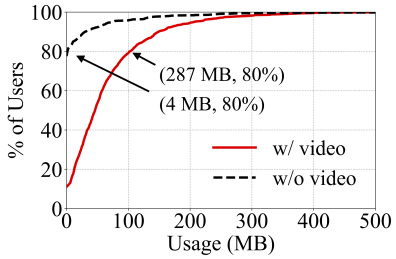

Fig. 20. Data Usage: w/wo "video"

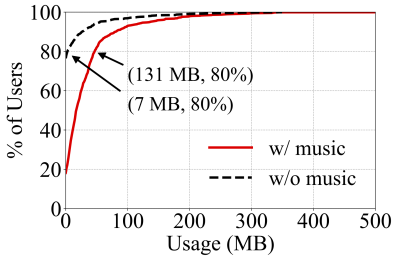

Fig. 21. Data Usage: w/wo "music"

4.3.3 Behavioral Tags. In the above sections, we have shown that different tags can be used as an effective indicator for different groups of users. We also show that users with similar tags have similar patterns including mobility patterns and usage patterns (indicated by Fig. 5 and Fig. 6). The observed patterns motivate us to involve the mobility and usage features gathering from users of different tags. Still, a user may have several tags, which pose a challenge for the consideration of the features. For example, the behavior of a "car" & "video" user may be quite different than a "car" & "WiFi" user, because they may have different cellular data usage patterns despite

possible similar mobility patterns. To start with, we calculate the average of mobility and usage features from the same group of users; we then consider the compound impact on a specific user by including information from different tags. To tackle the problem of implicit influences from different tags, we introduce the soft attention module to put weights on different tags [47]. The details can be found in the next part.

4.4 Prediction Model

In this part, we introduce the details of our prediction model in *CellPred*, including the mobility prediction part and usage prediction part. The overall architecture of the model is shown in Fig. 22.

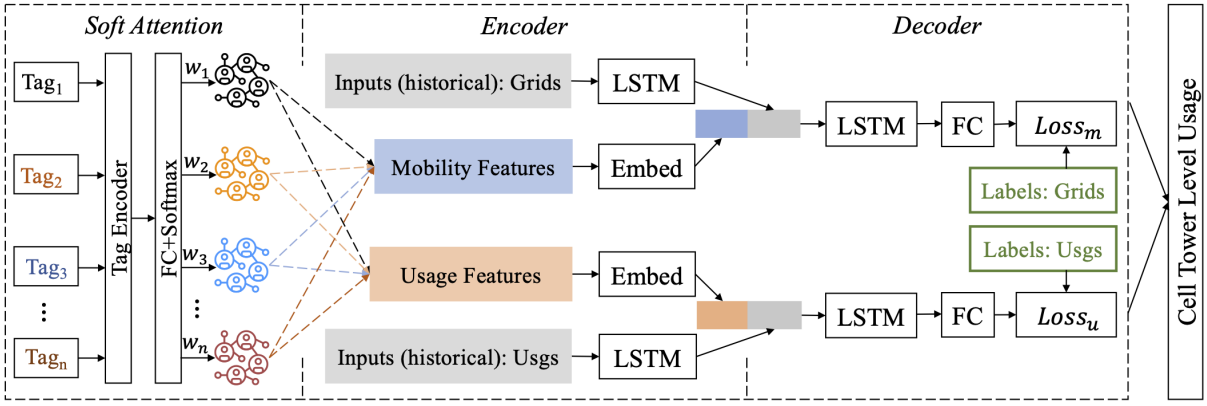

Fig. 22. Prediction Model Architecture of *CellPred*

4.4.1 CellPred: Mobility Prediction. As aforementioned, the historical mobility records are formatted as sequences of grid indexes. In detail, we cluster the cell towers into grids partitioned by the coordinates of cell towers. The two-dimension grid index is mapped into value by one-to-one correspondence, indicated as *Grids* in Fig. 22. We also have extra features extracted from the historical records of users. To achieve a satisfying prediction performance, both the inputs and the features are considered in *CellPred*. We utilize deep learning methods here for the mobility status inference [56], which are suitable for our large-scale and multi-day data sets compared to traditional machine learning solutions such as Hidden Markov State based method [28]. Specifically, we consider Long short-term memory (LSTM) here as our cells in encoder and decoder for mobility status prediction [17], which are widely implemented in works related to mobility prediction, location inference[12, 33] or broad sense of spatiotemporal data investigation [54].

A illustration of LSTM cell is shown in Fig. 23. The processing can be observed from the figure and abstracted as the following equations:

$$g_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{7}$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{8}$$
$$\widetilde{U}_t = tanh(W_U \cdot [h_{t-1}, x_t] + b_U) \tag{9}$$
$$U_t = g_t * U_{t-1} + i_t * \widetilde{U}_t \tag{10}$$
$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{11}$$
$$h_t = f_t * tanh(U_t) \tag{12}$$

We denote the current cell as the cell processing input from time t, the inputs of the cell on the left are from the cell of last epoch, i.e., U_{t-1} and h_{t-1} of epoch $t-1$, and the outputs of the cell denoted as U_t and h_t, also working as the inputs for the cell of next epoch $t+1$. The core characteristic of LSTM is the "gate" consists of sigmoid layer and pointwise product (denoted by the red rectangle in the Fig. 23), which controls the information

flow, i.e., decides how much input information can be through for further consideration. This process is reflected in Equation 7. The processing in the unit starts from the inputs of epoch t, including last epoch output h_{t-1} and current epoch input x_t, which are considered by Equation 8 as i_t; and a potential cell state $\widetilde{U}_t$ calculated by the $tanh$ layer in Equation 9 is combined with i_t for the update of previous cell state U_{t-1}. The cell state information preserving and dropping are completed by Equation 10 and the output h_t is determined by Equation 11 and Equation 12. More details can be found in [17]. The structure with such a subtle mechanism of preserving and dropping information makes LSTM widely used in sequential data processing, such as spatiotemporal sequential data [54] and natural language sequential data [44].

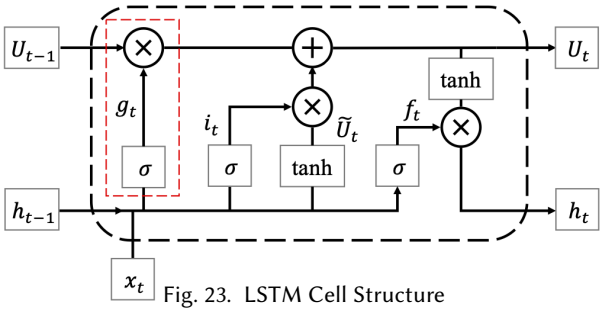

Fig. 23. LSTM Cell Structure

Our mobility part modeling in *CellPred* is also based on the LSTM cells, shown by the upper half of Fig. 22. To be short, for a certain user, we take inputs from the historical mobility records and mobility-related features. Regarding the mobility features, we leverage the mobility features from different groups of users. For all the users, we consider them as different groups by different tags they have. Different features are calculated based on these groups, e.g., we calculate the average trip entropy for all the trips of a group of users, distinguished by a certain tag. Notice that a user may have multiple tags at the same time, i.e., s/he belongs to multiple groups at the same time. Different tags can bring different impact on the mobility and usage of the user. To tackle this problem, we implement a soft attention module in our framework, indicated by the left part in Fig. 22. Firstly, different tags of the user are encoded to vectors by `word2vec` as dimension N. Then these vectors are concatenated as dimension $N * n$, where n is the total number of tags this user has. Then the layer is passed to feed the forward neural networks with fully connected layers (FC), combined with *ReLu*, the outputs are through *softmax* as w_1, w_2, ..., w_n. Define the mobility features from different groups as $mf_1, mf_2, ..., mf_n$, then the final mobility features for this user can be calculated as $\sum_{i=1}^{n} w_i * mf_i$.

After this, all the mobility features are then embedded into high dimension space as tensors with dimension $M * 1$. At the same time, the mobility sequences are truncated as the same length D, where each value represents the location of the user during a time slot, indicated by the partitioned grid. The first $D - i$ values will be used as inputs, and the rest i values will be used as labels for loss calculation. The mobility sequence will be encoded by an LSTM layer with the matching dimension as embedded features, i.e., $M * N$, which can grasp the spatiotemporal correlations in users' trips. Then the embedded features and encoded mobility sequence are concatenated to feed the decoder, including another LSTM layer and layers of FC to transform the high-dimension outputs to final predicted values. The predicted values and real values will be used to calculate the loss and backpropagate it forward. Considering we focus on the accuracy of prediction, we use Mean Absolute Percentage Error (MAPE) as the loss function for mobility status prediction ($Loss_m$), details of this loss function will be given in Section 5.

4.4.2 CellPred: Usage Prediction. For the usage modeling, the structure is generally similar to the mobility modeling. The features are now related to cellular data usage, as we summarized previously. Users with the same tag will be considered as a group in the feature calculation because users with similar tags display similar

behavioral patterns, as demonstrated in Section 2. The common features extracted from users may help enhance the performance of prediction. Besides, users with sparse data, e.g., only a few days of data, can benefit from the features of trips from similar users. Final usage features for a user can also be expressed as the weighted combination of features from different groups of users, i.e., $\sum_{i=1}^{n} w_i * uf_i$, where w_i is the weight for usage features uf_i from the group of users with the tag_i. The usage sequences are with values denoting the cellular data consumed during a certain time slot.

Similarly, we separately embed the usage features into the higher dimension as $M * 1$ vectors and encode the usage sequence by the LSTM layer, which can learn the temporal dependence of users' usage records. Extracted high dimensional features from the inputs are then concatenated for the next layer of LSTM, followed by FCs to transform results to predicted usage values. We also use MAPE as the loss function for usage prediction ($Loss_u$), the detailed implementation is slightly different than the mobility prediction and will be elaborated in Section 5.

4.4.3 CellPred: Individual to Cell Tower. After we have the predicted mobility status $grid_{(i,u,t)}$ and usage status $usage_{(j,u,t)}$ from the user u at the time of t (i and j are the values of $grid$ and $usage$), we then sum all the usage from users who are predicted as associated to the same grid $grid_i$ at the time t. In this way, we can have aggregated data usage at the cell tower. Then the final result is compared to the real data usage consumed at the cell tower level for accuracy calculation. More details will be shown in Section 5.

5 EVALUATION

In this section, we comprehensively evaluate the framework *CellPred* via real-world signaling data set together with subscription tag data. We first introduce our overall evaluation methodology, including how we manage the large-scale data sets, the details about our evaluation setting, metrics we use in the evaluation, the spatial and temporal granularity of the framework implementation, and details about baselines. Then we show the evaluation results, including the performance of multiple baselines, and impact from contextual factors, such as time slot length used in the sequence truncation, the time of the sequence during the day, the day of the sequence during the week, the proportion of data, and location in prediction, etc.

5.1 Evaluation Methodology

5.1.1 Evaluation Data Management. The evaluation is based on the three weeks long signaling data introduced in Section 3. Due to the large size of signaling data (with 20GB data generated per day), we process the data on a high-performance cluster by using big data frameworks, including Hadoop and Spark. Details about this cluster are given here: (i) 12 Hewlett-Packard machines with 2 Tesla K80c on each; (ii) 10 Dell machines with 4 Tesla K80c on each; (iii) 4 Xeon E5-2650 with a half TB memory each; (iv) A series of 800GB SSD and 15TB of HDD. We utilize various techniques to deal with real-world data issues in the data preprocessing, such as spatial and temporal outliers, etc. We also make use of parallel processing to speed up data processing. After data formatted as described in Section 4, we implement *CellPred* by *PyTorch* with GPU to accelerate the training.

5.1.2 Evaluation Setting. We implement the cross-validation scheme for all the evaluations. In detail, we separate three weeks of data into training data (70%), validation data (15%), and testing data (15%) after we have all the mobility and usage sequences for all the users across different days. We evaluate the predicted results by aggregating individual usage up to cell tower level and comparing them to the real usage at the specified time for a specific cell tower. Both the mobility pattern and usage are formatted into sequences of values, where the mobility-related values are quantified by the values mapped from the grid indexes and the cellular data usage related data are quantified by the number of *sRequest* records.

5.1.3 Evaluation Metrics. We measure the prediction accuracy of mobility and usage by one metric: MAPE (Mean Absolute Percentage Error), i.e., the deviation of the predicted results from the true values. The MAPE function is

also working as the loss function in *CellPred*, as mentioned in Section 4. However, the calculation for MAPE of mobility and usage is slightly different. For mobility, the predicted value is formatted into one value representing the hashed grid index. Then we specify the two-dimension grid indexes by the constructed hash table. After that, we calculate MAPE by considering the difference between predicted grid indexes and the truth indexes by using the Euclidean norm. Specifically, the MAPE of mobility prediction is calculated as follows:

$$\begin{aligned} x_{pred},\ y_{pred} &= Hash_{1Dto2D}(grid_{pred}) \\ MAPE_{mobility} &= \frac{\sqrt{(x_{pred}-x_{real})^2+(y_{pred}-y_{real})^2}}{\sqrt{x_{real}^2+y_{real}^2}} \end{aligned} \tag{13}$$

The x_{pred} and y_{pred} represent the predicted 2-dimension grid index, i.e., longitude index and latitude index, x_{real} and y_{real} denote the corresponding real 2-dimension grid index. The MAPE first considers Euclidean norm measuring the distance from the predicted grid the user is associated to the true grid, then it is normalized by the location of the real grid.

The MAPE of usage prediction is straightforward, which is calculated based on the predicted usage value and real usage value, shown as below:

$$MAPE_{usage} = \frac{|usage_{pred}-usage_{real}|}{usage_{real}} \tag{14}$$

Further, we assign the predicted cellular data usage of the grid amount to the cell towers in the grid. If there is only one cell tower in the grid, then we treat the value as cellular data usage of this tower; if there are multiple cell towers, we distribute the cellular data usage evenly to towers within the grid. After we finish that, then we add up the predicted cellular data usage at cell tower level, on which we further calculate the final MAPE to quantify the prediction performance, similarly defined as below:

$$MAPE_{usage} = \frac{|usage_{cell,pred}-usage_{cell,real}|}{usage_{cell,real}} \tag{15}$$

All the results in the later part be measured by this MAPE to illustrate the performance.

5.1.4 Granularity. Given the fact the signaling data records contain information from both temporal and spatial perspectives, i.e., when and where it happens, here we introduce the temporal and spatial granularity of our evaluation. The goal of a proper granularity setting is to evaluate the prediction performance of cellular data consumption effectively. For the temporal granularity, we partition a day (24 hours) into multiple time slots by setting a fixed time slot length as 1 minute, which is even more fine-grained than the average updating frequency of signaling data (2.85 minutes). We can also make the time slot interval longer; we only use 1 minute as an example here to explain how our framework works in a specific setting, longer time slot setting will also be evaluated. In total, there are 1,440 time slots for each day. For the spatial granularity, we partition the Chinese city Hefei into grids with the unit size as $300m \times 300m$, considering that 71% of cell towers have a covering radius as 300 meters according to our cell tower properties data. Using coarser spatial granularity directly may not be reasonable since cell towers cannot be specified within a larger grid, which could lead to error in inferring cell tower level cellular data consumption. A finer granularity setting with sophisticated localization algorithms with extra data and work may be feasible while it is not the focus of this work. In total, there are $407 \times 582 = 236,874$ grids, meaning there are 407 different indexes horizontally (longitudinal) and 582 different indexes vertically (latitudinal). Based on this spatiotemporal partition, we implement our framework for the cellular data usage prediction.

5.1.5 Baseline. There are extensive related works focusing on predicting cellular data usage and mobility inference of users in the cellular networks [20, 48]. Here we implement the following baselines to compare the performances with our *CellPred*. To begin with, we assume the data are preprocessed and well-formatted into

temporal series with values representing cellular data consumed at the cell tower level, with the same one-minute temporal granularity setting, on a daily basis.

- **Naive model**: In the naive model, we first prepare the usage data well at cell tower level, then we directly use historical data from previous days to predict the future day by averaging all the data in corresponding time slots. For example, suppose we want to predict the cellular data consumption of one cell tower in a given minute (e.g., 18:00-18:01) during one certain day, we directly averaging all the data usage during 18:00-18:01 from the previous days and treat the averaged values as predicted values for this cell tower.
- **ARIMA model**: The ARIMA (Autoregressive integrated moving average) model is vastly used in time series related analysis and inference problems, including data usage prediction [25]. The ARIMA model can capture the temporal pattern existing in sequential data, and it works well when there are clear patterns in data and few outliers. Similar to the naive method, we implement the ARIMA model at the cell tower level as a solid baseline to compare with our *CellPred*. The ARIMA is implemented by the model provided in *statsmodel* of Python.
- **D-STN**: We also consider one state-of-the-art deep learning method in the data usage prediction, proposed by *Zhang et al.* [53], using Double Spatio-Temporal Neural Network technique (D-STN) to predict long term mobile data usage. The details of the framework of D-STN can be found in [53]. Considering the difference in the application scenario of their work (spatiotemporal frames as inputs) and our work (individual user sequences as inputs), small changes such as input dimensions are made to ensure the model's effectiveness.
- **CellPred-wo**: We also implement our CellPred at the user level **w**ith**o**ut considering user behavior-related features, i.e., we simply implement the model with temporal series directly inferred from raw signaling data at individual user level without any consideration of behaviors, e.g., mobility regularity. This baseline is for comparison to *CellPred* to present the impact of individual user behaviors.

5.2 Evaluation Results

5.2.1 Overall Performance. In this section, we show our comprehensive evaluation results. For *CellPred* model, we set the prediction time slot (temporal granularity) to 1 minute, the trip duration is set to 30 minutes, i.e., the first 29 values of the trip are inputs, and the last one value is treated as the label. We also tried different trip duration such as 10, 20, and longer duration such as 60; the performance varies little. Here we only use 30 as an example to show the performance. The `word2vec` dimension is set by default as 100, the embedding dimension and LSTM encoding dimension are set as 200, the loss functions we use for mobility and usage are based on MAPE with specification introduced previously. The optimizer we use is Adam optimizer. Learning rate is set as 0.01 for most cases, while we tune it accordingly in the training, e.g., training upon data of different hours are trained separately.

Firstly, we show the overall comparison of our method with different baselines during the different time of day in Fig. 24. Here we separate data into different hours according to their timing information and use all the data during a specific hour as training and testing data, e.g., the results at 1:00 AM are obtained by all data ranging from 1:00 AM to 2:00 AM. Several conclusions can be observed and summarized here: i) we can see that *CellPred* has the lowest MAPE for all 24 hours of the day, denoting the highest prediction accuracy compared to other baselines. The *CellPred-wo* and *D-STN* achieve comparable performance across the different time of day; ii) we can observe a diurnal pattern from baselines (Naive and ARIMA), similar to the distribution of the number of active users and signaling data, as shown in Fig. 9, i.e., a higher MAPE during the day time and lower MAPE during the night. The reason for this phenomenon may be that data usage is more dynamic during day time for cell towers, brought by more diverse user distribution and data usage during that time, e.g., around 4 PM people are gradually off from work and commuting to home or restaurants; iii) for *CellPred* and *CellPred-wo*, the diurnal pattern is not that obvious, but we can still observe a slight MAPE drop during the night. We think one reason is

that in the late night, the activeness in terms of mobility and cellular data usage decreases, another reason is that we observe a relatively large decrease of the signaling data amount in the late night, which both cause the prediction accuracy to increase. We also summarize the statistics of prediction results in Fig. 25. Each bar in the figure represents the mean value of prediction results from a different time of day. The cap of the bar indicates the standard deviation (SD) of the prediction MAPE, e.g., the top of the cap has the value as the sum of mean value and SD. Our *CellPred* model here has the lowest average of MAPE compared to other baselines; also the methods with small SDs suggest the performance are not largely affected by the different time of the day.

We are also curious about which part of the features affect *CellPred* more, e.g., the mobility features or the usage features. Thus, we implement *CellPred* without mobility features or usage features, results are shown in Fig. 26, where *CellPred* without mobility features is denoted as "wo/ mobility features" and *CellPred* without usage features is denoted as "wo/ usage features". We can see that in general, the black line is higher than the green line, meaning that *CellPred* without mobility features performs worse than *CellPred* without usage features. The possible reason is that the mobility prediction is worse without the aid from mobility features, and it follows that the individual user usage is assigned to the wrong grids, leading to worse performance at the aggregated cell tower level.

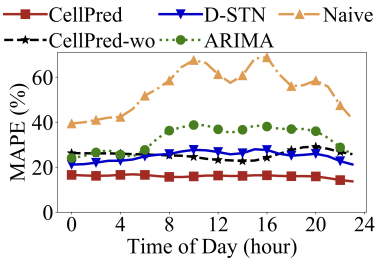

Fig. 24. MAPE Distribution

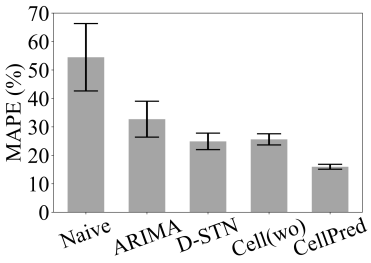

Fig. 25. MAPE Statistics

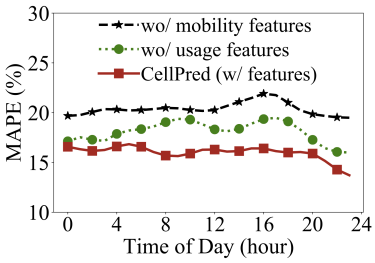

Fig. 26. Impact of Features

5.2.2 Contextual Factors. We also investigate different contextual factors related to human activities and quantify their impact upon the cellular data prediction, including the amount of data we use in the implementation of *CellPred*, different temporal granularity we use in the prediction, different time slot steps used in the prediction, location of the data used in the prediction. These factors may affect the performance of cellular data usage prediction in one way or another, for instance, a coarser temporal granularity is expected to have a better prediction performance since the data dynamics is lower when a longer time slot is used.

Impact of Time Slot Length The time slot length we mentioned in the evaluation methodology can be adjusted, which indicates the temporal granularity we use in the model implementation. By default, we set the time slot length as 1 minute. But different time slot length we use for the prediction may have different results associated with it, here we explore the relationship between time slot length in prediction and cellular data prediction accuracy. Specifically, we compare the performances by setting the time slot length as 1 minute, 5 minutes, 10 minutes, and 20 minutes, respectively. For example, a 5 minutes slot length means we have a user's location every 5 minutes, and we also have his/her cellular data usage amount updated every 5 minutes. The results are shown in Fig. 27, where we can see that MAPE slightly decreases as we increment the time slot length, meaning the prediction accuracy increase, but the change is trivial (within 3%). We also compare it with *CellPred*-wo, in which we can observe a similar pattern. Besides, in consistence with the previous results, the *CellPred* performs better than *CellPred*-wo.

Impact of Proportion of Data Another interesting factor to consider is the proportion of data we use in the model implementation. Since we can see how different proportions of the training data affect the performance of our model, and whether a small proportion of data can achieve a relatively satisfying result. Notice that we directly sample the data of individual users. Different proportions of data are used in the model implementation,

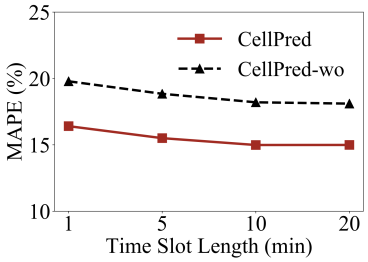

Fig. 27. Impact of Time Slot

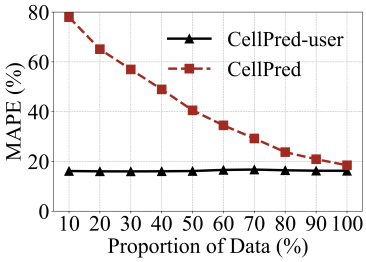

Fig. 28. Impact of Data Proportion

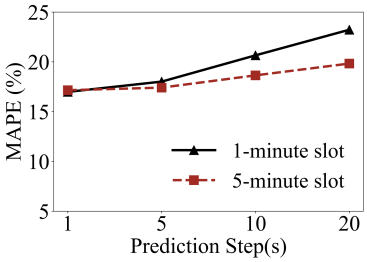

Fig. 29. Impact of Prediction Step

and we show the results in Fig. 28. The black line with triangle markers shows the results of individual user level MAPE, i.e., MAPE of individual user mobility and usage prediction, and the red line with square markers show the final MAPE on the cell tower level. From the results, we can know that different proportions of data used in the data won't affect the individual level prediction too much since the number of users used in training is still substantial (e.g., 1% of 3 million users is still 30 thousand users). But the prediction accuracy of the cell level usage is profoundly affected by the data proportion since the final results on cell towers are aggregated from individual users, the performance is expected to be satisfying only if 100% of users' data are considered.

Impact of Prediction Step The length of prediction steps can reveal more about the performance of the model. We utilize different prediction steps to compare the performance of *CellPred*, e.g., predicting the data usage for the next 1 minute, 5 minutes, and 10 minutes. The results are shown in Fig. 29, where we testify the effect of different prediction steps based on different time slot length (1-minute and 5-minute). We can see that the MAPE grows as we predict more prediction steps, i.e., further future data usage, for both 1-minute and 5-minute time slot length.

Impact of Day of Week Different day of the week is another temporal factor which may affect the performance of prediction, i.e., performance may vary due to the different day in a week. Here we separate data into two categories by the date, i.e., weekdays and weekends, and show the prediction results in Fig. 30, by implementing the model on separated data. We can see that in the late night (12:00 AM to 8:00 AM), prediction performances are evenly matched. While during the daytime, the performance during the weekdays is worse than the weekends. We think it is brought by the fact that people tend to be frequently using cellular networks during weekdays, e.g., using cellphones during commuting from work to home. We can also observe that in the night (8:00 PM to 12:00 AM), the MAPE during weekdays drops fast and is lower thann weekends when it is approaching midnight. A possible reason for this is that people tend to rest early and prepare for work of the next day, while they tend to be actively using cellular networks during weekends nights.

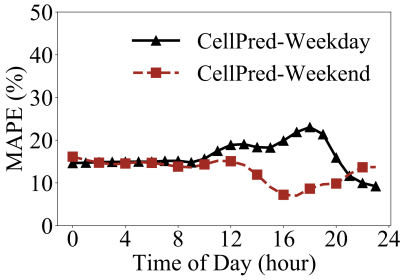

Fig. 30. Impact of Week Time

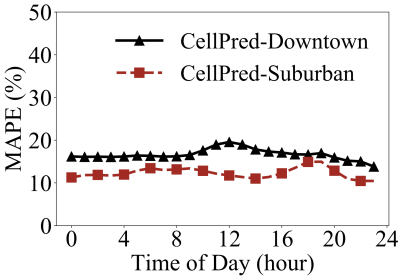

Fig. 31. Impact of Region

Impact of Region Location Moreover, we consider the impact of different regions. Here we divide the whole city into downtown and suburban by the "N 2nd Ring Road", which is the contour of a highway surrounding the downtown region. We consider the region within the surrounding highway as the downtown region and

region outside the highway as the suburban region. Then we implement the model by data from different regions, corresponding results are shown in Fig. 31. From the results, we can see that, in general, the MAPE values in the downtown region are higher than the suburban regions, suggesting that the prediction performance in the suburban region is better than the downtown region. We think this result is caused by the fact that cell tower distribution in suburban regions is sparser than downtown regions, as shown in Fig. 7. Thus it is easier for us to predict the mobility status of users, i.e., their connected cell towers. Once we have the correct prediction on connected cell towers, then it will indirectly improve the final prediction results on cell tower level data usage. Another reason could be that mobility and usage patterns are more clear in suburban regions since there are fewer people and sparser road networks, which simplifies the users' movements and cellular data usage patterns.

6 DISCUSSIONS

In this section, we discuss several major issues related to our work, including lessons learned, limitations, potential applications, privacy issues, and details of public data access.

Lessons Learned: We summarize lessons learned from our work as follows.

- The prediction performance from the individual user level is better than methods focusing on aggregate cell tower level, and the idea of considering user behaviors enhances the mobility and usage pattern prediction, as indicated in Fig. 24, Fig. 25, and Fig. 26.
- The users with similar tags are likely to have similar mobility and usage patterns, as shown in Fig. 5 and Fig. 6. Considering groups of users with similar patterns may be a good approach to boost the prediction performance. It also motivates us to dig deeper into fundamental similarities among users based on signaling data in the future work.
- Different tags related to users' behaviors tend to indicate different usage and mobility patterns, as shown in Fig. 16, Fig. 17, Fig. 18, Fig. 19, Fig. 20, and Fig. 21. Behavioral differences suggested by the tags can be utilized to aid users' mobility and usage prediction.
- The performance of *CellPred* varies little as we use different temporal granularity in the prediction (as shown in Fig. 27). This suggests that we have the potential to achieve a real-time prediction by shortening the time slot length without deteriorating the prediction performance too much.
- It is also found that the performance of *CellPred* changes in different time of day (Fig. 24), day of week (Fig. 30), and different regions (Fig. 31). Patterns in changing prediction performance are consistent with user's behaviors, as we explained in Section 5. It also further indicates how behaviors of users can impact the cellular data usage investigation.

Limitations: The limitations of this work are discussed here: i) one limitation is about the quantification of cellular data usage. In this work, we use the number of *sRequest* from the signaling data to quantify the cellular data usage consumption, which is not entirely equivalent to real cellular data usage. But we argue that this would not affect the efficacy and quality of our work since the traffic consumption data can be directly used to implement our framework; ii) for the mobility prediction part, we did preprocessing to make the mobility status more predictable, e.g., outlier filtering, while this might slightly affect the performance of *CellPred* since we have to distribute the individual user usage to cell towers. But we argue that the preprocessing is necessary, and the evaluation shows the effectiveness of the framework; iii) the balance of training efficiency and performance is subtle. The algorithm we select for the prediction is deep learning, which requires dedicate selection over optimizer function, loss function, learning rate, hidden layer number, hidden layer size, etc. There could be a better combination of hyperparameters existing, while it is also possible for them to cost more training time; iv) another limitation of this work is that it is only based on the signaling data from a particular operator. However, it has been shown in previous work that the cellular networks will generate similar signaling data with a minor variance

based on the technology they are using and users in their networks [11]. Given the data from another service provider, our user behavior based modeling can be implemented accordingly without significant modification.

Potential Applications: The cellular networks have tremendously improved the quality of life by enabling ubiquitous connectivity for cellular users that are physically mobile [11, 55]. In turn, the cellular networks are based on the mobility characterization of network users, i.e., mobility models, e.g., users' mobility characteristics influence network measurement [34], modeling [43], performance prediction [46] and mobility management in cellular networks [31]. Thus, our *CellPred* with user mobility and usage investigation and prediction functions can enable broad applications; data usage prediction in the evaluation at cell tower level is shown as one concrete example. We provide more application examples of individual user level mobility and usage inference here:

- Based on our user-level and tower-level analysis on the understanding of user mobility and activeness, cellular network operators can either temporarily close some cellular base stations to reduce operational costs if reduced usage demand is predicted based on real-time cellular user behaviors, or deploy portable base stations to cope with predicted increasing demand due to irregular real-time user behaviors brought by events [34]. These activities have been widely used in both developing countries, e.g., China, and developed countries, e.g., the United States.
- User behaviors can be used for user profiling, i.e., showing users' interests towards items or services, which are commonly used for marketing purposes, e.g., data plan promotion [58]. In detail, users showing interest in data-heavy services such as video watching can be potential customers of data packages. The user profiling can also be considered for technical purposes, e.g., synthetic data generation [29], where users' characteristics are considered in data generation.
- Users mobility and usage patterns can also be used for service coverage gap modeling [22]), e.g., determining the spatiotemporal coverage of cellular networks; or anomaly detection in cellular networks (e.g., heavy hitter [5], social events [13], traffic anomaly [37]), which can help the operators better manage cellular networks.
- In the coming 5G market, user mobility and usage investigation can also provide guidance for local caching [4], 5G applications related to virtual reality [30] and augmented reality [10]. Since these applications require fast and stable cellular networks connections.

Privacy and Ethics: Our research vision is aligned with maximizing the benefits of the users while minimizing the risk of the users. In particular, our research objective is to work with the cellular providers to improve their cellular services by understanding the cellular demand based on the data they collected legally from their users to benefit their users. For example, metadata of cellular users have been used to perform analysis on load balancing, tower deployment, anomaly detection, malicious cellular usage detection/blocking, access patterns, business opportunities, etc. In particular, cellular data usage modeling and prediction can enable many applications significantly related to cellular service quality, such as load balancing. There is little risk for the users when we are achieving this objective. All user IDs have been hashed into global identifiers by the cellular providers' staff, and we cannot leverage these identifiers to trace back to any individual users. During our project, we respect the privacy of users, and the raw data have not been moved from the cellular providers' internal server. We did not seek any unnecessary data, e.g., payment methods or billing addresses. The users are aware of the potential usage of their information by the operators, and we envision the majority of customers may not be against this work since we aim to maximize their benefits and to minimize the risk during the analyses. As for the voluntary consent, we provide a link of the user privacy policy of operators as in [7–9] (we cite from all three major cellular operators for commercial privacy purpose). This consent form has been provided to a user when signing up for cellular services. If any user does not agree, he or she cannot be granted any service and will not be in the data set we analyzed. Otherwise, all the users in the data set have agreed and consented the data collection process and resultant analyses. Snippets of the policy are provided here: "To ensure that you can enjoy our services, and for the

purpose of accurately charging, we will automatically collect the log data via cell towers, switchboard, terminals, including but not limited to subscription information, orders of services (voice service, short message service, searching, connection records), usage records, purchases records, bills, location information, terminal information, etc. Subscription information mainly covers the communication services and Internet data services, including the registration information and contracts of subscriptions, changes, and terminations actions, together with information on data packages purchasing and extra services registering. Purchasing information mainly includes payment records, arrears information, balance update information, account information, credit information, etc. Terminal information includes the type of terminal, MAC address, OS version, unique identifier (e.g., IMEI). Location information includes GPS coordinates. We will use your information under the following circumstances: To improve your service experience and service quality, or deal with your consulting, complaints, reports, etc., or to recommend better or more suitable services, we will use the collected information for the purpose of our products or services under the constraint of this privacy policy."

Public Data Access: Accessing empirical cellular data sets is vital to mobile computing researches, but such data sets are usually not available for fellow researchers due to the various real-world issues. As an initial step, we will release our sample data in aggregate forms with privacy protection schemes [3] to enable potential following works built upon our work. Details of the releasing data set are disclosed as below: i) period: one week data from 2017-06-13 to 2017-06-19; ii) data amount: covering 2k cellular users, together with their subscription tags; iii) data format: encoded user ID, timing of the record, associated grid ID, record type (specified as *sRequest*); iv) data link: https://www.cs.rutgers.edu/~dz220/data.html.

7 RELATED WORK

A lot of works regarding cellular networks data usage have been performed by using different data sets, concentrating on various topics in cellular networks. We classify the state-of-the-art works by two criteria, i.e., (i) whether they focused on the aggregate level or the individual level, and (ii) whether they consider user behaviors or not. Table 3 shows the work in four disjoint categories, which positions our work compared to others. In the following parts, we discuss the related works based on the behavior criteria.

Behavior-agnostic works: In this category, researchers are mainly focusing on network diagnose problems to optimize the design of network systems or protocols. *Li et al.* [27] make use of smartphone collected data at the network control protocol level to implement parallel processing in protocols, achieving a data access latency reduction. In [26], researchers design and implement their tools to collect protocol data from both the control plane and data plane, in which hidden problems are revealed and insights are given. There are also several papers working on data usage prediction at aggregated cell tower level by different learning-based neural networks models [45, 53]. These works generally focused on the aggregated level. Another direction is to work on the individual level, such as modeling human mobility with call detailed records [12, 15, 23, 36]. However, all those works only directly models on the data without considering individual behaviors.

Table 3. Cellular Networks Based Works

	Aggregate Level	Individual Level	
Behavior-agnostic	[26, 27, 45, 53]	[12, 15, 23, 36]	
Behavior-aware	[14, 35, 38, 43]	App Usage [39, 42]	Data Usage **Our work**

Behavior-aware works: In this category, papers usually leverage collected detailed data to solve user behavior-related problems. *Tu et al.* [38] analyze the dependency and coupling effect upon *CSFB* based voice services and IP based data sessions in 4G LTE networks, possible fixes are also proposed to this issue. Measurements about how users interact with their devices are provided through controlled experiments in [14], and some insights are shown based on the measurements. A prototype measurement service based on the insights is also given. In [35], researchers unveil fine-grained geospatial and temporal correlations existing in the cellular networks via data

from RAN (Radio Access Network) and CN (Core Network). Similarly, in [43], the authors focus on spatiotemporal modeling and prediction upon cellular networks through deep learning techniques. These works generally study the behaviors at the aggregate level. In contrast, recently works [39, 42] analyze individual behaviors such as mobile application usage based on internet access records. Our work differs from the previous works in that we focus on behavior-aware cellular traffic prediction, which has not been explored before.

In short, despite all the above-mentioned state-of-the-art works, there is no previous work in predicting individual cellular usage with consideration of user behaviors. To our knowledge, our paper is the first paper in this category, i.e., prediction cellular traffic by uncovering individual user behaviors patterns to advance the state of the art.

8 CONCLUSION

In this paper, we design and implement a framework called *CellPred* to predict the cellular data usage consumption in cellular networks. Specifically, we focus on making use of individual level mobility patterns, cellular data usage patterns, and behavior tags, which distinguish us from the previous works. Our evaluation is based on a three-week-long signaling data set with a daily amount of 267 million records, with a daily size of 20GB from more than 3 million users, together with data subscription tags. Compared to existing works, our study is built upon from a behavior-aware individual level, and we provide better performance compared to the existing aggregate cell tower level cellular data prediction. Based on our results, we also present some unique patterns and insights, potential applications, which can be beneficial for cellular operators and fellow researchers in cellular networks and human behavior related areas.

ACKNOWLEDGMENTS

This work is partially supported by NSF 1849238 and 1932223. We thank the associate editor and reviewers for the detailed and insightful feedback for this work. We especially thank Dr. Dashan Guo, who provided insight and expertise that greatly assisted the research.

REFERENCES

[1] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C González. 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation research part c: emerging technologies* 58 (2015), 240–250.

[2] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. 2018. Human mobility: Models and applications. *Physics Reports* 734 (2018), 1–74.

[3] Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. 2015. A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Scientific data* 2 (2015), 150055.

[4] Federico Boccardi, Robert W Heath, Angel Lozano, Thomas L Marzetta, and Petar Popovski. 2014. Five disruptive technology directions for 5G. *IEEE communications magazine* 52, 2 (2014), 74–80.

[5] Pedro Casas, Pierdomenico Fiadino, and Alessandro D'Alconzo. 2016. Machine-Learning Based Approaches for Anomaly Detection and Classification in Cellular Networks.. In *TMA*.

[6] Bin Bin Chen and Mun Choon Chan. 2009. MobTorrent: A framework for mobile Internet access from vehicles. In *INFOCOM 2009, IEEE*. IEEE, 1404–1412.

[7] China Mobile. 2017. China Mobile Personal Information Privacy Policy. http://gd.10086.cn/gmccapp/webpage/protocol/safePro.html. (2017).

[8] China TeleCom. 2017. China TeleCo Personal Information Privacy Policy. https://www.189.cn/tj/sy_ycgg/100506.html. (2017).

[9] China UniCom. 2017. China UniCom Personal Information Privacy Policy. http://www.10010.com/net5/front/include/index_a/privacypolicy.html. (2017).

[10] Melike Erol-Kantarci and Sukhmani Sukhmani. 2018. Caching and computing at the edge for mobile augmented reality and virtual reality (AR/VR) in 5G. In *Ad Hoc Networks*. Springer, 169–177.

[11] Zhihan Fang, Fan Zhang, Ling Yin, and Desheng Zhang. 2018. MultiCell: Urban population modeling based on multiple cellphone networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 106.

[12] Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. 2018. Deepmove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 1459–1468.

[13] Rosario G Garroppo and Saverio Niccolini. 2018. Anomaly detection mechanisms to find social events using cellular traffic data. *Computer Communications* 116 (2018), 240–252.

[14] Aaron Gember, Aditya Akella, Jeffrey Pang, Alexander Varshavsky, and Ramon Caceres. 2012. Obtaining in-context measurements of cellular network performance. In *Proceedings of the 2012 Internet Measurement Conference*. ACM, 287–300.

[15] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *nature* 453, 7196 (2008), 779.

[16] Hadi Habibzadeh, Zhou Qin, Tolga Soyata, and Burak Kantarci. 2017. Large-scale distributed dedicated-and non-dedicated smart city sensing systems. *IEEE Sensors Journal* 17, 23 (2017), 7649–7658.

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[18] Wenlu Hu, Ziqiang Feng, Zhuo Chen, Jan Harkes, Padmanabhan Pillai, and Mahadev Satyanarayanan. 2017. Live Synthesis of Vehicle-Sourced Data Over 4G LTE. In *Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems*. ACM, 161–170.

[19] Tiziano Inzerilli, Anna Maria Vegni, Alessandro Neri, and Roberto Cusani. 2008. A location-based vertical handover algorithm for limitation of the ping-pong effect. In *2008 IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*. IEEE, 385–389.

[20] Shenggong Ji, Yu Zheng, and Tianrui Li. 2016. Urban sensing based on human mobility. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1040–1051.

[21] Yunhan Jack Jia, Qi Alfred Chen, Zhuoqing Morley Mao, Jie Hui, Kranthi Sontinei, Alex Yoon, Samson Kwong, and Kevin Lau. 2015. Performance characterization and call reliability diagnosis support for voice over lte. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. ACM, 452–463.

[22] Shan Jiang, Joseph Ferreira, and Marta C Gonzalez. 2017. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Transactions on Big Data* 3, 2 (2017), 208–219.

[23] Chaogui Kang, Stanislav Sobolevsky, Yu Liu, and Carlo Ratti. 2013. Exploring human movements in Singapore: a comparative analysis based on mobile phone and taxicab usages. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. 1.

[24] Farooq Khan. 2009. *LTE for 4G mobile broadband: air interface technologies and performance*. Cambridge university press.

[25] Hyun-Woo Kim, Jun-Hui Lee, Yong-Hoon Choi, Young-Uk Chung, and Hyukjoon Lee. 2011. Dynamic bandwidth provisioning using ARIMA-based traffic forecasting for Mobile WiMAX. *Computer Communications* 34, 1 (2011), 99–106.

[26] Yuanjie Li, Chunyi Peng, Zengwen Yuan, Jiayao Li, Haotian Deng, and Tao Wang. 2016. Mobileinsight: Extracting and analyzing cellular network information on smartphones. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 202–215.

[27] Yuanjie Li, Zengwen Yuan, and Chunyi Peng. 2017. A control-plane perspective on reducing data access latency in LTE networks. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. ACM, 56–69.

[28] Wesley Mathew, Ruben Raposo, and Bruno Martins. 2012. Predicting future locations with hidden Markov models. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. ACM, 911–918.

[29] Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, Kolar Purushothama Naveen, and Carlos Sarraute. 2017. Mobile data traffic modeling: Revealing temporal facets. *Computer Networks* 112 (2017), 176–193.

[30] Jason Orlosky, Kiyoshi Kiyokawa, and Haruo Takemura. 2017. Virtual and augmented reality on the 5G highway. *Journal of Information Processing* 25 (2017), 133–141.

[31] Utpal Paul, Anand Prabhu Subramanian, Milind Madhav Buddhikot, and Samir R Das. 2011. Understanding traffic dynamics in cellular data networks. In *INFOCOM, 2011 Proceedings IEEE*. IEEE, 882–890.

[32] Zhou Qin, Zhihan Fang, Yunhuai Liu, Chang Tan, Wei Chang, and Desheng Zhang. 2018. EXIMIUS: A measurement framework for explicit and implicit urban traffic sensing. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. ACM, 1–14.

[33] Amin Sadri, Flora D Salim, Yongli Ren, Wei Shao, John C Krumm, and Cecilia Mascolo. 2018. What will you do for the rest of the day?: An approach to continuous trajectory prediction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 186.

[34] M Zubair Shafiq, Lusheng Ji, Alex X Liu, Jeffrey Pang, Shobha Venkataraman, and Jia Wang. 2016. Characterizing and optimizing cellular network performance during crowded events. *IEEE/ACM Transactions on Networking (TON)* 24, 3 (2016), 1308–1321.

[35] M Zubair Shafiq, Lusheng Ji, Alex X Liu, Jeffrey Pang, and Jia Wang. 2015. Geospatial and temporal dynamics of application usage in cellular data networks. *IEEE Transactions on Mobile Computing* 14, 7 (2015), 1369–1381.

[36] Xuan Song, Hiroshi Kanasugi, and Ryosuke Shibasaki. 2016. DeepTransport: Prediction and Simulation of Human Mobility and Transportation Mode at a Citywide Level.. In *IJCAI*, Vol. 16. 2618–2624.

[37] Kashif Sultan, Hazrat Ali, and Zhongshan Zhang. 2018. Call detail records driven anomaly detection and traffic prediction in mobile cellular networks. *IEEE Access* 6 (2018), 41728–41737.
[38] Guan-Hua Tu, Chunyi Peng, Hongyi Wang, Chi-Yu Li, and Songwu Lu. 2013. How voice calls affect data in operational LTE networks. In *Proceedings of the 19th annual international conference on Mobile computing & networking*. ACM, 87–98.
[39] Zhen Tu, Runtong Li, Yong Li, Gang Wang, Di Wu, Pan Hui, Li Su, and Depeng Jin. 2018. Your Apps Give You Away: Distinguishing Mobile Users by Their App Usage Fingerprints. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies archive* 2, 3 (2018), 138.
[40] Maarten Vanhoof, Willem Schoors, Anton Van Rompaey, Thomas Ploetz, and Zbigniew Smoreda. 2018. Comparing Regional Patterns of Individual Movement Using Corrected Mobility Entropy. *Journal of Urban Technology* 25, 2 (2018), 27–61.
[41] Guang Wang, Wenzhong Li, Jun Zhang, Yingqiang Ge, Zuohui Fu, Fan Zhang, Yang Wang, and Desheng Zhang. 2019. sharedCharging: Data-Driven Shared Charging for Large-Scale Heterogeneous Electric Vehicle Fleets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–25.
[42] Huandong Wang, Yong Li, Sihan Zeng, Gang Wang, Pengyu Zhang, Pan Hui, and Depeng Jin. 2019. Modeling Spatio-Temporal App Usage for a Large User Population. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–23.
[43] Jing Wang, Jian Tang, Zhiyuan Xu, Yanzhi Wang, Guoliang Xue, Xing Zhang, and Dejun Yang. 2017. Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach. In *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*. IEEE, 1–9.
[44] Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 225–230.
[45] Xu Wang, Zimu Zhou, Fu Xiao, Kai Xing, Zheng Yang, Yunhao Liu, and Chunyi Peng. 2018. Spatio-temporal analysis and prediction of cellular traffic in metropolis. *IEEE Transactions on Mobile Computing* (2018).
[46] Xu Wang, Zimu Zhou, Zheng Yang, Yunhao Liu, and Chunyi Peng. 2017. Spatio-temporal analysis and prediction of cellular traffic in metropolis. In *2017 IEEE 25th International Conference on Network Protocols (ICNP)*. IEEE, 1–10.
[47] Yequan Wang, Minlie Huang, Li Zhao, et al. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 606–615.
[48] Jing Wu, Ming Zeng, Xinlei Chen, Yong Li, and Depeng Jin. 2018. Characterizing and Predicting Individual Traffic Usage of Mobile Application in Cellular Network. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, 852–861.
[49] Xiaoyang Xie, Yu Yang, Zhihan Fang, Guang Wang, Fan Zhang, Fan Zhang, Yunhuai Liu, and Desheng Zhang. 2018. coSense: Collaborative Urban-Scale Vehicle Sensing Based on Heterogeneous Fleets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 196.
[50] Fengli Xu, Yuyun Lin, Jiaxin Huang, Di Wu, Hongzhi Shi, Jeungeun Song, and Yong Li. 2016. Big data driven mobile traffic understanding and forecasting: A time series approach. *IEEE transactions on services computing* 9, 5 (2016), 796–805.
[51] Yu Yang, Xiaoyang Xie, Zhihan Fang, Fan Zhang, Yang Wang, and Desheng Zhang. 2019. VeMo: Enabling Transparent Vehicular Mobility Modeling at Individual Levels with Full Penetration. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.
[52] Demetrios Zeinalipour-Yazti, Christos Laoudias, Constandinos Costa, Michail Vlachos, Maria I Andreou, and Dimitrios Gunopulos. 2012. Crowdsourced trace similarity with smartphones. *IEEE Transactions on Knowledge and Data Engineering* 25, 6 (2012), 1240–1253.
[53] Chaoyun Zhang and Paul Patras. 2018. Long-term mobile traffic forecasting using deep spatio-temporal neural networks. In *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 231–240.
[54] Chuanting Zhang, Haixia Zhang, Jingping Qiao, Dongfeng Yuan, and Minggao Zhang. 2019. Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data. *IEEE Journal on Selected Areas in Communications* 37, 6 (2019), 1389–1401.
[55] Desheng Zhang, Tian He, and Fan Zhang. 2017. Real-time human mobility modeling with multi-view learning. *ACM Transactions on Intelligent Systems and Technology (TIST)* 9, 3 (2017), 1–25.
[56] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Thirty-First AAAI Conference on Artificial Intelligence*.
[57] Kai Zhao, Sasu Tarkoma, Siyuan Liu, and Huy Vo. 2016. Urban human mobility data mining: An overview. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 1911–1920.
[58] Liang Zheng, Carlee Joe-Wong, Chee Wei Tan, Sangtae Ha, and Mung Chiang. 2017. Customized data plans for mobile users: Feasibility and benefits of data trading. *IEEE journal on selected areas in communications* 35, 4 (2017), 949–963.