User Mobility from the View of Cellular Data Networks

Ying Zhang
Ericsson Research
Email: ying.zhang@ericsson.com

***Abstract*—Understanding the user mobility is essential to resource optimization and algorithm evaluation in mobile networks, such as network planning, content distribution, and evaluation of hand-over mechanisms. Existing human mobility models focus on extracting mobility patterns from Call Detail Records (CDRs) or WiFi traces. While the former only captures movements during phone calls, the latter does not provide direct answers to mobility of cellular network users in a large scale. In this paper, we take the first step to investigate if the mobility properties derived from cellular data traffic is different from the previous findings using other data source, especially the commonly used CDR based approach. We present a comprehensive characterization of the mobility patterns from the cellular data networks' perspective, using a set of systematic methods. We find that the data network records can provide finer granularity of location and movement information. Three different temporal movement patterns are identified. Furthermore, we propose a new method for predicting future application usage given the mobility patterns and show promising results.**

I. Introduction

Cellular data networks have become the fundamental component of mobile access, and have become ubiquitous, fueled by the user-friendly mobile devices (*e.g.,* cellphones, netbooks, and tablet devices) as well as the plethora of mobile applications. Indeed, mobile data traffic has surpassed voice on a global basis and the pronounced growing trend continues [1]. Given its pervasive usage, cellular data access potentially provide more information about the user than the traditional voice traffic, for example, for identifying the locations and movements of each user.

User mobility modeling has myriad usage in resource optimization, infrastructure planning, real-time network service provisioning, and new mechanism evaluations. For example, it can be used to study the effect of the Mobility Areas (MAs) and Tracking Areas (TAs) design and evaluate algorithms of reducing communication signals for handover. The location and mobility information also drives a vibrant ecosystem of location-based services, *e.g.,* personalized proximity advertising, itinerary recommendation, destination based contextualized reminders, and pre-caching of data.

The proliferation of mobile devices such as cellphones makes it possible to understand the human movements. In the past two decades, many attempts have been made to representatively model the real people's mobility. A number of methods can be used to identify a mobile device's physical location, such as cell tower look up, wifi access point lookup, triangulation of cell tower and wifi access point, and Global Positioning System (GPS), with varying accuracy levels and power consumptions. One body of existing modeling work relies on logging locations via these techniques on the phones. They are limited to a small group of participating users and a small geographic area such as a university campus [2], [3], [4]. Other attempts use the the Call Detail Records (CDRs) collected by cellular network operators to deduce models for large geographic areas with diverse populations. CDRs contains the identity of the cell tower the phone was associated when a voice call was placed or a text message was received. By correlating the cell IDs with the geographic locations of the towers, the CDR based methods can accurately capture many aspects of the human mobility [5], [6], [7], [8], [9]. More recently, researchers have use location information from the location sharing services (LSS), *e.g.,* Facebook, Foursquare, that allows users to enter their location information from GPS and share with friends on social networks [10].

978-1-4799-3360-0/14/$31.00 ©2014 IEEE

Despite the promising results demonstrated by the CDR based approach, because of the dominance and growth of cellular data traffic, in this paper, we seek answers to the following question, *will the mobility models exhibit different properties by analyzing the user data network records?* The hypothesis of the presence of such difference stems from the following three observations. First, evidently people use the data service increasingly more often than the voice service [11]. CDRs are only generated when a phone engages in a voice call or text messages. But people uses apps for music, news, gaming, and even chatting more frequently, even during a trip. Second, the usage pattern of data services is different from that of voice services. On the one hand, many apps switch from cellular to WiFi connectivity whenever the latter is available. It implies more hidden locations from the view of cellular data networks. On the other hand, many apps may have periodic network activities in the background [12], providing information even when users are not actively using the phone. Third, if such distinction exists, then understanding the difference can help better analyze the mobility's impact on cellular data networks, cache content for real time data services, and plan resources for data network infrastructure.

More specifically, in this paper, we aim at answering the question: is the mobility characteristics derived from the cellular data network usage different from those derived from other data sources, *e.g.,* CDRs and location sharing services? In other words, we examine the mobility patterns from the view of cellular data networks: constructing human movement traces from the sequence of cell tower IDs which provide data services to the users. Despite that many recent work starts to look into the cellular data traffic [13], [14], we make the

following contributions in this work.

- To the best of our knowledge, we are the first to systematically study the mobility patterns from the cellular data network usages. We conduct comprehensive characterization of the mobility properties and compare them with the CDR based and LSS results, among which we emphasize more on CDR results due to the more available published results. Note that the goal of this paper is not to propose a new mobility model, but to examine the common properties used in mobility models, such as the dominant locations, radius of gyration, and the location entropy. Though many blanks need to be filled in to construct a new mobility model, we believe these analysis, offering new insights, is one important step towards finding the new model.
- We propose a systematic analysis methodology that considers inaccuracies from this type of data. For the purposes of load balancing and signal attenuation, cell towers usually have overlapping coverage, which then causes a device to oscillate between several alternative towers even though the device remains still. Such oscillation introduces false positives in detecting true movements. We develop a probability based method to reduce such false positives. We applied our methodology on one week of data on two large European cities in order to understand the temporal and spatial diversities. Although we understand our results may not be applicable to other regions with significantly different population and density, we believe that our methodology can be applied to analyze this type of data in general.
- Finally, we explore the predictability of the locations of a user, the trip patterns, and the application usage associated with given locations. Furthermore, we propose a prediction method that forecasts which application will be used in the next location for each user. We can achieve over 90% accuracy for 80% of the users, showing high potential for using this method in content prefetching.

Among a number of interesting results that we present, the key ones are summarized below.

1) The data network records provide more information about the users' locations, with relatively smaller silence periods, during which the locations are unknown. The main reason for this finding is the more frequent usage of mobile data networks, compared to making voice calls. Though maybe expected, our finding confirms and quantifies such differences and its impact on mobility studies.
2) We identify three groups of users with distinct temporal access and mobility patterns. One group uses the data services evenly across time, the second group favors noon time, while the third group mostly use the network after work in the evening. We further found that these three groups of users have different distributions of dominating cells, but with no significant difference in terms of traveling distance.
3) From the predictability analysis, we found that overall the predictability is higher for locations in our approach, compared to the CDR based approach. The predictability of application usage in a given location diverse, depending on the application types and the importance of the locations.
4) Finally, it is feasible to predict the future application usage, based on the locations and the previous application usage. The prediction accuracy is significantly improved if considering the previous application usage in the model. Moreover, users with similar physical and cyber appearance can also be used to improve the accuracy.

The rest of the paper is organized as follows. Section II introduces our data set, data preprocessing and analysis methodology. The characterization of mobility from cellular data network is presented in Section III as well as the comparison with other models. From the observations, we propose a model to predict the application usage in Section IV. Finally, we discuss the related work in Section V and conclude in Section VI.

II. Methodology

In this section, we will provide details of the datasets. Then we introduce our definitions, methods to identify locations, as well as movements, and metrics of predicability.

A. Data set

We extracted locations for each user from two sources of data. First, before a mobile device and a network server can communicate, the network set up a Packet Data Protocol (PDP) context (*i.e.,* a link or connections that allows them to communicate). During the PDP context establishment, the BS that connects to the mobile device is sent in the GTP-C message. However, this is not updated afterwards. The second source of the location is the General Performance Event Handling (GPEH) logs [15], which records individual cell change events in the Radio Network Controller (RNC). Since we focus on the mobility from the data network usage's perspectives, we correlate the above data sources of locations with the GTP-U traffic. We kept the a summary of the users' data traffic (*i.e.,* the applications used by this user within this minute), as well as the cell level locations in this minute. If there is no data network activity within the minute, then we do not have a record.

Our data is collected at two large HSPA networks in two large cities in Europe in 2011, with one metropolitan area in southern Europe and the other one also a large city in the Scandinavian region. Dataset 1 contains 285K unique IMSI (International mobile subscriber identity), and 12K cell towers, lasting for two weekdays. Dataset 2 contains 152K IMSI and 9K base stations, lasting for eight days with both weekdays and weekends. Note that, we are aware that the mobility patterns may vary across regions, thus, we analyze the two datasets separately and only emphasize on common observations between the two dataset. Though trying our best, we admit that some observations may still be specific to our dataset and may not be generalized. But our analysis methodology should be applicable to any data.

The cell level location contains the mobile country code (MCC), mobile network code (MNC), Location Area Code (LAC), and the Cell ID (CID). We further map them to its longitude and latitude using The Google Maps Geolocation API [16].

B. Definitions and pre-processing

Here, we explain how we extract mobility patterns from the trace. In the first step, we organize our data into a sequence of N three-tuple records for U users. The i^{th} record for each user u is:

$$r_i^u =< t_i^u, c_i^u, A_i^u >, i = 1 \dots N \quad (1)$$

The fields are the timestamp, the cell ID, and the set of applications that this user uses, respectively. If we cannot identify the traffic from neither the agent fields nor the HTTP requests, the field $A_i^u = \emptyset$. We use a consistent notation to capture the silence period, *i.e.,* when the user does not have any data traffic, which is denoted as $< t_i^u, *, * >$.

Next, we group consecutive records according to the rule below. We define the stillness duration of user u at time t to be:

$$d_i^u = t_j^u - t_i^u, \forall i \le i' \le j, c_{i'}^u = c_i^u \& c_{j+1}^u \neq c_j^u \quad (2)$$

Essentially it is the duration of staying at cell c_i^u since time t_i, until u moves to a different cell ID c_{j+1}^u. Note that it also covers the scenario of user going offline, *i.e.,* $c_{j+1}^u = *$, the duration of previous cell terminates before $(j+1)^{th}$ record. In the following, we use c_{j+1}^u to represent the first different cell after c_i^u.

Finally, we identify travel paths from a sequence of records.

$$P_{i \to j}^u = r_i^u, r_{i+1}^u, \dots, r_j^u, \forall i \le i' \le j, d_{i'}^u < \delta \quad (3)$$

Here δ is a pre-defined threshold of the maximum stillness in a trip. This definition allows transient silence periods interleaved with data service periods. Since we do not have the explicit data session termination signals in our data, we identify the silence period if no data records are observed for a relatively long period (δ).

C. Handling oscillation

The oscillation phenomenon happens when a user is within the coverage of two or more cells. It could be caused by the fluctuation of signal strengths from these cells or load balancing purpose. For example, a user was associated with cell x_1 at time t, and switched to cell x_2 located 2km away after 1 second and immediately back to x_1 in the next second. However, the intuition of detecting oscillation is that the patterns typically differ when the user is staying still but pingponging between cells, and when he is moving between two places that are sufficiently far apart.

We take three steps to eliminate the noise introduced by oscillation or pingpong phenomenon.

- For each cell, compute the set of neighboring cells that could be oscillation candidates. For each cell m, we compute the conditional probability for any other cell n:

$$P(n \mid m) = \frac{\sum_i (Pr\,(c_{i+1} = n \mid c_i = m))}{U}$$

  Then n is considered as an oscillating alternative of cell m, if $P(n \mid m) > p_\delta$. p_δ is a pre-defined threshold based on the probability distribution.

- Remove candidates with far distance. Oscillation can only occur between cells close to each other. Thus, we further reduce the candidate set: n is an oscillating alternative of cell m, if $dist(m,n) < dist_\delta$. We use $dist_\delta = 70km$, because the maximum radius of a cell defined by 3GPP is 35 km. After these two steps, we obtain a candidate set for each cell. The distance between two base stations is computed using the Google Maps Geolocation API [16].

- Remove oscillation for each trip $P_{i \to j}^u$. For each cell in the path, we check if any of its candidate cells that is also in this path.Then we pick the cell which occurs most often in this path and replace its other oscillating cells. For example, if $P = \{x_1, x_2, x_1, x_3, x_4\}$ and the candidate set of x_1 is x_2. We replace x_2 with x_1 in this sequence, because x_1 occurs more in this path than x_2.

There are systematic methods [17] that focus on solving the oscillation problem. In this work, we use this simple heuristics that are efficient to apply to large-scale data. In the future, we plan to compare with other methods.

D. Important locations

Previous work has shown that people spend most of their time at a few important places [6]. Most previous models focus on the two most important locations, "home" and "work". They extract these two locations either from other data source, *e.g.,* census [5], or using the time-of-day heuristics: considering a location as a users' home if he spends most of his time between 10 PM and 6 AM at this location [13].

From the view of the data network, however, it is not clear if home and work will exhibit to be the most dominant locations, primarily because of the availability of WiFi connectivity. Therefore, we take a general approach to investigate this problem. We define the important places of a user from two dimensions of significances: *the duration and the occurrence*. Important locations like home and work are possibly score high in both dimensions. Another type of locations may have high occurrence but low duration, *e.g.,* coffee shops or schools that the user visits regularly but does not stay long. Possibly, a user may have even higher data network usage in these locations. In Section III we also evaluate all these metrics.

E. Predictability

We use a simple yet powerful metric, entropy, to quantify the degree of predictability for a time series. To compare with CDR based models, we follow the same entropy definitions used in [8], except for the last one. [8] contains more details of the definitions. The last entropy is proposed by this work, capturing the predictability of application usage in a given location.

- Location diversity entropy: $S_0 = log_2 N^u$, N^u is the number of unique locations visited daily by user u.

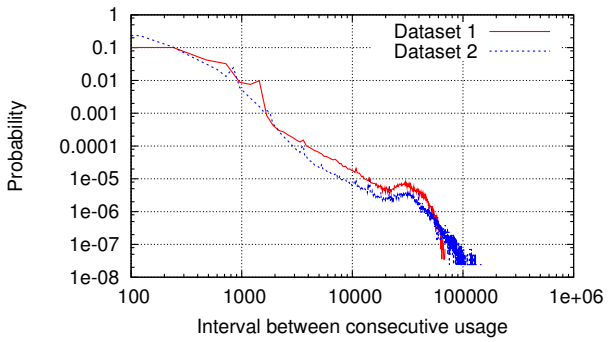

Fig. 1. User occurrence distribution.

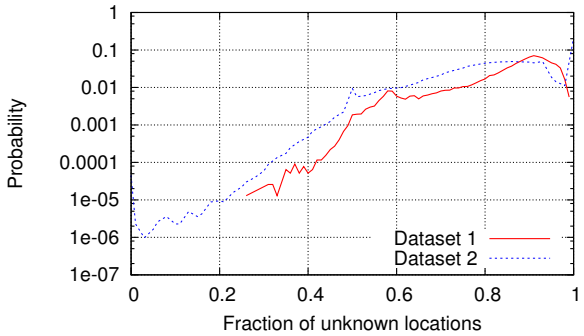

Fig. 2. Silence period distribution.

- Location recurrence entropy:
  $S_1 = -\sum_{j=1}^{N^u} p_j^u log_2 p_j^u$, p_j^u is the probability of user u visiting location j.
- Path entropy:
  $S_2 = -\sum_{r_{j(sub)}^u \subset r_j^u} p(r_{j(sub)}^u) log_2 p(r_{j(sub)}^u)$. Here r represents each path. For each path of user u, r_j^u, we take its subset $r_{(sub)j}^u$ and then compute the probability of this subsequence. Overall, S_2 means the likelihood of visiting a location from a particular route.
- $S_3 = -\sum_{apps} \sum_j p(app_k^u \mid j^u) log_2 p(app_k \mid j)$. Here it is a summation over all users and all applications in $apps$. $p(app_k^u \mid j^u)$ is the probability of user u using app_k at location j.

III. Mobility Analysis

In this section, we present our results from four aspects, the basic characterization of our data set, the mobility patterns, the temporal and spatial characteristics, and finally, the predictability. For each analysis, we compare the observations with the results from CDR-based approaches [8], [5], [9], and with models obtained from location-sharing services [10]. The CDR based approach extracts the cell tower information when a phone call is made or an SMS is sent. The location-sharing services (LSS), *e.g.,* Foursquare, Gowalla, and Facebook checkins, allow users to "check in" at venues that they want to share with friends on the social networks. The service

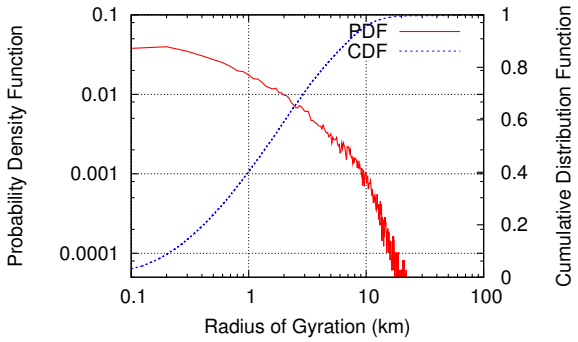

Fig. 3. Radius of gyration.

sharing service provider records such location data and uses them to perform mobility modeling. In this section, we aim at finding the unique properties provided by the cellular data network. Since we do not have access to CDR and LSS data, we perform comparison with existing published results. In the rest of this paper, we use *CDR* to stand for the CDR based approach, *LSS* for the location sharing services, and DNR for the data network records (our approach).

A. Basic characterization

Though users carry their phones with them most of the time, mobile phone can only provide location information to the network when the user uses voice or data services. We first quantify how often users use these network service. We group continuous records of one user together into one usage instance, and then compute the time interval between any two consecutive usages of the same user. Figure 1 shows the interval in seconds (x-axis) and its corresponding probability distribution (y-axis). The users tend to use their phones in short bursts, the dominant fractions are within 1000 seconds. It has a long-tail distribution, suggesting long periods with no activity. Both data sets have similar trends. Compared with the intervals between consecutive calls in the CDR based approach [8], the gap here is much smaller. For example, 0.1% of the intervals is larger than 2000 seconds in Figure 1, but corresponds to 10000 seconds in LSS [8]. There is no exact plots in [10], but the hourly check-in rate is 0.01, translating to over 3000-second intervals.

Although users use data services more often, during those long durations in the tail, we have no information about the user's location. We call it silence period. We directly measure the fraction of intervals when the user's location is unknown per hour and plot its distribution for all users in Figure 2. The distribution are more spread out compared to the unknown locations in CDR. Some users are very active, *i.e.,* more than 50% of their time have network activity. Most users are between 0.6 and 0.9, meaning 10% to 40% of their locations are known. These two figures both show that DNR provides more data points daily for analysis than CDR and LSS, as expected. The variance across users is large, suggesting one could draw more conclusion from heavy users.

Since we have similar observations from both data sets, for analysis below, we will combine both datasets for the ease of

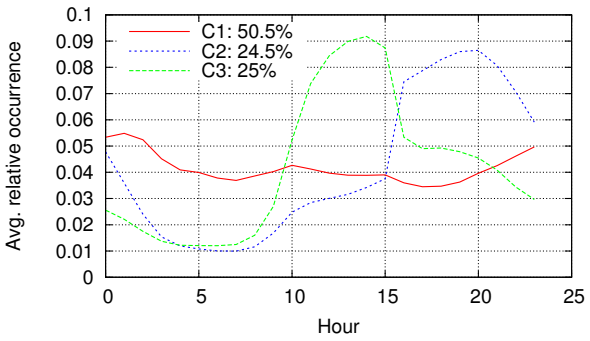

Fig. 4. Three clusters of users based on occurrence.

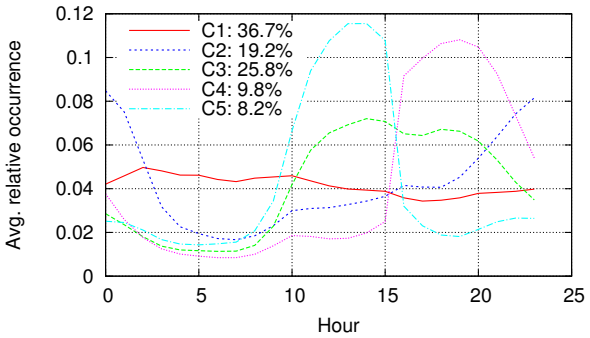

Fig. 5. Five clusters of users based on occurrence.

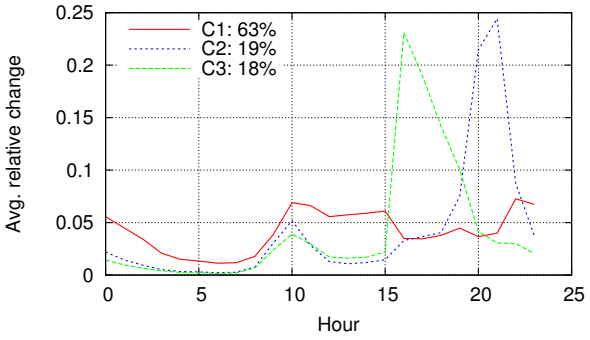

Fig. 6. Three clusters of users based on movements.

presentation. Since users with few occurrences do not provide much information for their mobility, for later mobility analysis, we remove inactive users who occurs less than 0.5 times per hour in average.

B. Movement patterns

One common metric to quantify how far a user moves over time is the radius of gyration [9]. It is the standard deviation of distances between the user's locations and the user's center of mass, which is a measure of how frequently and how far a user moves. Figure 3 follows a power-law distribution with a fat tail. Most user's daily movements are confined to an area of 10 kms, while a few users travel more than 20 kms. Comparing with CDR and LSS, it follows the same distribution model, but with even smaller radius. It could be a consequence of more frequent usage on the apps on smartphones today. Comparing with LSS, the radius we observe is much smaller. [10] found that 34% of all users display a radius of gyration of less than 10 miles, while only 14% have a radius of gyration larger than 500 miles. The information entered into these social networks are often of significant meaning to users, especially for interesting locations that the user does not often visit. For example, a user would like to inform his friends that he visits another city, rather than goes to a grocery store closely. However, it is a small radius that reflects people's routine life. In summary, these comparison results show that although different data sources discover similar models (power-law distribution with truncated tail), the value of the parameters is different. It depends on the temporal and spatial granularity that the data source can capture regarding the human movement.

Next, we examine the temporal property of user occurrence and movements. We examine how many times each user uses the network and how many times he moves per hour, and normalize them by the total appearance/movements. Similar to [5], we perform clustering on all the users using X-means [18]. It outputs 3 clusters as the best clustering choice and 5 clusters as the second best choice.

Figure 4 shows the distribution of average relative occurrence for three clusters. The population of users in each cluster is shown in the caption of the figure. Interestingly, the largest cluster C_1 (with 50.5% of users) has relatively flat distribution across time. C_2 and C_3, each containing a quarter of the population, have peaks at different times. C_2 users are most active around noon while C_3 users more likely go online after work.

We also present the 5-cluster result (the second best choice) in Figure 5: C_1, C_4 and C_5 are similar to the 3 clusters in Figure 4, C_3 is the group of users who are active in general when they are awake (from 10am to 11pm). Surprisingly, C_2, containing a nontrivial fraction of users, have an unexpected distribution: being more active at night, with the peak during mid-night. We suspect that the activities for these users during mid-night are application automatic updates or backups. It could also be mapped to people who work the night-shifts. Another unexpected result is C_1, where the users constantly use the data and move around. It could be a result of some applications with periodical heart-beat communications.

Similar to the occurrence, we also compute the vectors of relative movements per hour for each user, and then perform clustering. X-means determines 3 clusters as the best answer, shown in Figure 6. C_1 users' mobility is wide spread across time, with two peaks from 10am to 3pm and after 9pm. The two peaks match with the people with day-shifts and people work the night-shifts. C_2 and C_3 both have significant spikes at 4PM and 21PM respectively, which are possibly mapping to the commuting back home and the entertainment activity at home at night. Here we can see that the flat usage of C_1 in Figure 4 and the mid-night back-up group C_2 in Figure 5 go away. One reason could be that C_2 in Figure 5 are likely to be back-up activities that do not involve movements. In our

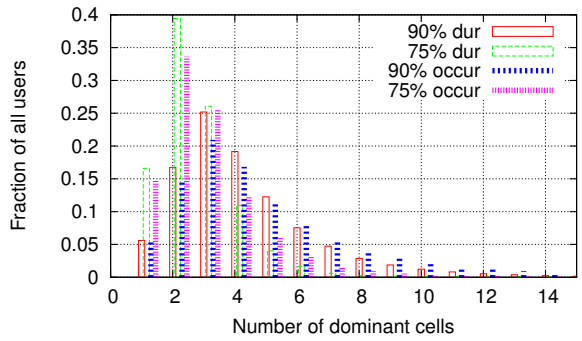

Fig. 7. Dominant cells per user

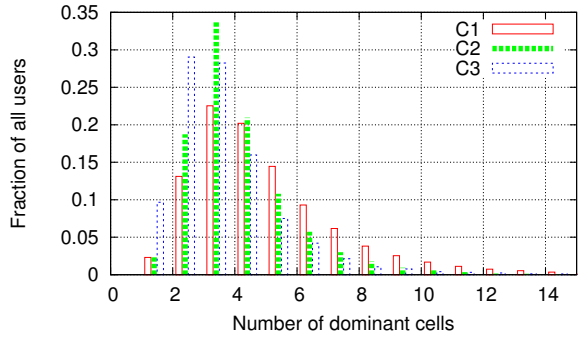

Fig. 8. Dominant cells for three clusters.

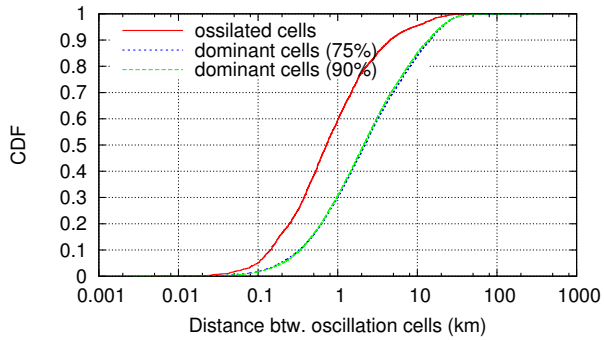

Fig. 9. Distance distribution between oscillating cells and dominant cells.

later analysis, we use this clustering results to perform further breakdowns.

Overall, these findings of temporal distinctions between users are very different from CDR and LSS. In CDR (Figure 4 of [5]), two clusters are detected: one group has more afternoon calls and the other makes more calls in the evening. This is a key finding of the DNR approach, as besides the expected time-of-day dynamics, DNR contains more regular access from the users, thus it allows us to capture more movements than CDR. LSS only identifies one group of users, who are more active in off-work hours than working hours.

C. Important locations

Previous work has shown that most users' activity are around a few important locations [5]. We define important locations as the cells where a user spends x% of his time on it (the duration curves in Figure 7), or the cells involved in top x% of his movements (the occurrence curves in Figure 7). We show x=90 and x=75 in Figure 7 to check different dominant levels. For 40% of the users, the top two cell towers account for 75% of the their time with cellular services. It is similar when measuring on movements, about 34% of the users. Although 90% of the time and movements are covered by the top 5 cell towers for most users, comparing to the CDR-based models, we can see that the number of important locations is much larger and more spread out across users. While CDR found that most mobility are accounted for with just the top two cell towers with which a user is associated [9], here most users are around 2-4 important locations, and a small number of users spend their time more evenly across up to 15 locations.

In Figure 6, we identify three clusters of users with distinct temporal movements. Now we study if these three groups of users also show different spatial property. The number of dominant cells are shown in Figure 8 for the number of cells that cover 90% of the duration. Interestingly, we found that C_1 have more dominant locations, probably because C_1 subscriber uses data services pervasively, thus, he is served by a number of different cell towers. On the other hand, C_3 are users who mostly use data services at night, therefore, he more likely just wanders around his home location. In the figure we see that C_3 has the smallest number of dominant cells.

We not only examine the number of dominant locations, but also look into what they are, how far they are apart. For each user, we obtain a list of his important locations, and then compute the distance between any pair of them, shown in the green and blue curves in Figure 9. Most of them are within 10km, probably the distance between home and work.

Using the method to detect oscillations in the trace, we smooth out the noise introduced by the "ping-ponging" effect. However, we do not have ground truth to validate our approach. We examine the distance between the oscillating pair of cells we detected and compare it with the normal distance between dominant cells. Though we use the distance threshold 75km as a part of the oscillation detection, from Figure 9, we can see that most of the oscillation detected are much closer, i.e. 60% are within 1 km and 90% of them are within 5 kms. Intuitively, these are the distance where oscillation most likely will happen. Comparing to the distance between dominant cells, we can see that the distance for oscillation is much smaller. It also shows that our oscillation detection is not sensitive to the choice of distance threshold.

D. Predictability

Finally, we examine whether user movements follow simple reproducible patterns, or whether they are predictable. To this end, we apply the metrics proposed in [8] so that we can directly compare the results with CDR. Figure 10 shows the entropy distribution for S_0, S_1, and S_2, as defined in Section II. For example, for S_0, the peak in 2.5 means that 2.5 bits are needed to represent the randomness of the location distribution for a given user. Similarly, 1.9 bits are needed to encode

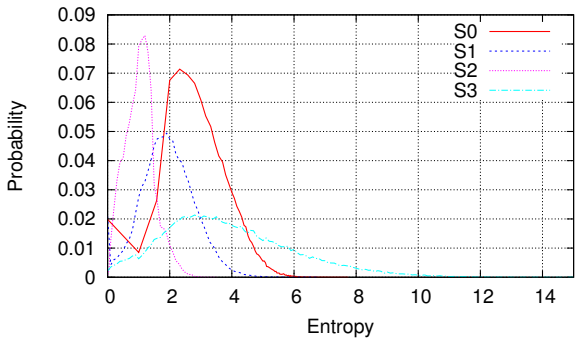

Fig. 10. Entropy distribution

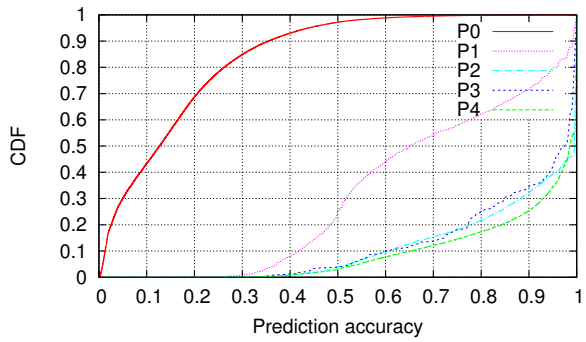

Fig. 11. Prediction accuracy

the randomness of temporal distribution among locations. The smaller the entropy is, the more predictable it is. Comparing with [8], the entropy range is much smaller. For example, the peak of S_0 in [8] is 6 and S_1's is 3. It means that the locations and movements from the data network's view are much more predictable. The S_2 distribution is even smaller. This is also consistent with CDR. It means that predicting the next location given a sequence of known locations is promising, which is consistent with [17] of using Markov chain for prediction. Finally, S_4 captures the predictability of using a particular application at a given location. Here we can see that the entropy is quite distributed. It suggests that the predictability depends on application type and the location. [13] found that the usage of application is more predictable in hot-spots. We plan to further investigate the prediction in the future work.

To this end, we summarize the key findings from the massive amount of analysis.

- The locations and movements observed from the data network have larger variance across users and the usage is flatter than the observations from CDR.
- Surprisingly, the user mobility radius is smaller from DNR, which is probably because that DNR captures more fine-grained trajectory of the user movement. For instance, if a user uses Google Map application while driving, it records every base stations during his route. The duration of using apps is likely to be longer than making a phone call.
- The number of important locations identified in DNR is generally higher than CDR, suggesting that the previous home/work based mobility model may need to be revisited.
- Clusters of users have been identified when measuring the key metrics of mobility behavior, which could be caused by different job types and habits of using smartphones. This suggests that new mobility models should take the job type into considerations.

IV. Prediction of Application Usage

The data network records contain rich information of the correlations between content and the locations, *i.e.,* the affiliation between mobility and one's cyber interests. While

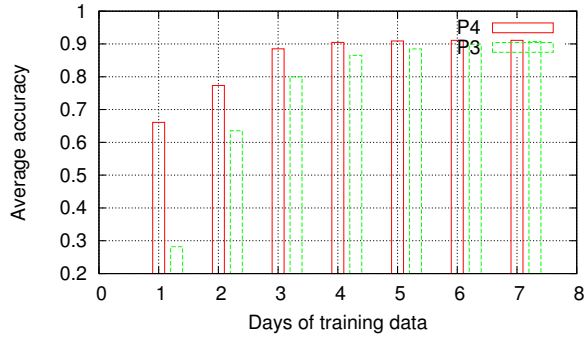

Fig. 12. Impact of training data

many existing work focus on prediction models for trajectory, instead, we focus on building a prediction model for the usage of application given its mobility information.

A. Predicting the application usage at a given location

The question we aim at answering in this section is, can we predict the usage of an application given one's location? We aim at building a prediction model for application usage given one's mobility state. Such a model is useful for many location-based applications. For instance, knowing that a user will use a specific application in a particular location, we can prefetch the content to a predicted future cell tower, such that the user can retrieve the content upon entering the range of the cell tower.

We describe our prediction method in the following. We argue that the correlation between cyber and physical space differs across users, depending on their diverse habits of using the mobile Internet. Thus, we start with a model for each individual user. Given the current location c_i^u for user u, we compute the probability of user u using application a as his next application, A_{i+1}^u, at his next location c_{i+1}^u as the following:

$$Pr(A_{i+1}^u = a \mid c_i^u) = \sum_{c_{i+1}^u} Pr(c_{i+1}^u = c \mid c_i^u) Pr(a \mid c) \quad (4)$$

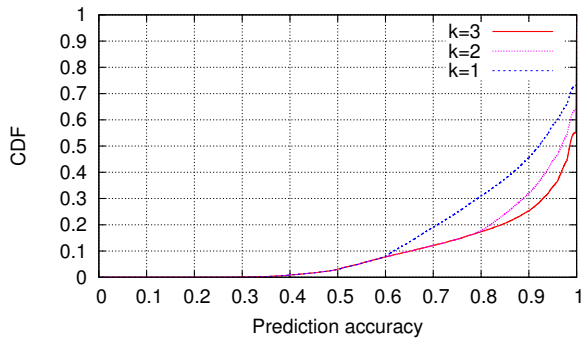

Fig. 13. Impact of parameter k

The first part of the equation, ($Pr(c^u_{i+1} = c \mid c^u_i)$), is well-studied in the prediction of trajectory. Like [3], we use a second-order Markov chain model. The second part can be computed as $Pr(a \mid c) = \frac{count(a,c)}{count(c)}$.

This basic model is intuitive and easy to implement. However, from our experiments, we found that it can be improved from two aspects. On the one hand, we found that the knowledge of application used in the current state can improve the prediction accuracy. Intuitively, one may have the routine of reading news followed by checking emails on his way to work, or one may be watching YouTube on his way home. Thus, we extend the basic model in Eq 4 with the current application usage A^u_i, as below.

$$Pr(A^u_{i+1} = a \mid c^u_i, A^u_i) = \sum_{c^u_{i+1}} Pr(c^u_{i+1} = c \mid c^u_i) Pr(A^u_{i+1} = a \mid c, A^u_i) \quad (5)$$

It is also a summation over all the possible next locations, same as Eq 4. The second half is computed as $Pr(A^u_{i+1} = a \mid c^u_{i+1}, A^u_i) = \frac{count(a,c,A^u_i)}{count(c,A^u_i)}$.

After improving the accuracy of the basic model, we realize that both Eq 4 and Eq 5 require obtaining sufficient data points for each individual users, which limits its capability to predict for a wider range of users. To this end, we leverage the similarity between users to relax this restriction. We first perform clustering to assign users to groups: users in the same group share similar trajectory and cyber interests. More specifically, for each user, we compute the frequency of using each app as well as the frequency of appearance at each location. Thus, each user is described as a vector of these frequencies in both physical and cyber domains. Then we use the x-means clustering method to generate $s_0, s_1 \ldots s_k$ for all the users. The cluster for user u is denoted as $S(u)$, then we can use a larger set of observations to compute:

$$Pr(A^u_{i+1} = a \mid c^u_i, A^u_i) = \sum_{c^{S(u)}_{i+1}} Pr(c^{S(u)}_{i+1} = c \mid c^u_i) Pr(A^{S(u)}_{i+1} = a \mid c, A^u_i) \quad (6)$$

B. Evaluation

Below we evaluate the accuracy of our prediction model and show the benefit of our extensions on the basic model. We separate the 8-day data to be 7 days for training the model and 1 day for testing. For each user, at any moment, we compute the probability of each application that he will use at his next location. Then we select the top k applications as the predicted set, which is configurable and is chosen to be 3 in the following experiment. If the observed next application is within this predicted set, then it is a hit. We compute a ratio of successful prediction for each user. Figure 11 shows the cumulative distribution of the success rate under different models. P_0 is the naive model without differentiating users: $P_0 = Pr(A_{i+1} = a \mid c_i)$, which is computed by combining all the data points for all users. P_1 is Eq 4, which considers the state for each user separately. P_2 and P_3 represent results computed using Eq 5, while P_2 uses a second-order Markov chain to compute $Pr(c^u_{i+1} = c \mid c^u_i)$ and P_3 uses a third-order Markov model. Finally P_4 is the results with Eq 6. We can clearly see the accuracy improvements in the enhanced models. The second-order and third-order Markov chain do not make much difference. Overall, our final model, Eq 6, achieves over 90% accuracy for 80% of the cases.

Two parameters in our model has impact on the accuracy. The first is the length of the training data. Thus, we vary the length of the training data from 1 day to 7 days and always use the 8^{th} day for testing. Figure 12 shows the average hit rate across all users. Initially the rate increases with more training data until the 4^{th} day, after which the improvement is not significant. This result demonstrates that our model can reach a stable state with a certain amount of the training.

The other parameter is the size of the predicted set k. The larger k is, the more likely we predict successfully. But a too large k will make the prediction less valuable. We evaluate different choices of k in Figure 13. We can see that even when $k = 1$, 70% of the cases can still achieve above 80% accuracy.

V. Related Work

Recently there is an increasing attempt to use cellular network data in various domains, such as urban planning, mobility modeling, and healthcare. Below, we discuss the subset of that work related to mobility modeling.

The predictability of human mobility has been shown to be high [7], [8] using CDRs from the cellular network. It is shown that the entropy of locations does not increase much beyond certain duration, and the accuracy of predicting an individual's location can be up to 93%, regardless how far he travels [7]. [9] further pointed out the regularity of human mobility in both temporal and spatial domains. Moreover, many work consistently found that users spend a significant fraction of time in their top locations only, especially home and work [13], [5]. Our work differs from the above in that we aim at uncovering the view of human mobility from a different data source and thus can draw different conclusions.

There are other ways to study the human mobility besides CDR, such as GPS, WiFi traces. Rhee et al. [19] studied movement patterns of 44 participants with GPS devices and built a model of Levy flights, meaning that the distribution is random walk with a heavy-tailed distribution of step lengths.

Kim et al. [2] studies human movements from data in WiFi access points in a university campus. Another analysis focusing on coarse-grained traces have also been done in a similar setting [3]. Again, we use a different types of data, which is more common, easier to gather, and can scale to a large population. [20] uses the cellular data network for mobility analysis but it is done half a decade ago. The mobile Internet usage has changed significantly in the recent years. Thus, there is a need for revisiting the problem.

Recently, a number of studies started to investigate the limitations of using CDR based approach for mobility inference. For instance, [21] pointed out the bias towards most significant locations in the CDR based approach. Schulz*et al.* [14] discovered that GSM data tends to underestimate the user radius of gyration compared with the GPS data. Our study follows the same spirit but focuses on the comparison with data network records. 3G signaling and hand-off data are also used for mobility studies. Among them, [22] characterizes the vehicular mobility from the 3G signaling data. The mobility pattern has been used for various applications such as route classification and prediction [23]. These are different aspects of mobility patterns, which is orthogonal to our work and we plan to investigate similar aspects using our dataset in the future.

The work most related to ours is the serendipity measurement, which is also using cellular data traffic [13]. They focus on analyzing the relationship between mobility patterns and usage of certain applications in cyber domain. They found a strong correlation between application affinity and locations. Though this work also presents basic characteristics of mobility observations, their focus is not on the mobility model itself. Dong *et al.* [17] very recently proposes a leap graph based model to accurately detect true movements, and to assist prediction. Our work differs in that we focus on analyzing the mobility properties from data network. But we could potentially be apply their method to improve accuracies of oscillation detection.

VI. Conclusion

In this paper, we focus on the mobility characteristics from the cellular data network's perspective, and then compare the findings with the CDR based approach and the finds from a location sharing service. We identify significant difference across data sources. In general, data network records are more fine-grained both temporally and spatially. We observe three classes of users, having distinct usage and movement patterns across time in a day. It could be a result of both different daily schedule and different availability of WiFi connectivities. It suggests that future modeling of the mobility from data network should take this factor into consideration. One important usage of the mobility model is content prefetching in mobile data networks. Thus, we propose a prediction method to forecast the future application usage and show promising results.

References

[1] Ericsson, "Mobile Milestone: Data Surpasses Voice Traffic." http://gigaom.com/2010/03/24/mobile-milestone-data-surpasses-voice-traffic/, 2010.

[2] M. Kim, D. Kotz, and S. Kim, "Extracting a mobility model from real user traces," in *Proc. IEEE INFOCOM*, (Barcelona, Spain), IEEE Computer Society Press, April 2006.

[3] J. Yoon, B. D. Noble, M. Liu, and M. Kim, "Building realistic mobility models from coarse-grained traces," in *MobiSys*, pp. 177–190, 2006.

[4] M. A. Bayir, M. Demirbas, and N. Eagle, "Discovering spatiotemporal mobility profiles of cellphone users," in *10th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, WOWMOM 2009, Kos Island, Greece, 15-19 June, 2009*, pp. 1–9, 2009.

[5] S. Isaacman, R. A. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger, "Human mobility modeling at metropolitan scales," in *MobiSys*, pp. 239–252, 2012.

[6] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *Proceedings of the 9th international conference on Pervasive computing*, pp. 133–151, 2011.

[7] H. Zang and J. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks.," in *MOBICOM*, pp. 123–134, 2007.

[8] C. Song, Z. Qu, N. Blumm, and A.-L. Barabsi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[9] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, June 2008.

[10] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services," in *ICWSM*, 2011.

[11] Ericsson, "Ericsson's 2012 Mobility Report." www.ericsson.com/ericsson-mobility-report?, 2012.

[12] F. Qian, Z. Wang, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck, "Profiling resource usage for mobile applications: a cross-layer approach," in *Proceedings of the 9th international conference on Mobile systems, applications, and services*, 2011.

[13] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Measuring serendipity: connecting people, locations and interests in a mobile 3g network," in *Proc. ACM SIGCOMM IMC*, pp. 267–279, 2009.

[14] D. Schulz, S. Bothe, and C. Körner, "Human mobility from gsm data – a valid alternative to gps?," in *Proc. of the Mobile Data Challenge Workshop*, 2012.

[15] Ericsson, "Real-time performance monitoring and optimization of cellular systems." http://www1.ericsson.com, 2002.

[16] A. Popescu, "Geolocation API Specification." http://dev.w3.org/geo/api/spec-source.html, 2012.

[17] W. Dong, N. G. Duffield, Z. Ge, S. Lee, and J. Pang, "Modeling cellular user mobility using a leap graph," in *Proc. International Conference of Passive and Active Measurement*, 2013.

[18] D. Pelleg and A. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in *Prof. International Conf. on Machine Learning*, 2000.

[19] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM Trans. Netw.*, vol. 19, pp. 630–643, June 2011.

[20] E. Halepovic and C. Williamson, "Characterizing and modeling user mobility in a cellular data network," in *Proceedings of the 2nd ACM international workshop on Performance evaluation of wireless ad hoc, sensor, and ubiquitous networks*, PE-WASUN '05, 2005.

[21] G. Ranjan, H. Zang, Z.-L. Zhang, and J. Bolot, "Are call detail records biased for sampling human mobility?," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 16, December 2012.

[22] P. Fiadino, D. Valerio, F. Ricciato, , and K. A. Hummel, "Steps towards the extraction of vehicular mobility patterns from 3g signaling data," in *4th International Conference on Traffic Monitoring and Analysis*, pp. 66–80, Springer Verlag, Mar 2012.

[23] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "Route classification using cellular handoff patterns," in *Proceedings of the 13th international conference on Ubiquitous computing*, UbiComp '11, 2011.