Inferring Correlation between User Mobility and App Usage in Massive Coarse-grained Data Traces
List of changes are appended at the end of the paper (page 22)

ZHENG LU, University of Tennessee
YUNHE FENG, University of Tennessee
WENJUN ZHOU, University of Tennessee
XIAOLIN LI, Nanjing University
QING CAO, University of Tennessee

With the rapid growth in smartphone usage, it has been more and more important to understand the patterns of mobile data consumption by users. In this paper, we present an empirical study of the correlation between user mobility and app usage patterns. In particular, we focus on users' moving speed as the key mobility metric, and try to answer the following question: are there any notable relations between moving speed and the app usage patterns? Our study is based on a real-world, large-scale dataset of 2G phone network data request records. A critical challenge was that the raw data records are rather coarse-grained. More specifically, unlike GPS traces, the exact locations of users were not readily available. We inferred users' approximate locations according to their interactions with nearby cell towers, whose locations were known. We proposed a novel method to filter out noises and perform reliable speed estimation. **We verify our methodology with out of sample data and show its improvement in speed estimation accuracy. We then examined several aspects of mobile data usage patterns, including the data volume, the access frequency, and the app categories, to reveal the correlation between these patterns and users' moving speed. Experimental results based on our large-scale real-world datasets revealed that users under different mobility category not only have different smartphone usage motivations but also have different ways of using their smartphones.**

ACM Reference format:
Zheng Lu, Yunhe Feng, Wenjun Zhou, Xiaolin Li, and Qing Cao. 2010. Inferring Correlation between User Mobility and App Usage in Massive Coarse-grained Data Traces
List of changes are appended at the end of the paper (page 22). *ACM J. Comput. Cult. Herit.* 9, 4, Article 39 (March 2010), 21 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In the past decade, the use of smartphones has grown tremendously among consumers. According to a recent report [5], there were 3.4 billion smartphone users worldwide, and the accumulated mobile data traffic reached 120 exabytes in 2015. Alongside this explosive growth in mobile data traffic is the popularity of smartphone apps, such as those served by Google Play and Apple Store, whose number has exceeded 1.5 million by July 2015 [28]. It is estimated that people spend as much as 30 hours monthly on average on these apps, a growth of over 65 percent compared to 2013 [18].

DOI: 10.1145/nnnnnnn.nnnnnnn

Recent research has invested considerable efforts to understand smartphone app usage behavior, which can inform app developers, mobile advertisers, and network service providers [38, 41]. For example, both temporal patterns (e.g., individual app usage histories) and spatial patterns (e.g., location contexts) have been extensively studied [16]. Their results have enabled novel applications, such as smartphone app launching prediction services [40] and location-aware event recommendations.

These studies, in spite of their usefulness, usually require fine-grained data. When only coarse-grained datasets are available, smart algorithms are needed to extract useful information. In particular, as the cellular data are commonly collected, it would be useful to design processing algorithms to extract and exploit useful mobility features from these datasets. At a large-scale (i.e., city-wide), we would like to understand user mobility, and investigate how this feature correlates with the usage patterns of smartphone apps. **(R2-1, R4-1) Understanding such correlations, if any, can provide useful insights on how users are using their smartphones, such as the unique characteristics of smartphone usage for users on different mobility modes, including stationary, walking, driving, or taking buses. Knowing such correlations can not only help app designers analyze and improve their apps in a more guided manner but also provide much-needed assistance to network operators to optimize in-the-field operations. Finally, our study can also provide valuable results for commercial use. For example, by knowing the users' app preference on different mobility modes, advertisers can deliver more relevant ads based on the current mobility mode of a user.**

Unfortunately, previous work on this topic has only investigated this problem in highly limited and controlled contexts, taking into account the usage history of a small set of users. For example, a few works have addressed the problem of transportation mode inference, with location information collected using more accurate hardware (e.g., GPS, sensors) from a small group of users in controlled experiments [3, 20, 21, 24, 29, 31, 36, 43]. Later work suggested that it may be possible to use cell tower communications to monitor users' mobility indirectly [23], where efforts have been focused on inferring users' trajectories [10, 17] or transportation mode [2, 33] using cell-phone traces (e.g., Call Detail Records, handover data) that do not directly contain location information. A limitation of these approaches, however, is that they are usually small-scale by nature, and the data collected were much cleaner given the controlled environment, making the proposed techniques impractical for real-world large-scale data. Our work follows the latter line of research, using large-scale cell-phone tower traces, with significant difference. First, our dataset consists of a truly large population, where we have access to mobile data access histories of millions of users in three cities that cover thousands of square miles. Second, our research goal is to reveal large-scale, population-level correlations, if any, between user mobility and app usage patterns, a goal that has not been addressed in any of previous research work.

We addressed the following two challenges in our work. First, to infer user mobility with cell-phone traces, we needed to filter the location history to obtain accurate estimates. In our dataset, the only location information available was the location of each cell tower. By communication principles, we know that a user's phone typically contacts towers with the best signal reception, which usually are the nearest ones. After surveying previous work on estimating trajectories based on similar datasets [9, 10, 17, 27, 37] or finding mobility motifs [6, 35], we could not find a technique that suited our needs as their results were clearly still too coarse-grained for our data. For example, one study was based on users who performed daily commute or took city-to-city long-distance trips. In contrast, our data were in dense urban areas where users employed a mixture of transportation modes ranging from walking, bicycles, to buses and cars (railway transportation was not present in our dataset). Therefore, we needed to develop a novel method to estimate more complicated user mobility behaviors for our dataset. Second, to effectively correlate app usage history with mobility patterns, we had to strike a tradeoff between the most popular apps and the sparingly used ones. More specifically, we found that a majority of users used "heavy-hitter" apps irrespective of their mobility modes. Inferring such correlations were less meaningful. Therefore, we focused on app groups where data exhibit differentiated popularity for users with different moving speeds, a task that was considerably more challenging than simply performing correlation analysis between all apps and all users without differentiation. We emphasize, however, that due to the data limitations (i.e., being extremely coarse-grained and heterogeneous and the absence of ground truth), all our conclusions are, at best, educated guesses that are based on real-world data. We believe such results are meaningful and insightful for a wide range of people: app developers, advertisers, network operators, and smartphone users.

The main contributions of this paper can be summarized as follows. We designed and evaluated a novel method to infer user speeds with cell-phone traces with low location accuracies. Compared to existing approaches, this method achieved far better and fine-grained estimation with adjustable confidence levels. Specifically, to overcome the problem of location accuracy, our method involved steps to segment traces by pass-boundary events. When a user establishes a new connection with a different tower, and performs intra-cell level zooming and analysis to calculate distance estimates. This method was also robust against issues caused by the uncertain nature of wireless communications, e.g., a user located in the overlapped communication coverage area of multiple towers may randomly communicate with each tower, causing cell oscillations that other simple methods cannot easily address. With the more accurate speed estimates, we were able to study the correlation of user mobility with app usage patterns in the real-world environment. Our results revealed correlations of user speed and mobile data access patterns, including data volumes, access frequency, and app choice. The results are novel in that no previous work, to the best of our knowledge, has gained similar insights or reported findings in this aspect.

The rest of this paper is organized as follows. In Section 2, we describe previous works on user mobility inference and geospatial app usage patterns. Section 3 defines our problem and provides details on the mobile data access trace we use in this paper. We describe our speed estimation methodology and design in Section 4. **An objective evaluation of our speed estimation algorithm is presented in Section 5.** Section 6 explains our findings on the correlation of user speeds and mobile data access patterns. **We discuss the limitations of our work in Section 7.** Finally, we conclude our work in Section 8.

2 RELATED WORK

In this section, we summarize recent literature on smartphone apps, user mobility, and geospatial analysis of mobile phone apps data.

To study the smartphone app usage behavior of a large group of users, previous work has analyzed mobile data traces generated by smartphone apps in studies of various scales. [41] studied the mobile user behavior by focusing on data usage, mobility pattern and application usage. In [38], the aggregated spatial and temporal prevalence, locality and correlation of smartphone apps at a national scale is investigated, by analyzing the mobile data generated by smartphone apps. Unlike our work, these previous work have not studied the relation of mobile user behavior with more complex user mobility, i.e., user speed.

Using GPS [3, 20, 21, 24, 29, 31, 36, 43] and embedded sensors [8, 15, 20, 21, 26, 30, 34], a separate body of research is able to use smartphones to infer user mobility patterns accurately in small-scale, controlled experiments, such as inferring transportation modes. Most of these works formulate the problem as a classification problem, where common challenges involve data segmentation [3, 20, 31, 43], feature selection [3, 29, 34, 43]. Multiple methods, such as SVM or linear regressions, are developed to achieve the best accuracy.

Although GPS and sensors are well suited for small-scale experiments, they are not scalable as users typically do not want their GPS traces to be shared with others. In recent work [23], it is revealed that there is a great potential for using cell-phone data traces such as Call Detail Records (CDRs) for user mobility inference. A large body of research literature exists applying this method for inferring user's trajectories [2, 9, 10, 14, 17, 27, 37] or mobility motifs [6, 35]. For example, [10, 17] inferred user trajectories from cell-phone traces based on how likely a specific route can lead to similar tower access sequences stored in the data traces. In another work [33], it aims to classify a user's transportation mode by clustering travel time distribution. Finally, researchers [2] also proposed approaches that can deal with common zig-zag problems in inferring user mobility from smartphone traces.

Different from these existing methods, however, our approach take advantage of the calability of cell-phone data traces and achieves fine-grained user mobility inference on top of it.

Studying correlations between app usage and features extracted from phone traces is not new in the literature. Previous work has studied relations of human mobility and social networks using geo-spatial features. For example, a work [4] found that the short-ranged travels are periodic and not likely to be related to the social network structures, while long-distance travels are heavily related to the social network status of a user. Based on these findings, a model was proposed to predict dynamics of future human movement with a high accuracy. Follow-up works such as [19] studied a similar problem with a different dataset. [25, 41] studied the geospatial relation of the app usage volume. Their works mostly studied the spatial correlation of the smartphone usage, while the user mobility's impact on app usage is still a missing piece of these works. [16] studied how the proximity, the location and individual differences (e.g., personality) can effect the user's mobile data usage. Finally, [42] showed the apps access pattern under various user mobility properties such as number of visited cell phone towers and radius of gyration. However, analysis of much complex user mobility such as user speed is still a missing piece in these works.

(M5) Another aspect addressed in the literature is related to the tradeoffs adopting either cellular network data or WiFi data. Although recent studies [1, 13] have shown that WiFi handles half of the mobile data traffic, it has been observed that cell phone networks typically have much better coverage and user mobility diversity [32, 39]. Furthermore, it is practically impossible to collect city-level WiFi data due to the heterogeneous nature of access points and privacy concerns. Therefore, it is more meaningful to study large-scale user mobility under cellular traffic, as is the methodology followed by this paper.

3 DATA DESCRIPTION

In this section, we provide a description of the dataset, followed by an example of a user's data.

3.1 Dataset

(M1) Our dataset contains mobile data access history of all active users (during a Sunday evening from 6 pm to 9 pm) of a major mobile carrier in three cities of China in September 2014. The data was collected by the carrier at the cell tower side. For each user, all data request records during the study period are available, where each record consists of the user ID (a hashed value for anonymity), the tower ID (from which we were able to look up its geo-coordinates), the timestamp, the app identifier, and other data access features such as data volumes.

The dataset includes more than 58 million mobile data access records with a total volume of more than 720 gigabytes, which covers all cell phones that were actively exchanging data with a total of 5199 cell towers in the area during the observation period. The number of unique users included in this dataset is identified as around 900 thousand. The total active time of all users accumulates to more than 1 million hours. Fig. 1 shows a heatmap of the mobile data access in a city area of our dataset. The dataset contains both user initiated network access and background network access.

3.2 Data Preprocessing Findings

We first preprocess the data and analyze the characteristics of mobile data access patterns. Distributional characteristics are visualized in Fig. 2. In particular, the number of records per user, the average time intervals between consecutive records, and the number of towers visited. We found that our dataset has a highly skewed distribution of the number of records per user, as shown in Fig. 2(a), and the time intervals between consecutive records, as shown in Fig. 2(b). Here, a higher record density, i.e., more records for a user in a time unit, indicates a better performance to infer user mobility even when the trip length is very short, as we can obtain a better

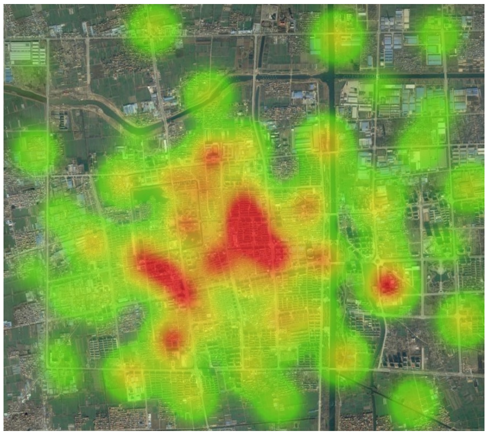
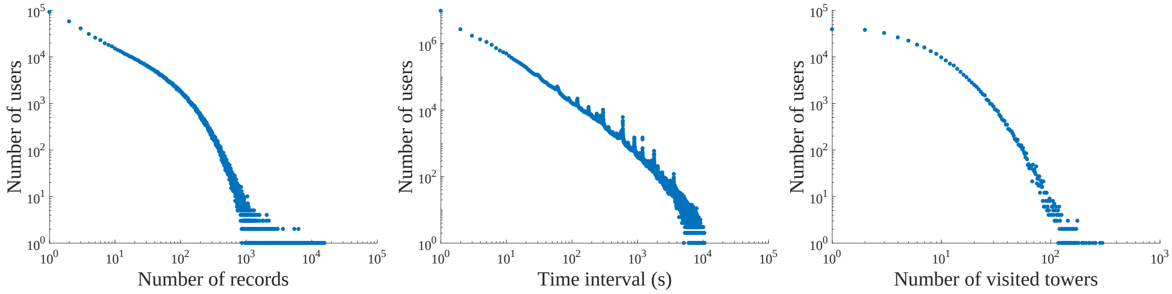


Fig. 1. Communication density in a city area.

granularity by analyzing these records. Actually, it is the case that most user only traveled a very short trip in terms of the number of visited towers according to Fig. 2(c).



(a) Number of Records (b) Avg. Time Interval (c) Number of Towers Visited

Fig. 2. Dataset characteristics.

Note that our dataset differs from commonly used mobility datasets used in existing work. Compared to moving trajectories like those captured by GPS, we do not know the exact locations of the users, and we only know a user is located nearby a tower to communicate with it. Furthermore, compared to other datasets with call detail records (CDR), our dataset is drawn from a region with more densely populated customers, where each may adopt different mobility methods such as walking, driving, or taking buses. Such differences make it harder to accurately estimate user speed based on existing methods.

3.3 An Example User's Traces

To provide a clear view of our data, we visualize a user's records from our dataset in Fig. 3 as a running example. Suppose that the user was taking the path (while using the cell phone) shown with the dashed line. In particular, she started by walking from location 1, to location 2 where she waited for the bus. After a few minutes, she got onto the bus, which took the path towards location 3. Even though we did not know the actual path of the user, her locations could be inferred by the nearby towers to which her communication data were sent to.

Since the locations of cell towers are known, we illustrate the tower locations on a map. We use markers to show tower locations and arrowed lines to show the sequence of visiting. The bottom part

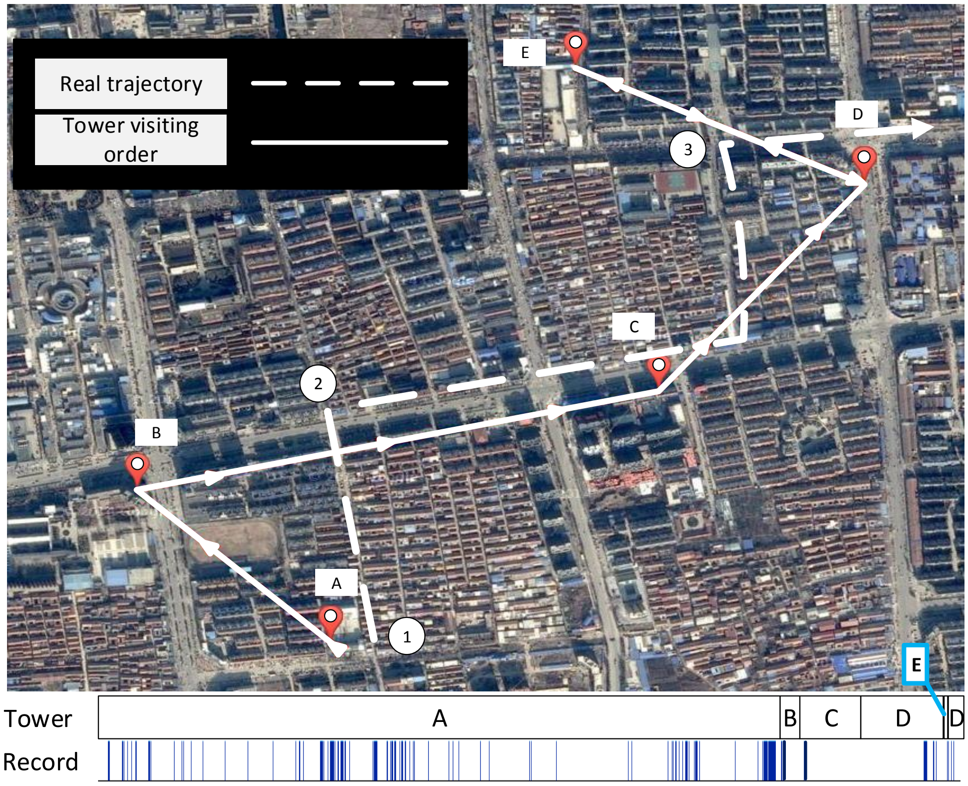


Fig. 3. Example data access activities of a user.

of Fig. 3 shows the timeline of the user's data access records. We also show the tower with which the user has communicated for each mobile data access record, by providing tower labels above. For this particular user, she communicated with tower A for a while, and shortly connected to tower B before switching to tower C. After that, the user was found again in tower D's coverage area. Then she was connected to tower E for a very short time, indicating a possible cell oscillation, and finally, she switched back to tower D.

4 SPEED ESTIMATION

In this section, we systematically describe our methodology for estimating user mobility speed using coarse-grained tower communication records and timestamps.

4.1 Methodology Overview

Our methods consist of multiple steps, where we first decompose traces of each user into segments to zoom into intra-cell speed estimation. Next, we estimate the distance and the travel time for each segment, where we employ a distance lower bound to filter out low-confidence estimates. In practice, such estimates are usually too noisy to be meaningful or reliable. Finally, we demonstrate how to compensate for speed estimation errors. Fig. 4 shows the structural overview of this methodology. The raw data parser on one end of this figure gathers data access records by users and sorts records of each user by time. On the other end, a list of tower locations from the mobile data access traces is extracted. Note that the system assigns a list for each city during processing steps.

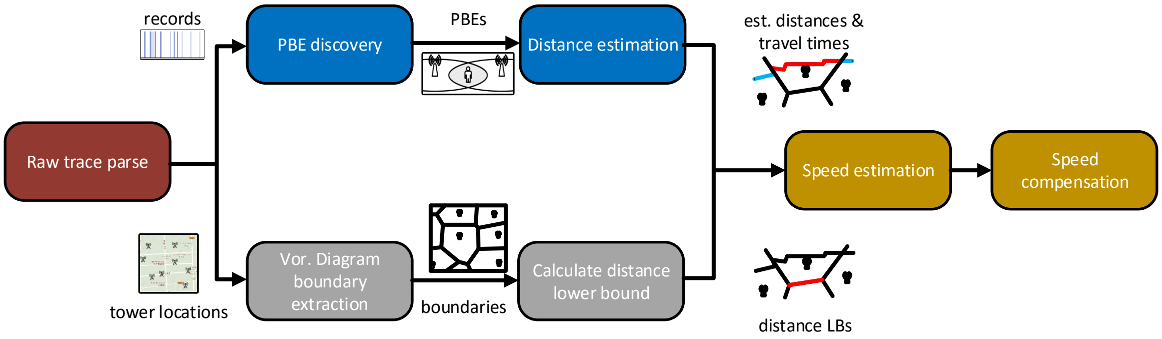


Fig. 4. Speed estimation system overview.

After we have parsed raw traces, we next process them in different steps in parallel. In one of the next steps, we analyze traces of each user and generate pass-boundary events (PBE) with the timestamp and location estimates of each record. Based on these events, we can estimate intra-cell travel distances and time accordingly.

In the second sequence of steps, we process the tower list for each city, by generating a Voronoi diagram based on tower coordinates. Here, Voronoi diagrams are used to simulate the tower coverage map, based on which we calculate all intra-cell boundary-to-boundary distance lower bounds. We keep such bounds in a separate list for lookup needs.

Finally, at the end of both processing sequences, we aggregate their results to estimate each user's speed distributions. We observe that for some segments, we do not have sufficient location information to accurately estimate a particular user's speed. Under such scenarios, we develop a compensation step where we try to infer the most likely speed based on speed distributions of this user in adjacent segments. The assumption is that one user will not change speed too much in short distances. We next discuss each component in more details in the following sections.

4.2 Data Segmentation by Identifying Pass-Boundary Events

As a user could be anywhere inside the tower's coverage area, we need to infer their speeds by exploiting multiple coverage areas. To this end, we first decompose the trace into segments, which we call "pass-boundary events" (PBE).

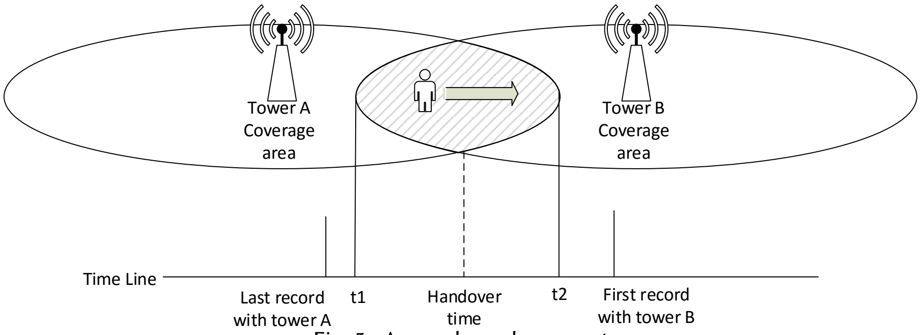


Fig. 5. A pass-boundary event.

Formally, a PBE is defined as when a user moving from one tower's coverage cell into an adjacent tower's coverage cell. There are two properties related to a PBE: first, each PBE has a boundary area, which is the overlapping coverage area of two towers; second, each PBE is associated with a time period of the user spent on crossing the boundary area. For example, in Fig. 5, the PBE event is associated with a boundary as the shadowed area, while its associated time period is (t_1, t_2) for entering and leaving this shadow area.

We now describe the algorithm on extracting PBEs from the mobile data access traces. For arbitrary two consecutive records r_i and r_j of a user with l_i and l_j as their location estimates respectively, if $l_i \neq l_j$, we define a PBE, denoted by $P_{i,j}$ as follows:

$$P_{i,j} = (r_i, r_j)$$

We denote the boundary of $P_{i,j}$, which is the overlapped area, as (l_i, l_j). We use $({t_i}^{last}, {t_j}^{first})$ as an estimate of the time period of entering and leaving $P_{i,j}$, where ${t_i}^{last}$ is the last record timestamp in l_i, and ${t_j}^{first}$ is the first record timestamp in l_j. The length of the time period of $P_{i,j}$ is bounded by ${t_j}^{first} - {t_i}^{last}$. Once PBEs are defined, we use them as reference points to decompose mobile data records of each user into segments.

ALGORITHM 1: Data segmentation

Data: *Trace*: **mobile data trace arranged by users u with record entries e sorted by time. Each e has location estimate l_e and timestamp t_e. We use l_c to represent location of current entry.**

Result: R: segments r arranged by user and time. Its location l_r is defined by location of two adjacent PBE. Its time t_r is defined by the time of first data entry and time of last data entry in this segment.

```
for each $u$ in $Trace$ do
    $l_c \leftarrow \emptyset$;
    $r \leftarrow \emptyset$;
    $e_{lastend} \leftarrow \emptyset$;
    $e_{start} \leftarrow \emptyset$;
    $e_{end} \leftarrow \emptyset$ ;
    for each $e$ in $Trace[u]$ do
        if $l_c \neq \emptyset$ and $l_c \neq l_e$ then
            $l_r \leftarrow (l_{e_{lastend}}, l_c, l_e)$, $t_r \leftarrow (t_{e_{start}}, t_{e_{end}})$ ;
            append $r$ to $R[u]$ ;
        end
        if $l_c = \emptyset$ or $l_c \neq l_e$ then
            $e_{lastend} \leftarrow e_{end}$, $l_c \leftarrow l_e$, $e_{start} \leftarrow e$
        end
        $e_{end} \leftarrow e$ ;
    end
    $l_r \leftarrow (l_{e_{lastend}}, l_c, l_e)$, $t_r \leftarrow (t_{e_{start}}, t_{e_{end}})$ ;
    append $r$ to $R[u]$ ;
end
return $R$
```

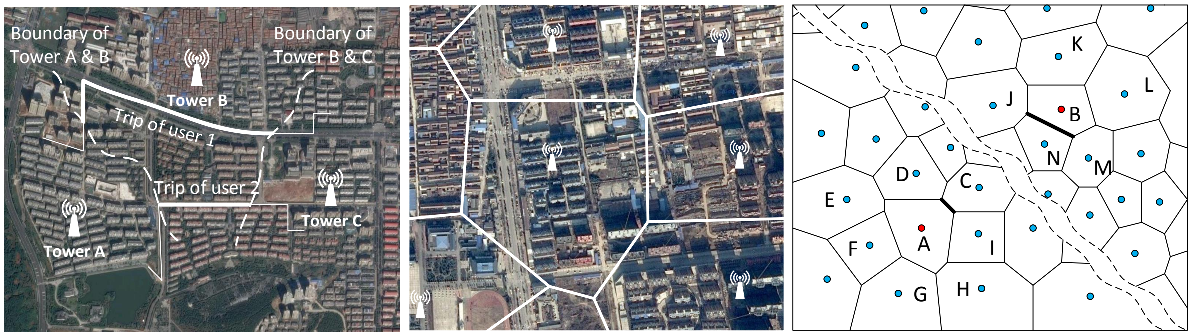
The detailed segmentation algorithm is shown in Algorithm 1. Specifically, the decomposition works as follows: we first generate the PBEs for each user, and then we consider all records of a user between two consecutive PBEs as one single stretch of *continuous stay* as such records should be communicating with the same tower. Therefore, they should share the same location estimate. Since we do not have observations on the intra-cell trajectories of user mobility, we consider the user to have the single constant speed for each stretch within a cell, i.e., between two PBE events. To estimate this speed, we use two consecutive PBEs. Note, however, that as the first and the last stretches of records only have one PBE each, they will not have speed estimates.

4.3 Distance and Time Estimation

We next describe how we estimate the speed between two PBEs. Specifically, we need to estimate the intra-cell boundary-to-boundary distance and the travel time. As the only available information for distance estimation is tower coordinates and tower visiting orders, for a segment with two PBEs $P_{i,j}$ and $P_{j,k}$, we use a straight line trajectory $l_i \rightarrow l_j \rightarrow l_k$ that passes all three tower l_i, l_j and l_k as an estimated trajectory. With the coordinates of towers, the euclidean distance between towers, $d(l_i, l_j)$ and $d(l_j, l_k)$, can be calculated. Since the boundaries are perpendicular bisectors of lines connecting towers (as we use Voronoi diagrams to represent cell coverage areas), the travel distance can be estimated by $\frac{d(l_i,l_j)+d(l_j,l_k)}{2}$. Note that if more information such as the underlying road networks are provided, the road trajectories that has the maximum likelihood to match visited tower sequences can also be used instead of the straight line trajectories.

The travel time of a segment is calculated by the time difference of two related PBEs. Since each PBE has a time interval associated with it for entering and leaving the overlapping area, we can calculate a range of possible values for travel time estimation, including both a tight bound and a relaxed bound. The former one suggests the shortest possible travel time to move through the area, while the latter one indicates the longest possible travel time. For example, for two PBEs $P_{i,j}$ and $P_{j,k}$ with a time interval of $(t_i{}^{last}, t_j{}^{first})$ and $(t_j{}^{last}, t_k{}^{first})$, respectively, we can easily derive the tight bound as $\Delta t_{tight} = t_j{}^{last} - t_j{}^{first}$ and the relaxed bound is $\Delta t_{loose} = t_k{}^{first} - t_i{}^{last}$.

4.3.1 Distance Lower Bounds. In this section, we introduce the concept of distance lower bounds. This is motivated by the observation that it is usually hard to accurately estimate the true distances of users using the coverage areas of given towers for all possible trajectories and represent them with a single distance estimate. To see this, we show an example in Fig. 6(a).



(a) (b) (c)

Fig. 6. Distance estimation methodology. (a) Common cases where a single distance estimate would fail. (b) Voronoi diagram to represent communication coverage of each tower. (c) Dealing with virtual boundaries.

As shown in Fig. 6(a), the area is divided into coverage areas of three towers A, B, and C. Solid lines represent real user trajectories while dashed lines represent the boundaries of towers. Observe that both user 1 and user 2 pass the three towers in the same order $A - B - C$. The real distance differences, however, are missing due to the limited location estimation accuracy of using tower locations. Therefore, in such cases, a single distance estimate will have to fail due to the wide variety of possible trajectories that can lead to the same tower visiting orders.

Faced with this challenge, our next goal is to filter out distance estimates that are not likely to occur in real world scenarios, and provide the trajectory that is most likely as the solution. The major step here is to evaluate the confidence levels of different distance estimates based on estimated trajectories and tower locations so that such confidence levels can be used as measures for evaluating differences in multiple trajectory lengths. Specifically, for two consecutive boundary events $P_{i,j}$ and $P_{j,k}$, the confidence level of a distance estimate d_{est} is defined as $C_{d_{est}} = \frac{d_{lb}}{d_{est}}$, where d_{lb} is the boundary-to-boundary distance lower bound, i.e., the minimum required distance to travel from the boundary of $P_{i,j}$ to the boundary of $P_{j,k}$, which serves as a conservative estimate for the shortest distance a user may travel. Intuitively, the longer an estimated distance is compared to this lower bound, the less likely it should be as it requires a more complex trajectory shape to be feasible.

In order to calculate the distance lower bound, we first simplify the tower coverage model with the Voronoi diagram. Then, based on the Voronoi diagram formed by towers' locations, we calculate the Voronoi cell shapes with their vertex locations. Fig. 6(b) shows an example of the Voronoi diagram construction with five towers. Each region in the Voronoi diagram represents the coverage area of one tower, while the edges in Voronoi diagrams are central focus lines of overlapping coverage area of towers (such areas are hidden in simple Voroni diagrams, but they widely exist in real-world tower communications). The shortest travel distance between boundaries is therefore transformed into the shortest distance between two Voronoi edges, and can be solved using simple geometric methods. The detailed algorithm is shown in Algorithm 2.

ALGORITHM 2: Distance lower bound estimation

```
Data: $TC$: a list of tower coordinates.
Result: $D_{lb}$: a list of all boundary-to-boundary distance lower bounds estimated from Voronoi diagram.
$TL \leftarrow TC$ ;
$P \leftarrow TL$ ;
Build Voronoi diagram $VD$ with $P$ ;
for each $edge1$ in $VD$ do
    $pset1 \leftarrow$ Voronoi points of $edge1$ ;
    for each $edge2$ in $VD$ do
        $pset2 \leftarrow$ Voronoi points of $edge2$ ;
        if $pset1 \cap pset2 \neq \emptyset$ then
            $D_{lb}[(pset1, pset2)] \leftarrow |edge1 - edge2|$;
        end
    end
end
return $R$
```

4.3.2 Virtual Boundaries. A fundamental limitation of using cellphone-tower communication datasets is that records are only collected when mobile data accesses are happening. If the user is keeping silent, there is no way for us to know their locations. In such cases, if the user has traveled across multiple boundaries, we may encounter the following observation: we analyze the consecutive records for this user and find out that their l_i and l_j may be far away from each other and do not necessarily share a common boundary area. If l_i and l_j are adjacent to each other, we say to $P_{i,j}$ has a real boundary. Otherwise, we refer to it as a virtual boundary.

Different from real boundaries that are treated as an edge in the Voronoi diagram, virtual boundaries are actually distance estimates themselves as users have passed the coverage area of several towers during a PBE within a virtual boundary. Since we do not have any information regarding which towers the user has visited in between, to calculate the distance lower bound of a virtual boundary, we instead use the shortest distance of all possible boundary pairs of l_i and l_j as the best estimate.

We now give an example in Fig. 6(c), where we analyze two consecutive records: r_i for tower A and r_j for tower B. Since tower A and tower B do not share physical boundaries, they only have a virtual boundary between them.

To calculate their shortest distance, we calculate the distance from each boundary of tower A to each boundary of tower B and use the shortest one of all boundary pairs as the estimated distance. In this example, the distance between boundary (A, C) and boundary (B, N) is used as the distance lower bound of the virtual boundary (A, B).

Returning to our earlier analysis, for segments that have PBEs with virtual boundaries, we merge them with adjacent segments if they exist. The distance estimate and lower bound of a segment are the sum of distance estimates and distance lower bounds of both records, and if any, the virtual boundaries between them. Note that as we calculate the sum of distance lower bounds, the resulting distance lower bound is still the minimum distance required to reach one real boundary from the other, even this requires that the trajectory should pass through virtual boundaries between consecutively visited towers.

4.4 Speed Estimation

Now that we have a distance estimate d_{est}, a distance lower bound d_{lb}, and a range of possible travel time represented as $(\Delta t_{tight}, \Delta t_{loose})$ for each segment, we can infer the travel time of a segment estimated by $\Delta t_{est} = \frac{\Delta t_{tight} + \Delta t_{loose}}{2}$. We denote this by Δt_{est}. We next calculate the confidence levels for both distance estimates and travel time estimates as follows:

$$C_{d_{est}} = \frac{d_{lb}}{d_{est}} \tag{1}$$

$$C_{\Delta t_{est}} = \frac{\Delta t_{est}}{\Delta t_{loose}} \tag{2}$$

By setting a threshold for both confidence levels, we can filter out estimates that are not accurate enough. Although we can filter out more inaccurate speed estimates with a much stricter threshold in both confidence levels, we may end up with a limited number of records that have qualified speed estimates. Finally, after setting proper threshold for confidence levels, the speed of the user can be estimated as the following:

$$s_{est} = \frac{d_{est}}{\Delta t_{est}} \tag{3}$$

The detail of the speed estimation algorithm is shown in Algorithm 3.

4.4.1 *Cell Oscillation and Speed Compensation.*

The distance lower bounds can also help to eliminate the cell oscillation problem, i.e., when a user near boundary area randomly communicates with two or more towers in short periods, generating a sequence of false pass-boundary events. Since the user keeps passing the same boundary, the distance lower bound for such scenarios should always be 0. Therefore, the confidence level of distance estimates will also be 0, which means that we can detect them and filter them out. **The distance lower bounds can also help when the user is near the boundary of multiple towers as the distance lower bound will also be very low in such cases.**

Since segments between these false PBEs usually have very short durations due to the nature of how they are generated, we estimate the speed for such segments based on the assumption that a user's speed does not change dramatically in a very short time period. Therefore, for a segment between false PBEs, if there is a segment that happens to be very close to it and has a qualified speed estimate, we will use its speed estimate as the speed estimate for the segment with false PBEs. Other kinds of low confident level speed estimations can also be compensated by the nearby segments with high confidence levels as long as the confidence level and time period are properly handled.

5 (M3) PERFORMANCE EVALUATION

ALGORITHM 3: Speed estimation

```
Data: $R$ from Algorithm 1 and $D_{lb}$ from Algorithm 2.
Param: $T_{C_d}$, $T_{C_t}$: confidence level threshold for distance estimates and travel time estimates.
Result: $S$: Speed estimates for each segment.
for each $r$ in $R$ do
    //Check if $r$ has real boundary and find its distance lower bound
    if $((l_{pre}, l), (l, l_{post}))$ is not in $D_{lb}$ then
        combine $r$ with next record ;
        continue ;
    else
        $d_{lb} \leftarrow D_{lb}[(l_{pre}, l), (l, l_{post})]$ ;
    end
    //Calculate travel time estimates
    $\Delta t_{est} \leftarrow \frac{\Delta t_{tight} + \Delta t_{loose}}{2}$ ;
    //Calculate confidence level
    $Cd_{est} \leftarrow \frac{d_{lb}}{d_{est}}$ ; $C\Delta t_{est} \leftarrow \frac{\Delta t_{est}}{\Delta t_{loose}}$ ;
    //Estimate speed if meet threshold
    if $Cd_{est} \geq T_{C_d}$ and $C\Delta t_{est} \geq T_{C_t}$ then
        $s_{est} \leftarrow \frac{d_{est}}{\Delta t_{est}}$ ; $S[r] \leftarrow s_{est}$ ;
    end
end
return $S$
```

In this section, we perform an evaluation of our speed estimation algorithm based on a similar but much smaller dataset [12] collected by Intel Placelab. This small dataset contains both cell phone data and GPS traces. We use the GPS as ground truth for the evaluation. The basic motivation for this study is that by proving our approach is effective for the small dataset with ground truth available, it is more likely for our estimations for large datasets to be accurate even if we cannot prove so without ground truth data for such large-scale datasets.

5.1 Dataset and Experimental Setup

The dataset we used was collected in the Seattle area in September 2004. Both cellphone data and GPS data were collected with a Nokia 6600 cellphone and a separate GPS unit. Similar to our dataset, the cellphone data contains only the IDs of the cell towers. Note that as the dataset was collected more than ten years ago, we can only obtain accurate coordinations for a subset of all tower IDs contained in the dataset. We thus discarded part of the dataset that contains tower IDs with unknown coordinates, as we cannot recover the ground truth for these towers. For the remaining valid data, the segments for user traces used in our evaluation contains several trips with a total aggregate period of 40 minutes and 2,007 records.

To make the comparison fair, we used the same parameter settings as those used in our large-scale dataset analysis. The thresholds of both the distance ratio d_{ratio} and the duration ratio Δt_{ratio} are set as 0.6. We are able to obtain speed estimations for 1,955 records out of the 2,007 records. After the filtering procedure, we have 571 records that meet both standards. We also disable both the distance

lower bound filter and the travel time filter by setting thresholds to 0 and use the raw speed estimates as our baseline for comparison. Note that, as we previously stated, if more information such as the underlying road networks is provided, our distance and travel time filters can also be applied to road trajectories to obtain the most likely ones that match the visited tower sequences instead of the straight line trajectories.

5.2 Speed Estimation Accuracy

Fig. 7 shows the CDF (cumulative distribution function) of errors for both the filtered speed estimates and the raw speed estimates. Here, we define the speed estimation errors as $e = |(s_{est} - s_{gt})|/s_{gt}$, where s_{est} is the speed estimated by cell phone data and s_{gt} is the ground truth speed calculated by GPS data. By using the absolute value, e is always a positive value. As the value of e can reach very high due to cell oscillations, we set upper limits of the estimated speed, hence on e, in Fig. 7 to better illustrate the difference of speed estimation errors that fall into reasonable error ranges.

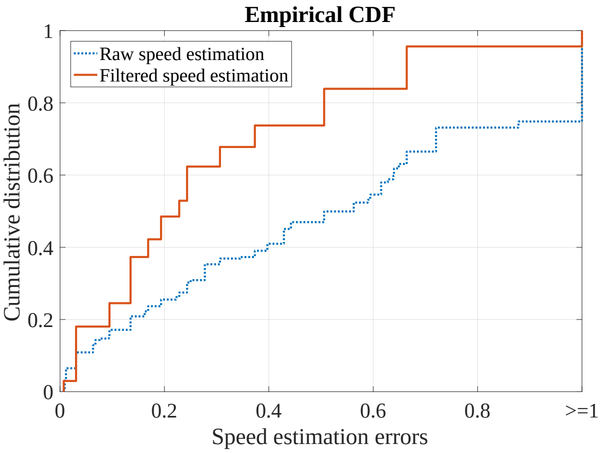


Fig. 7. The CDF of raw speed estimation and filtered speed estimation

As we can see in Fig. 7, the filtered speed estimation significantly outperforms the raw speed estimation in terms of accuracy. About 50% of the filtered speed estimates have error rates less than 0.2, while only 25% of raw speed estimates achieve the same level of accuracy. More than 80% of filtered speed estimates have error rates less or equal to 0.5, while for raw speed estimates this number is only less than 50%. On the other end of this figure, observe that there is only less than 5% of filtered speed estimates with error rates higher than 1, compared to more than 15% of raw speed estimates.

5.3 Speed Filtering

Fig. 8 shows the number of records in each speed estimation error range, from 0.1 to 1 with an interval of 0.1. Observe that our speed estimation algorithm works well compared to raw estimates. For example, we can observe that approximately 50% of records that have raw speed estimation errors

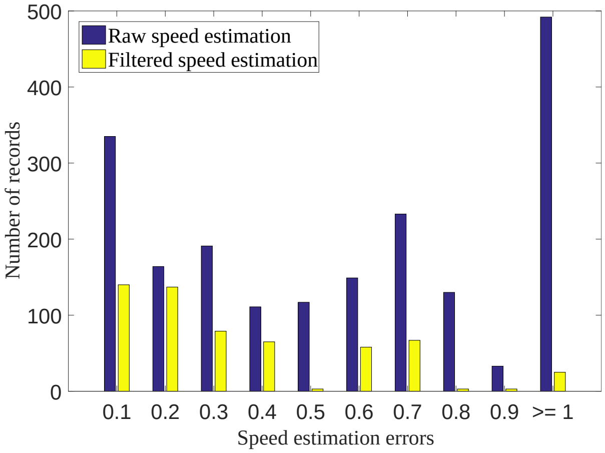


Fig. 8. Number of records in each error bracket for raw speed estimation and filtered speed estimation

less than 0.5 have been filtered out, while more than 85% of records that have raw speed estimation errors greater than 0.5 have been filtered out.

6 LARGE-SCALE DATA ANALYSIS

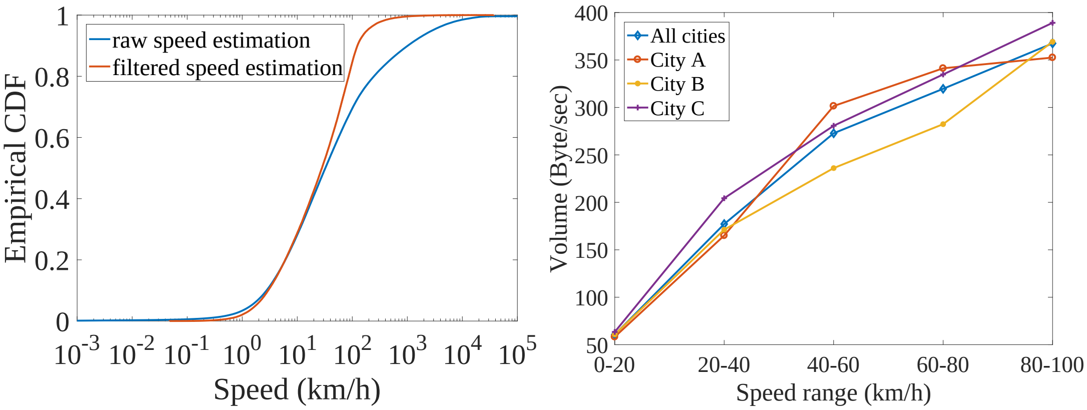
With our methodology on speed estimates, we next explain our findings on correlations between user mobility and mobile data access patterns in this section. We start with the correlation of the speed and the average mobile data access volumes. Then we reveal the relation of speed and average time intervals between consecutive mobile data accesses. Finally, we illustrate the correlation between speed and the types of app usage that are responsible for generating the corresponding mobile data traffic.

(R1-4) We implement our speed estimation algorithm and mobile data usage pattern analysis algorithm in Python. We use the Voronoi package from Scipy [11] to construct Voronoi maps with tower coordinates and Shapely [7] for geometry calculations. All analysis is carried out on a single Cloudlab [22] c8220 server with two 10-core 2.2GHz E5-2660 processors and 256GB memory.

6.1 Speed and Data Volumes

To estimate the speed, our algorithm requires a user has visited at least 3 towers consecutively. In the dataset, we find that around 13 million records out of 58 million records can be utilized. Although the dataset contains both user initiated network access and background network access, we find it very hard to separate them reliably. In our experiments, to balance the accuracy of speed estimates and the number of mobile data access records that have qualified speed estimates, we set the threshold of both distance ratio d_{ratio} and duration ratio Δt_{ratio} empirically as 0.6. After the filtering, we have around 1 million records out of total 13 million records that meet both criteria. Fig. 9(a) shows the cumulative density function (CDF) of both raw speed estimates without filtering and filtered speed estimates. As we can see that the filtered speed estimates are more realistic compared to raw speed estimates. Most of the false high speed estimates and low speed estimates are filtered out by setting thresholds of confidence levels for distance estimates and travel time estimates. In the following experiments, we only show results in the speed range from 0 km/h to 100 km/h, since there are very few records with a speed estimate above 100 km/h for any meaningful insights.

Fig. 9(b) shows the results of the correlation of user speed and the average mobile data access volumes per user per second. We demonstrate the data from all three cities combined and each city respectively. The figure

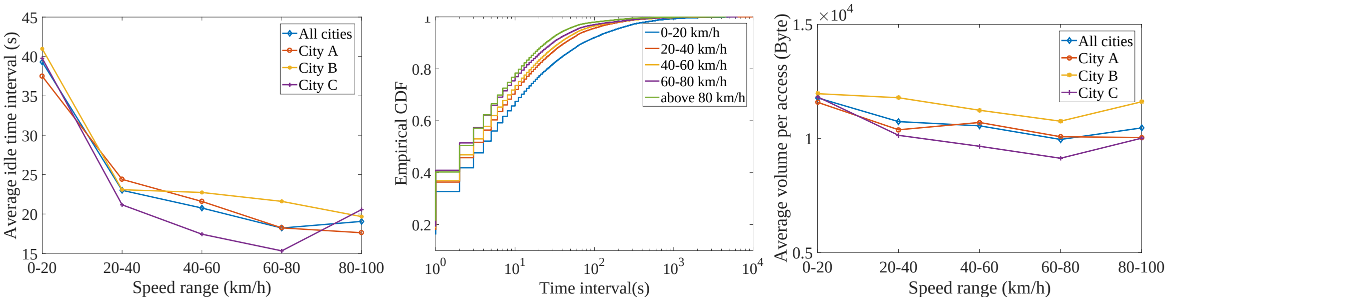


(a) Empirical CDF of speed estimates.

(b) Correlation of speed estimates with data volumes.

shows a clear trend that users are more active in accessing mobile data as the speed increases and the trend holds true for all three cities. In fact, a user with speed estimates of 80-100 km/h could reach an average data volume of 6 times of a low-speed user. Similarly, this trend also holds true for all the cities. Note that these results only show an increase in the mobile data access volume as user speed increases. It does not suggest lower speed users access online contents less frequently. Actually, we believe one reason might be that a large portion of a low-speed user's online needs is already fulfilled by various kind of high-speed connections such as Wifi hotspots. To this end, we reach similar findings with previous work [41] on the correlation of user mobility and mobile data access volume, except that the previous work used the number of towers visited by a user as the indicator of user mobility.

6.2 Speed and Access Frequency



(c) Average Time Interval

(d) Time Interval Empirical CDF

(e) Average Volume per Data Access

Fig. 9. The correlation of speed estimates with (a) average idle time interval between consecutive data access, (b) idle time interval between consecutive connections, (c) average data access volume for each data access.

Fig. 9(c) shows the correlation of speed and average idle time intervals between consecutive mobile data access records. **Here a mobile data access record is defined as a single record entry in our dataset.** The CDF of data time intervals for various speed ranges of all three cities are also shown in Fig. 9(d). Note that since the time precision of our data trace is seconds, so there are steps in Fig. 9(d). The decrease in time intervals as speed increases suggests that high-speed user accesses mobile data more frequently than low-speed users. A user with a speed estimate of 80-100 km/h access mobile data almost twice more frequently than a user with a speed estimate of 0-20 km/h on average. The trend holds for all three cities except that there is an odd point at 80-100 km/h for one city, which may be caused by the lacking of available data.

We show the average volume for each data access in Fig. 9(e). As the user speed increases, there is no apparent correlation with average volume for each data access. This suggests that increasing in the average volume which is shown in Fig. 9(b) is mainly cause by the increased data access frequency, not the volume for each data access.

6.3 Speed and App Usage

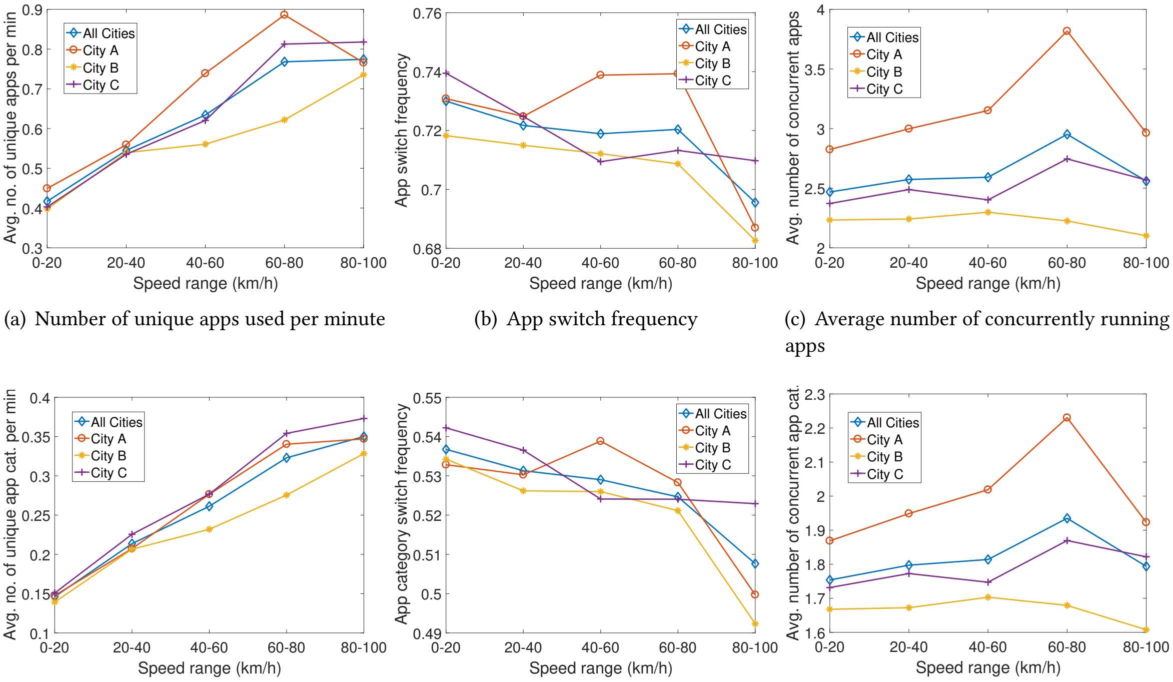
According to the mobile service provider, each app in our dataset was assigned to one of 18 categories, as shown in Table 1.

App Category	# Apps	Volume (GB)
Instant Messages	30	97.3
Reading	101	17.6
Microblog	43	13.0
Navigation	38	10.8
Video	63	45.2
Music	33	27.4
App Market	45	37.0
Game	106	9.2
Payment	18	1.2
Comic	12	0.8
Email	10	1.5
P2P	8	3.9
VOIP	17	0.3
Multimedia Messages	2	0.3
Browsing	558	353.5
Finance	25	0.7
Security	22	5.2
Others	244	95.8

Table 1. App categories

Fig. 10(a) and Fig. 10(d) show the correlation between user speed and the average number of unique apps and app categories being used for each user during each data segment per minute. The trend clearly shows that as the speed goes up, the app usage diversity increases rapidly. A user with a speed estimate of 80-100 km/h could use as many as 2 times apps per unit of time compared to a low-speed user. An explanation might be that for users with high mobility, they may use their phones more often, use more kinds of apps, and be less likely to focus on one app for prolonged periods of time.

In Fig. 10(b) and Fig. 10(e) we show the frequency of app switch and app category switch. They both follow the same trend that lower speed users tend to switch apps and app categories more frequently. Combined with Fig. 10(a) and Fig. 10(d), we can see that although low speed users switch apps more frequently, they only switch among fewer number of apps.



(a) Number of unique apps used per minute

(b) App switch frequency

(c) Average number of concurrently running apps

(d) Number of unique app categories used per minute

(e) App category switch frequency

(f) Average number of concurrently running app categories

Fig. 10. Correlation of user speed and the number of unique apps and app categories used

We then further extended our study by investigating the correlation of the number of concurrently running apps and app categories with estimated user speed. The results are shown in Fig. 10(c) and Fig. 10(f). Although the actual launch time and closing time for each running app are unavailable, we use the network access data to infer such information. For each app, we treat the time of the first data access as the launch time of an app. If the app stops accessing network for a certain period (referred to as the closing threshold), we record the last network access time as the closing time for the app. We chose 10 seconds as the closing threshold to plot Fig. 10(c) and Fig. 10(f), but we had tested various other thresholds and found similar trends. Interestingly, in most cities, the peak happens at the 60-80 km/h and decreases as the speed further increases.

Finally, we further investigated the trend of the contribution of various app categories on the total mobile data access as the user speed increases. The contribution was defined as the mobile data access of one category versus all categories. Focusing on apps that contributed the most to the total mobile data access volume, we selected the top 8 app categories. The correlation between the user speed and the contribution of each category is shown in Fig. 11.

Among the top 8 categories, the Microblog, Navigation and Music categories show a clear upward trend as the speed increases. The impact of navigation has the most steady increase due to the increased needs for such apps when driving. The impact almost doubles for users with speed estimates of 80-100 km/h compared to users with speed estimates of 0-20 km/h. Instant message, Video and App market show a downward trend as the speed increases. The reason could be the users are cost sensitive and strictly control the data usage for large app

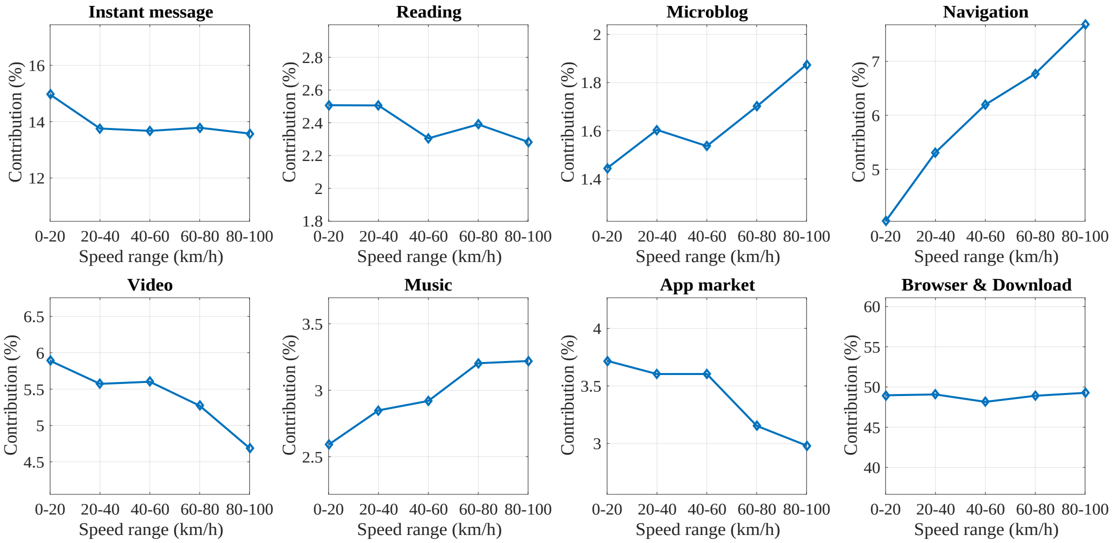


Fig. 11. Correlation of user speed and contribution of app categories.

downloading and video streaming. Browser & Downloading and Reading show a quite stable impact that does not change a lot as the speed increases.

7 DISCUSSIONS AND LIMITATIONS

Despite the results based on our analysis of the large-scale dataset, we acknowledge that our study still has a number of limitations. It is pivotal to discuss them so that our results can be interpreted in a meaningful manner.

(M2-a, M4) First, our speed estimation algorithm is solely based on the coarse-grained location information from cell phone data access traces. We have not integrated more fine-grained WiFi data for the reason that cell phone network still has far better coverage compared to WiFi networks. Moreover, users on cell phone networks usually have a higher degree of speed variations, whereas users using WiFi are more likely to be stationary.

On the other hand, we discuss briefly on how to integrate WiFi data to improve speed estimation, if such data were available. We observe that the coverage patterns of WiFi access points (APs) are very different from cell phone towers. They are usually heavily overlapped with each other. Furthermore, WiFi APs usually have smaller coverages than cell phone towers. Therefore, if we have WiFi traces, we can estimate the locations of users more accurately by using triangularization methods. Such locations can serve as calibration records for the estimated trajectories. For example, when estimating the intra-cell boundary-to-boundary distances, instead of using a straight line trajectory that passes all adjacent tower as an estimated trajectory, we can use the trajectory that passes all recorded WiFi coverage areas in between as the better estimates. However, in this case, the distance lower bound should also be the shortest distance between two boundaries that passes all recorded WiFi coverage areas in between. Note that the distance lower bound might now be a straight line in this case.

(M2-b, R1-3, R4-3) A second limitation is that our dataset has limited temporal and spatial coverage. Therefore, we are not able to perform in-depth studies on temporal or spatial trends. Further, although our dataset contains data from three cities, they only represent patterns of densely populated areas. We believe that our speed estimation methods can be easily applied to similar datasets from other areas. For example, in our evaluation section, we use a smaller dataset from a whole different area to verify the performance of our speed estimation algorithm.

(M2-c) Finally, our speed estimation algorithm utilizes passing-boundary events as building blocks. This technique requires that users have visited at least three nearby towers. This is because of the fact that, with tower coordinates as the only location information, it is almost impossible to infer user locations in a specific cell phone tower coverage area based on this tower alone. Therefore, if a user's trace is recorded by fewer than three towers, the uncertainty of speed estimation will undoubtedly increase.

8 CONCLUSIONS

In this paper, we studied the correlation between user mobility and app usage patterns. In particular, we focused on users' moving speed as the key mobility metric. A key challenge addressed by our methodology is to estimate speeds accurately with high confidence and reliability. **We verify our methodology with out of sample data. The results shows great improvement in estimation accuracy.** Based on the speed estimations, we are able to reveal the correlation of user mobility with mobile data usage patterns including the data volume, the data access frequency, and the traffic share of apps on the total mobile data traffic. Results showed that with users that have high speed estimation tend to user smartphone more frequently and generate more traffic on the mobile data network. Furthermore, the user speed also played an important role in the contribution of each smartphone app categories on the total mobile data traffic.

REFERENCES

[1] Paul Baumann and Silvia Santini. 2014. How the availability of Wi-Fi connections influences the use of mobile devices. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 367–372.

[2] Shlomo Bekhor and I Blum Shem-Tov. 2015. Investigation of travel patterns using passive cellular phone data. *Journal of Location Based Services* 9, 2 (2015), 93–112.

[3] Filip Biljecki, Hugo Ledoux, and Peter Van Oosterom. 2013. Transportation mode-based segmentation and classification of movement trajectories. *International Journal of Geographical Information Science* 27, 2 (2013), 385–407.

[4] Eunjoon Cho, Seth A Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *KDD*. 1082–1090.

[5] Ericsson. 2016. Ericsson Mobility Report. (Feb. 2016). http://www.ericsson.com/res/docs/2016/mobility-report/ericsson-mobility-report-feb-2016-interim.pdf.

[6] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2012. Next place prediction using mobility markov chains. In *MPM*. 3.

[7] Sean Gillies and others. 2013–. Shapely. (2013–). https://pypi.python.org/pypi/Shapely [Online; accessed ¡today¿].

[8] Samuli Hemminki, Petteri Nurmi, and Sasu Tarkoma. 2013. Accelerometer-based Transportation Mode Detection on Smartphones. In *SenSys*. 13:1–13:14.

[9] Sahar Hoteit, Stefano Secci, Stanislav Sobolevsky, Carlo Ratti, and Guy Pujolle. 2014. Estimating human trajectories and hotspots through mobile phone data. *Computer Networks* 64 (2014), 296–307.

[10] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. 2013. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *UrbComp*.

[11] Eric Jones, Travis Oliphant, Pearu Peterson, and others. 2001–. SciPy: Open source scientific tools for Python. (2001–). http://www.scipy.org/ [Online; accessed ¡today¿].

[12] Anthony LaMarca, Yatin Chawathe, Jeffrey Hightower, and Gaetano Borriello. 2004. CRAWDAD dataset intel/placelab (v. 2004-12-17). Downloaded from http://crawdad.org/intel/placelab/20041217/placelab. (Dec. 2004). DOI:http://dx.doi.org/10.15783/C7KS30 traceset: placelab.

[13] Kyunghan Lee, Joohyun Lee, Yung Yi, Injong Rhee, and Song Chong. 2010. Mobile data offloading: How much can WiFi deliver?. In *Proceedings of the 6th International COnference*. ACM, 26.

[14] Ilias Leontiadis, Antonio Lima, Haewoon Kwak, Rade Stanojevic, David Wetherall, and Konstantina Papagiannaki. 2014. From cells to streets: Estimating mobile paths with cellular-side data. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*. ACM, 121–132.

[15] Vincenzo Manzoni, Diego Maniloff, Kristian Kloeckl, and Carlo Ratti. 2010. *Transportation mode identification and real-time CO_2 emission estimation using smartphones*. Technical Report. SENSEable City Lab, Massachusetts Institute of Technology.

[16] Lei Meng, Shu Liu, and Aaron D Striegel. 2014. Analyzing the impact of proximity, location, and personality on smartphone usage. In *INFOCOM WKSHPS*. 293–298.

[17] Mathé Young Mosny. 2006. *Path Estimation Using Cellular Handover*. Bachelor of Science Thesis. Princeton University.

[18] Nielsen. 2014. SMARTPHONES: SO MANY APPS, SO MUCH TIME. (July 2014). http://www.nielsen.com/us/en/insights/news/2014/smartphones-so-many-apps--so-much-time.html.

[19] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. 2011. An Empirical Study of Geographic User Activity Patterns in Foursquare. In *ICWSM*. 570–573.

[20] Hiroki Ohashi, Takayuki Akiyama, Masaaki Yamamoto, and Akiko Sato. 2014. Automatic trip-separation method using sensor data continuously collected by smartphone. In *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*. IEEE, 2984–2990.

[21] Sasank Reddy, Min Mun, Jeff Burke, Deborah Estrin, Mark Hansen, and Mani Srivastava. 2010. Using Mobile Phones to Determine Transportation Modes. *ACM Trans. Sen. Netw.* 6, 2 (2010), 13:1–13:27.

[22] Robert Ricci, Eric Eide, and The CloudLab Team. 2014. Introducing CloudLab: Scientific Infrastructure for Advancing Cloud Architectures and Applications. *USENIX ;login:* (2014).

[23] Geoff Rose. 2006. Mobile phones as traffic probes: practices, prospects and issues. *Transport Reviews* 26, 3 (2006), 275–291.

[24] Jason Ryder, Brent Longstaff, Sasank Reddy, and Deborah Estrin. 2009. Ambulation: A tool for monitoring mobility patterns over time using mobile phones. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, Vol. 4. IEEE, 927–931.

[25] M Zubair Shafiq, Lusheng Ji, Alex X Liu, Jeffrey Pang, and Jia Wang. 2012. Characterizing geospatial dynamics of application usage in a 3G cellular data network. In *INFOCOM*. 1341–1349.

[26] Dongyoun Shin, Daniel Aliaga, Bige Tunçer, Stefan Müller Arisona, Sungah Kim, Dani Zünd, and Gerhard Schmitt. 2015. Urban sensing: Using smartphones for transportation mode classification. *Computers, Environment and Urban Systems* 53 (2015), 76–86.

[27] Zbigniew Smoreda, Ana-Maria Olteanu-Raimond, Thomas Couronné, and others. 2013. Spatiotemporal data from mobile phones for personal mobility assessment. *Transport Survey Methods: Best Practice for Decision Making* 41 (2013), 745–767.

[28] Statista. 2016. Number of apps available in leading app stores as of July 2015. (2016). http://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/.

[29] Leon Stenneth, Ouri Wolfson, Philip S Yu, and Bo Xu. 2011. Transportation mode detection using mobile phones and GIS information. In *GIS*. 54–63.

[30] Carlo Tacconi, Sabato Mellone, and Lorenzo Chiari. 2011. Smartphone-based applications for investigating falls and mobility. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011 5th International Conference on*. IEEE, 258–261.

[31] Karol Waga, Andrei Tabarcea, Minjie Chen, and Pasi Franti. 2012. Detecting movement type by route segmentation and classification. In *Collaborative computing: networking, applications and worksharing (CollaborateCom), 2012 8th International Conference on*. IEEE, 508–513.

[32] Daniel T Wagner, Andrew Rice, and Alastair R Beresford. 2014. Device Analyzer: Large-scale mobile data collection. *ACM SIGMETRICS Performance Evaluation Review* 41, 4 (2014), 53–56.

[33] Huayong Wang, Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. 2010. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *ITSC*. 318–323.

[34] Shuangquan Wang, Canfeng Chen, and Jian Ma. 2010. Accelerometer based transportation mode recognition on mobile phones. In *APWCS*. 44–46.

[35] Tingting Wang, Cynthia Chen, and Jingtao Ma. 2014. Mobile phone data as an alternative data source for travel behavior studies. In *Transportation Research Board 93rd Annual Meeting*.

[36] Peter Widhalm, Philippe Nitsche, and Norbert Brändie. 2012. Transport mode detection with realistic smartphone sensor data. In *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 573–576.

[37] Peter Widhalm, Yingxiang Yang, Michael Ulm, Shounak Athavale, and Marta C González. 2015. Discovering urban activity patterns in cell phone data. *Transportation* 42, 4 (2015), 597–623.

[38] Qiang Xu, Jeffrey Erman, Alexandre Gerber, Zhuoqing Mao, Jeffrey Pang, and Shobha Venkataraman. 2011. Identifying diverse usage behaviors of smartphone apps. In *IMC*. 329–344.

[39] Kuldeep Yadav, Amit Kumar, Aparna Bharati, and Vinayak Naik. 2014. Characterizing mobility patterns of people in developing countries using their mobile phone data. In *Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on*. IEEE, 1–8.

[40] Tingxin Yan, David Chu, Deepak Ganesan, Aman Kansal, and Jie Liu. 2012. Fast app launching for mobile devices using predictive user context. In *MobiSys*. 113–126.

[41] Jie Yang, Yuanyuan Qiao, Xinyu Zhang, Haiyang He, Fang Liu, and Gang Cheng. 2015. Characterizing User Behavior in Mobile Internet. *IEEE Transactions on Emerging Topics in Computing* 3, 1 (2015), 95–106.

[42] Lin Yang, Mingxuan Yuan, Wei Wang, Qian Zhang, and Jia Zeng. 2016. Apps on the move: A fine-grained analysis of usage behavior of mobile apps. In *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on*. IEEE, 1–9.

[43] Yu Zheng, Yukun Chen, Quannan Li, Xing Xie, and Wei-Ying Ma. 2010. Understanding transportation modes based on GPS data for web applications. *ACM Transactions on the Web* 4, 1 (2010), 1.

Response to the Review Comments

Dear Editors and Reviewers,

Thank you very much for your insightful feedback and constructive suggestions. We have thoroughly addressed your comments in this revision. In the following, we will first summarize main changes made to the manuscript, and then provide a point-to-point response to comments from both the meta-review and each reviewer.

For each point of change, we assigned a symbol that corresponds to a specific review comment. For example, M3 means Comment 3 in the meta-review, and R1-3 means Comment 3 from Reviewer 1, and so on. Changes to the manuscript are highlighted in blue.

In this revision, we primarily made the following changes.

1. (M3) We performed additional evaluations of our speed estimation algorithm using a dataset that contains ground-truth information (see Section 5 on Page 12).
2. We added a section to discuss the limitations of our work (see Section 7). We highlighted when our system would be useful and valid, and when the results would require additional assumptions. More specifically,
    - (M4) We discussed how to integrate WiFi data into our analysis when such data are available;
    - (M2b, R1-3) We discussed possible limitations of our data with respect to the coverage of time and geographic areas;
    - (M2c) We discussed the limitations introduced by coarse-grained location information and irregular time intervals in the dataset.
3. We added more information about our study as suggested by the reviewers.
    - (M1) We provided more background information and description about our dataset (see Section 3 on Page 4);
    - (M5) We added and discussed additional references (see Section 2 on Page 4);
    - (R1-4) We added more details about implementation and environment settings (see Section 6 on Page 13).
4. (R2-1, R4-1) We revised the motivation of our study to highlight the need and usefulness of our approach (see Section 1 on Page 2).

5. We re-arranged the figures to make them close to the corresponding text, and also fixed several naming issues.

6. We proofread the paper thoroughly to make sure that there is no more language or formatting issue.

```
>> RESPONSE TO META-REVIEW <<
```

```
>  1. (R1) Describe how the data was collected. Who were involved? How long
>  the data was collected? When the data was collected? Is it cell data the
>  only collected data or was WiFi also collected?
```

(M1) Thank you for this comment. Our dataset was collected by a major mobile carrier in China as the operator for cell phone towers. All active GSM data access traces during a three-hour period from 6pm to 9pm on a Sunday in September 2014 were recorded. We added this background information to the paper, and marked the related changes in the paper as (M1). See Section 3 on Page 4.

This dataset only contained cellular data and did not contain any WiFi data. Despite this limitation, we emphasize that the large data size allows us to carry out a study on the real-world app usage with a larger scale and better accuracy than related work.

```
> 2. (R1 and R2) The authors should address the generality of the proposed
> algorithm as a limitation. Generally, smartphone users use both cellular
> and wifi networks in other areas. The current analysis model excludes
> many contexts of users such as location, time, individual context, etc.
> In order to understand the resulting data, it is very important to know
> about moving speed with respect to users and their culture.
```

Thank you for pointing out the generality issue. In this revision, we added a separate section to discuss these issues and possible limitations.

a. (M2a) On the limitation of using WiFi data: We recognize that although WiFi handles half of the mobile data traffic and more than half of time users are connected to WiFi, cell phone networks typically have much better coverage and user mobility diversity. It is more meaningful to study user mobility under cellular traffic from the operators' perspective to understand large-scale user preferences and behavioral patterns. We do acknowledge that the lack of WiFi data is a limitation of our work and that our current speed estimation algorithm cannot be directly applied to WiFi data. However, it is very hard to find WiFi data on a whole city scale, and to our knowledge, no such large-scale dataset has yet collected. Therefore, we highlight this as a limitation due to such practical concerns. We mark the related changes in the paper in Section 7 on Page 17.

b. (M2b) Time and Location contexts: Due to the short time period of our dataset, it was not feasible for us to perform a thorough temporal analysis for much longer periods of time, such as days or weeks. Nevertheless, the speed estimation method can be easily applied to similar datasets from other areas of any given time period to explore time and location contexts. For example, in our evaluations section, we used a different dataset in the U.S. to verify our speed estimation algorithm. We discussed this issue in the limitations section. We mark the related change in the paper in Section 7 on Page 18.

c. (M2c) Individual contexts: We agree that our results only represent those users who frequently moved and use mobile network. Because the dataset is fundamentally coarse-grained, we can only analyze limited location information and irregular time intervals. We state this limitation in our new limitation section in Section 7 on Page 18.

```
>  3. (R2) Speed estimation algorithm is not proved yet. The authors should
>  demonstrate that their approach to estimate intra-cell movement speed is
>  accurate and therefore can be reliable used for the correlation analysis.
>  Many assumptions in estimating moving speed are made without much
>  explanation. These assumptions should be proved and discussed in detail.
>  In addition, the classes of moving speed should be analyzed and discussed
>  in detail.
```

(M3) Following one of the reviewer's suggestions, we used a dataset collected by Intel Placelab to perform additional evaluations of our speed estimation algorithm. This dataset contains both cell phone data and GPS traces. We used the GPS data as the ground truth for evaluating our speed estimation algorithm that utilizes cell phone data only. We also added more details regarding the assumptions on estimating the moving speeds in our algorithm design. The moving speed classes are also analyzed and discussed with more details.

The changes are available in Section 5 on Page 12.

```
> 4. (R1 and R2) The authors should discuss or demonstrate the implications
> of considering cellular traffic only on the results presented in this
> work. WiFi usage could be a major limitation of the generalization of the
> study.
```

Please see our response listed under meta-review Comment 2a (M2a) above.

```
>  5. Suggested References:
>  [1] Kyunghan Lee,Joohyun Lee,Yung Yi,Injong Rhee, and Song Chong. 2013.
>  Mobile Data Offloading: How Much can WiFi Deliver? IEEE/ACM Transactions
>  On Networking 21, 2 (2013), 536-551.
>
>  [2] Paul Baumann and Silvia Santini. 2014. How the availability of Wi-Fi
>  connections influences the use of mobile devices. In Proceedings of the
>  2014 ACM International Joint Conference on Pervasive and Ubiquitous
>  Computing Adjunct Publication - UbiComp ’14 Adjunct. (2014), 367-372.
```

Thank you. We have added all suggested references and discussed them in the related work section. These changes can be found in Section 2 on Page 4 of the paper.

```
>> RESPONSE TO REVIEWER 1 <<
```

```
> Although the authors shows interesting results, there are several
> limitations to be impored. First of all, I'm wondering how the data was
> collected. The authors mainly describes the volume of dataset. However,
> it is very important to know about how the data was collected. The
> population of the dataset might be affected by the culture  and types of
> users.
```

Please see our response to meta-review comment 1.

```
>  Second, the proposed methodology and analysis are too general to be
>  applied to understand smartphone app usage. This is because the contexts
>  users use their smartphones are blurred with moving speed. Generally,
>  user behaviours are divided into moving, sitting and staying at home or
>  office and also related to time of day. Users spend much time on using
>  smartphone when waiting  or sitting for something. The usage patterns are
>  also related to locations such as home or office or stores. Furthermore,
>  the results could not represent the significant part of users since usage
>  patterns are very diverse according to users. As the authors mentioned
>  with Fig 1, the data was mainly from those who frequently moved and used.
>  I think that the moving speed should be related  to more contexts such as
>  time and location. It might be also interesting to see the difference
>  between inter-cell moving and staying within a cell.
```

Please see our response to meta-review comment 2.

```
> Third,  I think that the results are very limited to specific areas.
> Although the data were  from three cities and the size of data is
> comparable with other countries, the results are only for the China. I
> think that the Fig 13 ~ 15 are not meaningful for understanding general
> usage patters, but for three cities of China.  The authors need to
> describe the limitation of their analysis and method.
```

(R1-3) We agree with this comment. We discuss this in our new limitation section, and we acknowledge that some of the analysis results, e.g., speed estimation results, may be specific the particular population in those areas where data are available. However, we consider our algorithm design itself to be general enough, and the speed estimation method can be easily applied to similar datasets from other areas. For example, our methods are also applied to an additional dataset that was collected in the US. The results in our estimation match the ground truth data (GPS) well in that smaller scale dataset. Related changes were marked in the paper in Section 7 on Page 18.

```
> Lastly, I would like to see how the authors dealt with the massage
> smartphone usage data.
```

(R1-4) We use Python to analyze our dataset. More specifically, we used the Voronoi package from Scipy to construct voronoi maps with tower coordinates. We used another Python package called shapely to calculate various geometry values in our computation of distance lower bounds. All analysis are carried out on a single Cloudlab c8220 server with two 10-core 2.2GHz E5-2660 processors and 256GB memory. We added such details in Section 6 on Page 13.

```
>> RESPONSE TO REVIEWER 2 <<

>  ## 1. Motivation for how the results of this work can be used and their
>  generalization.
>
>  The authors present correlations between user movement speed and other
>  features such as number of applications used, traffic volume, idle time,
>  etc.
>  The authors motivate such findings with: "Understanding such
>  correlations, if any, could provide useful contextual information for
>  relevant and accurate app recommendation and ad delivery. For example, if
>  we find out hiking hobbyists use certain apps considerably more often,
>  then such apps may be more useful venues for ad delivery for equipment
>  makers for hiking activities."
>
>  However, from the findings presented in this work I only see that Fig16
>  supports the aforementioned motivation by providing insights about the
>  correlation between movement speed and app categories.
>  Providing stronger / additional arguments why all the other observations
>  are relevant would help the authors to increase the value of this work.
```

(R2-1) We appreciate the insightful suggestions. We have added additional arguments on the observations and their support for the conclusions in the paper. The changes can be found on page 2 in section 1.

```
>  Furthermore, as mentioned in the paper (however, only once I guess), the
>  traces used for this work cover cellular communication only.
>  As the authors have pointed out, part of the communication might be
>  handled over Wi-Fi.
>  In recent studies, authors have observed that over 60% of the time users
>  are connected to Wi-Fi [1] and that half of the traffic is typically
>  handled over Wi-Fi [2].
>
>  So the important question at this point is: how representative are the
>  results presented in this work given the fact that they consider cellular
>  app usage and traffic only?
```

Thank you for your comments. Please see response to meta-review comment 2a.

```
> ## 2. My second concern is about the performance in estimating movement
>  speed that has a direct influence on the results presented in this work.
>
>  To estimate intra-cell speed, the authors propose a novel algorithm that
>  is based on a set of assumptions such as that users move with a constant
>  speed (4.1) or with a "straight line trajectory" (4.3).
>  The resulting computation of the movement speed is therefore a direct
>  consequence of the aforementioned assumptions and the introduced
>  approach.
>  Therefore, the results presented in this work rely on the quality of this
>  computation.
>
>  The results presented in this work show a positive correlation (Fig10a)
>  between movement speed and traffic volume / sec. as well as a negative
>  correlation (Fig10b) between the idle time and speed.
>  The conclusion is that the faster a user passes a given cell, the more
>  bytes/s she will generate and the smaller the idle time intervals are.
>  So to make sure that these insights indeed cover a valid collection
>  between movement speed and other features, it is mandatory to show that
>  the aforementioned assumptions hold, in general, and the movement speed
>  estimation is accurate, to some extent, allowing to make conclusions from
>  the experiments.
>
>  Addressing this shortcoming might for instance include an evaluation of
>  the approach on a data set that contains GPS and cellular data.
>  There are several publicly available data sets that might be helpful at
>  this stage (reality mining, nokia, lifemap, etc.).
>  Alternatively, running a small custom study to verify the assumptions and
>  get quantitative evidence that movement speed estimations are accurate,
>  might also be an option.
>
>  Without showing that the novel approach presented in this work to compute
>  movement speed produces reliable estimates, it is at least possible that
>  to some extent the insights presented in this work result from the
>  inaccurate movement speed estimations.
```

We really appreciate your constructive comments. We indeed agree that we need to demonstrate that the speed estimations to be correct. To achieve this goal, following this comment, we used a dataset collected by Intel Placelab to perform additional evaluations of our speed estimation algorithm. This dataset contains both cell phone data and GPS traces. We used the GPS data as the ground truth for evaluating our speed estimation algorithm that utilizes cell phone data only. Our results show that even though the dataset was collected from a different country (US), our speed estimation algorithm worked well with same set of parameters as those used in our experiment section. We add a new evaluation section to show the results. The changes are available in Section 5 on Page 12. We hope that by showing that the estimation algorithm can be applied to different datasets, we are able to say with more confidence that the conclusions drawn from the city-level datasets are more reliable.

```
> ## Minor comments:
>  - Please try to put Figs and Listings on the same page as the text
>  which refers to them
>  - Inconsistent writing of PBE in Sec 4.1 and Sec 4.2
>  - Potential naming (variables) inconsistencies between Alg1 and Sec 4.2
>  - Fig10a y-axis label: is it not supposed to be "bytes/sec" as
>  explained in the corresponding section?
>  - Sec 5.3: how do you define "a data access". is it a single CDR in
>  the data set?
>  - What is the overall value / take-home message of Fig11 and Fig12? Is
>  it not better to have relative values if the number of instances differ?
>  - Fig14 and Fig15 should be a bit smaller to match the font size of the
>  text
```

Thank you for identifying these problems. We have corrected these problems accordingly.

```
>  Suggested References:
>  [1] Kyunghan Lee,Joohyun Lee,Yung Yi,Injong Rhee, and Song Chong. 2013.
>  Mobile Data Offloading: How Much can WiFi Deliver? IEEE/ACM Transactions
>  On Networking 21, 2 (2013), 536551.
>
>  [2] Paul Baumann and Silvia Santini. 2014. How the availability of Wi-Fi
>  connections influences the use of mobile devices. In Proceedings of the
>  2014 ACM International Joint Conference on Pervasive and Ubiquitous
>  Computing Adjunct Publication - UbiComp '14 Adjunct. (2014), 367372.
```

Please see response to meta-review comment 5. We have added all suggested references and discussed them in the related work section. These changes can be found in Section 2 on Page 4 of the paper.

```
>> RESPONSE TO REVIEWER 4 <<
```

```
>  the use case presented (that of targeted advertising) needs evidencing
>  or reframing. There are many ways to target user interests and it is not
>  obvious that an approach based on speed is appropriate. If there is no
>  evidence to support the use of speed in targeted adverts then I would
>  have preferred to have seen the approach based as a generic analysis with
>  a range of possible applications being suggested.
```

(R4-1) We appreciate the insightful suggestions. We agree that there are many ways to target user interests. However, our primary goals in this paper aim to study whether speed differences of users will have a significant impact on their usage patterns in a large scale population. Although other factors may also have an impact on usage patterns, we consider that they are out of the scope of this paper. We have added additional arguments on the observations and their support for the conclusions in the paper. The changes can be found on page 2 in section 1.

```
>  clearly the data captured is only part of a user's overall data
>  consumption. This point is made in the paper but should be made more
>  explicitly at the start.
```

Please see response to meta-review comment 1.

```
>  the length of the data trace (3 hours) seems short - it would be good
>  to explain why the authors feel this is an appropriate amount of data to
>  analyse.
```

(R4-3) Thank you for your comments. As we mentioned in the summary of responses, this is all the data that available to us. We have discussed this limitation of our dataset in the new limitation section. Even though our dataset is relatively short, the speed estimation method can be easily applied to similar datasets of any given time period. Similar procedures can be used to study usage patterns in those datasets without much difficulty. We have made changes to the paper to include this limitation in section 7 on page 17.

```
>  the speed estimation approach contains a very large number of
>  assumptions and there is no evidence presented that it actually works.
>  This is the biggest weakness in the paper - it *must* provide some
>  evidence of success. Even a simple study with 20 users where ground truth
>  and cell records were collected would be sufficient. Without that there
>  is simply not enough to convince the reader that the algorithm presented
>  works.
```

Please see our response to meta-review comment 3.

```
>  the analysis would be much stronger if it was backed up with some
>  evidence (e.g. a survey or a focus group or some observations studies)
>  that trued to explain the patterns seen.
```

Thank you for your comments. Actually, in a recent work [1] which also study correlation of user mobility and mobile data access, similar trends as shown in Figure 9b have been found. Another work on geospatial relation of the app usage [2] showed similar smartphone app usage distribution as shown in Table 1 and Figure 11.

Related references:

[1] Jie Yang, Yuanyuan Qiao, Xinyu Zhang, Haiyang He, Fang Liu, and Gang Cheng. 2015. Characterizing User Behavior in Mobile Internet. IEEE Transactions on Emerging Topics in Computing 3, 1 (2015), 95–106.

[2] M Zubair Shafiq, Lusheng Ji, Alex X Liu, Jeffrey Pang, and Jia Wang. 2012. Characterizing geospatial dynamics of application usage in a 3G cellular data network. In INFOCOM. 1341–1349.