Mining correlation of user speed and app usage behavior in massive mobile data traces

ABSTRACT

In this paper, we investigate the prediction of mobile app categories based on where the mobile user is, and what the mobile user is doing. A large dataset of 2G data via smartphone apps has been collected from hundreds of towers in a city during a three hour period in the evening (6pm to 9pm local time). We examine several useful features that may be correlated with app types, and combine them into a unified model for prediction.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; See http://acm.org/about/class/1998/ for the full list of ACM classifiers. This section is required.

Author Keywords

Authors' choice; of terms; separated; by semicolons; include commas, within terms only; required.

INTRODUCTION

Smartphone has seen a rapid growth during last decades. 74.9 percent of mobile subscribers use smartphone in the U.S in early 2015. Accompanied with the rapid growth of smartphone is the explosive growth of smartphone apps. Number of apps in both Google Play and Apple App Store exceed 1.5 million by July 2015. The time people spent on smartphone apps has achieved more than 30 hours monthly and seen a growth over 65 percent compared to 2013.

The increasing importance of smartphone apps attracts a large research body to study smartphone app usage behavior [30, 32]. Both temporal pattern (individual app usage history) and spatial pattern (context correlation) have been well studied. These studies not only help us understand how people using smartphone apps that is useful in re, but also enable applications such as smartphone app launching prediction [31] and customized smartphone app recommendation [need another reference]. However, the user mobility, which also plays an critical role in users' app usage, yet has not received well attention on their correlation with users' app usage behavior. The reason could be that user mobility are usually not directly available in cell-phone traces and are not very easy to acquire.

Previously, inference of user's mobility such as transportation mode are highly rely on additional hardware (e.g. GPS, sensors) or surveys. Both suffer from availability and scalability issues. [16] indicates there is great potential of using cell-phones to monitor users' mobility. Later, several paper have studied the problem of inferring users' trajectory [12, 9] or transportation mode [25, 4, 1] from various cell-phone traces (e.g. Call Detail Records, handover data). Compared to previous methods, such an approach does not require additional resources and have excellent coverage.

The location data conatined in most mobile phone traces are quite limited, usually only the cell phone tower ID with which it communicated. So the localization accuracy of these traces are very poor. As a result, only limited user mobility can be extracted from the data, i.e. approximated trajectory [20, 8, 29, 12, 9] or mobility motif [27, 6]. And the trajectory inferred from such data are in a quite coarse grained way. In our work, we use passing boundary events combined with distance lower bound estimation to overcome the above issues to robustly estimate the speed of each user.

Previous work on geospatial smartphone app usage mainly focus on spatial correlation of smartphone app usage volume [18, 32]. Limited work on correlation of user mobility and mobile phone app usage have been done. In this paper, by analyzing the data traffic collected at three cities in China, we reveal the correlation of user's speed and several aspects including data volume, access frequency, market share of apps in smartphone app usage.

Our main contribution is:

- Reveal correlation of user mobility and smartphone app usage pattern
- Improved user mobility inference to meet specific need of revealing correlations.

The rest of paper are organized as follows:...

DATA DESCRIPTION

The trace in our settings contains mobile data access history of a set of users. Each user has a set of records $R = r_1, r_2, ..., r_i, ...r_n$. Each mobile data access record r has the following fields:

$$\langle UserID, TowerLocation, TimeStamp, DataAccess, \dots \rangle$$

where

- $UserID$ is identifier of a user, a hashed value for anonymity
- $TowerLocation$ is the location (latitude and longitude) of cell phone tower with witch the user communicated, denoted by l

- *TimeStamp* is the time stamp of a mobile data access record, denoted by t
- *DataAccess* is the mobile data access of the app for this records, including app identifier and data volume, denoted by *DA*
- ... represents many other attributes not of focus here.

We assume the set of records R for each user have already been sorted by timestamp t.

The mobile data access patterns is the mobile phone usage bahavior of each individual. The patterns in our study include average mobile data access volume, average time intervals between consecutive mobile data access and various kinds of apps that generate these traffic.

Our approach takes two steps to solve the proposed problem. The first step is to estimate user speed s from the location l and time t information provided by the trace. The second step is to study the correlations between estimated user speed and various mobile data access patterns mentioned above.

Before we proceeding to the details of our approaches, we briefly introduce our dataset in this section. The dataset is mobile data access records provided by a cellular network operator in China. It was collected from three mid-size cities, including both urban and suburban area, during a three-hour period in the early evening (6pm - 9pm). 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 5199 cell towers in the area during the observation period. The number of unique users included in this dataset is 0.9 million. And the total active time of all user is more than 1 million hours.

Compared to similar data such as Call Detail Records (CDR), our dataset share same trends in several key statistics but is more dense temporally. For example, our dataset also have highly skewed distribution on number of records per user and time intervals between consecutive records as shown in Fig. 1. Finer grained in temporal means better potential to infer user mobility even when the trip length is very short. Actually it is the case that most user only travelled a very short trip in terms of number of visited towers according to Fig. 1.

An example user's trace

To show a clear view of our trace and serve as a running example, we have selected a random user from the dataset and show his mobile data access trace in fig. 2. The figure contains two parts. The top part is a map shown the towers that were visited by the user. We use markers to show tower locations and arrowed lines to show the sequence of visiting. The bottom part of the figure shows the time line of the user's data access records with pulses, each pulse represent a mobile data access record and its position on the time line shows its relative time. We also show to which tower the user is communicated for each mobile data access with tower labels above the pulses. So for this particular user, he communicate with tower A for a quite long time, and shortly connected to tower B then swithed to tower C. After a while, the user was found in tower D's coverage area. After a while, he connected to tower E for a very short time, and then switch back to tower D.

ESTIMATE SPEED

In this section, we introduce our approach to estimate speed from the given dataset. We first define passing boundary events and use it to distill more accurate location estimates from the mobile data access trace. Then we develop a mechanism based on the distance lower bound to filter out speed estimates with low confidences. We will discuss them in more details below. After we extract speed estimates from the trace with high confidence, we will reveal their correlation with the mobile data access patterns of users in the next section.

To estimate user speed, the main challenge we deal with is the redundant but coarse grained location estimations, i.e. multiple records have the same location estimation which is the coordinate of the towers with which users communicate. The location estimation accuracy is the whole coverage area of towers in this case. Consider that most users only have very short trips as shown in Fig. 1, this level of accuracy is critically low for the purpose of speed estimation. The unstable nature of wireless communication aggravates the problem, e.g. even a static user may have communicated with more than one tower (cell oscillation).

Structure overview

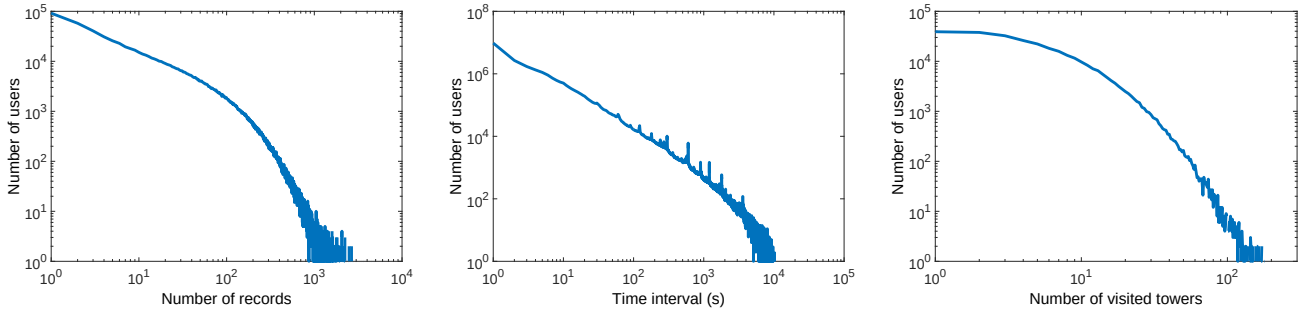
Fig. 3 shows the structural overview of our approach. The raw data parser first gather data access records and tower locations from the mobile data access trace. Then the passing boundary events (PBE) are extracted from the data access records. Based on these passing boundary events, traveled distances and durations are estimated. With tower locations, a Voronoi diagram is built and Voronoi edges are collected to be used to approximate the communication coverage boundaries among towers. And distance lower bounds for each user to pass a tower's coverage area are estimated based on the approximated boundaries. With distance estimations, distance lower bounds and duration estimates, the system can estimate the user's speed and filter out inaccurate speed estimates with criterion based distance lower bounds and duration estimates. For some records that do not have sufficient location information to accurately estimate the user's speed, the system will also compensate their speed estimation. We will discuss each component in more details in the following sections.

Passing boundary event

We define passing boundary events and shows how to distill finer grained location information from our trace with it. For arbitrary two consecutive records r_i and r_j from the sorted mobile data access records of a user, if they have different related tower location, we define them as a passing boundary event (PBE), denoted by $P_{i,j}$:

$$P_{i,j} = (r_i, r_j), where l_i \neq l_j$$

For example, in fig. 4, a user moved from tower A's coverage area to tower B's coverage area. The switch from tower A to tower B should happen in the overlapped area which is shown with shadows. Although we don't know exactly when the



(a) Number of records per user (b) Time interval between consecutive records (c) Number of visited towers

Figure 1. dataset statistics

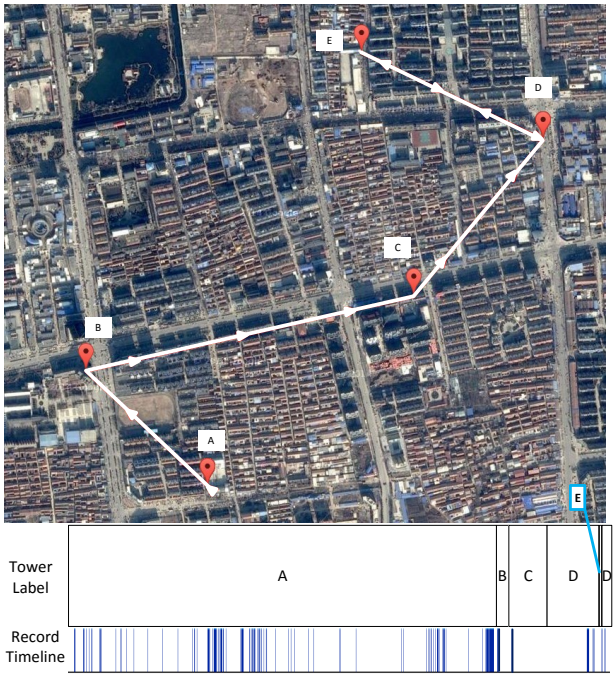


Figure 2. A typical data access session of a user.

switch happened, but the time should be bounded by the time of last record with tower A and the time of first record with tower B. So for a PBE, the time of the event is a time interval defined by the time of the two related records (t_i, t_j). Note that since the mobile data access of a user are not continuous temporally, i.e. user may communicate with towers that are far away from each other in two consecutive records. So for the boundary related to the event, if the two towers are adjacent with each other, i.e. their communication coverage overlap, then the boundary of the even is the overlapped boundary area and we refer to it as a real boundary. Otherwise, the boundary of the event is not exist and we say the PBE has a virtual boundary.

PBEs with real boundaries have better location estimation accuracy than a single record. The location estimation accuracy of an arbitrary record r_i is the whole coverage area of the related tower l_i. For a PBE $P_{i,j}$ with a real boundary, the location accuracy of the boundary is the overlapped boundary area of the two related towers. Since the boundary area is only a sub-area of whole coverage areas of both towers. The location accuracy of PBE $P_{i,j}$ is better than location accuracy of both related records r_i and r_j. By combining location information in two consecutive records that have different location estimates, we can achieve a better location estimation accuracy.

Note that the better location accuracy only stands for PBEs with real boundaries, not PBEs with virtual boundaries. Since there are no overlaps for two towers of a virtual boundary, it's hard to make any assumptions of the size of boundary area compared to the size of coverage area of related towers. The possible boundary area of a virtual boundary could be much larger than the coverage area of both towers. In some cases, it may including the communication areas of multiple towers.

Due to the redundancy in the location estimations, we rearrange our data by aggregating mobile data access records between two consecutive PBEs as a single unit called aggregate mobile data access record. All records belonging to the same aggregate record are communicate with the same tower. An aggregate mobile data access record is the minimum unit when we estimate the speed, that is, all records belonged to the same aggregate record will have the same speed estimate with our algorithm. Each aggregate record has one, e.g. the first and the last session, or two PBEs related to it.

distance lower bound estimation

Even with the finer grained location information in PBEs, it is still hard to accurately estimate the traveled distance of a user that pass the coverage area of a given tower. To see this, we show an example in fig. 5. The whole area is divided into three coverage area of three towers A, B and C. Solid lines represent real user trajectories while dashed lines represent boundaries of towers. Both user 1 and user 2 psss the three towers in the same order $A-B-C$. The difference of the trip length is lost in the records due to the limited location estimation accuracy. Although we can calculate an estimated traveled distance to pass a tower's coverage area with existing approaches [20, 8, 29, 12, 9, 4, 1] easily. The variety of trajectories that can lead to same tower visiting orders makes it impossible to come

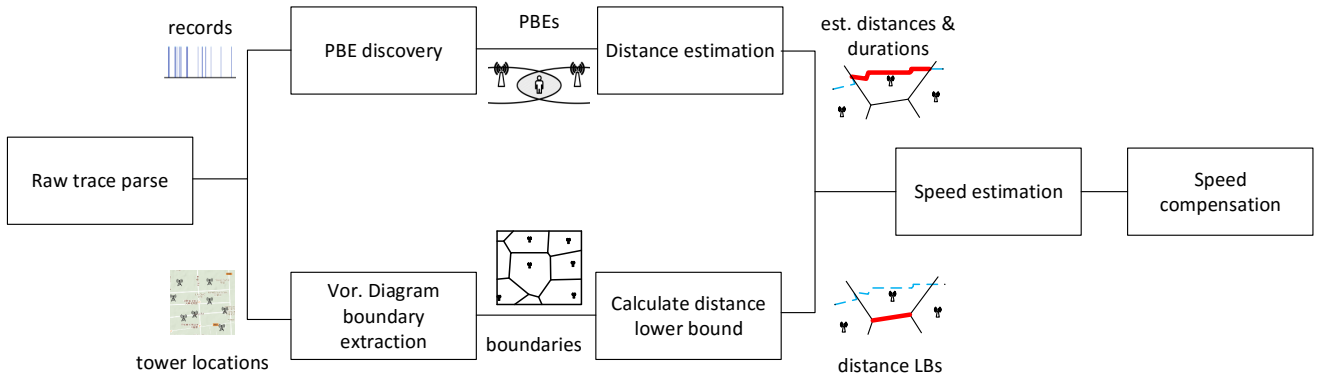


Figure 3. Speed estimation system overview

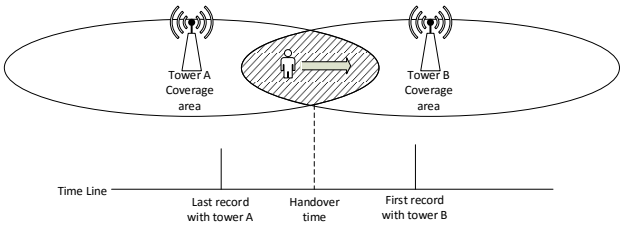


Figure 4. Passing boundary event.

Figure 5. Common cases where a distance estimate will fail

up with a single traveled distance estimations that is accurate for all the trajectories in such a case. So we can only have a distance estimate with low confidence.

The goal of our approach is to filter out distance estimates with low confidence as shown in case fig. 5. The main challenge to achieve this is to decide the confidence level of each distance estimates we have. Since the low confidence in estimation accuracy mainly caused by possible distance difference for various trajectories, a measure of difference in possible trajectory length can help us estimate the confidence level. We define the measure as following:

[define goes here]

In large scale dataset like our traces, it is very costly to , but we also want to know what is the minimum distance required to travel from one boundary area to the other boundary area. We call this distance as the distance lower bound for a pair of boundaries.

To easily calculate the distance lower bound we first simplify the tower coverage model by assuming the cell phones only communicate with the nearest tower. With this assumption, we can use equirectangular projection to reduce the tower coverage map to a Voronoi diagram with each tower's location as Voronoi points. Fig. 6 shows an example of the Voronoi diagram containing five towers. Each region in the Voronoi diagram represents the coverage area of the related tower. Boundaries of regions in Voronoi diagram represent the overlapped boundaries area of towers. Then the shortest distance required to travel from one boundary to another boundary can be simplified as the shortest distance of two Voronoi boundaries.

If we have an estimated distance that is much larger than the distance lower bound, then it is likely that the two boundaries could have paths of various distances. So using one distance estimation to represent the distance of all possible path may not be accurate. On the contrary, if the estimated distance is very close to the distance lower bound, then the estimated distance should be able to represent the distance of most paths between two boundaries. The distance lower bounds can also help to eliminate the problem of false passing boundary events. Since the user keeps passing the same boundary, the distance lower bound for such scenarios is always 0.

For the distance estimates, other than estimating with the trajectory that has the maximum likelihood with visited tower sequence, which require the knowledge of underlying road network. We use a very simple scheme that only require tower coordinates to estimate distances of two boundaries. Suppose one PBE is from l_i to l_j, and the other one is from tower

Figure 6. Voronoi diagram using Voronoi region to represent communication coverage of each tower

l_j to l_k. We first calculate straight line distance $d(l_i, l_j)$ and $d(l_j, l_k)$ by using tower's coordinates. Since the boundaries are perpendicular bisector of straight lines connecting towers, then the travel distance can be estimated by $\frac{d(l_i,l_j)+d(l_j,l_k)}{2}$.

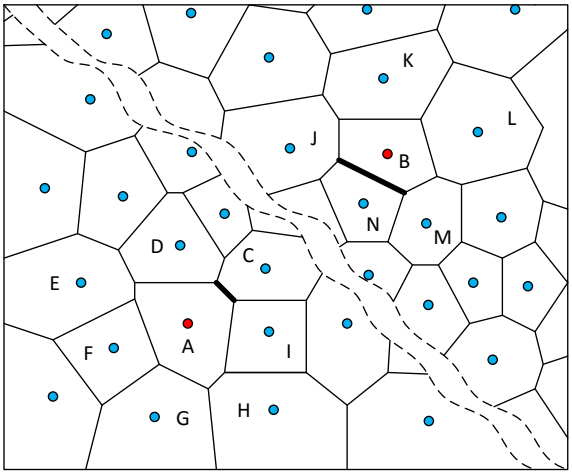


Figure 7. Deal with virtual boundaries

Remember that a real boundary means overlapped communication coverage areas while a virtual boundary means there is no overlap between the communication coverage areas of the towers. Different from real boundaries that are treated as a line which does not have distance for itself in distance estimations, virtual boundaries actually have distance estimates due to the fact that user have passed the coverage area of several towers. But we do not have the information of which boundaries does the user pass through for a virtual boundary. So to calculate the distance of a virtual boundary that connect tower l_i to l_j, we use the shortest distance of all possible boundary pairs of l_i and l_j. For example in fig. 7, suppose two consecutive records r_i is the user's last record in tower A, r_j is the user's first record in tower B. Since tower A and tower B does not share a boundary, so the related PBE has a virtual boundary. To calculate the distance, we calculate the distance from each boundary of tower A to each boundary of tower B, and use the distance of the shortest distance of all boundary pairs. In this example, the distance between boundary (A, C) and boundary (B, N) is used.

speed estimation

To infer user's speed during each aggregate record, the PBEs with real boundaries are used as reference points since they have better location accuracy as mentioned above. For aggregate records that have PBEs with virtual boundaries, we merge them with adjacent aggregate records if there is any. The distance estimates and distance lower bound of the merged record is the sum of distance estimates and distance lower bound of both records and the virtual boundary between them. Note that when we sum up distance lower bounds, the result is still the minimum distance required to reach one real boundary from the other one that passing through virtual boundaries in between following visited tower sequence in the trace. We denote real boundaries by b. Suppose the two PBEs are $P_{i,j}$ with real boundary bi, j and $P_{k,l}$ with real boundary $b_{k,l}$. Then the distance estimate and the distance lower bound between them are denoted by $d_{est}(b_{i,j}, b_{k,l})$ and $d_{lb}(b_{i,j}, b_{k,l})$.

For the duration between two reference points (PBEs with real boundaries), we can simply use the time difference of the PBEs. Note that for each PBE, the time related to it is not a time point but a time interval, We will have two durations, a tight duration which is the time difference of the first and last record belonging to the aggregate record between two reference points and a loose duration which is the time difference of two records that does not belong to the aggregate record between the reference points. For example, for two PBEs $P_{i,j}$ and $P_{k,l}$ with time interval (t_i, t_j) and (t_k, t_l) respectively. Suppose $P_{i,j}$ happens before $P_{k,l}$, then $t_i \leq t_j \leq t_k \leq t_l$. We denote the tight duration by $\Delta t_{tight} = t_k - t_j$ and loose duration by $\Delta t_{loose} = t_l - t_i$. So the estimated duration of the aggregate record between $P_{i,j}$ and Pk, l can be calculated by $frac\Delta t_{tight} + \Delta t_{loose}2$, we denote it by Δt_{est}.

Large differences between d_{est} and d_{lb} or between Δt_{tight} and Δt_{loose} indicate inaccuracy in distance estimate or duration estimate respectively. So before we estimate the speed, we set up a set of criterion to filter out these records with possible inaccurate estimates:

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

$$\Delta t_{ratio} = \frac{\Delta t_{tight}}{\Delta t_{loose}} \tag{2}$$

By setting a threshold for both criterion, we can filter out speed estimates that are not accurate enough. Although we can filter out more possible inaccurate speed estimates with

very strict threshold in both criterion, we may end up with limited number of records that have qualified speed estimates.

For aggregate records which meet both criterion, we calculate their speed estimates s as following:

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

speed compensation

Due to the false passing boundary event mentioned in left part of fig. 5, a large number of records will have a distance lower bound with a value of 0. And they will eventually been filtered by out distance criterion so that they will not receive any speed estimates. Since these aggregate records usually have very short duration due to the nature of how they are generated. One way to estimate the speed for such records are based on the assumption that a user's speed does not change dramatically in a very short time period. So for a aggregate record with false passing boundary event, if there is an aggregate record that are happened very close to them and have a qualified speed estimates, then we will use its speed estimates as the speed estimates for the record with false passing boundary event.

FINDINGS - REVEAL CORRELATION OF SPEED AND MOBILE DATA ACCESS

With the speed estimates, we show and explain our findings on correlations of user mobility and mobile data access patterns in this session. We start with the correlation of speed and average mobile data access volume. Then we revealed the relation of speed and average gaps between consecutive mobile data access. In the last, we show the correlation between speed and the apps that are used to generate mobile data traffic.

experiment settings

Our algorithm can only estimate speed when a user has visited more than 3 towers, so only 13 million records out of 58 million records have a speed estimate. In our experiments, to balance the accuracy of speed estimates and the volume 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} at 0.6. After the filtering, we have around 1 million records out of total 13 million records that meet both criterion. Fig. 8 shows the histogram of speed estimates.

In the following experiments, we only show results in the speed range of 0 km/h to 100 km/h, since there is very few records have speed estimates above 100 km/h to gain any meaningful insights.

correlation of speed and mobile data access volume

Fig. **??** shows the results of the correlation of speed and average mobile data access volume per user per second. We show data from all three cities combined and each city respectively. The figure shows a clear trend that users are more active in access mobile data as the speed increases. A user with speed estimates of 80-100 km/h could reach a average data volume of 6 times of a low speed user. And this trend holds for all the cities. Note that this does not suggest lower speed users do not access online contents less frequently since they have more

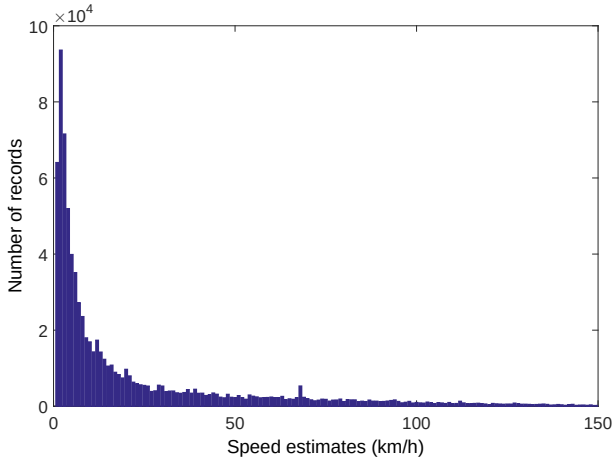


Figure 8. Histogram of speed estimates.

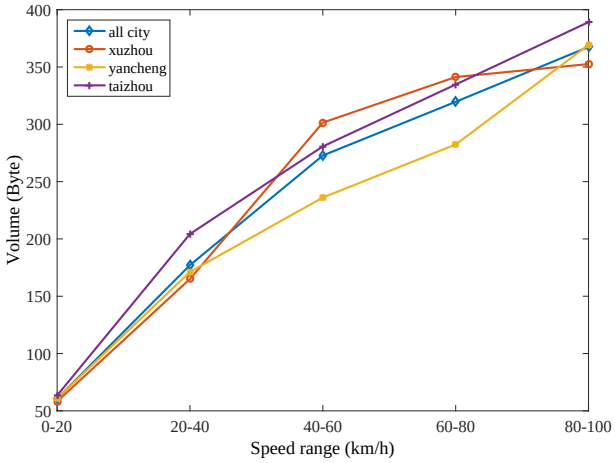


Figure 9. Correlation of access volume and user speed.

sources to reach online contents than high speed users, i.e. WIFI, Ethernet. Previous work [32] reaches similar results while using number of towers visited by user as indicator of user mobility.

speed and mobile data access frequency

Fig. 10 shows the results of the correlation of speed and time intervals between consecutive mobile data access records. The decrease in the time interval as speed increases suggest that high speed user access mobile data more frequently than low speed users. A user with speed estimate of 80-100 km/h access mobile data almost 2 times more frequently than a user with speed estimates 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 the city 'taizhou', which may caused by the lacking of available amount of data.

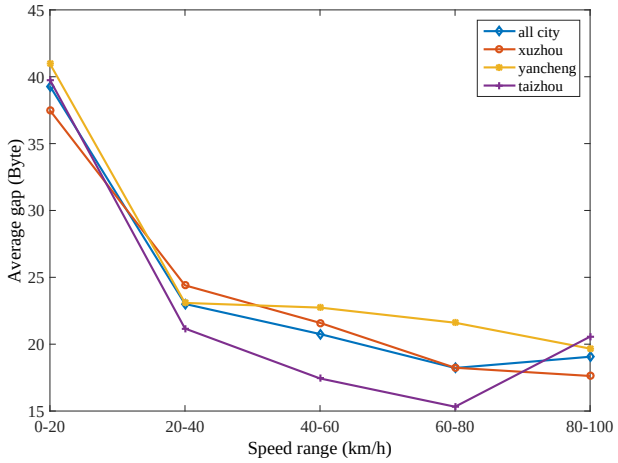


Figure 10. Correlation of time interval between consecutive data access and user speed.

App category	number of apps	data 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
Online 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
Browser & Download	558	353.5
Finance	25	0.7
Security	22	5.2
Other1	237	74.7
Other2	7	21.1

Table 1. App categories

speed and mobile data access pattern

App category information

According to the mobile service provider, all apps in our trace are grouped into 19 categories. We showed the name, the number of apps and total volume in our data trace of each category in Table 1.

In this section we study how the impact of various app categories change for different speed range. The impact is defined as the mobile data access of one category versus all categories. As we shown in Table 1 that the volume of data for each category is not even, among all 19 categories, we only interested in the app categories that contribute most to the total mobile data access volume. Note that apps in other1 and other2 are those can hardly classify to any other 17 categories. Since they do not share common properties, thus we do not take them into consideration. We select the top 8 app categories with most impact on mobile data access and show their impact changes in fig. 11.

Among the top 8 categories, microblog, navigation, music shows an clear trend of increasing as 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 shows a trend of decreasing as speed increases. The reason could be the users are cost sensitive and dose not want to spend mobile data on large app downloading or video streaming. Brower & downloading and reading shows a quite stable impact that does not changes a lot as speed increases.

RELATE WORK

Smartphone app usage

Smartphone app usage have draw attention of a large research body. To study the smartphone app usage behavior of large group of users, previous work usually analyze mobile data traces generated by smartphone apps. [30] comprehensively shows the aggregated spatial and temporal prevalence, locality and correlation of smartphone apps at a national scale by analyzing mobile data generated by smartphone apps. [32] studied the smartphone app usage patterns of various mobile user groups. Although correlation of user mobility and data volume generated by apps have been briefly studied in this paper, limited results have been presented compared to our work.

User mobility

Using GPS [14, 24, 33, 2, 22, 17, 28, 15] and embedded sensors [14, 26, 19, 10, 23, 15, 7] of smartphones to inferring user mobility such as transportation mode have been extensively studied. Most of these works form the problem as a classification problem. Common challenges includes data segementation [14, 24, 33, 2] where the data are segmented so that each segment only contains one transportation mode, and feature selection [33, 2, 26, 22] that proper features enable the classifiers separate various similar classes, i.e. car and bus.

Although GPS and sensors are well suited for user mobility inference and can infer user mobility such as transportation mode where even the speeds of various modes are the same. They require additional energy cost and does not scale well. [16] revealed the great potential of using cell-phone data traces such as Call Detail Records (CDRs) for user mobility inference. There is a large research body in the literature that studied methods to inferring user's trajectory [20, 8, 29, 12, 9, 4, 1] or mobility motif [27, 6]. [12, 9, 4] infer user trajectory from cell-phone traces based on how likely a specific route can lead to similar tower access sequences stored in the data traces. Besides, there is a great uncertain about a user's location when the user is not active. So previous work also study several different interpolation methods [8, 5] to fill in the uncertain location when the user is not active.

[25] does not try to estimate a user's exact trajectory from smartphone traces, instead it classify a user's transportation

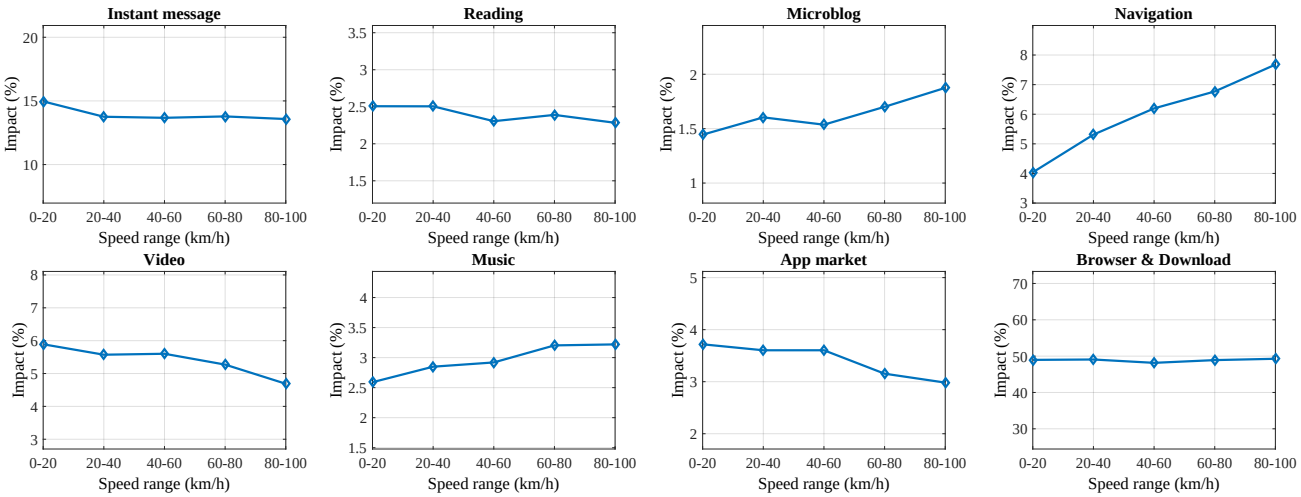


Figure 11. Correlation of access pattern and user speed.

mode by clustering on travel time distribution. [] proposed an approach that can deal with common zig-zag problems in inferring user mobility from smartphone traces. Earlier works also use signal strength received at mobile phone to estimate user's speed [21], but this approach suffers with the same problem as using GPS or sensor.

Geospatial app usage

Previous works also studied relations of human mobility and social networks. [3] found that the short-ranged travels are periodic and not related to the social network structure much, while long-distance travels are heavily related to the social network. Based on these findings, a model is proposed to predict dynamics of future human movement with high accuracy. Follow up work such as [13] studied a similar problem with a different dataset. [18, 32] studied the geospatial relation of app usage volume. Their works mostly studied the spatial correlation of smartphone usage and user mobility's impact on app usage is still a missing piece of these works. [11] studies how proximity, location and individual differences (e.g., personality) can effect user's mobile data usage.

CONCLUSIONS

REFERENCES

1. S Bekhor and I Blum Shem-Tov. 2015. Investigation of travel patterns using passive cellular phone data. *Journal of Location Based Services* (2015), 1–20.
2. 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.
3. Eunjoon Cho, Seth A Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1082–1090.
4. John Doyle, Peter Hung, Damien Kelly, Seán McLoone, and Ronan Farrell. 2011. Utilising mobile phone billing records for travel mode discovery. (2011).
5. Michal Ficek and Lukas Kencl. 2012. Inter-call mobility model: A spatio-temporal refinement of call data records using a Gaussian mixture model. In *INFOCOM, 2012 Proceedings IEEE*. IEEE, 469–477.
6. Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2012. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*. ACM, 3.
7. Samuli Hemminki, Petteri Nurmi, and Sasu Tarkoma. 2013. Accelerometer-based Transportation Mode Detection on Smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems (SenSys '13)*. ACM, New York, NY, USA, Article 13, 14 pages. DOI: `http://dx.doi.org/10.1145/2517351.2517367`
8. 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.
9. 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 *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. ACM, 2.
10. Vincenzo Manzoni, Diego Maniloff, Kristian Kloeckl, and Carlo Ratti. 2010. Transportation mode identification and real-time CO2 emission estimation using

smartphones. *SENSEable City Lab, Massachusetts Institute of Technology, nd* (2010).

11. Lei Meng, Shu Liu, and Aaron D Striegel. 2014. Analyzing the impact of proximity, location, and personality on smartphone usage. In *Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on*. IEEE, 293–298.

12. Mathï£¡ Young Mosny. 2006. *Path Estimation Using Cellular Handover*. Ph.D. Dissertation. Princeton University.

13. Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. 2011. An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. 570–573.

14. H. Ohashi, T. Akiyama, M. Yamamoto, and A. Sato. 2014. Automatic trip-separation method using sensor data continuously collected by smartphone. In *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*. 2984–2990. DOI: http://dx.doi.org/10.1109/ITSC.2014.6958169

15. 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, Article 13 (March 2010), 27 pages. DOI: http://dx.doi.org/10.1145/1689239.1689243

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

17. J. Ryder, B. Longstaff, S. Reddy, and D. 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. 927–931. DOI: http://dx.doi.org/10.1109/CSE.2009.312

18. 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, 2012 Proceedings IEEE*. IEEE, 1341–1349.

19. 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.

20. 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.

21. Timothy Sohn, Alex Varshavsky, Anthony LaMarca, Mike Y Chen, Tanzeem Choudhury, Ian Smith, Sunny Consolvo, Jeffrey Hightower, William G Griswold, and Eyal De Lara. 2006. Mobility detection using everyday gsm traces. In *UbiComp 2006: Ubiquitous Computing*. Springer, 212–224.

22. Leon Stenneth, Ouri Wolfson, Philip S Yu, and Bo Xu. 2011. Transportation mode detection using mobile phones and GIS information. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 54–63.

23. C. Tacconi, S. Mellone, and L. Chiari. 2011. Smartphone-based applications for investigating falls and mobility. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011 5th International Conference on*. 258–261.

24. K. Waga, A. Tabarcea, Minjie Chen, and P. Franti. 2012. Detecting movement type by route segmentation and classification. In *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2012 8th International Conference on*. 508–513. DOI:http://dx.doi.org/10.4108/icst.collaboratecom.2012.250450

25. Huayong Wang, Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. 2010a. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*. IEEE, 318–323.

26. Shuangquan Wang, Canfeng Chen, and Jian Ma. 2010b. Accelerometer based transportation mode recognition on mobile phones. In *2010 Asia-Pacific Conference on Wearable Computing Systems*. IEEE, 44–46.

27. 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*.

28. P. Widhalm, P. Nitsche, and N. BrÃd'ndie. 2012. Transport mode detection with realistic Smartphone sensor data. In *Pattern Recognition (ICPR), 2012 21st International Conference on*. 573–576.

29. 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.

30. Qiang Xu, Jeffrey Erman, Alexandre Gerber, Zhuoqing Mao, Jeffrey Pang, and Shobha Venkataraman. 2011. Identifying diverse usage behaviors of smartphone apps. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 329–344.

31. Tingxin Yan, David Chu, Deepak Ganesan, Aman Kansal, and Jie Liu. 2012. Fast app launching for mobile devices using predictive user context. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*. ACM, 113–126.

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

33. 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 (TWEB)* 4, 1 (2010), 1.