Analysis of Mobility Patterns for Urban Taxi Cabs

Mohammad Asadul Hoque[1]; Xiaoyan Hong[2]; Brandon Dixon[3]
Department of Computer Science
The University of Alabama
Tuscaloosa, USA
{mhoque[1]; hxy[2]; dixon[3]}@cs.ua.edu

***Abstract*—This paper analyzes urban taxi mobility traces obtained from San Francisco Yellow cabs. The paper presents a rigorous analysis of taxi mobility pattern with the instantaneous velocity profile, spatio-temporal distribution, connectivity of vehicle communications, clustering, hotspots and other characteristics like trip duration and empty cruise interval. The empirical data analyses presented here can be a helpful resource for wireless researchers, government organizations, taxi companies and even for the drivers or passengers. While wireless researchers can estimate the capabilities and constraints of vehicular communication from connectivity and mobility patterns, government can plan and work on issues related to implementing proper DSRC infrastructure. Finally, taxi companies and drivers can benefit from maximizing the trip revenue and minimize empty cruise time though balanced loaddistribution and awareness of the hotspots.** *(Abstract)*

Keywords-Hotspots; Cruise time; connectivity; traffic trace; Clustering, Partitioning, V2V, Hotspot.

I. INTRODUCTION

The GIS based computer aided taxi dispatching (CAD) systems provide an easy way to track the movement of each individual taxi and monitor the occupancy status of the vehicle. In order to distribute the load fairly among the fleet, it is nevertheless important for the taxi companies to have a prior idea about the demand and availability statistics based on historical data that can be generated from the archived GPS trace records in their systems. On the other hand, to maximize daily trip revenue and minimize empty cruise time, it is also necessary for the drivers to have a sound idea about the geographical distribution of taxi hotspots for passenger pickup and drop off which varies along time.

More important, the historical archived data of mobility traces can provide significant information, such as geographical distribution and time varying density of the road traffic, for helping vehicle communications, for implementing Intelligent Transportation Systems applications, and for planning of deploying DSRC infrastructure. Recent research have shown studies on many interesting facts related to Vehicular Ad hoc Network (VANET) like urban mobility models, vehicle-to-vehicle (V2V) connectivity etc.

A remarkable initiative of San Francisco Exploratorium [4] is the Cabspotting project [5], which is intended as a living framework to use the activity of commercial cabs to explore the economic, social, political and cultural issues that are revealed by the realistic GPS traces. In this paper, we present our analysis on the traces available through this project provided by San Francisco Yellow Cabs[3]. Our analysis dealt with 536 cabs generating over 10 million mobility traces over a period of one month. Our results show new interesting factors about taxi cab mobility, passenger data and communication potentials. Here our analysis of taxi mobility pattern emphasizes the following characteristics:

i) Instantaneous velocity profile
ii) Spatio-temporal distribution of cabs
iii) Frequency distribution of pickup and drop off
iv) Identification of hotspots
v) Trip duration and empty cruise interval
vi) V2V connectivity
vii) Network partitions and Clustering

The subsequent sections are organized as follows: We discuss related works in Section II, followed by our analysis model and data collection methodology in Section III. Sections IV and V presents the results for a single cab and monthly averages for whole fleet respectively. Section VI provides a detail analysis on vehicle connectivity as well as clustering of the mobile nodes. Finally, we conclude in Section VII.

II. RELATED WORK

Several interesting works related to taxi mobility patterns has been addressed by the researchers. Most of these works are based on analyzing GPS traces from different taxi cab companies to explore hidden characteristics of urban mobility models. Some of these researchers tend to reveal new mobility models while others focus on clustering and hot spot identification.

Piorkowski*et. al*[9] utilized the Cabspotting data archived over a month to propose a parsimonious mobility model called Heterogeneous Random Walk (HRW) which captures some of the important mobility characteristics observed from the macroscopic level. A key feature of the model is that nodes follow independent and statistically equivalent mobility patterns, despite the presence of long-term clusters. They also evaluate the predictive power of the HRW model in the context of epidemic dissemination, which is one of the most prominent paradigms for routing in DTNs. Their work motivates the vehicular networking community to deeply investigate the taxi mobility traces for further research.

Shin et al [8] used real-life location tracking data collected from the Taxi Telematics system developed in Jeju, Korea. Their analysis aimed at obtaining meaningful moving patterns of taxi cabs .They have extracted some interesting statistical factors such as taxi's driving type, driving time, driving area, pickup rate etc. Lee et. al [7] analyzed a pick-up pattern of taxi

service in the same geographical area aiming at clustering the pickup and drop off locations to develop a location recommendation service for empty taxis. The same author in another paper [9] analyzed both spatial and temporal statistics of taxi's waiting spots from the movement history. These works provide an insight to the possible dimensions of utilizing location tracking data for the purpose of taxi industry.

III. System Model and Data Collection

The Cabspotting project tracks San Francisco's taxi cabs as they travel throughout the Bay Area. The data is transmitted from each cab to a central receiving station once in every minute, and then delivered in real-time to dispatch computers via a central server. This system broadcasts the cab call number, location and whether the cab currently has a fare. The cab locations are not stored by Yellow Cab, but only used in real-time to aid dispatch. Cabspotting server communicates to the Yellow Cab server and stores the data in a database, encoding the call number for privacy. The patterns traced by each cab create a living and always-changing map of city life. This project is intended for researchers to explore these issues in the form of a small experiment, investigation or observation. One of the most important component of this project is the API [10] that allows real time tracking information of individual cabs. Two other mentionable applications belonging to this project is the CabTracker [11] which averages the last four hours of cab routes into a map and the Time Lapse[12]which reveals time-varying patterns such as rush hour, traffic jams, holidays and unusual events.

A. Trace Record

Each mobility trace record contains the following fields:

1) Latitude & Longitude: Two floating point values of the current GPS position of the cab.

2) Occupancy status: A binary value indicating the passenger occupany status. A value of 0 indicates that the cab is free while 1 means hired by passenger.

3) Timestamp: Unix timestamp of the trace reception time.

B. Accumulation of Trace Records

Using the API we accumulated real time traces of these cabs over a time frame of more than 24 hours starting from July 17, 2011 11:01:09 PM to July 18, 2011 11:57:08 PM. A total of 2063trace records were captured within this time frame.

We also collected previously archived data for a period one month from CRAWDAD [13] that was acquired though the same procedure. The archived records summed up to a total of more than 10 million traces organized in individual ascii files for each of the 536 licensed yellow cabs. These trace files were simulated using our own developed application. We analyzed the traces both from the perspective of a single cab as well as from the perspective of the whole fleet.

C. Calculation of Geographical Distance

Previous work with GPS trace data and distances mostly considered Euclidian distance between two points. However, this calculation completely ignores the fact that the earth is round yielding incorrect results. The difference between Euclidian distance and a correct approach can be described in Figure 1. According to the Euclidian distance, the distance between two points P_1 and P_2 would be equal to the cord P_1P_2, whereas the actual distance would be along the circular arc.

Figure 1. Euclidian distance vs. actual geographical distance

In our work we investigated two algorithms, namely, the Spherical Law of Cosines and Equi-rectangular approximation, in calculating a geographical distance between two trace locations. Our implementations and usage of the two schemes suggests that, for more accurate precision level, the spherical cosine is better than the Equi-rectangular approximation. But for faster system performance the latter is preferred. In our mathematical analysis, we used the latter in case of averaging one month's data for all the cabs, which contained over 10 million records. While working with a single cab over 24 hour time span we used Spherical Cosine Law to get an accuracy level of less than one meter. Below we mentioned the mathematical equations for both the approaches.

1) Spherical Law of Cosines:

$$d = cos^{-1}(sin(lat_1).sin(lat_2) + cos(lat_1).cos(lat_2).cos(long_2 - long_1)).R$$

2) Equirectangular approximation

$$x = \Delta lon.\cos(lat)$$
$$y = \Delta lat$$
$$d = R.\sqrt{x^2 + y^2}$$

Here, R is the radius of earth (6371 km).

IV. Mobility Analysis for a Single Taxi

A. Instantaneous Velocity

Figure 2 shows the velocity profile of a single cab within a day. Vertical axis shows the calculated instantaneous speed of the cab in km/hour. Two axis along the horizontal plane denote latitude and longitude. The figure demonstrates some logical findings from the urban traffic perspective.The average speed in downtown area is calculated to be less than 40 km/h or approximately 25 mile/hour. On the other hand, average speed on the freeways is above 100 km/hour or more than 65 mile/hour. The averages speed of the taxi cab over the whole day calculated from the trace records was 43.81 km/h.

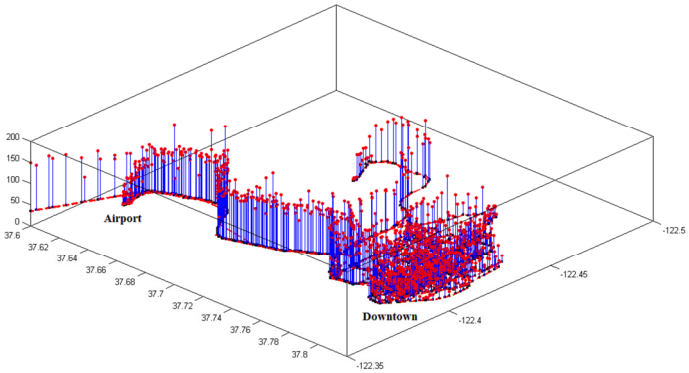
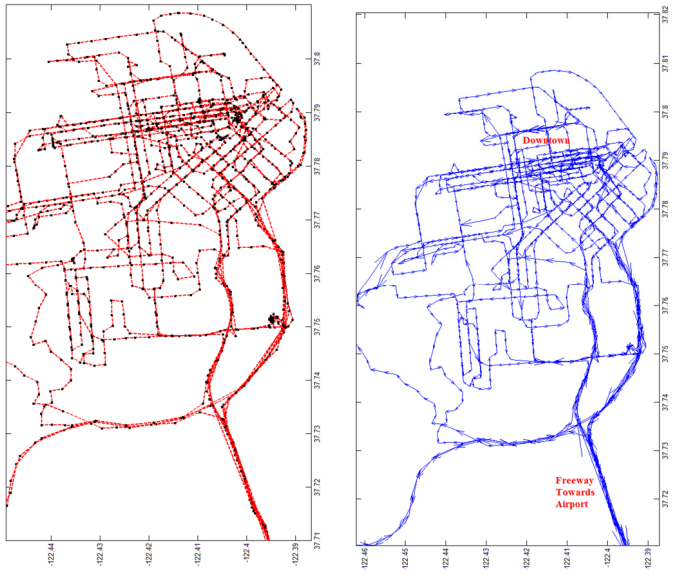


Figure 2. Velocity Profile of a single taxi cab in one day

B. *Trace Locations and Mobility Pattern*

Figure 3(a) below shows the geographical locations of trace points where we calculated instantaneous velocity and moving directions of the taxi cab within a time frame of one day. Here we only show a partial plot near the San Francisco downtown area. The black dots denote the location points where trace records were generated by the taxi cab and received by the GPS satellite. The average distance between two trace points were found to be 253.27 meter which is close enough to derive the direction of the movement. Within the downtown area the gap is mostly less than 100 meter due to slow traffic, whereas in the freeways these trace points are more than 1 km apart. The maximum distance found between two trace points was 1.79 km. The average time gap between two trace records wascalculated about 43.34 sec. Figure 3(b) describes the instantaneous velocity in different trace points.



(a) (b)

Figure 3. (a) Spatial Distribution of trace locations for a single cab over 24 hours (b) Instantaneous direction of mobility and velocity

V. STATISTICS FOR THE FLEET

We have analyzed archived data for a duration of one month containing traces of all 536 cabs. Some of the key findings are mentioned in the following subsections.

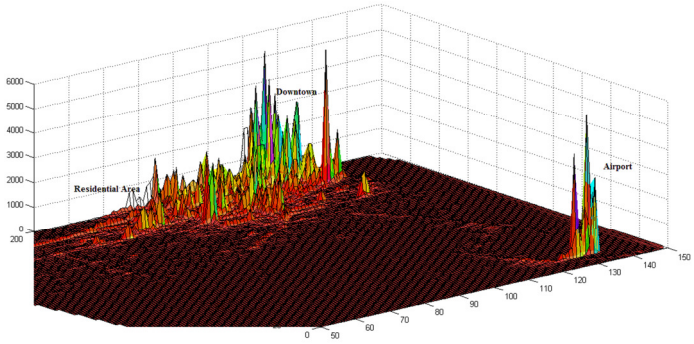


Figure 4. Frequency Distribution of passenger pickup and drop off locations

A. *Passenger pickup and drop off locations*

Figure 4 describes the frequency distribution of pickup and drop off over the whole month. Vertical axis shows number of pickup and drop offs in the geographical location. Frequency of pickup and drop off is much higher in downtown and airport area than residential area. Figure 5 shows the spatial distribution of pickup locations where a single red dot corresponds to a single pickup incident. Figure 5 is initially plotted using MATLAB on white background and then superimposing on Google Map.

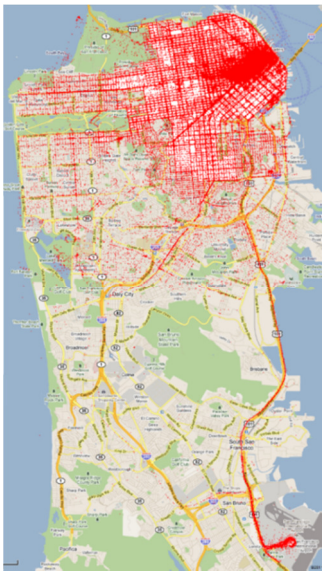


Figure 5. Spatial distribution of taxi hotspots

B. *Passenger Trip Duration and Driver's Cruise Time*

Table II shows the frequency distribution of passenger trip duration over one month where almost 48% trips take less than 10 minutes and about 4% trips take above an hour. Table III shows the frequency distribution of driver's empty cruise time which is defined as the time gap between consecutive drop off and pickup event. The statistics shows that almost more than half of the times drivers manage to get another passenger within 10 minutes of previous drop off.

TABLE I. FREQUENCY DISTRIBUTION OF PASSENGER TRIP DURATION

Trip Duration	*Frequency*	*Cumulative %*
5	97821	21.33%
10	122133	47.97%
15	77198	64.80%
20	49753	75.65%
25	33318	82.92%
30	20794	87.45%
35	13436	90.38%
40	9035	92.35%
45	6374	93.74%
50	4689	94.76%
55	3562	95.54%
60	2695	96.13%
More	17754	100.00%

TABLE II. FREQUENCY DISTRIBUTION OF CRUISE TIME

Cruise Time	*Frequency*	*Cumulative %*
5	124765	36.75%
10	65519	56.05%
15	36944	66.94%
20	24861	74.26%
25	17247	79.34%
30	11797	82.82%
35	8318	85.27%
40	6216	87.10%
45	4908	88.54%
50	3892	89.69%
55	3206	90.64%
60	2827	91.47%
More	28961	100.00%

VI. ANALYSIS OF COMMUNICATION FEATURES

A. *Degree of Connectivity*

We define the *Degree of Connectivity (DoC)* as the total number of nodes reachable from a particular node via any wireless path not longer than a given number of hops.

The *Average Degree of Connectivity (ADoC)* is the metric that characterizes the reachability of any random node with the network. Mathematically, *ADoC* specifies the average number of reachable nodes from a single source within a given path length. Hence, *ADoC* is defined by,

$$ADoC = \frac{\sum_{i=1}^{n} DoC}{n}$$

It is obvious that, increasing the wireless transmission range will have a significant impact on the *Degree of Connectivity*. Figure 6 demonstrates the impact of increasing path length as well as transmission range on *Average Degree of Connectivity (ADoC)*.The *ADoC* graph shown here corresponds to the snapshot of the whole taxi fleet at a particular time (Figure 8). The experimental time chosen was at 2:30 pm on June 5, 2008 which was a working day.

From the figure, it can be clearly observed that the average degree of connectivity is minimum for single hop connection, while longer transmission range corresponds to higher degree of connectivity. As we gradually increase the path length (hop count), more and more source-destination pairs become reachable via multi-hop communication which ultimately increases the *ADoC* of the network. All the curves show a near-linear rate of connectivity increase with the increment of path length up to a certain point when the curve becomes horizontal. This corresponds to the state when no more nodes can be explored with further hop increase. We refer this point as the saturation point. However, the slope of the curve depends on the transmission range, which implies that the longer the range the less number of hops are required to achieve maximum possible connectivity. The *ADoC* of a network after saturation indicates the portion of the fleet that can be reached from an arbitrary source node using multi-hop communication. The graph can also describe the percentage of the wireless coverage after a specific number of hops for any transmission range which may provide an estimate for the QoS provisioning of delay-sensitive applications.

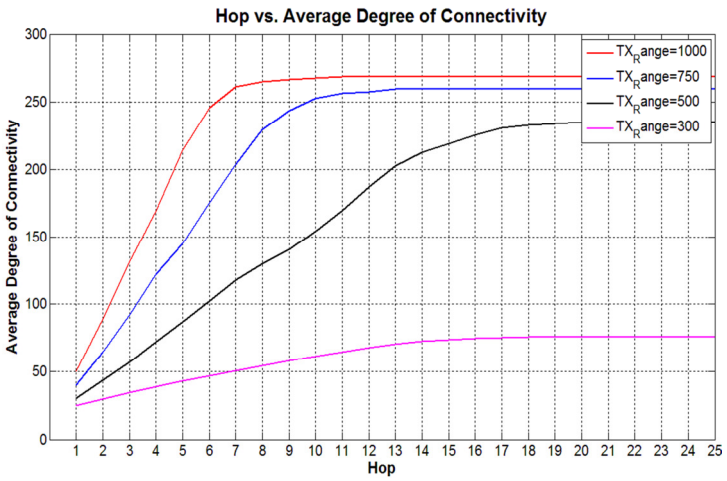


Figure 6. Impact of path length and transmission range on *ADoC*

B. *Network Partitioning*

We attempt to identify the network partitions of the whole fleet of cabs based on the instantaneous positions at a certain time. Using the same snapshot as the previously analysis, the mobile taxi nodes distributed all across the city of San Francisco can be partitioned into various clusters based on their wireless connectivity between other nodes. For a specific transmission range of 300m, it was found that, out of total 536 nodes, more than 20% of the nodes were isolated or disconnected from any other node. 16 clusters were found having 2 nodes and 9 with 3 nodes. The largest cluster found included 155 nodes, which is located in the downtown area. The second largest cluster with 120 nodes was found in the airport vicinity. Table III shows the distribution of nodes in various sizes of clusters for 300m transmission range.

TABLE III. PARTITIONING OF NODES FOR 300 METER TX RANGE

Cluster Size (# of nodes)	**Number of Clusters**	**Total Nodes**
1	110	110
2	16	32
3	9	27
4	1	4
5	3	15
6	1	6
11	1	11
22	1	22
34	1	34
120	1	120
155	1	155
Total	**145**	**536**

As the Degree of Connectivity varies along with transmission range, the clustering also changes. Table IV shows the distribution of nodes in different sizes of clusters for different transmission ranges. It is quite natural that, the number of isolated nodes (clusters with size 1) decreases as the transmission range increases. Also the total number of clusters is reduced at the same time. Figure 7 shows the average size of clusters for different transmission ranges. The average cluster size is less than 4 in case of 300m range whereas in case of 1000m it goes above 10.

TABLE IV. PARTITIONING FOR DIFFERENT TRANSMISSION RANGES

Cluster Size	Number of Clusters			
	Range 300m	Range 500m	Range 750m	Range 1000m
1	110	69	46	31
2	16	11	10	10
3	9	7	3	1
4	1	1	2	1
5	3	0	0	1
6 to 10	1	1	3	3
11 to 20	1	0	0	1
21 to 30	1	1	1	1
31 to 50	1	1	1	1
51 to 100	0	0	0	0
101 to 150	1	0	0	0
151 to 350	1	0	0	0
350+	0	1	1	1
Total	**145**	**92**	**67**	**51**

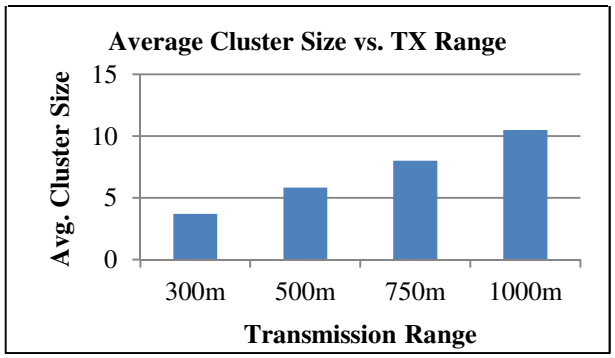


Figure 7. Average size of clusters for different transmission ranges.

C. *K-Mean Clustering*

We also simulated a K-mean clustering analysis using MATLAB which resulted in nine clusters covering 512 cabs at that particular moment. In figure 8 we show the nine clusters for the taxi cabs. The black cross mark denotes the cluster head and red dots denote taxi cab (both empty and occupied). It can be noted that, this clustering has a different objective compared with the previous one which is based on wireless connectivity and reflects the network topology. Here, the K-means clustering attempts to minimize the total distances of the nodes from each of the centroids. The following table V shows the results of K-Means clustering:

TABLE V. CLUSTERING OF NODES

Cluster number	Centroid Location		Number of nodes
	Latitude	Longitude	
1	37.6184	-122.392	45
2	37.7747	-122.404	62
3	37.7884	-122.403	70
4	37.7515	-122.395	152
5	37.6319	-122.397	54
6	37.7806	-122.423	40
7	37.743	-122.474	22
8	37.7913	-122.41	62
9	37.7832	-122.445	33

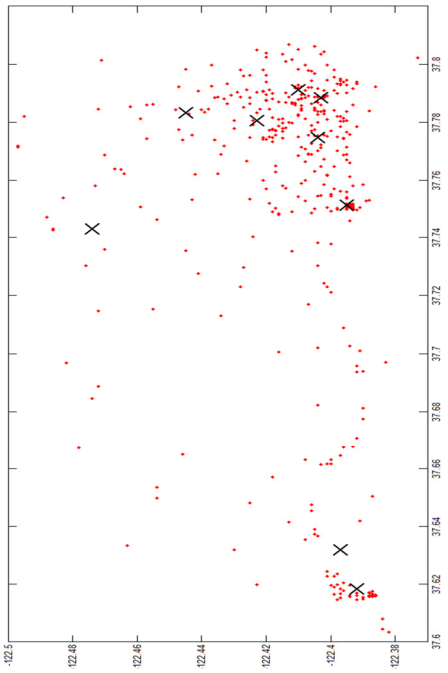


Figure 8. KMeans Clustering results showing clusterheads

VII. CONCLUSION

The analytical data presented in this paper revealed many new and useful features that can be helpful for wireless researchers, government organizations, taxi companies and even for the drivers or passengers. Our future work will explore the clustering feature of mobility for V2V communications and for DSRC infrastructure configurations.

VIII. REFERENCES

[1] NYC Taxi cab fact book

[2] http://www.lta.gov.sg/corp_info/doc/Taxi%20Info%20for%20LTA%20Website%2007.pdf

[3] http://www.yellowcabsf.com/

[4] http://www.exploratorium.edu/

[5] http://cabspotting.org/

[6] Junghoon Lee, Inhye Shin, Gyung-Leen, ParkAnalysis of the Passenger Pick-Up Pattern for Taxi Location Recommendation, Fourth International Conference on Networked Computing and Advanced Information Managemen,2008

[7] Junghoon Lee, Analysis on the Waiting Time of Empty Taxis for the Taxi Telematics System, ICCIT, 2008

[8] In-Hye Shin and Gyung-Leen Park, Association Analysis of Location Tracking Data for Various Telematics Services, ICCSA 2010

[9] M. Piorkowski, N. Sarafijanovoc-Djukic and M. Grossglauser, A Parsimonious Model of Mobile Partitioned Networks with Clustering, IEEE COMSNETS 2009, Bangalore, India.

[10] http://cabspotting.org/api

[11] http://cabspotting.org/client.html

[12] http://cabspotting.org/timelapse.html

[13] http://www.crawdad.org/