START: Status and Region Aware Taxi Mobility Model for Urban Vehicular Networks

Haiquan Wang*†, Wenjing Yang*†, Jingtao Zhang*†, Jiejie Zhao*†, Yu Wang‡ *School of Software, Beihang University, Beijing, China
†Beijing Key Laboratory of Network Technology, Beijing, China
‡Department of Computer Science, University of North Carolina at Charlotte, Charlotte, USA

***Abstract*—The mobility model in urban vehicular networks is one of the most important factors that impacts the evaluation of any vehicular networking protocols via simulations. However, to obtain a realistic mobility model in the dynamic urban environment is very challenging. Recently, several studies extract mobility models from large-scale real data sets (mostly taxi GPS data) without consideration of the statuses of taxi. In this paper, we discover three simple observations related to the taxi status via mining of real taxi traces: (1) the behavior of taxi will be influenced by the statuses, (2) the macroscopic movement is related with different geographic features in corresponding status, and (3) the taxi load/drop events are varied with time. Based on these three observations, a novel taxi mobility model (START) is proposed with respect to taxi statuses, geographic region and time. The simulation results illustrate that proposed mobility model has a good approximation with reality in the contact characteristics, trace samples and distribution of nodes in four typical time period.**

I. Introduction

Vehicular networks play a critical role in building smart cities and supporting comprehensive urban informatics. There is a growing commercial and research interest in the development and deployment of vehicular networks in urban environment. However, due to the high cost of real deployment of vehicular network systems, simulations are usually used to conduct evaluations prior to actual deployment. One of the key factor that impacts the performance of vehicular networks is the mobility model (i.e., mobility pattern of vehicles, including speed and direction). Therefore, it is necessary to obtain realistic mobility models. In addition, realistic mobility model can also used for city planning, traffic control, and other important tasks of smart cities.

Mobility models for mobile networks [1], [2] have been well-studied, and they can be classified into free space and constrained models based on the degree of randomness. For the free space scenario, the random way point (RWP) model [3] is the most commonly used, which identifies a pause time, a speed range from zero to the maximum, and a random destination in each round. An early study [4] shows that RWP in many cases is a good approximation of the vehicular mobility model based on real street maps. The constrained mobility models [4], [5], [6] are closer to the realistic mobility by taking the geographic structure (such as the street layout, traffic rules, and multi-lane roads) into consideration. Recently, there is also a new trend to extract the vehicular moblity model from real vehicular trace data (mainly taxi GPS trace data) [7], [8]. For example, [8] propose mobility models by estimating three parameters (turn probability, road section speed and travel pattern) from Shanghai taxi trace data. However, all these constrained mobility models are complicated and strongly related to the simplified maps, and the existing taxi-based mobility models ignore the statuses of taxi (vacant or occupied).

In this paper, we argue that the taxi behavior and geographic features are strongly related to the status of the taxi. In addition, time is another factor effecting the taxi behavior because the passenger flow volume varies with time. Meanwhile, the passenger flow volume effects the quantity of working taxies. We validate such claims by statistical analysis over a large-scale Beijing taxi trace data.

Based on these discoveries, we propose a *STAtus and Region aware Taxi mobility model* (START). In the macro scope, instead of simply dividing the area into coarse-grain regions, START divides the area into two set of regions according to the density of passenger load or drop events for each time period.

When a taxi takes a passenger, the current location is selected from the set of load-event regions. The destination region, where the drop event happens, is selected in the set of drop-event regions.

We investigate the relationship between load-event regions and drop-event regions and use it to decide the start point and destination. Routes from the start points to destinations are found by Dijkstra algorithm. For microscope, the speed of the taxi is generated based on its status, which is learned via statistical analysis. Extensive simulations are carried out to compare the similarity of node trace characteristics and contact characteristics. The results show that START model has a good approximation of the real scenario in trace samples, in terms of distribution of nodes and the contact characteristics. To the best of our knowledge, our work is original to develop mobility models by investigating taxi behavior and geographic features

Yu Wang is the corresponding author (e-mail: yu.wang@uncc.edu).

This research has been partially supported by the US National Science Foundation (NSF) under Grant No. CNS-1319915 and CNS-1343355, the National Natural Science Foundation of China (NSFC) under Grant No. 61300173, 61428203, and 61170295, the Project of Aeronautical Science Foundation of China under Grant No.2013ZC51026 and 2011ZC51024, the Fundamental Research Funds for the Central Universities under Grant No. YWF-12-LXGY-001, and the State Key Laboratory Software Development Environment and Network Information and Computing Center of Beihang University.

of different statuses.

The rest of this paper is organized as follows. Section III provides the statistical results from real data to validate two important assumptions for START model. Section IV presents the detail of START model. Simulation results are reported in Section V. Finally, Section VI concludes this paper.

II. Related Works

For the free space scenario, the random way point (RWP)[3] movement model is the most commonly used. The movement model identified a pause time, speed range from zero to the maximum, and movement area where the model select a random destination. Amit Kumar Saha, el at. [4] found that RWP in many cases the Random Waypoint mobility model is a good approximation of the vehicular mobility model based on real street maps. The constrained mobility models show closer relevance to the realistic. Literature [**?**] demonstrated that graph structure is close related with inter-contact time distribution in both random and social mobility on grid-based graphs. Some models [4], [**?**], [8], [5], [6] take the geographic structure into consideration.

In this section, we sum up the researches on mobility models.Mobility model can be classified into free space and constrained models[1], [2] based on the degree of randomness. Random walk, Random Waypoint and Random Direction mobility model are three classical free style mobility models. Those models defined simple mobility patterns, which is good for us to create mobility models and analysis. But they also have notable disadvantages, e.g., they are out of reality, because too many factors are ignored.

In order to increase the degree of reality, mobility models are constrained or relayed in many respects. Some researchers build models based on geographic models. Manhattan models are a typical models which models the city as a Manhattan style grid, with a uniform block size across the simulation area, while all streets are two-way with a lane in each direction which constrained car movements [5], and nodes can move straight forward or turn direction at a cross road. Other models import the map information into mobility models. In 2004, Saha, A.K[20] model the vehicle networks based on the real map. It compared with the RWP models, a frequently used model in vehicular networks, to find out the difference between the RWP and real traces. In 2005, literature [22] proposed a comprehensive models on wireless vehicular networks and transportation. This paper simplifies the real map to estimate the network performance in ad hoc and proposes the mobility model STRAW. Besides, some other mobility focus on the microscopic characteristics of mobility, they introduce the transportation features into mobility, such as the traffic lights and accretion on some road segments. The information of the geography can improve the reality. In 2007, Atulya Mahajan, et al. [**?**] accounted for the street layout, traffic rules, multi-lane roads, acceleration-deceleration, and radio frequent (RF) attenuation due to obstacles, and further evaluated the synthetic maps by comparing with real maps. In 2008, David R. Choffnes, et al.[6] developed their movement model based on a realistic vehicular traffic model on road defined by real street map data. but too many microscopic detail can not be applied to scenarios with large scale nodes.

In recent years, vehicular sensors or handheld devices spread rapidly, that makes it possible to collect and analyze the real traces of large amount of nodes and helps us to improve the traffic and network macroscopically. In 2010 and 2012, Huang H, et al. [8], [**?**] proposed mobility models based on taxi trace data in Shanghai, China. They designed three parameters, i.e., turn probability, road section speed and travel pattern, which can be estimated by analyzing the data statistically. But it is complicated to re-implement this model, for the model is strongly related to the map they simplified from the real map. To the best of our knowledge, our work is original to develop mobility models by investigating taxi behavior, time and geographic feathers of different statuses.

III. Statistical Analysis of Beijing Taxi Traces

In this section, we focus on the statistical analysis on the speed, duration, and taxi event characteristics of the Beijing taxi data set, a large-scale urban vehicular trace data.

A. Trace Dataset: Beijing Taxi Traces

A real-world GPS data set is used for our analysis, which was generated by $12,455$ taxis in Beijing, China within five days from March 3 to March 7, 2011. In the data set, each entry includes a base station ID, company name, taxi ID, timestamp, current location (including longitude and latitude), speed, event, status, et al. Of all the fields, the taxi ID, time stamp, and current location, status and event are used for study. There are five types of events and four types of statuses in the data set, which are summarized in Table I. We only focus on the vacant and occupied statuses (and corresponding load and drop events) in this paper. Note that GPS traces from taxis have been used recently for inferring human mobility [9] and modeling city-scale traffics [10]. Therefore, we believe that they are also suitable to be used to build mobility models in large-scale urban scenario.

TABLE I: Events and statuses in Beijing taxi traces

Event	Explanation
0 (drop)	a taxi's status changes to vacant.
1 (load)	a taxi's status changes to occupied.
2	set up defense.
3	cancel defense.
4	no event happened.
Status	Explanation
0 (vacant)	a taxi is vacant.
1 (occupied)	a taxi is occupied.
2	a taxi is setting up defense.
3	stop running.

B. Three Claims on Taxi Behaviors

The following two claims are foundations of the proposed taxi mobility model, which are also reasonable based on the experience in our daily life:

- **Claim 1:** The behavior of a taxi changes when its status updates. When a taxi is occupied, its destination is certain, and the vehicular speed of a occupied taxi accelerates relatively. In contrast, when a taxi is vacant, it slows down or even stops to search for potential passengers along the road. Therefore, taxi behavior characteristics, such as speed and status duration, vary consequently.
- **Claim 2:** The taxi behavior associates with time. The quantities of load/drop events may vary with time comforming to certain rules. For example, the quantity of passengers late in the night is relatively fewer than that of passengers during the daytime.
  1) The hotspots of load/drop events vary with time.
  2) For the same time period during a day, the load/drop events distribute similar.
- **Claim 3:** The movement behavior of taxis associates with geographic features. When a taxi is occupied, the destination may be tend to certain geographic places, such as the airport. Meanwhile, when a taxi is vacant, its driver tends to look for some hot spots, where more people want to take a taxi, such as downtown areas. Therefore,
  1) The destination selection of a taxi is influenced by different regions.
  2) Events occur in different regions un-evenly, passenger drop and load events are distinct.

Next, we analyze the speed, duration and passenger load/drop events distribution over the Beijing taxi traces to validate the three claims above.

C. Taxi Speed

We first investigate the average speed $\overline{speed}$ in each status. If taxi i drives in occupied status for a distance d using time t, then its average speed in this status is d/t. From March 3 to 7, 2011, $\overline{speed}_{vacant} = 3.627m/s$, while $\overline{speed}_{occupied} = 7.083m/s$. Clearly, occupied taxis drive faster. To further investigate the cumulative speed distribution, proportion for every $\overline{speed}$ section is calculated and plotted in Fig. 1. Here, a point at $(5, 0.2)$ means 20% records fall in the range $[0, 5)km/h$. We also fit the speed to model the microscope behavior (will be discussed in Section IV). Fig. 1 shows that $\overline{speed}$ distribution differs for each status and with strong regularity for each status.

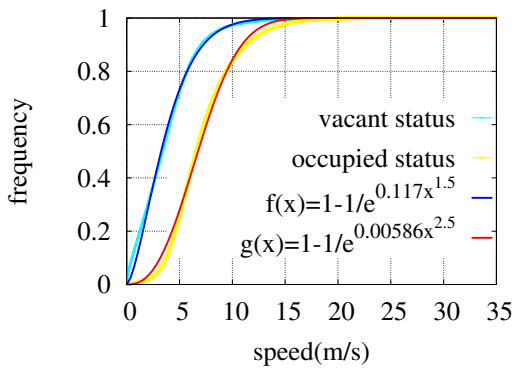


Fig. 1: Speed distributions for vacant and occupied statuses.

D. Taxi Status Duration

The duration distribution for each status are shown in Fig. 2. Status duration represents the time length of a taxi staying in a certain status. The red line presents the duration time distribution for vacant status, and the green one is for occupied status. Note that the red line (vacant status) approaches to one earlier than the blue line (occupied status). So the value of vacant duration is smaller than the value of occupied duration. This is reasonable since drivers tend to shorten the waiting time to raise their incomes.

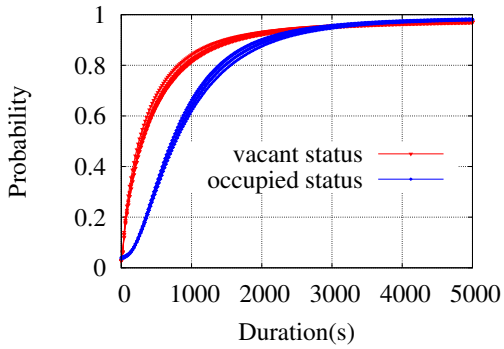


Fig. 2: Status duration distributions.

Overall, the statistical results for both speed and status duration are consistent with *Claim 1*, that is, the behaviors of taxis are similar within each status while differ between the two statuses.

E. Event distribution varied with time

In this section, we analysis how many load and drop event happened for each hour. We divide the regions into $200m \times 200m$ grids, and count the load and drop events happened in every hour. As table II, the analysis results show that the total volumes of load and drop events for a week are similar, close to 2.7million. And the time periods are identical for the peak and valley of the different event quantity. Shown as figure 3, the curves of event quantity varied with time of the two kinds of event are similar, too. Which is consistent with our experiences, because the load and drop quantity should be in balance. Whats more, the event quantity varied with time show strong regularitywhen the quantity decreases or increases is relatively certain.

TABLE II: Events quantity varied with time

item	drop event quantity	load event quantity
total quatity for a week	2,679,385	2,707,290
maximum of an hour	28,583	28,130
minmum of an hour	861	918
time of the peak value	2011/11/04 19:00-20:00	2011/11/04 19:00-20:00
time of the valley	2011/11/03 4:00-5:00	2011/11/03 4:00-5:00

From table II, the peak of the event quantity happened at the same time period. In this case, we investigate the load and drop events further. By filtering the cells whose event quantities are lower than 5 per hour, we found that although

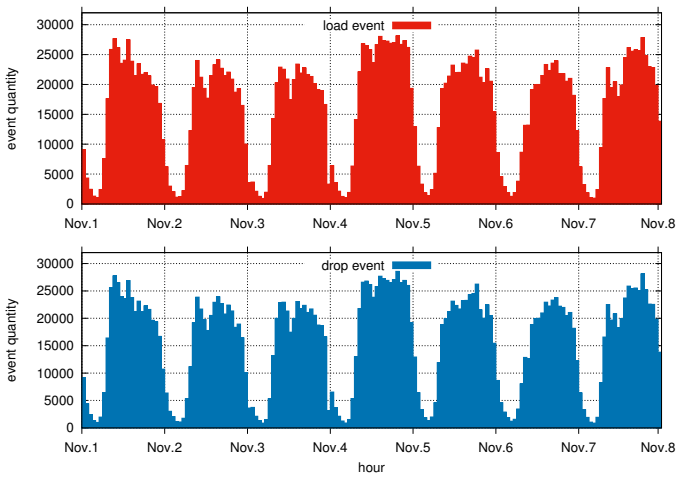


Fig. 3: Taxi event varied with time.

the event quantities and time periods are similar, the load and drop events tend to happen in different places.

From 4:00 to 5:00, a hot spot of the drop events is the beijing west railway station.This may be caused by catching morning trains. From 19:00 to 20:00, both the load and drop events happens frequently. We can figure out the main road of Beijing in figure 5 (b) and (d).

- For drop and load event, the total event number for a week are very close.
- The peak and valley occurs in similar time. From the data, we can find that the peak and valley happened at same time, coincidentally.
- The maximum event number is much larger than the minimum event number. The quantity of event number changes over time. The curves for the load and drop events follows similar rules. The quantities at the same time range are similar, too.
- the hotspots of the load and drop events are relatively fixed during a time peroid.

The analysis results fit with our daily experience: (1) the amount of passengers decreases early in the morning, (2) the load event quantity equilibrates with the drop event quantity, and (3) the event quantity at certain time shows certain regularity.

F. Taxi Event Distribution

In this section, we try to find out whether the load and drop events distributes with geographical perference. Fig. 5 shows the load/drop events hotspots (more than 20 events happened in one hour) in same time periods, from 19:00 to 20:00. We choose two workdays and a weekends. Comparing the load and drop event hotspots, we can find the events distribute much even for loading passengers. For each kind of events, some peaks occurs repeatly, as highlighted in the red circles. Through, the event amounts are different from the workdays and the weekend. The position of those hotspots are still similar. This phenomenon may be caused by the load-event spots are mainly at homes of the citizens, while the drop-event spots tend to gather together at workplaces, shopping malls, railway stations or scenic spots. Overall, the amount of loading/dropping passengers in each cell shows geographic features: the distribution is uneven, and the difference between load/drop-event distributions illustrates the load/drop-event regions are different. All of these support *Claim 3*.

IV. START Mobility Model

A. Overview of START

Movement model defines the mobility pattern of nodes, which can be represented as a collection of path segments, say $Paths :< p_1, p_2, ..., p_n >$. To generate a p_i, START takes two steps: destination selection and moving process (from current location to the selected destination).

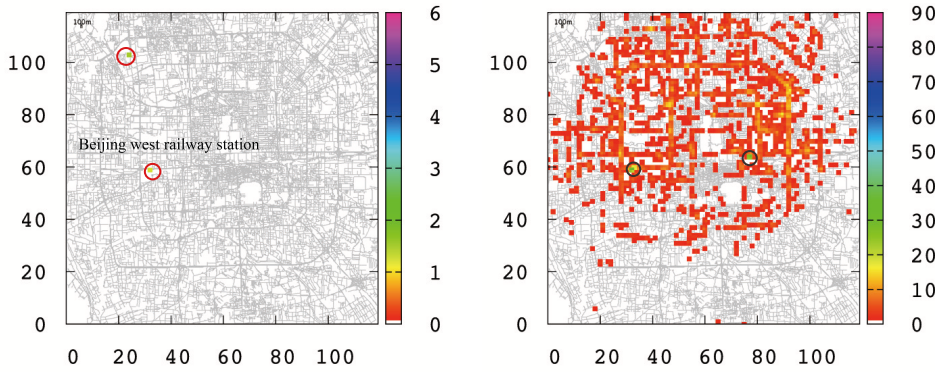
Destination Selection: In START, the selection of a destination of a node is closely related to not only its current location but also its current status. A travel path of a taxi can be simplified as a multi-hop process, in which a hop indicates an load/drop event happened. Seeing that, we define a *region transition probability* to figure out the probability of the next hop falling in a certain region from the current region. Particularly, when two successive events are different: one load and one drop. We divide the area into regions by the density of passenger load/drop events at different time, respectively. We will show how to divide the area in the next subsection. For a time period t, let

$$\mathbf{R}_t^{load} = \{R_{i,t}^{load}\} \tag{1}$$

$$\mathbf{R}_t^{drop} = \{R_{j,t}^{drop}\} \tag{2}$$

denote the set of regions of load and drop events, respectively. i and j are the region id of the load and drop event regions and t is a integer which denotes the hour of current time, such as t=7 means the time is in the range of 7:00:00-7:59:59. The union of the load or drop region set for a t is the whole area, that is $\bigcup\{R_{i,t}^{load}\} = \bigcup\{R_{j,t}^{drop}\} = Area$. Then, for every time period, two transition probability matrixes are calculated: one is the transition probability from a passenger drop region $R_{i,t}^{drop}$ to a passenger load region $R_{j,t}^{load}$ or $R_{j',t+1}^{load}$, while the other is the transition probability from a passenger load region $R_{j,t}^{load}$ to a passenger drop region $R_{i,t}^{drop}$ or $R_{i',t+1}^{drop}$. From the status duration distribution, we find that the more than 95% status duration for a taxi status is no longer than one hour. In this case, the transition probability from a region in time t to another region at time t, where t is larger than t+1, will be ignored. If the status of a taxi changes to being vacant, its current location locates and time in a $R_{i,t}^{drop}$. Consequently, a destination region in $\mathbf{R}_t^{load}$ or $\mathbf{R}_{t+1}^{load}$ will be selected by querying the transition matrix. Then, START will randomly select a map node in the region as the destination. As to the status of a taxi changing to being occupied, the destination selection process is similar except for that the transition matrix from $\mathbf{R}_t^{load}$ to $\mathbf{R}_t^{drop}$ or $\mathbf{R}_{t+1}^{drop}$ is used instead. In summary, during this process, the destination of drop/load location is randomly selected based on the region transition matrix corresponding to the current status.

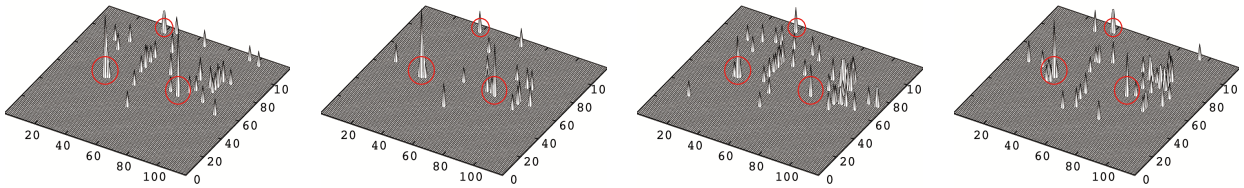
Moving Process: When the source location (current location) and destination location are given or selected, the next



(a) drop events at 4:00-5:00 (b) drop events at 19:00-20:00

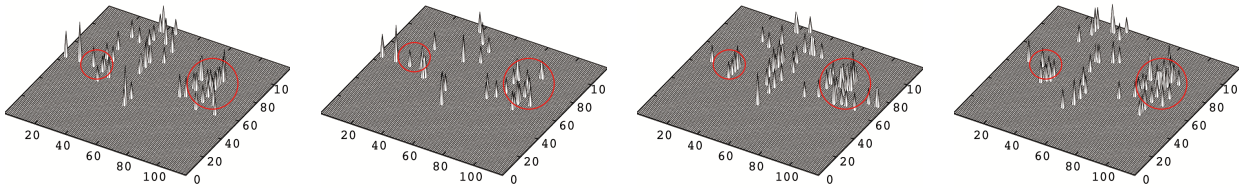
(c) load events at 4:00-5:00 (d) load events at 19:00-20:00

Fig. 4: Taxi density for load/drop events in one hour.



(a) Nov.1th, Tuesday (b) Nov.3th, Thursday (c) Nov.5th, Saturday (d) Nov.6th, Sunday

hotspots of drop events



(e) Nov.1th, Tuesday (f) Nov.3th, Thursday (g) Nov.5th, Saturday (h) Nov.6th, Sunday

hotspots of load events

Fig. 5: Taxi density for load/drop events in one hour.

step is to find a path to connect them. To simplify the process, we adopt the Dijkstra algorithm, which will find a shortest path from the source to the destination based on the map. The speed of the path then is assigned to *speed* based on the current status. Here, the value of *speed* is drawn from the average speed distribution of corresponding status, which will be introduced in the last subsection of this section.

B. Region Transition Probability

For event distribution of the load and drop events are different with each other and varied with time, the region $R_{i,t}^{load}$ and $R_{j,t}^{drop}$ are recognized by different metrics, that is, drop or

TABLE III: region recognization parameters

Item	0:00-8:59	9:00-12:59	13:00-20:59	21:00-23:59
η_{drop}	556	84	180	51
η_{load}	58	84	182	51
_top	200	200	200	200
clusterSize	500	500	500	500

load event distribution during each time period. For instance, if the taxi is currently occupied, then the next hop event is the drop one. Hence, choosing a target region from a region set obtained based on drop event distribution is more logical.

Before defining the region transition probability, we first describe our **region recognition process**. Firstly, we divide the area into small grids of side length 200 meters, and define them as cells (as shown in the following equation) where lon and lat are relevant longitude and latitude, X and Y are side length of the grid/cell).

$$C_{x,y} ::= \{(lon, lat)|x \leq \frac{lon}{X} < x+1, y \leq \frac{lat}{Y} < y+1\}.$$

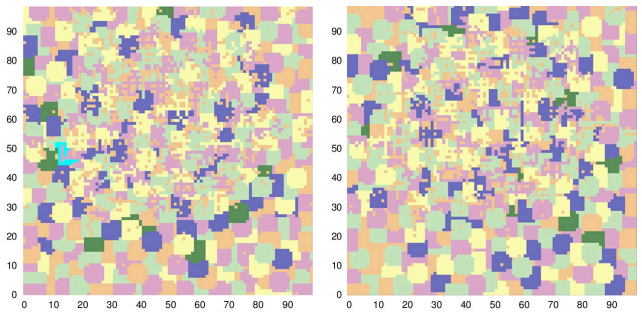
Then, we consider a region as a union of adjacent cells, as following.

$$R_m ::= \{C_{i,j}|if C_{x,y} \in R_m \Rightarrow \|x-i\| \leq 1, \|y-j\| \leq 1\}.$$

m denotes the region identifier. The main idea of clustering cells to regions is merging adjacent cells whose event density is larger than an event threshold η into a same region. For drop/load event, this process is conduct respectively. Three parameters, threshold η, Cluster Size and range $\_top$, are used in this process. Cluster Size defines the size of region, i.e., it is a limitation on the size of a region, saying $\|R_i\| \leq ClusterSize$. We only consider the top $_top$ regions, in which event density of each cells are larger than threshold η. For each time period, we set different threshold by its average events number in each cell at that time, that is, η equals to twice the average event number.

The overall clustering algorithm is shown as follows. We sort the all the cells by event density in descending order, and begin with the first cell to search its neighbors whether to join the same region or not using breadth traversal. After the top regions are formed, the other cells which do not belong to the top _top regions will also be clustered into regions, whose size should still be small than Cluster Size. Consequently, every cells will be clustered into regions and the size of each region are not larger than Cluster Size. By clustering cells into regions, two region sets, R^{load} and R^{drop}, can be recognized from the data set for every time period. As can be seen from Fig. 6, the differences among load region and drop regions for the Beijing taxi data set are clear. In this figure, every colored block presents a region.In addition, Cluster Size = 200, range t = 200 (range is the same with Cluster Size by coincidence) and threshold $\eta = 121$ are set for load event and η equates with 141 for drop event (these are set by the average event density of the top 5000 cells order by its event density).

Calculation of region transition probability: we define a region transition probability to figure out the probability of the next hop falling in a certain region from the current region. A transition probability from a load region i to a drop region j from time t is denoted as:



(a) drop event regions (b) load event regions

Fig. 6: Region recognition

$$p_{i \to j,t}^{load \to drop} \tag{3}$$

Similarly, A transition probability from a drop region j to a load region i from time t is denoted as:

$$p_{j \to i,t}^{drop \to load} \tag{4}$$

After clustering cells into regions, the transition probability from R_i^{load} to R_j^{drop} and the one from R_i^{drop} to R_j^{load}, donated as $p_{i \to j}^{load \to drop}$ and $p_{i \to j}^{drop \to load}$, can be caculated. Since both transition probability can be calculated similarly, we only introduce the detailed one of $p_{i \to j}^{load \to drop}$.

To calculate a $p_{j \to i,t}^{drop \to load}$, we should find out all the taxies in the drop region j of time t, denoted as $V_{j,t}^{drop}$. For every taxi $v \in V_{j,t}^{drop}$, its next load event record, which marks where when it load passengers, can be extracted. we map every next load event records to according region of corresponding time range and find out how many records fall in the region i. The transition probability equals to the ratio between the total taxies and the amount of taxies whose next load events happen in region i. We restrict the time from current drop event record to next load event cannot be across more than one hour, that is the region i belongs to the region set of time t or time $t+1$ and we ignore the records whose hour of timestamp is more than $t+1$ hour. For example, for $t = 7$, the record with timestamp 9:00:00 is invalid, while the record whose timestamp is 8:59:59 is valid. The method to calculate a $p_{i \to j,t}^{load \to drop}$ is similar.

C. Speed Distribution

To obtain the speed distribution of each status, we fit the cumulative average speed distribution to get the cumulative probability distribution function, and then take a derivative with it to obtain the speed probability distribution. From Fig. 1, the $\overline{speed}$ distribution shows exponential law. Given that, we set the function form as follows:

$$\begin{cases} f(x) = 1 - 1/exp(a_1 x^{1.5}) \\ g(x) = 1 - 1/exp(a_2 x^{2.5}) \end{cases} \tag{5}$$

Here, $f(x)$ is the function form for the $\overline{speed}$ distribution of vacant status, and the other one $g(x)$ is for that of occupied status. The *root mean square* (rms) of residuals for each fit are reported in Table IV. The smaller rms of residuals means better fitting. In this table, the values are all less than 0.02, showing good similarity.

TABLE V: Simulation Parameter

type	Trace	START	SP	RWP
mobility model	External	START	ShortestPathMapBased	RandomWaypoint
simulation time (sec)	7200			
node scale	1000/3000			
contact range (meter)	200			
speed range (m/s)	-	-	[0,16.67], [0,22.22] [0,33.33]	
external file	traces of 1000 nodes from 6:00 to 8:00 in Nov.8th,2011 traces of 3000 nodes from 11:00 to 13:00 in Nov.8th,2011 traces of 3000 nodes from 17:00 to 19:00 in Nov.8th,2011 traces of 3000 nodes from 22:00 to 24:00 in Nov.8th,2011	-	-	-
map file	-	Beijing Map		-

TABLE IV: Parameters and rms of residuals of fitting curves

Fitting curves	rms of residuals
$f(x)$ with $a_1 = 0.117$	0.0113159
$g(x)$ with $a_2 = 0.0586$	0.0137029

V. Model Verification

In this section, START mobility model is validated on the aspects of node distribution and contact characteristics compared with existing mobility models and the real traces. We pick two simple mobility models for comparison: one free space - Random Way Point (RWP) model, the other is constrained model, Shortest Path (SP). SP mobility model is based on the underlying map of Beijing where vehicles move along the map roads by Dijkstra algorithm to random destinations. Both models take no consideration of the node statuses and geographical distributions. All mobility models are implemented on Opportunistic Networking Environment (ONE)[11].

In our simulations, vehicles are deployed in an area of $24,000 \times 24,000m^2$, including fourth ring roads in Beijing. The speed range of RWP and SP need to be configured. To ensure the accuracy, we choose three different speed ranges $[0, 16.67]m/s$, $[0, 33.3]m/s$ and $[0, 22.2]m/s$ (the upper bounds of speed match the speed limits 80,120,60 km/h. The parameters for the simulations are shown asV. To evaluate the time feature of START, 4 time periods are chosen: in the morning from 6:00:00 to 7:59:59, at noon from 11:00 to 13:00, at afternoon from 17:00 to 19:00 and lat in the evening from 22:00 to 24:00. Accordingly, we extract the real traces at corresponding time period of 21st November 2011. Because the taxi quantity is much small in the morning, traces of 1000 taxies are extracted, while for the other three time period, traces of 3000 taxies are randomly selected.

A. Traces and Node Distributions

Trace samples and their node distribution snapshots from different mobility models are reported in Fig. 7 and Fig 8. Fig. 7 shows the trace in one day. The traces of the real data and START only cover some parts of the area, while the traces of SP and RWP almost go through the whole area. Recall that SP and RWP select a destination randomly in the area, while START takes the associations between current region and destinations into consideration (which satisfies the movement rules of taxis). In Fig. 8, real trace, START and SP exhibit the road structures, while the node distribution of RWP is much uniform. As to START, the destination section process decides that it tends to select a destination in the regions with higher load/drop event probability. Therefore, with the decline of the randomness, the snapshot of START becomes much clear and centralized on the main roads, which matches real traces very well.

Since the node distribution has a great impact on the transport and network performance, a good understanding of it can help to route and control. However, nodes are dynamic leading to a dynamic node distribution. In order to quantify the changing node distribution, we introduce the in/out degree. The in/out degree figures out how many taxies moving in or out from a region in a time period. In/out degree defines how many nodes moveing in or out a area during a period of time. We divide the simulation scenario into grids of $400m \times 400m$ to investigate the in/out degree, and the time period to measure the in/out degree is as two hours according to the simulation time.

Firstly, we compare the in/out degrees of different days at same time period. From figure XXX and XXX, we can find the relative error with their average in/out degree is less than XXX. This suggests that the in/out degree for multiple days are consistent and we can compare our START traces with it.

Figure XXX demonstrates the in degree of START, SP, RWP and the average in degree distribution. The relative error of START compared with the average one is XXX, that of SP is XXX, and RWP is XXX.

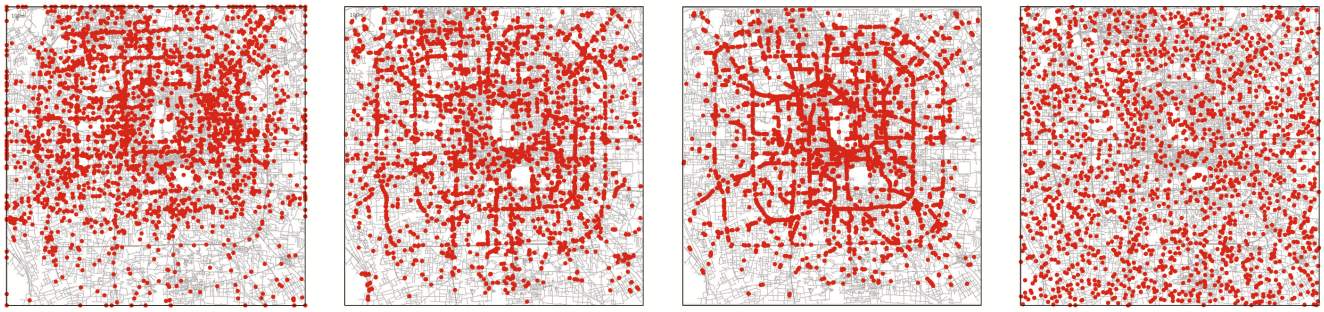
B. Contacts Characteristics

Contact defines an opportunity to communicate, a contact range should be assigned to quantify the max distance we consider two nodes contact with each other. In our simulations, the transmit ranges of high-speed interfaces and blue tooth interface are both set to $200m$ for potential contacts.

The contact time and inter contact time among vehicles are also evaluated as the indicators to validate the similarity. Fig. 11 and fig. 10 reports the contact time and inter-contact distributions, which shows the probability of the contact or inter-contact time smaller than certain time length. To substantiate, a point $(25, 0.5)$ in the plots means the probability is 0.5 when contact or inter-contact time is shorter than $25s$. From fig. 11 the speed will effect the contact distribution of

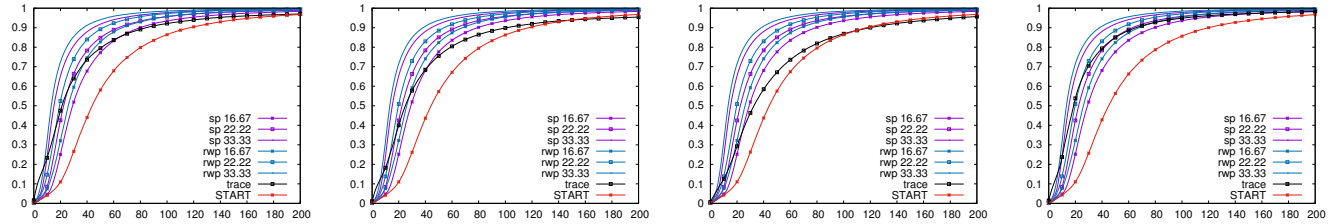
(a) Real Trace (b) START (c) SP (d) RWP

Fig. 7: Trace samples.



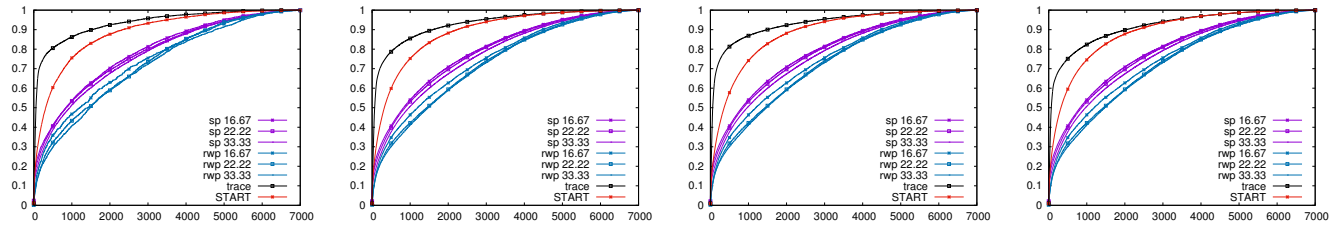
(a) Real Trace (b) START (c) SP (d) RWP

Fig. 8: Nodes distribution snapshots.



(a) 1000 nodes from 6:00 to 8:00 (b) 3000 nodes from 11:00 to 13:00 (c) 3000 nodes from 17:00 to 19:00 (d) 3000 nodes from 22:00 to 24:00

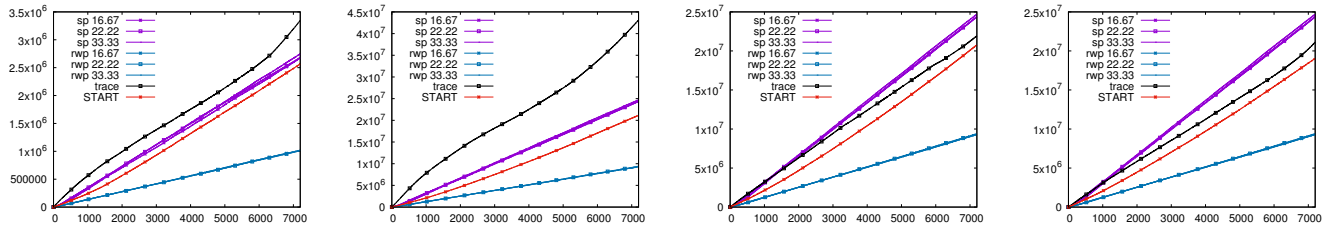
Fig. 9: Contact times distribution



(a) 1000 nodes from 6:00 to 8:00 (b) 3000 nodes from 11:00 to 13:00 (c) 3000 nodes from 17:00 to 19:00 (d) 3000 nodes from 22:00 to 24:00

Fig. 10: Inter-contact times distribution

SP and RWP. The speed in a low speed range will makes the contact time distribution of SP and RWP approach that of real traces. For START, in the time from 17:00 to 19:00, the model matches with the real traces best. Fig. 10 also show the sum of contact time regarding to the simulation time. For the inter contact time, the curves of SP (RWP) are close with each other, which means the speed range dose not make much difference on the inter contact time. Compared with the distribution of contact time, it may because that the inter contact time are much longer than the contact time, the inter contact time is more stable to present the performance of nodes. In Fig. **??**(a), the three purple lines of SP and the three blue lines of RWP



(a) 1000 nodes from 6:00 to 8:00 (b) 3000 nodes from 11:00 to 13:00 (c) 3000 nodes from 17:00 to 19:00 (d) 3000 nodes from 22:00 to 24:00

Fig. 11: Time vs. total contact time.

are overlapping with each other, too. The curves of the total contact time vs. time present a liner law. From the analysis above, for SP and RWP, the differences of speed ranges show little influence on these curves.

To conclude, by comparing the node distribution and contact characteristics, the evaluation results confirm that START mobility model achieves great similarities with the real data. START takes the usage of speed and geographic features related with taxi status, while SP employs the map information and RWP is a random model taking use of no realistic data.

VI. Conclusion

Since the mobility model is important for vehicular networks and other smart cities applications, a new mobility model START based on real taxi GPS data is proposed. By assuming the taxi behavior is related with its statuses and geographic features, statistical experiments are conducted to verify those assumptions using the real trace data. With carefully estimations of the average speed distribution of each status and the region transition probability between drop and load event regions, START considers both macroscopic and microscopic movements. For the macroscopic movements, a node moves and switches between load-event regions and drop-event regions. Then the microscopic movements (such as speeds for each status) can be applied. START is implemented and evaluated in ONE simulator by comparing with the real trace, RWP and SP mobility models. For both node distribution and contact features, START shows better performance than the other two mobility models. This demonstrates that START has a good approximation with reality and can be used for urban vehicular network research and applications.

Appendix

Algorithm 1 Clustering

INPUTs: $Cells = \{C_{x,y}\}$, the event threshold η, $CLUSTERSCALE$, and $REGION\_ID\_SEED = 1$.
$ClusterQueue = \varnothing$ & $UsedCells = \varnothing$
Sort $Cells$ by events in descending order
for $CELL_{x,y} \in Cells$ **do**
  if $CELL_{x,y} \notin UsedCells$ **then**
    $CELL_{x,y}.region = REGION\_ID\_SEED$
    $size = 1$
    $REGION\_ID\_SEED = REGION_ID_SEED + 1$
    $ClusterQueue.enqueue(CELL_{x,y})$
    $UsedCells.add(CELL_{x,y})$
    while $ClusterQueue \neq \varnothing$ **do**
      $CELL_{x,y} = ClusterQueue.dequeue()$
      if $REGION\_ID\_SEED \leq \_top$ **and** $CELL_{x,y}.events \geq \eta$ **then**
        $enqueueNeighbor(CELL_{x-1,y})$
        $enqueueNeighbor(CELL_{x-1,y-1})$
        $enqueueNeighbor(CELL_{x-1,y+1})$
        $enqueueNeighbor(CELL_{x+1,y})$
        $enqueueNeighbor(CELL_{x+1,y-1})$
        $enqueueNeighbor(CELL_{x+1,y+1})$
        $enqueueNeighbor(CELL_{x,y-1})$
        $enqueueNeighbor(CELL_{x,y+1})$
      else
        $enqueueNeighborOthers(CELL_{x-1,y})$
        $enqueueNeighborOthers(CELL_{x-1,y+1})$
        $enqueueNeighborOthers(CELL_{x-1,y-1})$
        $enqueueNeighborOthers(CELL_{x+1,y})$
        $enqueueNeighborOthers(CELL_{x+1,y-1})$
        $enqueueNeighborOthers(CELL_{x+1,y+1})$
        $enqueueNeighborOthers(CELL_{x,y-1})$
        $enqueueNeighborOthers(CELL_{x,y+1})$
      end if
    end while
  end if
end for

References

[1] X. Lu, Y.-c. Chen, I. Leung, Z. Xiong, and P. Lio, "A novel mobility model from a heterogeneous military MANET trace," in *Proc. of 7th International Conference on Ad-hoc, Mobile and Wireless Networks (ADHOC-NOW)*, 2008.

[2] S. Ahmed, G. C. Karmakar, and J. Kamruzzaman, "An environment-aware mobility model for wireless ad hoc network," *Computer Networks*, vol. 54, no. 9, pp. 1470–1489, 2010.

[3] J. Broch, D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva, "A performance comparison of multi-hop wireless ad hoc network routing protocols," in *Proc. of the 4th annual ACM/IEEE international conference on Mobile computing and networking*, 1998.

Algorithm 2 $enqueueNeighbor(CELL_{x,y})$

if $CELL_{x,y}.events \geq \eta$ **and** $size < clusterSize$ **and** $CELL_{x,y} \notin UsedCells$ **then**
  $ClusterQueue.enqueue(CELL_{x,y})$
  $CELL_{x,y}.region = REGION\_ID\_SEED$
  $UsedCells.add(CELL_{x,y})$
  $size = size + 1$
end if

Algorithm 3 $enqueueNeighborOthers(CELL_{x,y})$

if $size < CLUSTERSCALE$**and** $CELL_{x,y} \notin UsedCells$ **then**
  $ClusterQueue.enqueue(CELL_{x,y})$
  $CELL_{x,y}.region = REGION\_ID\_SEED$
  $UsedCells.add(CELL_{x,y})$
  $size = size + 1$
end if

[4] A. K. Saha and D. B. Johnson, "Modeling mobility for vehicular ad-hoc networks," in *Prox. of the 1st ACM International Workshop on Vehicular Ad Hoc Networks*, 2004.

[5] F. J. Martinez, J.-C. Cano, C. T. Calafate, and P. Manzoni, "Citymob: a mobility model pattern generator for vanets," in *Proc. of IEEE International Conf. on Communications Workshops*, 2008.

[6] D. R. Choffnes and F. A. N. E. Bustamante, "An integrated mobility and traffic model for vehicular wireless networks," in *Proc. of the 2nd ACM international workshop on Vehicular ad hoc networks*, 2005.

[7] M. Kim, D. Kotz, and S. Kim, "Extracting a mobility model from real user traces." in *Proc. of 25th IEEE International Conference on Computer Communications (INFOCOM)*, 2006.

[8] H. Huang, Y. Zhu, X. Li, M. Li, and M.-Y. Wu, "Meta: A mobility model of metropolitan taxis extracted from gps traces," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, 2010.

[9] R. Ganti, M. Srivatsa, A. Ranganathan, and J. Han, "Inferring human mobility patterns from taxicab location traces," in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2013.

[10] J. Aslam, S. Lim, X. Pan, and D. Rus, "City-scale traffic estimation from a roving sensor network," in *Proc. of the 10th ACM Conference on Embedded Network Sensor Systems (SenSys)*, 2012.

[11] A. Keraen, J. Ott, and T. Karkkainen, "The ONE simulator for DTN protocol evaluation," in *Proc. of the 2nd International Conference on Simulation Tools and Techniques*, 2009.