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

Haiquan Wang
Wenjing Yang
Jingtao Zhang
Jiejie Zhao
School of Software, Beihang University, Beijing, P.R.China
Beijing Key Laboratory of Network Technology, Beijing, P.R.China

Yu Wang*
Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA
corresponding author

Abstract

Using a realistic mobility model will enhance the validity of simulations, while the difficulty lies in discovering rules from large amounts of data and applying those rules. Researchers have been working on mobility model extracting from real data set, whereas the taxi behavior difference between different taxi statuses was ignored in previous works. As statistical analysis results of real taxi traces demonstrate certain distribution law of speed and duration for each status and geographical features, a novel taxi movement model START is proposed with respect to taxi statuses and regions. Micro-scope and macro-scope are both considered. In micro-scope, speed and duration cumulative distribution are fitted to estimate the speed of node and the nodes duration in corresponding status. And in macro-scope, a transition probability matrix is calculated to define the nodes movement from region to region. Instead of simply dividing the area into squares, we cluster the squares with higher taxi dense quantity and define two types of regions – dense region and sparse region. Simulations are carried out to display the trace generated by our model to compare the similarity of generated trace features and contact characteristics. In order to verify our assumptions that the taxi statuses affect the accuracy of models, a simplified movement model S-START based on START is implemented, but it ignores the taxi statuses. The relevance of START, S-START, ShortestPathMovement (based on the map of Beijing) and Random Way Point model with real trace are compared. The simulation results illustrate that the taxi statuses and geographic characteristics have obvious influence on the accuracy of models and the proposed mobility model has a good approximation with reality.

keyword

mobility model, taxi status, Markov process, dense region, sparse region

1 Introduction

In vehicle ad hoc networks (VANETs) [17], realistic movement model is an important way to improve route planning, control traffic situations, or solve the vehicle-to-vehicle communication problems. However, movement models will influence simulation performance, since movement model defines the nodal movement pattern including speed and direction. It is necessary to work on realistic mobility models.There are some obstacles to create realistic mobility. Firstly, it is difficult to utilize large amount of data directly, because simulation scenarios are changeable. Some researchers [10, 8] modeled the vehicular movement, extracting different feathers from real data sets. But in previous works, it is ignored that taxi behavior changes with taxi status.

In our work, a STatus and Region Aware Taxi mobility model, START, is proposed based on the real taxi GPS data, which involves 12,455 taxis in Beijing, China and 74,175,360 records from March 3rd, 2011 to March 7th,2011. Four taxi statuses from 0 to 3 are given by the data set, status 0 indicates a taxi is vacant and status 1 refers that a taxi is occupied. The other two statuses(defense status and stop running status) will not be discussed in this paper, because the amount of data is small and the behavior characteristics are not certain. Two assumptions, one assumes that the taxi behavior differ with the taxi status and the other assumes that taxi movement has geographic features, are proposed in section 3. They are validated to be reasonable by the statistical analysis of the data set. The mobility model is developed on microscopic and macroscopic aspects. For microscope, the *speed* and *duration* for the two taxi statuses are discussed respectively. For an instance, *speed* distribution for vacant status (status 0) indicates the probability for a nodes running at a certain speed, so that the speed of nodes can be assigned according to this rule. In the macro scope, instead of simply dividing the area into squares, we cluster the area according to the node density. By

dividing entire area into grids, cells adjacent to each other and with higher node density are grouped into dense regions while other girds are classified into one sparse region. Then the transition probability between regions are calculated, so that the macroscopic movement can be defined. Simulations are carried out to compare the similarity of node trace characteristics and contact characteristics. In order to estimate the assumption that the taxi statuses cannot be ignored, a simplified model S-START based on START is implemented, which ignore the taxi status difference. The performance of START, S-START, Shortest Path movement model and Random Way Point model are compared with that of real trace data. The results show that the taxi status have obvious effect on the taxi behavior and further influence the simulation results. Our mobility model has a good approximation with the real scenario.

The rest of our paper is organized as follows: Section 2 provides an overview of related works on mobility models. Section 3 proposes two assumptions which are further validated by statistical results of real data. Section 4 presents the modeling process. Simulation results are demonstrated in Section 5. Finally, Section 6 concludes this paper.

2 Related Works

In recent years, vehicular ad hoc networks (VANETs) have drawn growing interest. Mobility model is crucial to simplify scenario and capture the main characteristics of VANETs by simulation, even though real data sets are available. We briefly review the mobility models below. Based on the restrictions of environmental attributes, mobility model can be classified into free space and constrained models[11, 1], while the restrict models can also be further categorized into models restricted by geographic structure, trace data, or both. 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. Upon reaching the destination, the node pauses again for a pause time, selects another destination, and repeats the previous procedure again [3] . Amit Kumar Saha, el at. [16] 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 have a close relevance to the real-world movement. Literature [14] demonstrated that graph structure is close related with inter-contact time distribution in both random and social mobility on grid-based graphs. Some models [16, 15, 8, 13, 4] take the geographic structure into consideration. Manhattan model (MM) 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 [13]. Downtown models add traffic density to the manhattan models to increase the accuracy in simulation. In 2007, Atulya Mahajan, et al. [12] 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.[4] developed their movement model based on a realistic vehicular traffic model on road defined by real street map data.

With the development in Vehicular communication equipment, large amount of data can be collected and utilized by researchers. Many mobility models take advantage of the trace data to achieve better accuracy and performance. In 2007, Zhang X, Kurose J, Levine B N, et al. [18] analyzed the bus trace taken from UMass DieselNet, consisting of Wi-Fi nodes attached to buses. They found that inter-contact times aggregated at a route level exhibit periodic behavior. Thus, they construct generative route-level models that capture the above behavior. However, the mobility model based on bus data with periodic behavior may lack in its universality in VANET.

Other researchers considered the graph factor and trace comprehensively. In 2006, by extracting wireless network traces required from 10,000 users at Dartmouth College over several years, Kim, al et. [10] developed a movement model based on their discoveries that the speed and pause time follow a log-normal distribution and the direction of movements closely reflects the direction of roads and walkways [10]. While the user data can reflect less mobility, we cannot apply it to VANET directly. In 2010 and 2012, Huang H, et al. [8, 7] 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. Whereas the mobility models existing have captured many characteristics in VANETs, the phenomenon that the attributes of vehicles, i.e., speed, turn probability, varies a lot from area to area, is ignored, limiting the accuracy of models. To the best of our knowledge, our work is original to develop mobility models by investigating the inequality in geographical distribution.

3 Assumptions and Statical analysis of Taxi Trace

In this section, we mainly focus on statistical analysis on the speed and duration characteristics on the data set. Firstly, the data set will be introduced in section 3.1.

CentreId	CompanyId	TaxiId	Time	Lon	Lat	Lon2	Lat2	Speed	Direction	Status	Event	Altitude
33210	$JYJ	13301104001	20110303000002	116.2887	39.86338	428709111	146956948	0	184	0	4	50
46978	$JYJ	13301104001	20110303000052	116.2887	39.86337	428709182	146956948	0	184	0	4	50
59974	$JYJ	13301104001	20110303000207	116.2887	39.86337	428709208	146956921	0	184	0	4	50
60080	$JYJ	13301104001	20110303000157	116.2887	39.86337	428709206	146956945	0	184	0	4	50
49643	$JYJ	13301104002	20110303000108	116.4492	39.93999	429301182	147239830	6	244	0	4	50
48840	$JYJ	13301104003	20110303000101	116.0196	39.88398	427717410	147033049	0	50	2	4	50
62463	$JYJ	13301104003	20110303000205	116.0196	39.88398	427717429	147033050	0	50	2	4	50
60812	$JYJ	13301104007	20110303000156	116.6839	39.88454	430165372	147034696	0	314	2	4	50
44364	$JYJ	13301104009	20110303000041	116.3647	39.85738	428989806	146935329	0	86	0	4	50
58047	$JYJ	13301104009	20110303000144	116.3647	39.85737	428989828	146935315	0	174	0	4	50
41813	$JYJ	13301104010	20110303000042	116.3572	39.87213	428962239	146989668	0	326	2	4	50

Figure 1: an example of the dataset

Table 1: Explanation of Events

Event	Explanation
0	A taxi's status change to vacant.
1	A taxi's status change to occupied.
2	Set up defense.
3	Cancel defense.
4	No event happened.

Table 2: Meanings of Statuses

Status	Explanation
0	A taxi is vacant.
1	A taxi is occupied.
2	A taxi is setting up defense.
3	Stop running.

Then, two assumptions are proposed and validated in the following section 3.2.

3.1 Trace Dataset: Beijing Taxi Traces

A real-world GPS data set is used in this paper, which was generated by 12,455 taxis in Beijing, China within 5 days from March 3rd,2011 to March 7th,2011. This data set has been used by several key researches and application programs of Intelligent Transportation Systems (ITS) in Beijing, China. The number of participated taxis (12,455 taxi node) is 18% of the total taxis in the city. Each taxi is equipped with a GPS device and upload its information (including location, speed, direction) about every 60 seconds. There are about 1.22×10^8 records in total. Figure 1 shows a sample of the data set. Each row includes a base station ID, company name, taxi ID (id), timestamp (t), current location (l, including longitude and latitude), location of 54 format, speed, acceleration, status of the taxi, event, and height. Of all the fields, taxi ID, timestamp, and current location, status and Event are used in this paper. Note that GPS traces from taxis have been used recently for inferring human mobility [5] and modeling city-scale traffics [2]. Therefore, we believe that they are suitable to characterize the contact patterns among vehicles in large-scale urban scenario.

Especially, there are five types of events and four types of status. The explanations are as table 1 and 2. We only discuss the status 0 and 1 in this paper.

3.2 Assumptions and Statistical Analysis of Taxi Data

According to experience in daily life, the following two assumptions are given:

Assumption 1 The behavior of a taxi will change when its status changes. When a taxi is taking passengers, its destination is fixed, and the speed of it is relatively faster. On the contrary, when a taxi is vacant, it will slow down or even stop, because drivers need to search for potential passengers along the road. Thus, taxi behavior characteristics, such as speed and the status duration varies consequently.

Assumption 2 Movement behavior associates with geographic feathers. Drivers waiting for passengers tend to some special regions (such as railway station, airport, shopping mall), in order to reduce the waiting time and increase their earnings. Consequently, the distribution of taxi is not uniform.

Therefore, speed and duration distribution law for each status are studied to support *assumption 1*. And the taxi density distribution is investigated to evaluate the *assumption 2*.

3.2.1 Speed analysis

In this section, average speed and speed distribution for each status are investigated.

The average speed will increase if taxis are occupied. It varies from 25.48 to 27.978 km/h when taxis are occupied, while it ranges from 9.977 to 11.272 km/h when taxis are vacant.

To further investigate the speed distribution, proportion for every speed section is calculated, in fig.2. For example, dot(19,0.0245) means 2.45% records fall in the speed range $[19,20)km/h$. Because the proportion for 0km/h reaching up to 20% is too large compared with the proportion of other speed section, we do not show it in figures. Fig. 2 shows that speed distribution differs for status 0 and 1. Most speed records are in the speed section $[0,100]km/h$. For status 0, in the speed section (0,40], it follows a linear distribution. But for status 1, there occurs a peak in the $[8,15]km/h$. Fig. 2 also demonstrates that the speed distribution is with strong regularity for each status.

3.2.2 Status duration analysis

The duration distribution for each status are shown in fig.. Status duration represents the time length of a taxi staying in a certain status. The red line presents the lasting time distribution for status 0, and the green one is for status 1. A dot of the line means the proportion of the

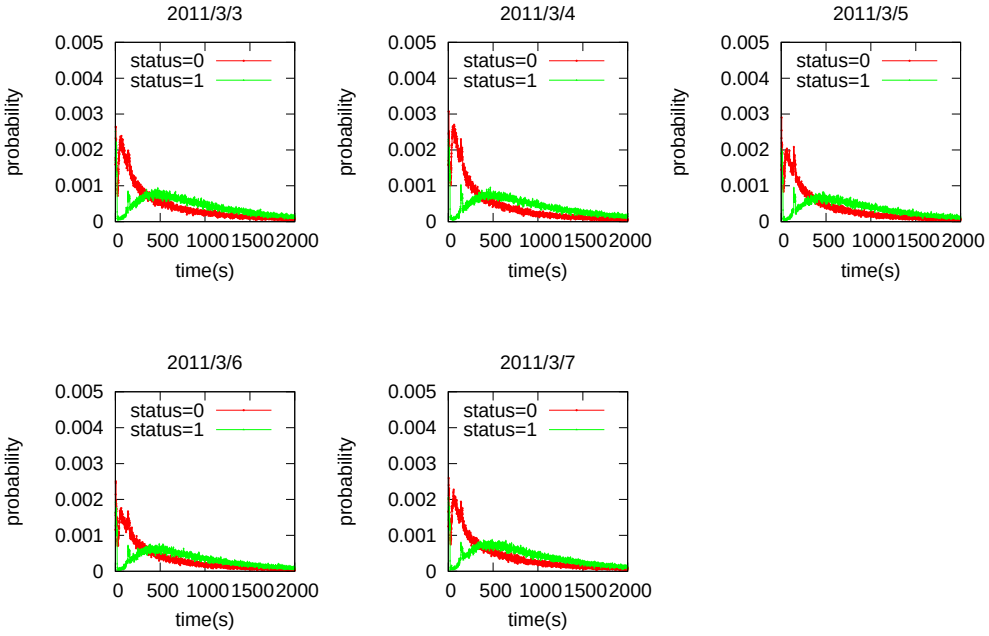


Figure 3: Status duration distribution.

duration. A peak exists in each line, and it is obvious that the peak of the red line is earlier than that of the other line. And the value of duration for status 0 trends to be smaller. It accords to the realistic situation, because drivers tend to shorten the waiting time to raise their income.

The statistical results are consistent with the *assumption 1*, that is, the behaviors of taxis are similar within each status while it differs between the two statuses.

3.2.3 Taxi density distribution

To validate *assumption*2, we quantitatively analyze vehicles density in one hour. By dividing the whole network into $100 * 100$ grids, taxi density distributions for event 0, 1 and 4 are computed in each cell. As shown in table 1, event 4 means no event, that is, a taxi uploads its GPS information every 60 minutes and set the event as 4. As fig.4.(d), (e), (f), it is obvious that the taxi density for event 4 is higher than for other events reaching up to $600 vehicles/cell$. The taxi density of cells on roads stayed in a relatively high level. Nevertheless, for event 1, the taxi density of most cells is less than 5, which means that less than 5 taxis during the time period took passengers in those locations, while the high dense cells occurred sparsely. By further exploring some high dense cells, it can be found that hot cells for event 0 and 1 include Beijing Train station, the junction of the south 4th Ring Road and subway line one in Beijing, China.

The taxi density distributions for event 1 from March 4th to March 6th are plot in fig.5. The crest value is about 50. And the position of peaks are similar. Fig.5.(c) demonstrates some difference. It may be caused by that it is Sunday and many people should return to the city and prepare for work next day.

The amount of loading passengers in each cell shows geographic feathers and the distribution is uneven, which support the *assumption*2.

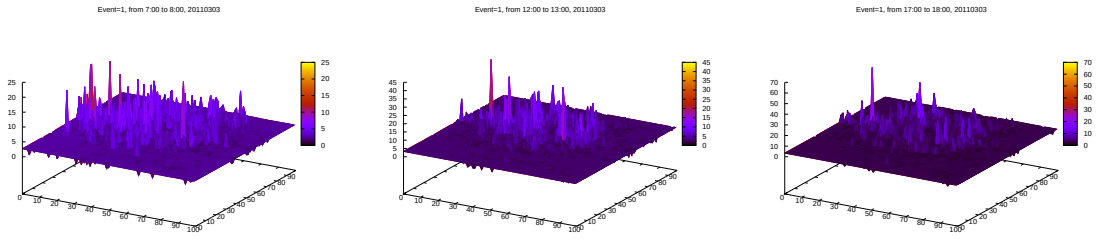
4 Modeling

The modeling process and parameter estimation of S-TART movement model will be introduced in this section.

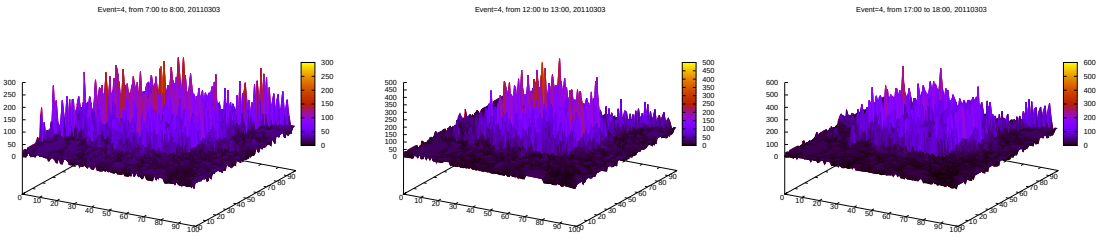
4.1 Modeling process

START is modeling in both microscopic and macroscopic aspect. In the microscopic aspect, we model the taxi speed and duration for status 0 and 1 respectively. During the period for status 0(or 1), nodes's speed will follow the speed distribution of status 0(or 1) and stay in this status for a certain time, which will be assigned by the distribution of lasting time for status 0(or 1). In this case, we abstract the different taxi behavior of the two state.

In the macroscopic aspect, researchers [7] have explored the travel pattern from one region to another. In

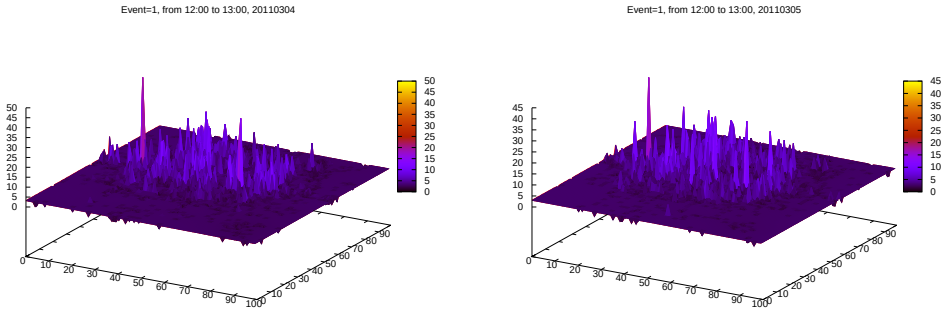


(a) *event* $= 1, 7:00-8:00$ (b) *event* $= 1, 12:00-13:00$ (c) *event* $= 1, 17:00-18:00$

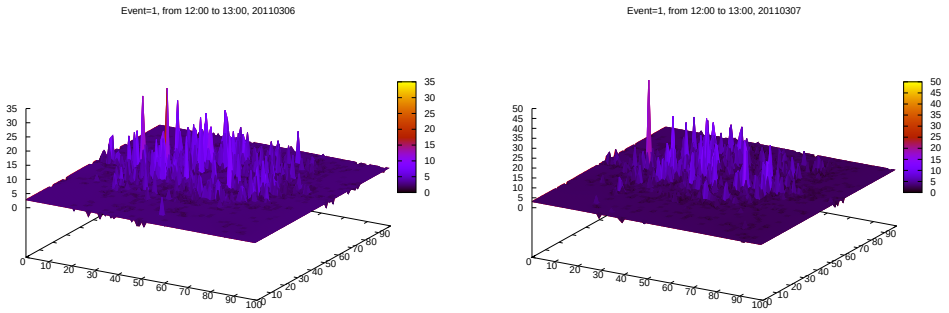


(d) *event* $= 4, 7:00-8:00$ (e) *event* $= 4, 12:00-13:00$ (f) *event* $= 4, 17:00-18:00$

Figure 4: Taxi density for event 1 and 4 in one hour

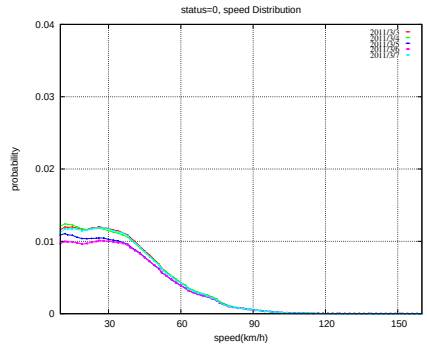


(a) *event* $= 1, 12:00-13:00, 2011/03/04$ (b) *event* $= 1, 12:00-13:00, 2011/03/05$

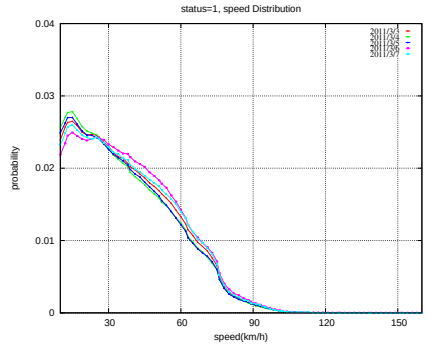


(c) *event* $= 1, 12:00-13:00, 2011/03/06$ (d) *event* $= 1, 12:00-13:00, 2011/03/06$

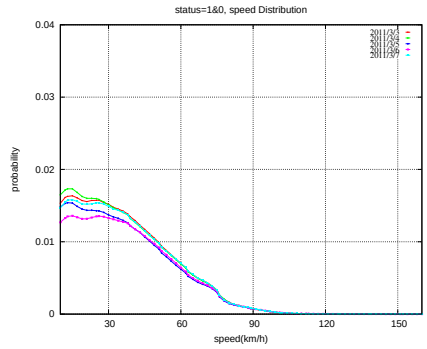
Figure 5: Taxi density for event 1 in one hour from 03/04 to 03/07, 2011



(a) status 0



(b) status 1



(c) status 1

Figure 2: Speed Distribution for status 0 and 1.Here, red line presents the speed distribution for status 0, and the green one is for status 1. A dot of the line means the proportion of the speed. For example, dot(19,0.0245) means 2.45% records fall in the range $[19,20)km/h$.

that case, we can assign the travel origination, path and destination based on the regularity instead of assign them randomly. Instead of simply dividing areas into $n*m$ cells. In this work, we think taxi drivers prefers to go to some hot regions, such as train station or film theaters. So we divide whole area into cells and define taxi dense areas and sparse area as follows:

We define following concepts:

Cell the area is divided into m*n grids, which can be notated as $(x,y,length_{left},length_{top},length_{right},length_{bottom});0 \leqq x < m, 0 \leqq y < n$. x and y are the index of cells.

Region We define region as a collection of cells,that is $\{cell\}$.

Event density for event 1 event density is the number of event 1 happened in certain cell during a period.

Event density threshold to define "higher event density", a threshold is set. A cell is with higher event density, if the event density is higher than the threshold.

Dense region Cell sequence adjacent with each other and event density in them is higher than the event density threshold.

Sparse region other cells expect for the cells in dense regions.

As a taxi's trace can be regarded as a Markov process, the One-step transition probability matrix between regions is calculated. A trace can be defined as a collection of path $\{path_i\}$, while $path_i = (< c_0^i, c_1^i, c_2^i, ..., c_{dest}^i >, speed_i)$, c means a coordinate,(x,y). For each node, a initial position will be assigned in recognized dense region. $path_i$ will set c_0^i as its current position and then determine the c_{dest}^i by the area transition probability, to which the node distribution will conform. A node move from c_j^i to c_{j+1}^i by choosing an adjacent cell, defined in section 4.1, with highest area transition probability. After a path is found, node will move towards the destination along the path.

For every path, a speed will randomly assigned by the speed distribution for its status. Specially, if a node cannot reach the destination in its current status duration time, the path will be divide into some sub-pathes with specific speed values. Status duration is also randomly generated following the corresponding status duration distribution.

With the transition probability matrix, the macroscopic movement of nodes can be developed.

4.2 Parameter estimation of speed distribution

In this section, we modeling the speed distribution.Because the probability will be influenced by the length of speed range, so we choose the cumulative distribution for modeling. Then we fit the cumulative distribution to get the cumulative probability distribution function, and then take a derivative with it to obtain the speed probability distribution.

Fig.1 plots the cumulative distribution of speed. From the figure we can get following information:

- In the speed range from about 0 to 40 km/h, the distributions show a linear relationship. While after the range, an exponential relationship can be observed.
- For the status 0, the distributions are similar. But on March.5 and 6, 2011, the curves show more similarity than on the other days. We assume that for March.5 and 6, 2011 are weekend, the speed tend to be smaller than on the weekdays because of the heavy traffic.
- For the status 0, the speed distribution differs with each other evenly.

We choose a segmented function to fit the cumulative distributions and estimated the parameters. Especially, we respectively discuss the distribution for status 0 on workdays and weekend. The fit formulas are given as formulas 1.

$$\begin{cases} g(x) = a*x+b & x \in [0,40] \\ f(x) = 1-exp(-c*x-d) & x \in (40,180] \end{cases} \quad (1)$$

Fig. 6 also plots the fit results. For status 0, we discuss the condition for that on workdays and weekend. The blue lines represent the fitting results for the speed range $[0,40]km/h$. And the red lines plot the fitting results for speed range $(40,180]km/h$. We also fit the cumulative speed distribution for status 0 and 1, shown as the right bottom in fig.6.

The rms of residuals for each fit are as table 3.The smaller rms of residuals means better fitting. In the table, the values are all less than 0.01, showing good similarity.

4.3 Parameters estimation of status duration distribution

Similar to the modeling process of speed, we explore the cumulative distribution of lasting time. Then we fit the cumulative distribution to get the cumulative distribution formular and its estimated parameters. The fit formulas are given as formulae 2, and the fitting results are shown as fig. 7.

Table 3: The rms of residuals of fitting curves

Categories	rms of residuals
$g^{workday}_{status=0}(x)[0,40]$	0.00272264
$g^{weekend}_{status=0}(x)[0,40]$	0.00386982
$f_{status=0}(x)(40,180]$	0.00148225
$g(x)_{status=1}[0,40]$	0.0176819
$f(x)_{status=1}(40,180]$	0.00760913
$g(x)_{status=0,1}[0,40]$	0.0160319
$f(x)_{status=0,1}[40,180]$	0.00299414

$$\phi(x) = 1-exp(-e*x) \quad x \in [0,43200] \quad (2)$$

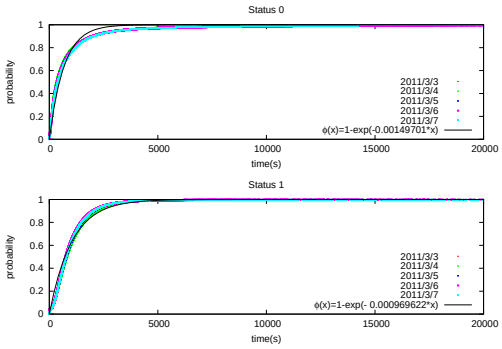


Figure 7: The fit result of the cumulative status duration distributions.

The fitting result for status 0 is $\phi(x)_0 = 1-exp(-0.00149701*x)$, and the rms of residuals is 0.0347105. The fitting result for status 1 is $\phi(x)_1 = 1-exp(-0.000969622*x)$ and the rms of residuals is 0.0337265.

4.4 Region transition probability

Based on the definition of dense region and sparse region are recognized. By dividing the region within the Fifth Ring Road, Beijing, China, about $750km^2$, into $100*100$grids and utilizing the GPS records from March 3,2011 to March 7,2011, regions are recognized. Every cell is regarded as square with length about $274m$. Since area size will affect the area transition probability,a single dense area cannot be larger than 100 cells. The node dense threshold is set as 121, which is the average node number in the top 5000 dense grids. Shown as fig.8,156 dense areas are recognized and marked by different colors, and the white area is regarded as the sparse area.

A car's trajectory can be regarded as a markov process, and the markov chain nodes are the areas in the city. In

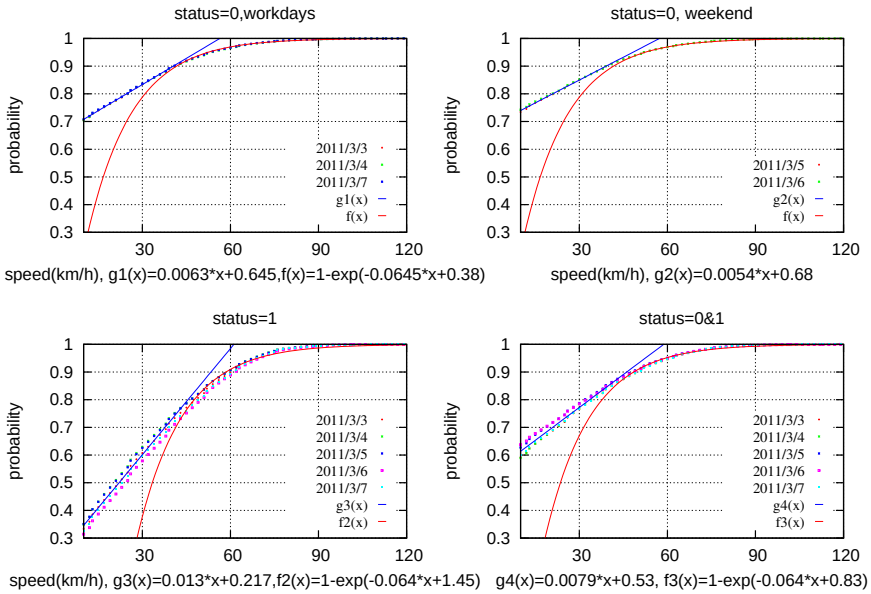


Figure 6: The fit result of the cumulative speed distributions.

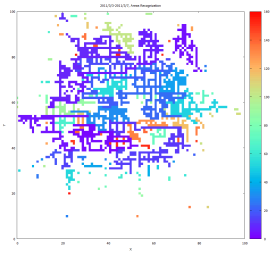


Figure 8: Dense area recognition. A dense area is considered as the grids connecting with each other and of same color. The white part is the sparse region. There are 156 dense regions here.

this section, we will introduces how to adopt step transition probability matrix from the vehicles' real trajectory data set.

After region recognition, n regions including dense regions and a sparse region, can be adopted and marked as A_i. For every short time $\triangle t$, a snapshot of the trajectory is token, which indicates the position information and their area id at time t, So that the taxi set for A_i at time t can be found, donated as $TaxiSet_i^{(t)}$. All the taxi set for k time slots and n areas can be presented as a matrix as matrix (5) .

$$\begin{pmatrix} TaxiSet_1^{(1)} & TaxiSet_1^{(2)} & \cdots & TaxiSet_1^{(n)} \\ TaxiSet_2^{(1)} & TaxiSet_2^{(2)} & \cdots & TaxiSet_2^{(n)} \\ \cdots & \cdots & \cdots & \cdots \\ TaxiSet_k^{(1)} & TaxiSet_k^{(2)} & \cdots & TaxiSet_k^{(n)} \end{pmatrix} \quad (3)$$

The One-step transition probability from area i to area$j$, donated as p_{ij}, can be calculated by the matrix (5). $|TaxiSet_i^{(t)}|$ means the taxi number in area i at time t and $TaxiSet_i^{(t)} \bigcap TaxiSet_j^{(t+1)}$ is the node number moving from area i to area j in one step$(1 \leq m \leq k-1, m \in N^+)$. In that case, p_{ij} during k time slots can be calculated by the formula (4):

$$\begin{aligned} p_{ij} &= \frac{|TaxiSet_i^{(1)} \bigcap TaxiSet_j^{(2)}| + \cdots + |TaxiSet_i^{(k-1)} \bigcap TaxiSet_j^{(k)}|}{|TaxiSet_i^{(1)}| + |TaxiSet_i^{(2)}| + \cdots + |TaxiSet_i^{(k-1)}|} \\ &= \frac{\sum_{m=1}^{k-1} |TaxiSet_i^{(m)} \bigcap TaxiSet_j^{(m)}|}{\sum_{m=1}^{k-1} |TaxiSet_i^{(m)}|} \end{aligned} \quad (4)$$

Note that , $\sum_{j=1}^{n} p_{ij} = 1$ and $p_{ij} \neq p_{ji}$. Then, One-step transition probability matrix is donated as (5) :

$$\boldsymbol{P} = (p_{ij})_{n*n} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p1n \\ p_{21} & p_{22} & \cdots & pnn \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & pnn \end{pmatrix} \quad (5)$$

5 Model Verification

In this section, START mobility model is validated on the aspects of node distribution and contact characteristics. All movement models are implemented on Opportunistic Networking Environment (ONE)[9].

In order to validate that the behavior difference for each status will affect the accuracy of mobility models, a simplified model, S-START, is developed based on

START by modifying the speed generator according the speed distribution $g_{status=0,1}(x)$ and $f_{status=0,1}(x)$. Because status difference is neglected, there is no need to discuss the status duration.

Shortest Path movement model based on the map in Beijing is an other comparison, which is implemented by ONE. It moves will find shortest path from source to destination by Dijkstra algorithm. The RWP model is another comparison, because it is proved to be an efficient model modeling the nodal movement in VANETs. It's simple and effective, but takes no consideration of the node statuses and geographical distribution.

The START, S-START, Shortest Path and RWP mobility model are compared with the real trace of Beijing, China. In simulations, Node number is set as 4000 and scenario in area $24445 * 23785m^2$ (a sub-map of the whole area), including fourth ring roads in Beijing. The simulation time is three hours and the warm up time for reports is one hour, so that the nodal movement and position will not be affected by its initial position. The communication range is $200m$.

5.1 Traces and distribution of nodes

Trace samples and their snapshots are demonstrated in this section, shown as fig. 9 and fig.10.

From fig.10, Real trace and Shortest Path movement model exhibit the road structure, while START and S-START display the geographic feathers defined in section 4.1. However, the node distribution of RWP is much uniform.

By dividing areas into $10 * 10$ grids, (length of each cell is about 2400 meters), the node density distributions are investigated, shown in fig.11. An interpolation process is conducted on RWP traces, because it will not generate a position data until it changes its direction. Under this condition, middle point in the straight line created by two original points of continues time stamp is inserted for every 5 seconds. In each cell of grids in 20 seconds, the distinct nodes occurring are counted, that is, if a node prints its location in a cell twice, it will not add up to the node density in this square. Consequently, the interpolation process on RWP trace will not influence the results. Moreover, the upper limit of speed is 33.3 m/s, so that for every node, it can occur in no more than 3 cells in 20 seconds. In table. 4, the average of node densities for models and real traces are close. Whereas, the variances change a lot from 4848 of RWP to 107423.3 of S-TART. The variance of Real trace is large, because nodes in reality during a short time is not evenly distributed. Although, Shortest Path movement model takes the geographic characteristics into consideration, the difference from different road segments is disregarded.

Table 4: Average and variance of node density

type	average	variance
Real Trace	41.39175	107401.1
START	44.28421	107423.3
S-START	45.3	86084.9
ShortestPath	43.41414	52820
RWP	43.2	4848

5.2 Contacts characteristics

The contacts (connections) and inter contact time (ICT) [6] are evaluated as the indicator to compare the similarity. During simulation time, there are 1744093 contacts in real trace scenario, 1808621 contacts of START, 1271044 contacts of S-START model and 1007756 contacts of RWP model.

The contact frequency proportion distribution is shown as in fig.12. In fig. 12. (a), the contact times of a node vs. the node frequency is illustrated. For an instance, a point (500,44) of the real trace presents that there are 44 nodes contacts 100 times with other nodes. For real trace $(740, 52)$, most nodes (740nodes) contact 52 times. The peak occurs in $(83, 880)$ for START, $()$ for S-START, $(1370, 41)$ for ShortestPath movement model and $(134, 500)$ for RWP.

Fig. 12. (b) demonstrate the neighbors of nodes. If a $node_i$ contacts with a $node_j$, $node_i$ and $node_j$ are neighbors. Based on the trace data set, 709 nodes are with about 8 neighbors. START shows most similarity with real trace with a peak of $(7, 709)$. For ShortestPath movement model, 1875 nodes have 4 neighbors and For RWP, 2448 nodes contact with other 2 nodes. If the node distribution is even, the peak occurs early, because most nodes are similar. On the contrary, unevenly movement of nodes will cause that some nodes contact with many other nodes but some nodes only contact with a few neighbors.

ICT is also widely used in VANETs to forecast contacts and assist routing decision. The cumulative ICT distributions are further explored, shown in fig.12(c). From 0 to 120 sec, the cumulative probability of ICT increases rapidly for the real trace and START. For the other two models, the growth rate varies slightly. In reality, taxies can contact twice in a short time, because the geographical distance are quite close after the first contact. In that case, the ICT tender to be short, causing the cumulative probability of ICT increases rapidly.

To conclude, by comparing the node distribution and contact characteristics, the START mobility model performs good similarity with the real data set. Although the S-START model only ignores the statuses by modifying the parameters of the speed generator, the statistical results show obvious dissimilitude from S-START and the reality. By comparing with Shortest Path movement

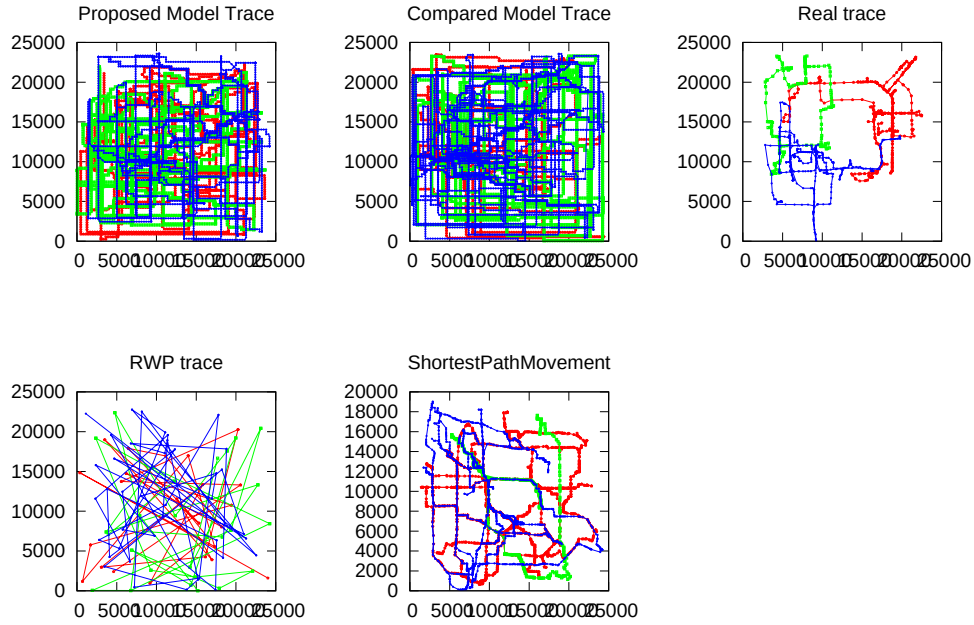


Figure 9: Trace samples

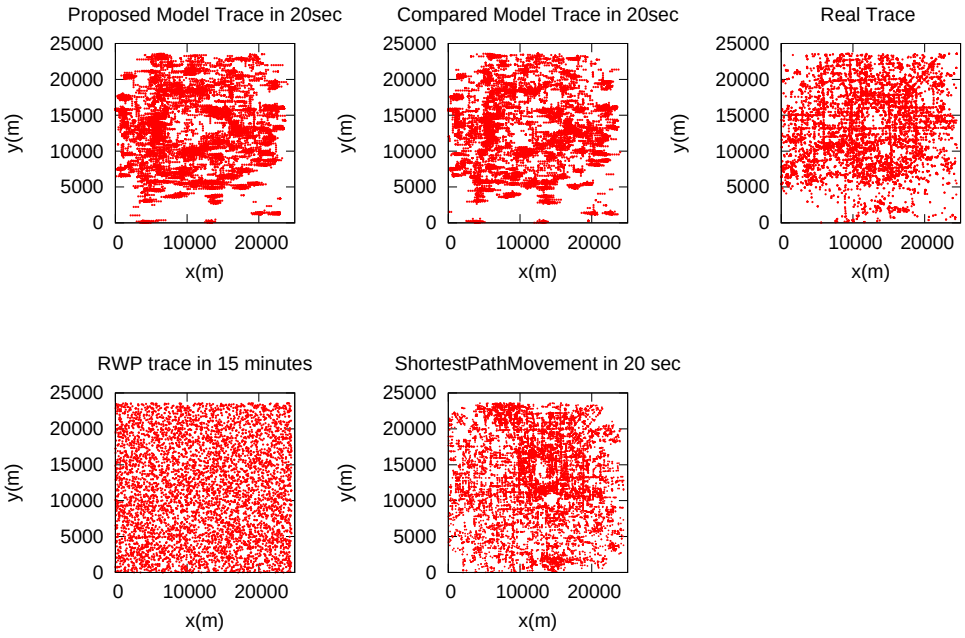
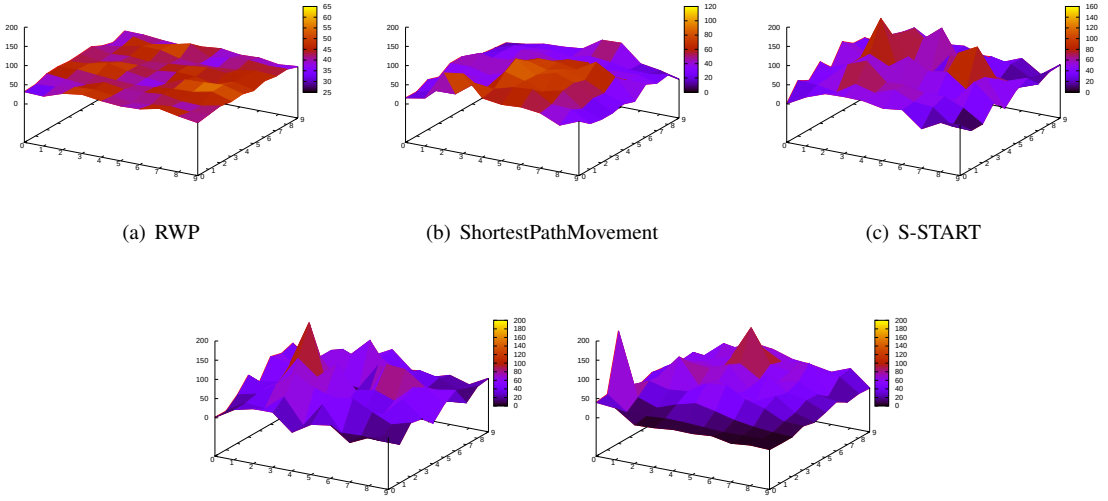


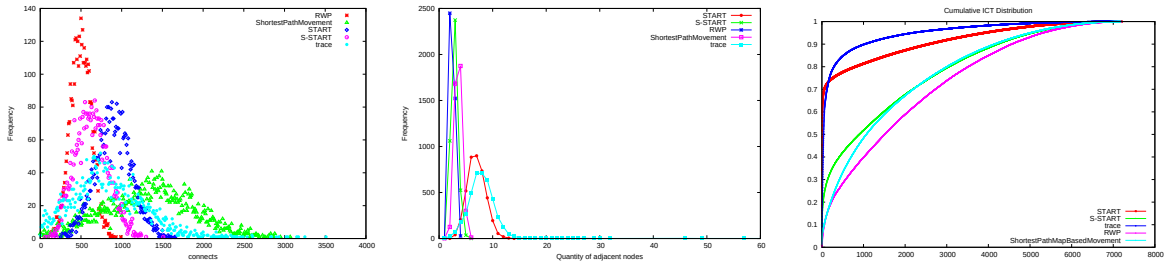
Figure 10: Nodes distribution snapshoots



(a) RWP (b) ShortestPathMovement (c) S-START

(d) START (e) Trace

Figure 11: Node density in 20 seconds



(a) Contact times vs. frequency (b) neighbors vs. frequency (c) Cumulative ICT distribution

Figure 12: Contact times distribution and Cumulative ICT distribution

model, the *assumption*2 that the unevenly distribution of taxies influences the accuracy of models can be supported in certain degree.

6 Conclusion

Since the mobility model is essential for mobile network, a novel mobility model START based on real GPS trace data is proposed. By assuming the taxi behavior is related with its statuses and taxi distribution in area is uneven, statistical experiments are conducted to verify those assumptions using the real trace data. Further, its parameter–speed and duration in each status are estimated respectively, and the one step transition probability matrix is calculated. In this case, macroscopic (from one region to another) movement and microscopic movement(speed and duration for each status) can be defined. Finally, the START is implemented on ONE simulator and estimate it by comparing with the real trace, RWP and Shortest Path movement Model. A simplified model, S-START, based on START is created, which modifies the parameters of speed distribution, and sets the parameters according to the speed distribution of both status 0 and 1. By comparing the START and the S-START, it can come to the conclusion that the taxi behavior in different status influences the node distribution and contact characteristics. In return, it proves our *assumption*1.

Shortest Path movement Model, based on the real map of Beijing and the Dijkstra algorithm, reflects the road structure of the city better than START. However, comparing the node distribution and contact feathers, S-TART shows better performance. Shortest Path movement model takes geographic feathers into consideration, but it neglects uneven geographic distribution. RWP also disregards this feather. Those simulation results above validate the *assumption*2 in certain degree.

Simulation results demonstrate that START has a good approximation with reality.

7 Acknowledgment

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 and No. 61170295, the Project of Aeronautical Science Foundation of China under Grant No.2013ZC51026 and No.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.

References

[1] AHMED, S., KARMAKAR, G. C., AND KAMRUZZAMAN, J. An environment-aware mobility model for wireless ad hoc network. *Computer Networks 54*, 9 (2010), 1470–1489.

[2] ASLAM, J., LIM, S., PAN, X., AND RUS, D. City-scale traffic estimation from a roving sensor network. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems* (New York, NY, USA, 2012), SenSys '12, ACM, pp. 141–154.

[3] BROCH, J., MALTZ, D. A., JOHNSON, D. B., HU, Y.-C., AND JETCHEVA, J. A performance comparison of multi-hop wireless ad hoc network routing protocols. In *Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking* (1998), ACM, pp. 85–97.

[4] CHOFFNES, D. R., AND BUSTAMANTE, F. A. N. E. An integrated mobility and traffic model for vehicular wireless networks. In *Proceedings of the 2nd ACM international workshop on Vehicular ad hoc networks* (2005), pp. 69–78.

[5] GANTI, R., SRIVATSA, M., RANGANATHAN, A., AND HAN, J. Inferring human mobility patterns from taxicab location traces. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (New York, NY, USA, 2013), UbiComp '13, ACM, pp. 459–468.

[6] HU, Y., WANG, H., XIA, C., LI, W., AND YANG, Y. On the distribution of inter contact time for DTNs. In *Local Computer Networks (LCN), 2012 IEEE 37th Conference on* (2012), IEEE, pp. 152–155.

[7] HUANG, H., ZHANG, D., ZHU, Y., LI, M., AND WU, M.-Y. A metropolitan taxi mobility model from real gps traces. *J. UCS 18*, 9 (2012), 1072–1092.

[8] HUANG, H., ZHU, Y., LI, X., LI, M., AND WU, M.-Y. Meta: A mobility model of metropolitan taxis extracted from gps traces. In *Wireless Communications and Networking Conference (WCNC), 2010 IEEE* (2010), pp. 1–6.

[9] KERAEN, A., OTT, J., AND KARKKAINEN, T. The ONE simulator for DTN protocol evaluation. In *Proceedings of the 2nd International Conference on Simulation Tools and Techniques* (2009), ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), p. 55.

[10] KIM, M., KOTZ, D., AND KIM, S. Extracting a mobility model from real user traces. In *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings* (2006), vol. 6, pp. 1–13.

[11] LU, X., CHEN, Y.-C., LEUNG, I., XIONG, Z., AND LI O, P. A novel mobility model from a heterogeneous military manet trace. In *Ad-hoc, Mobile and Wireless Networks*, Ad-hoc, Mobile and Wireless Networks. Springer, 2008, pp. 463–474.

[12] MAHAJAN, A., POTNIS, N., GOPALAN, K., AND WANG, A. Modeling vanet deployment in urban settings. In *Proceedings of the 10th ACM Symposium on Modeling, Analysis, and Simulation of Wireless and Mobile Systems* (2007), pp. 151–158.

[13] MARTINEZ, F. J., CANO, J.-C., CALAFATE, C. T., AND MANZONI, P. Citymob: a mobility model pattern generator for vanets. In *Communications Workshops, 2008. ICC Workshops '08. IEEE International Conference on* (2008), pp. 370–374.

[14] MAYER, C. P., AND WALDHORST, O. P. On the impact of graph structure on mobility in opportunistic mobile networks. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on* (2011), pp. 882–887.

[15] PENG, W., DONG, G., YANG, K., SU, J., AND WU, J. A random road network model for mobility modeling in mobile delay-tolerant networks. In *Mobile Ad-hoc and Sensor Networks (MSN), 2012 Eighth International Conference on* (2012), pp. 140–146.

[16] SAHA, A. K., AND JOHNSON, D. B. Modeling mobility for vehicular ad-hoc networks. In *Proceedings of the 1st ACM International Workshop on Vehicular Ad Hoc Networks* (2004), pp. 91–92.

[17] YOUSEFI, S., MOUSAVI, M., AND FATHY, M. Vehicular ad hoc networks (vanets): Challenges and perspectives. In *ITS Telecommunications Proceedings, 2006 6th International Conference on* (June 2006), pp. 761–766.

[18] ZHANG, X., KUROSE, J., LEVINE, B. N., TOWSLEY, D., AND ZHANG, H. Study of a bus-based disruption-tolerant network: mobility modeling and impact on routing. In *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking* (2007), pp. 195–206.