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

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 differences between different taxi statuses have been 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, dense region, sparse region

I. Introduction

In vehicle ad hoc networks (VANETs) [**?**], 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 [**?**], [**?**] modeled the vehicular movement, extracting different feathers from real data sets. But taxi status is ignored in the previous works.

In this 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, taxi status 0 (*vacant status*) and status 1 (*occupied status*) are considered. The other two statuses(defense status and stop running status) will not be discussed in this paper, because the amounts 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

*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.

in section III. 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 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 II provides an overview of related works on mobility models. Section III proposes two assumptions which are further validated by statistical results of real data. Section IV presents the modeling process. Simulation results are demonstrated in Section V. Finally, Section VI concludes this paper.

II. Related Works

relatied works are not ready.

III. 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 III-A. Then, two assumptions are proposed and validated in the following section III-B.

A. 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.

TABLE I
EXPLANATION OF EVENTS AND STATUS

Event	Explanation
0	A taxi's status change to vacant, called off event.
1	A taxi's status change to occupied, called on event.
2	Set up defense.
3	Cancel defense.
4	No event happened.
Status	**Explanation**
0	A taxi is vacant.
1	A taxi is occupied.
2	A taxi is setting up defense.
3	Stop running.

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. 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 [**?**] and modeling city-scale traffics [**?**]. 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 I and **??**. We only discuss the vacant status and occupied status in this paper.

B. Assumptions

According to the 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. In contrast, when a taxi is vacant, it will slow down or even stop to search for potential passengers along the road. Thus, taxi behavior characteristics, such as speed and the status duration varies consequently.
- **Assumption 2:** The movement behavior of taxies associates with geographic feathers.
  1) The destination selection will be influenced by different regions.
  2) events occurs in different regions un-evenly, passenger off and on events are distinct.

Therefore, we analysis the speed,duration and passenger on/off events distribution to estimate our assumptions.

IV. Modeling

Movement model defines the mobility pattern of nodes, which can be represented as a collection of pathes$Paths :< p_1, p_2, ..., p_n >$, so dose START. A p_i includes two steps, destination selection and moving process from source location to destination.

Destination selection: In START, to select a destination of a node is closely related to note only its current location but also its current status. Dividing the area into regions by the density of passenger on/off events or loading passenger events respectively, two transition probability matrixes are calculated, one is the probability from passenger off event regions $\{REGION_{m,off}\}$ to passenger on events regions $\{REGION_{n,on}\}$. Note that $\bigcup\{REGION_{m,off}\} = \bigcup\{REGION_{n,on}\} = AREA$. If the status of a taxi changes to vacant, its current location determine a $REGION_{i,off}$. Consequently, a destination region in $\{REGION_{n,on}\}$ will be selected by querying the transition matrix from $\{Region_{m,off}\}$ to $\{Region_{n,on}\}$. Then, START will randomly select a map node in the region as the destination. As to the status of a texi changing to Occupied, the destination selection process is similar. During this process, the region transition matrix will be utilized according to the current status.

Moving process: When the source location (current location) and destination location is given, next step is to find a path and set the speed. To simplify, we adopt the Dijkstra algorithm ,which will find a shortest path from source to the destination, to route on map. The speed is assigned by the average speed distribution of the corresponding status.

Based on the design above, we model the movement on the speed, duration and region transition matrix respectively.

A. Region transition probability

A travel path of a taxi can be simplified as a multi-hop process, in which a hop indicates an on/off event happened. Seeing that, we define a *region transition probability* to figure out the probability of the next hop falling in a definite region j from the current region i. Particularly, two successive events are different. Likewise, the region i and j are recognized by different metrics, that is, off or on event distribution. It is more reasonable. For an instance, if the taxi is occupied, the next hop event is the off one. Hence, choosing a target region from a region set divided by off event distribution is more logical.

To calculate the region transition probability, the **region recognition process** should be executed in advance.

Firstly, we divide the area into 100×100 grids, and define cells in it as equation 1. Then, we consider region as adjacent cells as equation 2.

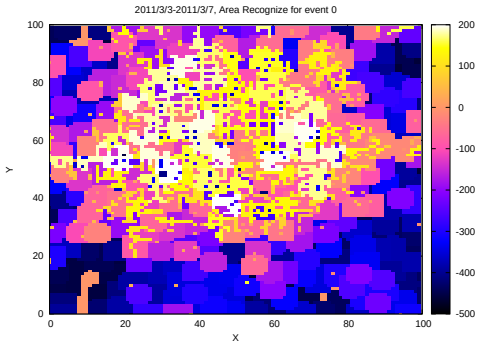
$$CELL_{x,y} ::= \{(lon, lat) | x \leq \frac{lon}{len_x} < x+1, \; y \leq \frac{lat}{len_y} < y+1\} \tag{1}$$

$$REGION_m ::= \{CELL_{x,y} | \exists CELL_{i,j} \in REGION_m \Rightarrow \|x-i\| \leq 1, \|y-j\| \leq 1\} \tag{2}$$

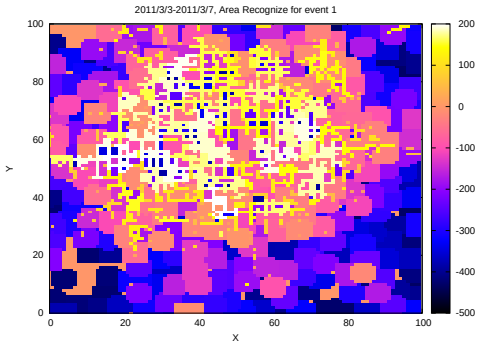
By clustering cells into regions, two set of region $\{REGION_m^{on}\}$ and $\{REGION_n^{off}\}$ can be recognized. The clustering idea is to put adjacent cells with event density larger than the event threshold η into one cluster. To avoid the size of the region become too large or too small, we set a $CLUSTERSIZE$ to restrict a region, say $\|REGION_i\| \leq CLUSTERSIZE$, and only limit the top 200 regions, in which $CELL_{x,y}.events \geq \eta$. After that, the other cells not belong to the top 200 regions, will be cluster into regions, while $\|REGION_j\| \leq CLUSTERSIZE$. We sort the cells by event density in descending order, and begin with the first cell to search is neighbors whether to join the same region using breadth traversal. The region recognition results for on/off events are shown in figure 1. The detail clustering algorithm is presented in the appendix. In addition, the $CLUSTERSIZE = 200$, $\eta = 121$ for on event and $\eta = 141$ for off event set by the average event density of the top 5000 cells order by its event density.

The calculate process of the region transition probability: After clustering cells into regions, the transition probability from $REGION_i^{on/off}$ to $REGION_j^{off/on}$, donated as $p_{i \to j}^{on \to off / off \to on}$. To substantiate, the calculate process of $p_{i \to j}^{on \to off}$ will be introduced in detail. Clearly, the records of $event = on$ in $REGION_i^{on}$ can be required from the data set. the record amount is donated as $\|RECORDS_i^{on} = \{record_{REGION_i^{on}}\}\|$. For $record \in RECORD_i^{on}$, the next event and location can be easily required. Therefore, the record can be associated with the its next hop information to $(taxiid, location_{current}, event, event_{next}, location_{next})$. The $RECORDS_{i \to j}^{on \to off} = \{record | event = on \cap event_{next} = off \cap location_{current} \in REGION_i^{on} \cap location_{next} \in REGION_j^{off}\}$ will be obtain.

$$p_{i \to j}^{on \to off} = \frac{\|RECORDS_{i \to j}^{on \to off}\|}{\|RECORDS_i^{on}\|} \tag{3}$$



(a) passenger off event regions



(b) passenger on event regions

Fig. 1. Region recognition

$$P^{on\to off} = (p_{i\to j}^{on\to off})_{m*n} \tag{4}$$

$$P^{off\to on} = (p_{i\to j}^{off\to on})_{n*m} \tag{5}$$

B. 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.6 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 vacant status, the distributions are similar. But on March.5 and 6, 2011, the curves show more similarity than on the other days.
- For vacant status, the speed distribution differs with each other evenly.

TABLE II
THE RMS OF RESIDUALS OF FITTING CURVES

Categories	rms of residuals
$g_{status=0}^{workday}(x)[0,40]$	0.00272264
$g_{status=0}^{weekend}(x)[0,40]$	0.00386982
$f_{status=0}(x)(40,120]$	0.00148225
$g(x)_{status=1}[0,40]$	0.0176819
$f(x)_{status=1}(40,120]$	0.00760913
$g(x)_{status=0,1}[0,40]$	0.0160319
$f(x)_{status=0,1}[40,120]$	0.00299414

A segmented function is chosen to fit the cumulative distributions and estimated the parameters. Especially, we respectively discuss the distribution for vacant status on workdays and weekend. The fit formulas are given as formulas 6.

$$\begin{cases} g(x) = a*x+b & x\in[0,40] \\ f(x) = 1-exp(-c*x-d) & x\in(40,120] \end{cases} \tag{6}$$

Fig. 2 also plots the fit results. For vacant status, 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 each status, shown as the right bottom in fig.2.

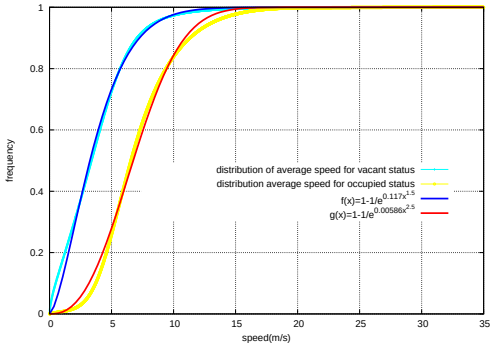


Fig. 2. The fit result of the cumulative speed distributions.

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

V. 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)[**?**].

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 $200m$.

A. Traces and distribution of nodes

Trace samples and their snapshots are demonstrated in this section, shown as fig. 3 and fig.4.

From fig.4, Real trace and Shortest Path movement model exhibit the road structure, while START and S-START display the geographic feathers defined in section **??**. 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.5. 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. III, 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 START. 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 III
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

B. Contacts characteristics

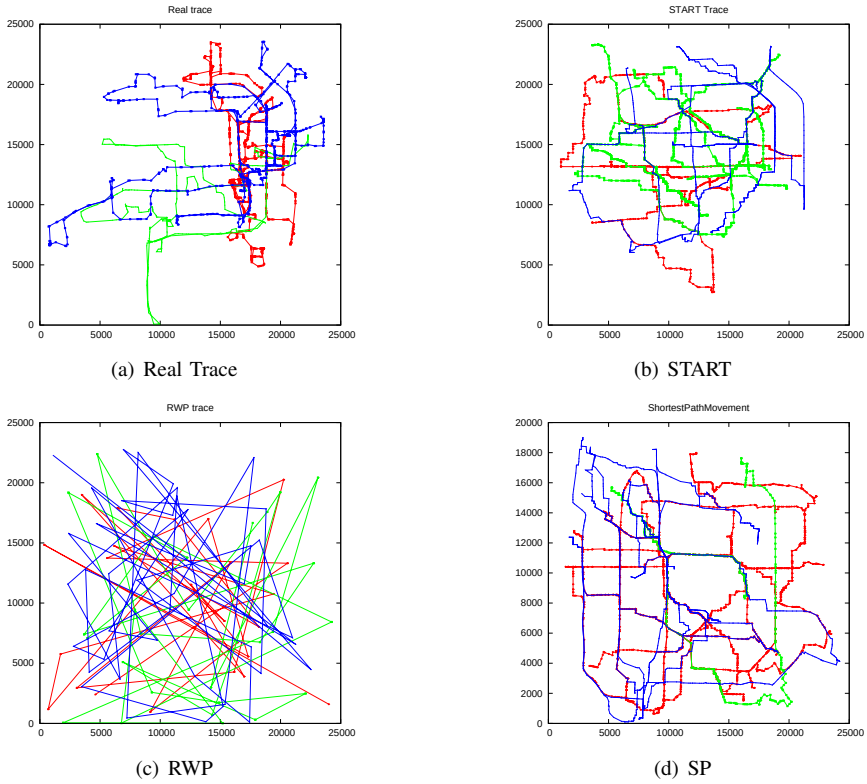
The contacts (connections) and inter contact time (ICT) [**?**] are evaluated as the indicator to compare the similarity. During simulation time, there are 1744093 contacts in real trace scenario, 1808621 contacts of START, 1271044 contacts of S-START model and 1007756 contacts of RWP model.

The contact frequency proportion distribution is shown as in fig.6. In fig. 6. (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. 6. (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.6(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



(a) Real Trace (b) START

(c) RWP (d) SP

Fig. 3. Trace samples

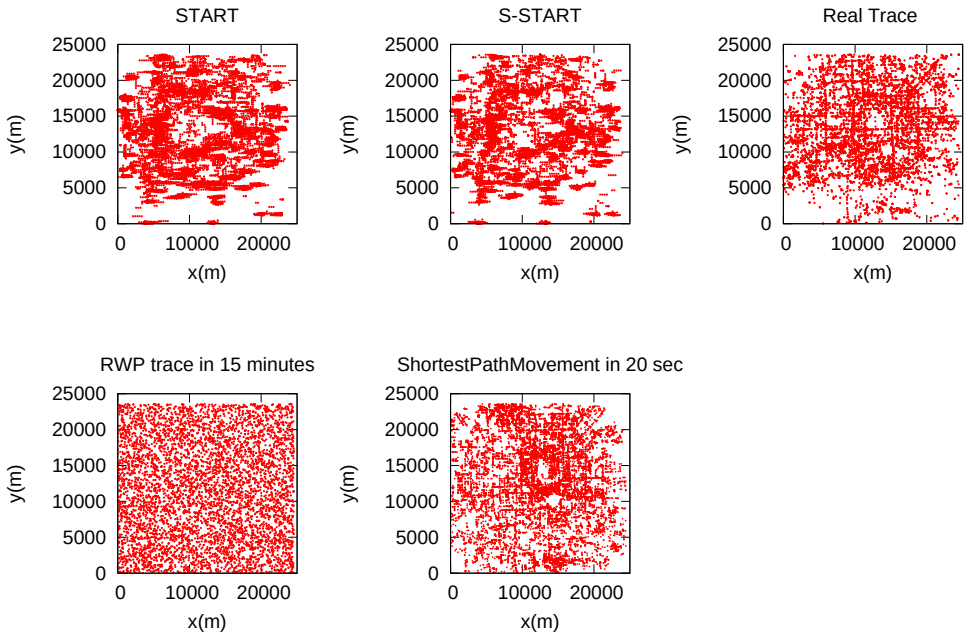
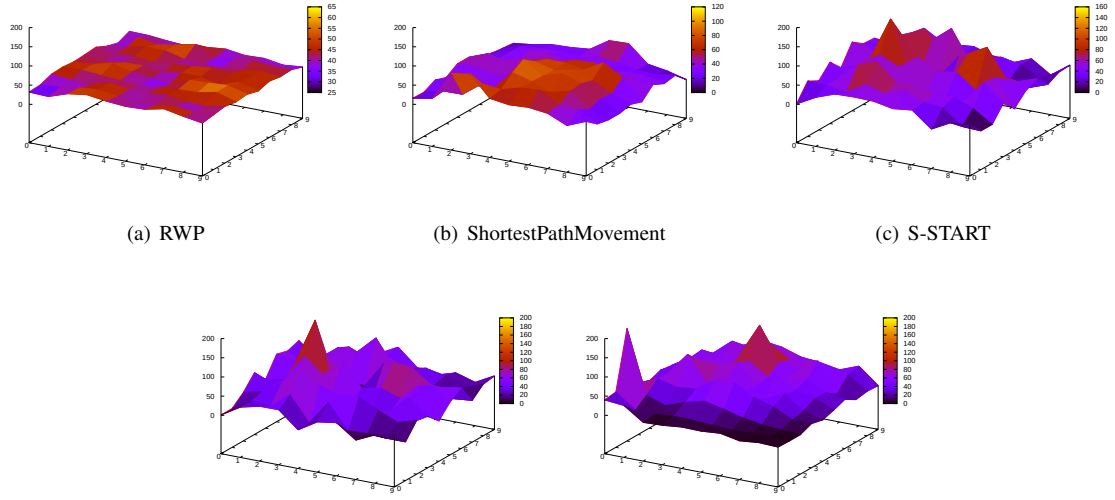


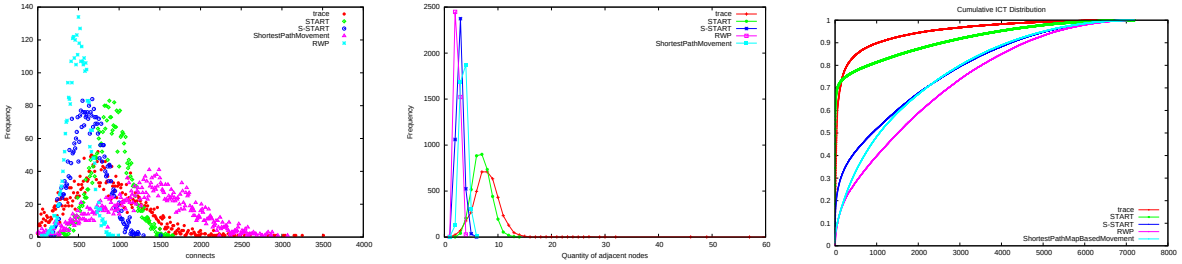
Fig. 4. Nodes distribution snapshoots



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

(d) START (e) Trace

Fig. 5. Node density in 20 seconds



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

Fig. 6. Contact times distribution and Cumulative ICT distribution

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

VI. 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 of each status are estimated respectively, and the region transition probability matrix is calculated. In this case, macroscopic (from one region to another) movement and microscopic movement(speed 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 each status. 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, START 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.

Appendix

Algorithm 1 Clustering

Require: $Cells = \{CELL\}$
  η ##events threshold
  $CLUSTERSCALE$
  $REGIONSEED = 200$
  $ClusterQueue = \varnothing$
  $UsedCells = \varnothing$
  Sort $Cells$ by events DESC
  for $CELL_{x,y} \in Cells$ **do**
    if $CELL_{x,y} \notin UsedCells$ **then**
      $CELL_{x,y}.region = REGIONSEED$
      $size = 1$
      $CLUSTERSEED = CLUSTERSEED - 1$
      $ClusterQueue.enqueue(CELL_{x,y})$
      $UsedCells.add(CELL_{x,y})$
      while $ClusterQueue \neq \varnothing$ **do**
        $CELL_{x,y} = ClusterQueue.dequeue()$
        if $REGIONSEED \geq 0$ **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
          $enqueueNeighborOthers(CELL_{x-1,y})$
          $enqueueNeighborOthers(CELL_{x-1,y+1})$
          $enqueueNeighborOthers(CELL_{x-1,y-1})$
          $enqueueNeighborOthers(CELL_{x+1,y})$
          $enqueueNeighborOthers(CELL_{x+1,y-1})$
          $enqueueNeighborOthers(CELL_{x+1,y+1})$
          $enqueueNeighborOthers(CELL_{x,y-1})$
          $enqueueNeighborOthers(CELL_{x,y+1})$

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

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

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

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