# 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 two simple observations related to the taxi status via mining of real taxi traces: (1) the behavior of taxi will be influenced by the statuses, and (2) the macroscopic movement is related with different geographic features in corresponding status. Based on these two observations, a novel taxi mobility model (START) is proposed with respect to taxi status and geographic region. The simulation results illustrate that proposed mobility model has a good approximation with reality in trace samples, distribution of nodes and the contact characteristics.**

## 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. A 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. 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. 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 term 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 of different statuses.

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

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.

## II. 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 characterize the contact patterns among vehicles 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. Two 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 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 two 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 III). Fig. 1 shows that $\overline{speed}$ distribution differs for each status and with strong regularity for each status.

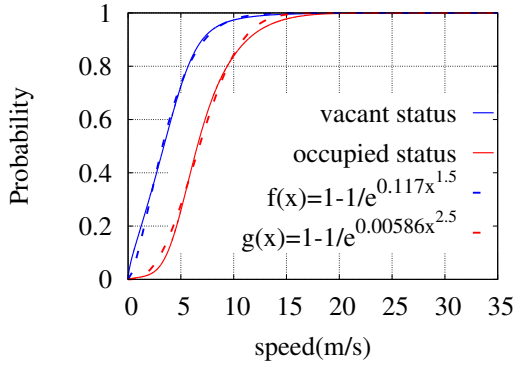



Fig. 1: Speed distributions for vacant and occupied statuses.

### 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 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.

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. Taxi Event Distribution

To validate *Claim 2*, we quantitatively analyze vehicles density in one hour periods. By dividing the whole network into $100 \times 100$ grids, taxi density distributions for load and drop event are computed in each cell. Fig. 3 shows the load/drop events distributions in three time periods. In the morning one, people begin to get out from home, so the load-event distribution is more even than that of drop-event. This

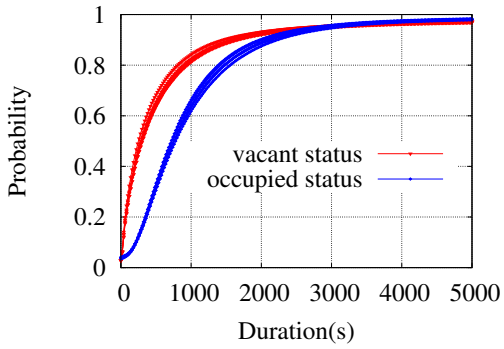



Fig. 2: Status duration distributions.

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, railway stations or scenic spots. Besides, in the evening time period, the dispersion of load-event is of lower degree than that of drop-event, for people coming home at that time. 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 2*.

## III. 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. Dividing the area into regions by the density of passenger load/drop events, respectively. We will show how to dividing the area in the next subsection. Let $\mathbf{R}^{load} = \{R_i^{load}\}$ and $\mathbf{R}^{drop} = \{R_j^{drop}\}$ denote the set of regions of load and drop events, respectively. We assume that $\bigcup\{R_i^{load}\} = \bigcup\{R_j^{drop}\}$ which is the whole area. Then, two transition probability matrixes are calculated: one is the transition probability from a passenger drop region $R_i^{drop}$ to a passenger load region $R_j^{load}$, while the other is the transition probability from a passenger load region $R_i^{load}$ to a passenger drop region $R_j^{drop}$. If the status of a taxi changes to vacant, its current location locates in a $R_i^{drop}$. Consequently, a destination region in $\mathbf{R}^{load}$ will be selected by querying the transition matrix from $\mathbf{R}^{drop}$ to $\mathbf{R}^{load}$. Then, START will randomly select a map node in the region as the destination. As to the status of a taxi changing to occupied, the destination selection process is similar except for that the transition matrix from $\mathbf{R}^{load}$ to $\mathbf{R}^{drop}$ 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 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

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 $R_j$ from the current region $R_i$. Particularly, when two successive events are different: one load and one drop. Therefore, the region $R_i$ and $R_j$ are recognized by different metrics, that is, drop or load event distribution. For an 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.

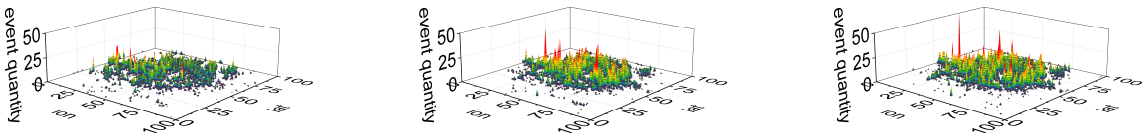
Before defining the region transition probability, we first describe our **region recognition process**. Firstly, we divide the area into $100 \times 100$ small grids, and define them as cells (as shown in the following equation) where $lon$ and $lat$ are relevant longitude and latitude, $len_x$ and $len_y$ are side length of the grid/cell).

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

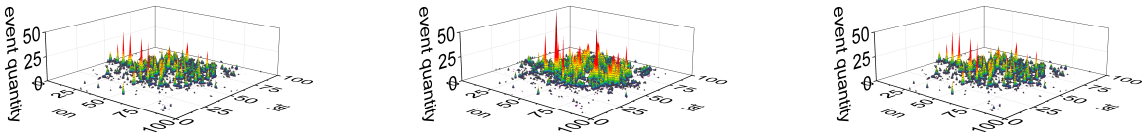
Then, we consider a region as a union of adjacent cells, as following.

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

The main idea of clustering cells to regions is to put adjacent cells with event density larger than an event threshold $\eta$ into a same region. To avoid the size of a region become too large or too small, we set a limitation on the size of a region, say $\|R_i\| \leq ClusterSize$, and also limit the number of final region is less than or equal to 200. Thus, we only consider the top 200 regions, in which $C_{x,y}.events \geq \eta$. After that, the other cells, who do not belong to the top 200 regions, will also be clustered into the 200 regions, while $\|R_j\| \leq ClusterSize$. The overall clustering algorithm is simple. We sort the $100 \times 100$ cells by event density in descending order, and begin with the first cell to search its neighbors whether to join the same region using breadth traversal. Consequently, every cells will be clustered into regions and the size of every region is not larger than $ClusterSize$. By clustering cells into regions, two region sets, $\mathbf{R}^{load}$ and $\mathbf{R}^{drop}$, can be recognized from the data set. The region recognition results for load/drop events are shown in Fig. 4 for the Beijing taxi data set. In this figure, every colored block presents a region. In addition, $ClusterSize = 200$, and $\eta = 121$ for load event and $\eta = 141$ for drop event ( these are set by the average event density of

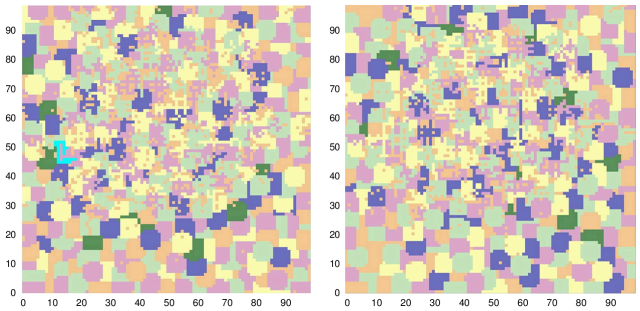



(a) $load, 7:00-8:00$ (b) $load, 12:00-13:00$ (c) $load, 17:00-18:00$




(d) $drop, 7:00-8:00$ (e) $drop, 12:00-13:00$ (f) $drop, 17:00-18:00$

Fig. 3: Taxi density for load/drop events in one hour.



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

Fig. 4: Region recognition

the top 5000 cells order by its event density). You can see clearly differences among load region and drop-off regions.

**Calculation of region transition probability**: 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 $p_{i\to j}^{load\to drop}$.

First, we count all records in $R_i^{load}$ from the data set, i.e. $REC_i^{load} = \{r \mid r.location \in R_i^{load} \text{ and } r.event = load\}$. Let $\|REC_i^{load}\|$ be the amount of such records. For record $r \in REC_i^{load}$, the next event and location can be easily obtained from the data set. Thus, we can get the set of such records whose next event is drop and next location is $R_j^{drop}$, i.e., $REC_{i\to j}^{load\to drop} = \{r \mid r.event = load,\ r.event_{next} = drop,\ r.location_{current} \in R_i^{load}, \text{ and } r.location_{next} \in R_j^{drop}\}$. Let $\|REC_{i\to j}^{load\to drop}\|$ be the amount of such records. We can then easily obtain the transition probability from $R_i^{load}$ to $R_j^{drop}$, as follows,

$$p_{i\to j}^{load\to drop} = \frac{\|REC_{i\to j}^{load\to drop}\|}{\|REC_i^{load}\|}.$$

TABLE II: 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 |

### C. Speed Distribution

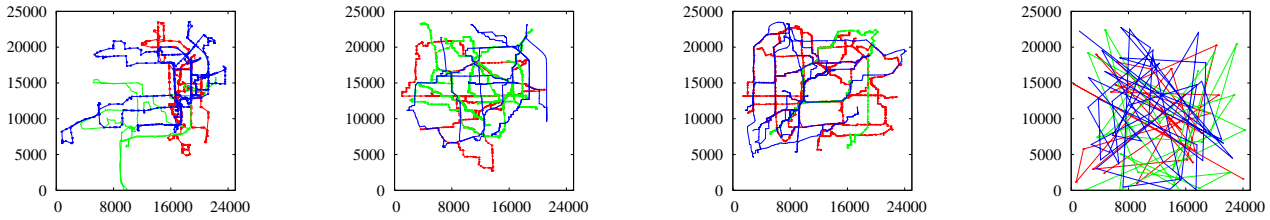
To obtain the speed distribution of each status, we fit the cumulative average speed distribution to get the cumulative probability distribution function, and then take a derivative with it to obtain the speed probability distribution. From Fig. 1, the $\overline{speed}$ distribution shows exponential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

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

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 II. The smaller rms of residuals means better fitting. In this table, the values are all less than 0.02, showing good similarity.

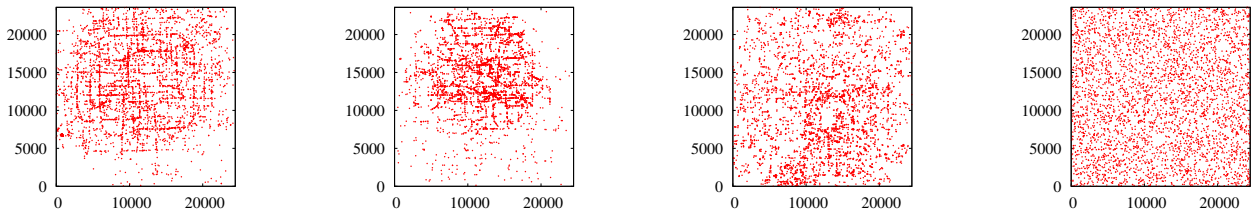
## IV. Model Verification

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




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

Fig. 5: Trace samples.




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

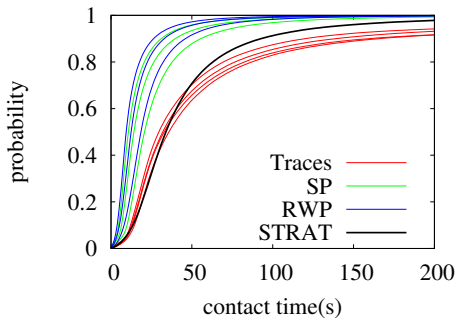
Fig. 6: Trace snapshots.

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].

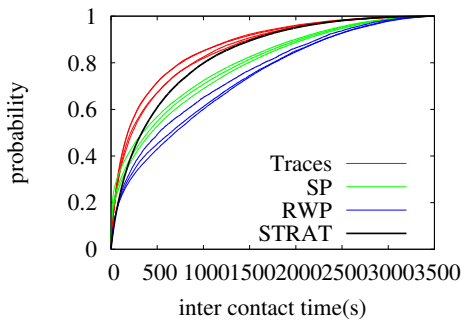
In our simulations, $4,000$ vehicles are deployed in an area of $24,445 \times 23,584 m^2$ (a sub-map of the whole area, including fourth ring roads in Beijing). The speed range of RWP and SP need to be configured. To ensure the accuracy, we choose three different speed ranges $[0, 44.4] m/s$, $[0, 33.3] m/s$ and $[0, 22.2] m/s$ (the upper bounds of speed match the speed limits 80,120,160 $km/h$. The simulation time is three hours and the warm up time for reports is one hour, so that the nodal movement and positions will not be affected by its initial position. The communication range between vehicles is set to $200m$ for potential contacts.

### A. Traces and Node Distributions

Trace samples and their snapshots from different mobility models are reported in Fig. 5 and Fig 6. Fig. 5 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. 6, 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.

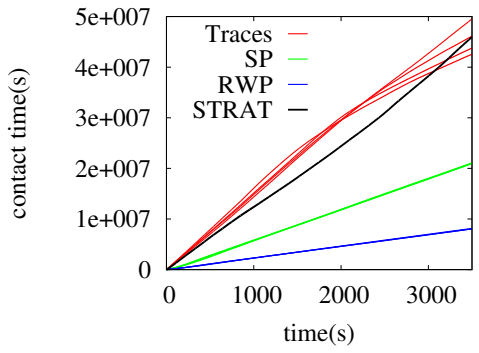



(a) cumulative contact time distribution

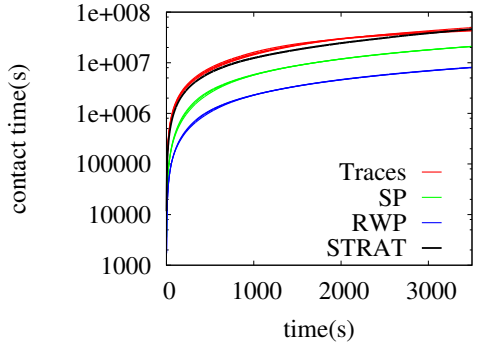



(b) cumulative inter contact time distribution

Fig. 7: Contact time, inter contact time distribution.




(a) time vs. total contact time




(b) time vs. total contact time in logscale

Fig. 8: Time vs. total contact time.

### B. Contacts Characteristics

The contact time and inter contact time among vehicles are also evaluated as the indicators to validate the similarity. Fig. 7 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 $25s$. Clearly, START matches the real traces best among three mobility models. Fig. 8 also show the sum of contact time regarding to the simulation time. In Fig. 8(a), the three green lines of SP and the three blue lines of RWP are overlapping with each other. To recognize the differences, we set y axis as log-scale in Fig. 8(b). The curves of the total contact time vs. time present a liner law. For SP and RWP, the differences of speed ranges show little influence on these curves. Clearly, the rank of the contact characteristic similarity with the real data is START $>$ SP $>$ RWP.

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.

## V. 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.

## 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.

[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.