Metropolitan-Scale Taxicab Mobility Modeling

Lei Zhang, Maryam Ahmadi, Jianping Pan, and Le Chang
University of Victoria, Victoria, BC, Canada

Abstract—Taxicabs, as one of the major transportation platforms in metropolises, are of great interest in vehicular communications and networking research. The mobility of taxicabs, determined by driver behaviors and passenger destinations, has attracted a lot of attention in recent years. Among different mobility models, trace-driven taxicab mobility models preserve many details but often bring too much overhead, while simple, random mobility models largely miss the needed social and geographical features. In this paper, we follow a new approach to capture the social behaviors (e.g., transition among regions) and geographical features (e.g., hot spots) of taxicabs in a metropolis, and build a hierarchical taxicab mobility model to strike a better balance between fidelity and tractability. The performance study shows that the synthesized mobility model can well capture the original trace data while being simple and extensible. It also reveals the difference between taxicab and pedestrian mobility, the latter of which is also often used for vehicles in the literature.

Index Terms—Taxicabs, trace-driven mobility models, hierarchical mobility models, social behaviors, geographical features, vehicular ad hoc networks

I. INTRODUCTION

In recent years, with the fast development of mobile devices, the market trend has attracted a lot of research efforts in mobile applications, systems and services. However, less attention has been paid to the modeling of realistic user mobility, which has a huge impact on the performance of mobile systems [1]. The purpose of studying user mobility is to reveal the regularity and characteristics of user motions, which not only help us to develop more effective communication and networking schemes, but also allow us to generate realistic synthetic data trace for testing or verification purposes due to the lack of real-world trace data.

In Mobile Ad hoc NETworks (MANETs), because of the limited availability of real trace data, most researchers have to rely on synthetic data for their simulation study. To generate such synthetic data, quite a few mobility models have been developed. These synthetic models capture the human mobility properties to different levels. Vehicular Ad hoc NETworks (VANETs), as one of the most important subsets of MANETs, are characterized by their high node speed, strict path-following movement and huge traffic dynamics. The mobility pattern of VANETs is actually very complicated. First, macroscopically speaking, the movement of a vehicle in fact reflects the social behavior of the driver, e.g., traveling among the regions of interest at a certain time. Second, from a microscopical viewpoint, a vehicle’s movement is highly restricted by its surroundings, including the roads, other nearby vehicles and traffic lights. The time could affect as well, as we can observe the difference between daytime and nighttime traffics. Considering these factors, traditional mobility models for human beings used in MANETs are not suitable for VANETs.

The model proposed in this paper falls into the trace-driven synthetic category. The main contribution of this paper is that it proposes a hierarchical taxicab mobility model, which can well capture the taxicab motion patterns in a large modern city and is able to reflect the traffic distribution over geographical locations. To achieve these goals, we mined the GPS traces of Shanghai taxis and derived the transition probabilities among different hot-spot regions for the macroscopic movement of taxicabs. Within each region, we modeled the microscopic movement to imitate the traffic distribution revealed from the real trace data. To be accurate, we have also taken the time-of-a-day factor into account. In our work, however, we do not aim at capturing every motion detail of Shanghai taxis. Instead, we strike a balance between catching the traffic properties of the real trace data and maintaining the general usability for any city-like scenarios, where the traffic distribution due to social behaviors and geographical features is of the most concern. The second contribution is that we discovered the residence time of each taxicab within a given region follows an exponential distribution, which is different from the log-normal distribution observed in the previous work for pedestrians in a campus scenario [7]. This observation actually reflects the intrinsic difference between pedestrian and vehicle mobility.

The rest of the paper is organized as follows. We first introduce the related work in Section II, followed by the preprocessing technique used on the raw trace data in Section III. In Section IV, we describe the modeling work in detail from both macroscopic and microscopic points of view. Then we evaluate the synthetic data generated by our model and discuss related issues in Section V. At last, we conclude the paper with future work in Section VI.

II. RELATED WORK

Synthetic modeling and trace-driven modeling are two dominant approaches to modeling human mobility. A considerable number of synthetic models have been proposed, among which many rely on certain stochastic assumptions of the user motion speed or direction, such as Random Waypoint Model [3], Smooth Random Model [2], Point Group Model [4], etc. These stochastic assumptions make such models simple and tractable, and thus popular. However, the simplification on temporal or spatial domains also makes them too idealistic. These models usually do not take geographical information and time effect into account, and thus can hardly reflect the motion patterns over space and time domains, such as the change of user density at different locations or during different time stages.
On the other hand, trace data collected from mobile nodes in the real world offer another opportunity to investigate user mobility from a more realistic perspective. However, most existing trace-driven models are proposed for modeling the human mobility in a small geographical area, e.g., campus or conference scenarios [7], [8], [10]. Due to the limited space and time span of these traces, such models cannot be simply adopted in VANETs. Nevertheless, important mobility metrics, such as transition probability and time, etc. from these papers, can be defined for VANETs as well.

As to the trace-driven modeling for VANETs, Zhang et al. explored the bus system contact traces obtained from UMass DieselNet [13] and developed a generative mobility model through the investigation of the inter-contact time between different bus-pairs [11]. The limitation of this work is that it is only applicable when the mobile nodes, e.g., buses, follow certain periodic schedules. Concerning taxi systems, which intrinsically have more randomness, Huang et al. constructed a mobility model using the GPS data from the taxis in Shanghai [5]. Only with conclusive discrete GPS data, is their model capable of reproducing the taxi trajectories on the map. However, the heavy dependence on the availability of the GPS data places a limit on the usability of their model. In [6], the authors focused on the microscopic movement of taxis, but they only investigated downtown areas, which are much smaller than the actual range that a taxicab usually covers each day. Thus their model is not able to capture the complete mobility features of taxis in metropolises. Other taxicab trace studies are also available [9], [12]. Nevertheless, these papers either studied traffic sensing from the traffic management point of view, or explored the user inter-contact time for the purpose of improving DTN-related performance metrics, and modeling the taxicab mobility was not their main focus.

III. GPS TRACE DATA ANALYSIS

In order to achieve a realistic modeling of VANETs, we first studied the GPS trace data [14] of taxis in Shanghai, a large modern city, and extracted the general traffic patterns therein.

The traces used in this paper include more than 4,000 taxis in Shanghai, a metropolis in China, in a duration of one month. A variety of information was captured in each record, including the record ID, taxi ID, geographical coordinates, speed, direction, timestamp and occupation status. The majority of the reports were captured in Shanghai’s main urban area, which is around 3,181.43 km² as shown in Fig. 1.

A. Pre-Processing and Error Filtering

To simplify the data processing, we grid Shanghai map into small unit square regions of size 1 km × 1 km each. We treat each unit square as the minimum unit for location analysis.

During the mining process, we observed that the traces contained errors, which had to be filtered out before further analysis. A common error found in these traces is similar to the “pingpong” phenomenon in wireless networks [10]. Assume a mobile node is within the coverage of two or more Access Points (APs), and is associated with AP1 at time \( t \). However, a few seconds later it could be associated with AP2, which is 200 m away. Similarly in our case, within a very short time interval, a taxicab reports two consecutive locations that are too far away from each other. However, the “pingpong” phenomenon is caused by GPS reporting errors, i.e., the distance exceeds the maximum possible distance traveled at the maximum allowed speed. In this work, we set the maximum allowed travel speed at 120 km/h, a practically enforced speed up-limit in Shanghai. We omit all the records which contain such GPS reporting errors.

B. Traffic Load of Unit Squares

“Hot” regions refer to the geographical regions that exhibit remarkable motion properties and are composed of unit squares with very noticeable traffic load. To locate such unit squares, we adopt two traffic metrics to express the traffic load in each unit square: the Vehicle Kilometers Traveled (VKT) and the Accumulative Residence Time (ART).

VKT is a widely-used traffic evaluation metric in transportation engineering, which refers to the distance traveled by the vehicles on the roads. It is usually considered as an indicator of the traffic pressure (or traffic demand) and is used to describe mobility patterns and travel trends. The VKT value for each unit square is calculated as:

\[
VKT_i = \sum_{k=1}^{N_i} v_k \cdot t_k,\]

where \( N_i \) is the total number of taxis once appeared in unit square \( i \), and \( v_k \) and \( t_k \) are the average travel speed and time duration taxi \( k \) spends in square \( i \), respectively. Higher speed and longer staying time imply more traffic pressure.

In modern cities, different areas may exhibit different traffic behaviors. Some areas, e.g., downtown, may have higher traffic flow rates, which make them more dynamic. On the other hand, areas such as airports usually demonstrate more static behaviors, since taxis tend to stay until they are hired by new customers. To reflect the static side of taxicab mobility, we calculate the ART of each unit square \( i \), which is the sum of the residence time for all taxis appeared in square \( i \).

After calculating VKT and ART values for all the unit squares, we can identify the traffic “backbone” of Shanghai,
where the majority of the recorded traffic is reported. Figure 1 gives an overview of the taxi traffic distribution over the city of Shanghai in terms of VKT. A warmer color (e.g., red over yellow) implies higher traffic load. As demonstrated, the aforementioned “hot” regions contribute to the most of the taxicab traffic in the city and become the areas we are more interested in when considering the taxicab mobility.

Besides the spatial taxicab traffic distribution, we have also considered the impact of time in a day. Figure 2(a) and (b) illustrate the VKTs in daytime hours (6 am to 6 pm) and nighttime hours (6 pm to 6 am), respectively. While Fig. 2(c) and (d) show the ARTs in daytime and nighttime, respectively. As expected, less traffic is reported during nighttime hours. From the figures we can see that, for either VKT or ART, the traffic distributions over the map are quite consistent in daytime and nighttime hours. The stable distribution indicates the consistency of the region division over time in the next section.

IV. TAXICAB MOBILITY MODELING

The mobility pattern of a taxicab system is composed of each taxicab’s travel trajectory: what areas a taxicab has come across; which area it will enter next; how long it spends in each region, at what speed, following what route, etc. A good mobility model should have a quality reflection of these behaviors. In this section, we present our model and show how it reveals the motion features of the taxis in Shanghai.

A. Identifying “Hot” Regions

We concentrate on popular regions having a large amount of traffic, i.e., with remarkable VKT or ART values. By clustering the unit squares combined with the geographical and social information of Shanghai, such as the locations of large commercial districts or transportation hubs, we convert the map of unit squares into a graph, whose nodes are referred to as independent “hot” regions in terms of traffic density. Table I gives each region’s name and its corresponding sequence number used in our modeling. As shown in Fig. 3, we plotted the contour of the VKT and ART values and picked 12 regions with considerable traffic densities. Most “hot” regions are active in both VKT in Fig. 3(a) and ART in Fig. 3(b). However, region 11 (Gaojin Zhen) and region 12 (Pudong Airport) appear more distinct in terms of ART than their VKTs, which means the taxicab mobility in these two regions is more static. This is quite reasonable since both of the regions are separated from the central popular areas, and the taxi drivers prefer to stay longer to pick up new customers instead of heading back empty immediately.

B. Region Transition Patterns: a Macroscopic View

We first use the transition residence time and transition probability to characterize the mobility patterns among regions, i.e., from a macroscopic perspective.

1) Transition residence time between regions: The transition residence time is defined as the travel time within one region before the taxi leaves for the next one, which does not necessarily mean that the vehicle is staying stationary with the speed 0 m/s. To investigate the distribution of the transition residence time, we first use the statistics toolbox in Matlab to generate some distributions fitting our data samples, among

TABLE I: Identified Hot Regions

<table>
<thead>
<tr>
<th>ID #</th>
<th>Region Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Xingzhuang District</td>
</tr>
<tr>
<td>2</td>
<td>Hongqiao Airport</td>
</tr>
<tr>
<td>3</td>
<td>Xinjing Zhen</td>
</tr>
<tr>
<td>4</td>
<td>Shanghai Railway Station</td>
</tr>
<tr>
<td>5</td>
<td>South Railway Station</td>
</tr>
<tr>
<td>6</td>
<td>Downtown</td>
</tr>
<tr>
<td>7</td>
<td>Wujiacang District</td>
</tr>
<tr>
<td>8</td>
<td>Pudong District</td>
</tr>
<tr>
<td>9</td>
<td>Chuansha District</td>
</tr>
<tr>
<td>10</td>
<td>Taxi Company</td>
</tr>
<tr>
<td>11</td>
<td>Gaojin Zhen</td>
</tr>
<tr>
<td>12</td>
<td>Pudong Airport</td>
</tr>
</tbody>
</table>

Fig. 2: Traffic Load during Daytime and Nighttime.

Fig. 3: The Division of Popular Regions.
which both exponential and log-normal distributions show good fits. However, as we can observe from Fig. 4, exponential distribution performs better than log-normal in fitting those samples with small residence time but high frequency, i.e., the first data bin in the figure. The samples with higher frequency dominate the whole distribution, and thus we prefer to adopt the exponential distribution as the transition residence time distribution approximation. We also perform the Chi-square tests to verify our hypothesis. The results show that the fit of our sampled data to an exponential distribution is accepted at the level of significance $\alpha = 0.05$. Different from the log-normal distribution observed in [7] for pedestrian mobility on campus, we believe exponential distribution is more capable of capturing taxicabs’ motion property. As for pedestrians, when they enter a region, i.e., a building, they probably stay there for a while until finishing working or shopping, etc. So the data bin having the highest frequency residence time falls in somewhere between the minimum and maximum values. But for taxis, it is more probable that they just pass through a region in a short time, if they do not need to take new customers. Thus the frequency of small residence time samples is much higher.

To understand the relationship between the transition residence time and the taxicab travel trajectory, we plot the transition residence time from one region to another in Fig. 5. These figures tell us the average transition residence time that the taxis spend in the current region before moving to the next (y coordinate of the circle center). In other words, by looking at this figure we can compare the transition residence time between different region pairs. We observe that the transition residence time within one region depends on the destination regions and we can also consider it as the inter-region transition residence time. For instance, consider taxis in region $i$ are leaving for region $j$, and the transition residence time in region $i$ follows distribution $P_t_{ij}$. We find that $P_t_{ij}$ is not identical to $P_t_{ik}$, if $j \neq k$, i.e., in Fig. 5, the size of circle$(i, j)$ is different from the size of circle$(i, k)$. We believe this is the first model disclosing such a phenomenon. So far in the research community, an identical transition residence time have been assumed for each region regardless of the taxicab transition decisions. Such a phenomenon in fact reflects the geographic and social features of different regions. If the traffic flow between two regions is very smooth, e.g., with less chance of traffic jams, we can expect a shorter transition residence time; otherwise, the transition residence time will be longer. Moreover, we compare the transition residence time in different time periods, and find that the transition residence time during nighttime hours is smaller than that in daytime. This indicates the change of the traffic load during a day. During the day time, with more vehicles on the roads, traffic jam is more likely to happen, which increases the travel time within regions.

2) Transition probability between regions: To reflect the motion pattern of taxis traveling among different regions, we count the total number of transitions between any two regions, and plot them in Fig. 6. In this figure, the radii of circles represent the number of such transitions. The number of transitions demonstrates similar patterns between daytime
and nighttime, despite the fact that there are much fewer transitions during nighttime. If such transition numbers are normalized by the total number of all transitions from the same region, we have the transition probabilities shown in Fig. 7, which are expressed by bars. The lengths of the bars at the same x coordinates add to 1. According to our results, the transfer probabilities remain quite stable during the daytime and nighttime, so we summarize all the statistics in one figure.

C. Taxicab Trajectory within a Region: a Microscopic View

After exploring the macroscopic movement of taxis such as the transition residence time and probabilities, we now focus on the trajectories within each region. Once a taxi enters a region, two cases are possible: 1) the current region is the destination region of the taxi; 2) it is a transit region. In the first case, the taxi first travels to the destination point, and then it either leaves for the next region or stays in the region waiting for a new customer. In the second case, the taxi crosses the current region and heads for the next one directly. In the real world, the trajectories can be much more complicated. For instance, the road restrictions or traffic congestion can greatly affect a taxi’s movement. However, it is challenging to cover all the details when modeling vehicle mobility behaviors at a city scale. Therefore, we omit such region-specific details and focus on the most general cases, which exhibit the major statistical properties close to the real scenarios.

We assume that in a synthetic mobility model, once a taxi leaves a region, after a certain amount of time, which is calculated using the distance between the two regions and its average travel speed, it arrives at the other region. In that region, the taxi always starts at a point near the boundary where it comes from. We call such a point the entry point. After that, it first determines a destination region based on the transition probabilities, and then chooses a temporary point, called the transit point, towards the destination region but in the current region, and moves towards it. Finally, it travels to a departure point which is close to the next region and leaves the current region from there. A sample trajectory is shown in Fig. 8, and the circles are the contours for the traffic load, where smaller circles imply heavier traffic load.

The setting of a transit point has very practical meanings. As mentioned above, this point is either the destination point of the current trip, or a representative point on its path across the region. Another important role of the transit point is to form the traffic load distribution in the space domain. As observed in Fig. 3(a), within each region, the traffic load is not uniformly distributed. Heavy traffic appears in certain locations while the traffic of the surrounding areas decreases gradually. One approach to model such a spatial distribution is to make the assumption that the traffic load over X and Y axes each follows a certain distribution, where the 2-D distribution over the region fits the trace data, i.e., at least one peak traffic area appears within the range of the heavy traffic areas in the trace. For simplicity, we choose Poisson distribution as the traffic load distribution for the verification in the next section. Other more complicated distributions can be adopted as needed. Such distributions are usually not identical for every region, due to the existence of various geographic constraints.

Besides the spatial properties, our model also considers the temporal property of the motion within one region. As pointed out in Section IV-B1, the transition residence time of the taxis spent in a region follows an exponential distribution. From the trace, we retrieve the distribution parameter $\lambda$ for each region, and use it to generate the synthetic transition residence time of taxis for the corresponding regions. We assume that the taxis have symmetric motion status along the two paths: one from the entry point to the transit point, and the other one from the transit point to the destination point. Based on this assumption, the transition residence time is evenly shared between the two paths. As the distance and the travel time in this region are known for a taxi, the speed can be easily calculated. The last step is to check whether the speed falls into a reasonable range, i.e., 0 km/h to 120 km/h. If the speed is valid, the motion properties of the entry point, transit point and departure point, will be added to the synthetic trace.

V. SYNTHETIC TRACE GENERATION AND VERIFICATION

In the previous section, we gave detailed description of our model, i.e., how we characterize the route of a taxi among and within regions and how we determine the time spent in these regions. We can describe our model as a five-input system, where the inputs are: transition probabilities among regions, transition time distribution within each region, traffic load distribution within each region, and travel time and speed between regions. To generate the synthetic trace data for a single taxi, we start from a randomly selected location at the beginning of a day, i.e., 0 am, and use the transition probabilities to determine its macroscopic route in the city (i.e., among regions). Within each region, we use the transition time
distribution and the traffic load distribution to determine the amount of time the taxi spends, and the microscopic trajectory within the region. As for the transition time distribution, the exponential distribution parameter is calculated from the data samples. From the real traces, we observed that the speed and the travel time of taxis between any pair of regions are quite stable, as most regions are geographically close to their neighbors. Therefore, when generating synthetic traces, we use the average speed as the travel speed, and the average travel time as the travel time between two regions. These approximations are reasonable according to the trace and also reduce the complexity of the model.

Following the process described above, we generate synthetic trace data including 4,000 taxis within one day. For the generated traces, VKT, as the main traffic metric, is also calculated for each unit square and the traffic load distribution is plotted in Fig. 9. Compared with the real data traces in Fig. 2(a), our generated traces capture the main characteristics of the taxis motion. They correctly reflect the main traffic areas and the traffic load distribution over the space domain. However, small differences can be observed due to: 1) exclusion of the areas with less significant traffic load; 2) approximation of the traffic load distribution (Poisson distribution) within each region.

We also verify our synthetic traces in terms of the taxicab density over time in the twelve regions shown in Fig. 10. Four sample times are selected, and we compare the normalized user densities of the real and the synthetic traces for all regions. These figures show that our synthetic data match the real trace quite well with a maximum deviation of 8%, which implies that our mobility model can successfully capture the taxi motion features in a city-scale among different popular regions. Users can determine their own model inputs based on their pre-defined city scenarios, i.e., customized transition probabilities and residence time. Our model can then be easily used to generate traces while keeping the realistic relationship between taxi motion, city geographic and social features.

VI. Conclusions

In this paper, we proposed a trace-driven hierarchical taxicab mobility model for large modern cities. Our model distinguishes metropolitan taxicab mobility from pedestrian mobility. Because the synthetic traces generated by the model match well with the real ones, we believe that the model can well capture the taxicab motion features on the social and geographical aspects at a city-scale with both macroscopic and microscopic considerations. What deserves to be mentioned is that our model keeps a balance between fidelity and tractability. The approach is not confined to the Shanghai taxicab traces, but can be easily extended to other city-like scenarios as well. As for the future work, we will put efforts on modeling the transition process. According to our observation from the trace, the taxi’s transition among regions does not demonstrate Markovian property, which is the reason that Markov process cannot be applied directly. High-order Markov process [12], or heterogeneous Markov process can be possible solutions.

Acknowledgment

This work is supported in part by the NSERC of Canada.

References