*Journal of Universal Computer Science, vol. 18, no. 9 (2012), 1072-1092
submitted: 1/12/11, accepted: 28/4/12, appeared: 1/5/12 © J.UCS*

A Metropolitan Taxi Mobility Model from Real GPS Traces

Hongyu Huang
College of Computer Science of Chongqing University
High Performance Networking Research Center of Chongqing University
Chongqing, China
hyhuang@cqu.edu.cn

Daqiang Zhang
Department of Computer Science of Nanjing Normal University
Nanjing, China
dqzhang@njnu.edu.cn

Yanmin Zhu
Department of Computer Science of Shanghai Jiao Tong University
Shanghai, China
yzhu@cs.sjtu.edu.cn

Minglu Li
Department of Computer Science of Shanghai Jiao Tong University
Shanghai, China
li-ml@cs.sjtu.edu.cn

Min-You Wu
Department of Computer Science of Shanghai Jiao Tong University
Shanghai, China
wu-my@cs.sjtu.edu.cn

Abstract: The past few years have witnessed the growing interest in vehicular ad hoc networks (VANETs) and their potential applications for Internet of Things (IoT). Since the mobility model is crucial to simulation based researches of VANET, using a realistic mobility model can ensure the consistency between simulation results and real deployments. Although there are many mobility models characterizing the movement of mobile nodes, none of them consider the behavior of vehicles in a metropolitan scenario. In this paper, we present our study of extracting a mobility model for VANET from a large amount of real taxi GPS trace data. In order to capture characteristics of the urban vehicle network from microscopic to macroscopic aspects, we design three parameters and extract their values from the GPS trace data. Using this mobility model, we can generate the synthetic trace to simulate the movement of taxis in the urban area of a metropolis. The validation is carried through extensive comparisons between the synthetic trace and the real trace. The results show that our mobility model has a good approximation with the real scenario.

Key Words: Vehicular Ad Hoc Network, Mobility Model, Taxi GPS Data

Category: C.2, C.2.m

1 Introduction

In recent years, Internet of Things (IoT) has attracted much attention because of its wide applications. As a killer application of IoT, vehicular ad hoc networks (VANETs) also have witnessed the growing interest from academy and industry. Supposing that each vehicle on the road is equipped with wireless devices and sensors, these vehicles can communicate each other or with the roadside infrastructures to transmit data for sharing information or accessing Internet. This IoT on wheel can help to build many intelligent transportation applications such as traffic congestion relief, traffic monitoring, and prediction of bus arrival time.

Since a large-scale testbed of VANETs is difficult to deploy, a mobility model is needed to generate traces of vehicles so that people can study VANETs by simulations. For example, Fiore and Harri [Fiore and Harri 2008] investigated the topology of VANET based on some existing mobility model and then explain why different mobility models lead to dissimilar network protocol performance. In order to make results of theoretical analysis suitable for real scenario, it is necessary to build these work on realistic mobility models. For example, Hongzi et. al [Zhu et al. 2011] validated that the node inter contact time in real VANET mobility follows exponential because of the existence of traffic influxes. Similarly, Xu et. al [Xu et al. 2009] evaluated the performance of traffic monitoring using VANET nodes which move according to realistic taxi GPS traces. Hence a realistic mobility model is crucial to guarantee the consistency between simulation results and real deployments.

Although there have been many mobility models, ranging from theoretical models [Johnson and Maltz 1996] [Royer et al. 2001] to realistic models [Zhang et al. 2007] [Burgess et al. 2006], none of them consider the continuous geographical mobility, i.e., node locations are given by coordinates from time to time, of vehicles in metropolitan scenarios. Since such a mobility model is very important to the VANET research, e.g., the performance study of VANET routing protocols, we study how to extract a mobility model from a large amount of empirical taxi GPS trace data.

In our work, we collected real GPS data which were reported by over 4,000 taxis running in the Shanghai urban area for three months from February to April of 2007. The GPS data includes not only the coordinates and timestamp, but also the moving direction and the running status. By investigating the GPS data, we find that the taxi mobility shows an obvious regularity. For example, taxis prefer to turn to the same direction at some road intersections and their travels appear some patterns. In order to capture the regularity of the taxi mobility, we propose three parameters, i.e., *turn probability*, *road section speed* and *travel pattern*. *Turn probability* characterizes the behavior of a taxi at a road intersection. *Road section speed* defines the running speed of taxis on a given road section. *Travel pattern* depicts the regularity of long run trips from origination to destination. We find that these parameters are able to capture characteristics of taxi mobility. Therefore we propose the mobility model of MEtropolitan TAxis (META) which can be used to generate the synthetic trace for the movement of taxis in an urban area.

In order to validate our mobility model, we generate the node trace according to META and reconstruct the taxi trace from the GPS data, which are called as META trace and real trace respectively in the rest of this paper. The validation was carried out by comparing these two traces in terms of several important metrics including trace characteristics, the network topology and the performance of a routing protocol. In addition, we further add a trace generated according to the random waypoint mobility model (RWP) [Johnson and Maltz 1996], called as RWP trace, into the comparison to reveal the difference between realistic models and random models. The results show that META trace has good approximation with the real trace which also validates the effectiveness of model parameters.

The rest of the paper is organized as follows. Section 2 discusses some related work. Section 3 describes the preliminary processing on the GPS data and then the methodologies of how to extract the mobility model are presented in Section 4. Section 5 shows the validation methods and results. Finally, Section 6 presents our conclusions.

2 Related work

Early researches of mobility models stress on simplicity and theory. Some well known mobility models such as random waypoint and random direction [Royer et al. 2001] make the node randomly choose a destination or direction to move and then travel with a randomly chosen speed. The advantages of these models are simple and good to analyze, but they also have notable disadvantages. For example, the average speed of nodes in the random waypoint mobility model is found to decrease with the progress of the simulation [Yoon et al. 2003]. Although many modifications have been considered to make these simple models more realistic [Jardosh et al. 2003] and more stable [Boudec and Vojnovic 2006], the consistency between the research results draw from these models with the reality still needs to be investigated.

In recent years, researchers were inspired to collect data from the real world and extract mobility models. Jain and Lelescu [Jain et al. 2005] [Lelescu et al. 2006] derived empirical models from the trace of registrations of wireless users. Kim et al. [Kim et al. 2006] also proposed a methodology to extract mobility model from real WiFi user traces. Recently, Hsu et al. [Hsu et al. 2007] made use of the Dartmouth WLAN dataset [CRAWDAD 2007] to model time-variant user mobility. The methodology in these researches is similar to ours, which is first revealing characteristics of real traces such as registration pattern, user pause time and location visiting preferences of users, and then extracting model parameters from these traces. Using this methodology, the model is guaranteed to generate traces which can capture features of the real world. Nevertheless, these traces are from human mobility so that they cannot be used directly in the VANET research. As is known that the movement of vehicles is constrained to running on the streets, traffic rules and status, so a realistic vehicular mobility model should come from the traffic data.

Actually, few researchers had studied on how to extract mobility model from real traffic data. A recent work comes from Zhang et al. [Zhang et al. 2007]. They studied the trace collected from UMass DieselNet [Burgess et al. 2006], which is a DTN testbed consisting of WiFi nodes attached to buses. Based on the analysis of the deterministic inter contact time for bus pairs running on route pairs, they constructed route-level models to capture the periodic inter contact time pattern. Since their mobility model is based on the interactions between a pair of buses, the ignorance of traces between interactions limits this model for a wider applicability.

In fact, it is more relevant to compare our model with the microscopic traffic simulator because both of them aim to generate continuous geographical traces. The common used traffic simulators include the SUMO [Krajzewicz et al. 2002] and the VISSIM [Vissim 2007]. Although they can generate realistic vehicular traces, their simulation parameters, such as the traffic light and the vehicle length, are too complex to specify for the VANET simulation so that they are more suitable to be used in transportation and traffic science. In our work, we showed that three parameters are enough to capture the characteristics of taxi mobility which can be used in the network simulation.

3 Background on data processing

Before we can extract the mobility model, there must be preprocess on the GPS data. In this section, we first introduce the information recorded in the real GPS data and the road network which our mobility model is based on. Since the GPS data needs to be mapped on the road network to reconstruct the trace, so the process of map-matching is also presented. For the sake of space limitation, we only briefly introduce the data preprocessing, please refer to our previous work [Huang et al. 2007] for more details.

3.1 Data collection

From February to April of 2007, we collected GPS data from over 4,000 taxis which were equipped with GPS devices, one of them is shown in Figure 1. These taxis reported their running status to a data center through the GPRS system in real time. Every piece of GPS data is defined by a 5-tuple $D(V_{ID}, T, \Psi, \Omega, \Theta)$, where V_{ID} is the unique ID of the taxi, T is a timestamp, Ψ is its geographical location represented by longitude and latitude coordinates, Ω is its direction of headway represented by a geographical angle clockwise from the due north, and Θ is the taxi status which tells whether the taxi is carrying passengers or not. Note that the GPS data also records speed information, but we have to discard it because it is inaccurate in most cases.

Since taxi GPS devices were originally deployed to monitor and schedule taxis and the GPRS messages that carry the GPS data will be charged, taxis preferred to report the GPS data in a long time interval. Although the report interval is not regular, varies from 30 seconds to few minutes, the average value is about 60 seconds.

Figure 1: A GPS device equipped on the taxi

3.2 Road network

Except for the taxi GPS data, there is a digital map of Shanghai city which records all of the roads according to the shape file format [ESRI 1998]. The basic unit of the road map is the road section which is between two consecutive intersections of a road. A road section is defined as a polyline in the shape file. Each polyline is enclosed by a bounding box which defines the maximum and minimum coordinates.

There are totally over 30 thousand road sections in the road map of the whole Shanghai city. In order to reduce the complexity of computation, we focus on the city center area of Shanghai, which is shown in Figure 2.

The city center is about 20 $km^2(5km * 4km)$. This area is chosen because it is the most popular area of vehicles and the vehicle density in this area is very high. This road network has 1522 road sections which include both arterial and inferior roads, so that it is representative for a common urban road network. In our work, we regard all the road sections are bidirectional and divide each road section into two road sections, so we actually have 3044 road sections.

3.3 Map-matching

Before making use of GPS data, we need to map them onto the road map which is called map-matching. In reality, due to various types of errors, e.g., tall buildings, the GPS data itself is noisy, which means the geographical location Ψ is not accurate and can be

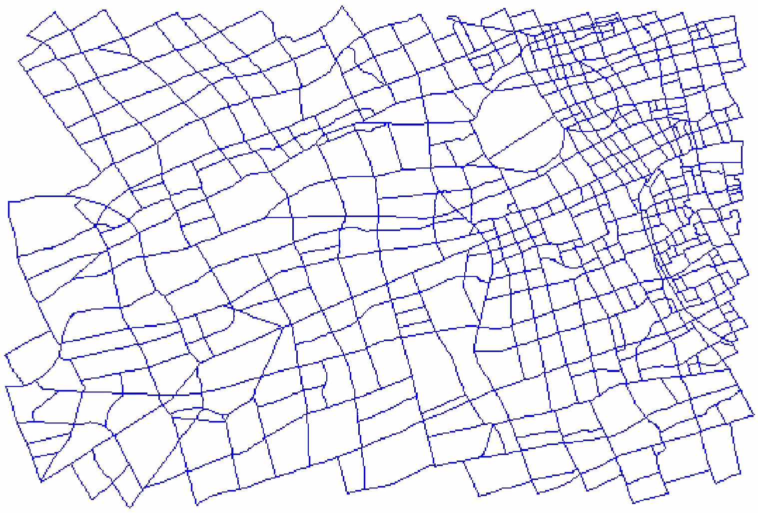

Figure 2: A small road network of Shanghai

away from its real location as far as about 60 meters. In order to map the GPS data onto the road network, we use a heuristic algorithm which considers both the distance and direction between the taxi and the nearby road sections. This heuristic works fine in case that the GPS data locates between two parallel road sections. So we also consider the historical information of the former matching result. If one of the parallel road sections belongs to the road that the former data mapped on, it is then determined as the result of this matching, or otherwise the nearest one is chosen to be the result.

In our field test of the map-matching algorithm, we took a GPS equipped taxi and record its traveling path. Then we retrieve the GPS data of this taxi collected during the period we were traveling and compare the map-matching result with the real location. The results of over 100 field tests show that the map-matching algorithm has a accuracy higher than 95% which guarantees the quality of our mobility model.

4 Model parameters

In this section, we present our methodology of how to extract the mobility model from the historical GPS data. The META mobility model includes three parameters, *turn probability*, *road section speed* and *travel pattern*. The former two parameters describe the microscopic behavior of vehicles and the latter focuses on the macroscopic feature.

4.1 Turn probability

One of the basic behaviors of a vehicle is its turn at the exit intersection of a road section because this reflects the common sense of drivers and sometimes the traffic rules. In order to obtain *turn probability*, the traveling path of a taxi needs to be determined in advance. If the interval between a pair of consecutive reports is long, the exact path that the taxi traveled is hard to determine and the taxi behavior along the series of intersections cannot be obtained. Therefore we define a turn pattern which is a definite path between two consecutive GPS data.

We define the following concepts which is illustrated in Figure 3.

Intersection: the intersections are depicted as crosses in Figure 3, which are labeled as $A1$, $B2$, etc.

Road section: a road section is a link between two adjacent intersections, labeled as $L_{I1,I2}$ where $I1$ and $I2$ are intersections defined above. Since all road sections are regarded as bidirectional, so $L_{I1,I2}$ and $L_{I2,I1}$ are different. On the road section $L_{I1,I2}$, the vehicle must travels from $I1$ to $I2$ and vice versa. We call $I1$ is the entrance intersection of $L_{I1,I2}$ and $I2$ is the exit intersection.

Path: a path is a sequence of road sections that the vehicle travels between two consecutive data, labeled as $P(L_1, L_2, ...)$ where $L_n, n = 1, 2, ...$ are road sections.

Turn pattern: the turn pattern of a road section is a pair of road sections (L_1, L_2) where the first GPS data locates on L_1 and the second data locates on L_2. The turn pattern is a kind of relationship between two consecutive GPS data from which we can definitely determine the path between the data without confusion. In Figure 3, the turn pattern of $L_{C2,C3}$ includes 10 road section pairs of $(L_{C2,C3}, L_{C3,B3})$, $(L_{C2,C3}, L_{C3,C4})$, $\ldots$, and $(L_{C2,C3}, L_{D3,D2})$ which represent data pairs $(s, e1)$, $(s, e2)$, $\ldots$, and $(s, e10)$. We call $(L_{C2,C3}, L_{C3,B3})$, $(L_{C2,C3}, L_{C3,D3})$ and $(L_{C2,C3}, L_{C3,C4})$ as direct turn pattern because $L_{C3,B3}$, $L_{C3,D3}$ and $L_{C3,C4}$ are adjacent to $L_{C2,C3}$. Then the other road section pairs are called indirect turn pattern.

The turn probability of every road section is obtained by searching through the GPS data of the three months. A turn probability from one road section to an adjacent road section is defined as

$$TP(L_i, L_j) = n/N, \tag{1}$$

where N is the number of data pairs which match the direct turn pattern, and n is the number of data pairs that match the turn pattern (L_i, L_j). If a data pair matches the indirect turn pattern (L_i, L_k), it can be split into two direct patterns (L_i, L_j) and (L_j, L_k). So it can be easily obtained that

$$TP(L_i) = \sum_{j=1}^{n} TP(L_i, L_j) = 1, \tag{2}$$

where (L_i, L_j) is direct turn pattern.

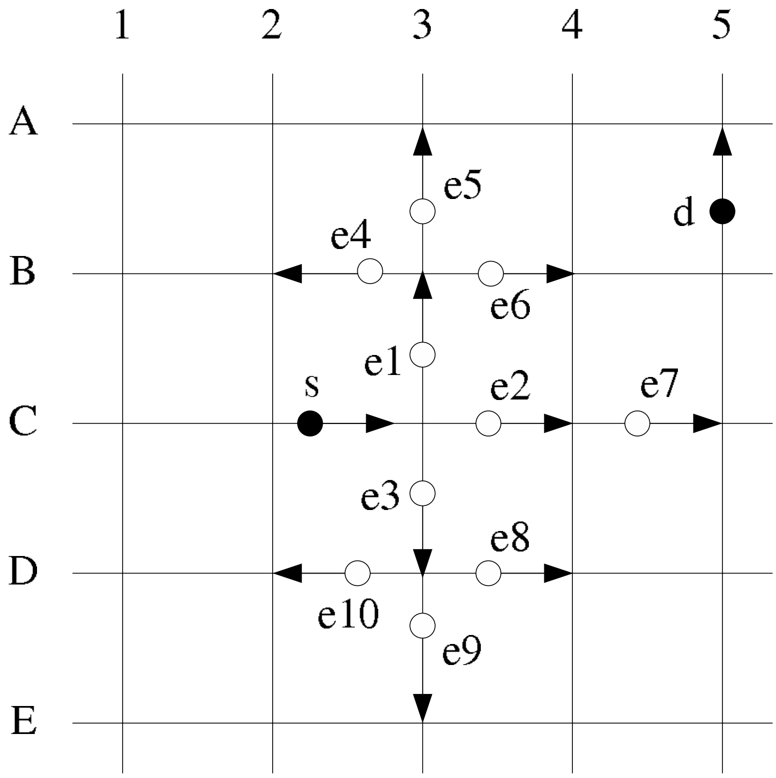

Figure 3: The turn pattern

When the turn probability for all road sections to their adjacent road sections is computed, we can use it to estimate the reasonable path for a long interval data pair. For example, to estimate the path between two consecutive data s and d, we first compute the metric λ which is defined as

$$\lambda_i = \prod_{j=1}^{n-1} TP(L_j, L_{j+1}). \tag{3}$$

In Equation 3, the λ_i is the path probability of the i^{th} path, n is the number of road sections along the path and $TP(L_j, L_{j+1})$ is the turn probability from road section L_j to L_{j+1}. We determine the path between s and d as the one which has maximum path probability and call this path as the maximum turn probability path. Now, for a data pair which cannot match the turn pattern, the maximum turn probability path between them can be regarded as the actual path that the node has traveled.

In order to check the correctness of the maximum turn probability path, we intendedly remove some data from a series of GPS data and find the maximum turn probability path among the remained data to see whether it can match the actual path. The results of over 10000 tests show that when the number of road sections between a data pair is less than six, the accuracy of maximum turn probability path can achieve 100%. Based on this accuracy, we make use of the data pairs between which the number of road sections of the maximum turn probability path is less than six. This promote the usage of historical GPS data from 13%, which match the turn pattern, to 79%. This method not only enables us to utilize more GPS data, but also help reconstruct real traces of taxis which can be used to validate META trace.

4.2 Road section speed

The second parameter of META is *road section speed* which is defined as the average speed of a road section during a time period, specified as five minutes in our work. Different from the random mobility model, the speed of a vehicle is restricted by the traffic status which is represented by *road section speed*. Because of the long interval of the GPS data, the instant speed for taxis at every second is difficult to estimate so the average speed is considered. The average speed between two consecutive GPS data is easy to compute from s/t where s is the length of the path and t is the interval of the two data. In order to convert the taxi speed to the road section speed, the average speed of a taxi during a path is assigned to road sections along that path. For example, in Figure 3, after the speed between s and $e1$ is computed, it is assigned to $L_{C2,C3}$ and $L_{C3,B3}$ as one of their speed elements during the time period. For every time period, we compute the average value of these speed elements of a road sections as its average speed. If there are no taxis pass a road section, its average speed during that time period is left to be blank.

Figure 4 shows the change of the speed of a road section, whose ID is 292431, during a week from March19 to March 23, 2007. Each point in this figure represents a road section speed of five minutes. The road section speed in this figure appears regular: It climbs up to the peak during midnight and 4 am. Then it drops sharply after 6 am and maintains at a low level during 8 am to 8 pm after when it rises and returns to the peak. This change also represents the regularity of the traffic status in an urban area.

From Figure 4, we can also see that the road section speed has a large variation, some values are very high and some are very low. Most of the variations come from the case that only one taxi passed the road section during a time period, its individual behavior will make the average speed inaccurate. If the taxi passes the road section without waiting for the traffic light, the average speed will be very high, otherwise it becomes low.

In order to get rid of oscillatory speeds and fill blank speeds of some time periods, we make use of the Fast Fourier Transform (FFT) filter tool of the Origin software [Origin 2009] to smooth the speed curve. To use the FFT filter, a parameter for the cutoff frequency needs to be specified which is defined as

$$F_{cutoff} = 1/n\Delta T, \tag{4}$$

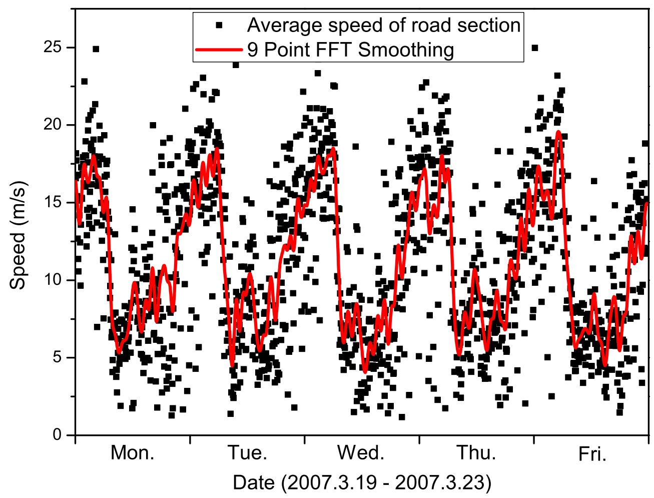

Figure 4: The change of the average speed of road section 292431 during March 19 and March 23, 2007

where F_{cutoff} is the cutoff frequency and n is the number of data which is used in the FFT filter and ΔT is the time spacing between two adjacent data points, which is a time unit of five minutes. The red line in Figure 4 represents the smoothed curve of n=9. A larger value of n will cut off the higher frequencies and generate a greater degree of smoothness. We specify n to be 9 because it has a low Sum of Squared Error (SSE) between the original data and the smoothed data while capturing the trend of the speed change.

Considering the fact that in most of the VANET researches, the node trace of one day duration is enough for simulations, so we assemble the data of all weekdays from February to April 2007 into one day and smooth it using FFT filter to obtain *road section speed* which is used in the META mobility model to generate synthetic traces. An example of the average value of assembled data is shown in Figure 5 where the red line is also the smoothed speed.

4.3 Travel pattern

The taxi mobility not only has regular microscopic behavior, but also appears macroscopic features. For example, a taxi locates in a certain area prefers to go to another area as the destination of a travel. In order to characterize the macroscopic feature of taxi mobility, we propose *travel pattern* which is defined as a probability that a node travels from one area to another area.

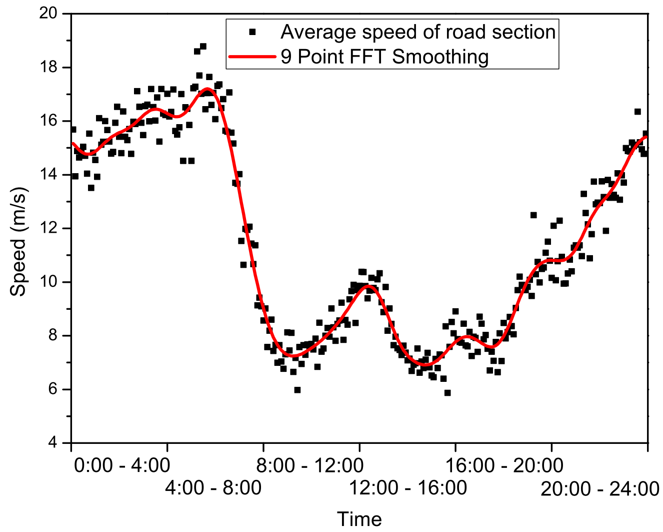

Figure 5: The change of the average speed of road section 292431 during a day with assembled data

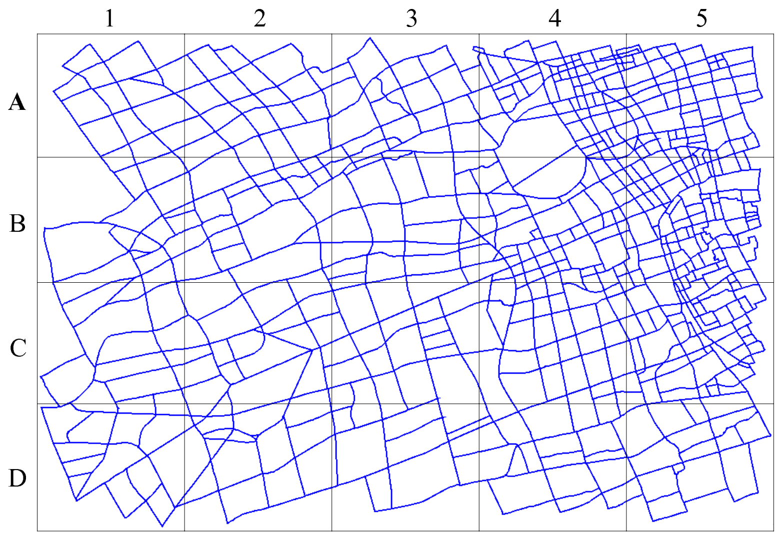

Figure 6: The road network with travel grids

In order to recognize the distribution of originations and destinations of taxi travels, the road network needs to be divided into different areas. For simplicity, we consider rectangular area and divide the road network into $n * m$ grids, which are called travel grids and shown in Figure 6. A large size of the travel grid will merge different travel patterns. And a small size of the travel grid cannot capture the characteristic of the travel pattern. As a result of tradeoff, we divide the road network into $4 * 5$ grids so each grid is about $1km * 1km$. Then the travel pattern of taxis is recognized by investigating their origination and destination. Recall that the GPS data records the status of the taxi of whether it carries passengers. Searching through the taxi GPS data by the order of timestamp, the first and last data of a series of consecutive data which have the same status can be identified. We regard the first data as the origination of this travel and the last data as the destination. Given the coordinate of the GPS data, we can easily tell which travel grid it locates in and then the travel pattern can be written as $(G1, G2)$, where $G1$ is the grid ID that the origination locates in and $G2$ is for the destination. Processing the historical GPS data in this way, the probability matrix for every travel grid is defined as

$$\mathbf{D_{C3}} = \begin{bmatrix} P_{A1} & \dots & P_{A5} \\ \vdots & \ddots & \vdots \\ P_{D1} & \dots & P_{D5} \end{bmatrix}. \tag{5}$$

The Equation 5 means if a source is in grid $C3$, its destination, D_{C3}, is chosen based on the probability in the matrix, listed from P_{A1} to P_{D5}. Note that the sum of the elements in the matrix is equal or lesser than 1. When it is lesser than 1, it means there is a probability of traveling out of the road network. Although the travel is identified based on the status transition of taxis, the travel itself is independent on whether the taxi is loaded with passengers or not. So only one probability matrix is enough for a taxi to plan travels.

4.4 Other considerations

Besides the model parameters aforementioned, we need to consider the number of mobile nodes in the road network and their initial distribution. To get the number of mobile nodes, we also assembled the data of weekdays during the three months and compute the average number of taxis. Figure 7 shows the change of the average number of nodes in the network during a day.

We can see from Figure 7 that the number of taxis in the network varies significantly during a day, from 200 in the morning to 500 in the afternoon. For simplicity, however, we do not consider the variation of the number of nodes in META so it has a constant number of nodes which run in the network. This constant value is determined based on the time of the trace to be generated. When a node leaves the network, we insert a new node on a road section which locates in the outer travel grids such as grid $A1$ or $D3$ in Figure 6. Finally, in order to determine the initial location of nodes for the

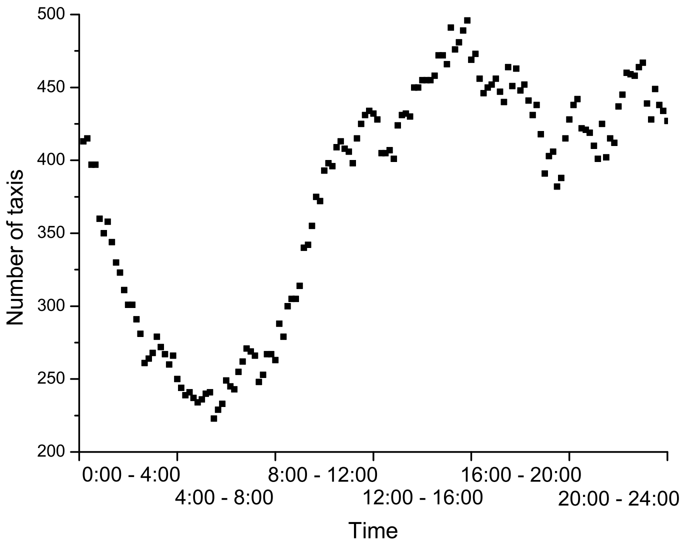

Figure 7: Change of the number of taxis in the network during a day

synthetic trace, we directly take a network snapshot, which is the locations of all nodes in the network at a certain time, from the real trace where the number of nodes in this snapshot is equal to the aforementioned constant number.

5 Model validation

In order to validate META, we compare META trace with the real trace. As is mentioned before, to reconstruct the real trace, we first map the GPS data on the road map. Then the maximum turn probability path between every two consecutive data is determined. Finally we interpolate the GPS data for every one second with the constant speed of s/t where s is the distance between the two data and t is the time interval.

We also generate META trace according to the META mobility model. Before a node begins to move, it randomly chooses a travel grid in the network according to the travel pattern and a road section in that grid as the destination of the travel. Then the node finds the maximum turn probability path between the origination and the destination. Note that such a path is not always available. Although every road section in the network is regarded as bidirectional, some of them are actually one-way road so one of the two road sections is impossible to reach. If a road section which locates on the reverse side of a one-way road is chosen as the destination, the node will repeat the processes to find a reachable destination. After the path is found, the node began to move towards the destination along the path. On each road section along the path, the taxi speed is randomly chosen from $[(1-\alpha)*v,(1+\alpha)*v]$ where v is the road section speed and α is the variation which is set to be 15% in our work.

For simplicity, we simulate the scenario of the vehicular network from 2 pm to 5 pm on the 21st March 2007. So the real trace is reconstructed using the GPS data collected during that time. Figure 7 shows the number of taxis in the network during the time period varies between 450 and 500, we choose to simulate 500 nodes when generating META trace which means there are always 500 nodes in the network. Since nodes can enter and leave network, the simulation result shows that 3161 nodes are generated in META trace. Meanwhile the real trace also recorded totally 2687 taxis during the three hours.

The META mobility model is validated from three aspects: trace characteristics, network topology and performance of the routing protocol.

5.1 Trace characteristics

In order to analyze the approximation of the two traces from a macroscopic aspect, we give the spatial and temporal evaluation of trace characteristics.

A straightforward metric for the spatial measurement is the distribution of nodes. Since the number of road sections is much more than the number of nodes, even if we simulate 500 nodes, the distribution of nodes in a single snapshot is meaningless. So we assemble all of the network snapshots of the three hours and analyze the distribution of how many nodes pass a road section during these three hours which can reflect the popularity of road sections. Figure 8 shows the complementary CDF (CCDF) of the number of nodes which pass a road section.

Figure 8 shows that the real trace and META trace follow the same trend and in both traces there are 9% road sections which have over 240 nodes passed them. The difference between the real trace and META trace is no more than 15% when the number of nodes that pass a road section is less than 420. In the real trace, there are some popular road sections, about 7% of the road sections, which have more than 420 nodes passed them.

From the temporal aspect, we consider the travel time of a node in the network which is defined as the time interval from the time a node enters the network to the time it leaves. Although META does not consider the temporal parameters, all of the three parameters can affect the travel time of a node. For example, longer turn probability path or lower road section speed will increase the node travel time.

Figure 9 shows the CCDF of the travel time of the real trace and META trace. Both the real trace and META trace follow the exponential distribution and the largest difference between them is no more than 15%.

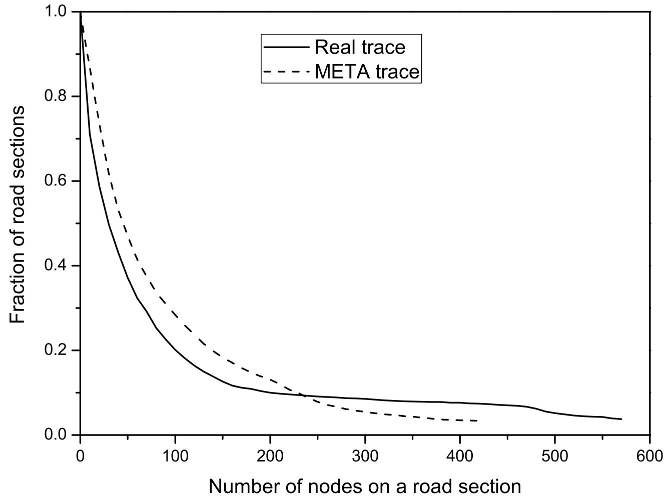

Figure 8: The CCDF of the number of nodes which pass a road section

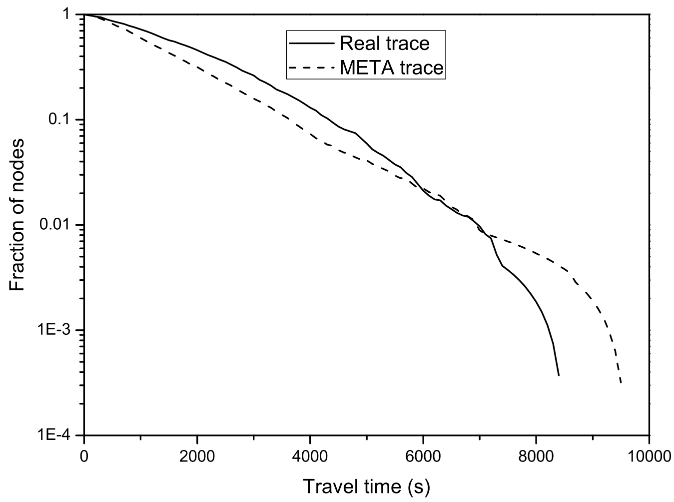

Figure 9: The CCDF of the travel time

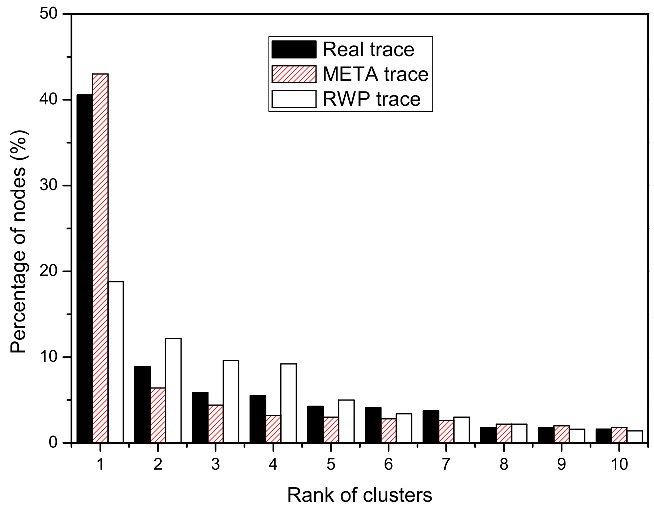

Figure 10: The sizes of the top 10 clusters in a network snapshot

5.2 Network topology

Since the network topology has a great impact on the performance of routing performance, a good understanding of it can help to design an applicable routing protocol. In order to investigate the network topology, we need to specify the communication range for nodes. A pair of nodes can communicate if and only if their distance is smaller than the communication range. There are some existing work which studied the performance of inter-vehicle communication. For example, Singh et al. [Singh et al. 2005] conducted field tests and claimed that the communication range of 400 meters is achievable under suburban environment. Considering the interference in urban scenario, we assume the communication range of a node is 200 meters.

To compare with the real trace and META trace, we also generate a trace according to the RWP model. We simulate 500 nodes which move in a square area of 5km * 4km for three hours. Each node randomly chooses a destination in the network and a speed from $(0km/h, 60km/h]$ and then starts to move. When arriving at the destination, the node starts another travel immediately.

We first analyze the size of clusters, which are connected subnetworks. Since the communication range is not large enough to make the network to be full connected, there can be many clusters in the network. We randomly take a snapshot from each trace and find out all clusters. For each trace, we sort the clusters by their sizes. Figure 10 shows the sizes of the top 10 largest clusters of each trace. Both the real trace and

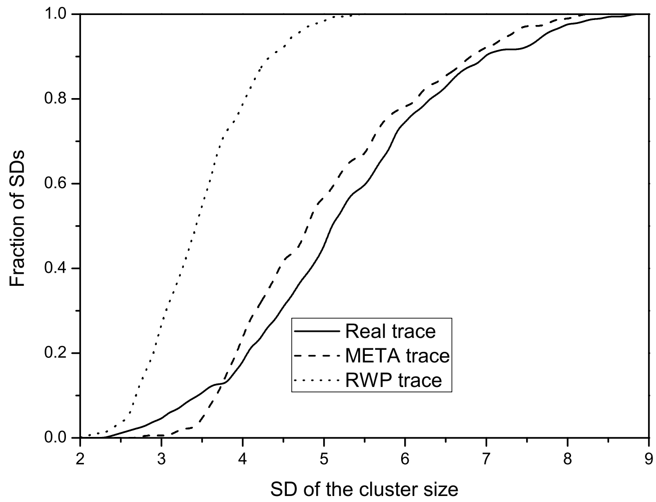

Figure 11: The CDF of the standard deviation of cluster size

META trace have a major cluster which connects over 40% of the nodes in the network. Whilst in RWP trace, all clusters have similar size because of the uniform distribution of nodes. We then compute the standard deviations (SD) of the cluster size for the real, META and RWP trace and the result are 5.0, 4.9 and 2.6 respectively.

Since the SD of cluster size represents the skew degree of cluster sizes, so a larger SD implicates a major cluster in the network and a smaller SD implicates the existence of some relatively large clusters. Hence we also investigate the distribution of the SDs. We choose 180 snapshots from each trace and every two snapshots have an interval of one minute. For each snapshot, the SD of the cluster size is computed and Figure 11 shows the CDF of the 180 SDs. We can see in Figure 11 that 90% SDs of RWP trace are distributed between 2.6 and 4.5. Meanwhile, 90% SDs of META trace are between 3.0 and 7.1 and 90% of SDs of the real trace are between 3.6 and 7.6.

Another important metric of the network topology is the distribution of contact time because it represents the dynamical characteristics of the network. Furthermore, since a longer contact time allows more data transmission during a contact, the knowledge of average contact time determines data transmission strategies [Huang et al. 2007]. For each trace, we randomly choose 3000 contacts which appear between 2 pm and 3 pm and compute the contact time. Figure 12 shows the CCDF of the contact time.

It can be seen from Figure 12 that META trace is close to the real trace and RWP trace appears a different trend. The average contact time of RWP trace is 17 seconds

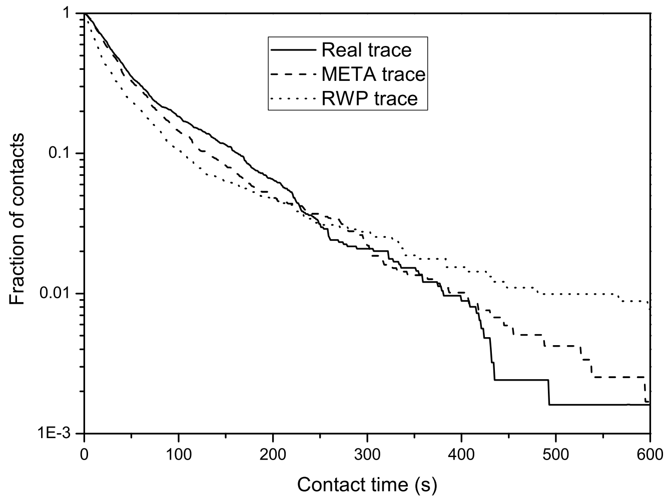

Figure 12: The CCDF of the contact time

while the average contact time of the real trace and META trace are 33 seconds and 30 seconds respectively. Although RWP trace has shorter average contact time, it has more contacts which can last a long time. For example, there are about 1% of the contacts in RWP trace can last over 600 seconds while both of the real trace and META trace have less than 0.2% of contacts which can last such duration.

5.3 Network performance

Finally, we evaluate the network performance by simulating a routing protocol and compare the data delivery ratio. Since the network is partially connected, we simulate the Epidemic [Vahdat and Becker 2000] routing protocol and evaluate the delivery ratio of different traces. We assume the communication between two nodes is duplex, but a node can only communicate with one of its neighbor at one time. We also assume that the size of data packet equals the network bandwidth so that when two nodes communicate, they can transmit a packet per second. The communication range of nodes is set to be 200 meters. Currently, we do not consider the signal collision and only focus on the data exchange. Before the simulation begins, 200 nodes are randomly chosen, 100 of them are sources and the other 100 nodes are destinations. Each source node generates a data packet for the destination node so 100 data packets are totally generated.

Figure 13 shows the CDF of the average delivery ratio of 10 simulations. It can be seen that all data packets can be delivered in 1500 seconds. However, RWP trace has

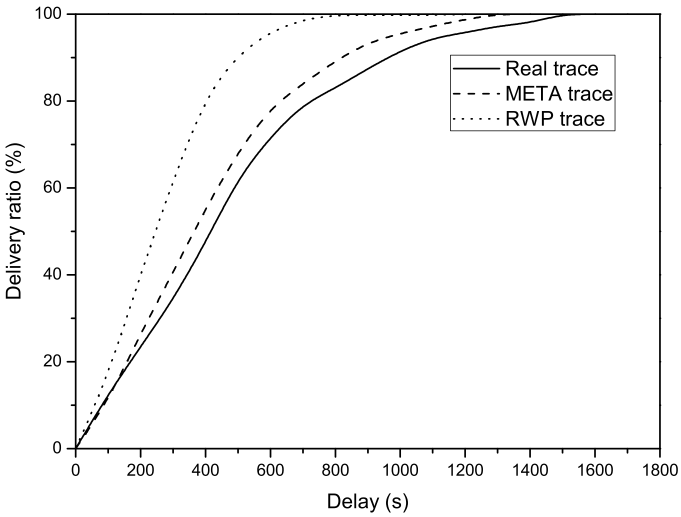

Figure 13: The delivery ratio of the Epidemic routing protocol in different traces

a shorter delay because the nodes in RWP do not have the constraint from streets. The average delay in RWP trace is around 240 seconds while META trace and the real trace have the average delay of 350 seconds and 400 seconds respectively.

6 Conclusion

In this paper, we have presented our study of extracting the META mobility model from the taxi GPS data. In order to characterize the regularity of taxi movement, we defined three model parameters, *turn probability*, *road section speed* and *travel pattern*. We first compute the turn probability at every intersection based on the turn pattern. Then the maximum turn probability path between two GPS data is found to estimate the road section speed. Since the GPS data records the taxi status of whether it is loaded with passengers, we can also estimate its travel pattern. In order to validate the effective of these model parameters, we compare META traces with the real traces which were reconstructed from the GPS data. From the aspects of trace characteristics, the network topology and the performance of Epidemic routing protocol, the validation shows that our model has a good approximation with real scenario.

Although META can precisely capture the characteristic of taxi movement, there is still space for improvement. For example, we intend to capture the feature that some of the popular road sections by refining the grid based travel pattern to a road section level pattern. In addition, we aim to enlarge the network scale to the whole metropolitan area of Shanghai so that people can do more applicable research based on META.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (Grant Nos. 61003247), the Fundamental Research Funds for the Central Universities (Grant Nos. CDJRC10180007 and CDJZR10180010), the Natural Science Foundation of Chongqing (Grant No. CSTC2010BB2210), the Post-doctoral Science Foundation of China (No. 2012M510932) and the Post-doctoral Science Foundation of China (No. YUXM201103013). Also, this work is supported by the National "Qian Ren Plan"of China.

References

[Boudec and Vojnovic 2006] Boudec, J.-Y. L., Vojnovic, M.:"The random trip model: stability, stationary regime, and perfect simulation" IEEE/ACM Transactions on Networking, 14, 2 (2006) 1153-1166.

[Burgess et al. 2006] Burgess, J., Gallagher, B., Jensen, D., Levine, B. N.: "Maxprop: routing for vehicle-based disruption-tolerant networks" Proc. IEEE INFOCOM, Barcelona, Spain (Apr 2006).

[CRAWDAD 2007] CRAWDAD: A Community Resource for Archiving Wireless Data At Dartmouth. (2007) http://crawdad.cs.dartmouth.edu/index.php.

[ESRI 1998] http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf.

[Fiore and Harri 2008] Fiore, M., and Harri, J.: "The Networking Shape of Vehicular Mobility" Proc. of the 9th ACM international symposium on Mobile ad hoc networking and computing (MobiHoc), pp. 261-272, Hong Kong, China(May 2008).

[Hsu et al. 2007] Hsu, W.-J., Spyropoulos, T., Psounis, K., Helmy, A.:"Modeling time-variant user mobility in wireless mobile networks" IEEE INFOCOM, Anchorage, USA (May 2007).

[Huang et al. 2007] Huang, H. Y., Luo, P. E., Li, X., Li, M., Shu, W., Wu, M. Y.:"Performance evaluation of SUVnet with real-time traffic data" IEEE Transactions on Vehicular Technology, 56, 6 (Nov 2007) 3381-3396.

[Jain et al. 2005] Jain, R., Lelescu, D., Balakrishnan, M.:"Model T: an empirical model for user registration patterns in a campus wireless LAN" Proc. ACM MobiCom, Cologne, Germany (Sep 2005).

[Jardosh et al. 2003] Jardosh, A., Belding-Royer, E. M., Almeroth, K. C., Suri, S.:"Towards realistic mobility models for mobile ad hoc networking" Proc. ACM MobiCom, San Diego, USA (Sep 2003).

[Johnson and Maltz 1996] Johnsno, D. B., Maltz, D. A.: "Dynamic source routing in ad hoc wireless networks"; Tomasz Imielinski and Hank Korth, editors, Mobile Computing, volume 353, pages 153-181. Kluwer Academic Publishers, 1996. Chapter 5.

[Kim et al. 2006] Kim, M., Kotz, D., Kim, S.:"Extracting a mobility model from real user traces" Proc. IEEE INFOCOM, Barcelona, Spain (April 2006).

[Krajzewicz et al. 2002] Krajzewicz, D., Hertkorn, G., Rossel, C., Wagner, P.:"SUMO (Simulation of Urban MObility): an open-source traffic simulation" 4^{th} Middle East Symposium on Simulation and Modelling 2002.

[Lelescu et al. 2006] Lelescu, D., Kozat, U. C., Jain R., Balakrishnan M.:"Model T++: an empirical joint space-time registration model" Proc. ACM MOBIHOC, Florence, Italy (May 2006).

[Origin 2009] http://www.originlab.com/index.aspx?s=8\&lm=115\&pid=78.

[Royer et al. 2001] Royer, E., Melliar-Smith, P.M., Moser, L.: "An analysis of the optimum node density for ad hoc mobile networks"; Proc. IEEE International Conference on Communications (ICC), Helsinki , Finland (Jun 2001).

[Singh et al. 2005] Singh, J. P., Bambos, N., Srinivasan, B., Clawin, D., Yan, Y.:"Empirical observation on wireless LAN performance in vehicular traffic scenarios and link connectivity based enhancements for multihop routing" Proc. IEEE WCNC, New Orleans, USA (Mar 2005).

[Vahdat and Becker 2000] Vahdat, A., Becker, D.:"Epidemic routing for partially connected ad hoc networks" Tech. Rep. CS-200006, Department of Computer Science, Duke University, Durham, NC, 2000.

[Vissim 2007] http://www.ptv.de/cgi-bin/traffic/traf_vissim.pl.

[Xu et al. 2009] Li, X., Shu, W., Li, M., Huang, H. Y., Luo, P. E., and Wu, M. Y.: "Performance Evaluation of Vehicle-Based Mobile Sensor Networks for Traffic Monitoring" IEEE Transactions on Vehicular Technology, 58, 4 (May 2009) 1647-1653.

[Yoon et al. 2003] Yoon, Y., Liu, M., Noble, B.: "Random waypoint considered harmful" Proc. IEEE INFOCOM, San Francisco, USA (Apr 2003).

[Zhang et al. 2007] Zhang, X., Kurose, J., Levine, B., Towsley, D., Zhang, H.: "Study of a bus-based disruption tolerant network: mobility modeling and impact on routing" Proc. ACM MobiCom, Montreal, USA (Sep 2007).

[Zhu et al. 2011] Zhu, H., Li, M., Fu, L., Xue, G., Zhu, Y., and Ni, L. M.: "Impact of Traffic Influxes: Revealing Exponential Inter-Contact Time in urban VANETs" IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, 22, 8 (August 2011) 1258-1266.