Location Prediction for Large Scale Urban Vehicular Mobility

Siyu Chen*, Yong Li*,Wenyu Ren*, Depeng Jin*, Pan Hui†
*Tsinghua National Laboratory for Information Science and Technology.
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.
†Deutsche Telekom Laboratories/TU-Berlin, Ernst-Reuter-Platz 7, Berlin 10587, Germany.
Email: liyong07@tsinghua.edu.cn

***Abstract*—Knowledge of where vehicles will be in near future helps users in daily planning, traffic monitors in vehicles scheduling, advertisers in fixed point advertising, and especially helps in communication network source provisioning. In this paper, we analyze the predictability of taxi mobility based on their locations and time period records and we present a prediction method of taxis for their next locations in 15 seconds using Markov predictor. The historical location trace of each taxi is used to train the transition probability matrix of next location for our predictor, and we use 3 different scenarios to predict. Based on records from over 2,000 taxis in Shanghai, and over 14,000 taxis in Beijing, we are able to predict the next vehicular location with an accuracy of 82%.**

***Index Terms*—vehicular network, mobility prediction, Markov chain**

I. Introduction

With the ever-increasing vehicles on the roads and the arising concept of Ambient Intelligence's smart city, vehicular networks that enable wireless communications for vehicles to obtain content from Internet resources or transfer mobile data in them have become increasingly important. Large-scale vehicular cellular access is expected to be available with an increasing number of vehicles equipped with devices to support communication in vehicular network, and interests on vehicular communication networks have grown significantly. Governmental and industrial dedications have been made, for example, in the USA, Federal Communications Commission (FCC) has allocated 75 MHz in spectrum for dedicated near-field vehicular communications, and IEEE is also drafting related standard description. Many consortia and standardisation bodies are actively developing technologies and protocols for information transmission between vehicles and Roadside Unit (RSU) infrastructure equipments, known as Vehicles to Infrastructures (V2I), and between vehicles, known as Vehicles to Vehicles (V2V).In such vehicular networks, a lot of smart applications based on real-time traffic information, such as taxi scheduling, traffic avoidance, touring guidance, and accident administration, become possible.

However, there are many challenges remaining before future driver assistance systems can effectively offer excellent service on vehicle's communication and support traffic situation interpretation, such as the road crossing scenario [1]. One of the most difficult challenges is the vehicular mobility prediction. In terms of the transportation system itself, if the present taxis' states and their prediction of future locations are known, better journey planning and scheduling can be completed. For example, a tourist who arrives at transit airport and plans to have a trip inside the transit city with time limitation such as 8-hour visa in Dubai, will enjoy benefits from the prediction service to arrange taxi rides in series of places around the city. This is one of the various applications in smart transportation. The main problem here is how to handle the vehicular dynamic within the capacity of existing road system by predicting and guiding the vehicular traffic. Besides, in terms of the mobile service provisioning, clients' mobility, especially when they are in vehicular networks causes obvious service loss during handover. Time-continuous services such as audio/video streaming suffer from temporary connection drop during clients handover from one access locality to another one, which is very common in vehicular networks as all the users as well as part of service providers are moving. Therefore, effective and accurate real-time predictions of vehicular mobility are needed. On the other hand, in terms of designing vehicular networks, mobility on communication that is specific to vehicular is gradually explicitly taking into account in research efforts on the development of communication protocols and mobility models of vehicular networks. Thus, we need to better predict vehicular mobility to increase the efficiency of communication in vehicular networks.

At present, the importance of mobility prediction in vehicular networks has received wide acknowledgment in the thesis [2]–[4]. However, most of current prediction works focuse on human mobility [5], [6]. In these works, mobility patterns and prediction limit of human beings are studied. Examples include using Markov chain predictor to study the individual behavior for the purpose of improving the service provisioning in WLAN [7]. Although the concept of predicting the next location based on the previously visited locations is widely used in individual vehicular mobility estimation and network handoff patterns evaluation, the prediction limits are not explicitly explained and remain unknown. Moreover, there is no work about vehicular network predictability limits and prediction verification on large-scale vehicular mobility data to analyze the performance in real urban scenarios.

In this paper, we use two groups of one-month real taxis mobility traces from Beijing and Shanghai, two of the biggest

978-1-4673-2480-9/13/$31.00 ©2013 IEEE

cities in China, which include over 14,000 taxis in Beijing and 2,000 taxis in Shanghai. With our preprocess on the original data, the mobility trace denotes associated history of each vehicle by the cell number in the network under a simple mobility model assumption. Based on this model, we propose a two-stage Markov process based mobility prediction algorithm. We use the designed Markov predictor under 3 different time scales to predict the future locations. The results show that when we use comparatively small number of historical trace records, such as 5,000, to train Markov predictor, the prediction accuracy decreases to 70%. However, when we only use 2,000 previous records to train Marcov predictor the prediction accuracy does not decrease significantly and remains 65%. These obtained results from real large-scale data prediction demonstrate that our proposed algorithm are very near to the limit if we use historical traces to predict.

The rest of the paper is organized as follows. After presenting the related work in Section II, we explain about our vehicular mobility traces in Section III, and give the system models for the vehicular location prediction in Section IV. In Section V, we provide the method to obtain the vehicular mobility prediction limitation, and design a Markov-based prediction algorithm that tries to achieve the obtained prediction limitation. Section VI introduces the experimental environments for performance evaluation and presents the simulation results of different prediction considerations. We conclude the paper in Section VII.

II. Related Work

In recent years, with wide deployment of pervasive technologies associated with location and time points, a huge increase of people's footprints records has been produced. These digital records of human's individual mobility pattern have motivated an increasing interest on researches in human mobility such as prediction of human mobility [4], event-driven traveling pattern [8]. Different from our work they focus on the mobility of human beings rather than moving vehicles.

GPS-capacitated vehicle moving data such as taxi traces, access point records have been collected to analyze rules behind them. In the work of John Krumm [9] a method called Predestination which uses a history of a driver's destinations was developed based on data collected from 169 distinct subjects who drove 7,335 trips. Huang et al. [10] explored the feasibility of vehicle future trajectory prediction in their experimental results where sufficient accuracy for application was achieved. Sebastien et al. developed the MMC algorithm to address the issue of predicting the next location of an individual based on the observation of his mobility behavior over some period of time and the recent locations that he has visited. Their work are usually based on a small amount of data and they lack analysis of the moving patterns in vehicular networks.

III. Trace Introduction and Pre-processing

We first introduce two large-scale urban vehicular mobility motion traces, which play a crucial role in motivating our work. They were collected separately in Shanghai and Beijing. Firstly, we give a brief description of them.

Shanghai trace was collected by SG project [11], in which 2,058 operational taxis continuously covered the whole month of February 2007 without an interruptions in Shanghai city. In this trace, a taxi sends its position reported by GPRS to the central database every 1 minute when it is vacant and every 15 seconds when it has passengers for the aim of real-time scheduling. However the relatively small number of 2,000 taxis and 1 minute duration may not be sufficient to record the statistical features of mobility in a large high-speed urban environment. Furthermore, the different reporting frequency may distort the records of the physical movements of the taxis.

As for *Beijing* trace [12], 27,000 taxis participated in the data collection carrying GPS receivers during May 2010. The reason we choose taxis as vehicular devices lies in that taxis have more sensitive reflections to changes in urban environments in terms of traffic control and urban planning and underlying road topology, and they have broader coverage of operation time and space compared with buses and private cars. The specific information contained in the every 15-seconds report from taxis to data center includes: taxis' ID, longitude and latitude coordinates of the taxi's locations, as well as time stamps. Beijing trace is the largest vehicular data trace available.

We obtain the taxis' moving trace and their moving variation from the longitude and latitude coordinates. Since these locations are measured by GPS devices, the noise may impact the accuracy of the location. Furthermore, coordinates are not suitable to grasp the movement rule of vehicles as the time intervals and report frequency to report locations are not the same for the traces. Thus we need to process the data trace to obtain the accurate locations of all the taxis. To achieve this goal, we first use the city maps of Shanghai and Beijing for the respective traces to correct the taxi's locations so that they are in the regions of related city roads. Secondly, we use linear interpolation (LI) to insert location points so that all the taxis have location information at every 15-second interval. Give a simple example of this method, consider that we have the location of one taxi in the original trace with the location $l_1, l_2, ..., l_n$ recorded at the time points $t_1 < t_2 < \vdots < t_3$. If we want to insert the location information of cell c_t at the time point t which is calculated according to the 15-second frequency. We need to find the time period where $t_n \leq t < t_{n+1}$ and then estimate the location of cell l_t by the following LI

$$l_t = \frac{t_{n+1} - t}{t_{n+1} - t_n} l_n + \frac{t - t_n}{t_{n+1} - t_n} l_{n+1}.$$

To differentiate the location of taxis the whole map is divided into different cells according to the road crossing. In our processing, crossings are regarded as points. Consider a set of coplanar points P. For each point P_x in the set P, we can draw a boundary which encloses all the points lying closer to P_x than to other points in the set P. The parameter seq which ranges from 0.1 to 0.9, decides how big one cell is.

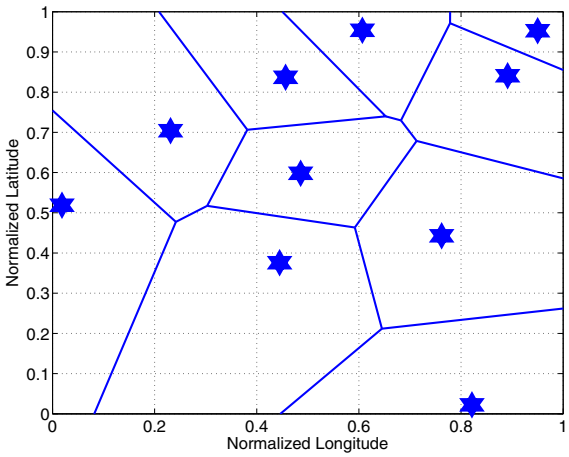


Fig. 1. The method to divide map into different cells

The smaller the seq is, the larger the cell is. To illustrate this method we use a figure (FIG. 1) in which the map is divided into different cells according to a set of points. With the help of LI, the accurate time point when the vehicle enters another cell is known.

IV. MODEL

Before we describe the predictor, we first give an overview of envisions about the vehicles, the data available, and the system for training and predicting. We assume a wireless network where the vehicles are regarded as different nodes traveling from cell to cell. Every node has a device to record its coordinates at any time, as well as its cell number. Besides there is a centralized mechanism to collect the history of every node and perform predictors. Therefore, the trajectory of the vehicles is denoted by a sequence of cell numbers, which together form an abstract model for prediction calculation.

V. PREDICTION METHODOLOGY

We use Markov predictor to predict the destination of the vehicles. The predictor uses transition probability matrix trained from each user's historical or whole trajectory to give prediction with given most recent location sequence. The order-k or $O(k)$ Markov predictor uses a sequence of symbols $(s_1, s_2, ..., s_n)$ as history record to predict the next symbol s_{n+1} from current context, in other words, the sequence of the k most recent symbols in the history $(s_1, s_2, ..., s_n)$. Assume history H=$s_1 s_2 ... s_n$ and subsequence H (i, j)=$s_i s_{i+1} ... s_j$ for any $1 \leq i \leq j \leq n$. As there are explicit states representing the locations of taxis, a multi-dimensional transition probability matrix $M(Y(i, j), s)$ encloses all the predicting information. Consider Y to be a random variable symbol and $Y(i, j)$ to be a sequence representing discrete random variates sequence $Y_i Y_i + 1 ... Y_j$ for any $1 \leq i \leq j \leq n$. Define the context $c = H(n-k+1, n)$. Let $\mathcal{S}$ be the set of all possible symbols. If Y has order-k stationary Markov distribution, for all $s \in \mathcal{S}$ and $i \in 1, ...n-k$, its distribution satisfies

$$\begin{aligned} &P(Y_{n+1} = s | Y(1, n) = H) \\ &= P(Y_{n+1} = s | Y(n-k+1, n) = c) \\ &= P(Y_{i+k+1} = s | Y(i+1, i+k) = c) \end{aligned}$$

Then we can estimate the $M_{Y(i,j),s}$ in transition probability matrix to s as $P(Y_{n+1} = s|H) \approx \widehat{P}(Y_{n+1} = s|H) = \frac{N(cs,H)}{N(c,H)}$ where $N(t', s)$ denotes the number of times the subsequence t' occurs in the sequence t . The markov predictor returns the most possible next symbol as :

$$Y_{n+1} = \arg\max_{s \in S}(P(Y_{n+1} = s))$$

Moreover,we define the $O(k)$ fall-back Markov predictor. Whenever the $O(k)$ predictor doesn't have enough information to predict, namely when the current context has never appeared before, $O(k)$ falls back to $O(k-1)$ Markov predictor. As a result we specially define the "order-0" Markov predictor, which always returns the symbol that occurs most frequently in history H. The early work of [13] shows that $O(2)$ Markov with fall-back performed the best. In the transition probability matrix, 0 stands for sequences or next symbols that never appear. In addition we use the parameter history length to decide how long the training history is when a predictor is requested, and this stands for the jitter of the behavior of vehicles.

Algorithm 1 Predicability Evaluation.

```
Initial the Node Number and Node records
for Node Number i do
  Denote T_i=trace of i
  for Trace T_i do
    if Trace sequence T_i' has appeared then
      Add the destination to this sequence class;
    else
      Add a new sequence class to sequence counter
    end if
  end for
  for T_i do
    Predict the next destination cell according to the previous 2 steps of trajectory, T_2' and use the most frequently appear cell number as the predict cell
    Check the prediction accuracy
  end for
  E_i^rand = log_2 N_i,
  E_i^unc = -sum_{j=1}^{N_i} p_i(j) log_2 p_i(j),
  E_i = -sum_{T_i' ⊂ T_i} P(T_i') log_2 [P(T_i')]
  solve Π^max where E = H(Π^max) + (1 - Π^max) log_2 (N - 1)
end for
```

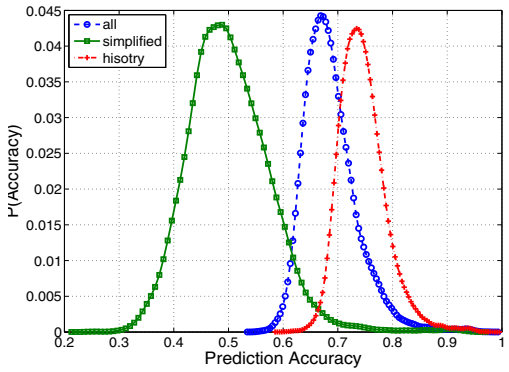
VI. PERFORMANCE EVALUATION

Meaningful performance evaluation metrics are essential requirements for proper estimation description. In many ap-

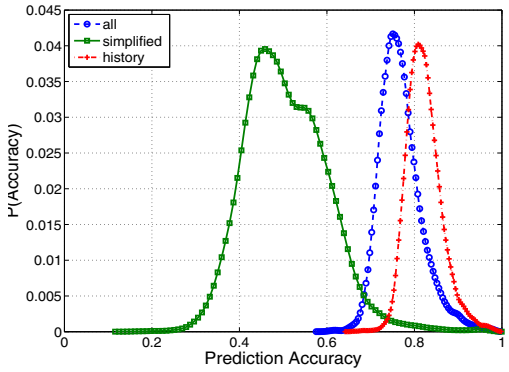
plication based experiments, application-specific metrics are naturally regarded as the evaluation metrics. However, the metrics were affected by many factors other than prediction accuracy, such as channel reservation policy in a cell phone communication application. Thus application metrics were not able to provide direct insight into predictor's quality. We develop the accuracy of the prediction as the main metric to evaluate the performance of predictor at location prediction. We define the ratio between the number of correct predictions and the number of all predictions as the accuracy metric. In the first scenario the prediction process starts from the first location of cell to the last location of cell in the trace, the prediction process lasts during the whole lifetime of the traces of vehicles, which is referred to "all information prediction". In the second scenario, only 2,000 most recent sequence is used to train the predicting result, which is referred to "simplified quick prediction". In the last scenario, all the history record of one node is used to train the prediction of next cell, which is referred to as "history based prediction".

For prediction performance evaluation, the accuracy of each taxi under 3 scenarios are shown in the following figures (FIG. 2, FIG. 3 where the $seq = 0.1, 0.3, 0.5$ separately. From the real large-scale data, we can see in both *Beijing* and *Shanghai* dataset the third scenario has the highest prediction accuracy which peaks at about 0.7; the first scenario, has the second highest prediction accuracy which peaks at about 0.65; the second scenario, also referred to as "simplified prediction" has the worst prediction results, peaking at about 0.48, because in this scenario we sacrifice accuracy for predicting speed. To our surprise, the "history based prediction" has better accuracy than "all information prediction", it is against our intuitive conjecture. We find the main difference between the two scenarios appears at the beginning period of the predicting process, because when at the later period we get most of the all information as the history records, the two scenarios are approximately very similar. However, at the beginning, "history based prediction" has less information but more related to the beginning period, while "all information prediction" has more information but in term with this period it brings more noise. That's why more information may not result in higher accuracy. In addition, the results agrees with the economic assumption that the drivers want to spend the least time when taxi is vacant and usually drive passengers around a certain area, which means that their behavior is of more predictability. To be more specific, the drivers may drive in one section for some days and gradually move to another section as passengers they drive want to go to. On the other hand, the taxis also face passengers who want to have a long ride to another distant place, in this situation the predictability of the next cell is much smaller.

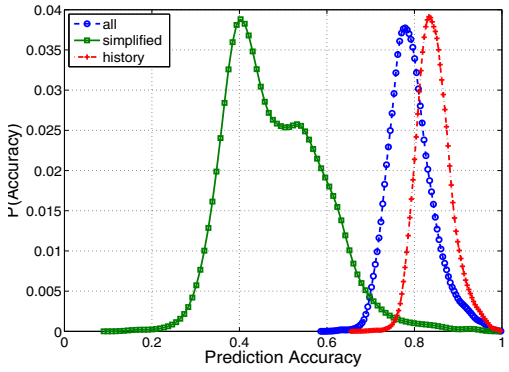
The results obtained from *Beijing* traces are shown in Fig. 2. There is an obvious trend that the bigger the seq is the bigger point the $P(Accuracy)$ under 3 scenarios peaks at. The $P(Accuracy)$ for history based prediction peaks at 0.74($seq = 0.1$), 0.81($seq = 0.3$) and 0.82($seq = 0.5$). Contrary to our conjecture, if we divide the map into more



(a) *seq=0.1*



(b) *seq=0.3*

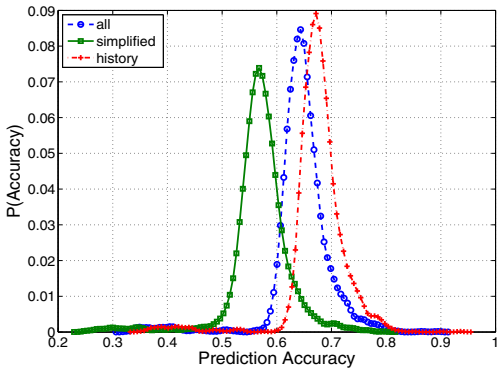


(b) *seq=0.5*

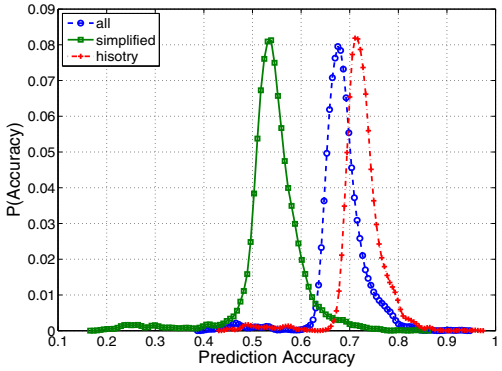
Fig. 2. Distributions of prediction accuracy under 3 scenarios with *Beijing* dataset when *seq* is different.

parts, the prediction accuracy increase, which reflects the deep predictability movement pattern in taxi behavior.

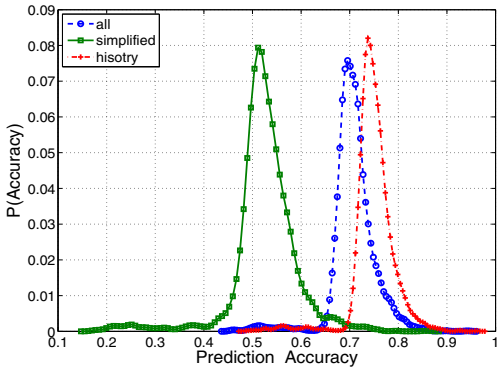
The results obtained from *Shanghai* are shown in Fig. 3. Different from the those results in *Beijing*, the prediction accuracy in 3 scenarios of *Shanghai* are more aggregate, as the "all information prediction" peaks at 0.64, the "simplified quick prediction" peaks at 0.57, and the "history based prediction" peaks at 0.67 when $seq = 0.1$. Similar to the results in *Beijing*, the seq and the prediction accuracy have some inner relationship, that the bigger the seq the higher the



(a) *seq=0.1*



(b) *seq=0.3*



(b) *seq=0.5*

Fig. 3. Distributions of prediction accuracy under 3 scenarios with *Shanghai* dataset when *seq* is different.

prediction accuracy. The highest prediction accuracy appears when $seq = 0.5$ under history based prediction with the accuracy of 77%.

To summarize, the results consistently show that a deep-rooted regularity behind vehicles' daily mobility. The prediction accuracy results in real large scale urban city dataset range from 57% to 82%.

VII. Conclusion

This paper explores the feasibility of vehicular network mobility prediction and the predictability of its movement behavior. The Markov predicting algorithm for next cell prediction based on previous movement records is proposed; and the prediction based on large-scale taxi dataset is studied. The prediction accuracy in 2 city *Beijing* and *Shanghai* ranges from 0.56 to 0.82 under 3 different scenarios: "all information", "simplified quick" and "history based". In order to further understand the mobility pattern in vehicular networks, we are planning as future work to analyze the behavior pattern of taxi drivers searching for possible passengers in large scale data.

Acknowledgment

This work is supported by National Basic Research Program of China (973 Program Grant No. 2013CB3291005), National Nature Science Foundation of China (Grants No. 61171065, No. 61021001 and No. 61133015), National High Technology Research and Development Program (Grants No. 2013AA010601 and No. 2013AA010605), and Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT).

References

[1] C. Hermes, C. Wohler, K. Schenk, and F. Kummert, "Long-term vehicle motion prediction," in *Intelligent Vehicles Symposium, 2009 IEEE*, pp. 652–657.

[2] I. Okutani and Y. Stephanedes, "Dynamic prediction of traffic volume through kalman filtering theory," *Transportation Research Part B: Methodological*, vol. 18, no. 1, pp. 1–11, 1984.

[3] H. Mahmassani, "Dynamic network traffic assignment and simulation methodology for advanced system management applications," *Networks and Spatial Economics*, vol. 1, no. 3, pp. 267–292, 2001.

[4] C. Song, Z. Qu, N. Blumm, and A. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[5] M. Shah, P. Verma, S. Merchant, and U. Desai, "Human walk aware mobility resistant efficient clustering for data gathering in cell phone based wireless sensor networks," in *Wireless and Optical Communications Conference (WOCC), 2011 20th Annual*. IEEE, pp. 1–6.

[6] M. Zhao, L. Mason, and W. Wang, "Empirical study on human mobility for mobile wireless networks," in *Military Communications Conference, 2008. MILCOM 2008. IEEE*, pp. 1–7.

[7] L. Song, U. Deshpande, U. Kozat, D. Kotz, and R. Jain, "Predictability of wlan mobility and its effects on bandwidth provisioning," in *Proceedings of INFOCOM*, vol. 6, 2006.

[8] F. Calabrese, F. Pereira, G. Di Lorenzo, L. Liu, and C. Ratti, "The geography of taste: analyzing cell-phone mobility and social events," *Pervasive Computing*, pp. 22–37, 2010.

[9] J. Krumm and E. Horvitz, "Predestination: Inferring destinations from partial trajectories," *UbiComp 2006: Ubiquitous Computing*, pp. 243–260.

[10] J. Huang and H. Tan, "Vehicle future trajectory prediction with a dgps/ins-based positioning system," in *American Control Conference, 2006*. IEEE, pp. 6–pp.

[11] H. Zhu, M. Li, L. Fu, G. Xue, Y. Zhu, and L. Ni, "Impact of traffic influxes: Revealing exponential intercontact time in urban vanets," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 22, no. 8, pp. 1258–1266, 2011.

[12] Y. Li, D. Jin, P. Hui, L. Su, and L. Zeng, "Revealing contact interval patterns in large scale urban vehicular ad hoc networks," in *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*, pp. 299–300.

[13] L. Song, D. Kotz, R. Jain, and X. He, "Evaluating location predictors with extensive wi-fi mobility data," in *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, vol. 2, pp. 1414–1424.