# On Geocasting over Urban Bus-Based Networks by Mining Trajectories

Fusang Zhang, Beihong Jin, *Member, IEEE*, Zhaoyang Wang, Hai Liu, Jiafeng Hu, and Lifeng Zhang

*Abstract*—Bus networks in cities have distinctive features such as wide coverage and fixed bus routes so that they show the potential of forming the communication backbone in vehicular ad hoc networks (VANETs). This paper focuses on the geocast in bus-based VANETs and presents a geocast routing mechanism named Vela. Specifically, Vela analyzes and mines historical bus trajectories and characterizes spatial–temporal patterns (i.e., bus travel-time patterns and bus spatial encounter patterns) in a moderate granularity of road segments, which makes the mined patterns both accurate and steady. Furthermore, Vela exploits these acquired patterns to build a probabilistic spatial–temporal graph model and provides the available routing paths with the best possible quality-of-service levels for data delivery requests. Moreover, Vela also employs a two-hop aware strategy that utilizes the real-time spatial–temporal relationships between buses to increase the chances of forwarding the data. The results of the experiments on the real and synthetic trajectories show that Vela performs much better in terms of delivery ratio and delay and has stronger scalability than the other solutions.

*Index Terms*—Vehicular ad hoc networks, bus-based routing, trajectory mining, time series analysis.

## I. Introduction

A Vehicular Ad hoc Network (VANET) is a kind of mobile and self-organizing wireless communication network. Combined with the other technologies (such as sensor technologies, positioning technologies), a VANET has been envisioned to be an underpinning of many valuable applications, such as emergency message delivery [1], geographic advertising [2], and various location-based services [3] for drivers and passengers. Geocast [4], delivering messages to specific geographic locations, is an essential communication paradigm to help to achieve these applications in a VANET. Compared with 3G and Wi-Fi networks, multi-hop vehicular *ad hoc* communications through Dedicated Short-Range Communications (DSRCs) have a cheap deployment cost and a relatively high bandwidth capacity for free [5]. Moreover, some real-time messages can be detected and disseminated efficiently via VANETs. These advantages enable VANETs to be the promising message exchange platforms in modern cities and play significant roles in different application domains.

As one of urban public transport systems, the bus network covers almost entire city, where the buses move along the fixed routes on the relatively regular schedules [6]. Therefore, the buses equipped with the wireless communication units are suitable for constituting a self-organizing network to deliver data. However, achieving the fast and reliable geocast in a bus-based *ad hoc* network is still a challenge due to the following reasons. First, because of the high speeds of vehicles, a VANET suffers from rapid topology changing, communication intermittency and network partitions. Second, the urban traffic environments varied with the time of a day worsen the message delivery quality, which may give rise to the loss of data delivery opportunity. Third, how to mine the models of bus travel time and patterns of the bus encounters and effectively apply to the data routing mechanism is still a problem that has not been solved perfectly. Up to now, some research work focuses on geocast over VANETs. Unfortunately, any dedicated routing mechanism which can distinguish between time-varying delivery conditions by mining real bus trajectory data has not been seen.

We focus on the appealing applications which need disseminate location-based service (LBS) messages (such as geographic advertisements, gas prices, traffic conditions, states of tourist points of interest, and so on) to a geographic region such as a point of interest (POI, referring to a rectangular area covering at least one road segment in this paper). Fig. 1 shows the scenario where an LBS message is disseminated to a geographic region (i.e., a POI) via a bus-based network. The whole message dissemination process can be divided into three phases. First, a mobile vehicle Veh1 initiates an LBS dissemination request and posts the message to a nearby bus of route $A$. Second, the bus of route $A$ carries the message until encountering a bus of route $B$ going to the POI. Thus, message delivery can be achieved by the coordination between the buses of different bus routes. Finally, the bus of route $B$ broadcasts the LBS message within the POI. As a result, all vehicles within
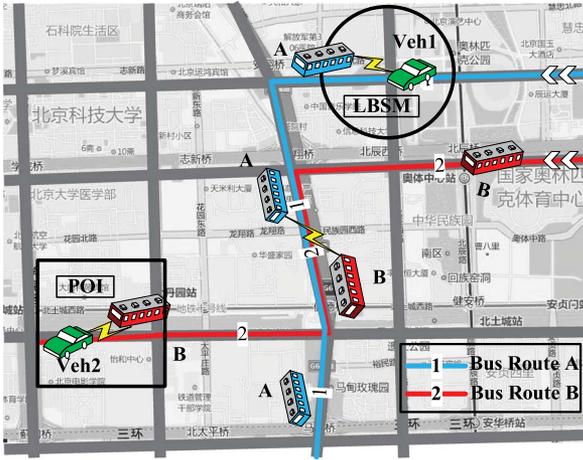
Fig. 1. Example of application scenarios.

the POI can receive the message. Some existing approaches [7], [8] have designed the protocols for the last phase to reliably broadcast messages within the POI. Therefore, we focus on the second phase of the above scenario, that is, delivering a message from a source node to the destination road segment with the goal of making the delay as low as possible and the reliability as high as possible. This routing mechanism not only is a key to realizing Vehicle to Vehicle (V2V) communications but also can be applied to Infrastructure to Vehicle (I2V) or Vehicle to Infrastructure (V2I) communications. Here, in consideration of the deployment cost of a large number of Road Side Units (RSUs), any RSU deployment, including deploying the RSUs only at bus-stops or intersections, is not assumed.

In this paper, we propose a novel geocast routing mechanism named Vela, which elegantly mines and utilizes the spatial-temporal patterns of buses for fast routing. We highlight our main contributions of this paper as follows:

- By observing the topology of a real-world bus network, we discover some inherent characteristics of the bus trajectories. Further, to reveal the spatial-temporal patterns of bus trajectories, we build the Auto Regressive and Moving Average (ARMA) models for travel time patterns and find the fitting polynomials for spatial encounter patterns, which enable us to make accurate predictions for travel times and encounters of buses.
- For the first time, we characterize the spatial-temporal patterns of buses in a moderate granularity of road segments. The resulting patterns not only overcome the shortcoming of low accuracy of the coarse-grained (i.e., bus line level) patterns, but also avoid the low stability of the fine-grained (i.e., bus level) patterns.
- Upon these, we propose a novel geocast routing mechanism for bus-based *ad hoc* networks, which builds a probabilistic spatial-temporal graph model for data delivery, chooses the path with the highest reliability from the top-$k$ minimum delay paths as the message delivery path, and provides an optimized two-hop aware strategy. Thus, the novel mechanism both exploits the spatial-temporal fea-

tures implied in the historical trajectories and utilizes the real-time spatial relations between buses on the roads.
- We conduct the experiments on real-world and synthetic trajectories, respectively. Compared to existing routing mechanisms over bus-based networks, the experimental results show that our routing mechanism has better performance and higher scalability.

The remainder of this paper is organized as follows. Section II contains a review of the related work. Section III gives some analyses of real-world bus trajectories. In Section IV, we describe a probabilistic spatial-temporal model for a bus-based network. Section V reveals the details about the proposed geocast routing mechanism which is based on the probabilistic spatial-temporal model; and in Section VI, we evaluate the performance of Vela and analyze the simulation results. Finally, we conclude our paper in Section VII.

## II. RELATED WORK

### A. Message Delivery in Vehicular Networks

The basic strategy to deliver messages over a VANET is the carry and forward, i.e., the vehicles carry the messages until encountering other vehicles to which they can forward the messages. To select an appropriate relay vehicle, various real-time and historical information, including geographical locations of vehicles, historical traffic statistics, vehicle trajectories, and inter-vehicle encounter patterns, is exploited [5], [9]–[13].

As for geocast over VANETs, the traditional geocast schemes for MANETs (such as GeoTORA [4] and GeoGRID [14]) are not suitable for VANETs because they usually require to frequently update their knowledge of the network, which leads to an unaffordable burden for a VANET. Some recent work focuses on geocast over VANETs. GeoMob [15] uses the k-means clustering on the GPS reports of vehicles to generate the regions, each of which contains different traffic volumes. On this basis, a message is first forwarded to the region where the destination locates, and then forwarded to the destination by exploiting the mobility patterns of individual vehicles within the region. We note that the message delivery between regions mainly depends on the taxies or buses which move across the regions frequently. Therefore, GeoMob may miss many cooperative delivery opportunities and lead to a long delay. Besides, GeoMob has no scheme of adjusting the regions with the dynamic change of traffic flows in a day, but in general such a scheme is supposed to help improve the effectiveness of routing. [16] views the trajectories provided by the on-board navigation systems as the vehicular future trajectories. Based on the future trajectory of the vehicle itself and the vehicles to be encountered, [16] calculates the coverage capability of the vehicle over the target region of a message to be delivered, and then makes the decision on message forwarding in terms of the coverage capability. However, the above approach is not applicable to a bus-based VANET, because the travel times of buses do not follow the gamma distribution (see Section III-C) which is the premise of the approach in [16].

## B. Routing in Bus-Based Networks

We classify the existing routing mechanisms for the bus-based networks into two categories by the granularity at which the underlying bus mobility pattern is explored.

The first category is to explore mobility patterns of individual buses (bus level) to support routing [17], [18]. [17] deploys 30 buses and builds a bus-based network called UMassDiesel-Net in the UMass Amherst campus. It utilizes the encounter probabilities of individual vehicles to estimate the costs of message delivery for a routing selection. [18] investigates the vehicles with repetitive motions and decides the routing paths through the expected minimum delay of messages. [18] also conducts experiments on UMassDieselNet traces. However, the approach in [18] is not suitable for urban bus systems, because while compared with the urban bus routes, the campus bus line usually has few bus stops and a reliable schedule. In brief, in the realistic bus networks, there is no evident regularity for individual buses so that the application scenarios which can apply the first category of methods are limited.

The second category is to investigate the mobility patterns of bus routes (bus line level) to facilitate the routing. BLER [19] utilizes the patterns at the bus line level for routing over bus-based networks for the first time. The routing selection is based on maximizing contact lengths between different bus routes. When the message reaches the destination bus line, a zigzag process is triggered to route it to the destination bus, that is, a bus transmits the message only to another bus running on the same line. R2R [20] analyses the bus pair encounter frequency and finds that around 44% bus pairs encounter only once over five weekdays. R2R [20] utilizes the same method as BLER but modifies the route selection metric into the encounter frequency of bus routes to achieve better performance than BLER behaves.

Besides, [21] takes advantage of the feature of bus stops, i.e., the buses from different bus routes may stop at the same bus stops. Therefore, [21] deploys wireless communication units at the bus stops apart from buses, and then provides bus-stop communications to relay the data from one bus line to another. However, under such a mechanism, the encountered buses cannot exchange information directly.

Different from the fine-grained (i.e., bus level) approaches [17], [18] and coarse-grained (i.e., bus line level) approaches [19], [20], we propose Vela which elegantly captures spatial-temporal patterns of bus-based networks at an intermediate granularity and obtains the better performance and scalability than the existing approaches.

## III. BUS TRAJECTORY ANALYSES

### A. Collecting Bus Trajectory Data

First of all, to obtain a macroscopic picture, we project onto the map the 146 bus routes operated outside of the second ring road of Beijing, China in March 2013. Fig. 2 shows the coverage status of the road network resulting from these bus routes. It shows that a real-world bus network indeed can be a VANET communication backbone. However, the bus transportation system is usually affected by traffic flows and emergency accidents. From a microscopic point of view, the
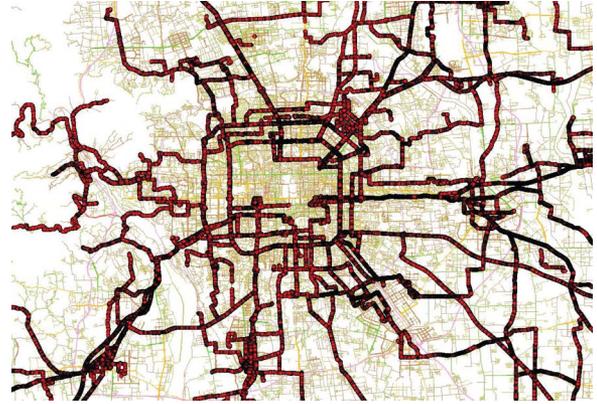


Fig. 2. Bus routes in the road map of Beijing.

TABLE I
GPS REPORT OF THE BUS

| Item | Value |
|---|---|
| *Timestamp* | 2013-03-01 21:49:36 |
| *Bus ID* | BJ G31279 |
| *Bus line number* | 939 |
| *Longitude* | 116.494115 |
| *Latitude* | 40.057195 |
| *Speed (km/h)* | 29 |
| *Azimuth angle* | 192 |
| *Up/down line mark* | down |
| *Next stop number* | 14 |

operations of a bus system often tend to deviate from the expectations of bus scheduling. To exploit the bus mobility and encounter patterns, we analyze the real bus trajectory data so as to discover the spatial-temporal patterns.

We choose three typical bus routes: bus route No. 939, No. 944 and No. 983 which have an average of 20, 60, and 42 buses per day respectively. All of the three routes go through the trunk road from Beichen Bridge to Anhui Bridge with the road length of 1920 meters, which is hereinafter referred to as a sampling road segment. Each day from 6:00 A.M. to 10:00 P.M., each bus sends back a GPS report every 20 seconds. Table I shows an example of GPS reports. The specific information contained in such a report includes: timestamp, bus ID, bus line number, the longitude and latitude of current location, moving speed, azimuth angle, up/down line mark, etc.

Due to the interference and loss of wireless signals, we first preprocess the collected bus GPS data, including amending the drifted GPS data, inserting the lost GPS data. Second, we extract the trajectory data of buses of a specified bus route on the sampling road segment according to the bus ID and up/down line mark. Finally, we explore the spatial-temporal characteristics of bus trajectories by statistics analysis.

### B. Topology Analysis

As a matter of fact, message delivery is based on the connectivity of buses. In urban bus-based networks, we should first figure out some basic problems. For example, are the bus-based networks well connected or highly partitioned? Can buses form connected cliques and which size can the cliques of multi-hop
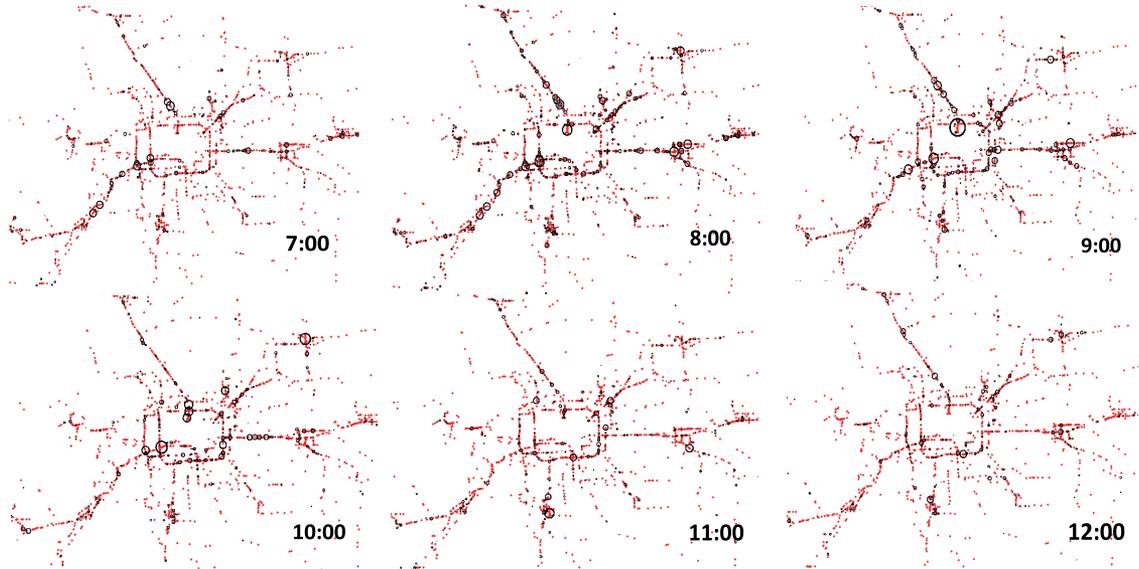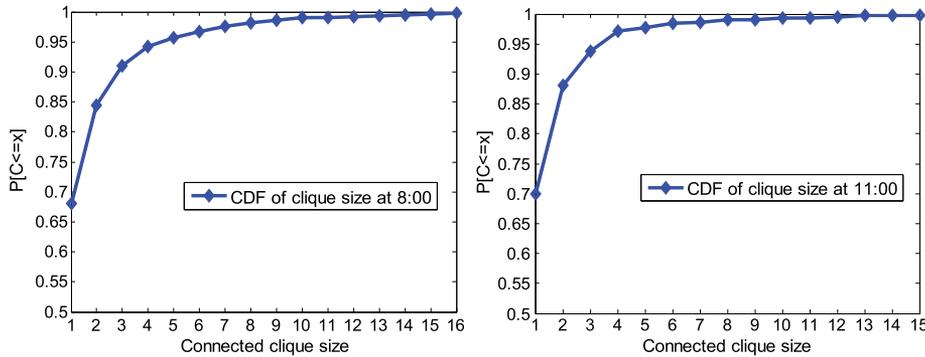
Fig. 3.   The topology snapshots with $R = 250$ m.



Fig. 4.   The cumulative distribution functions of C(t).

connected buses attain? How does the connected clique change as the communication range of buses increases? How do the connectivity features vary in time? The answers to these questions not only directly show the topology features of the bus network but also drive us to improve the design of geocast routing protocol. Therefore, we capture the real-world topologies from the bus trajectory data of Beijing and analyze them from the following three aspects.

*1) The Topology Snapshots Over Time:* The level of connectivity of the buses-based network is characterized through the connected clique, denoted as $C(t)$, that is a cluster of buses which can reach each other via multi-hop communication at time $t$. The clique size is the number of buses in the clique. The number of connected clique reflects the level of network fragmentation.

As shown in Fig. 3, on the topology snapshots at 6 different time points (from 7:00 to 12:00) which contain 1781, 1942, 2181, 1978, 1804, and 1591 buses respectively, we plot the connected cliques. We set communication range $R$ to 250 m, as this is the value usually used by the existing work [22], [23]. The black circles in Fig. 3 represent the connected cliques, which show that the bus-based network is highly partitioned

into thousands of separate individuals. Hence, the routing in bus-based network should base on the carry-and-forward strategy. During the peak hour from 8:00 to 9:00, the number of buses increases and tends to form connected cliques. The cumulative distribution functions (CDFs) of the clique size are depicted in Fig. 4. It shows that the network is largely made of small cliques and above 65% of vehicles is isolated. The cliques whose sizes are equal to and greater than 2 account for 32% of all the cliques at 8:00 and 30% at 11:00, respectively. However, the distributions of cliques are different at the different time points. Inferring from the above topology observations, we can say both the bus travel time on the same road and the connectivity are time-varied. Such temporal features are important and should be exploited where the bus travel time is employed to decide the routing.

*2) The Connected Cliques Affected by Different R:* We now observe how different communication ranges impact the connected cliques of a topology. We find from Fig. 5, when $R$ is set to 500 m, the connected cliques are formed easier and the sizes of cliques are larger than the ones while $R$ is 250 m. The ratios of cliques (size > 2) are 42% at 8:00 and 43% at 11:00, respectively. Therefore, the communication

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

ZHANG *et al.*: ON GEOCASTING OVER URBAN BUS-BASED NETWORKS BY MINING TRAJECTORIES
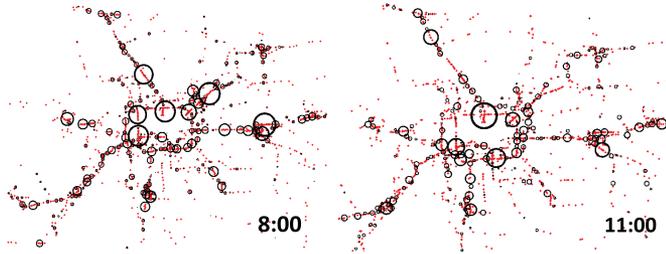
5

Fig. 5. The topology snapshots with $R = 500$ m.

range has a nonnegligible impact on connectivity, furthermore it will impact on the performance of the routing schemes. In Section VI, the experiments are conducted to give us a detailed quantitative analysis of routing performance in terms of the communication range.

*3) The Distance Distribution Between Buses:* The distance distributions for all buses and the buses of same lines are provided by Fig. 6. The distance distribution implies the condition of network partition and indicates that the traditional routing schemes [24], [25] are inefficient. In Fig. 6(b), we note there are 14.1% of buses that can direct a deliver message to the front bus of the same line if R is set to 250 m. Meanwhile, there are 8.7% of buses that can deliver messages to the front bus of the same line by two hops via another bus. This characteristic is utilized to enhance the efficiency for routing in Section V.

### C. Observing Spatial-Temporal Patterns

In this subsection, we focus on exploring two kinds of patterns from bus trajectories, i.e., the travel time patterns and the spatial encounter patterns. This is because in a bus-based *ad hoc* network, buses follow relatively stable schedules, the uncertainties of their mobilities are greatly decreased, and mobility patterns of the buses have the potential to play a significant role in a geocast routing. Especially, the routing performance can be improved greatly with the help of choosing an appropriate mining granularity and a full utilization of the spatial-temporal patterns between the buses.

*1) Travel Time Patterns:* The travel times of buses on a road segment, which may vary in one day with varied traffic conditions, reflect the rough transmitting times of messages in the carry-and-forward paradigm.

Fig. 7 shows the average travel times of buses on the sampling road segment from Mar. 4, 2013 (Monday) to Mar. 10, 2013 (Sunday). In Fig. 7, the travel times on weekdays demonstrate obvious fluctuation in different time periods of a day. In particular, the bus travel times from 7:00 A.M. to 9:00 A.M. are two times as long as those of the other time periods. Meanwhile, the travel times around 6:00 P.M. increase slightly. However, although the travel times vary during a day, the changing trend of travel times on the weekdays follows a similar pattern. By contrast, the travel times at the weekend are relatively stable. Especially, the standard deviation is only 11.7 seconds.

Learned from the previous work [26], [27], the vehicular travel times of a road segment are supposed to follow a gamma distribution or a log-normal distribution. In order to observe whether the bus travel times follow these distributions, we conduct an empirical study with the real trajectories of buses. From the trajectories of 7 days from Mar. 4, 2013 to Mar. 10, 2013, we collect 1565 time values of different buses travelling on the sample road segment and plot the histogram in Fig. 8. We use the gamma distribution to fit the travel time values and obtain the parameters via the maximum likelihood estimation. However, the Kolmogorov-Smirnov (K-S) test rejects the hypothesis with a significant level of 95 percent. This indicates the travel time values of the buses do not follow the gamma distribution. We also examine whether the travel times fit a log-normal distribution by a similar way. The result of the K-S test shows that the travel times do not follow log-normal distribution either. We attribute this unexpected result to the fact that the driving manner of buses is obviously different from the ordinary vehicles. As for a bus, it has to stop at specified bus stops, stays for a short time and continues to drive. Therefore, we have to adopt another method to capture the pattern of travel times in different time periods (see Section IV-B).

*2) Spatial Encounter Patterns:* We make an assumption that two buses could have a chance to exchange a message (called an encounter) if the distance between them is within a given communication range. Next, we give the formal definition of the probability of the encounter between two bus routes as follows.

*Definition 1 (Encounter Probability):* The encounter probability between bus route $A$ and bus route $B$ on a road segment $r$, denoted as $P(A_r, B_r)$, is defined as the ratio of the encounter frequency of bus route $A$ and $B$ to the travel frequency of bus route $A$ within a certain period of time, namely:

$$P(A_r, B_r) = \frac{\sum_{ts=s}^{d} f_{ts}(A_r, B_r)}{\sum_{ts=s}^{d} f_{ts}(A_r)} \qquad (1)$$

where $ts \in [s, d]$ is a time slice, and for a specific time slice $ts$, the indicator function $f_{ts}$ is defined as

$$f_{ts}(A_r, B_r) = \begin{cases} 1 & \text{if a bus of } A \text{ contacts a bus of } B \text{ on } r \\ 0 & \text{otherwise} \end{cases}$$

$$f_{ts}(A_r) = \begin{cases} 1 & \text{if a bus of } A \text{ appears on } r \\ 0 & \text{otherwise.} \end{cases}$$

As indicated from the definition, the binary relation $P(A_r, B_r)$ is non-commutative (i.e., $P(A_r, B_r) \neq P(B_r, A_r)$), and $P(A_r, B_r)$ reflects the success probability that a bus of route $A$ forwards a message to a bus of route $B$.

We calculate the encounter probability of the bus route No. 944 and bus route No. 939 on the sampling road segment. The time slice $ts$ is set to 20 seconds, which is the same as the GPS reporting interval. Fig. 9 shows the encounter probabilities of two bus routes on the weekdays of four consecutive weeks (i.e., 4th–8th, 11th–15th, 18th–22th, and 25th–29th March, 2013). We try to examine whether the encounter probabilities during four weeks show periodic variations. So we convert the encounter probabilities into the instantaneous frequencies by the Fourier transformation. The resultant power spectrum is

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

6                                                                                                                    IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS
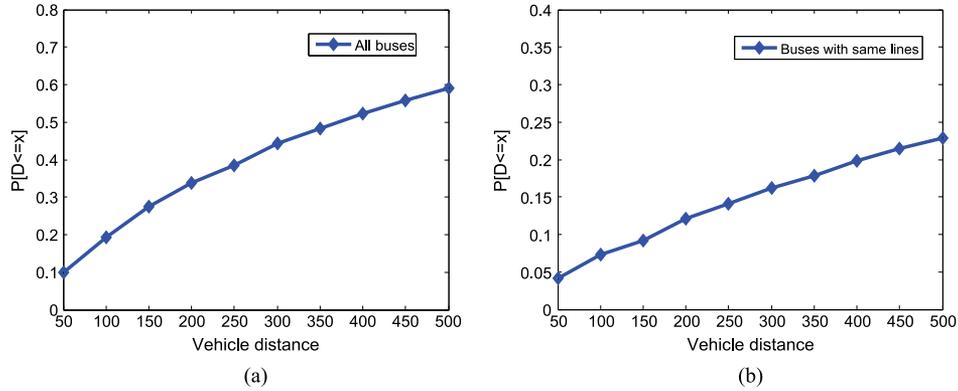


Fig. 6.   The cumulative distribution functions of distance. (a) All buses. (b) buses of same lines.
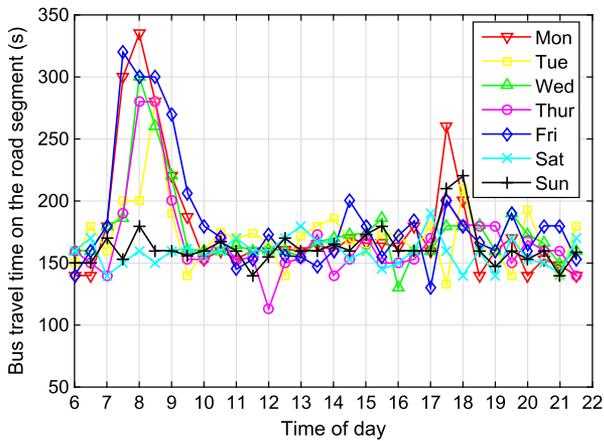


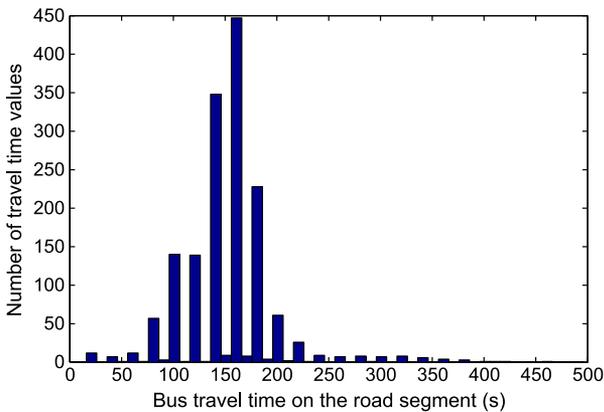Fig. 7.   Bus travel times on the sampling road segment.



Fig. 8.   Distribution of bus travel times.

shown in Fig. 10. The spikes can be seen from the graph of the power spectrum, which indicates that the encounter probability serials do be periodic.

From the above analyses, we reckon that the patterns of the travel times and the patterns of spatial encounters existing in the trajectories, but these patterns differ from those ones of ordinary vehicles. Driven by this viewpoint, we manage to build the models of bus travel times and explore the spatial-temporal patterns of the bus encounters, and then apply them to the routing mechanism.

## IV. PROBABILISTIC SPATIAL-TEMPORAL MODELS

### A. Overview

We next give the definition of a probabilistic spatial-temporal graph which acts as an abstract layer over a bus-based network for message delivery.

*Definition 2 (Probabilistic Spatial-Temporal Graph):* A probabilistic spatial-temporal graph is constructed based on the bus trajectories over a set $\omega$ of bus routes, denoted as $G = <V, E>$, including a vertex set $V$ and a directed edge set $E$. Specifically, a vertex $A_{r_i}$ in $V$ denotes the bus route $A$ which passes through the road segment $r_i$, and a directed edge $e \in E$ from $A_{r_i}$ to $A_{r_{i+1}}$ (TYPE I) indicates that a bus of route $A$ can drive from the road segment $r_i$ to the adjacent segment $r_{i+1}$. However, a directed edge $e \in E$ from $A_{r_i}$ to $B_{r_i}$ ($A \neq B$) (TYPE II) indicates that the road segment $r_i$ is a common road segment of bus route $A$ and $B$ and the buses of route $A$ and $B$ may encounter and exchange messages on the road segment $r_i$.

As introduced in Definition 1, $P(A_{r_i}, B_{r_i})$ is the probability that the buses of bus route $A$ and $B$ encounter on road segment $r_i$. The message forwarding opportunities depend on the encounter probability $P$. The probability of transition from vertex $A_{r_i}$ to $A_{r_{i+1}}$ is a tautology proposition, and then we have the addition of Definition 1:

$$P\left(A_{r_i}, A_{r_{i+1}}\right) = 1.$$

Fig. 11 shows an illustration of probabilistic spatial temporal graph, large circles represent bus routes and the TYPE I edges connect the vertexes within the same bus route. TYPE II edge bridged two large circles is the transit between bus route A and B.

*Definition 3 (Weight Tuple):* The weight of edge $e$ is denoted as a tuple $< P, D >$, consisting of the encounter probability $(P)$ and delay $(D)$. The encounter probability $P$ has been discussed in Definition 1. The delay refers to the transmission latency between vertices. There are two types of latencies, corresponding to the two types of edges. For the edge $e$ of TYPE I, that is, the message is carried by a bus from road segment $r_i$ to $r_{i+1}$, the delay $D(A_{r_i}, A_{r_{i+1}})$ is the average travel time of all the buses on road segment $r_i$ in the specified time period. It will be estimated by mining the correlation
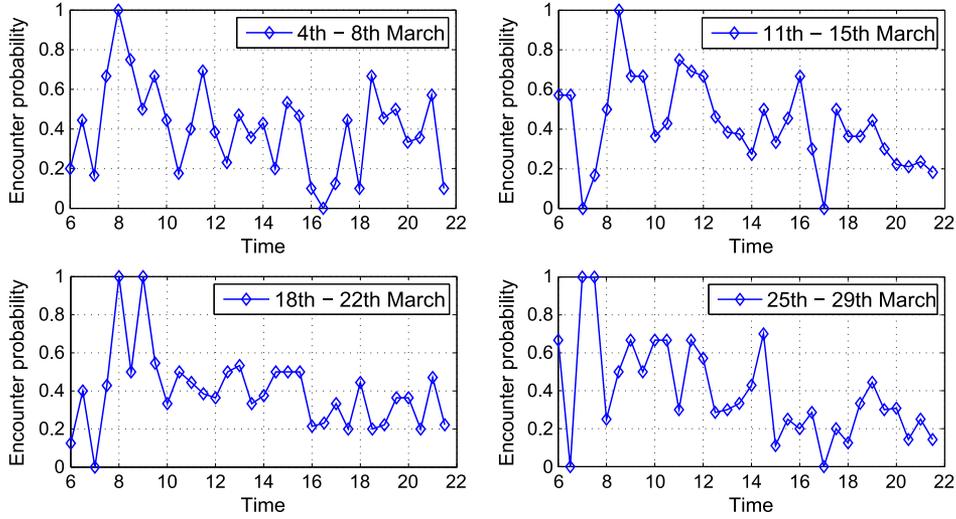
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

ZHANG *et al.*: ON GEOCASTING OVER URBAN BUS-BASED NETWORKS BY MINING TRAJECTORIES

7



Fig. 9. Encounter probabilities of two bus routes.



Fig. 10. Power spectrum of encounter probabilities.



Fig. 11. Illustration of probabilistic spatial temporal graph.



Bus route A: 1-4-9-11
Bus route B: 2-4-9-12
Bus route C: 1-2-5-10-11-12-8-3-1

Fig. 12. Bus routes.



Fig. 13. Spatial-temporal graph.

of historical bus trajectories (See Section IV-B). For the edge $e$ of TYPE II, which means that the message is transmitted between two buses, $D(A_{r_i}, B_{r_i})$ is set to 0. It is because the transmission time is short enough compared to the carrying time and therefore it could be ignored.

As an example to demonstrate the construction of a probabilistic spatial-temporal graph, we give a Manhattan road network consisting of 9 intersections and 12 road segments in Fig. 12. In detail, three bus routes (i.e., $A$, $B$ and $C$) cover the road network. The route of bus route $A$ is 1-4-9-11; the route of bus route $B$ is 2-4-9-12; the route of bus route $C$ is 1-2-5-10-11-12-8-3-1. The corresponding probabilistic spatial-temporal graph is shown in Fig. 13. Three large circles represent bus route $A$, $B$ and $C$, respectively. Note that the first element of the weight tuple of a TYPE-I edge is equal to 1, such as $P(A_4, A_9) = 1$ and the second element of the weight tuple of a TYPE-II edge equals zero, such as $D(A_4, B_4) = 0$.

## B. Predicting Travel Times of Buses

In order to predict the bus travel times, we treat the historical bus travel times as the time series data, and adopt the linear time series analysis method 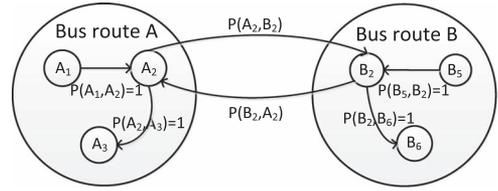[28]. The motivation behind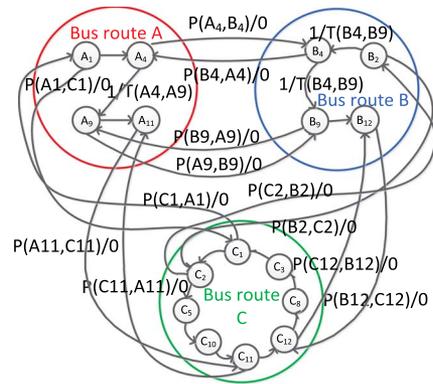 it is that the amount of historical bus data is not large and the time series techniques can be immediately applied to establish fairly accurate mathematical models in the case of limited sample

data. On the other hand, from the observation of real-world bus trajectories, we find that the bus travel times on the same road segment are different during the different time periods of a day. To adapt to the time-varying feature in bus travel times, we deal with the data on weekdays and at weekends, separately, and then partition a day into several time periods and construct a suitable ARMA model for the historical bus data of every time period, so as to forecast the bus travel time from the specified road segment to its adjacent road segment in the specified time period.

More specifically, given a data set containing bus travel times in the past $x$ weekdays, we first evenly divide the entire bus operating time (from 6:00 A.M. to 10:00 P.M.) into several fixed time periods ($TPs$) with a length of $D$ as an interval. Here $D$ is set to 20 minutes, therefore, the operating time of one day is divided into 48 $TPs$, and the data in the data set are viewed as 48 time series, which are denoted as $\{T_t^{TP}, t \in [1, x]\}$. For the sake of simplicity, if a specified time period $y \in TP$ is given, then the time series will be denoted as $\{T_t, t \in [1, x]\}$. Then, for every $T_t$, we conduct the stationary test [29], model identification and parameter estimation.

Considering that the distribution of travel times is unknown and the runs test [30] is a non-parametric statistical test that checks a randomness hypothesis without the need of any assumption of data distribution, we analyze $\{T_t, t \in [1, x]\}$ by the runs test and find that all the 48 time series are stationary.

As for the model identification, it is essential to analyze the autocorrelation of a time series. Therefore, we divide a time series $\{T_t, t \in [1, x]\}$ of travel times into two subseries with the length of $x - k$, namely, $\{T_1, \ldots, T_{x-k}\}$ and $\{T_{k+1}, \ldots, T_x\}$. Obviously, the two subseries have the same mean, denoted as $\overline{T}$. Then, the value of the autocorrelation function (ACF) can be calculated as follows:

$$\rho_k = \frac{\gamma_k}{\gamma_0} \quad (2)$$

where

$$\gamma_k = E\left[(T_t - \overline{T})(T_{t+k} - \overline{T})\right]$$
$$= \frac{1}{x-k}\sum_{t=1}^{x-k}(T_t - \overline{T})(T_{t+k} - \overline{T})$$
$$k = 0, 1, \ldots, x - 1.$$

ACF takes into account the influence of $T_{t+1}, T_{t+2}, \ldots,$ $T_{t+k-1}(t \in [1, x - k])$, and reflects the travel time correlation between two days with $k$ days apart, e.g., the correlation between $T_t$ and $T_{t+k}$. If $\rho_k$ equals zero when $k$ is greater than $q$, then this indicates that the time series of travel times has a $q$-step correlation. As thus, the $q$ is the order of the Moving Average (MA) model.

Next, we analyze the partial autocorrelation which indicates the time correlation between $T_t$ and $T_{t+k}$ after ignoring the influence of $T_{t+1}, T_{t+2}, \ldots, T_{t+k-1}(t \in [1, x - k])$. The partial autocorrelation function (PACF), denoted as $\varphi_k$, can be ob-

tained by a Yule-Walker equation whose matrix representation is defined as follows:

$$\begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \varphi_{k1} \\ \varphi_{k2} \\ \vdots \\ \varphi_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{bmatrix}. \quad (3)$$

Substituting $k$ with the positive number such as 1, 2, 3 $\ldots$ into the equation, the PACF ($\varphi_k = \varphi_{kk}$) can be solved as follows: $\varphi_{11} = \rho_1, \varphi_{22} = (\rho_2 - \rho_1^2)/(1 - \rho_1^2), \varphi_{33} = \begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix} / \begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_2 & 1 & \rho_1 \\ \rho_1 & \rho_1 & 1 \end{vmatrix}, \ldots$ If the value of the PACF $\varphi_k$ is equal to zero when $k$ is greater than $p$, that is, the PACF value of the time series is zero at lag $p$, then the $p$ is the order of the Auto Regressive (AR) model.

Due to the randomness of the time series, the estimation of $\rho_k$ or $\varphi_k$ may have possible deviation and fluctuate around zero when $k > p$, or $k > q$. However, if $\rho_k$ and $\varphi_k$ are not cut off but asymptotically converge to zero, then $T_t$ is also considered to follow the ARMA($p, q$) model, where $p$ and $q$ can be determined based on whether the values of the correlation functions exceed the corresponding confidence bounds.

Consequently, the built forecast function is as follows:

$$T_t = \varphi_1 T_{t-1} + \varphi_2 T_{t-2} + \cdots + \varphi_p T_{t-p} + \alpha_t$$
$$- \theta_1 \alpha_{t-1} - \cdots - \theta_q \alpha_{t-q} \quad (4)$$

where $t > x$, $p$ and $q$ are the orders of the Auto Regressive model and Moving Average model, which are obtained by the above model identification, $\alpha_t \backsim N(0, \delta^2)(i = 1, 2, \ldots)$ is a white noise series, and the coefficient $\varphi_i(i = 1, 2, \ldots, p)$ and $\theta_j(j = 1, 2, \ldots q)$ are estimated by the method of moments estimation.

Finally, we conduct the Ljung-Box Q test [31] to ensure that the sequence $\sigma_t$ of residuals between the measured values and the estimated values $\hat{T}_t$, i.e., $\sigma_t = T_t - \hat{T}_t(t = 1, 2, \ldots, x)$, is a white noise sequence.

When the test is passed, we get the ARMA models for $T_t(t > x)$ which can predict the message delivery delay from road segment $r_i$ to $r_{i+1}$ during the specified period of the $t$-th ($t > x$) day. The similar models can also be built for the time series of travel times at weekends to obtain the prediction values.

### C. Validating the Prediction Method

In this subsection, we use real GPS trajectories of buses to validate the effectiveness of prediction methods proposed in the previous subsection.

We extract the trajectories on the sampling road segment from Beichen Bridge to Anhui Bridge of four weeks from Mar. 4 to Mar. 29, 2013 (6:00 A.M.–10:00 P.M.), and divide these trajectories into two groups by weekdays and weekends. And then, in the light of the method in the last subsection, we analyze the time series data of weekdays of first three weeks to

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

ZHANG *et al.*: ON GEOCASTING OVER URBAN BUS-BASED NETWORKS BY MINING TRAJECTORIES
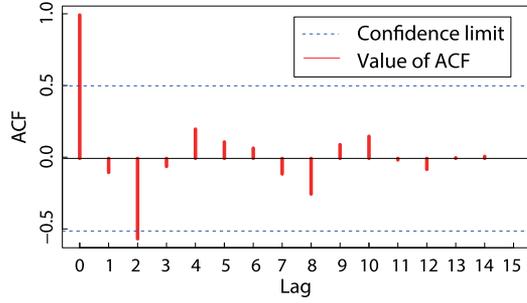
9
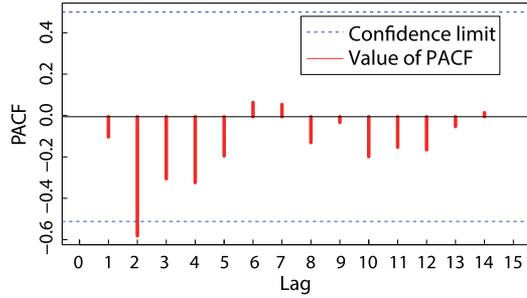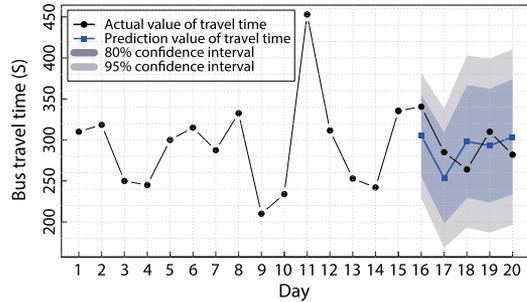


Fig. 14. ACF values.



Fig. 15. PACF values.



Fig. 16. Prediction results of travel times.

predict the travel times of the last weekday at the specified time period (i.e., 8:00 A.M. to 8:20 A.M.).

First, we examine the values of ACF and PACF of bus travel times. As shown in Fig. 14, the ACF at lag 2 is beyond the boundaries of the confidence and then gradually dies out to zero. The PACF shown in Fig. 15 has a large and significant spike at lag 2 beyond the boundaries of the confidence. It indicates that an ARMA(2,2) model is suitable for travel time prediction.

Then, the parameters of the forecast function are determined by moments estimation, and the travel time values (at the specified time period in the fourth week) are predicted 5-step ahead. As shown in Fig. 16, the residuals are few. Moreover, the residuals are a white noise series which is validated by the Ljung-Box Q test.

In addition, to understand the travel times at different time periods of a day, we establish a forecast equation for each time period (48 in total) and execute the one-step-ahead prediction.

So the travel times of next Monday and next Saturday are obtained in Fig. 17(a) and (b), respectively.

Three error measurements (i.e., Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE)) [32] are employed to evaluate the prediction accuracy. Table II depicts the prediction accuracies of travel times at different time periods of a week, wherein the MAPEs of weekdays and weekends are 10.42% and 8.02%, respectively.

### D. Estimating Encounter Probabilities

Through the statistics of encounter probabilities for bus routes in Section III-C, we can see that the encounters of bus routes on the same road segments show similar trends in four successive weeks, so it is suitable to use the polynomial fitting to reflect the trend of average encounter probabilities. According to the definition of encounter probability, we calculate the historical average encounter probabilities $P_{t_1}, P_{t_2}, \ldots, P_{t_n}$ at different time periods (i.e., $t_1, t_2, \cdots, t_n$). In this way, the encounter probability $P$ at time $t$ can be obtained by the polynomial fitting as follows:

$$P(t) = \mu_0 + \mu_1 t + \mu_2 t^2 + \ldots + \mu_m t^m \quad (5)$$

where $m$ represents the order of the polynomial, $\mu_i$ $(i = 0, \ldots, m)$ is determined by the least square method.

We give an instance that demonstrates the fitting result of encounter probabilities. While the data in Section III-C are taken as the input and $m$ is set to 4, the resultant polynomial fitting function is the equation (6) and the corresponding polynomial curve is depicted in Fig. 18

$$p(t) = -0.00012t^4 + 0.00713t^3 - 0.15304t^2$$
$$+ 1.36139t - 3.68266. \quad (6)$$

### V. ROUTING MECHANISM

In this section, we propose a geocast routing mechanism. As a basic assumption in bus-based VANETs, all buses know their 1-hop neighbors by periodic heartbeat messages. In general, a heartbeat message contains the sender's bus information and its neighbor list. Now each bus is required to hold the probabilistic spatial-temporal graph. As thus, if a bus on a road segment (denoted as $V_{\text{source}}$) wants to deliver a message to a destination road segment through a multi-hop communication, then $V_{\text{source}}$ will execute the following steps.

*1) Identifying the Destination:* From the existing bus routes, the bus routes that pass through the destination road segment are identified and then extracted to form a node set, denoted as *goalSet*.

*2) Calculating the Top-$k$ Shortest Delay Paths:* For each node in *goalSet*, denoted as $V_{\text{dest}}$, the shortest path from $V_{\text{dest}}$ to $V_{\text{source}}$ (in terms of delay) is calculated by using Dijkstra's algorithm over the spatial-temporal graph, and then the shortest path from $V_{\text{source}}$ to $V_{\text{dest}}$ is stored in a top-$k$ path set. Next, $V_{\text{source}}$ node is expanded one step further, that is, let the nodes connected to $V_{\text{source}}$ be the starting points (denoted as
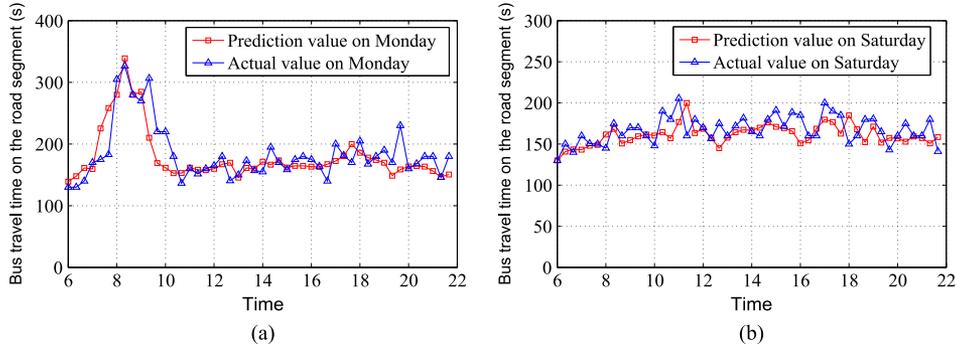
Fig. 17. Prediction values of the travel times. (a) Monday. (b) Saturday.

TABLE II
PREDICTION ACCURACIES OF BUS TRAVEL TIMES

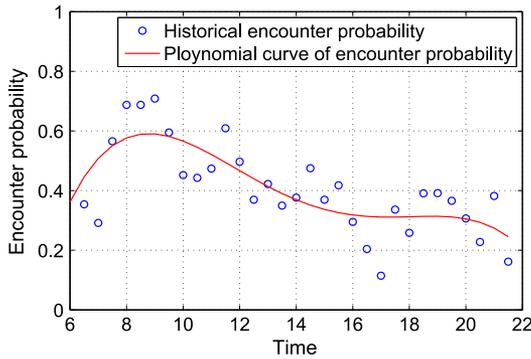|  | $MAE(s)$ | $MAPE(\%)$ | $RMSE(s)$ |
|---|---|---|---|
| $Weekdays$ | 19.81 | 10.42 | 28.5 |
| $Weekends$ | 13.56 | 8.02 | 17.12 |



Fig. 18. Polynomial fitting of encounter probabilities.

$V_{\text{nbr}}$). The shortest paths from $V_{\text{nbr}}$ to $V_{\text{dest}}$ are searched. The resulting paths plus the edge from $V_{\text{source}}$ to $V_{\text{nbr}}$ are also put into the top-$k$ path set. The same expanding procedure for the $V_{\text{nbr}}$ is repeated until there are $k$ paths in the top-$k$ path set.

*3) Calculating the Reliability of Message Delivery for Top-k Paths:* The reliability of path is computed by summing the encounter probabilities of the constituent road segments in spatial-temporal graph. Then, the path with the highest reliability among top-$k$ paths is selected as the message delivery route.

*4) Delivering a Message Along the Selected Path From $V_{\text{source}}$:* As for a bus of route $A$, if the current routing task derived from the predetermined routing path is to deliver the message to a bus of route $B$ on road segment $r$, then it will directly send the message to a bus of route $B$ while it arrives at road segment $r$ and finds that the bus of route $B$ is in its neighbor list.

*5) Optimizing to Enable Two-Hop Aware Routing:* For various reasons, a bus of route $A$ may not encounter a bus of route $B$. Fig. 19 shows such a scenario. At this point, a two-hop aware strategy is carried out. Taking the scenario in Fig. 19 as an example, if the bus of route $A$ searches its neighbor list and finds that one of its neighbors (e.g., a bus of route $C$ in Fig. 19) has a neighbor of route $B$, then the message is transmitted through a consecutive two-hop communication, i.e., first from
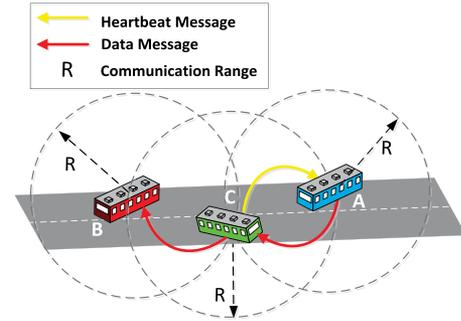


Fig. 19. Example of the two-hop aware strategy.

the source bus of $A$ to a bus of $C$ and then from the bus of $C$ to the destination bus of $B$.

*6) Handling the Delivering While Failing to Encounter the Bus:* If the bus that is carrying the message fails to encounter a bus of the predetermined route on the specified road segment, then after a while the bus will detect that the message is not delivered along the predetermined path. Next, the bus regards itself as the source node and re-calculates the route to the destination, so that the message can follow the new path to the destination.

The two-hop aware strategy only depends on the heartbeat messages, therefore it does not increase any other extra communication overheads. Meanwhile, the buses of a route only need to search their received heartbeat messages locally, so the strategy can achieve two-hop neighbor perception at a low computational cost. Moreover, the strategy can be used for buses on the same bus routes. When the buses from the same route perceive each other within two hops, the message sender can leverage the specific neighbor to deliver the message to another bus of the same route, thus reducing the time of message carrying, so that the message can be delivered faster. In short, the two-hop aware strategy enhances the feasibility of message routing in practical applications.

Fig. 20 shows a whole picture of Vela.

## VI. EVALUATION

### A. Experimental Setup

In order to evaluate the performance and scalability of Vela, we conduct extensive experiments to compare our routing mechanism with the several alternatives, i.e., MaxProp [17], BLER [19], R2R [20] and Vela-W.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

ZHANG *et al.*: ON GEOCASTING OVER URBAN BUS-BASED NETWORKS BY MINING TRAJECTORIES
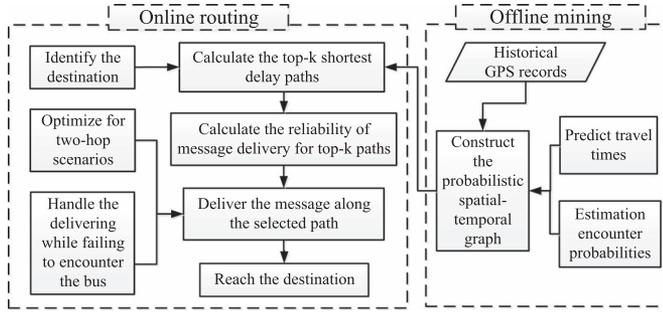
11



Fig. 20.   Framework of vela.



Fig. 21.   Real map of Haidian district, Beijing.

MaxProp is a bus level approach. It utilizes encounter probabilities of individual buses to estimate the cost of a virtual end-to-end path to the destination and uses the cost as the metric for routing decisions.

BLER is a bus line level approach. It builds a bus route graph where a vertex denotes the bus route, an edge linking two vertexes indicates at least one contact between two bus routes, and the weight of an edge is the sum of contact lengths of road segments. The routing path to the destination bus line is selected by the criteria of maximizing contact lengths, and then a zigzag process is triggered to route the message to the destination road segment, that is, a bus transmits the message only to another bus running on the same line.

R2R constructs a similar graph as BLER except that the weight of an edge is revised to the frequency of encounter between two bus routes.

Vela-W in fact is Vela without the two-hop aware strategy. Comparing Vela against Vela-W can give us the evidence of the effectiveness of the two-hop aware strategy in routing.

We adopt the delivery ratio and delivery delay as metrics to evaluate the performance and scalability of Vela and the other mechanisms. We use an open-source simulator NS-3 [33] for simulation. According to the specifications of IEEE 802.11p [34], the communication range is set to 250 meters, transmission speed is set to 6 Mbps, message size is set to 128 kilobytes and heartbeat message interval is set to 1 second. Based on the real map of Haidian district, Beijing (see Fig. 21), two groups of experiments on the real-world bus trajectories and the synthetic trajectories are conducted, respectively.

The first group of experiments is to observe the performance of message delivery. It is based on a 4000 m × 5000 m area of Zhongguancun district, Beijing, where the street layout is derived from OpenStreetMap [35]. In this area, there are 35 buses belonging to bus route No. 944, bus route No. 939, or bus route No. 881. We select the bus trajectories which occur between 10:30 and 12:00 on March 20, 2013 as the input of experiments. Before any experiment, real trajectories (reported once every 20 seconds) are converted into the tcl file format (where a GPS location per second is listed) using the interpolation method. During the experiment, the buses enter the selected area according to the real-time points of the bus trajectories and move along the real trajectories. All of the buses generate messages every one second and they randomly select the road segments on bus routes as destinations. 3500 messages
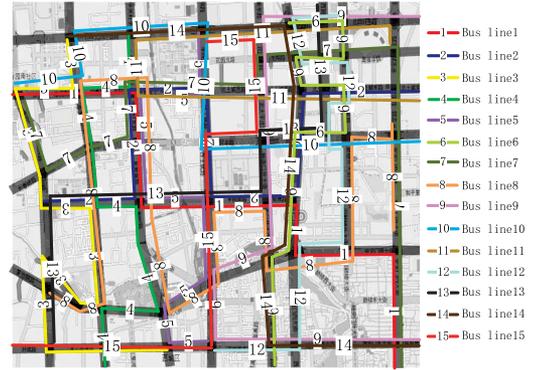


Fig. 22.   15 bus routes on the real map.

are sent in total. The experiment lasts 5400 seconds. During the experiment, all the messages arriving and their latencies are recorded.

The second group of experiments is to evaluate the scalability of the routing mechanisms by observing how the delivery ratio and delivery delay change with the increase of the number of bus routes. The vehicle mobility simulation tool SUMO [36] is used to generate the trajectories of 15 bus routes shown in Fig. 22. Specifically, the Intelligent Driver Model (IDM) is chosen for buses as their mobility model to reflect the features of bus movements in urban scenarios. Bus stops are set on road segments and bus dwell times at the bus stops are set randomly which range from 30 seconds to 120 seconds. The departing time intervals of buses at starting bus stops are set to 200 seconds. The messages are generated by the same way as the first group of experiments.

### B. Experimental Results and Analyses

As the preparation of the experiments, all the routing mechanisms need the existing bus trajectories to build their corresponding graphs.

For the first group of experiments, we use 20-day trajectories of 35 buses from 3 bus routes to build the graphs requested by the different mechanisms.

When we use the above trajectories to calculate the probability of encounter between individual bus pairs for MaxProp, we find that only a pair of buses (i.e., bus ID G62226 and
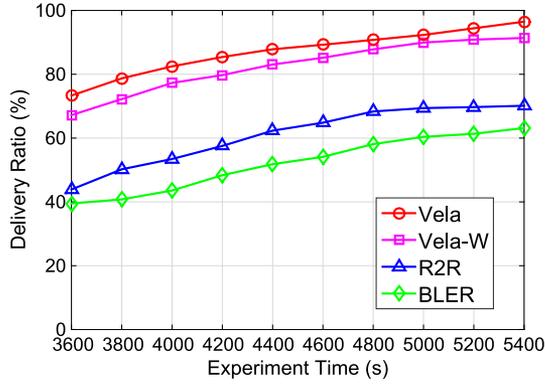
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

12                                                                IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS



Fig. 23.  Message delivery ratio.



Fig. 25.  Message delivery ratio.



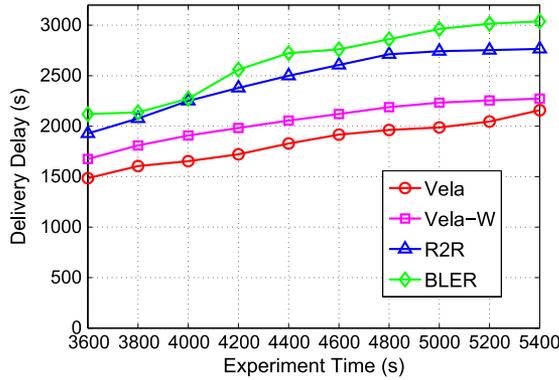Fig. 24.  Message delivery delay.



Fig. 26.  Message delivery delay.

bus ID G64904) has encountered twice and no other bus pair encounters each other during the 20 days. In other words, the contact graph needed by MaxProp cannot be built. Therefore, messages cannot be delivered to the destinations by MaxProp. The above results of applying the MaxProp to a real-world bus scenario demonstrate that using the encounter of individual buses to estimate the delivery likelihood for selecting the routing path is not feasible in practical and further the MaxProp is not appropriate to routing in the urban bus-based network. Hereinafter, we no longer mention the MaxProp due to its infeasibility.

In the first group of experiments, what we observe is the changes of delivery ratios and latencies in different mechanisms with the elapse of time (from 3600 seconds to 5400 seconds). In Fig. 23, all the mechanisms have the increasing delivery ratios as time goes on. It can be seen that the delivery ratio of Vela outperforms the other mechanisms. For instance, the delivery ratios of Vela and Vela-W reach 95.4% and 92.7% at 5400 seconds while the ratios for BLER and R2R are 73.5% and 63.2%, respectively. It exhibits that the methods which explore the spatial-temporal characteristics on road segments have more accuracy than the bus line based methods. Fig. 24 shows the message delivery delay of all the four mechanisms. The delay of Vela is always lower than that of the bus line level approaches, which indicates that Vela achieves its intended goal of latency reduction. In contrast, R2R and BLER have relatively low delivery ratios and high delivery delay. The reasons may come from the following three aspects. First, the contact length
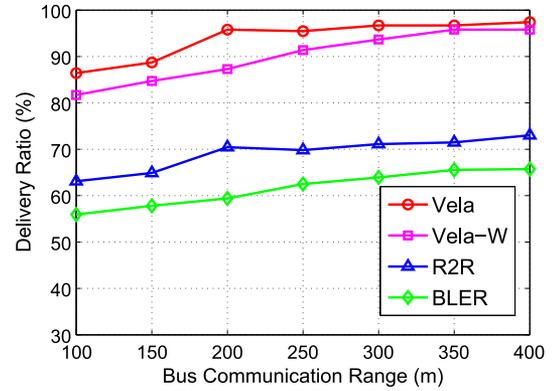
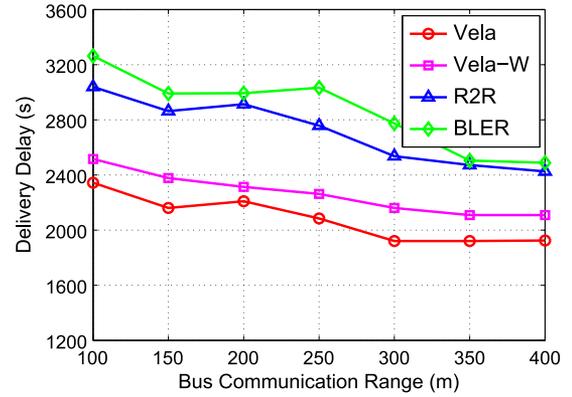and frequency are the metrics to decide the routing path in BLER and R2R, respectively. So they do not always bring the decrease of the delivery time since the message would be transmitted along some long-distance route. Second, a message may need to be carried a long time before it is forwarded to the buses that are required to satisfy some conditions (in terms of contact length or frequency), that is, many delivery opportunities may be missed since the bus encountered cannot meet the condition of data forwarding. Third, in comparison with Vela which delivers a message to the destination road segment directly, BLER and R2R need a zigzag process after the message reaches the destination bus line. This will no doubt increase the message delivery delay.

We also see that the average delivery ratio of Vela is 4.2% higher than that of Vela-W and the average delivery delay of Vela decreases by 221 seconds in comparison with Vela-W. It is because the two-hop aware strategy in Vela provides the chance to deliver the messages to the buses reachable by two hops.

Next, we conduct experiments with different communication ranges of buses, in order to investigate how the performance is changed under varying bus communication ranges. Fig. 25 shows that the delivery ratios of all the mechanisms are improved as the transmission range increases and with a relatively quick rise around the range of 200 m. Here, the delivery ratio of Vela is 95.8% with the range of 200 m, which is just the ratio of Vela-W with the range of 350 m. This indicates that two-hop aware strategy is effective to enhance the opportunity of encounter and expand the coverage of awareness. In
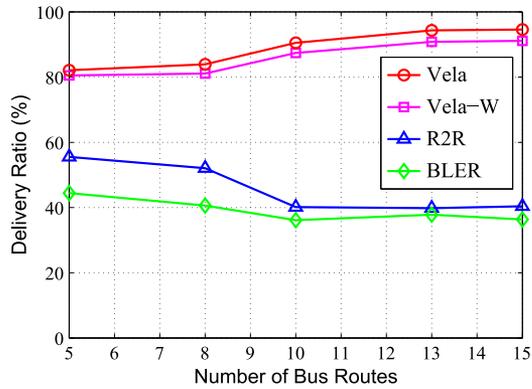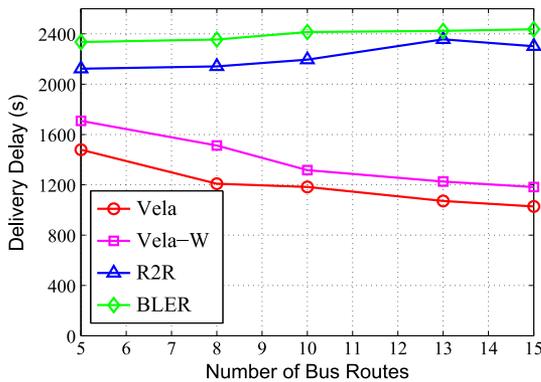
Fig. 27. Message delivery ratio.



Fig. 28. Message delivery delay.

addition, as shown in Fig. 25, the delivery ratios tend to be stabilized while the range is greater than 300 m. Fig. 26 shows the trends of message latencies as the communication range increases. While delivering a message, a large coverage range implies that the message can be transmitted to the next hop bus early, reducing the time to carry the message. The data in Fig. 26 prove this statement. Although the delay has a slight increase in the communication range of 200 m, the message delay, upon the whole, decreases and tends to be stable while the communication range is greater than 300 m.

For the second group of experiments, we use synthetic trajectories to build the graphs requested by Vela/Vela-W, BLER and R2R. The second group of experiments examines the effect of the varying number of bus routes. Fig. 27 shows that when the number of bus routes is increased from 5 to 15, the delivery ratio of Vela has the similar growth from 82.1% to 94.7%, which reflects the delivery ratio is well improved with the increase of bus lines. Fig. 28 depicts the delivery delay under the different number of bus routes. Apparently, Vela has the decreased delay with the increase of bus routes. These results demonstrate the scalability of Vela. On the other hand, as for R2R and BLER, the impact of the growth of bus routes is not obvious when the number of bus routes is low (from 5 to 8). But with the growth of bus routes (from 8 to 15), their performance gradually degrades. This is because R2R or BLER prefers to select a path with a long contact length or a high contact frequency while facing with multiple candidate paths. As a result, the chosen path would be a time-consuming one, the arrived messages

may have long latencies, and some messages may not reach at the destination within the experiment time. In brief, these experimental results prove that R2R and BLER lack scalability.

## VII. CONCLUSION

In this paper, we propose a geocast routing mechanism Vela for bus-based VANETs. Considering the unique characteristics of bus-based networks, Vela takes full advantage of the historical spatial-temporal relationships between buses for routing. It mines the historical trajectories of a bus network so as to capture the patterns of travel times on the road segments as well as the patterns of bus encounters on the same road segments. By combining the captured patterns with the bus network, Vela constructs a probabilistic spatial-temporal graph model for the bus-based VANET. As the result of the proper decision on the patterns to be mined and the appropriate mining methods, the graph model can accurately indicate the delivery delay and reliability over the bus-based VANET. Therefore, as for a message delivery request, the available routing path with best-possible QoS levels can be calculated from the graph model. On the other hand, Vela also utilizes the real-time spatial-temporal relationships between buses for routing. Vela provides a two-hop aware strategy to perceive the two-hop neighbors and increase the chances of forwarding the data along the pre-calculated path. We conduct the extensive experiments on the real and synthetic trajectories. The experimental results show that Vela has a high delivery ratio, a low latency and strong scalability in comparison with the other mechanisms.

## REFERENCES

[1] M.-C. Chuang and M. C. Chen, "DEEP: Density-aware emergency message extension protocol for VANETs," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 4983–4993, Oct. 2013.

[2] D. Zhang and C. K. Yeo, "Enabling efficient WiFi-based vehicular content distribution," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 3, pp. 479–492, Mar. 2013.

[3] M. Ayaida, M. Barhoumi, H. Fouchal, Y. Ghamri-Doudane, and L. Afilal, "Joint routing and location-based service in VANETs," *J. Parallel Distrib. Comput.*, vol. 74, no. 2, pp. 2077–2087, Feb. 2014.

[4] Y.-B. Ko and N. H. Vaidya, "GeoTORA: A protocol for geocasting in mobile ad hoc networks," in *Proc. IEEE ICNP*, 2000, pp. 240–250.

[5] Y. Zhu, Y. Wu, and B. Li, "Trajectory improves data delivery in urban vehicular networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 4, pp. 1089–1100, Apr. 2014.

[6] F. Zhang, H. Liu, Y.-W. Leung, X. Chu, and B. Jin, "Community-based bus system as routing backbone for vehicular ad hoc networks," in *Proc. IEEE ICDCS*, 2015, pp. 73–82.

[7] F. J. Ros, P. M. Ruiz, and I. Stojmenovic, "Acknowledgment-based broadcast protocol for reliable and efficient data dissemination in vehicular ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 11, no. 1, pp. 33–46, Jan. 2012.

[8] M. Slavik and I. Mahgoub, "Spatial distribution and channel quality adaptive protocol for multihop wireless broadcast routing in VANET," *IEEE Trans. Mobile Comput.*, vol. 12, no. 4, pp. 722–734, Apr. 2013.

[9] J. Jeong, S. Guo, Y. Gu, T. He, and D. Du, "TBD: Trajectory-based data forwarding for light-traffic vehicular networks," in *Proc. IEEE ICDCS*, 2009, pp. 231–238.

[10] F. Xu *et al.*, "Utilizing shared vehicle trajectories for data forwarding in vehicular networks," in *Proc. IEEE INFOCOM*, 2011, pp. 441–445.

[11] J. Zhao and G. Cao, "VADD: Vehicle-assisted data delivery in vehicular ad hoc networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 3, pp. 1910–1922, May 2008.

[12] H. Zhu, S. Chang, M. Li, K. Naik, and S. Shen, "Exploiting temporal dependency for opportunistic forwarding in urban vehicular networks," in *Proc. IEEE INFOCOM*, 2011, pp. 2192–2200.

[13] H. Zhu *et al.*, "ZOOM: Scaling the mobility for fast opportunistic forwarding in vehicular networks," in *Proc. IEEE INFOCOM*, 2013, pp. 2832–2840.

[14] W.-H. Liao, Y.-C. Tseng, K.-L. Lo, and J.-P. Sheu, "GeoGRID: A geocasting protocol for mobile ad hoc networks based on grid," *J. Internet Technol.*, vol. 1, pp. 23–32, Dec. 2000.

[15] L. Zhang, B. Yu, and J. Pan, "GeoMob: A mobility-aware geocast scheme in metropolitans via taxicabs and buses," in *Proc. IEEE INFOCOM*, 2014, pp. 1779–1787.

[16] R. Jiang, Y. Zhu, T. He, Y. Liu, and L. M. Ni, "Exploiting trajectory-based coverage for geocast in vehicular networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 12, pp. 3177–3189, Dec. 2014.

[17] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine, "MaxProp: Routing for vehicle-based disruption-tolerant networks," in *Proc. IEEE INFOCOM*, 2006, pp. 1–11.

[18] C. Liu and J. Wu, "Practical routing in a cyclic MobiSpace," *IEEE/ACM Trans. Netw.*, vol. 19, no. 2, pp. 369–382, Apr. 2011.

[19] M. Sede *et al.*, "Routing in large-scale buses ad hoc networks," in *Proc. IEEE WCNC*, 2008, pp. 2711–2716.

[20] L. Li, Y. Liu, Z. Li, and L. Sun, "R2R: Data forwarding in large-scale bus-based delay tolerant sensor networks," in *Proc. IET WSN*, 2010, pp. 27–31.

[21] U. G. Acer, P. Giaccone, D. Hay, G. Neglia, and S. Tarapiah, "Timely data delivery in a realistic bus network," *IEEE Trans. Veh. Technol.*, vol. 61, no. 3, pp. 1251–1265, Mar. 2012.

[22] N. Liu, M. Liu, W. Lou, G. Chen, and J. Cao, "PVA in VANETs: Stopped cars are not silent," in *Proc. IEEE INFOCOM*, 2011, pp. 431–435.

[23] W. Alasmary and W. Zhuang, "Mobility impact in IEEE 802. 11p infrastructureless vehicular networks," *Ad Hoc Netw.*, vol. 10, no. 2, pp. 222–230, Mar. 2012.

[24] C. Lochert, M. Mauve, H. Füßler, and H. Hartenstein, "Geographic routing in city scenarios," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 12, no. 4, pp. 69–72, Jan. 2005.

[25] F. Zhang, B. Jin, W. Zhuo, Z. Wang, and L. Zhang, "A content-based publish/subscribe system for efficient event notification over vehicular ad hoc networks," in *Proc. UIC*, 2012, pp. 64–71.

[26] A. Polus, "A study of travel time and reliability on arterial routes," *Transportation*, vol. 8, no. 2, pp. 141–151, Jun. 1979.

[27] H. Rakha, I. EI-Shawarby, M. Arafeh, and F. Dion, "Estimating path travel-time reliability," in *Proc. IEEE ITSC*, 2006, pp. 236–241.

[28] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley, 2008.

[29] W. Suwardo, M. Napiah, and I. Kamaruddin, "ARIMA models for bus travel time prediction," *J. Inst. Eng.*, vol. 71, no. 2, pp. 49–58, Jun. 2010.

[30] J. V. Bradley, *Distribution-Free Statistical Tests*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1968.

[31] G. M. Ljung and G. E. P. Box, "On a measure of lack of fit in time series models," *Biometrika*, vol. 65, no. 2, pp. 297–303, Aug. 1978.

[32] M. V. Shcherbakov *et al.*, "A survey of forecast error measures," *World Appl. Sci. J.*, vol. 24, pp. 171–176, 2013.

[33] NS-3 Tutorial, 2013. [Online]. Available: http://www.nsnam.org/ns-3-19/documentation/

[34] *IEEE 1609—Family of Standards for Wireless Access in Vehicular Environments (WAVE)*, STD98169, 2013. [Online]. Available: http://www.standards.its.dot.gov/factsheets/factsheet/80

[35] OpenStreetMap, 2014. [Online]. Available: http://www.openstreetmap.org/

[36] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of SUMO-simulation of urban mobility," *Int. J. Adv. Syst. Meas.*, vol. 5, no. 3/4, pp. 128–138, 2012.

**Beihong Jin** (M'12) received the B.S. degree from Tsinghua University, Beijing, China, in 1989 and the M.S. and Ph.D. degrees from the Chinese Academy of Sciences, Beijing, in 1992 and 1999, respectively, all in computer science. Since 1992, she has been with the Institute of Software, Chinese Academy of Sciences, where she is currently a Full Professor. She has obtained research support from the National Natural Science Foundation of China, the 863 Plan of China MOST, etc. She has published more than 90 research papers in international journals and conference proceedings and holds two China patents. Her research interests include mobile and pervasive computing, middleware, and distributed systems. She is a Senior Member of China Computer Federation and a member of the IEEE and the ACM.

**Zhaoyang Wang** is currently a Ph.D. candidate with the Institute of Software, Chinese Academy of Sciences, Beijing, China. His research interests include mobile and pervasive computing, middleware, and distributed systems.

**Hai Liu** received the B.Sc. and M.Sc. degrees in applied mathematics from South China University of Technology, Guangzhou, China, in 1999 and 2002, respectively, and the Ph.D. degree in computer science from City University of Hong Kong, Kowloon, Hong Kong, in 2006. He is currently a Research Assistant Professor with the Department of Computer Science, Hong Kong Baptist University, Kowloon. His research interests include wireless networking, cloud computing, and algorithm design and analysis.

**Jiafeng Hu** received the B.S. degree from Jilin University, Changchun, China, in 2011 and the M.S. degree from the Chinese Academy of Sciences, Beijing, China, in 2014. He is currently a Ph.D. candidate with the Department of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong. His research interests include mobile and pervasive computing, middleware, and distributed systems.
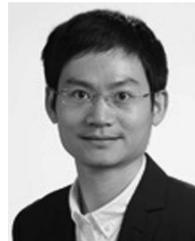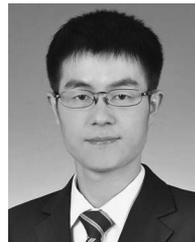
**Fusang Zhang** received the B.S. degree in 2010 from Xidian University, Xi'an, China, and the M.S. degree in 2013 from the Chinese Academy of Sciences, Beijing, China, where he is currently a Ph.D. candidate with the Institute of Software. His research interests include vehicular ad hoc networks, mobile and pervasive computing, middleware, and distributed systems.

**Lifeng Zhang** received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2014. He is currently an Assistant Professor with Ludong University, Yantai, China. His research interests include mobile and pervasive computing, vehicular ad hoc networks, and distributed systems.