Delay Minimization for Data Dissemination in Large-Scale VANETs with Buses and Taxis

Jianping He, *Member, IEEE*, Lin Cai, *Senior Member, IEEE*, Peng Cheng, *Member, IEEE*, and Jianping Pan, *Senior Member, IEEE*

Abstract—Minimizing the end-to-end delay for data dissemination in a large-scale VANET with both buses of fixed schedules and taxis of random schedules is a challenging issue, due to the scalability, high-mobility, and network heterogeneity concerns. Particularly, the mix of random taxis and fixed-scheduled buses makes the delay components along a path dependent and hard to estimate. In this paper, to address the scalability and high-mobility issues, we introduce a store-and-forward framework for VANETs with extra storage using "drop boxes", which function similar to network routers. Next, we propose an optimal link strategy which is independent of the message arrival time and can be executed in a distributed manner. Then, we derive the expected path delay, considering the dependence of the delay components along the path, and propose the optimal routing strategy to minimize the expected path delay. Trace-driven simulations have been used to validate the rigorous analysis, and demonstrate the superior performance of the proposed strategies, which result in a substantial delay reduction and a much higher delivery ratio when compared with the state-of-the-art solutions without drop boxes. The strategies can further improve the delay performance when compared with the over-simplified routing solutions which ignore the dependence of the delay components.

Index Terms—Mobile networks, VANETs, information forwarding, delay minimization

1 Introduction

Future vehicles are anticipated to be equipped with onboard units (OBU) to communicate through wireless links, and vehicle-to-vehicle communications can be of low cost thanks to the license-free spectrum usage and the minimum infrastructure requirement. Data dissemination is a promising application in Vehicular Ad hoc NETworks (VANETs), where messages are carried and forwarded by vehicles cooperatively toward their destinations at given locations. For example, hotels, restaurants, and tourist attractions can send their advertisement, promotions, and service availability to all the vehicles near the airport, train station, ferry, and major bus terminals, so passengers on board the vehicles can use the information for their immediate trip planning, etc. Different from the traditional Internet applications, the message destination in these VANET message dissemination applications is not a particular node/vehicle with a global ID or IP address, but any vehicle(s) in a particular region. How to find the right vehicles to carry the messages to a location can be viewed as a link selection problem. Extensive researches have been done for link selection in small-scale VANETs where the destination and the source are close to each other [1].

In this paper, we consider a more challenging issue, link selection in large-scale VANETs (e.g., covering a whole city) for data dissemination. A message, which can be data or multimedia files, can be piggybacked by a vehicle toward the message destination location. When the traveling path of the vehicle deviates from the direction to the destination of the message, the message should be forwarded to a more appropriate vehicle through wireless communications during the contact time of the two vehicles (i.e., when the two vehicles are within each other's transmission range). Considering the possible long distance between the source and the destination, multiple vehicles may be needed to carry the message which form a multi-hop path, and a large number of vehicles may encounter the message carrier in such a VANET Geocast scenario and thus participate in the routing process. The key issue is how to find a suitable path such that the message can be carried and forwarded to the destination with the minimal expected delay.

Although the traditional delay-tolerant networking (DTN) techniques, e.g., ring-based index, mobility patten evaluation, and social-based approach, may be used for the routing problem here, the performance is not satisfactory, given the new challenges in large-scale VANETs related to the scalability, high mobility, and network heterogeneity. First, given the large number of vehicles possibly involved, the solutions relying on flooding cannot scale well. Also, the assumptions often used that vehicles maintain the contact history to estimate the pair-wise contact probability or contact pattern may result in high communication and storage cost for the vehicles when the size of the network grows. Second, as vehicles have high mobility, the contact time

- J. He is with The State Key Laboratory of Industrial Control Technology and Innovation Joint Research Center for Industrial Cyber Physical Systems, Zhejiang University, China, and the Department of Electrical and Computer Engineering, University of Victoria, BC, Canada. E-mail: jphe@iipc.zju.edu.cn, jphe@uvic.ca.
- P. Cheng is with the State Key Laboratory of Industrial Control Technology and Innovation Joint Research Center for Industrial Cyber Physical Systems, Zhejiang University, China. E-mail: pcheng@iipc.zju.edu.cn.
- L. Cai and J. Pan are with the Department of Electrical and Computer Engineering, University of Victoria, BC, Canada. E-mail: cai@ece.uvic.ca, pan@uvic.ca.

Manuscript received 29 June 2015; revised 1 Sept. 2015; accepted 5 Sept. 2015. Date of publication 22 Sept. 2015; date of current version 29 June 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TMC.2015.2480062

between two vehicles is short, which limits both the amount of information to be exchanged and the time needed to make a routing decision. Thus, the previous approaches that require extensive computation are not desirable.

Third, there exist different types of vehicles with different mobility patterns. One typical type of vehicles has a fixed schedule and route, and the other has random ones. In this work, buses and taxis represent these two types of vehicles, respectively. If using buses as the message carriers only, given their fixed schedules and routes, it is simple to estimate and bound the delay. However, the coverage of buses is limited, and so is the network connectivity. If using taxis only, the opposite is true. Although we can leverage the coverage and flexibility of taxis and the deterministic schedule and path of buses for improving the network connectivity and delay performance, the heterogeneity also introduces an unexpected complexity to analyze the delay for optimal routing.

To first address the scalability and mobility issues, we propose a store-and-forward framework for VANET message delivery. The whole area covered by the VANET (such as a city) is divided into regions, each of which contains a hot-spot location with a large volume of vehicle traffic, e.g., major bus exchanges, shopping centers, airports, etc. We assume that a message can be stored temporarily in a “drop box” in the hot-spot until being forwarded to the right vehicle to reach the neighbouring region (the next hop). The idea of using drop boxes has been proposed in the literature for reducing delay [2] and enhanceing throughput [3] in delay-tolerant networks (DTNs). Using drop boxes can increase the network connectivity and reduce the uncertainty, and thus is useful to address the scalability and mobility problems in VANETs. In our work, a drop box can be a storage device installed in an access point which is relatively static, or in a vehicle temporarily parked in the hot-spot, and the drop-box in a region can be a cluster of communication and storage devices if the traffic volume becomes large. In the latter case, the content in the storage device will be forwarded to other drop boxes when the vehicle needs to move. Given the page limit of the paper, we do not further elaborate the management of drop boxes and assume that the drop box is always available at each hot-spot with sufficient capacity. To demonstrate the feasibility of the framework, we used real-world bus and taxi GPS traces in Shanghai, China, to identify the hot regions and extract the bus and taxi traffic statistics between these regions. The traffic information is used to find the optimal path toward the destination. With the addition of in-network storage at the hot-spots, network connectivity and reliability for large-scale VANETs can be improved. Given the limited number of regions, the amount of vehicle traffic information (e.g., bus routing schedule and average taxi arrival rates) needed for routing is limited, and thus this solution is scalable.

The next issue is to decide which path has the minimal expected delay. If the network uses buses only, this problem is straight-forward, and can be solved using the existing shortest-path algorithms. If the network uses taxis only, the delay between any two regions (or per-hop delay) is random, considering the random time to wait for a suitable taxi, and the random traveling time of the taxi to the next-hop region. Furthermore, it is reasonable to assume that these delay components by taxis only along the path are independent, so the end-to-end delay analysis and routing can be greatly simplified. When both types of vehicles are involved, the problem becomes much more complicated. First, for each hop, a decision should be made on which type of vehicle is preferred for carrying the message, and this decision may be changed over time given the random process of taxi arrivals and the known time to catch the next bus. Since the possible strategies at each hop are unlimited, it is difficult to solve this problem directly using general centralized algorithms such as enumeration algorithm. Second, the end-to-end optimal path selection relies on the expected delay of each path, which is a hard problem since the delay components along a given path are dependent to each other. Thus, this routing problem does not satisfy Bellman’s Principle of Optimality [4], so we cannot use dynamic programming approaches, such as Dijkstra’s or Bellman-Ford algorithm, to solve the shortest-path problem.

Therefore, in this paper, we first formulate a delay minimization problem for data dissemination in VANETs. We then provide the optimal link strategy and the exact delay estimation, followed by an optimal path selection algorithm to solve the problem. We also conduct trace-based simulation to demonstrate the feasibility of our solution. The main contributions of this paper are three-fold. First, we propose the region-based store-and-forward message dissemination framework for large-scale VANETs, using both buses and taxis as possible message carriers and using drop boxes as routers. Simulation results with the traffic traces in Shanghai show that the framework can improve the connectivity, reliability, and scalability, and achieve a substantial performance gain compared with the previous solutions without drop boxes. Second, we derive the optimal link strategy in terms of which types of vehicles to piggyback and how long the message should wait for them. Third, to the best of our knowledge, this is the first work that obtains the closed-form expression for the expected path delay of VANET data dissemination with both buses and taxis, considering the dependence of the delay components between them along the path. We also develop the optimal routing algorithm based on the theoretical delay analysis. The rigor of our analysis has been validated by extensive simulations. Trace-driven simulations also show that the proposed solution can outperform the shortest-path solutions that ignore the dependence of the delay components along the path.

The rest of this paper is organized as follows. Section 2 discusses the related work. In Section 3, we first introduce the system model and formulate the optimal routing problem. Section 4 presents the optimal per-hop strategy, and derives the expected delay along a given path and the optimal routing algorithm for minimizing the expected path delay. Simulation results are presented in Section 5, followed by the concluding remarks in Section 6.

2 Related Work

Data dissemination in VANETs is a challenging and important problem, and many efforts have been devoted to devising efficient routing solutions to improve the delivery ratio, minimize the latency and delivery delay, and reduce the overhead.

Most of the existing protocols achieve high-efficiency data dissemination by exploiting certain properties, e.g., considering the encounter probability and vehicular trajectories observed from real data traces. In [5], [6], the authors used the encounter probability for routing decision in intermittently connected networks and VANETs. MaxProp proposed in [6] uses the encounter information to model the cost of a virtual end-to-end path to the destination, and utilizes it as a metric for routing decisions. To achieve a higher performance, Zhu et al. [7] used high-order Markov chains to predict the vehicular trajectories, and derived the packet delivery probability with the predicted trajectories. Then they proposed an effective routing algorithm with a high delivery ratio and low cost. ZOOM proposed in [9] aims to achieve fast opportunistic forwarding in vehicular networks. It automatically chooses the most appropriate mobility information when deciding the next data-relays, which can reduce the end-to-end delay while reducing the network traffic. Based on some large sets of GPS traces of vehicles, [8] presents an opportunistic forwarding algorithm by exploiting the temporal dependence of inter-contact time, which can increase the delivery ratio and reduce the end-to-end delay. In [10], [11], [12], [13], [14], the authors concerned the performance evaluation of data dissemination protocols, e.g., delay, reliability and efficiency. Some properties, e.g., sociological orbit and dynamics of network connectivity, can be used to design the data dissemination protocol and improve the performance of the protocol.

Some existing works have already used the infrastructure or drop boxes or throwboxes to assist the data dissemination in DTNs and in VANETs, and they have shown that these road-side units can greatly increase the contact probabilities between vehicles and reduce the delivery delay [29], [30], [31], [32], [33]. For example, Banerjee et al. in [29] first examined the performance-cost tradeoffs for VANETs by considering assisted infrastructures. It demonstrated that using a small amount of infrastructures can significantly improve the performance of a routing protocol. Wu et al. in [32] pointed out that it is able to improve the data delivery ratio and reduce the delivery delay in VANETs by deploying infrastructures. Hence, these works motivated us to adopt the drop boxes to assist data dissemination in our routing protocol.

However, most of the routing strategies in previous works are heuristic and the mobility pattern of vehicles in their models is purely random. Although many types of vehicles have a random mobility, some others such as buses usually have a deterministic or predictable mobility pattern. When both the deterministic and random mobility patterns are considered for data dissemination in VANETs, it becomes a more complicated but practical problem, which is the main focus of this paper. Recently, Zhang et al. [15] used both buses and taxis as the message carriers in urban vehicular networks. They proposed a mobility-aware Geocast algorithm (GeoMob) for urban VANETs from the DTN perspective to better deal with the high mobility and transient connectivity issues. Compared with the existing works using both buses and taxis, the main novelties of this paper include the region-based store-and-forward network architecture design, the optimal link strategy, and the optimal routing algorithm to minimize the expected path delay.

3 System Model and Problem Formulation

3.1 Scenario and Assumptions

We consider the Geocast problem where a message is carried and forwarded by cooperative vehicles toward its destination with the assistance of drop boxes. Note that a vehicle will travel according to its travel plan determined by the driver and passenger(s) and the road and traffic condition, and the message is simply piggybacked on the vehicle. We assume that the travel plan is known to the OBU, e.g., the travel destination and route calculated by the GPS navigation unit can be sent to the OBU.[1] The destination is a known location, or within a certain known target region. If the latter, the message will be carried to the target region first and then flooded in the region to reach all vehicles in the destination. The flooding process is relatively simple, which has been heavily investigated [25], [26]. The main challenge we focus on is how to deliver the message to the target region by selecting suitable vehicles along a path.

To address the network scalability issue, we first divide the whole VANET area into regions, and select a hot-spot in each region with a large traffic volume, such as major bus exchanges, shopping centers, airports, central business districts, etc. A drop box is installed at each selected hot-spot. The drop box can be hosted in an idle vehicle's OBU and the content can be handed over to another idle vehicle if the current one needs to leave, or be hosted in a stationary road-side unit. These drop boxes may not have an Internet connection, or, even if a limited Internet access is available, we do not use it for data dissemination to save the cost. For example, the drop box can be a storage device in a bus which is parked in a hot-spot location, without Internet access. The drop boxes in the VANET function similar to the routers in the Internet but with larger storage. We assume that when a vehicle is within the drop box's coverage, the travel plan of the vehicle will be reported to the drop box. Also, it will transfer the data it carries to the drop box if the message needs to be carried by another vehicle. The drop box then stores the message until it finds a suitable vehicle and transfers the message to it. It is noted that since the transmission range of the drop-box is limited, the drop-box can only relay the message to those vehicles within its coverage for the next carrier. Hence, depending on the transmission range of the drop-box, only a portion of taxies may pass the coverage area of the drop-box. This is essentially a thinning process. We can handle the thinning process by multiplying the vehicle arrival rate by a ratio (the arrival rate in the area covered by the drop-box over the arrival rate in the whole region).

The key issues are how to select vehicles for each link and find the optimal path from the source to the destination with the minimal delay. These two issues are generally coupled. To make the problem more tractable, we narrow down to the case that only source routing is used, so the path information from the source to the destination (which contains a sequence of drop boxes along the path) is

[1] The vehicles without a travel plan or not willing to share the travel plan information should not be considered as a possible carrier, and how to provide incentive to vehicles to participate in the VANET will be an interesting research topic that is beyond the scope of this work.

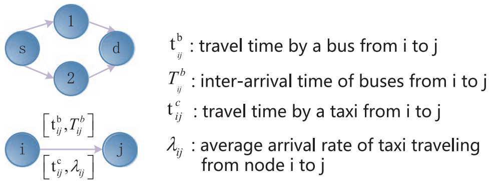

1
s
d
2
$[t_{ij}^{b}, T_{ij}^{b}]$
i
j
$[t_{ij}^{c}, \lambda_{ij}]$
t_{ij}^{b} : travel time by a bus from i to j
T_{ij}^{b} : inter-arrival time of buses from i to j
t_{ij}^{c} : travel time by a taxi from i to j
λ_{ij} : average arrival rate of taxi traveling from node i to j

Fig. 1. The graph of a sample vehicular network.

determined by the source node and the routing information is attached to the message.

3.2 Network Topology and Models

To solve the optimal routing problem, we first abstract the VANET to a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes, and $\mathcal{E}$ is the set of edges. The nodes denote the drop boxes where the messages can be stored and forwarded to other vehicles. A directed edge $< i, j > \in \mathcal{E}$ denotes a link from node i to node j (where i and j should be geographical neighbors to speed up the routing decision), which means that the message can be delivered by a bus or a taxi from node i to node j directly. Let nodes s and d be the source and the destination, respectively. Let $\mathcal{N}_i = \{j| < i, j > \in \mathcal{E}\}$ be the neighbor set of node i. A sample VANET is given in Fig. 1. There are two directed paths between s and d, and messages can be delivered following one of the directed paths in the graph, e.g., $s \rightarrow 1 \rightarrow d$.

In the VANET, there are two types of vehicles with different mobility patterns, represented by buses and taxis, respectively. Let t_{ij}^{b} be the travel time by a bus from node i to node j (if there are no buses from i to j, we set $t_{ij}^{b} = \infty$). [2] Let T_{ij}^{b} be the inter-arrival time of buses from i to j. For simplicity, we assume that t_{ij}^{b} and T_{ij}^{b} are two constants, and satisfy $T_{ij}^{b} = T$ and $\frac{t_{ij}^{b}}{T} \in \mathbf{N}^{+}$ for $i, j \in \mathcal{V}$. This assumption can be relaxed, and even if there is a fluctuation of T (e.g., T equals a constant plus a random number), the performance of our solution will not be affected substantially, as shown in Section 5. Let t_{ij}^{c} be the travel time by a taxi from i to j, which can be a random variable. We also assume that $\frac{\mathbf{E}\{t_{ij}^{c}\}}{T} \in \mathbf{N}^{+}$ for $i, j \in \mathcal{V}$.

The arrival process of taxis from node i to node j is assumed to follow a Poisson Process with the average arrival rate of $\lambda_{ij} \geq 0$ for $< i, j > \in \mathcal{E}$. λ_{ij} can be estimated from the history [22], [23], [24], and $\lambda_{ij} = 0$ if and only if (iff) there is no taxi from i to j. The Poisson arrival assumption has been widely used in the literature, e.g., [20] in DTNs and [21] in the multi-cast capacity analysis, and we have validated it by comparing with the real traces, where the trace information is given in Section 5. Using the real traces, first, we investigate the inter-arrival time distribution of taxis of each link and compare it with the exponential distribution. From the traces, we obtained the number of taxis traveling from one cluster to another in one month, and then calculated the average taxi arrival rate in this directed link. We also obtained the CDF of the inter-arrival time from the traces. The results show that the Poisson arrival assumption is acceptable, as the CDF of the inter-arrival time from the traces matches well with that of the exponential R.V. with the same average arrival rate.

Fig. 2a shows the CDF of the taxi inter-arrival time for the link from cluster 2 to cluster 32, and that for the link from cluster 18 to cluster 27. Meanwhile, we choose two representative time intervals, i.e., peak time (8:00am-6:00pm) and off-peak time (12:00am-6:00am) and study the taxi inter-arrival time in these two time intervals, and the corresponding CDF of the taxi inter-arrival time from cluster 32 to cluster 22 (random selected the clusters) is shown in Fig. 2b. Clearly, for both the peak and off-peak time, the Poisson arrival assumption between two clusters is acceptable. Moreover, we divide the one month data into two parts and study their CDF of the inter-arrival time, and the corresponding results are shown in Fig. 2c, where the CDF lines extracted from the first half month data is marked as History, and the CDF from the second half month data is marked as Test. As shown in Fig. 2c, we find that the CDF lines are close to each other, which means that the Poisson model based on historical data is reasonable to predict the future inter-contact time distribution. Results for other pairs show the similar tendency and are omitted due to the space limit.

Table 1 summarizes all the important notations in this paper for easy reference.

3.3 Problem Formulation

Assume that there exist N paths from s to d, denoted by $p_{sd}^{k}, k = 1, 2, \ldots, N$. We define $\Omega = \{p_{sd}^{k} | k = 1, 2, \ldots, N\}$, as the set of all paths from s to d. The optimal routing problem needs to consider two issues. First, when a message arrives at a node, a decision should be made on whether it should wait for and be carried by a bus or a taxi to reach the next hop, which is called the link strategy. Second, a decision should be made at s on which path in Ω should be selected.

Let s_i^t be a link strategy of node i. Under s_i^t, the message at node i will wait for a taxi in time interval $[t_i^a, t_i^a + t)$, and if it cannot encounter a suitable taxi in this time interval, it will be carried by a bus at time instant $t_i^a + t$. Obviously, $t_i^a + t$ is the time of a bus leaving from i toward j for $j \in \mathcal{N}_i$. Then, we define the link strategy set for node i as

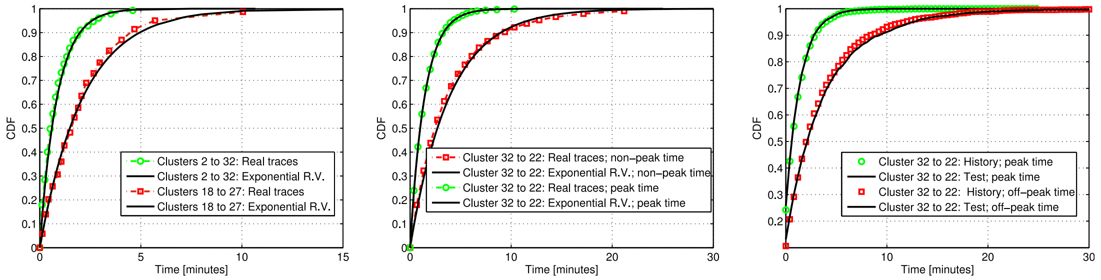
$$\Theta_i = \{s_i^t | t_i^w \leq t, t \geq 0\}, \quad i \in \mathcal{V}, \tag{1}$$

where t_i^w is the waiting time at node i, and t is the maximum waiting time under strategy s_i^t. Let $T_{sd}(k)$ denote the end-to-end delay of the message following path $p_{sd}^{k} \in \Omega$ from s to d. The delay, $T_{sd}(k)$, depends on the traffic flow of buses and taxis for the links along path $p_{sd}(k)$, and the strategy at each node. Since the traffic flow of taxis between two neighbors is random, $T_{sd}(k)$ is also random. Note that the delay is one of the most important metrics in the message dissemination problem [16], [17], [18], [19]. Hence, we formulate the optimal routing problem as to select the path with the minimal expected delay as follows:

$$\min_{p_{sd}^{k} \in \Omega} \quad \min_{s_i^t \in \Theta_i, i \in p_{sd}^{k}} \mathbf{E}\{T_{sd}(k)\}. \tag{2}$$

To solve this problem, it needs to find the best link strategy for each link along a given path, and also find the best path in

[2] Hereafter, $(\cdot)^b$ and $(\cdot)^c$ denote the parameters of buses and taxis, respectively.



(a) for the entire day

(b) only peak and off-peak time

(c) history and test data

Fig. 2. The CDF of taxi inter-arrival time.

terms of the minimal expected delay. The problem is non-trivial due to the complicated interaction of deterministic and random mobility patterns of buses and taxis, which results in the dependency of link delays along a path. We use the following example to illustrate the difficulty of this problem.

Example 3.1. Consider a VANET shown in Fig. 1. For each link, the travel time by buses is 30 minutes and that by taxis on average is 20 minutes. The average taxi arrival rates are $\lambda_{s1} = 0.2, \lambda_{1d} = 0.05, \lambda_{s2} = 0.05,$ and $\lambda_{2d} = 0.2$ per minute. There are other incoming and outgoing links to and from nodes 1 and 2, so the flow conservation law does not reflect for the partial link graph. Intuitively, these two paths may have the same expected delay as both of them have two links with the identical parameters in a reversed order. However, the message arrival time at 1 (or 2) will affect how long it needs to wait to meet the next bus, and the probability of encountering a taxi before any bus arrives. The message arrival time at 1 (or 2) depends on the decision made at s. As shown in later sections, if taking the best link strategy, one of the path actually has a lower expected delay than the other one.

In the following section, we will discuss how to estimate the delay and solve the optimization problem (2) in detail.

4 Optimal Routing to Minimize Expected Delay

The optimal solution needs to address two issues. First, given a path, what is the optimal link strategy in terms of how long the message should wait for a taxi (or a bus) to be carried to the next hop. Second, how to find the optimal path from the source to the destination with the minimal expected delay. In this section, we first obtain the optimal link strategy. Then we analyze the expected delay of each path and design an algorithm to select the optimal path.

TABLE 1
Important Notations

Symbol	Definition
T	inter-arrival time of buses
t_{ij}^b	travel time by a bus from node i to node j
t_{ij}^c	travel time by a taxi from node i to node j
t_i^w	waiting time at node i
t_i^a	message arrival time at node i
$\tilde{t}_i^a$	waiting time at node i until the first bus departure: $\tilde{t}_i^a = \lceil \frac{t_i^a}{T} \rceil T - t_i^a$
$\hat{t}_i^a$	time from the last bus departure to the message arrival at node i: $\hat{t}_i^a = T - \tilde{t}_i^a$
s^∞	strategy of always waiting for the first taxi
s^1	strategy of waiting for the first taxi unless the bus arrives first

4.1 Optimal Link Strategy

The difficulty of our problem is that the delays by buses of adjacent links are dependent to each other. This requires us to carefully formulate the optimal link strategy problem such that the optimal solution of each link will lead to the optimal solution for the whole path.

Let t_{ij} be the delay on link $< i, j >$ from the time a message arrivals at node i to the time it arrivals at j. Let t_{ij}^w be the waiting time of the message at node i before it is carried toward node j, which equals t_{ij} minus the travel time of the carrier, a bus or a taxi. As the taxi travel time t_{ij}^c is independent of the routing strategy and other delay components, e.g., t_i^w, along the path, we abuse the notation slightly by letting $t_{ij}^c = \mathbf{E}\{t_{ij}^c\}$ in the following. Denote by t_i^a the message arrival time at i. The time interval from the message arrival till the next bus arrival at i is denoted by $\tilde{t}_i^a = \lceil \frac{t_i^a}{T} \rceil T - t_i^a$.

Hereafter, we analyze the optimal link strategy. We first discuss the optimal link strategy without considering the dependence of the links' delays. We then have the following theorem whose proof is given in Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TMC.2015.2480062.

Theorem 4.1. *If $t_{ij}^b - t_{ij}^c \leq \frac{1}{\lambda_{ij}}$, then the best strategy to minimize the expected delay from i to j is that the message will wait for a taxi only before the first bus coming (denoted by s^1), i.e., the waiting time t_{ij}^w satisfies $t_{ij}^w \leq \tilde{t}_i^a$, and it will be carried by the first bus if no taxi arrives before the bus; if $t_{ij}^b - t_{ij}^c > \frac{1}{\lambda_{ij}}$, then the best strategy to minimize the expected delay is to always wait for a taxi (denoted by s^∞), i.e., the waiting time t_{ij}^w satisfies $t_{ij}^w \leq +\infty$.*

From the above theorem, the optimal link strategy is easy to obtain by comparing the values of t_{ij}^b (the lowest link delivery delay by buses) and $t_{ij}^c + \frac{1}{\lambda_{ij}}$ (the average link delivery delay by taxis) when the delay dependence of links is ignored. This theorem can be used for obtaining an approximate and suboptimal solution of problem (2), which will be discussed in detail in the simulation part.

Then, we consider the optimal link strategy with the link delay dependence. Note that under s^∞, we have $t = \infty$. The expectation of t_{ij} satisfies

$$\mathbf{E}\{t_{ij}\} = \lim_{t\to\infty}\left[\frac{1}{\lambda_{ij}} + \left(t_{ij}^b - \frac{1}{\lambda_{ij}} - t_{ij}^c\right)e^{-\lambda_{ij}t} + t_{ij}^c\right] = \frac{1}{\lambda_{ij}} + t_{ij}^c.$$

It is obvious that, with s^∞, the delay of the link is independent of those in the previous hops, which is an important feature that we will use later. However, under s^1, $\mathbf{E}\{t_{ij}\}$ depends on the arrival time t_i^a as the waiting time should satisfy $t_{ij}^w \le \lceil\frac{t_i^a}{T}\rceil T - t_i^a$. Given t_i^a, we have

$$\mathbf{E}\{t_{ij}\} = \frac{1}{\lambda_{ij}} + \left(t_{ij}^b - \frac{1}{\lambda_{ij}} - t_{ij}^c\right)e^{-\lambda_{ij}\tilde{t}_i^a} + t_{ij}^c,$$

where $\tilde{t}_i^a = \lceil\frac{t_i^a}{T}\rceil T - t_i^a$. Clearly, we have $\mathbf{E}\{t_{ij}\} = t_{ij}^b$ when $\tilde{t}_i^a = 0$, which is corresponding to the case that the message arrival time equals the departure time of the bus from i to j, so the message takes the bus directly under s^1. Hence, under s^1, $\mathbf{E}\{t_{ij}\} \ge t_{ij}^b$, and the equality holds iff $\tilde{t}_i^a = 0$.

Since under s^1, $\mathbf{E}\{t_{ij}\}$ depends on t_i^a which depends on the decisions of the previous hops, the per-hop optimal strategy obtained in Theorem 4.1 may not be an optimal strategy for problem (2). Hence, taking the dependence of the links' delays into consideration, we give the following theorem, which shows that there are only two possibly optimal link strategies, i.e., s^1 and s^∞, and gives a method to select the optimal link strategy to achieve the minimal expected path delay for solving problem (2).

Theorem 4.2. *Given a path from* s *to* d, *if* $t_{ij}^b + \mathbf{E}\{t_{jd}|\tilde{t}_j^a = 0\} \le \mathbf{E}\{t_{id}|\tilde{t}_i^a = 0, s_i^t = s^\infty\}$, *the optimal strategy for link* $< i,j >$ *is* s^1; *otherwise, the optimal strategy for link* $< i,j >$ *is* s^∞.

Proof. Since the taxi flow between i and j follows the Poisson distribution, if the first bus will not be selected, i.e., s^1 is not the optimal strategy for node i, then all the following buses will not be selected, i.e., s^∞ is the optimal strategy for node i. Therefore, there are only two per-hop optimal strategies for problem (2), i.e., s^1 and s^∞.

When $t_{ij}^b + \mathbf{E}\{t_{jd}|\tilde{t}_j^a = 0\} \le \mathbf{E}\{t_{id}|\tilde{t}_i^a = 0, s_i^t = s^\infty\}$ holds, selecting the first bus will have a smaller total expected delay than that for always waiting for taxi; consequently, s^1 is the best strategy. Similarly, s^∞ is the best strategy when $t_{ij}^b + \mathbf{E}\{t_{jd}|\tilde{t}_j^a = 0\} > \mathbf{E}\{t_{id}|\tilde{t}_i^a = 0, s_i^t = s^\infty\}$. □

According to the above theorem, in order to minimize the expected delay, one can obtain the optimal strategy of each link for a given path by comparing the values of the conditional expected delay of $t_{ij}^b + \mathbf{E}\{t_{jd}|t_j^a = 0\}$ and $\mathbf{E}\{t_{id}|t_i^a = 0, s_i^t = s^\infty\}$. How to calculate the exact value of the conditional expected delay is quite complicated, and in the following sections, we will analyze and design an algorithm to calculate it, as well as the expected end-to-end delay for each given path with Theorems 4.1 and 4.2.

4.2 Path Delay Analysis

In this section, we present the theoretical results on the delay of a given path, which are necessary for the design of the optimal routing algorithm.

From the discussion in the last section, for a given path, there are two per-hop optimal strategies, i.e., s^1 and s^∞, so we need to consider two cases. First, when all nodes along the path take s^∞, the expected delay for each link is independent of each other, and we call this the fully random case. Second and more difficult, when part of the nodes along the path take s^1, the expected delays for these hops depend on the previous hops and the arrival times of the message at these nodes, and we call it the mixed case. Note the case that all nodes in the path taking s^1 is a special case of the mixed case, so we skip it here due to the space limit.

Fully random case. We consider the fully random case first. Referring to the theoretical results in [22], [27], we have the following lemma which provides the PDF of the total waiting time for a given path.

Lemma 4.3. *Assume that a message is delivered from* i_0 *to* i_m *according to a path* $i_0 \to i_1 \to \cdots \to i_m$, *and the associated optimal strategy for each node is* s^∞. *We have*

$$f_{i_0 i_m}(t) = f_{i_0 i_1}(t) * f_{i_1 i_2}(t) * \cdots * f_{i_{m-1} i_m}(t), \tag{3}$$

where $*$ *is the convolution operator.*

Combining the definition of convolution with Theorem 2 of [22] yields

$$f_{i_0 i_m}(t) = \sum_{l=1}^{m} C_l^m \lambda_l e^{-\lambda_l t}, \tag{4}$$

where $\lambda_l = \lambda_{i_{l-1} i_l}$ and $C_l^m = \prod_{s=1, s\neq l}^m \frac{\lambda_s}{\lambda_s - \lambda_l}$ (when $\lambda_s = \lambda_l$, it is obtained by taking the limit). From (4), one infers that the expected delay of $t_{i_0 i_m}$ satisfies

$$\mathbf{E}\{t_{i_0 i_m}\} = \sum_{k=1}^{m}\left(t_{i_{k-1} i_k}^c + \frac{1}{\lambda_{i_{k-1} i_k}}\right). \tag{5}$$

Hence, for the fully random case, the above equation gives a closed-form expression of the expected delay.

Mixed case. Next, we consider the mixed case. When the per-hop optimal strategy is given for each node, we can divide a path into several segments, where each segment has its first m nodes taking s^∞ and the last n nodes taking s^1. We note that the delay of each segment is indeed independent to each other. Thus, by summing up the expected delay of each segment, the expected delay of the path can be obtained. Consequently, it is sufficient to analyze the delay of a path $p_{i_0 i_{m+n}}$ as follows

$$i_0 \to i_1 \to \cdots \to i_m \to i_{m+1} \to \cdots \to i_\ell,$$

where $\ell = m + n$, and the one-hop optimal strategy for the first m nodes is s^∞ and that for the last n nodes is s^1 ($m, n \ge 1$). Note that when with s^1, the delay analysis becomes much more difficult as the waiting time depends on the arriving time at the node. Thus, the conditional expectation should be considered for the delay computation.

To simplify the notation, in the following, we let $t_k^a = t_{i_k}^a$, $t_k = t_{i_{k-1} i_k}$, $\lambda_k = \lambda_{i_{k-1} i_k}$ and $f_{i_{k-1} i_k} = f_k$ for $k = 1, 2, \ldots,$

$m+n$. Assuming that the interval of the bus arrival and departure time at each node is negligible (how to remove this assumption will be discussed in Section 4.3), so under s^1, the message being carried by a bus will be carried by a bus in the next hop. Therefore, we only need to study the situation that for the first $m+k$ nodes, the taxies are selected, while at the remaining $n-k$ nodes, buses are selected as the message carriers for $0 \leq k \leq n$.

Let $t^k_{m,n}$ be the total travel time from i_0 to i_{m+n} (excluding the waiting time in each node), which is given by

$$t^k_{m,n} = \sum_{i=1}^{m+k} t^c_i + \sum_{j=n-k}^{n} t^b_{m+j}, \quad \text{for } 0 \leq k \leq n.$$

For $1 \leq i \leq m$ and $1 \leq k \leq n$, we define

$$\tilde{C}^{m,k}_i(j) = \begin{cases} \prod_{s=m+1}^{m+k} \frac{\lambda_s}{\lambda_s - \lambda_i}, & j = i; \\ \frac{\lambda_i}{\lambda_i - \lambda_j} \prod_{s=m+1, s\neq j}^{m+k} \frac{\lambda_s}{\lambda_s - \lambda_j}, & j \neq i. \end{cases} \tag{6}$$

There are six notations (see in Appendix B, available in the online supplemental material), including A and B (and its derivatives), defined for simplifying the statement of the theoretical results and their proof. All these notations can be obtained from simple computation. We then have the following two theorems whose proofs are given in Appendix B and Appendix C, available in the online supplemental material, respectively.

First, the following theorem is to study the PDF of time $\hat{t}^a_{m+n}(= \hat{t}^a_\ell)$ and the corresponding probability of that $\hat{t}^a_\ell = 0$ or $\hat{t}^a_\ell \neq 0$. This theorem is very important for computing the expected delay, and also used in the optimal path selection algorithm in the next section.

Theorem 4.4. *Suppose that a message is delivered following path $p_{i_0 i_{m+n}}$, and $\ell = m+n$, $t^a_0 = t_0$ and $\tilde{t}_0 > 0$. We have*

1) *the probability of $\hat{t}^a_\ell \in (0, T)$ is*

$$\Pr\{\hat{t}^a_\ell \in (0,T)\} = \dot{A}^{m,n}_0(\tilde{t}_0) + \dot{A}^{m,n}_1(T) \tag{7}$$

and the PDF of $\hat{t}^a_\ell$ is given by

$$\hat{f}^a_\ell(t) = \begin{cases} A^{m,n}_1(t), & 0 < t < \hat{t}_0, \\ f_{i_0 i_\ell}(t - \hat{t}_0) + A^{m,n}_1(t, \hat{t}_0), & \hat{t}_0 \leq t < T, \end{cases} \tag{8}$$

where $A^{m,n}_1(t, \hat{t}_0) = A^{m,n}_1(t) - A^{m,n}_1(\hat{t}_0)$;

2) *the probability of $\hat{t}^a_\ell = 0$ is*

$$\Pr\{\hat{t}^a_\ell = 0\} = 1 - \dot{A}^{m,n}_0(\tilde{t}_0) - \dot{A}^{m,n}_1(T). \tag{9}$$

Note that when $\tilde{t}_0 = 0$, we have $\hat{t}_0 = T$. From the above theorem, one infers that $\Pr\{\hat{t}^a_\ell \in (0,T)\} = \dot{A}^{m,n}_1(T)$ and

$$\hat{f}_\ell(t) = A^{m,n}_1(t), \quad \text{for } 0 < t < T, \tag{10}$$

and

$$\Pr\{\hat{t}^a_\ell = 0\} = 1 - \dot{A}^{m,n}_1(T). \tag{11}$$

Then, the next theorem provides the closed-form expression of the expected delivery delay when the message is delivered following path $p_{i_0 i_{m+n}}$.

Theorem 4.5. *Suppose that the message is delivered following path $p_{i_0 i_{m+n}}$ and $t^a_0 = t_0$. We have*

$$\mathbf{E}\{t_{i_0 i_{m+n}}\} = \sum_{i=0}^{n} \left(A^0_i + A^1_i + A^2_i\right), \tag{12}$$

where A^0_i, A^1_i and A^2_i are given in the proof, which are the functions of t_0, T and the Poisson parameters.

The computational complexity of each A^0_i, A^1_i and A^2_i is less than $O((m+n)^3)$ since the parameters in their expressions are less than $O((m+n)^2)$ as discussed earlier. Hence, the computational complexity of $\mathbf{E}\{t_{i_0 i_{m+n}}\}$ is less than $O((m+n)^4)$. Note that when $\tilde{t}_0 = 0$, one can easily obtain that $A^0_i = 0$ for $i = 0, 1, \ldots, n$. Hence, we have

$$\mathbf{E}\{t_{i_0 i_\ell}\} = \sum_{i=0}^{n} \left(A^1_i + A^2_i\right), \quad \text{if } \tilde{t}_0 = 0. \tag{13}$$

Meanwhile, for the above two theorems, when $m = 0$, it is equivalent to that setting $\lambda_1 = \infty$ and $t^b_{i_0 i_1} = t^c_{i_0 i_1} = 0$ for $m = 1$. Hence, we can obtain the associated results for $m = 0$ by taking limiting over λ_1, i.e., taking limiting on the RHS of (12) to obtain $\mathbf{E}\{t_{i_0 i_{m+n}}\}$. Similarly, when $n = 0$, we set $\lambda_{m+1} = \infty$ and $t^b_{i_m i_{m+1}} = t^c_{i_m i_{m+1}} = 0$.

4.3 Optimal Path Selection Algorithm

Using the theoretical results obtained above, we can design an algorithm to solve problem (2), which is given in Algorithm 1. Using Theorem 4.1, we can obtain the optimal strategy for minimizing the link delay. In Line 2 of the Algorithm, clearly, for the last link in a given path, its optimal strategy to minimize its link delay is also the optimal link strategy for minimizing the path delay. In Line 3, using Theorem 4.4, we can calculate the expected delay of the last link when the arrival time at it (node l) is given, which is a conditional expectation. In Lines 4 to 8, using Theorem 4.5, we can calculate the PDF of the arrival time at node l when the link strategy for $< j, l >$ and t^a_j are given, and then calculate the expected delay from j to d. Then, we can obtain the optimal link strategy of $< j, l >$ by using Theorem 4.2 to minimize the expected delay from j to the destination. By repeating this process backward on the previous links, we can obtain the optimal link strategy of each link in the path to minimize the expected delay for the whole path and to solve problem (2). Hence, in Line 9, we can obtain the optimal solution of problem (2) by comparing the delay of each path candidate.

According to Algorithm 1, one obtains the optimal strategy for each node in any given paths, and also the optimal path and its expected delay for problem (2). Note that in Algorithm 1, we calculate the expected delay hop-by-hop starting from the last hop in a given path, and then from Steps 3 to 6 we can set $m = 0$ and $n = 1$ in Theorems 4.4 and 4.5 to calculate the one-hop conditional expected delay, which means that it can take the non-zero interval of the bus arrival and departure time at each node into consideration. Hence, Algorithm 1 is general enough to remove the assumption that the interval of the bus arrival and departure time at each node is negligible.

Algorithm 1. Optimal Path Selection Algorithm

```
Input: the source and destination, the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and the statistics of each link including inter-arrival time and average travel time in the graph.
1: Let $t_s^a = 0$. Denote $\mathbf{E}_k\{t_{sd}\} = \mathbf{E}_k\{t_{sd}|t_s^a = 0\}$ to be the expected delay of the $k$th path, $p_{sd}^k$, for $k = 1, 2, \ldots, N$.
2: For $p_{sd}^k$, if $t_{ld}^b > t_{ld}^c + \frac{1}{\lambda_{ld}}$ for the last link, say $< l, d >$, let $s^\infty$ be the link optimal strategy of node $l$; otherwise, let $s^1$ be the link optimal strategy of node $l$.
3: Let $s^\infty$ be the link-strategy of node $j$. Calculate $\mathbf{E}_k\{t_{ld}|\tilde{t}_l^a\}$ and $\mathbf{E}_k\{t_{jd}|\tilde{t}_j^a\}$ with using (5) (when $s^\infty$ is for $< l, d >$) or (12) (when $s^1$ is for $< l, d >$). Then, set the optimal link-strategy of node $j$ based on Theorem 4.2.
4: Suppose in path $s \rightarrow \cdots \rightarrow i \rightarrow j \rightarrow \cdots \rightarrow d$, each link optimal strategy from $j$ to $d$ and $\mathbf{E}_k\{t_{jd}|\tilde{t}_j^a\}$ are obtained. Let $s^\infty$ be the link-strategy of node $i$.
5: Calculate the PDF $\hat{f}_{ij}(t|\tilde{t}_i^a)$ of $\hat{t}_d$ with (8) and (10), and the conditional probability $\Pr\{\hat{t}_j = 0|\tilde{t}_i^a\}$ with (9) and (11),
6: Calculate $\mathbf{E}_k\{t_{id}|\tilde{t}_i^a\}$ by
       $\mathbf{E}_k\{t_{id}|\tilde{t}_i^a\} = \Pr\{\hat{t}_j = 0|\tilde{t}_i^a\}\mathbf{E}_k\{t_{jd}|\tilde{t}_j^a = 0\} + \int_0^T \hat{f}_{ij}(t|\tilde{t}_i^a)\mathbf{E}_k\{t_{jd}|\tilde{t}_j^a = T - t\}dt + \mathbf{E}_k\{t_{ij}|\tilde{t}_i^a\}.$   (14)
7: Let $\tilde{t}_j^a = 0$, then compare the values of $t_{ij}^b + \mathbf{E}\{t_{jd}|\tilde{t}_j^a = 0\}$ and $\mathbf{E}\{t_{id}|\tilde{t}_i^a = 0\}$. If $t_{ij}^b + \mathbf{E}\{t_{jd}|\tilde{t}_j^a = 0\} \leq \mathbf{E}\{t_{id}|\tilde{t}_i^a = 0\}$, let $s^1$ be the optimal link strategy, i.e., $s_i^{t^*} = s^1$; otherwise, let $s^\infty$ be the optimal link strategy, i.e., $s_i^{t^*} = s^\infty$.
8: If $s_i^{t^*} = s^1$, recalculate $\mathbf{E}_k\{t_{id}|\tilde{t}_i^a\}$ by repeating Steps 5 and 6 under the setting of $s_i^t = s^1$.
9: If $i = s$, $\mathbf{E}_k\{t_{sd}\} = \mathbf{E}_k\{t_{id}|\tilde{t}_i^a = 0\}$; otherwise, go to Step 4.
10: Repeat Steps 2 to 9 until $k = N$. Compare $\mathbf{E}_k\{t_{sd}\}$ for $k = 1, 2, \ldots, N$ and let
       $k^* = \arg\max_{k=1,2,\ldots,N}\mathbf{E}_k\{t_{sd}\}.$
Output: the optimal path for problem (2) is $p_{sd}^{k^*}$.
```

To speed up the route selection process, we first use the generalized Dijkstra's shortest-path algorithm combined with the Approximation approach to select the K shortest paths. In the Approximation approach, the path delay is the sum of all the link delays without considering the interdependence of the delay components, and the link delay is simplified by $\lambda_{ij} + t_{ij}^c$ for s^∞ and t_{ij}^b for s^1, using the optimal link strategy in Theorem 4.1. K is typically a small integer, and is determined by the criteria that the delay of the Kth shortest path is much smaller than that of the $(K+1)$th path. Then, these selected K paths will be viewed as the elements of $\mathbf{\Omega}$ and let $N = K$, which will largely reduce the number of paths in $\mathbf{\Omega}$, and thus can reduce the computational complexity of Algorithm 1 and speed up the optimal path selection.

Algorithm 1 is an off-line algorithm which obtains the optimal path between the source node and node(s) in the destination location in terms of the average delivery delay. One can further design an online data forwarding process based on Algorithm 1. Note that in Algorithm 1, each node i will obtain $\mathbf{E}\{t_{id}|\tilde{t}_i^a\}$ for every path between i to d, i.e., node i can achieve the minimum delay $\mathbf{E}^*\{t_{id}|\tilde{t}_i^a\}$ when the $\tilde{t}_i^a$ is given. When there is a vehicle (assuming a taxi here) traveling from any node l (the forwarding data is stored) to node i, the vehicle should be selected as the carrier if

$$\mathbf{E}^*\{t_{id}|\tilde{t}_i^a\} + t_{li}^c \leq \mathbf{E}^*\{t_{ld}|\tilde{t}_l^a\},$$

which is the main idea that can be adopted to design an online data forwarding algorithm.

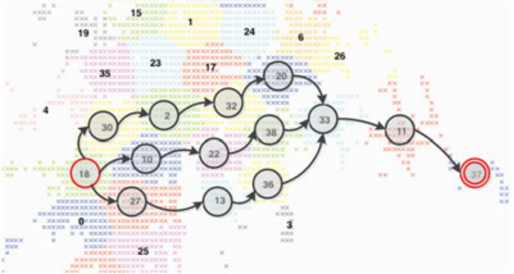


Fig. 3. VANET covering Shanghai city, with three paths from Hongqiao airport (cluster 18) to Pudong airport (cluster 37).

5 Performance Evaluation

To validate the analysis and compare with the state-of-the-art solutions, we conducted extensive simulations, considering the realistic topology using the real traces of taxis and buses in Shanghai. We first introduce some delay estimation approaches. The resultant path delay obtained from Approximation approach (described in Section 4.3) is named "Approximation". If using the link strategy according to Theorem 4.1, then the derived expected delay using Theorem 4.5 is named "Suboptimal". If using the optimal link strategy in Theorem 4.2, the derived expected delay using Theorem 4.5 is named "Theory". The average delay obtained from the Monte Carlo simulation is named "Simulation". By comparing these delay estimation methods, we will demonstrate the effectiveness of our analytical results ("Theory") in delay estimation and the optimal path selection, and reveal that using the simple approximation methods to calculate the delay and select the optimal path may reduce effectiveness. Then, we will also compare the performance of our approach with that of the state-of-the-art geocast solution in [15] in terms of the delivery delay, delivery ratio and overhead-ratio. In the simulation, we consider the east-bound vehicle traffic and the links only that constitute a subset of the graph, so we do not maintain the flow conservation law for this partial graph. The proposed optimal link strategy is used in the simulation as well. Since the transmission delay of a message is assumed to be much smaller than the vehicle travel time, we also ignore this delay component.

5.1 Simulation Settings

Trace information. To demonstrate the feasibility of our solution in a large-scale VANET, we consider a realistic setting using the real data traces (partially available at http://www.cse.ust.hk/scrg) collected from about 2,300 taxicabs and 2,500 buses in Shanghai from February to March 2007. Using the same approach as [15], we divided the area of Shanghai city into 40 regions using the travel distance-based clustering method, where the distance is obtained through online map services, e.g., Google maps. The clustering result is shown in Fig. 3 (same as Fig. 2 in [15]), where the regions

TABLE 2
Delay Comparison, Hongqian to Pudong (in Minutes)

	First Vehicle	Bus Only	Taxi Only	[15]	Simulation	Theory	Suboptimal	Approximation
P1	196.443	211.497	199.399	N/A	157. 879	159.484	159.484	158.389
P2	174.675	181.504	237.519	N/A	162.137	160.835	163.474	158.254
P3	173.175	181.499	399.317	N/A	167.274	166.048	166.048	157.985
OPT	173.175	181.497	199.399	196.819	157. 879	159.484	159.484	157.985

are numbered and colored. In each region, the location with the highest vehicle density was selected as the hot-spot location for installing the drop box.

Scenario settings. In the simulation, we used the real trace. Specifically, for each simulation run, we randomly selected a time from 12:00am-12:00pm as the starting time of a message to be forwarded in the source region. When the message arrives at a hot spot in each region, the optimal link strategy obtained form Theorem 4.5 is used for the next-hop data delivery strategy. For example, suppose that a data arrives at node i at time t_i^a and s_i^∞ is the optimal strategy from node i to node j, then the first taxi in time interval $[t_i^a, \infty)$ traveling from node i to node j (selected from the real trace) will be selected as the data carrier. Based on the simulation, we then obtain the statistics of the delay (e.g., expected delay) using the routing solution. The average travel times of each link for taxis and buses are 15 and 30 minutes with a fluctuation of up to ± 20 percent, respectively, and the bus interval is $T = 15$ minutes. In the simulation, we first select Hongqiao airport (cluster 18) and Pudong airport (cluster 37) as the source and the destination node, s and d, respectively. Then, we consider the case where more pairs of s and d (randomly selected among 40 nodes) are adopted to evaluate the performance of proposed scheme in this paper.

5.2 Simulation Results

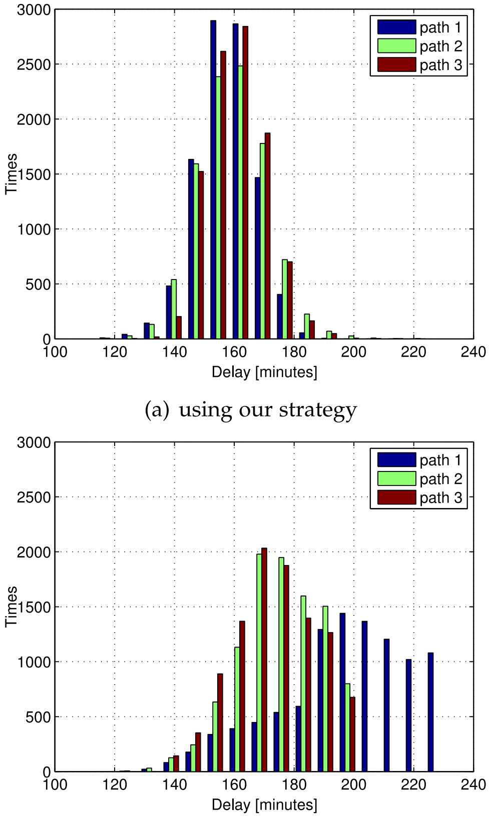
For comparison, we consider three simple strategies that use s^1 only, Bus only and s^∞ only, respectively, and the results are in the columns of “First Vehicle”, “Bus Only” and “Taxi Only” in Table 2. In the other columns of the table, we include the simulation results of the average delay using the state-of-the-art geocast solution in [15] and the simulation results using the proposed optimal link strategy, respectively, followed by the analytical expected delay of the three paths obtained using the optimal link strategy from Theorem 4.2, the link strategy from Theorem 4.1, and the simple approximation, respectively. In the following, we first selected one representative pair of source and destination nodes, representing Hongqiao airport and Pudong airport, which has a large traffic volume and fits very well for the application scenario considered in this paper. We have the following key observations and conclusions.

First, as shown in Table 2, if we use the first vehicle only or buses only or taxis only for data dissemination, the resultant shortest paths have the delay of 173.175, 181.497, and 199.399 minutes, respectively, which are much higher than our proposed optimal solution (157.879 minutes) using both buses and taxis with the optimal link strategy. The bus-only option has a smaller average end-to-end delay as the number of taxis traced is limited so the taxi arrival rates in certain links are low, which renders a long taxi waiting time. In addition, consider the delay of 196.443 minutes and 157.879 minutes in P1, there is a minimum traveling time (105 min) for both of them which cannot be reduced by any link strategy. By removing the traveling time, the reminder of the delay (which is the waiting time) is actually reduced from 91.443 to 52.879 minutes thanks to our link strategy. Hence, the reduction ratio of the waiting time is $\frac{91.443-52.879}{91.443} = 42.2\%$.

Second, using the approach in [15], the simulation results show that for the same source/destination pair and the same vehicle traffic density, the average hop count is six, similar to ours. However, using their solution, not only the average delay is much higher, but also the delivery ratio for Hongqiao to Pudong airport is 51 percent only. In our approach, with the assistance of drop boxes, all the messages can be delivered successfully and with a much lower average delay. Thus, the proposed framework makes a good tradeoff between a small cost of installing and maintaining drop boxes and the substantial performance gains in both reliability and delay. More importantly, our approach has a much lower overhead ratio compared with that of the approach in [15], where the overhead ratio is defined as the ratio of the total number of message transmissions to the number of transmissions for those successfully delivered messages. The average overhead-ratio is 275.0615 using the approach in [15] while it is less than 15 for each path under our approach. The reason that [15] requires a much higher overhead is that, without drop boxes, vehicles need to exchange messages frequently with many vehicles they encounter till the message reaches the destination. Obviously, when the vehicle network has a higher density or a larger coverage, the overheads will further increase. Thus, the proposed framework is more scalable.

Third, the analytical results (“Theory”) and the simulation results match well, which demonstrates the correctness of our analysis. In addition, since the approximation method ignores the dependency of the delays along the path, the delay estimations using the approximation contain non-negligible errors. Consequently, the optimal path obtained from the approximation method is Path 3, which is actually the worst path according to the simulation and theoretical results. Thus, we should not use the simple approximation method to calculate the delay and select the optimal path.

Finally, comparing “Theory” and “Suboptimal”, for the links in Path 1 and Path 3, the link strategies obtained from Theorem 4.1 are the same as the optimal link strategies obtained from Theorem 4.2. However, for Path 2, the optimal strategy for the link from cluster 22 to cluster 38 should be s^1 which is different from the link strategy from Theorem 4.1. This is why the path delay in “Theory” and in “Suboptimal” for Path 2 are different. The results suggest



(a) using our strategy

(b) using first-met vehicle to forward

Fig. 4. Histogram of the delivery delay from Hongqiao to Pudong.

that the relatively simple suboptimal approach may be adopted with cautions.

Fig. 4a shows the histogram of the delivery delay for the three paths with 10^4 simulation runs under our strategy. From the figure, all messages can be delivered with a delay less than 240 minutes, and Path 1 has a much higher chance of delay no greater than 180 minutes. Fig. 4b shows the histogram of the delay for the three paths with 10^4 simulation runs under the first vehicle strategy. In this case, it is observed that the three paths have a high chance of delay greater than 180 minutes especially for Path 1. Comparing the results in Fig. 4a with those in Fig. 4b, one sees that the message delivery delay using our strategy is much better than that under the first vehicle strategy.

Next, we further randomly (not artificially) selected five pairs of source and destination nodes which are separated by at least three hops for validation. The optimal average message delivery delays under the five pairs of s and d, i.e., clusters 18 and 37, clusters 33 and 10, clusters 28 and 18, clusters 20 and 18, and clusters 18 and 1 (selected randomly), are shown in Table 3. It is observed that the error between the simulation and theory is smaller than 1.25 percent. Hence, using our theory, we can obtain a highly accurate estimation of the expected message delivery delay, which validates the results in the obtained strategies. It should be pointed out that we have repeated the simulation with other source and destination pairs, and the results show the same tendency. Since our theoretical results have taken the delay dependence into consideration, they provide highly accurate delay estimation. Therefore, based on our theoretical results, one can select the optimal path and strategy, no matter which pair of source and destination nodes is selected.

TABLE 3
Delay Comparison between Simulation and Theory (in Minutes)

Clusters	18 to 37	33 to 10	28 to 18	20 to 18	18 to 1
Simulation	157.897	91.224	105.344	76.504	67.197
Theory	159.484	91.7122	104.441	77.065	67.214

5.3 Further Discussions

Utilizing the drop boxes or the similar devices (e.g., throwboxes) as relays in VANETs or DTNs can increase the data delivery probability, reduce the delivery delay, and enhance the network capacity, etc., since the contact probability between the vehicles or mobile nodes can be increased with them [2], [3], [32], [33], [34]. In the situation that the drop boxes are distributed in multiple parked vehicles, how to manage and coordinate them considering the vehicle mobility pattern requires further investigation.

Note that the communication capacity and storage capacity of any type of vehicles (e.g., bus-only) is limited, so it prefers to distribute the information to various types of vehicles to improve the network capacity. Thus, taking both buses and taxis into consideration, it can not only reduce the delivery delay (as shown in the above simulation), but also improve the network capacity. Meanwhile, we prefer vehicle-to-vehicle communications in order to distribute a large amount of delay-tolerant information in a low-cost and distributed manner. The application scenario can be advertisements sent to interested locations, e.g., the hotel, restaurant, attraction, and other local business advertisements and promotions need to be sent to passengers near the airport, ferry, coach bus terminals, etc., and such information (can be in a large volume and needs to be updated every several days) is targeted to vehicles in certain locations, instead of specific vehicles, so we need a low-cost solution for them. Thus the alternative solutions such as using cellular systems or Wi-Fi may not be desirable due to higher cost and/or limited capacity. It is possible to combine the several wireless technologies together and optimize the data delivery over them jointly, which is an interesting further research issue. In addition, when the storage capacity is limited, how to prioritize messages according to their waiting time is a very interesting issue. It is possible to formulate a utility-based optimization problem which will be left as our future work.

Given the store-and-forward VANET framework, there are many open issues worth further investigation. In this paper, source routing is used, so the path is selected by the source node and will not change along the path and over the time once selected. If we let each hop select the next hop dynamically, the problem becomes much more complicated. Each hop needs to make a decision whenever a taxi in its coverage will leave for another region. The optimal decision needs to consider the expected delay for the following link as well as those for all the links in all possible paths from the current node to the destination. Second, we now assume that message handover can only happen at hot spots, and it is also possible to allow a faster taxi to take the message from a slower bus along the road to further accelerate the process. Third, if the size of the message is large, the transmission delay should be considered and added into the per-link delay. Since the transmission delay is independent of other delay components, it is not difficult to consider it. Furthermore, how to deal with the situation that the taxi arrival follows other distributions and how to consider other performance metrics such as the delay jitter in optimal routing are important further research issues. Finally, as a first attempt to the delay analysis in the situation that random processes and deterministic processes co-exist, there are many possible applications that can leverage our analytical framework, e.g., for mobile social networks where both the deterministic and random mobility patterns exist, for multi-modal travel planning, etc.

6 Conclusions

In this paper, we have investigated how to minimize the expected end-to-end delay in a large-scale VANET with both fixed-scheduled buses and random-scheduled taxis. We first proposed a store-and-forward framework, by introducing drop boxes which function similarly to routers with extra storage. Second, we proposed an optimal link strategy which is independent of the message arrival time and can be executed in a distributed manner. Third, we derived the expected path delay, considering the dependency of the delay components along each path and proposed the optimal routing solution to minimize the path delay. Simulations driven by real traces in Shanghai have been used to validate the rigor of the analysis, and demonstrate the superior performance of the proposed solution, which results in a substantial delay reduction and a much higher delivery ratio compared with the state-of-the-art solutions without drop boxes. The solution further improves the delay performance compared with the over-simplified routing solutions which ignore the dependence of the delay components.

Acknowledgments

The authors would like to thank the editor and the anonymous reviewers for their helpful and constructive comments, which have helped us improve the manuscript significantly. This work was supported by 973 program under grant 2015CB352500, NSFC under grant 61533015 and 61503332, CPSF under grant 2014M551739 and 2015T80614, ZJSF under grant LY14F030016, and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] C. Maihofer, "A survey of geocast routing protocols," *IEEE Commun. Surveys Tuts.*, vol. 6, no. 2, pp. 32–42, Apr. 2004.

[2] M. J. Lok, B. R. Qazi, and J. M. Elmirghani, "Data dissemination with drop boxes," in *Proc. IEEE Int. Conf. Adv. Inf. Netw. Appl.*, 2009, pp. 451–455.

[3] W. Zhao, Y. Chen, M. Ammar, M. Corner, B. Levine, and E. Zegura, "Capacity enhancement using throwboxes in DTNs," in *Proc. IEEE MASS*, 2006, pp. 31–40.

[4] R. E. Bellman, *Dynamic Programming.* Princeton, NJ, USA: Princeton Univ. Press, 2003.

[5] A. Lindgren, A. Doria, and O. Schelen, "Probabilistic routing in intermittently connected networks," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 7, no. 3, pp. 19–20, 2003.

[6] J. Burgess, B. Gallagher, D. Jensen, and B. Levine, "MaxProp: Routing for vehicle-based disruption-tolerant networks," in *Proc 25th IEEE Int. Conf. Comput. Commun.*, 2006, pp. 1–11.

[7] 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.

[8] 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.

[9] H. Zhu, M. Dong, S. Chang, Y. Zhu, M. Li, and S. Shen, "Zoom: Scaling the mobility for fast opportunistic forwarding in vehicular networks," in *Proc. IEEE INFOCOM*, 2013, pp. 2832–2840.

[10] T. Luan, L. Cai, J. Chen, S. Shen, and F. Bai, "Engineering a distributed infrastructure for large-scale cost-effective content dissemination over urban vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1419–1435, Mar. 2014.

[11] H. Zhou, J. Chen, J. Fan, Y. Du, and K. Sajal, "ConSub: Incentive-based content subscribing in selfish opportunistic mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 669–679, Sep. 2013.

[12] S. Mao, S. Panwar, and Y. T. Hou, "On minimizing end-to-end delay with optimal traffic partitioning," *IEEE Trans. Veh. Technol.*, vol. 55, no. 2, pp. 681–690, Mar. 2006.

[13] X. Li, W. Shu, M. Li, P. Luo, H. Huang and M-Y. Wu, "Performance evaluation of vehicle-based mobile sensor networks for traffic monitoring," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 1647–1653, May 2009.

[14] A. Takahashi, H. Nishiyama, N. Kato, K. Nakahira, and T. Sugiyama, "Replication control for ensuring reliability of convergecast message delivery in infrastructure-aided DTNs," *IEEE Trans. Veh. Technol.*, vol. 63, no. 7, pp. 3223–3231, Sep. 2014.

[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. 1279–1787.

[16] N. Lu, T. Luan, M. Wang, X. Shen, and F. Bai, "Capacity and delay analysis for social-proximity urban vehicular networks," in *Proc. IEEE INFOCOM*, Mar. 2012.

[17] U. 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.

[18] E. H. Ngai, J. Liu, and M. R. Lyu, "An adaptive delay-minimized route design for wireless sensor-actuator networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, pp. 5083–5094, Nov. 2009.

[19] X. Zhu, P. Li, Y. Fang, and Y. Wang, "Throughput and delay in cooperative wireless networks with partial infrastructure," *IEEE Trans. Veh. Technol.*, vol. 58, no. 8, pp. 4620–4627, Oct. 2009.

[20] S. Ioannidis, A. Chaintreau, and L. Massoulie, "Optimal and scalable distribution of content updates over a mobile social network," in *Proc. IEEE INFOCOM*, 2009, pp. 1422–1430.

[21] U. Lee, S.-Y. Oh, K.-W. Lee, and M. Gerla, "Scalable multicast routing in delay tolerant networks," in *Proc. IEEE Int. Conf. Netw. Protocols*, 2008, pp. 218–227.

[22] W. Gao, Q. Li, B. Zhao, and G. Cao, "Multicasting in delay tolerant networks: a social network perspective," in *Proc. 10th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2009, pp. 299–308.

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

[24] F. Bai and B. Krishnamachari, "Spatio-temporal variations of vehicle traffic in VANETs: Facts and implications," in *Proc. 6th ACM Int. Workshop Veh. InterNETworking*, 2009, pp. 43–52.

[25] Y. Ko and N. Vaidya, "Flooding-based geocasting protocols for mobile ad hoc networks," *Mobile Netw. Appl.*, vol. 7, no. 6, pp. 471–480, 2002.

[26] I. Stojmenovic, A. Ruhil, and D. Lobiyal, "Voronoi diagram and convex hull based geocasting and routing in wireless networks," *Wireless Commun. Mobile Comput.*, vol. 6, no. 2, pp. 247–258, 2006.

[27] J. He, P. Cheng, L. Shi, and J. Chen, "Clock synchronization for random mobile sensor networks," in *Proc. IEEE 51st Annu. Conf. Decision Control*, 2012, pp. 2712–2717.

[28] B. Gu, and X. Hong, "Optimal routing strategy in throw-box based delay tolerant network," in *Proc 6th Int. ICST Conf. Commun. Netw. China*, 2011, pp. 501–506.

[29] N. Banerjee, M. D. Corner, D. Towsley, and B. N. Levine, "Relays, base stations, and meshes: Enhancing mobile networks with infrastructure," in *Proc. 14th ACM Int. Conf. Mobile Comput. Netw.*, 2008, pp. 81–91.

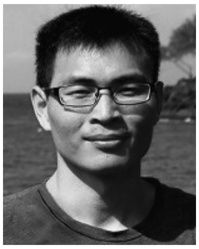
[30] Y. Ding, C. Wang, and L. Xiao, "A static-node assisted adaptive routing protocol in vehicular networks," in *Proc. 4th ACM Int. Workshop Veh. Ad Hoc Netw.*, 2007, pp. 59–68.

[31] Y. Xiong, J. Ma, W. Wang, and J. Niu, "Optimal roadside gateway deployment for VANETs," *Przeglad Elecktrotechnicznyt*, vol. 88, no. 7B, pp. 273–276, 2012.

[32] Y. Wu, Y. Zhu, and B. Li, "Infrastructure-assisted routing in vehicular networks," in *Proc. IEEE INFOCOM*, 2012, pp. 1485–1493.

[33] K. Wong, B. Lee, B. Seet, G. Liu, and L. Zhu, "BUSNet: Model and usage of regular traffic patterns in mobile ad hoc networks for inter-vehicular communications," in *Proc. IEEE 10th Int. Conf. Telecommun.*, 2003, pp. 102–108.

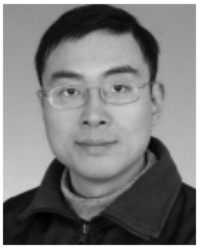
[34] H. Zhou, J. Chen, H. Zhao, W. Gao, and P. Cheng, "On exploiting contact patterns for data forwarding in duty-cycle opportunistic mobile networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 9, 4629–4642, Nov. 2013.



Jianping He (M'15) received the PhD degree in control science and engineering in 2013, from Zhejiang University, Hangzhou, China. He is currently a postdoctoral research fellow in both the State Key Laboratory of Industrial Control Technology at Zhejiang University and the Department of Electrical and Computer Engineering at the University of Victoria. He is a member of Networked Sensing and Control group (NESC). His research interests include the control and optimization of sensor networks and cyber-physical systems, the scheduling and optimization in VANETs and social networks, and the investment decision in financial market and electricity market. He is a member of the IEEE.

Lin Cai (S'00-M'06-SM'10) received the MASc and PhD degrees in electrical and computer engineering from the University of Waterloo, Waterloo, Canada, in 2002 and 2005, respectively. Since 2005, she has been with the Department of Electrical & Computer Engineering at the University of Victoria, where she is currently a professor. Her research interests span several areas in communications and networking, with a focus on network protocol and architecture design supporting emerging multimedia traffic over wireless, mobile, ad hoc, and sensor networks. She has been a recipient of the NSERC Discovery Accelerator Supplement Grants in 2010 and 2015, and the best paper awards of IEEE ICC 2008 and IEEE WCNC 2011. She has served as a TPC symposium co-chair for IEEE Globecom'10 and Globecom'13, and the associate editor for the *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Vehicular Technology*, *EURASIP Journal on Wireless Communications and Networking*, *International Journal of Sensor Networks*, and *Journal of Communications and Networks* (JCN). She is a senior member of the IEEE.

Peng Cheng (M'10) received the BE degree in automation, and the PhD degree in control science and engineering, in 2004 and 2009, respectively, both from Zhejiang University, Hangzhou, P.R. China. Currently, he is an associate professor with the Department of Control Science and Engineering, Zhejiang University. He serves as an associate editor for *Wireless Networks* and the *International Journal of Distributed Sensor Networks*. He is also the guest editor for the *IEEE Transactions on Control of Network Systems*. He served as the publicity co-chair for IEEE MASS 2013 and local arrangement chair for ACM MobiHoc 2015. His research interests include networked sensing and control, cyber-physical systems, and robust control. He is a member of the IEEE.



Jianping Pan (S'96–M'98–SM'08) received the bachelor's and PhD degrees in computer science from Southeast University, Nanjing, China, and he did his postdoctoral research at the University of Waterloo, Waterloo, Canada. He is currently a professor of computer science at the University of Victoria, Victoria, Canada. He also worked at the Fujitsu Labs and NTT Labs. His area of specialization is computer networks and distributed systems, and his current research interests include protocols for advanced networking, performance analysis of networked systems, and applied network security. He has been serving on the technical program committees of major computer communications and networking conferences including IEEE INFOCOM, ICC, Globecom, WCNC, and CCNC. He is the Ad Hoc and sensor networking symposium co-chair of IEEE Globecom 2012 and an associate editor of the *IEEE Transactions on Vehicular Technology*. He is a senior member of the ACM and IEEE.