Contact-Aware Data Replication in Roadside Unit Aided Vehicular Delay Tolerant Networks

Yong Li, *Member, IEEE*, Depeng Jin, *Member, IEEE*,
Pan Hui, *Senior Member, IEEE*, and Sheng Chen, *Fellow, IEEE*

Abstract—Roadside units (RSUs), which enable vehicles-to-infrastructure communications, are deployed along roadsides to handle the ever-growing communication demands caused by explosive increase of vehicular traffics. How to efficiently utilize them to enhance the vehicular delay tolerant network (VDTN) performance are the important problems in designing RSU-aided VDTNs. In this work, we implement an extensive experiment involving tens of thousands of operational vehicles in Beijing city. Based on this newly collected *Beijing* trace and the existing *Shanghai* trace, we obtain some invariant properties for communication contacts of large scale RSU-aided VDTNs. Specifically, we find that the contact time between RSUs and vehicles obeys an exponential distribution, while the contact rate between them follows a Poisson distribution. According to these observations, we investigate the problem of communication contact-aware mobile data replication for RSU-aided VDTNs by considering the mobile data dissemination system that transmits data from the Internet to vehicles via RSUs through opportunistic communications. In particular, we formulate the communication contact-aware RSU-aided vehicular mobile data dissemination problem as an optimization problem with realistic VDTN settings, and we provide an efficient heuristic solution for this NP-hard problem. By carrying out extensive simulation using realistic vehicular traces, we demonstrate the effectiveness of our proposed heuristic contact-aware data replication scheme, in comparison with the optimal solution and other existing schemes.

Index Terms—Mobile date dissemination, vehicular delay tolerant networks, communication contact, data replication

1 Introduction

1.1 Background

Nowadays, as more and more vehicles are equipped with devices to provide wireless communication capability, interests on vehicular communications and networks have grown significantly [1]. Newly emerged vehicular communication networks are seen as a key technology for improving road safety and building intelligent transportation system (ITS) [2]. Many applications of vehicular networks are also emerging, including automatic collision warning, remote vehicle diagnostics, emergency management and assistance for safe driving, vehicle tracking, automobile high speed Internet access, and multimedia content sharing [1]. In USA, Federal Communications Commission has allocated 75 MHz of spectrum for dedicated short-range communications in vehicular networks [3], and IEEE is also working on related standard specifications. Many consortia and standardization bodies are actively developing technologies and protocols for information transmission between vehicles and roadside unit (RSU) infrastructure equipments, known as vehicles to infrastructures (V2I), as well as between vehicles, known as vehicles to vehicles (V2V) [4]. Although the third generation and forth generation mobile cellular networks with broad coverage and high bandwidth are able to provide multimedia content downloading services for the moving vehicles, with the increase of the services and user demands, cellular networks will very likely be overloaded and congested in the near future. Especially during peak time and in urban central areas, cellular-based vehicular communications will face extreme performance hits in terms of low network bandwidth, missed calls, and unreliable coverage [12]. In terms of the data in the mobile content sharing, some data items are popular and needed by a large amount of users [30]. Thus, benefiting from the common interests among the users and properties of free of cost, efficient utilization of the spare capacity of local sharing links, V2I and V2V communications becomes more inevitable for the application of mobile content dissemination.

In vehicular networks, V2I communications between vehicles and RSUs play an important role in enhancing the networking performance since V2V communications are inherently opportunistic and stochastic [5]. In RSU-aided vehicular networks, RSUs deployed along the roadside offer more 'steady' connections for data transmission [6]. In such kind of vehicular networks, data traffic demands initiated from vehicles are random and bursty by nature, and RSUs act as gateways to the Internet and to other infrastructure systems, such as ITS. Vehicles transmit their access requests and information to RSUs, and RSUs then send responses to

- *Y. Li and D. Jin are with the State Key Laboratory on Microwave and Digital Communications, Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. E-mail: {liyong07, jindp}@tsinghua.edu.cn.*
- *P. Hui is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China, the Telekom Innovation Laboratories, Berlin, Germany, and Aalto University, Helsinki, Finland. E-mail: panhui@cse.ust.hk.*
- *S. Chen is with the Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, United Kingdom, and the King Abdulaziz University, Jeddah 21589, Saudi Arabia. E-mail: sqc@ecs.soton.ac.uk.*

Manuscript received 9 Dec. 2013; revised 10 Feb. 2015; accepted 24 Feb. 2015. Date of publication 24 Mar. 2015; date of current version 4 Jan. 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.2015.2416185

the Internet for querying the data and information on the behalf of vehicles [5]. Usually, the communication coverage of an RSU is limited in the case of utilizing the range-limited wireless technologies like WiFi or millimetre wave communications. On the other hand, vehicles are highly mobile and sometimes sparse by nature. Therefore, it is difficult to maintain a connected network between the RSUs and vehicles to communicate. Opportunistic contacts between RSUs and vehicles, however, are capable of providing high-bandwidth communication capacity for data transmission, which forms the basis of RSU-aided vehicular delay tolerant networks (VDTNs) [7], [8], [9], [10]. Within a VDTN, a RSU will store the data in its buffer, and forwards them to an appropriate vehicle when a transmission opportunity is available along this vehicle's movement, which is referred to as a communication contact [11]. By exploiting the delay-tolerant nature of non-realtime applications, the service providers can delay and even shift large amount of the data transmissions to VDTN. Compared with the broadly covered cellular networks, although this RSU-based VDTN approach may bring extra deployment cost and induce tolerable delay for the data dissemination, it helps to cope with the explosive traffic demands and mobile data growth with the limited cellular network capacity.

1.2 Challenging and Related Work

Many challenging and open problems exist in designing RSU-aided VDTNs [12], and currently many consortia and standardization bodies are actively developing technologies and protocols for efficient data transmission in VDTNs [2], [12]. Recent works have focused on how to deploy RSU infrastructure to handle the growing communication demands as the number of vehicles increases, and have proposed optimal RSU placement schemes with the consideration of the vehicular traffic and city structures [5], [12], [13], [14], [15]. With an optimal RSU deployment, which dramatically enhances the VDTN's performance in terms of data transmission delay and ratio [12], one of the major remaining problems is how to efficiently utilize the RSUs to improve the data dissemination performance. In vehicular sensor networks, existing works [16], [17], [18], [19], [20] investigate the schemes of data replication using RSUs. For example, Ref. [16] identified a set of design choices of content-addressed storage and mobility-assist storage to utilize the resources of RSUs, while Ref. [17] proposed multi-hop data replication schemes to deal with the opportunistic mobility. However, these works do not take the mobility patterns of the vehicular with the RSU into the consideration of data replication design. In a VDTN, data dissemination efficiency depends on how the RSUs replicate the mobile data and, furthermore, the vehicular mobility critically influences the opportunistic data transmission. Therefore, how the mobile data are replicated to the targeted RSUs by considering the vehicular mobility and data requirements as well as the RSUs' data storage policy is a critically important problem to be solved. For high-speed vehicles, the wireless link from a vehicle to a RSU is highly dynamic and subject to opportunistic contact, which makes the decision for the data replication among the deployed RSUs extremely challenging [1]. Therefore, a fundamental understanding of the underlaying patterns of the network dynamics existing in the vehicles to RSUs communications is essential to solve this challenging problem.

Thus, the mobile data dissemination performance of an RSU-aided VDTN depends on how often the opportunistic communication contacts between the vehicles and RSUs occur and how long these contacts last, which are referred to as contact rate and contact time, respectively. The contact rate defines how many times the vehicles can communicate with a RSU within a certain time window, which is related to the inter-contact time or contact interval. The contact time, also known as the contact duration, is another important factor that directly affects the amount of data that can be transferred between vehicles and RSUs when they can communicate. Therefore, contact interval and contact duration influence the throughput and capacity of RSU-aided VDTNs. Due to lack of realistic models for opportunistic contacts between RSUs and vehicles, the current proposed data forwarding protocols can only be investigated based on the synthetic mobility model [21], which may lead to the misleading results that are unachievable in realistic vehicular environments. Therefore, it is critical to design efficient data replication schemes based on an accurate understanding of the communication contacts for large-scale urban vehicular mobility environments.

1.3 Contribution Summary

In this paper, we investigate contact-aware data replication for RSU-aid VDTNs by considering the application of mobile data dissemination. More specifically, we study the problem of how the system replicates mobile data to the deployed RSUs to enhance the mobile data sharing and dissemination efficiency. In order to solve this problem, we first modeling the patterns of opportunistic communication contacts between vehicles and RSUs based on two large-scale urban vehicular mobility traces, and we then propose an efficient data replication scheme for the system to replicate mobile data. Our novel contribution is threefold which are summarized as follows.

- We collect real mobility traces from about 27,000 operational taxis for one month in Beijing city, which records the mobility contact patterns between the RSUs and vehicles in a large city. By analyzing the large volume of realistic urban vehicular mobility traces to investigate contact rates, we find that the distribution of inter-contact time is exponential, which reveals that the contact rate between the vehicles and RSUs exhibits strong Poisson property. In terms of the contact time, we find that it obeys an exponential distribution with high accuracy. This is contrast to the existing findings of a power law exhibited in human mobility.
- We formulate the communication contact-aware data replication into an optimization problem with the realistic VDTN settings that 1) the network contains heterogeneous vehicles in terms of data preference, 2) the mobile data items are multi-types of different size and delay sensitivity, and are erasure coded [22] to increase the dissemination efficiency, and 3) the RSUs' storages for content sharing are limited in size. These realistic conditions were not taken

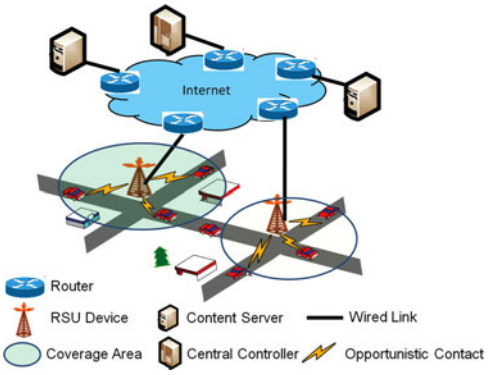


Fig. 1. Illustration of the mobile data offloading system integrating cellular network and opportunistic communications.

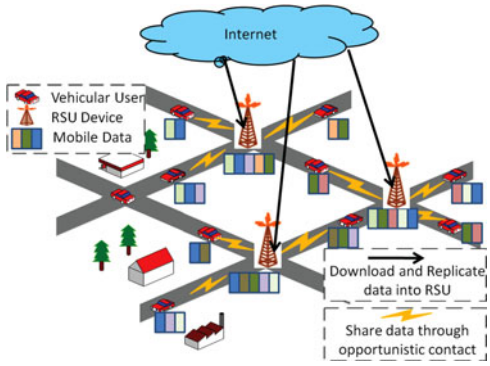


Fig. 2. Mobile data offloading in RSU-aided VDTNs.

into account in the previous works [6], [10] for simplicity reasons.

- We provide an efficient solution for this NP-hard problem of optimal contact-aware data replication by proposing a heuristic algorithm to replicate the mobile data to RSUs' buffers. By reformulating the problem, we also propose a benchmark algorithm that attains the optimal system performance, albeit at the cost of high computational complexity. Through extensive realistic trace-driven simulations, we demonstrate that our heuristic algorithm achieves excellent system performance in comparison to the optimal and several other existing schemes.

1.4 Paper Structure

The rest of the paper is organized as follows. Section 2 describes the communication contact-aware mobile data dissemination system and presents the associated optimization problem. In Section 3, we model the communication contact in terms of contact interval and contact duration. Based on the obtained communication contact model, we specify the optimization problem and design a heuristic algorithm to solve it in Section 4. In Section 5, we re-formulate the problem in order to obtain an optimal solution as the benchmark. Section 6 introduces the experimental environment for performance evaluation and provides extensive simulation results. Our conclusions are drawn in Section 7.

2 System Description and Problem Statement

2.1 System Overview

The network topology is shown in Fig. 1, where vehicles travel around the city roads, and RSUs are deployed, each providing coverage over a certain area. Since current optimal placement algorithms usually place the RSUs in the intersections of main roads [5], in our system, we assume that the RSUs are placed at the intersections. The RSUs are connected to the content servers in Internet through wired links. Vehicular stations requiring mobile data send their data requests to the corresponding content servers via the RSUs. Then, the requested data are proactively delivered from the corresponding content servers to some chosen or targeted RSUs via the wired links under the guidance of the data replication policy. Since usually these connected wired links provide relatively large bandwidth, the data could be delivered successfully before the delay-tolerant dissemination between the RSUs and vehicles. RSUs will further disseminate the mobile contents to the vehicular users that request the data through opportunistic communication which occurs when the vehicles move into the communication coverage of the RSU. After the vehicle stations receive the requested data, as usual, they will send acknowledgements to the content servers via the RSU.

A central controller is deployed for this integrated mobile data dissemination system, and it can communicate with the RSUs and content servers. The central controller is tasked to make data replication decisions for the content servers to distribute the data into the buffers of RSUs based on the vehicular mobility patterns and mobile data demands. The mobility patterns are related to the communication contact patterns between the vehicles and RSUs. Two important metrics are the average contact interval, which is related to how often a vehicle will enter into a RSU's coverage area, and the average contact duration, which is the average time that the vehicle will stay in the coverage area. Many vehicles travel on predetermined routes and schedules. Examples include city buses and people traveling by cars to and from work. Therefore, daily mobility patterns may exhibit certain regularity, and communication contact interval and duration are often quasi-statistic. Thus, the information of contact patterns can often be obtained by the central controller in advance with high accuracy. Moreover, new contact information can be collected regularly from vehicles. On the other hand, the content demands are sent by the vehicles to the central controller directly. Therefore, the central controller has the necessary information to make the data replication decision.

2.2 Data Replication and Networking Modelling

In our RSU-aided mobile data dissemination system depicted in Fig. 2, the content servers first replicate the mobile data to the RSUs, and these RSUs further disseminate the data to the vehicles that are interested in them by vehicle-to-RSU opportunistic communication. Specifically, in order to increase the data dissemination efficiency, mobile data items are encoded into dissemination packets by erasure coding [22]. Then, the encoded packets are replicated to the RSUs where they are buffered in the RSUs' local storages. Finally, when the vehicles meet the RSUs, they receive their requested data. We consider realistic multiple data dissemination, where the system disseminates multiple data items,

and a RSU need to store multiple data packets of different data items, depending on its buffer size, Furthermore, a vehicular user may also be interested in different data items.

In this RSU-aided mobile data dissemination system, there are $R+U$ nodes. More specifically, there are R RSUs buffering the data and U vehicular users requesting the data, respectively. We use $\mathcal{R}$ and $\mathcal{U}$ to denote the sets of RSUs and vehicles, respectively, where $|\mathcal{R}|=R$ and $|\mathcal{U}|=U$. We also refer to a vehicle as a vehicular user or simply a user throughout the discussion. For the RSUs, the system requires their storages to buffer the mobile data, which may include multimedia content of very large size, such as movies. Even though RSU devices may have large storage, it is impossible to ask a RSU to contribute all its storage solely for the data dissemination purpose. Therefore, we should take the storage that each RSU is willing to share as one of our constraints, which directly influences the number of data items that can be stored. Considering this realistic condition, we assume that RSU r, $r \in \mathcal{R}$, can at most buffer L_r size of data items.

2.3 Mobile Data Modelling

Since there are many different types of mobile data, for example, multimedia newspapers, weather forecasts, movie trailers, etc., we model the mobile traffics of C different data items, labelled as $\mathcal{C}$. For any $c \in \mathcal{C}$, its data length is l_c, and its lifetime is T_c, which means that all the RSUs will stop disseminating the mobile data c after the deadline T_c.

Because the contact duration between a vehicle and a RSU is usually very limited in VDTNs, a complete mobile data may not be transmitted in one contact. The most straightforward way is using simple fragmentation, but it will result in the coupon collector's problem [23] that will significantly decrease the data recovery efficiency. To mitigate this problem, we adopt the erasure coding technique [22], [24] to encode the mobile data into a large set of small coded packets, and any sufficient subset of the coded packets can be used to reconstruct the mobile data. Specifically, the coding process takes the original data c of size l_c and a given coding rate as the input, and outputs the packets of size g_c. Any $v_c=(1+\epsilon)l_c/g_c$ coded packets can be used to reconstruct the mobile data c back, where ϵ is a small constant determined by the exact erasure coding algorithm employed [24]. In general, $g_c < l_c$. Occasionally, a mobile data c may have a very large size, and in this case $g_c \ll l_c$. The size g_c is data dependent as different data c may be encoded with different packet sizes.

2.4 Interests Modeling

We now characterize the user behaviors of accessing to different data items in the mobile data dissemination system. In a system with multiple data items, a user will have different interests in different data items, and different users will have different dynamic accessing behaviors. Moreover, some data items are popular data that are interested by many users, while some other data items are not popular data which may only be interesting to a small number of users. We describe the user's interests to different mobile data by a subscriber profile, and model the popularity of mobile data by an interest distribution on keywords, which is a widely used approach to model user interest distribution on different data items in diverse applications [25]. Specifically, for all the mobile data, the system have K keywords, denoted by the set $\mathcal{K}$, to describe them. Any data item $c \in \mathcal{C}$ is described by a subset of keywords, denoted by $\mathcal{K}_c \subseteq \mathcal{K}$, and weight ϱ_{k_c} which indicates the importance of keyword $k_c \in \mathcal{K}_c$. In this way, we can define the popularity of mobile data items. Without loss of generality, we assume $\sum_{k_c \in \mathcal{K}_c} \varrho_{k_c}=1$. To model the interests of different subscribers on different data, we define P_u^k as the degree of how user $u \in \mathcal{U}$ is interested in keyword $k \in \mathcal{K}$. In this way, we can compare the interests of user u to two different keywords $k_1, k_2 \in \mathcal{K}$ by $P_u^{k_1}$ and $P_u^{k_2}$. Thus, the interest profile of user u is defined by the set $\mathcal{P}_u=\{P_u^k : k \in \mathcal{K}\}$. Without loss of generality, we assume $\sum_{k \in \mathcal{K}} P_u^k = 1$. The interest probability of user $u \in \mathcal{U}$ in mobile data $c \in \mathcal{C}$, defined by $w_{u,c}$, can then be obtained as

$$w_{u,c} = \sum_{k_c \in \mathcal{K}_c} \varrho_{k_c} P_u^{k_c}. \tag{1}$$

For any user $u \in \mathcal{U}$, we further let $w_u=[w_{u,1} w_{u,2} \cdots w_{u,C}]$ represent its 'affection' to all the data.

2.5 Communication Contact

A RSU can communicate with a vehicle to disseminate data only when the vehicle moves into its communication coverage, which is referred to as *communication contact*. During the communication contact, RSU can transmit the mobile data in the rate of η bytes per second to vehicles. Consider generic RSU and vehicle, denoted by r and u, respectively. The vehicle moves according to its mobility trajectory that covers the region $\Omega \subset \mathbb{R}^2$, and the RSU is deployed in a fixed position. Let $\chi_r \in \Omega$ and $\chi_u(t) \in \Omega$ be the positions of RSU r and vehicle u at time t, respectively, where t is in continuous-time scale. We assume that the RSU has a circular coverage with the radius of $R_{\mathrm{T}x}$. When the vehicle is within the communication range $R_{\mathrm{T}x}$ with the RSU, communication may occur. In a practical system, since usually the RSUs will buffer different mobile content, we need to consider the contact interval by telling the difference of each RSU. Thus, we define the contact interval of each RSU-vehicle pair, and then obtain the contact rate of each pairs. With the above consideration and notations, the contact interval and contact duration of RSU r and vehicle u can formally be defined as follows.

Definition 1 (Contact Interval). *The contact interval of RSU r and vehicle u, denoted as $CI_{r,u}$, is defined as the time interval that takes the vehicle to come within the coverage of the RSU again from the last time, denoted as t_0, when the vehicle was moving out of the coverage, that is,*

$$CI_{r,u} = \min_t \big\{ (t-t_0) : \|\chi_r - \chi_u(t)\| \le R_{\mathrm{T}x},\, t > t_0 \big\}.$$

Definition 2 (Contact Duration). *Assume that user u comes within the communication coverage of RSU r at time t_c, that is, $\|\chi_r - \chi_u(t_c^-)\| > R_{\mathrm{T}x}$ and $\|\chi_r - \chi_u(t_c)\| = R_{\mathrm{T}x}$, where t_c^- denotes the time before t_c. The contact duration of r and u is defined as the time during which the user is in contact with the RSU before moving out of its coverage, that is,*

TABLE 1
List of Commonly Used Notations and Variables

Notation/Variable	*Description*
R, U, C	The numbers of RSUs, vehicles and data items in the system, respectively.
$\mathcal{R}, \mathcal{U}, \mathcal{C}$	The sets of RSUs, vehicles and data items, respectively.
r, u, c	The indexes for RSUs, vehicles and data items, respectively.
l_c, T_c	The size and lifetime of data item $c \in \mathcal{C}$, respectively.
g_c, v_c	The size of a coded packet, and the number of the coded packets that can reconstruct data $c \in \mathcal{C}$.
L_r	The buffer size of RSU $r \in \mathcal{R}$.
$\theta_{u,r}$, $\lambda_{u,r}$	The contact rate and contact duration parameters between RSU $r \in \mathcal{R}$ and vehicle $u \in \mathcal{U}$.
η	The transmission rate between a RUS and a vehicle during communication contact.
$w_{u,c}$	The interest probability of vehicle $u \in \mathcal{U}$ in mobile data $c \in \mathcal{C}$.
$\mathbf{X} = (x_{r,c})$	The data replication policy of the system.

$$CT_{r,u} = t - t_c \text{ with } \min_{t-t_c}\{t : \|\chi_r - \chi_u(t)\| > R_{\mathrm{T}x}\},$$

where both t and t_c are in the continuous-time scale.

The contact interval and duration directly influence the throughput and capacity of the RSU-aided mobile data dissemination system. Higher contact rate and longer contact time usually result in higher network throughput and larger capacity. Clearly, how to replicate the mobile data to the RSUs depends on the underlying system's communication contact properties. Thus, revealing fundamental laws and properties of these communication contacts can greatly benefit the data replication problem analysis and provides important hints regarding the optimal system solution. After formulating the optimal mobile data replication problem in the RSU-aided VDTN, we will turn to modeling the communication contact with the aid of two large-scale realistic urban vehicular mobility traces. The result of this modeling will help us to solve the challenging mobile data replication problem.

2.6 Problem Formulation

For the mobile data dissemination system with R RSUs, U vehicles and C mobile data items, the system optimization goal is to maximise the expected interests satisfaction of all the users in this RSU-aided VDTN, which depends on the mobile data replication policy. Denote $\mathbf{X} = (x_{r,c})$, where $r \in \mathcal{R}$ and $c \in \mathcal{C}$, as the data replication policy, in which $x_{r,c} \in \{0, 1, \ldots, v_c\}$, and $x_{r,c} = a$, $a \neq 0$, indicates that RSU r stores a coded packets of date c in its buffer, while $x_{r,c} = 0$ indicates that r does not store any packet of data item c. Since a lifetime T_c is assigned to each data item c, if a user does not receive the required item from RSUs after the lifetime is expired, the data dissemination to this particular user fails or the mobile data dissemination system does not meet this user's interest. On the other hand, in the system with limited RSU buffer and limited communication contact, some of users will not be guaranteed to obtain the mobile data before its deadline, and it is impossible to meet all the users interests. Therefore, we set the objective as maximizing the expected overall interests satisfaction of all the users. Since this objective function depends on the data replication policy $\mathbf{X}$, it is denoted as $J(\mathbf{X})$.

Maximising the system's expected interests satisfaction for all the users and over all the mobile data items can be specified as the following optimisation problem

$$\begin{aligned} \max \quad & J(\mathbf{X}) \\ \text{s.t.} \quad & x_{r,c} \in \{0, 1, \ldots, v_c\}, \forall r \in \mathcal{R}, c \in \mathcal{C}, x_{r,c} \in \mathbf{X}, \\ & \text{and } \sum_{c \in \mathcal{C}} g_c x_{r,c} \leq L_r, \forall r \in \mathcal{R}, \end{aligned} \tag{2}$$

where $\sum_{c \in \mathcal{C}} g_c x_{r,c} \leq L_r$ is the buffer size constraint of RSU r. For the ease of reference, we now summarize the commonly used notations and variables throughout the paper in Table 1.

3 Modeling the Communication Contact

Although the model for the opportunistic contacts between vehicles is already proposed by Ref. [11], the contact patterns of the contacts between the vehicle and RSU is still unknown, and it is an open problem to model the opportunistic communication contacts between the vehicle and RSU. Thus, in this Section, we investigate it based on two large-scale urban vehicular mobility traces of *Beijing* and *Shanghai*. Specifically, we use the taxis mobility traces to study and model the contact patterns between the vehicles and RSU, which should be different from that of human mobility like walking.

3.1 Description of the Two Data Sets

Shanghai trace was collected by SG project [26], where 2,019 operational taxis continuously covered one month of February 2007 without any interruptions in Shanghai city. In this trace, a vehicle sends its position report by GPRS to the central database every 1 minute when it has passengers on-board and every 15 seconds when it is vacant for the reason of real-time scheduling. The different intervals of reporting however may distort the records of the physical movements of the vehicles, since most of vehicles are not vacant most of the time. Even though *Shanghai* trace is large among the existing vehicular mobility traces, the number of the vehicles involved still may not be sufficient to record the statistical features of contact interval and duration in a high-speed large urban environment.

In collecting *Beijing* trace, we used the mobility track logs obtained from 27,000 participating Beijing vehicles carrying GPS receivers during the whole May month in 2010. Most of the vehicles involved are taxis. The reason for us to utilize taxis as vehicular devices is that most taxis in Beijing already have GPS systems installed for service monitoring, and

collect their traces will not induce privacy and cost problems. Although taxis mobility may be different from other vehicles due to different driver behaviors, taxis are more sensitive to urban environments in terms of underlying road topology, traffic control and urban planning, and they have broader coverage in terms of space and operation time than buses and private cars. Thus, using the taxi to reveal the contact patterns and study the data dissemination problem is meaningful. Specifically, we utilized the GPS devices to collect the vehicles locations and timestamps and GPRS modules to report the records every 15 seconds for moving vehicles. The specific information contained in such a report includes: the vehicle's ID, the longitude and latitude coordinates of the vehicle's location, timestamps, instant speed and heading. *Beijing* trace is the largest vehicular data trace available.

3.2 Empirical Data Processing

To obtain an accurate contact interval and contact duration, we need to know the exact beginning and ending times of each opportunistic contact. However, GPS reports were collected in discrete time and they may be collected at different time intervals as in the case of *Shanghai* trace. Therefore, the times that an opportunistic contact starts and ends, respectively, may not be recorded in timestamps. Consequently, the traces need to be processed in order to extract the contact duration. In extracting the contact from a GPS trace, we assume that a vehicle is able to communicate with a RSU if their estimated locations are within the given communication range and within the same specific time duration, which is called a contact. In reality even if a vehicle is within a RSU's coverage range, they may not be able to successfully transfer data due to physical layer signal issues and MAC layer association time. However, vehicular communication protocols like IEEE 802.11p [36] aim to make the new standard more robust in the real-world vehicular scenarios [40]. For example, in IEEE 802.11p, the vehicles are allowed to transmit and receive data frames with the wildcard setting, which means the vehicles can immediately communicate with the RSU upon encounter without any additional overhead [40], [41]. Consequently, since we focus on the distribution of contact, we only investigate the potential communication opportunities between vehicles and RSUs, and we leave the issues related to how to ensure successful data transmissions during contacts as future work.

3.2.1 Location Adjustment

By collecting the GPS information of longitude and latitude coordinates, we obtain the vehicles' moving traces that indicate the vehicles' locations varying with time. Since these locations are measured by GPS devices, the data may be corrupted by noise due to the inaccuracy of GPS devices. Furthermore, the vehicles may not all report their location at the same time slots with the same fixed frequency, as in the case of *Shanghai* trace. Therefore, we need to process the data trace to obtain accurate locations of all the vehicles in the same time slots and with the same frequency, and also deal with the special taxis behavior of stopping at a place waiting for passages, which would influence the contact patterns when the locations of taxis just happen to be close to a RSU. In order to achieve these goals, we first use the city maps of Shanghai and Beijing for the respective traces to delete the traces of stopping taxis of both occupied and unoccupied, which both records the contact patterns between the vehicular mobility the RSU, by detecting their locations that are not in the regions roads, and then also correct all the moving vehicles' locations so that they are in the regions of related city roads. Then, we insert location points in the turning place based on the information of map to make sure any two sequenced location reports are in the same linear road. Finally, we use the method of linear interpolation (LI) to insert location points so that all the vehicles have location information at every 15-second interval. To illustrate how this LI method works, consider that we have the location information of a vehicle in the original trace with the locations $l_1, l_2, \ldots, l_n$ recorded at the time points $t_1 < t_2 < \cdots < t_n$, and we want to insert the location information l_t at the time point t which is calculated according to the 15-second frequency. We just need to find t_m that satisfies $t_m \leq t < t_{m+1}$, and then estimate the location l_t by the following LI

$$l_t = \frac{t_{m+1} - t}{t_{m+1} - t_m} \cdot l_m + \frac{t - t_m}{t_{m+1} - t_m} \cdot l_{m+1}.$$

In order to verify that this data preprocessing does not introduce artificial and inaccurate information into the original data trace, we use the data obtained by this preprocessing method for the one-day vehicles' locations to plot the trajectories of all the vehicles, and it can be concluded that the data sets are sufficiently large and even only using one-day data we can recover the whole city maps. In order to further demonstrate the accuracy of our data preprocessing, we compare the recovered maps of Beijing and Shanghai with the true Beijing and Shanghai maps. It can be verified that all the vehicles' trajectories determined by the preprocessing are in the related city roads, and the two city maps drawn by these one-day trajectories are very similar to the corresponding true city maps.

3.2.2 Contact Extraction

To extract communication contacts from the GPS traces, the work [11] uses a time window to measure the contact. However, as reported in [11], it is difficult to select a suitable time window, since a large time window may result in false contacts while a small window may miss some real communication opportunities. To overcome this difficulty, in this work, we also use the LI method to extract the opportunistic contact, which is illustrated in Fig. 3. Firstly, we select the time points near and inside the chosen communication contact, which should at least include one time point during the communication contact, one point before the contact and one after the contact. Then, we use the selected time points to estimate the communication contact beginning and ending times. In Fig. 3, we have selected t_1, t_2 and t_3 as the three such time points recorded in the trace. The communication contact's beginning time t_c and ending time t_d can be estimated by the LI expressed as follows

$$t_c = t_1 + \frac{\bar{\chi}(t_1) - R_{\mathrm{Tx}}}{\bar{\chi}(t_1) - \bar{\chi}(t_2)}(t_2 - t_1),$$

$$t_d = t_2 + \frac{\bar{\chi}(t_3) - R_{\mathrm{Tx}}}{\bar{\chi}(t_3) - \bar{\chi}(t_2)}(t_3 - t_2),$$

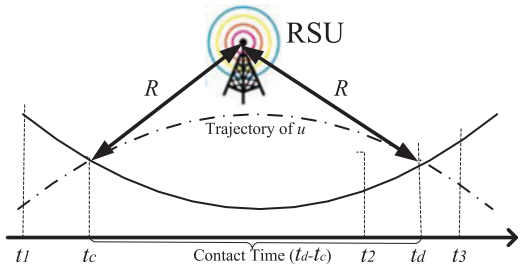


Fig. 3. Extract communication contacts from the GPS reports of RSU r and vehicle u. Two arc lines are the trajectory of u and the coverage boundary of r, and R is the communication range. t_1, t_2 and t_3 are three timestamps in the GPS records, and the positions of the vehicles at t_1, t_2 and t_3 are reported by GPS. t_c is the time that the communication contact begins, and t_d is the end of the communication contact. Since both t_c and t_d are not recorded, they have to be estimated to obtain the contact interval and contact duration.

with $\bar{\chi}(t_k) = \|\chi_r - \chi_u(t_k)\|$. Consequently, the contact duration can be obtained by $t_d - t_c$. Then, the contact interval can be obtained as the time interval of two successive communication contacts.

We note that the LI method is based on the approximate assumption that the distance between RSU and vehicle changes linearly closed to and during the communication contact. Since in real urban vehicular environments, a contact occurs often when the vehicle moves nearby the RSU, e.g., the vehicle is passing the intersection where the road is usually straight. Therefore, an LI is sufficient for us to extract the communication contact. In order to obtain more accurate results, we use several different sets of times points and three different communication ranges in the LI based estimation. More specifically, in the communication contact extraction, we choose three sets of time points near and within the contact concerned to obtain three LI estimates, and use the average value as the final contact time and inter-contact time. We also consider four different coverage ranges of 50, 100, 150 and 200 meters for both *Shanghai* and *Beijing* traces.

3.2.3 Acquisition of Results

To pre-process the empirical data and to extract the communication contacts, we place the RSUs in the intersections of the main roads, extract the contacts from the trace and investigate the distributions of the contact duration and contact interval. Specifically, we study the complementary cumulative distribution functions (CCDFs) of the contact interval and duration. In particular, we focus on the aggregate CCDF of all the vehicles and all the deployed RSUs, which is the CCDF per contact sample over all the distinct pairs of RSU and vehicle.

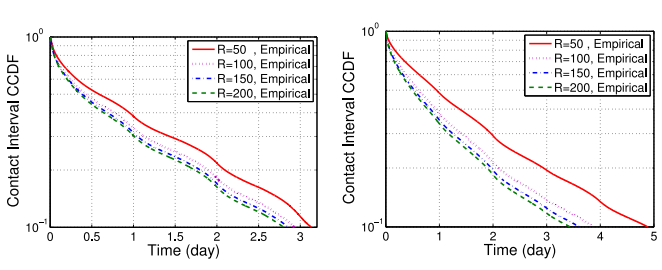


Fig. 4. Empirical contact interval patterns for *Beijing* and *Shanghai* traces in the linear-log coordinate: (a) Beijing, and (b) Shanghai.

TABLE 2
The Adjusted R-Square Statistics of the Exponential Fittings to the Empirical CCDFs of Contact Interval Obtained from *Beijing* and *Shanghai* Traces

Transmission range R	50 m	100 m	150 m	200 m
Beijing trace	98.36%	97.37%	96.90%	96.59%
Shanghai trace	98.61%	97.24%	96.79%	96.55%

3.3 Modelling Result of Contact Interval

The aggregated empirical distributions of contact interval between RSUs and vehicles are examined, and the obtained empirical CCDFs of contact interval during entire trace collection time are plotted in Figs. 4a and 4b for *Beijing* and *Shanghai* traces, respectively, given four different coverage ranges of $R = 50$, 100, 150 and 200 m. Here, we display 90 percent of the distribution in the linear-log scale. By observing the empirical distribution curves of Fig. 4, we find that they are close to straight lines. Since we plot the distribution in the linear-log scale, these empirical curves showing in Fig. 4 clearly offer the insight that the contact interval distribution exhibits exponential property.

In order to verify the accuracy of the exponential distribution model for the contact interval of urban vehicular mobility, we fit the exponential curves to all the 100 percent empirical CCDFs of contact interval obtained for both *Beijing* and *Shanghai* traces with the four different coverage ranges. The goodness of fit is measured quantitatively by the R-square statistics [27], which is defined as the percentage of the variation between the empirical CCDF and the fitted distribution. The results are summarised in Table 2, where the adjusted R-square statistics are computed with Matlab Curve Fitting Toolbox. It can be seen from Table 2 that the average adjusted R-square statistics are all over 96 percent for both *Shanghai* and *Beijing* traces. This confirms the accuracy of the exponential distribution model for contact interval. In Fig. 5, we compare the 100 percent empirical CCDFs with the corresponding exponential fittings for both *Beijing* and *Shanghai* traces, given the coverage ranges of $R = 50$ and 200 m.

3.4 Modelling Results of Contact Duration

We next examine the aggregated empirical distributions of contact duration between RSUs and vehicles extracted from the two vehicular mobility traces. The 90 percent empirical CCDFs of contact duration during entire trace collection time are displayed with the linear-log scale in Figs. 6a

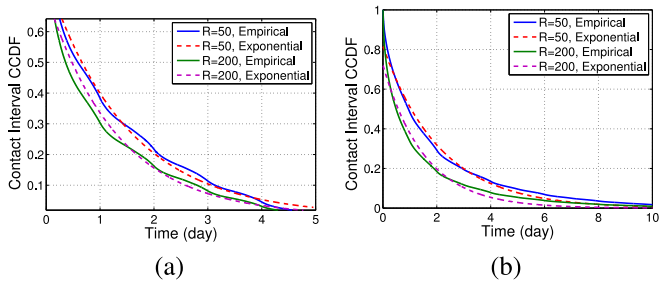


Fig. 5. Empirical contact interval patterns and their exponential fittings for *Beijing* and *Shanghai* traces in the linear-linear coordinate: (a) Beijing, and (b) Shanghai.

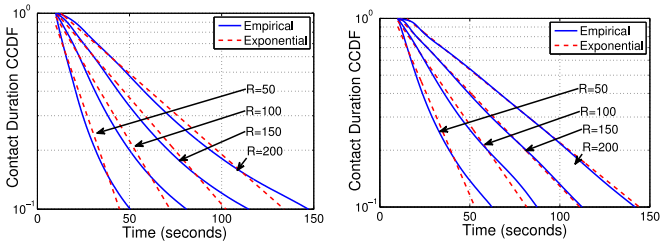


Fig. 6. Empirical contact duration patterns and their exponential fittings for *Beijing* and *Shanghai* traces in linear-log coordinate: (a) Beijing, and (b) Shanghai.

and 6b for *Beijing* and *Shanghai* traces, respectively, again given four coverage ranges of $R = 50$, 100, 150 and 200 m, where the exponential fittings to the corresponding empirical CCDF curves are also shown for comparison. The 100 percent empirical CCDF curves and their corresponding exponential fittings are depicted in Fig. 7 with the linear-linear scale. The results of Figs. 6 and 7 qualitatively show that the contact duration exhibits a very clear exponential distribution. To quantitatively measure the accuracy of the exponential model for the contact duration of RSU-aided VDTNs, Table 3 lists the adjusted R-square statistics of the exponential fittings to the empirical data of contact duration. It can be seen from Table 3 that the average adjusted R-square statistics are all over 97 percent for all the exponential fittings. This confirms the accuracy of the exponential model of contact duration.

4 Problem Specification and Heuristic Solution

Based on the results of modeling *Beijing* and *Shanghai* vehicular traces presented in the previous section, we can justifiably assume that the communication contact rate between RSU r and vehicle u obeys the Poisson process with contact rate $\theta_{u,r}$ and the duration of their contact follows an exponential distribution with rate parameter $\lambda_{u,r}$.

4.1 Problem Specification

In order to solve the optimization problem (2), we need the explicit expression for the objective function $J(\mathbf{X})$. We note that $J(\mathbf{X})$ depends on the system dynamics of communication contact as well as the data replication policy and buffer size of the RSUs. Recall that our goal is to maximize the expected interest satisfaction. Let us first define $Q_{u,c}$ as the probability that user u has successfully

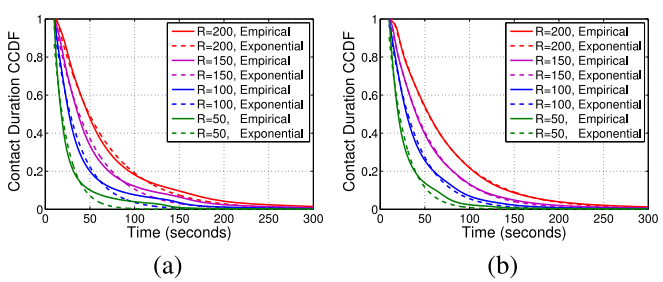


Fig. 7. Empirical contact duration patterns and their exponential fittings for *Beijing* and *Shanghai* traces in linear-linear coordinate: (a) Beijing, and (b) Shanghai.

TABLE 3
The Adjusted R-Square Statistics of the Exponential Fittings to the Empirical CCDFs of Contact Duration Obtained from *Beijing* and *Shanghai* Traces Shown in Fig. 7

Transmission range R	50 m	100 m	150 m	200 m
Beijing	97.83%	98.90%	99.34%	99.53%
Shanghai	97.91%	99.42%	99.92%	99.97%

received mobile data c before its lifetime. Then, we can express the objective function as

$$J(\mathbf{X}) = \sum_{u \in \mathcal{U}} \sum_{c \in \mathcal{C}} Q_{u,c}. \tag{3}$$

In other words, we need to calculate the probability for user u to successfully receive mobile data c before its lifetime.

Note that the condition for user u to recover mobile data c back from the encoded packets is that it has received at least v_c packets before the deadline of T_c. Therefore, we first consider how many packets on average a RSU $r \in \mathcal{R}$ can disseminate to vehicle $u \in \mathcal{U}$. Because u encounters r with the Poisson contact rate of $\theta_{u,r}$ and the contact event is independent on the user interests, we can model the dissemination opportunity as the Poisson process with rate $\theta_{u,r} w_{u,c}$ if r stores data c in its buffer. Consequently, the times that the communication contact between u and r occur for transmitting mobile data c before the deadline T_c, denoted by $N^c_{u,r}$, obeys the following probability function

$$P\left(N^c_{u,r} = j\right) = \frac{e^{-\theta_{u,r} w_{u,c} T_c} \left(\theta_{u,r} w_{u,c} T_c\right)^j}{j!}, \; 0 \le j < +\infty.$$

Consider a generic contact, say m, in the $N^c_{u,r}$ contacts, where $1 \le m \le N^c_{ur}$. Denote its distribution function as $\tau^c_{u,r}(m)$, which follows the exponential distribution of

$$f_{\tau^c_{u,r}(m)}(x) = \lambda_{u,r} e^{-\lambda_{u,r} x} x = \Gamma\left(1, \frac{1}{\lambda_{u,r}}\right),$$

where $\Gamma(1, \frac{1}{\lambda_{u,r}})$ is the Gamma distribution with parameters of 1 and $\frac{1}{\lambda_{u,r}}$. More generally, a random variable X that has the Gamma distribution with parameters $\alpha > 0$ and $\beta > 0$ is denoted by $X \sim \Gamma(\alpha, \beta)$, and its probability density function is given by

$$f_\Gamma(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, \; x > 0,$$

where $\Gamma(\alpha)$ is the ordinary gamma function defined as $\Gamma(\alpha) = \Gamma(\alpha, 0) = \int_0^\infty t^{\alpha-1} e^{-t} dt$. The cumulative distribution function of $X \sim \Gamma(\alpha, \beta)$ can be expressed as

$$F_\Gamma(x; \alpha, \beta) = \int_0^x f_\Gamma(u; \alpha, \beta) du = \frac{\gamma\left(\alpha, \frac{x}{\beta}\right)}{\Gamma(\alpha)},$$

where $\gamma(a, x) = \int_0^x t^{a-1} e^{-t} dt$ is the lower incomplete gamma function.

We now derive the total contact duration between u and r for transmitting mobile data c before its deadline T_c by considering all the N^c_{ur} contacts together. Since they are independent identically distributed, we can express the

total contact duration, denoted by $\mathrm{T}(N_{u,r}^c)$, as follow

$$\mathrm{T}\Big(N_{u,r}^c\Big) = \sum_{m=1}^{N_{u,r}^c} \tau_{u,r}^c(m) = \Gamma\bigg(N_{u,r}^c, \frac{1}{\lambda_{u,r}}\bigg).$$

After obtaining the cumulative contact duration, we turn to how many packets that u is able to receive from r. Denote $\pi_{u,r}^c$ as the number of coded packets for data c that RSU r can disseminate to user u before the lifetime T_c, where $0 \le \pi_{u,r}^c \le x_{r,c}$. For $0 \le i \le x_{r,c} - 1$, we have

$$\begin{aligned} &P\Big(\pi_{u,r}^c = i \Big| N_{u,r}^c = j\Big) \\ &= P\bigg(\frac{ig_c}{\eta} \le T\big(N_{u,r}^c\big) < \frac{(i+1)g_c}{\eta}\Big| N_{u,r}^c = j\bigg) \\ &= F_\Gamma\bigg(\frac{(i+1)g_c}{\eta}; j, \frac{1}{\lambda_{u,r}}\bigg) - F_\Gamma\bigg(\frac{ig_c}{\eta}; j, \frac{1}{\lambda_{u,r}}\bigg) \\ &= \frac{\gamma\Big(j, \frac{\lambda_{u,r}(i+1)g_c}{\eta}\Big) - \gamma\Big(j, \frac{\lambda_{u,r} i g_c}{\eta}\Big)}{\Gamma(j)}. \end{aligned} \tag{4}$$

For $i = x_{r,c}$, we have

$$\begin{aligned} &P\Big(\pi_{u,r}^c = x_{r,c} | N_{u,r}^c = j\Big) \\ &= P\bigg(T\big(N_{u,r}^c\big) \ge \frac{x_{r,c} g_c}{\eta}\Big| N_{u,r}^c = j\bigg) \\ &= 1 - F_\Gamma\bigg(\frac{x_{r,c} g_c}{\eta}; j, \frac{1}{\lambda_{u,r}}\bigg) \\ &= \frac{\Gamma(j) - \gamma\big(j, \frac{\lambda_{u,r} x_{r,c} g_c}{\eta}\big)}{\Gamma(j)}. \end{aligned} \tag{5}$$

From the law of total probability, we have

$$\begin{aligned} P\big(\pi_{u,r}^c = i\big) &= \sum_{j=0}^{+\infty} P\big(N_{u,r}^c = j\big) P\big(\pi_{u,r}^c = i | N_{u,r}^c = j\big) \\ &= \begin{cases} \sum_{j=0}^{+\infty} \frac{e^{-\varpi}\varpi^j}{j!} \frac{\gamma\left(j, \frac{\lambda_{u,r}(i+1)g_c}{\eta}\right) - \gamma\left(j, \frac{\lambda_{u,r} i g_c}{\eta}\right)}{\Gamma(j)}, \\ \qquad\qquad 0 \le i \le x_{r,c} - 1; \\ \sum_{j=0}^{+\infty} \frac{e^{-\varpi}\varpi^j}{j!} \frac{\Gamma(j) - \gamma\left(j, \frac{\lambda_{u,r} v_c g_c}{\eta}\right)}{\Gamma(j)}, \qquad i = x_{r,c}; \end{cases} \end{aligned} \tag{6}$$

where $\varpi = \theta_{u,r} w_{u,c} T_c$.

Let us define the following generating function

$$G_{u,r}^c(y) = \sum_{i=0}^{x_{r,c}} P\big(\pi_{u,r}^c = i\big) y^i.$$

In probability theory, $G_{u,r}^c(y)$ is called the probability generating function of $\pi_{u,r}^c$, where $P(\pi_{u,r}^c = i) = \frac{G_{u,r}^{c\,(i)}(0)}{i!}$ and $G_{u,r}^{c\,(i)}(y)$ is the ith derivative of $G_{u,r}^c(y)$. Define $\pi_u^c = \sum_{r\in\mathcal{R}} \pi_{u,r}^c$ as the total number of packets of data c that user u can receive from all RSUs before the lifetime T_c. The probability generating function of π_u^c can be written as

$$\prod_{r\in\mathcal{R}} G_{u,r}^c(y) = \sum_{i=0}^{n} a_i y^i,$$

in which n is the order of the power series expansion and

$$P\big(\pi_u^c = i\big) = \frac{\Big(\prod_{r\in\mathcal{R}} G_{u,r}^c\Big)^{(i)}(0)}{i!} = a_i. \tag{7}$$

Therefore, the probability for user u to receive the data c, $Q_{u,c}$, can be expressed as

$$Q_{u,c} = 1 - P\big(\pi_u^c < v_c\big) = 1 - \sum_{0\le i<v_c} a_i. \tag{8}$$

Thus, we can transfer the optimization problem (2) into the following form

$$\begin{aligned} \max \quad & J(\mathbf{X}) = \sum_{u\in\mathcal{U}} \sum_{c\in\mathcal{C}} \bigg(1 - \sum_{0\le i<v_c} a_i\bigg) \\ \text{s.t.} \quad & x_{r,c} \in \{0, 1, \ldots, v_c\}, \forall r \in \mathcal{R}, c \in \mathcal{C}, x_{r,c} \in \mathbf{X}, \\ & \text{and} \sum_{c\in\mathcal{C}} g_c x_{r,c} \le L_r, \forall r \in \mathcal{R}. \end{aligned} \tag{9}$$

4.2 Heuristic Algorithm Design

Since our objective function in the optimization problem (9) is the sum of some coefficients and each of these coefficients is the product of derivative of some probability generating function, we cannot derive a closed-form expression for this objective function. Consequently, it is difficult to analyze its properties, and the optimization problem (9) is NP-hard. Therefore, it is hard to design optimal algorithms for our optimal contact-aware data replication problem by solving the optimization problem (9), and furthermore, even an optimal algorithm is designed, its computational complexity may be unacceptable in practice. Thus, we first consider to design an heuristic non-optimal algorithm with an acceptable computational complexity to solve the problem (9). For this purpose, we define the objective function $J(\mathbf{X})$ over the subset of $\mathcal{R} \times \mathcal{C}$. For $\mathbf{A} \subseteq \mathcal{R} \times \mathcal{C}$, we define the data replication policy $\mathbf{X}$ as

$$\mathbf{X} = \mathcal{F}(\mathbf{A}) : \begin{cases} x_{r,c} > 0, & (r,c) \in \mathbf{A}; \\ x_{r,c} = 0, & (r,c) \notin \mathbf{A}. \end{cases}$$

Since $\mathcal{F}(\mathbf{A})$ is bijection, we have the objective function over the subset $\mathbf{A}$ given by $\widetilde{J}(\mathbf{A}) = J(\mathcal{F}(\mathbf{A}))$.

Greedy approach is employed as a heuristic solution to the optimisation problem (9) in the way that data allocation is determined one by one. When one more encoded packet of a data item is stored in a RSU, which is in accordance with the constraints, the objective function is enhanced. The gain of the objective function is generally different for different choices of data item and RSU. We use an allocation pair $\big(r, c|_p\big)$ to denote allocating one packet of data c to RSU r, where $c \in \mathcal{C}$ and $r \in \mathcal{R}$. We also define $\mathbf{G} = \big\{\big(r, c|_p\big) : r \in \mathcal{R}, c \in \mathcal{C}, L_r \ge g_c\big\}$, which is the collection of available allocation pairs satisfying the storage constraint. Note that we assume that the number of packets for each data is sufficiently large so that we do not run out of the packets to be allocated. We also use the sets $\mathbf{A}$ and $\mathbf{B}$ to denote the replication strategies, which are the collections of allocation pairs that we have chosen. There may be multiple identical pairs in $\mathbf{A}$ and $\mathbf{B}$, which indicates that multiple packets of the

same data are replicated to the same RSU. $\widetilde{J}(\mathbf{A})$ gives the objective function value according to allocation strategy $\mathbf{A}$.

As our first greedy strategy, we select the packet and RSU that maximise the gain on the objective function at each stage, that is, select $(r_0, c_0|_p)$ as

$$(r_0, c_0|_p) = \arg \max_{(r,c|_p)\in\mathbf{G}} (\widetilde{J}(\mathbf{A} \cup (r, c|_p)) - \widetilde{J}(\mathbf{A})). \quad (10)$$

The length of each data item packet is also important. This is because a data item packet may offer a large gain but it may also have a large length such that other items' packets cannot be stored. Thus, our second greedy strategy is to calculate the gain per unit data length for each choice of data packet and RSU, and select the pair $(r_0, c_0|_p)$ that maximizes this per-unit-length gain

$$(r_0, c_0|_p) = \arg \max_{(r,c|_p)\in\mathbf{G}} \frac{\widetilde{J}(\mathbf{B} \cup (r, c|_p)) - \widetilde{J}(\mathbf{B})}{g_c}. \quad (11)$$

Note that $\widetilde{J}(\mathbf{A} \cup (r, c|_p)) - \widetilde{J}(\mathbf{A}) = 0$ if the number of total packets of data c allocated in all the RSUs so far is less than v_c, since no user can recover data item c yet. Therefore, in the initialization phase, we allocate the first v_c packets of each data c randomly on the RSUs. Since v_c is much smaller than the number of total packets that can be allocated, this random initialization will hardly influence the performance of our greedy strategies.

Each of the two greedy strategies has its advantage and drawbacks. Thus, a combination of these two strategies is used to enhance the overall performance. In our algorithm, these two strategies are both performed and we choose the better result from the two solutions. We present this heuristic algorithm in Algorithm 1. In Line 1 of the algorithm, we first allocate the first v_c packets of each data c randomly and initialize the allocation sets to this initialisation result. Then, from Line 2 to 9, we choose the data packet and RSU that maximizes the gain on the objective function among all the legitimate choices, and save the result in the set $\mathbf{A}$. From Line 10 to 17, we select the data packet and RSU that maximizes the gain per unit data length, and save the result in the set $\mathbf{B}$. Finally, we choose a better solution by comparing the objective function values of $\mathbf{A}$ and $\mathbf{B}$ from Line 18 to 23. From the algorithm procedure, we can show that it is a pseudo-polynomial-time algorithm with the computational complexity in the order of

$$O\left(R^3\left(\sum_{c=1}^{C} v_c\right)^2 U\right). \quad (12)$$

This complexity is acceptable for current RSU-aided VDTNs, because the number of RSUs, R, is typically limited. For example, even in the case of most density deployment of large-scale cities of Beijing and Shanghai, the number of RSUs is about 200-300 [33]. Moreover, the central controller has the sufficient computational capacity with the current technologies of clod computing [34] to run this heuristic algorithm.

Remark 1. Our proposed heuristic algorithm is a centralized algorithm which is run at the central controller. The central controller requires the data related information, including the data packet lengths g_c, the data lifetimes T_c and the numbers of data packets v_c for recovering data, as well as the VDTN related information, including the storage sizes of RSUs L_r, the users' content interests $w_{u,c}$, the contact rates $\theta_{u,r}$, and the contact duration rate $\lambda_{u,r}$. Note that we do not collect these information via the VDTN. Since the central controller is connected to the content servers via the wired Internet, it can obtain the content-related parameters of g_c, v_c and T_c with ease, and the required communication overhead is negligible compared with the bandwidth of the wired Internet. Similarly, the central controller can directly obtain the storage sizes of RSUs L_r with little effort via wired links. On the other hand, vehicles are equipped with communication interfaces to connect to the wireless infrastructure network, such as the cellular network. Therefore, they are able to send the vehicle-related parameters of $w_{u,c}$, $\theta_{u,r}$ and $\lambda_{u,r}$ through the uplink channel of the infrastructure wireless network, not by the channels of transmitting the content items, to the central controller, and this signaling overhead is also insignificant compared with the bandwidth of the wireless infrastructure network. Basically, there already exist certain amount of signaling between a vehicle and the infrastructure network, and the required vehicle-related parameters may take "piggybacking" in these existing signalings. Moreover, g_c, v_c T_c, L_r and $w_{u,c}$ are statistic parameters, which are well-known by the content servers or the central controller in advance. Many vehicles travel on predetermined routes and schedules and their daily mobility patterns exhibit certain regularity. Therefore, the contact patterns of many vehicles with the RSUs are often quasi-statistic, which can often be obtained by the central controller in advance. Moreover, any new contact information may be collected regularly from vehicles. On the other hand, the current technology of cloud computing is able to provide powerful storage and computation ability by the data center network connected servers, which is easy to handle such kinds of information collection and storage [34].

5 Re-Formulation for Optimal Solution

In the previous section, we have obtained an heuristic algorithm to solve the NP-hard problem (2). Although the computational complexity of this heuristic algorithm is acceptable, how close this heuristic solution to the optimal solution of the problem (2) remains unknown. In order to quantitatively evaluate the goodness of our heuristic solution, it is necessary to design an algorithm that is capable of obtaining the optimal solution of the optimization problem (2), albeit the associated computational complexity will be very high. In the objective function analysis carried out in Section 4, we note that the data replication policy $x_{r,c}$ and the communication contact are tightly coupled. This makes it impossible to obtain an explicit and closed-form expression for $Q_{u,c}$, the probability for user u to successfully receive data c before the data lifetime T_c. However, if we only consider the communication contact, we can easily obtain the explicit expression of how many packets that each user can receive. Based on this explicit expression, we

can then further consider the data replication policy $x_{r,c}$ together with the optimization constraints. Following this idea and the insights of formulation in Ref. [8], we now reformulate the optimization problem (2) into an integer programming form.

Algorithm 1. Heuristic algorithm for the data replication.

1: Initialize: allocate the first v_c packets of each data c randomly and denote this random allocation result as $\mathbf{A}_{\mathrm{ri}}$. Then set $m = j = 0$ and $\mathbf{A} = \mathbf{A}' = \mathbf{B} = \mathbf{B}' = \mathbf{A}_{\mathrm{ri}}$;
2: **while** $m = 0$ or $\widetilde{J}(\mathbf{A}) - \widetilde{J}(\mathbf{A}') > 0$ **do**
3: $m = m + 1$
4: $(r_m, c_m|_p) = \arg\ \max_{(r,c|_p)\in\mathbf{G^A}}(\widetilde{J}(\mathbf{A}\cup(r,c|_p)) - \widetilde{J}(\mathbf{A}))$
5: $\mathbf{A}' = \mathbf{A}$
6: $\mathbf{A} = \mathbf{A}\cup(r_m, c_m|_p)$
7: $L^{\mathbf{A}}_{r_m} = L^{\mathbf{A}}_{r_m} - g_{c_m}$
8: Update $\mathbf{G^A}$
9: **end while**
10: **while** $j = 0$ or $\widetilde{J}(\mathbf{B}) - \widetilde{J}(\mathbf{B}') > 0$ **do**
11: $j = j + 1$
12: $(r_j, c_j|_p) = \arg\max_{(r,c|_p)\in\mathbf{G^B}} \dfrac{\widetilde{J}(\mathbf{B}\cup(r,c|_p)) - \widetilde{J}(\mathbf{B})}{g_c}$
13: $\mathbf{B}' = \mathbf{B}$
14: $\mathbf{B} = \mathbf{B}\cup(r_j, c_j|_p)$
15: $L^{\mathbf{B}}_{r_j} = L^{\mathbf{B}}_{r_j} - g_{c_j}$
16: Update $\mathbf{G^B}$
17: **end while**
18: **if** $\widetilde{J}(\mathbf{A}) > \widetilde{J}(\mathbf{B})$ **then**
19: $\mathbf{OPT}^* = \mathbf{A}$
20: **else**
21: $\mathbf{OPT}^* = \mathbf{B}$
22: **end if**
23: Return $\mathbf{OPT}^*$ and $\widetilde{J}(\mathbf{OPT}^*)$

Firstly, regardless of how many encoded packets of data c are stored in the buffer of RSU r, we consider the maximum number of packets that can be transmitted from r to user u within the data lifetime T_c, and denote it as $\pi^c_{u,r}$. Following the same derivation of (6) given in Section 4, we can obtain the following expression for its probability function

$$P(\pi^c_{u,r} = i) = \sum_{j=0}^{+\infty} P(N^c_{u,r} = j) P(\pi^c_{u,r} = i | N^c_{u,r} = j) = \begin{cases} \sum_{j=0}^{+\infty} \frac{e^{-\varpi}\varpi^j}{j!} \frac{\gamma\left(j, \frac{\lambda_{u,r}(i+1)g_c}{\eta}\right) - \gamma\left(j, \frac{\lambda_{u,r} i g_c}{\eta}\right)}{\Gamma(j)}, & 0 \le i < v_c, \\ \sum_{j=0}^{+\infty} \frac{e^{-\varpi}\varpi^j}{j!} \frac{\Gamma(j) - \gamma\left(j, \frac{\lambda_{u,r} v_c g_c}{\eta}\right)}{\Gamma(j)}, & i = v_c. \end{cases} \tag{13}$$

Each user u can receive the packets of data c from the R possible RSUs. Each $\pi^c_{u,r}$ related to RSU $r \in \mathcal{R}$ has the $v_c + 1$ possible integer values, ranging from 0 to v_c. Specially, as shown in (13), if the aggregated contact duration between u and r is sufficiently long to transmit enough packets for data recovery, $\pi^c_{u,r} = v_c$; otherwise, $\pi^c_{u,r} < v_c$. If we consider all the RSUs $\mathcal{R}$ together, the numbers of packets for data c that user u can receive from all the R possible RSUs form a $1 \times n$ vector denoted by $\boldsymbol{\phi}_{u,c} = [\pi^c_{u,1}\pi^c_{u,2}\ldots\pi^c_{u,R}]$, where $0 \le \pi^c_{u,r} \le v_c$. Note that there are total $(v_c+1)^R$ possible patterns of $\boldsymbol{\phi}_{u,c}$, and we define $\Phi_{u,c}$ as the set containing all these patterns. Since the contact process is independent, for any $\boldsymbol{\phi}_{u,c} = [\pi^c_{u,1} = i_1 \pi^c_{u,2} = i_2 \cdots \pi^c_{u,R} = i_R] \in \Phi_{u,c}$, the probability of observing it can be calculated as

$$P(\boldsymbol{\phi}_{u,c}) = \prod_{r\in\mathcal{R}} P(\pi^c_{u,r} = i_r). \tag{14}$$

Thus, we have the probability of the maximum number of packets for data c that can be transmitted from all the RSUs to user u, without considering the packet replicate policy. We now jointly consider this maximum number of packets $\boldsymbol{\phi}_{u,c}$ with the packet replicate policy $x_{r,c}$ together. For each $\boldsymbol{\phi}_{u,c} \in \Phi_{u,c}$, given the data replication solution $x_{r,c}$, define

$$y^c_{u,r,\boldsymbol{\phi}}(x_{r,c}) = \min\{\boldsymbol{\phi}_{u,c}(r), x_{r,c}\},$$

where $\boldsymbol{\phi}_{u,c}(r)$ is the rth element of $\boldsymbol{\phi}_{u,c}$ and $1 \le r \le R$. Clearly, $y^c_{u,r,\boldsymbol{\phi}}(x_{r,c})$ is the *actual maximum number* of packets for data item c that user u can receive from r. Thus, user u can receive $\sum_{r\in\mathcal{R}} y^c_{u,r\boldsymbol{\phi}}(x_{r,c})$ packets all together from all the R RSUs. Note that if user u receives less than v_c packets in total for data item c, it cannot recover data c and thus it fails to obtain data c. On the other hand, if it receives v_c or more packets together, it can successfully recover the data c from any of v_c encoded packets. Therefore, given the data replication solution $\mathbf{X}$, we can define the binary variable $h^c_{u,\boldsymbol{\phi}}(\mathbf{X})$ that indicates whether the contact pattern $\boldsymbol{\phi}_{u,c}$ can facilitate user u successfully obtaining the mobile data c, which can be expressed as

$$h^c_{u,\boldsymbol{\phi}} = h^c_{u,\boldsymbol{\phi}}(\mathbf{X}) = \mathbf{I}\left(\sum_{r\in\mathcal{R}} y^c_{u,r,\boldsymbol{\phi}}(x_{r,c}) \ge v_c\right). \tag{15}$$

Here $\mathbf{I}(X)$ is the indication function which is equal to 1 if X holds, otherwise it is equal to 0.

Further considering all the users and all the contents together, it can readily be seen that the objective function can be written as

$$J(\mathbf{X}) = \sum_{u\in\mathcal{U}}\sum_{c\in\mathcal{C}}\sum_{\boldsymbol{\phi}_{u,c}\in\Phi_{u,c}} P(\boldsymbol{\phi}_{u,c}) h^c_{u,\boldsymbol{\phi}}(\mathbf{X}). \tag{16}$$

Therefore, the optimal communication contact-aware data replication problem expressed in (2) can be re-defined as follows

$$\max \quad J(\mathbf{X}) = \sum_{u\in\mathcal{U}}\sum_{c\in\mathcal{C}}\sum_{\boldsymbol{\phi}_{u,c}\in\Phi_{u,c}} P(\boldsymbol{\phi}_{u,c}) h^c_{u,\boldsymbol{\phi}}(\mathbf{X}) \tag{17}$$

$$\text{s.t.} \quad x_{r,c} \in \{0, 1, \ldots, v_c\}, \forall r \in \mathcal{R}, c \in \mathcal{C}, x_{r,c} \in \mathbf{X}, \tag{18}$$

$$\sum_{c\in\mathcal{C}} g_c x_{r,c} \le L_r, \forall r \in \mathcal{R}, \tag{19}$$

$$h^c_{u,\boldsymbol{\phi}} \in \{0, 1\}, \forall u \in \mathcal{U}, c \in \mathcal{C}, \boldsymbol{\phi}_{u,c} \in \Phi_{u,c}, \tag{20}$$

$$\sum_{r\in\mathcal{R}} \min\{\boldsymbol{\phi}_{u,c}(r), x_{r,c}\} \ge v_c h^c_{u,\boldsymbol{\phi}}(\mathbf{X}), \quad \forall u \in \mathcal{U}, c \in \mathcal{C}, \boldsymbol{\phi}_{u,c} \in \Phi_{u,c}. \tag{21}$$

As stated in the original optimization problem (2), the objective is to maximize the average probability that the users'

interests are satisfied. The objective function given in (17) is expressed in an explicit and closed form, and it is monotonically increasing with the variable $h^c_{u,\phi}$. The constraints (18) and (19) are related to the number of stored packets and the buffer size of each RSU, which are the two basic constraints in the original optimization problem (2). The constrains (20) and (21) are transformed from the condition of $\sum_{r\in\mathcal{R}} y^c_{u,r,\phi}(x_{r,c}) \geq v_c$, and they ensure that $h^c_{u,\phi} = 1$ if and only if $\sum_{r\in\mathcal{R}} \min\{\phi_{u,c}(r), x_{r,c}\} \geq v_c$; otherwise, $h^c_{u,\phi} = 0$. Obviously, the optimisation problem specified in (17) to (21) is NP-hard.

Note that constraint (21) is nonlinear. In order to formulate the optimal communication contact-aware data replication problem into a linear programming form, we further replace this constraint by including the variable $y^c_{u,r,\phi}$ in the formulation and rewrite the above optimization problem, (17) to (21), as the following integer linear programming problem

$$\max \quad J(\mathbf{X}) = \sum_{u\in\mathcal{U}} \sum_{c\in\mathcal{C}} \sum_{\phi_{u,c}\in\Phi_{u,c}} P(\phi_{u,c}) h^c_{u,\phi}(\mathbf{X}) \tag{22}$$

$$\text{s.t.} \quad x_{r,c} \in \{0, 1, \ldots, v_c\}, \forall r \in \mathcal{R}, c \in \mathcal{C}, x_{r,c} \in \mathbf{X}, \tag{23}$$

$$\sum_{c\in\mathcal{C}} g_c x_{r,c} - L_r \leq 0, \forall r \in \mathcal{R}, \tag{24}$$

$$h^c_{u,\phi} \in \{0, 1\}, \forall u \in \mathcal{U}, c \in \mathcal{C}, \phi_{u,c} \in \Phi_{u,c}, \tag{25}$$

$$v_c h^c_{u,\phi} - \sum_{r\in\mathcal{R}} y^c_{u,r,\phi} \leq 0, \forall u \in \mathcal{U}, c \in \mathcal{C}, \quad \phi_{u,c} \in \Phi_{u,c}, \tag{26}$$

$$y^c_{u,r,\phi} \leq \phi_{u,c}(r), \forall u \in \mathcal{U}, c \in \mathcal{C}, \phi_{u,c} \in \Phi_{u,c}, \tag{27}$$

$$y^c_{u,r,\phi} - x_{r,c} \leq 0, \forall u \in \mathcal{U}, c \in \mathcal{C}, \phi_{u,c} \in \Phi_{u,c}. \tag{28}$$

As the above optimization is an integer linear programming problem, it can be solved using the existing optimization tool kits, such as CPLEX [28] and YALMIP [29]. As expected, solving this NP-hard problem is expensive, requiring the computational complexity at least in the order of

$$O\left(R!\left(\sum_{c=1}^{C} v_c\right)!U!\right), \tag{29}$$

which is unacceptable in practical system. In the following performance evaluation, we will use this optimal solution as the reference to assess the performance of our proposed heuristic algorithm.

6 Performance Evaluation

We used two realistic vehicular mobility traces to evaluate our proposed heuristic algorithm for the data replication, and the goals of our performance evaluation were: a) validating that the results obtained by our heuristic algorithm are close to the optimal solution for the integer linear programming problem specified in (22) to (28), b) demonstrating that our heuristic scheme of contact-aware data replication achieves competitive performance of data dissemination under real-world vehicular mobility traces, compared with other existing schemes, and c) analyzing the influence of different parameters on the system performance. In order to achieve these aims, we compared the performance of our heuristic scheme, labelled as *Heuristic Coding Data Replication* (HCDR), with the following schemes:

a) *Optimal Coding Data Replication* (OCDR), representing the optimal solution of the investigated problem, which is obtained by solving the optimisation problem specified in (22) to (28) using the tool kit CPLEX.

b) *Coding and Random Data Replication* (CRDR), in which the mobile data are coded and the coded packets are uniformly and randomly replicated to the RSUs' buffers until no more packets can be stored.

c) *Non-coding and Greedy Data Replication* (NGDR) [30], [31], where the system allocates the RSUs' buffers with the original uncoded mobile data by the heuristic based greedy algorithm. This algorithm represents the most up-to-date work in the area of data replication [30].

6.1 Simulation System Set Up

Our evaluation was conducted on the real vehicular mobility traces of *Beijing*. For these traces, by focusing on the downtown area, we sorted the intersections according to their average numbers of passing vehicles recorded in the GPS traces, and selected top 25 intersections as the positions to deploy the RSUs, which is a widely used RSU deployment strategy [5]. At the same time, since there were some vehicles that rarely appeared in the selected downtown area, we only selected the vehicles that had frequently contacts with the deployed RSUs as simulation nodes. In all the experiments, the network simulation of the whole trace was divided into the warm-up period and the data dissemination period. We used the first half of the trace as the warm-up period for the central controller to accumulate the contact statistics based on the contact counts and other necessary network information in order to obtain the contact durations and contact rates of all the RSU-vehicle pairs. During the warm-up period, the controller also collected the information of the vehicles' interests, content sizes and buffer sizes from the users, content servers and RSUs. After collecting all these information, the central controller made the decision on the network resource allocation using the data replication algorithm to be evaluated, and then obtained the mobile data storage or buffer allocation policy. In the process of mobile data offloading which consisted of the second half of the trace, the mobile data were generated from the content servers, stored in the RSUs' buffers, and transmitted during the data dissemination period. After the simulation of the entire trace, we collected the system performance results.

In practical vehicular networks, the physical layer specification for wireless environment is the dedicated short range communication (DSRC) technology [36], the infrastructure-to-vehicle communication standard of IEEE 802.11p (WAVE) [36]. According to the data transmission requirements of the emerging vehicular applications of in-vehicle entertainment [37] and augmented reality assisted safety driving [38], [39], we generate the sizes of data items randomly and uniformly distributed in the range of

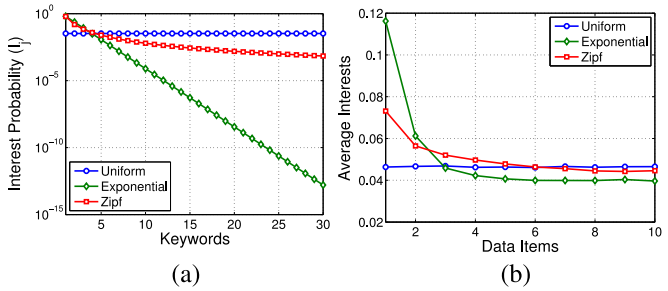


Fig. 8. (a) Interest probability of each keyword, and (b) average interest probability of each data item, where the number of keywords $K = 30$ and the number of data items $C = 10$.

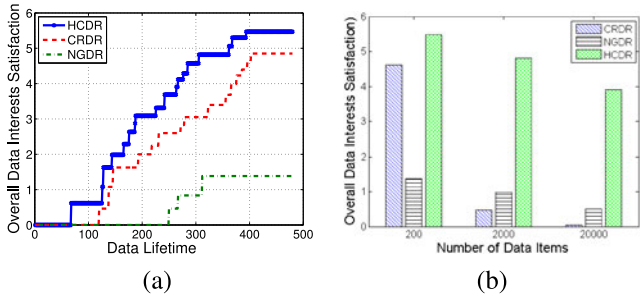


Fig. 9. Performance comparison of three different schemes under the vehicular environment of *Beijing* trace: (a) varying the data lifetime, and (b) different number of data items.

$[50, 500]$ MB. The RSUs' buffer sizes for the data buffering were randomly and uniformly generated in the range of $[1, 10]$ GB. We used the well-developed Tornado Z code [24] for erasure coding, which is a class of erasure codes that have extremely fast encoding and decoding algorithms. Since Tornado Z code has an average decoding inefficiency of 1.054, we set $\epsilon = 1.054$ for the erasure coding. As the vehicles' interests are highly heterogeneous in practice, we used the Zipf distribution to generate different user interests for each simulation. Zipf distribution is widely used in content population modeling [25], in which most of the users' interests concentrate on the popular data. To generate the Zipf interest distribution of realistic data sets, we set the number of keywords to $K = 205$ under the requirements of our interests model, which is comparable with the number of considered node, and we assumed that the keywords k_1 to k_K were ranked according to their popularity. For keyword k_j, we assumed that the average interest of all the users was I_j, and used the Zipf distribution with rank exponent 1 following true Zipf's law [35] to generate I_j, which was defined as $I_j = \frac{1/j}{\sum_{q=1}^{K} 1/q}$. For each data item c, $c \in \{1, 2, \ldots, C\}$, its describing subset of keywords was given by $\mathcal{K}_c = \{k_c, k_{c+1}, \ldots, k_{c+4}\}$ with the equal weighting $\varrho_{k_j} = 1/5$ for each keyword k_j. By using (1), we obtained the interest probability of user $u \in \mathcal{U}$ in data item $c \in \mathcal{C}$.

To illustrate the property of the Zipf interest distribution, we plot the interest probability of each keyword and the average interest probability of all the users on each data item in Figs. 8a and 8b, respectively, for a simple case of 30 keywords and 10 data items, where the results of the uniform and exponential distributions are also shown for comparison. From the results depicted in Fig. 8, we observe that for the exponential distribution, most of the user interests concentrate on the popular data items, while for the Zipf distribution, the difference between high popular data and low popular ones is smaller than the case of exponential distribution. Thus, the Zipf interest probability distribution lies between the uniform and exponential distributions. This type of user interest is also demonstrated to fit the realistic mobile data downloading applications well [32].

In a generic mobile content dissemination system, the most important performance metrics include the average probability of the users receiving their mobile contents and the data dissemination delay. For our RSU-aid vehicular delay tolerant data dissemination system, however, the data dissemination delay becomes a secondary issue. In our investigation, we consider the data dissemination delay as the system constraint by imposing the requirement of delivering each mobile data item c before its lifetime T_c. Therefore, we mainly focus on the overall data interests satisfaction (ODIS) of all the users, which is clearly the optimization objective in our problem (2). Thus, we define the observed utility function of overall data interests satisfaction in the simulation, defined by

$$\text{ODIS} = \sum_{\forall u \in \mathcal{U}, \forall c \in \mathcal{C}} w_{u,c} G_{u,c}, \tag{30}$$

as our performance evaluation metric, where

$$G_{u,c} = \begin{cases} 1, & \text{vehicle } u \text{ receives data } c \text{ before } T_c, \\ 0, & \text{otherwise.} \end{cases} \tag{31}$$

6.2 Results and Analysis

We first evaluated the performance gap between our proposed heuristic scheme HCDR and the optimal solution OCDR. It is worth pointing out again that the computational complexity of the OCDR is very high, as given in (29), and it may not be practical to implement it in real systems. We varied the number of deployed RSUs from 5 to 25 and set the data lifetime T_c to 100 and 120, respectively. From the simulation, we observe that the performance of the heuristic scheme HCDR is always close to that of the optimal solution OCDR under the various scenarios of different numbers of deployed RSUs and two different data lifetimes. To quantitatively analyze how close the HCDR solution to the optimal OCDR solution, we calculated the average performance deviation between the HCDR and the OCDR, which is under 7 percent over the utilized traces and two data lifetimes, which means that the performance of the heuristic scheme HCDR is only 7 percent worse compared with the optimal solution. This confirms the excellent performance achieved by the low-complexity HCDR scheme.

After the quantitatively evaluation to demonstrate that the performance our HCDR is very close to the optimal solution OCDR, we enlarged the number of deployed RSUs to 50, and compared the performance of the two existing CRDR and NGDR schemes with our HCDR by varying the data lifetime T_c under the vehicular environment of *Beijing* trace. The results obtained are shown in Fig. 9a. For each selected lifetime, we re-generate the user interest distribution according to the introduced approach

to capture the interest changes in the evaluation. As expected, as the data lifetime increases, the data interest satisfaction increases too. Intriguingly, each ODIS metric curve in Fig. 9a shows step increase as the data lifetime increases. The reason is because in a large-scale urban vehicular environment, the contact duration is much smaller than the contact interval, as can be clearly seen from the modelling results of Section 3. Data transmission occurs during the contact, which does not happen very often in time, and data transmission opportunities do not increase continuously with the increase of data lifetime. The results of Fig. 9a confirm that our HCDR significantly outperforms the two existing schemes, the CRDR and NGDR. More specifically, in terms of the average ODIS metric over the range of data lifetimes simulated, our HCDR is about 49.2 and 350 percent better than the CRDR and NGDR, respectively. It is worth noting that when erasure coding is applied, even the random data replication scheme, the CRDR, performs much better than the non-coding greedy based data replication scheme, the NGDR. Thus, applying erasure coding in a RSU-aided data dissemination system is a good idea. On the other hand, in order to investigate the influence of the number of data items on the system performance, we set it to be 200, 2,000, and 20,000, and obtain the performance of the three schemes in Fig. 9b. From the results, we can observe that when the number of data items is enlarged, the performance of CRDR drops quickly since it replicates the data uniformly and randomly; while NGDR achieves better performance since it utilizes the proposed greedy replication algorithm. Among them, our HCDR scheme is the most robust solution with the increase of the data items. These results further demonstrate that the efficiency of the combined utilization of erasure coding and greedy replication.

In order to further analyze the influence of some important parameters on the performance of the data dissemination system, we observe the system performance obtained by the proposed HCDR scheme under different numbers of deployed RSUs by varying the selected parameters. Since how much redundancy that the data are encoded is vital to the data dissemination as observed before, we first investigated the influence of the data coding redundancy. For a data $c \in \mathcal{C}$ with length l_c, after erasing coding with the given coding rate, it is coded to ψ_c packets with length of g_c. Obviously, $\psi_c \geq v_c$. Thus its data coding redundancy, denoted by ∂_c, can be expressed as $\partial_c = \psi_c g_c / l_c$. The overall system coding redundancy, denoted by ∂, can then be expressed as

$$\partial = \frac{\sum_{c \in C} \psi_c g_c}{\sum_{c \in C} l_c}. \tag{32}$$

We varied the system coding redundancy ∂ from 1.5 to 6, and depict the results obtained in Fig. 10a. From the results of Fig. 10a, we observe that the system performance is significantly enhanced when ∂ increases from 1.5 to 3, while for $\partial > 4$, the increase in the system performance is slow down, or even becomes saturated as in the case of 25 deployed RSUs. This reveals that there exists a limit in using the coding redundancy to increase the data dissemination efficiency. In designing practical systems, therefore,

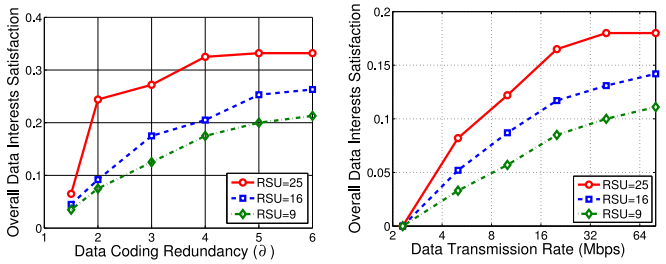


Fig. 10. The influence of (a) coding redundancy, and (b) data transmission rate on the system performance with different number of deployed RSUs under *Beijing* trace.

a suitable data coding redundancy should be chosen to trade off the performance gain and the data dissemination cost. Next, we investigated the influence of the data transmission rate η on the system performance by varying η from 2 to 64 Mbps, and the results obtained are shown in Fig. 10b. As expected, the higher the data transmission rate is, the higher the achievable data interest satisfaction. However, there seems to exist a performance enhancement limit with the increasing of η. For example, in the case of 25 deployed RSUs and with $\eta = 32$ Mbps, the link transmission capacity seems already to be sufficient to maximise the achievable system performance.

7 Conclusions and Future Works

We have investigated the problem of communication contact-aware mobile data replication for the mobile data offloading system in RSU-aided vehicular delay tolerant networks. We have studied this problem in a realistic VDTN environment, where the communication contact is modelled based on real vehicular mobility traces, the RSUs' buffers for storing the mobile data are limited, and the vehicles have different interests on different data items. Extensive simulation results obtained based on real large-scale vehicular traces have demonstrated that our heuristic scheme significantly outperforms other existing algorithms for mobile data offloading.

Our work has also revealed the exponential models for both inter-contact time and contact duration between the RSUs and vehicles in the large-scale urban vehicular mobility environments of Beijing and Shanghai, two of the largest cities in China. This is different from the existing results found in human mobility where the contact time obeys power law distribution. Thus, the contact time decays more quickly in RSU-aided vehicular delay tolerant networks than in human based networks. This result also suggests that the existing forwarding schemes based on power law distribution are possibly over pessimistic.

Our investigated mobile data replication scheme is based on the statistic distributions of the contact interval and duration between the RSU and vehicle. Thus, the designed data dissemination strategy is quasi-statistic. On the other hand, since vehicular mobility would exhibit patterns, using specific prediction algorithm is able to make explicit predication based on vehicular historical mobility trajectory. In this case, microscope mobility models are need by investigating these taxis mobility traces via geometric model or hidden Markov model based map matching to extract more

precision information about the inserted positions. With these kinds of models and algorithms to predict their mobility trajectory, RSUs can intelligently pre-fetch the mobile data to enhance the data dissemination efficiency. These are interesting problems for our future works.

ACKNOWLEDGMENTS

This work was supported by the National Basic Research Program of China (973 Program) (No. 2013CB329001), National Nature Science Foundation of China (Nos. 61301080, 91338203, 61171065, and 91338102). A small part of results was presented in IEEE ICC 2013. P. Hui is the corresponding author.

REFERENCES

[1] M. Khabazian, S. Aissa, and M. Mehmet-Ali, "Performance modeling of message dissemination in vehicular ad hoc networks with priority," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 1, pp. 61–71, Jan. 2011.

[2] G. Dimitrakopoulos and P. Demestichas, "Intelligent transportation systems," *IEEE Veh. Technol. Mag.*, vol. 5, no. 1, pp. 77–84, Mar. 2010.

[3] J. Zhao and G. Cao, "VADD: Vehicle-assisted data delivery in vehicular ad hoc networks," in *Proc. 25th IEEE INFOCOM*, Apr. 23-29, 2006, pp. 1–12.

[4] J. J. Blum, A. Eskandarian, and L. J. Hoffman, "Challenges of intervehicle ad hoc networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 4, pp. 347–351, Dec. 2004.

[5] A. Abdrabou and W. Zhuang, "Probabilistic delay control and road side unit placement for vehicular ad hoc networks with disrupted connectivity," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 1, pp. 129–139, Jan. 2011.

[6] X. Lin, R. Lu, X. Liang, and X. Shen, "STAP: A social-tier-assisted packet forwarding protocol for achieving receiver-location privacy preservation in VANETs," in *Proc. 30th IEEE INFOCOM*, Apr. 10-15, 2011, pp. 2147–2155.

[7] D. Camara, N. Frangiadakis, F. Filali, and C. Bonnet, "Vehicular delay tolerant networks," In *Handbook Research on Mobility Computing: Evolving Technologies and Ubiquitous Impacts*, M. M. Cruz-Cunha, and F. Moreira, Eds. Hershey, PA, USA: Information Science Reference, 2011, pp. 356–367.

[8] X. Zhuo, Q. Li, W. Gao, G. Cao, and Y. Dai, "Contact duration aware data replication in delay tolerant networks," in *Proc. 19th IEEE Int. Conf. Netw. Protocols*, Oct. 17–20, 2011, pp. 236–245.

[9] M. Doering, W. B. Pöttner, T. Pögel, and L. Wolf, "Impact of radio range on contact characteristics in bus-based delay tolerant networks," in *Proc. 8th Int. Conf. Wireless On-Demand Netw. Syst. Serv.*, Jan. 26–28, 2011, pp. 195–202.

[10] M. Johnson, L. De Nardis, and K. Ramchandran, "Collaborative content distribution for vehicular ad hoc networks," in *Proc. 44th Allerton Conf. Commun., Control, Comput.*, Sep. 27–29, 2006, pp. 751–760.

[11] H. Zhu, M. Li, L. Fu, G. Xue, Y. Zhu, and L. M. Ni, "Impact of traffic influxes: Revealing exponential inter-contact time in urban VANETs," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 8, pp. 1258–1266, Aug. 2011.

[12] J. Gozálvez, M. Sepulcre, and R. Bauza, "IEEE 802.11p vehicle to infrastructure communications in urban environments," *IEEE Commun. Mag.*, vol. 50, no. 5, pp. 176–183, May 2012.

[13] P.-C. Lin, "Optimal roadside unit deployment in vehicle-to-infrastructure communications," in *Proc. 12th Int. Conf. IEEE ITS Telecommun.*, Nov. 5–8, 2012, pp. 796–800.

[14] I. Filippini, F. Malandrino, G. Dán, M. Cesana, C. Casetti, and I. Marsh, "Non-cooperative RSU deployment in vehicular networks," in *Proc. 9th IEEE Annual Conf. Wireless On-Demand Netw. Syst. Serv.*, Jan. 9–11, 2012. pp. 79–82.

[15] S. I. Sou, "A power-saving model for roadside unit deployment in vehicular networks," *IEEE Commun. Lett.*, vol. 14, no. 7, pp. 623–625, Jul. 2010.

[16] U. Lee, E. Magistretti, B. Zhou, M. Gerla, P. Bellavista, and A. Corradi, "Efficient data harvesting in mobile sensor platforms," in *Proc. 4th IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, 2006, pp. 1–8.

[17] K. W. Lim and Y. B. Ko, "Multi-hop data harvesting in vehicular sensor networks," *IET Commun.*, vol. 4, no. 7, pp. 768–775, 2010.

[18] U. Lee, Z. Biao, G. Mario, M. Eugenio, B. Paolo, and C. Antonio. "Mobeyes: Smart mobs for urban monitoring with a vehicular sensor network," *IEEE Wireless Commun.*, vol. 13, no. 5, pp. 52–57, Oct. 2006.

[19] C. Zhang, R. Lu, X. Lin, P.H. Ho, and X. Shen, "An efficient identity-based batch verification scheme for vehicular sensor networks," in *Proc. 27th IEEE Conf. Comput. Commun.*, May 2008, pp. 1–6.

[20] X. Li, W. Shu, M. Li, H. Huang, and M. Y. Wu, "DTN routing in vehicular sensor networks," in *Proc. IEEE Global Telecommun. Conf.*, Nov. 2008, pp. 1–5.

[21] A. Rowstron and G. Pau, "Characteristics of a vehicular network," Univ. of California, LA, USA, Tech. Rep. UCAM-CL-TR-617, Feb. 2009.

[22] Y. Liao, K. Tan, Z. Zhang, and L. Gao, "Estimation based erasure-coding routing in delay tolerant networks," in *Proc. Int. Conf. Wireless Commun. Mobile Comput.*, Jul. 3–6, 2006, pp. 557–562.

[23] C. Fragouli, J. Widmer, and J.-Y. Le Boudec, "On the benefits of network coding for wireless applications," in *Proc. 4th Int. Symp. Modeling Optimization Mobile, Ad Hoc Wireless Netw.*, Apr. 3–6, 2006, pp. 1–6.

[24] J. W. Byers, M. Luby, and M. Mitzenmacher, "A digital fountain approach to asynchronous reliable multicast," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 8, pp. 1528–1540, Oct. 2002.

[25] W. Gao and G. Cao, "User-centric data dissemination in disruption tolerant networks," in *Proc. 30th IEEE INFOCOM*, Apr. 10–15, 2011, pp. 3119–3127.

[26] M. Li, H. Zhu, Y. Zhu, and L. M. Ni, "Ants: Efficient vehicle locating based on ant search in shanghaiGrid," *IEEE Trans. Veh. Technol.*, vol. 58, no. 8, pp. 4088–4097, Oct. 2009.

[27] B. S. Everitt, *The Cambridge Dictionary of Statistics*, 3rd Ed. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[28] CPLEX: Linear Programming Solver. [Online]. Available: http://www.ilog.com/, 2010.

[29] I. Löfberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in *Proc. IEEE Int. Symp. Comput. Aided Control Syst. Des.*, Sep. 2–4, 2004, pp. 284–289.

[30] Y. Li, M. Qian, D. Jin, P. Hui, Z. Wang, and S. Chen, "Multiple Mobile data offloading through disruption tolerant networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 7, pp. 1579–1596, Jul. 2014.

[31] A. Kulik, H. Shachnai, and T. Tamir, "Maximizing submodular set functions subject to multiple linear constraints," in *Proc. 20th Annu. ACM-SIAM Symp. Discrete Algorithms*, Jan. 4–6, 2009, pp. 545–554.

[32] W. Gao, G. Cao, A. Iyengar, and M. Srivatsa, "Cooperative caching for efficient data access in disruption tolerant networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 3, pp. 611–625, Mar. 2014.

[33] Y. Li, W. Ren, D. Jin, P. Hui, L. Zeng, and D. Wu, "Potential predictability of vehicular staying time for large scale urban environment," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 322–333, Jan. 2014.

[34] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A.-Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, pp. 50–58, Apr. 2010.

[35] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, Mar. 1999, pp. 126–134.

[36] *Family of Standards for Wireless Access in Vehicular Environments (WAVE)*. IEEE Std 1609, U.S. Department of Transportation, Apr. 13, 2013.

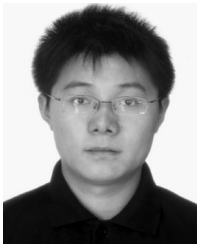
[37] U. Lee, R. Cheung, and M. Gerla, "Emerging vehicular applications," *Vehicular Networks: From Theory Practice*. London, U.K.: Chapman and Hall, 2009.

[38] S. Tachi, M. Inami, and Y. Uema, "Augmented reality helps drivers see around blind spots," *IEEE Spectrum*, 31 Oct. 2014.

[39] D. Wagner, G. Reitmayr, A. Mulloni, E. Mendez, and S. Diaz, "Mobile augmented realityTracking, mapping and rendering," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, Sep. 2014, pp. 383–383.

[40] D. Jiang and L. Delgrossi, "IEEE 802.11p: Towards an international standard for wireless access in vehicular environments," in *Proc. IEEE Veh. Technol. Conf.*, Apr. 2008, pp. 2036–2040.

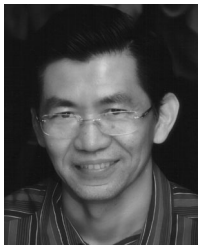
[41] M. Li, T. Wu, W. Lin, K. Lan, C. Chou, and C. Hsu, "On the feasibility of using 802.11p for communication of electronic toll collection systems," *ISRN Commun. Netw.*, vol. 2011, Article ID 723814, 2011.



Yong Li (M'09) received the BS degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2007, and the PhD degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. During July to August in 2012 and 2013, he worked as a visiting research associate in Telekom Innovation Laboratories (T-labs) and the HK University of Science and Technology, respectively. During December 2013 to March 2014, he visited the University of Miami, Florida, as a visiting scientist. He is currently a faculty member of electronic engineering at the Tsinghua University. His research interests are in the areas of networking and communications, including mobile opportunistic networks, device-to-device communication, software-defined networks, network virtualization, future Internet, etc. He has published more than 100 research papers and has 10 granted and pending Chinese and International patents. He has served as a technical program committee (TPC) chair for WWW workshop of Simplex 2013, served as the TPC of several international workshops and conferences. He is also a guest-editor for *ACM/Springer Mobile Networks and Applications*, Special Issue on Software-Defined and Virtualized Future Wireless Networks. Now, he is the associate editor of *EURASIP Journal on Wireless Communications and Networking*. He is the member of the IEEE.

Depeng Jin received the BS and PhD degrees from Tsinghua University, Beijing, China, in 1995 and 1999, respectively both in electronics engineering. He is an associate professor at Tsinghua University and a vice chair of the Department of Electronic Engineering. He was awarded National Scientific and Technological Innovation Prize (second class) in 2002. His research fields include telecommunications, high-speed networks, ASIC design, and future Internet architecture. He is the member of the IEEE.

Pan Hui received his Ph.D degree from Computer Laboratory, University of Cambridge, and earned his MPhil and BEng both from the Department of Electrical and Electronic Engineering, University of Hong Kong. He is currently a faculty member of the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology where he directs the HKUST-DT System and Media Lab. He also serves as a Distinguished Scientist of Telekom Innovation Laboratories (T-labs) Germany and an adjunct Professor of social computing and networking at Aalto University Finland. Before returning to Hong Kong, he has spent several years in T-labs and Intel Research Cambridge. He has published around 150 research papers and has some granted and pending European patents. He has founded and chaired several IEEE/ACM conferences/workshops, and has been serving on the organising and technical program committee of numerous international conferences and workshops including ACM SIGCOMM, IEEE Infocom, ICNP, SECON, MASS, Globecom, WCNC, ITC, ICWSM and WWW. He is an associate editor for the *IEEE Transactions on Mobile Computing* and *IEEE Transactions on Cloud Computing*. He is a senior member of the IEEE.



Sheng Chen (M'90-SM'97-F'08) received the BEng degree from the East China Petroleum Institute, Dongying, China, in January 1982, and the PhD degree from the City University, London, in September 1986, both in control engineering. In 2005, he was awarded the higher doctorate degree, Doctor of Sciences (DSc), from the University of Southampton, Southampton, United Kingdom. From 1986 to 1999, he held research and academic appointments at the Universities of Sheffield, Edinburgh and Portsmouth, all in United Kingdom. Since 1999, he has been with electronics and computer science, the University of Southampton, United Kingdom, where he currently holds the post of a professor in intelligent systems and signal processing. He is a distinguished adjunct professor at King Abdulaziz University, Jeddah, Saudi Arabia. He is a chartered engineer (CEng) and a fellow of IET (FIET). His recent research interests include adaptive signal processing, wireless communications, modelling and identification of nonlinear systems, neural network and machine learning, intelligent control system design, evolutionary computation methods and optimisation. He has published more than 470 research papers. He is an ISI highly cited researcher in the engineering category (March 2004). He is a fellow of the IEEE.