Contact-Aware Optimal Resource Allocation for Mobile Data Offloading in Opportunistic Vehicular Networks

Xiangming Zhu, Yong Li, *Senior Member, IEEE*, Depeng Jin, *Member, IEEE*, and Jianhua Lu, *Fellow, IEEE*

***Abstract*—To cope with increasing vehicular traffic and extensive application demands on vehicular cellular networks, opportunistic vehicular networks are used to disseminate mobile data by high-capacity device-to-device communication contacts when vehicles come into the communication range of each other, which offloads significant traffic from the cellular network. The current opportunistic vehicular data transmission models often do not take into account the fact that the contact duration is usually very short, due to the high mobility of vehicles, which will limit the amount of data that can be transmitted during one opportunistic contact. In this paper, we consider a mobile data offloading system that integrates cellular network and vehicular opportunistic communications. Based on this proposed system, we establish a mathematical framework to study the problem of contact-aware optimal resource allocation for mobile data offloading by the explicit consideration of the contact duration. Based on theoretical analysis of this problem, we derive an optimal scheme for mobile data offloading to optimally allocate the network resources in terms of opportunistic contacts and offloading helpers' storage. By carrying out extensive simulation using two realistic urban vehicular traces, we demonstrate the effectiveness of our contact duration aware optimal offloading scheme, in comparison with a wide range of existing schemes.**

***Index Terms*—Contact duration, mobile data offloading, resource allocation, vehicular opportunistic networks.**

I. Introduction

WITH the ever-increasing number of vehicles on roads, traffic jams and accidents have become a serious and widespread problem [1]. To alleviate this serious problem, recently there is a strong interest in developing vehicular networks to enable wireless communications for vehicles to obtain information and content from the resources located in Internet. At the same time, mobile Internet access is getting increasingly popular for providing various services and applications, including video, audio and images. From the latest data, Cisco forecasts that mobile traffic will increase eight-fold between 2015 and 2020 and is expected to grow to 30.6 exabytes per month by 2020 [2]. Among this traffic, about 75% will be mobile video data [2]. Mobile cellular networks provide the most popular method of mobile access today. With the increase of mobile services and user demands, however, cellular networks will very likely be overloaded and congested in the near future. Especially during peak time and in urban areas, vehicular communication will face non-ignorable performance hits in terms of low available network bandwidth, missed calls, and unreliable coverage. To cope with this explosive traffic demands and mobile data growth with the limited cellular network capacity, it is an important agenda for cellular providers to provide quick and promising solutions in vehicular network. Opportunistic contact between vehicles offers higher bandwidth communication capacity for data transmission, which is known as opportunistic vehicular networks, or Vehicular Delay Tolerant Networks (VDTNs) [3]–[7]. 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. Among vehicular networks, large amounts of data traffic will be hot news and popular contents, for which most users will have no specific preference. Thus these contents are regular and commonly-popular. Also, some popular contents may be required by many subscribers that share the same interests, such as popular video of games and shows. Benefiting from these common interests of users, providers only need to deliver the information to a small number of users, and the data will be further disseminated by the selected users through VDTN communications. This kind of offloading is very attractive to the operators as it is the quickest way, at the smallest cost, to support the exponential growth of mobile data, which otherwise could not be supported even if all the operators update their cellular network infrastructures to 4G [8].

In this work, we consider an integrated cellular and opportunistic communication network architecture to better utilise the available resources for supporting mobile data traffics. In this integrated network, as usual, the mobile traffic services are provided on demand to the vehicles that request the mobile data. The cellular network with its seamless coverage supports the control channels for service requests and system information collection to make the offloading decision. As the cellular network is connected to the content servers, it is also responsible for

Manuscript received May 25, 2015; revised December 3, 2015, June 14, 2016, and October 1, 2016; accepted January 21, 2017. Date of publication February 13, 2017; date of current version August 11, 2017. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2013CB329001 and in part by the National Nature Science Foundation of China under Grant 91338102 and Grant 61321061. The review of this paper was coordinated by Dr. M. Dianati.

The authors are with the Department of Electronic Engineering, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: zhu-xm10@qq.com; liyong07@tsinghua.edu.cn; jindp@tsinghua.edu.cn; lhh-dee@mail.tsinghua.edu.c).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2017.2668396

transmitting the mobile data to a small number of the selected or targeted vehicles. Then, the mobile data are disseminated via vehicle-to-vehicle opportunistic communications to the vehicles that are requesting them. By offloading non-realtime mobile data to the VDTN, the cellular network can avoid congestion and free sufficient resources for other realtime applications. In such an integrated system, the data offloading efficiency depends on how the targeted vehicles store the mobile data and, furthermore, the vehicular mobility critically influences the opportunistic content transmission. Therefore, how the cellular network transmits the mobile data to the targeted vehicles, by considering the vehicular mobility and content requirements as well as the targeted vehicles' data storage policy, is an important problem to be solved. As high-speed vehicles travel along the roads in the city, the wireless link from a vehicle to another is highly dynamic and subjects to opportunistic contact, which makes the decision for the data storage among the targeted vehicles quite challenging [1].

Currently, potential applications for vehicular networks are a mix of news, entertainment, travel and business. For example, ads companies periodically broadcast multimedia advertisements of local businesses to potential vehicles, and traffic authorities may disseminate an accurate update of a geographic map for scenic area plus some traffic and accident information for intelligent navigation. For these applications, popular content should be delivered to a potential number of users, and these popular contents usually have potentially large file sizes. Thus, for our investigated VDTN based mobile data offloading, the data are usually large multimedia contents, such as video, audio and images, and it requires a relatively long time to transmit the whole data content [3]. However, in VDTN, network nodes are high speed vehicles. Even if they communicate via WiFi which has a relatively long transmission range, the contact duration is very limited and, consequently, the communication capacity of one contact is very limited as well [9]. Therefore, a single mobile data item may need more than one contact to complete the transmission from one node to another. Thus, we need an efficient offloading system and an optimal scheme for mobile data offloading. Although some protocol and system design works take the limited contact duration into consideration by splitting large data into packets or chunks that are individually transmitted [33]–[35], many current resource allocation approaches still simply assume that the contact duration is sufficient long for the complete transmission of one data item, which is unrealistic. For example, RAPID has considered the limited contact duration, but it focuses on the design of routing protocols [43]. A similar work in human mobility [9] does consider the contact duration, but assumes it following a power law distribution, which is not suitable for vehicular network. By collecting the real mobility traces from about 27 000 operational taxis during one month in Beijing city as well as using the existing mobility traces recorded in Shanghai city [27], we analyse the vehicular contact duration characteristics [28]. Our study find that different from the existing results for human mobility, in vehicular mobility, 80% of the total distribution of contact duration obeys an exponential distribution. Furthermore, the study [9] focuses on a fully distributed algorithm, which fails to obtain the optimal system performance. In our proposed integrated mobile data offloading system, by contrast, the cellular network provides seamless coverage to support control channels for collecting system information and making offloading decisions. Therefore, a more efficient centralized algorithm is required in order to better utilise the available resources and to increase mobile data offloading efficiency.

In this paper, based on the proposed integrated mobile offloading system, we investigate the problem of contact duration aware optimal network resource allocation in the environment of vehicular based mobile data offloading. We explicitly take into account the key limitation of very short contact duration and propose an optimal system resource allocation algorithm for the cellular network to distribute the content to the targeted vehicles. Our novel contribution is threefold which are summarized as follows.

1) We propose a mobile offloading system that includes the control and data transmission channels by integrating VDTN with cellular network, and we establish a mathematical framework for the contact duration aware optimal mobile data offloading problem, where vehicles have limited storages, as an utility maximisation problem under linear constraints.
2) We provide an optimal system resource allocation solution for this challenging problem by proposing an efficient algorithm to allocate the limited buffers of targeted vehicles to the mobile data that are waiting for offloading. Based on detailed theoretical analysis, we prove that our algorithm achieves the optimal system performance.
3) Through extensive real trace-driven simulations, we demonstrate that our algorithm achieves excellent system performance in challenging opportunistic vehicular network environments, in comparison with several existing schemes.

The rest of the paper is organized as follows. After presenting the related works in Section II, we describe the integrate mobile data offloading system and formulate the associated optimization problem in Section III. In Section IV, we analyse this optimization problem and design the related optimal algorithm to obtain the system solution. In Section V, we introduce the experimental environment for performance evaluation and provide extensive simulation results. Finally, we conclude the paper in Section VI.

II. Related Works

Recently, the mobile data offloading problem is investigated under the general context of offloading mobile data from the overloaded cellular networks to other networks [4], [8], [10], [12], [38]. These works can be divided into three categories according to their offloading destination networks. One category, known as broadcast offloading, uses mobile broadcasting networks to offload the cellular traffic [10]. The second more popular approach uses freely available WiFi networks [8], [12], [13], which we refer to as WiFi offloading. Higgins *et al.* [13] proposed *Intentional Networking*, a mechanism for using network diversity to improve the application-level aggregate bandwidth. By using the measurements obtained in a WiFi and cellular 3G network coverage area, it is shown in [13] that the system latency is enhanced significantly. Lee *et al.* [8] studied daily mobility patterns of humans, and found that WiFi could offload about 65% of the total mobile data traffic from mobile 3G networks and save 55% of battery power without using any delayed transmission. Balasubramanian *et al.* [12] studied the problem of augmenting mobile 3G using WiFi in a moving vehicular environment. The designed system, called Wiffler, augments mobile 3G capacity by leveraging delay tolerance and fast switching. It showed that, for a delay tolerance of 60 seconds, 45% of traffics can be offloaded to WiFi [12].

In this paper, we focus on another type of offloading, which transmits the traffic by opportunistic communications between vehicles, namely, the opportunistic vehicular network offloading. In general, opportunistic offloading utilises opportunistic communications [4], [38]. Han *et al.* [4] exploited opportunistic communications among people to facilitate offloading by peer-to-peer sharing after one user obtained the content, and they showed that the designed heuristic algorithm can offload cellular traffic by up to 73.66% for the studied traces. However, the work in [4] is for the human based Delay Tolerant Network (DTN) offloading and uses multi-hop opportunistic forwarding, in which a set of targeted users are selected to disseminate the traffic to all other users in the network. This multi-hop forwarding requires all the users to cooperate in traffic offloading by using their own resources, which may not be applyed to all networks [20]. Moreover, it is not suitable to offload large data items. In our study, we focus on the VDTN scenario, and our solution relies only on a small subset of users who are willing to participate in the offloading, which is a more practice scenario for the vehicular mobile data offloading.

Storage allocation problems are also addressed in the area of vehicular-based or general DTN-based content distribution and routing [5], [7], [14]–[16]. However, there are significant differences between the offloading and the content sharing and routing. Firstly, in the application of mobile data offloading, mobile data originate from the Internet, and most data consist of large files with very different sizes. Existing works on DTN content sharing do not take the content size into consideration [15], [16], which however has a big impact on vehicular mobile data offloading. A content dissemination framework that harnesses ad hoc communication opportunities to minimize the load on the wireless infrastructure was proposed by [46]. However, this work mainly focused on the strategies to determine how many copies of the content should be injected, when, and to whom, and didn't consider the influence of content size and storage constraints. In our work, we explicitly consider different data sizes and storage constraints. Secondly, in mobile data offloading, the latency of data matters, since it impacts the user experience. Thirdly, in an offloading problem, the system is more concerned with how much data are offloaded from the cellular network and how much capacity can be saved. Finally, in offloading, nodes are usually mobile users, which can use for example the 3G network to communication, and the control channel of the cellular system can collect and obtain the overall system metrics of node contact rates and other information without bring more signaling overhead to the networks. Therefore, centralized algorithms work better in this case, and it is not necessary to design distributed algorithms. By contrast, in DTN content sharing, distributed algorithms are usually required. In both these works of content distribution and mobile data offloading, how to describe the relation between the content and users is a basic modeling problem. Until now, the most widely used approach is utilizing the users' interests, which depend on the content popularity, to model the users' requirements on different content [17], [18], [29]. This model is also introduced into vehicular networks by [17], which proposed a protocol for data dissemination according to vehicles' subscriptions indicating the users' interests. In our work, we also use this most recognized model of user interest to investigate the contact duration aware optimal network resource allocation for vehicular based mobile data offloading application.

In vehicular network design, some works [33]–[35] take the limited contact duration into consideration by splitting large mobile data into packets or chunks so that they can be transmitted separately. However, these works focus on the protocol and system design, and do not investigate the optimal resource allocation problem by taking the limited contact duration into consideration. Balasubramanianhas *et al*. considere the influence of the limited contact duration, but what they focus on is the routing protocols [43]. Besides, Yao *et al.* study the minimum offloading problem for bulk data dissemination in VDTNs [47]. What they concentrate on is how to determine the initial offloading points and the dissemination scheme for offloaded traffic and they also do not investigate optimal resource allocation problem. In the past two years, researchers have started to pay attention to the effects of the limited contact duration in the resource allocation algorithms and performance evaluation. Qiu *et al.* [19], [37] present a framework to model the message propagation process and give a detailed expression of average information dissemination delay by considering the contact duration. Different from their work, which focuses on performance evaluation, we consider how to utilise the limited contact duration to design the network algorithm. Zhuo *et al.* [9], [36] recognise the deficiency of the existing data replication schemes that treat the complete data item as the replication unit, and propose a contact duration aware data replication scheme for human based DTN. However, as mentioned before, different from this work for human based DTN, we study the mobile data offloading problem in vehicular networks, and our aim is to derive an optimal system solution for VDTN based mobile data offloading.

III. System Overview and Problem Formulation

A. System Overview

The network topology is shown in Fig. 1, where vehicles travel around the city roads, and a cellular network provides a seamless coverage over the region. All the vehicle stations are dual mode, namely, they can connect to the cellular network, and they also form the VDTN via opportunistic communications. The base stations of the cellular network are connected to the content servers in the Internet through wire-line links. A central controller is deployed for the integrated mobile data offloading

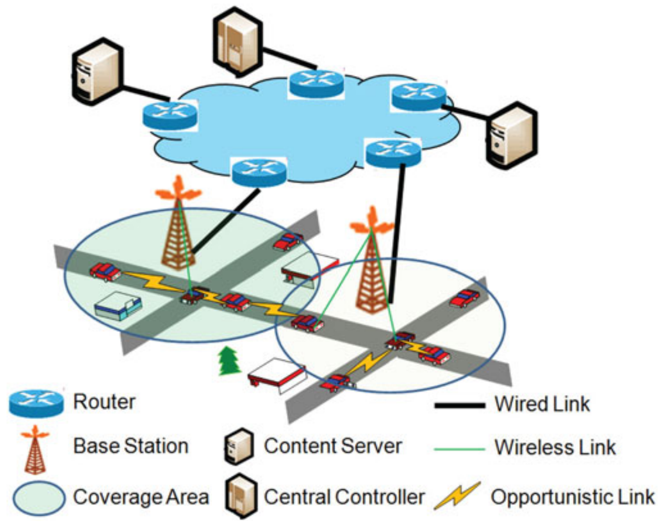


Fig. 1. Mobile data offloading system integrating cellular network and opportunistic communications.

system, and it can communicate with the cellular network, vehicle stations, and content servers. The central controller acts as the intermediary between any cellular vehicular user and any remote content server. Vehicle stations requiring mobile data send their data requests to the central controller via the cellular network firstly, and then the controller implements the offloading scheme to decide whether to send these demands to the content server. Then, the requested data are first delivered from the corresponding content servers to some chosen or targeted vehicles via the wireless links of the cellular network under the guidance of the offloading storage policy. A targeted vehicle will further disseminate the mobile contents to the corresponding vehicles that request them through opportunistic communication which occurs when two vehicles move into the communication range of each other. In this kind of opportunistic communication relaying on limited contact duration, data transmission occurs among the node pairs in the sparse network. Even in the 802.11p based vehicular communication interfaces with large interaction range, the usually utilized directional antenna for opportunistic communication should reduce the contention significantly [38]. Thus, for the MAC layer of the opportunistic content transmission, we assume no contentions exist between different ongoing transmissions. However, the process of discovering all neighbors in a device's communication range, is of great importance and particularly challenging when devices have directional antennas instead of omni-directional ones. There are already many works concentrated on this problem, and some protocols and algorithms have been proposed which can achieve considerable performance [44]. After the vehicles receive the requested data, as usual, they will send acknowledgements to the content servers via the cellular network. Therefore, the content servers know which vehicles have not received their requested data. After the deadline of delivering the mobile data, a content server can directly send the requested data to those vehicles that have not received them yet.

The central controller is tasked to make the content storage decisions for the content servers to distribute the data into the buffers of selected vehicles, based on the vehicular mobility patterns and mobile data demands. The mobility patterns are related to the contact patterns of vehicular pairs, occurring when vehicles move into the communication range. Two important metrics are the average contact rate, which is how frequently they will meet with each other on average, and the average contact duration, which is the average time that they meet in each contact. Many vehicles travel on predetermined routes and schedules, and examples include city buses and people traveling by cars to and from work. Therefore, daily mobility patterns exhibit certain regularity, and the contact rates between vehicles are often quasi-statistic. Although the realtime contact between vehicles may change significantly, the statistical results over a long time are of high probability to converge to a stable value due to the regular mobility patterns. Thus, the information of contact patterns can often be obtained by the central controller in advance with high accuracy. Moreover, the 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 required information to make the storage allocation decision.

Between any cellular vehicular user and any remote content server, the central controller acts as the intermediary. To keep track of requests from the vehicular users, the main requirement for the central controller is the computational resources since it needs to implement the offloading scheme for all the nodes. As what will be discussed later, we use the mathematic tool proposed in reference [30] to solve the optimization problem, whose running time is polynomial. Thus the computational complexity of our scheme is also polynomial, which can be achieved without too much difficult. 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]. Also, the central controller needs to collect requests from all users, which is concerned with the problem of privacy. It is assumed that the central controller works in cooperation with the content servers. Thus it has access to the content requested by the individual vehicles. Specifically, our offloading scheme mainly deals with the commonly-popular contents such as news, which are not concerned with private information. For the private information, it will be transmitted directly from the content server via the cellular network.

B. Content Offloading and Networking Modelling

In our system, users are vehicles and, therefore, we do not distinguish the term "user" and the term "vehicle". In this proposed VDTN-based mobile data offloading system, some chosen vehicles or road side communication equipments, referred to as *helpers*, will participate in the offloading. Incentives for these users can be given by using some micro-payment scheme, or the operator can offer the helper a reduced cost for the service or better quality of service [20]–[22]. A full analysis of such incentives is beyond the scope of this paper. Basically, the service provider chooses some users that are willing to participate in data offloading, and transmits mobile data to these chosen

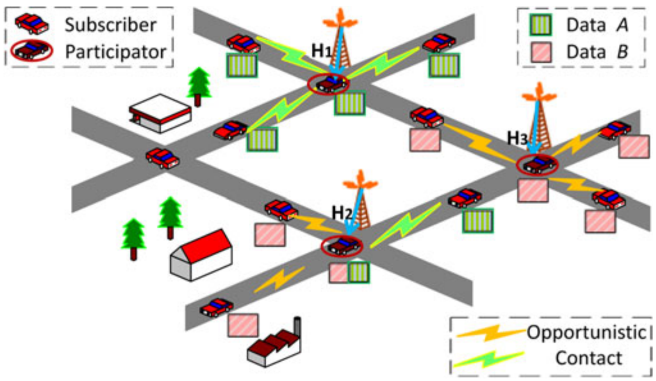


Fig. 2. Mobile data offloading in opportunistic vehicular networks.

users through the cellular network. Then these users further propagate the data to other users that are interested in them by short range device-to-device communication. As mentioned before, for those users that have not received their requested data from the helpers after some "tolerable" duration which is related to the data lifetime, the cellular network can opt to directly send the data to them. For application of news or entertainment, the lifetime may be relatively long since these applications are not time sensitive. For applications of safety or accident information, the lifetime can be controlled within the requirements of delivery time to ensure timeliness of information.

Note that even when users have additional storage and energy to help with the data offloading, some of them may not be willing to act as helpers due to their selfish behaviours or privacy concerns. From the service provider's perspective, on the other hand, even if many users want to be helpers, it can only choose a limited number of them as helpers, as the provider needs to pay directly or indirectly to these users. The exiting incentive schemes [20]–[22] are available to select vehicles in the system to act as offloading helpers. Hence, in such a system, all the vehicles can be divided into the two groups of offloading helpers and mobile data subscribers, respectively. Specifically, the helpers participate the mobile data offloading and disseminate the data to the subscribers. We note that the cellular network can always deliver the data directly to the helpers to meet their requests. Thus, in our investigation we only consider the data requests from the subscribers, and the goal of the mobile data offloading system is to satisfy the subscribers' interests.

Fig. 2 illustrates our opportunistic vehicular network based mobile data offloading scheme. Mobile data items are first transmitted to the helpers where they are buffered in the helpers' local storages. Then, when the helpers meet other vehicles, they transmit the data to those users who are interested in these data items. In our system, we consider multiple data offloading, where the system disseminates multiple data items, and a helper may need to store more than one packet, depending on its buffer size. Furthermore, a user may also be interested in different data items. For example, in Fig. 2, the service provider transmits mobile data A to helpers H_1 and H_2 as well as mobile data B to helpers H_2 and H_3. Other vehicles can obtain the data from these helpers according to their subscribing interest to data by the opportunistic communication paradigm.

More specifically, in our proposed offloading system, there are $S + H$ mobile users, labelled as $i \in \{1, 2, \cdots, S + H\}$. These mobile users are moving vehicles or roadside equipments. Since in reality 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 the lifetime T is assigned, which means that all the helpers will stop disseminating the mobile data c after the deadline T. We further use $\mathcal{H}$ to denote the set of helper users that are willing to participate in the offloading and use $\mathcal{S}$ to denote the subscriber users, where $|\mathcal{H}| = H$ and $|\mathcal{S}| = S$. For the offloading helpers, the system requires their storages to buffer the mobile data, which may include multimedia content of very large size, such as movies. Even the 32 GB of storage available on current devices can only store a limited amounts of video content. Furthermore, it is impossible for a user to contribute all its storage for offloading because of the privacy and energy considerations. Therefore, we should take the storage that each helper is willing to share as our constraint, which directly influences the number of data items that can be stored. Considering this realistic condition, we assume that helper h, $h \in \mathcal{H}$, can at most buffer L_h size of data items.

C. Contact Modelling

Vehicles can communicate with each other only when they move to within the transmission range, which is referred to as a *communication contact*. During the communication contact, nodes can transmit the mobile data in the rate of v bytes per second. We assume that the communication contact between vehicles i and j obeys the Poisson process with contact rate $\gamma_{i,j}$. Poisson distributed contact rate has been observed from real vehicular traces and is widely used to model opportunistic vehicular systems [23], [24]. Although the Poisson contact rate assumption is a little strong compared to the empirical justifications on them in [23], [24], it enables the tactical analysis for the optimal resource allocation solutions, which will be demonstrated by the performance evaluation that compares our proposed optimal solution under this assumption with a wide range of existing schemes. The mobility of a helper has much more significant influence on the mobile data offloading system considered than a subscriber, and different helpers have different mobility patterns. For simplification, we assume that the contact rate of one specific helper $h \in \mathcal{H}$ and any subscriber is the same: $\gamma_{h,s} = \gamma_h$, $\forall s \in \mathcal{S}$. For vehicles that travel on predetermined routes and schedules, there exist many similarities between the mobility patterns of these vehicles, especially for people travelling by cars to and from the work in the same area. Thus this simplified model does not lose too much generality while simplifying the derivation of the optimal resource allocation solutions, which is also demonstrated by the performance evaluation. Alternatively, γ_h may be defined as the average contact rate of helper $h \in \mathcal{H}$ with the subscriber set $\mathcal{S}$: $\gamma_h = E_{s\in\mathcal{S}}\big[\gamma_{h,s}\big]$. Therefore, we define the contact rate of helper $h \in \mathcal{H}$ by γ_h and γ_h will be different for different helpers $h \in \mathcal{H}$.

Because a contact duration is very limited in vehicular networks, a complete mobile data may not be transmitted in one contact. To better understand the VDTN based mobile data offloading problem, an accurate modelling of the contact duration of vehicular networks is critical. In the literature, there have been some studies on the characteristics of the contact duration in human mobility environments by empirical results [25], [26]. These studies validated that the contact duration of human mobility follows a power law distribution, and the Pareto distribution can usually be used to model it. However, there exists no result reported on the vehicular environment. If we investigate the VDTN mobile data offloading problem by simply adopting the contact duration model of human mobility [3], [9], we may obtain either over-pessimistic or over-optimistic results due to the significant differences between human and vehicular mobilities. Therefore, it is necessary to understand the true characteristics of the contact duration in opportunistic vehicular data transmission. By collecting the real mobility traces from about 27 000 operational taxis during one month in Beijing city, which records the key mobility features of urban vehicles in a large city as well as using the existing mobility traces recorded in Shanghai city [27], we analyse the vehicular contact duration characteristics [28]. Our study finds that different from the existing results for human mobility, in vehicular mobility, 80% of the total distribution of contact duration obeys an exponential distribution. Therefore, in this study, we assume that the contact duration for the helpers to transmit the content to the subscribers follows an exponential distribution with parameter λ.

D. Interests Modelling

We now characterise the subscriber behaviours in accessing to different data items in a mobile data offloading system. In a system with multiple data items, a subscriber will have different interests in different data items, and different subscribers will have different dynamic behaviours. Moreover, some data items are popular data that are interested by many subscribers, while some other data items are not popular data which may only be interesting to a small number of subscribers. In this work, we describe the subscriber's interests to different mobile data by a subscriber profile, and model the popularity of mobile data by an interest distribution on key words, which is a widely used approach to model the user interest distribution on different data items in diverse applications [29]. 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 v_{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 $\sum_{k_c \in \mathcal{K}_c} v_{k_c} = 1$. To model the interests of different subscribers on different data, we define P_s^k as the degree of how subscriber $s \in \mathcal{S}$ is interested in keyword $k \in \mathcal{K}$. In this way, we can compare the interests of subscriber s to two different keywords $k_1, k_2 \in \mathcal{K}$ by $P_s^{k_1}$ and $P_s^{k_2}$. Thus, the interest profile of subscriber s is defined by the set $\mathcal{P}_s = \{P_s^k : k \in \mathcal{K}\}$. Without loss of generality, we assume $\sum_{k \in \mathcal{K}} P_s^k = 1$. Finally, the interest probability of subscriber $s \in \mathcal{S}$ in mobile data $c \in \mathcal{C}$, defined by $w_{s,c}$, can be obtained as follows:

$$w_{s,c} = \sum_{k_c \in \mathcal{K}_c} v_{k_c} P_s^{k_c}. \tag{1}$$

E. Problem Formulation

Denote $\mathbf{X} = (x_{h,c})$, $h \in \mathcal{H}$ and $c \in \mathcal{C}$, as the storage allocation policy, in which $x_{h,c} \in \{0, 1\}$ and $x_{h,c} = 1$ indicates that helper h stores item c in its buffer. Since a lifetime T is assigned to each data item, if subscribers do not receive the required item from helpers after the lifetime is expired, they will receive it directly from the wireless infrastructure through cellular networks, which means that this mobile item is not offloaded or in other words the mobile data offloading system does not meet subscribers' interest. Therefore, the objective of the mobile data offloading system is to maximise the expected overall interests satisfaction of all the subscribers. Obviously, this objective function depends on the storage allocation policy $\mathbf{X}$ and is denoted as $U(\mathbf{X})$.

Thus, maximising the system's expected interests satisfaction for all the mobile data items and over all the subscribers can be specified as the following optimization problem:

$$\begin{aligned} &\max U(\mathbf{X}) \\ &\text{s.t. } x_{h,c} \in \{0,1\}, \forall h \in \mathcal{H}, c \in \mathcal{C} \\ &\text{and} \sum_{c \in \mathcal{C}} l_c x_{h,c} \le L_h, \forall h \in \mathcal{H} \end{aligned} \tag{2}$$

where $\sum_{c \in \mathcal{C}} l_c x_{h,c} \le L_h$ is the buffer size constraint of dissemination helper h. To solve this optimisation problem, we use the mathematic tool proposed by Kulik [30], which develops a general mathematic method for maximising submodular functions subject to linear constraints.

Here, we summarize the commonly used variables throughout the paper in Table I.

IV. Problem Analysis and Algorithm Design

A. Optimisation Problem Analysis

In the last section, we have formulated the optimisation problem of $\max U(\mathbf{X})$ with the buffer size constraints, where $\mathbf{X}$ denotes the storage allocation policy and $U(\mathbf{X})$ denotes the objective function, depending on the storage allocation policy $\mathbf{X}$. In the optimisation problem, we have defined the storage allocation policy that need to be optimised and the buffer size constraints when performing the optimisation. However, the explicit expression for $U(\mathbf{X})$, the objective of the optimisation, has not been defined yet. Thus, in order to solve the optimisation problem (2), we need the explicit expression for $U(\mathbf{X})$. We note that $U(\mathbf{X})$ depends on the system dynamics of communication contact as well as the storage allocation and buffer size in the helpers. Recall that our goal is to maximise the expected interest satisfaction. Let us first define $Q_{s,c}$ as the probability that subscriber s has successfully received the mobile data c before its lifetime. Then, we can express the objective

TABLE I
LIST OF COMMONLY USED VARIABLES THROUGHOUT THE PAPER

Variable	Description
$\mathcal{S}$ and S	The set and the number of subscriber users.
$\mathcal{H}$ and H	The set and the number of offloading helpers.
$\mathcal{C}$ and C	The set and the number of mobile data items.
$\mathcal{K}$ and K	The set and the number of keywords.
l_c	Length of data item c, $c \in \mathcal{C}$.
T	Lifetime of date items.
L_h	Buffer size of helper h, $h \in \mathcal{H}$.
v	Mean mobile data transmission rate.
$\gamma_{i,j}$	Communication contact rate of vehicles i and j.
γ_h	Communication contact rate of helper h, $h \in \mathcal{H}$, with subscriber.
λ	Exponential distribution parameter of contact duration.
P_s^k	The degree of how subscriber $s \in \mathcal{S}$ is interested in keyword $k \in \mathcal{K}$.
v_{k_c}	The importance of keyword k_c.
$w_{s,c}$ and W_c	Interest probability of subscriber $s \in \mathcal{S}$ in mobile data $c \in \mathcal{C}$, $W_c = \sum_{s \in \mathcal{S}} w_{s,c}$.
$x_{h,c}$	$x_{h,c} \in \{0, 1\}$ and $x_{h,c} = 1$ indicates that helper h stores item c in its buffer.
$U(\mathbf{X})$	System objective function depends on the storage allocation policy $\mathbf{X}$, $\mathbf{X} = (x_{h,c})$.
$Q_{s,c}$ and $G(c)$	Probability that subscriber s has successfully received the mobile data c before its lifetime, $G(c) = Q_{s,c}$.
D_c	The accumulated contact duration that subscriber s encounters the helpers with data c.
z_h	The accumulated contact duration that subscriber s encounters helper h before the data lifetime T.
d_c	System allocated communication resource for mobile data c.
r_i	The duration time of the ith contact.
t_c	The time that is needed for subscriber s to receive the mobile data c.
$\widehat{G}(d_c)$	Rewritted function of successful receiving probability $G(c)$.
d_{used}	The allocated resource running in the processing of resource allocation algorithm.
d_{all}	The total communication resource.

function $U(\mathbf{X})$ as

$$U(\mathbf{X}) = \sum_{s \in \mathcal{S}} \sum_{c \in \mathcal{C}} w_{s,c} Q_{s,c} \tag{3}$$

by explicitly combining the subscribers' interests. In our system, a helper will contact any subscriber with the same contact rate. Therefore, we can drop the index s in $Q_{s,c}$ and re-define it as $G(c) = Q_{s,c}, \forall s \in \mathcal{S}$. Then, the objective function (3) becomes

$$U(\mathbf{X}) = \sum_{c \in \mathcal{C}} W_c G(c) \tag{4}$$

where $W_c = \sum_{s \in \mathcal{S}} w_{s,c}$.

For a subscriber s to successfully obtain the mobile data c, it must encounter the helpers with the data c for sufficient time in order to receive the whole mobile data successfully. We denote the accumulated contact duration that subscriber s encounters the helpers with data c as D_c. Based on the definition of $G(c)$, we have

$$G(c) = P\{D_c \cdot v \geq l_c\} = P\left\{ \sum_{h \in \mathcal{H}} x_{h,c} z_h \geq \frac{l_c}{v} \right\} \tag{5}$$

where $P\{\bullet\}$ denotes the probability, v is the mean data transmission rate, and the variable z_h is the accumulated contact duration that subscriber s encounters helper h before the data lifetime T. Since our object is to complete the transmission of data c to subscriber s before lifetime T, the exact data transmission rate at any time is not important and what we care more is the mean data transmission rate. As long as the total transmission during the contact duration exceeds the data size, the data item can be successfully disseminated to the subscriber. For this reason, we consider the mean data transmission rate in our work. We note that subscriber s encounters helper h with the Poisson contact rate of γ_h, and this Poisson encounter event is independent. Therefore, the accumulated time that subscriber s encounters all the helpers with data c follows the Poisson distribution with parameter

$$d_c = \sum_{h \in \mathcal{H}} x_{h,c} \gamma_h T.$$

As expected, d_c depends on $x_{h,c}$, the buffers allocated to mobile data c, and γ_h, the contact rates of helpers. The parameter d_c is the mean meeting times that subscriber s encounters all the helpers with data c before the data lifetime T. Since we assume that the contact duration for the helpers to transmit the content to the subscribers follows an exponential distribution with parameter λ, the probability that data c can be successfully offloaded is mainly determined by the meeting times. Consequently, we refer to d_c as the system allocated communication resource for mobile data c. Therefore, d_c, $c \in \mathcal{C}$, are the equivalent decision variables to $\mathbf{X}$, and this allows us to transform the original optimal storage allocation problem (2) into our optimal communication resource allocation problem to be given explicitly blow. First, we can have the following expression:

$$G(c) = P\left\{ \sum_{h \in \mathcal{H}} x_{h,c} z_h \geq \frac{l_c}{v} \right\} = P\left\{ \sum_{i=1}^{N(d_c)} r_i \geq \frac{l_c}{v} \triangleq t_c \right\} \tag{6}$$

where $N(d_c)$ is the Poisson distribution with parameter d_c, r_i is the duration time of the ith contact, and $t_c = \frac{l_c}{v}$ is the time that is needed for subscriber s to receive the mobile data c. We note that $\{r_i\}$ are independently identically exponentially distributed with parameter λ. Therefore, $\sum_{i=1}^{N(d_c)} r_i$ is a compound Poisson process. As is well-known, we cannot obtain an explicit and exact expression for the compound Poisson process. Therefore, we use some approximation method to obtain $G(c)$.

Lemma 1: The normal distribution with the expectation $\frac{d_c}{\lambda}$ and the variance $2d_c\lambda^2$, denoted as $N\left(\frac{d_c}{\lambda}, 2d_c\lambda^2\right)$, is a reasonable approximation to the distribution of $\sum_{i=1}^{N(d_c)} r_i$ when λ is large.

See the appendix for proof.

Since the contact duration time is very limited in VDTN, consequently, the exponential parameter λ is relatively large. According to Lemma 1, we can obtain the expression of $G(c)$

as follows:

$$G(c) = \int_{t_c}^{\infty} \frac{1}{\sqrt{2\pi 2 d_c \lambda^2}} e^{-\frac{(x - d_c/\lambda)^2}{4 d_c \lambda^2}} dx$$
$$= \int_{\frac{t_c}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(x - \frac{\lambda\sigma}{2}\right)^2}{2}} dx = 1 - \phi\left(\frac{t_c}{\sigma} - \frac{\sigma\lambda}{2}\right), \quad (7)$$

where $\sigma = \sqrt{2 d_c \lambda^2}$, $d_c = \frac{\sigma^2}{2\lambda^2}$, and $\phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$. Noting that $G(c)$ is a function of d_c with $c \in \mathcal{C}$, we rewrite this function as $\widehat{G}(d_c)$, and further denote the objective function $U(\mathbf{X}) = \sum_{c\in\mathcal{C}} W_c \widehat{G}(d_c)$ as $\widehat{U}(\mathbf{d})$, where $\mathbf{d} = \left(d_c\right)$, $c \in \mathcal{C}$, is the decision variable vector that contains all the elements d_c for $c \in \mathcal{C}$. Therefore, we can transfer the optimization problem (2) into the following form:

$$\begin{aligned} &\max\ \widehat{U}(\mathbf{d}) = \sum_{c\in\mathcal{C}} W_c \widehat{G}(d_c) \\ &\text{s.t.}\quad d_c \in \mathbb{R}, \forall c \in \mathcal{C};\ \text{and} \sum_{c\in\mathcal{C}} d_c \le d_{\text{all}} \end{aligned} \quad (8)$$

where the constant

$$d_{\text{all}} = \sum_{c\in\mathcal{C}} \sum_{h\in\mathcal{H}} x_{h,c}\gamma_h T = \sum_{h\in\mathcal{H}} \gamma_h T \sum_{c\in\mathcal{C}} x_{h,c} = \sum_{h\in\mathcal{H}} \gamma_h T L_h$$

is obtained by taking the equality sign in the inequality $\sum_{c\in\mathcal{C}} x_{h,c} \le L_h$. We refer to d_{all} as the total communication resource. Then d_c is the part of the communication resource allocated to mobile data c. In the above optimisation problem, we note that the decision variables $\left(d_c\right)$ are related to the storage allocation policy $\left(x_{h,c}\right)$ and the communication contact rates γ_h.

For the optimisation problem (8), we observe that we need to decide how to optimally allocate the communication resource d_{all} to each mobile data. We have the following lemma revealing the property of the objective function in the optimization (8).

Lemma 2: $\widehat{U}(\mathbf{d})$ is a monotonically increasing concave function of $\mathbf{d}$ when $d_c \ge \lambda t_c, \forall c \in \mathcal{C}$.

See the appendix for proof.

B. Algorithm for Optimal Contact-Aware Resource Allocation

Based on Lemmas 1 and 2, we now design an optimal algorithm, summarised in Algorithm 1, to solve the optimisation problem (8). This algorithm uses watering approaching which is widely used in solve concave function maximisation problem. We have proved in the appendix that the optimal solution $\mathbf{d}$ meets the condition that $\widehat{G}'(d_c) \cdot W_c = \text{Constant}$ if $d_c > 0$. In order to find the value of the constant, we use the bisection procedure to obtain the optimal results through iteration. We first initiate the upper bound Φ_{up} and the lower bound Φ_{down} of the constant. In each iteration, we let the temporary constant $\Phi_{\text{mid}} = \frac{\Phi_{\text{up}} + \Phi_{\text{down}}}{2}$. Then, for all data $c \in \mathcal{C}$, we calculate the value of d_c if $\widehat{G}'(d_c) \cdot W_c = \Phi_{\text{mid}}$ can be satisfied, or we set $d_c = 0$. After obtaining all values of d_c, if the used communication resource is larger than the total communication resource, it means the optimal constant is smaller than Φ_{mid}, and we set $\Phi_{\text{up}} = \Phi_{\text{mid}}$, or we set $\Phi_{\text{down}} = \Phi_{\text{mid}}$. Using the updated upper bound and lower bound, we continue the iteration until the

Algorithm 1: Optimal Resource Allocation Scheme, where the system utility function $\widehat{G}(d_c)$ depends on the contact rates and durations.

Initial $\Phi_{up} = \max\{W_c\}$, $\Phi_{\text{down}} = 0$, $\Phi_{\text{mid}} = 0$ and precision parameter $\varpi = 10^{-3}$
while $|\Phi_{up} - \Phi_{\text{down}}| > \varpi$ **do**
  $\Phi_{\text{mid}} = \frac{\Phi_{up} + \Phi_{\text{down}}}{2}$, $d_{\text{used}} = 0$
  for $c \in \mathcal{C}$ **do**
    if $\Phi_{\text{mid}} \le \widehat{G}'(d_c)|_{d_c \text{ is inflexion point}} \cdot W_c$ **then**
      Calculate d_c, s.t. $\widehat{G}'(d_c) \cdot W_c = \Phi_{\text{mid}}$
      $d_{\text{used}} = d_{\text{used}} + d_c$
    else
      $d_c = 0$
    end if
  end for
  if $d_{\text{used}} > d_{\text{all}}$ **then**
    $\Phi_{\text{mid}} = \Phi_{\text{up}}$
  else
    $\Phi_{\text{mid}} = \Phi_{\text{down}}$
  end if
end while

results meet the precision requirement and obtain the optimal solutions. The optimality of Algorithm 1 is proved in the following theorem.

Theorem 1: The solution obtained by Algorithm 1 is the optimal solution for the optimization problem (8).

See the appendix for proof.

C. Heuristic Algorithm for Buffer Allocation

We have now obtained the optimal contact-aware communication resource allocation solution. For each data $c \in \mathcal{C}$, we have now obtained the allocated communication resource d_c. The allocation of d_c depends on the buffer allocation policy. However, for the same allocation of d_c, the buffer allocation policy can be different as long as the total meeting times reach the requirement of d_c. Thus, based on this optimal communication resource allocation result, we now need to allocate the mobile data to the storages or buffers of helpers. Consider the optimisation problem defined in (2), we can express the corresponding storage allocation problem of deciding $\{x_{h,c}\}$ according to the given $\{d_c\}$ as the following problem:

$$\left\{ x_{h,c}, c \in \mathcal{C}, h \in \mathcal{H} \,\middle|\, \sum_{h\in\mathcal{H}} x_{h,c}\gamma_h = d_c/T, c \in \mathcal{C}; \sum_{c\in\mathcal{C}} l_c x_{h,c} \le L_h, h \in \mathcal{H} \right\}.$$

This problem is to decide the $H \times C$ variables $\{x_{h,c}\}$ according to the C known variables $\{d_c\}$. Thus, it is an indeterminate problem or in other words, there exists many possible solutions.

Since the constraints of this problem are loose, heuristic algorithm that can quickly obtain a solution works well here. For example, we can use the helper with the largest remaining buffer size to store the data with the largest size first, while meeting the constraint for the corresponding d_c. That is, we loop the helpers whose remaining buffer sizes are the largest to choose the largest data to buffer until all the helpers' buffers are full, while meeting the constraints for $\{d_c\}$ all the times. We refer to this buffer allocation scheme as the *heuristic algorithm* for buffer allocation.

In the next section of performance evaluation, we will simulate different buffer allocation schemes based on the optimal resource allocation result obtained by Algorithm 1 to investigate the influences of different buffer allocation schemes on the overall achievable system performance.

V. Performance Evaluation

Under the environment of MATLAB (Matrix Laboratory) [45], we compared the performance of our proposed contact-aware optimal resource allocation scheme, which considers the contact duration in solving the optimal resource allocation problem, with the performance of the several existing schemes, including some contact *Duration Aware* schemes, denoted by DA for short, which also explicitly take into account the contact duration in obtaining solutions, as well as some contact *Duration Unaware* schemes, denoted by DU for short, which do not take into account the contact duration in obtaining solutions. Specifically, we compared the following DA schemes.

1) *DA-Optimal-I:* This is our optimal contact-aware network resource allocation design, namely, the system allocates the communication resources according to Algorithm 1. Given the optimal communication resource allocation solution, the heuristic algorithm described in Section IV-C is used to allocate the mobile data to the buffers of the helpers.
2) *DA-Optimal-II:* This is also our optimal contact-aware network resource allocation design. However, given the optimal communication resource allocation solution obtained by Algorithm 1, it adopts a random buffer allocation scheme that randomly chooses the mobile data to buffer at each helper, while meeting the constraints for $\{d_c\}$.
3) *DA-DARA [9]:* As mentioned in the introduction part, the contact Duration Aware Replication Algorithm (DARA) scheme [9] operates in a fully distributed manner to allocate the buffer for each mobile data, and it fails to obtain the optimal system solution. It is the latest existing scheme that explicitly considers the contact duration in design making.
4) *DA-Greedy [30]:* By taking into account the contact duration, the system allocates the buffer for the mobile data using a heuristic based greedy algorithm proposed in [30]. More specifically, this scheme runs a *suboptimal* greedy algorithm based on our system utility function that depends on the contact duration.
5) *DA-Proportion:* In this scheme, the number of the stored copies of each data is uniformly proportional to its interests W_c, and the data with higher W_c are stored into the helpers that have longer contact durations with the subscribers.

We also compared our DA-Optimal-I and DA-Optimal-II with the following DU schemes.

1) *DU-Optimal:* This scheme ignores the contact durations by setting the contact durations of all node pairs to infinity. It then adopts our optimal Algorithm 1 to obtain the "optimal" communication resource allocation solution and then uses the same heuristic buffer allocation scheme presented in Section IV-B to place the mobile date into the helpers' buffers.
2) *DU-Greedy [31]:* The system allocates the buffers for the mobile data by a heuristic based greedy algorithm [31], which represents the most up-to-date work in the area of DTN offloading without explicitly considering the contact duration.
3) *DU-Random:* This scheme does not consider the contact duration at all. Each helper simply chooses the data items randomly to fill its buffer until no more item can be stored.

It can be seen that the benchmark schemes chosen include the recently proposed scheme that takes into account the contact duration (DA-DARA) as well as some most up-to-date heuristic schemes. Our comparative simulation study therefore is capable of offering an objective evaluation for our optimal duration awareness solutions (DA-Optimal-I and DA-Optimal-II).

A. Simulation System Setup

1) Vehicular Mobility Traces: Our evaluation was conducted on the two real vehicular mobility traces, *Shanghai* trace [24] and *Beijing* trace, which recorded the positions of vehicles carrying GPS devices. Specifically, about 2100 operational taxis were running for the whole month of February 2007 in Shanghai city to collect *Shanghai* trace [24]. In collecting *Beijing* trace, we used mobility track logs obtained from 27 000 participating Beijing taxis carrying GPS receivers during the whole May month in 2010. *Beijing* trace is the largest vehicular data trace available. Generally speaking, private vehicles move more regularly than taxis, since private vehicles usually follow fixed routes from home to work while taxis tend to move randomly. Thus the property of the taxis network is more complex due to its randomness. For private vehicles, it is easier to implement the data offloading scheme because we can obtain the routes of private vehicles in advance and do the data offloading between vehicles that have the same route. Also, the data trace of private vehicles is difficult to be obtained because of privacy problem and the trace from taxis is the only data trace of large scale that can be obtained for research. Since Beijing and Shanghai are two typical big cities worldwide, the results obtained from these two traces can reveal the common property of the typical city traffic worldwide. More specifically, we utilised the GPS devices to collect the taxi locations and timestamps and GPRS modules to report the records every one minute for moving taxis. Then, the real GPS locations of individual vehicles, as a function of time, were used to simulate the offloading using the proposed allocation algorithm. For any pair of vehicle X and Y, the contact began when they moved into the communication range and ended when they moved out of the communication range according to their real-time positions. The real mobility traces are quite complex, and we cannot use any simple distribution to depict them. This is why we use assumption and approximation when doing the theoretic analysis to obtain our offloading scheme. We now applied the scheme to the real mobility traces to evaluate our scheme. If the system performed well, it can prove that the assumption and approximation are reasonable and our offloading scheme is effective and valid for real mobility traces.

In both vehicular traces, there were some nodes that rarely met with others. For these nodes, most of their data still need to be obtained via cellular networks. However, there still exist a large number of nodes that contact with each other frequently. The objective of our scheme is to cope with this explosive traffic demands and mobile data growth by offloading these data traffic from the cellular network to the vehicular opportunistic network. Although we cannot include all the nodes in our scheme, implementing data offloading scheme for these nodes have already relieved considerable burden of the cellular network. Moreover, in big cities like Beijing and Shanghai, the mobility patterns of most vehicles are limited in a small range according to the administration partition. Thus it is more reasonable to consider the problem of data offloading regionally instead of taking the whole network as an entirety. Therefore, we selected the nodes that had frequently contacts with others as the typical example of the vehicles in the same region to evaluate the performance of our proposed scheme. Specifically, for each of these two traces, we selected 200 nodes by ranking their contact times. Among these nodes, the most contacted 50 nodes were set as helpers, and the rest were used as subscribers. In all the experiments, the network simulation of the whole trace was divided into the warm-up period and the offloading period. We used the first half of the trace as the warm-up period for the content controller to accumulate the average contact rates of every helpers based on the contact counts and other necessary network information, such as the contact durations and the contact interval rates of all the helper-subscriber pairs. During the warm-up period, the controller also collected the information of the subscribers' interests, content sizes and buffer sizes from the subscribers, content server and helpers. After collecting all these information, the content controller made the decision on the communication resource allocation using Algorithm 1, and then obtained the mobile data storage or buffer allocation policy. In the process of the data offloading which consisted of the second half of the trace, the mobile data were generated from the content server, stored in the buffers of all the helpers, and offloaded during the offloading period. All the helpers will stop disseminating the mobile data $c \in \mathcal{C}$ after the lifetime T and subscribers will receive it directly from the wireless infrastructure through cellular networks. After the simulation of the entire trace, we collected the system performance results.

2) Mobile Data, Lifetime, Helpers' Buffers and Subscribers' Interest Profiles: In the simulation, there were $C = 30$ mobile data items. According to the data transmission requirements of the emerging vehicular applications of in-vehicle entertainment [39] and augmented reality assisted safety driving [40], [41], the sizes of data were generated randomly and uniformly in the range of $[100\,\text{MB},\ 200\,\text{MB}]$. Each data item was assigned a lifetime T randomly and uniformly in the range of [600 s, 3600 s]. The helpers' buffer sizes were randomly and uniformly generated in $[0,\ 2l_\text{a}\ \text{GB}]$, where l_a was the average buffer size. It might happen that a helper has a very high contact rate but a very small buffer size. To avoid accidents, we will average our results over 30 runs. Thus the influence of the extreme case will be little. In order to define the subscribers' interests, we set the number of keywords to $K = 35$, and assumed that the keywords k_1 to k_{35} were ranked by their popularity. For each data item c_i, $i \in \{1, 2, \cdots, 30\}$, its describing subset of keywords was given by $\mathcal{K}_{c_i} = \{k_i, k_{i+1}, \cdots, k_{i+4}\}$ with equal weight $v_{k_i} = 1/5$ for each keyword. We used the normal distribution to generate the interest profile $\mathcal{P}_s$ for subscriber s. All of these parameters above will stay the same in all the later simulations if no special explanation is provided. For keyword k_i, we assumed that the average interest of all the subscribers was I_i, and used the following three distributions to generate I_i in order to obtain different interest distributions.

1) Exponential distribution: $I_i = \frac{e^{-i}}{\sum_{j=1}^{35} e^{-j}}$. In this case, most of the subscribers' interests concentrated on the popular data.
2) Zipf distribution with exponent 2: $I_i = \frac{1/i^2}{\sum_{j=1}^{35} 1/j^2}$. In this case, most of the subscribers' interests also concentrated on the popular data, but the difference between high popular data and low popular ones was smaller than the case of exponential distribution.
3) Uniform distribution: $I_i = 1/35$. In this case, the subscribers' interests were uniformly distributed among the data.

We simulate the mobile data offloading system with these different interest distributions to verify the efficiency of our proposed optimal solutions under different interest distribution settings.

B. Contact and Data Transmission

In the data offloading process, the communication range between vehicles was set to 200 m, which is the typical setting for using the current Dedicated Short Range Communication (DSRC) technology [32]. The accurate number of transmission range is not the problem considered in this paper, which is the work of physical layer. For different transmission range, we can use the same optimal scheme for mobile data offloading. Under this setting, when a helper vehicle and a subscriber vehicle, traveling according to the GPS trajectories recorded in the trace, move into the communication range of 200 m, the mobile data dissemination occurs if the subscriber is interested in any data stored in the helper. Referring to [42], 802.11p (WAVE) is currently considered for Dedicated Short Range Communications (DSRC), which works in 5.86-5.925 GHz and has the transmission rate of 27 Mbps within the range of 300 m. Based on this, the mean mobile data transmission rate was set to $v = 20$ Mbps, and the amount of the transmitted data in one contact depended on the contact duration. When a subscriber meets a helper with the data that it is interested in, it may only receive part of the data item, since the contact duration is limited. In this case, this subscriber will store the received amount of the data item in

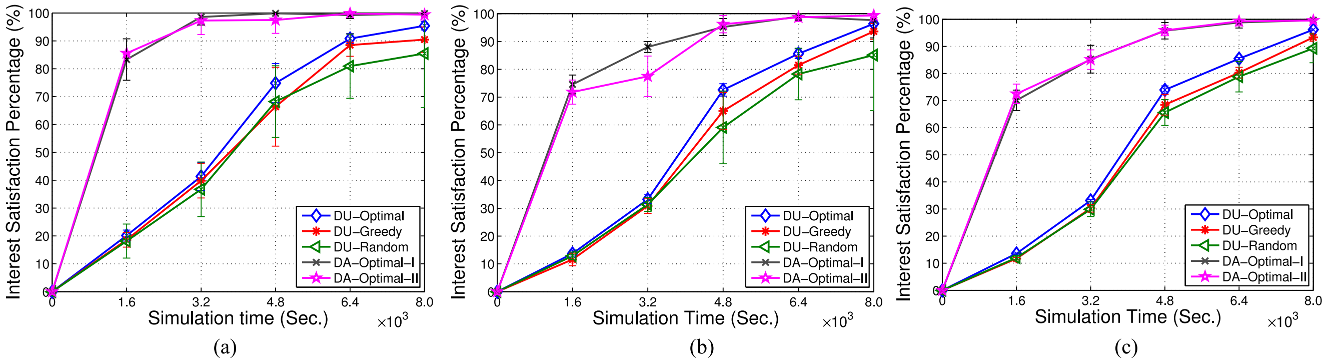


Fig. 3. Performance comparison of our two DA optimal schemes with three existing DU schemes for *Beijing* trace with the average buffer length of 3.5 GB, where the results are averaged over 30 runs, and the vertical bars indicate the related standard deviations. (a) Exponential interests. (b) Zipf interests. (c) Uniform interests.

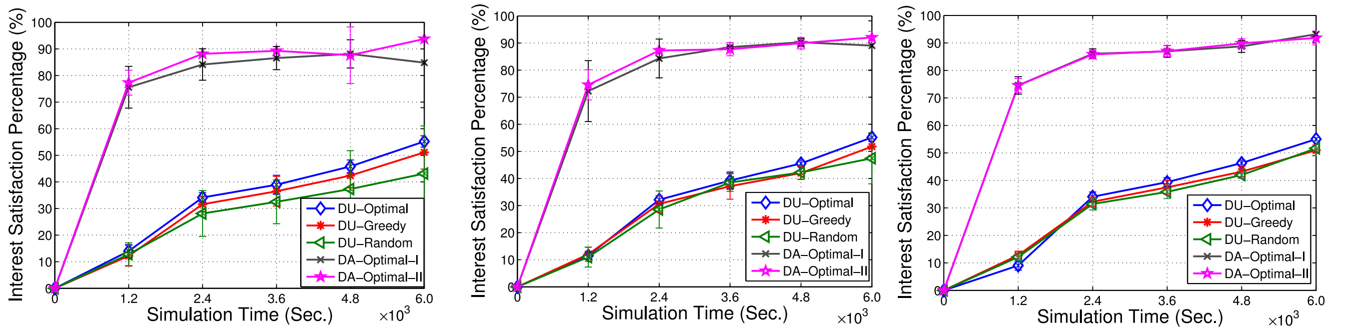
its buffer, and then requests the remaining part of the data item from other helpers when new contacts occur. When a helper with more than one data meets a subscriber who are interested in these data, the data will be transmitted in the order of their interests to the subscriber.

C. Result Analysis

1) Comparison of Our DA Optimal Schemes with Existing DU Schemes: To demonstrate the efficiency of our two proposed DA optimal schemes, the DA-Optimal-I and DA-Optimal-II, over the existing DU schemes, the DU-Optimal, DU-Greedy and DU-Random, we first set the average buffer length to $l_{\mathrm{a}} = 3.5$ GB and randomly and uniformly generated the helpers' buffer sizes in [0, 7 GB]. In this simulation, the relatively large buffers were set for the helpers. The subscribers' interest satisfaction results obtained by these five schemes for *Beijing* trace are shown in Fig. 3, where the results were averaged over 30 simulation runs by simulating the system with the buffer allocation strategy of each algorithm. As expected, the interest satisfaction percentage increased as the simulation or offloading time increased, since more packets were disseminated to the subscribers owing to more communication contacts occurring. Specifically, Fig. 3 (a) presents the results for the case of exponentially distributed interests, where it can be seen that our two DA optimal schemes outperformed the three DU schemes considerably. At the very beginning of the system offloading process, such as 1 minute, the interest satisfaction percentage was less than 10%, which is relatively quite low. Thus our offloading scheme mainly deals with the non-realtime mobile data that is delay-tolerant. When the system offloading time reached 1.6×10^3 s, about 20% of the total simulation time, the performance of the three DU schemes were only about 20%, while our two DA optimal schemes already met about 85% interest satisfaction, which represented improving the system performance by 4 times over the three DU benchmarks. When the simulation time reached 4.8×10^3 s, about the three fifths of the total simulation time, our two schemes almost satisfied all the subscribers' interests, while the DU-Optimal, DU-Greedy and DU-Random schemes only met about 75%, 68%, and 69% of interest satisfactions, respectively. When the offloading ended at the simulation time of 8×10^3 s, the DU-Optimal achieved about 95% interest satisfaction, higher than 90% and 85% achieved by the DU-Greedy and DU-Random schemes, respectively. The helpers' buffer spaces were ample in this simulation, and our two optimal schemes achieved 100% interest satisfaction well before the end of the offloading time. Additionally, our two optimal schemes could meet a required level of performance much earlier than the three DU benchmarks. Therefore, we can conclude that our two DA optimal schemes significantly enhanced the data offloading performance, in terms of subscribers' interests satisfaction percentage and offloading delay.

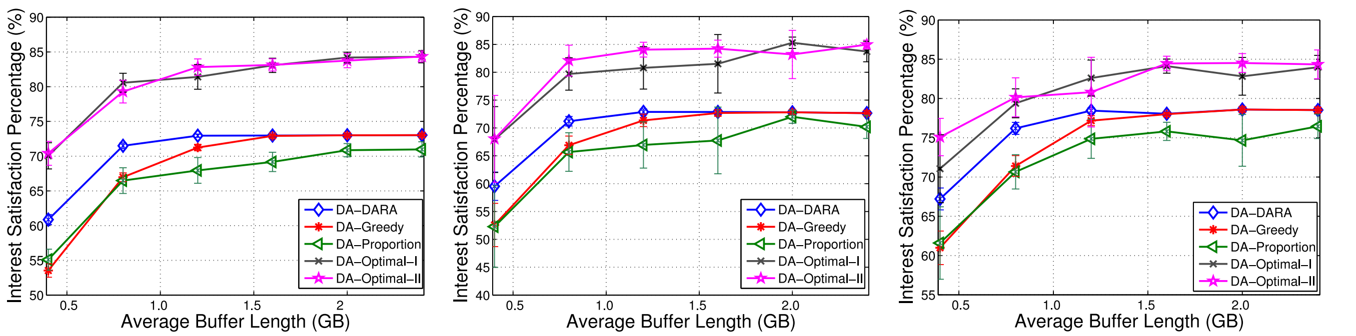
Similar observations can be drawn from the results under the Zipf and uniform distributed interests, depicted in Fig. 3 (b) and (c), respectively. From the results shown in Fig. 3 (a) to (c), we note that our two optimal schemes reached the maximum achievable performance the earliest in the case of exponential distributed interests. This is because the Zipf and uniform distributed interests are more similar, compared with the exponentially distributed ones. We can infer that our algorithms work more efficiently in the case of heterogeneous interest distribution. Additionally, we observe that our two DA optimal schemes attained the similar performance. The difference between the DA-Optimal-I and DA-Optimal-II is in the buffer allocation policy. The both designs are based on the optimal communication resource allocation solution obtained by Algorithm 1. Therefore, we can also infer that the system's achievable performance is mainly determined by the communication resource allocation policy.

We now implement the same analysis for the trace of *Shanghai*. The average vehicle-vehicle contact time and contact rate will differ between different traces. However, the differences of contact time and contact rate will influence the efficiency of data dissemination among the vehicular network. For traces with longer contact time and higher contact rate, the data dissemination will be more efficient. Thus it needs smaller buffer lengths to store the data, and completes the data dissemination earlier. We next set the helpers' average buffer length to $l_{\mathrm{a}} = 2.5$ GB, which is a little smaller than that of *Beijing* trace, and obtained the results for *Shanghai* trace as illustrated in Fig. 4,



(a) (b) (c)

Fig. 4. Performance comparison of our two DA optimal schemes with three existing DU schemes for *Shanghai* trace with the average buffer length of 2.5 GB, where the results are averaged over 30 runs, and the vertical bars indicate the related standard deviations. (a) Exponential interests. (b) Zipf interests. (c) Uniform interests.



(a) (b) (c)

Fig. 5. Performance comparison of our two DA optimal schemes with three existing DA schemes for *Beijing* trace with the offloading time of 8×10^3 s, where the results are averaged over 30 runs, and the vertical bars indicate the related standard deviations. (a) Exponential interests. (b) Zipf interests. (c) Uniform interests.

where again the simulation results were collected over 30 runs by simulating the system with the buffer allocation strategy of each algorithm. Similar observations to those for *Beijing* trace can be drawn, namely, our two DA optimal schemes significantly outperformed the three existing DU schemes, in terms of subscribers' interests satisfaction percentage and offloading delay. We note that the maximum achievable performance of every schemes are lower than those for *Beijing* trace. For example, our two DA optimal schemes achieved maximally about 90% interest satisfaction, instead of 100% as for *Beijing* trace. This indicates that the buffer resources in this case were not as ample as in the case of *Beijing* trace.

2) Comparison of Our DA Optimal Schemes with Existing DA Schemes: We next focused on comparing our two DA optimal schemes, the DA-Optimal-I and DA-Optimal-II, with the three existing DA benchmark schemes, the DA-DARA, DA-Greedy and DA-Proportion. In particular, we concentrated on the influence of the helpers' buffer sizes on the system's achievable performance by varying the average buffer size l_{a} from 0.4 GB to 2.4 GB, while setting the simulation or offloading time to 8×10^3 s and 6×10^3 s for *Beijing* and *Shanghai* traces, respectively. We also used the exponential, Zipf and uniform distributions of interests in the investigation. Note that for the average buffer size of $l_{\mathrm{a}} = 0.4$ GB, the helpers' buffer sizes were randomly and uniformly chosen from [0, 0.8 GB], while for the average buffer size of $l_{\mathrm{a}} = 2.4$ GB, the helpers' buffer sizes were randomly and uniformly generated from [0, 4.8 GB]. Figs. 5 and 6 present the results obtained for *Beijing* and *Shanghai* traces, respectively. From the results shown in Figs. 5 and 6, we observe that our two DA optimal schemes achieved the similar performance and they outperformed the three existing DA schemes considerably, which again demonstrated the optimality of our designs. Also, we can see that the performance gaining of our DA optimal schemes is larger in percentage when the helpers' buffers are small for both the two traces. Therefore, we can infer that our DA optimal schemes perform more efficiently when the system resource is very limited. Among the three existing DA benchmarks, the distributed DA-DARA scheme performed better than the other two schemes, particularly when the system resource in terms of helpers' buffers was very limited. However, when the average buffer size was larger than 1.6 GB, the DA-Greedy algorithm achieved the same performance of the DA-DARA, indicating that the advantage of the distributed DA-DARA solution over the DA-Greedy solution diminished when the buffer resource was relatively ample. The results of Figs. 5 and 6 show that our two DA optimal schemes consistently outperformed the three existing benchmark algorithms over the wide range of

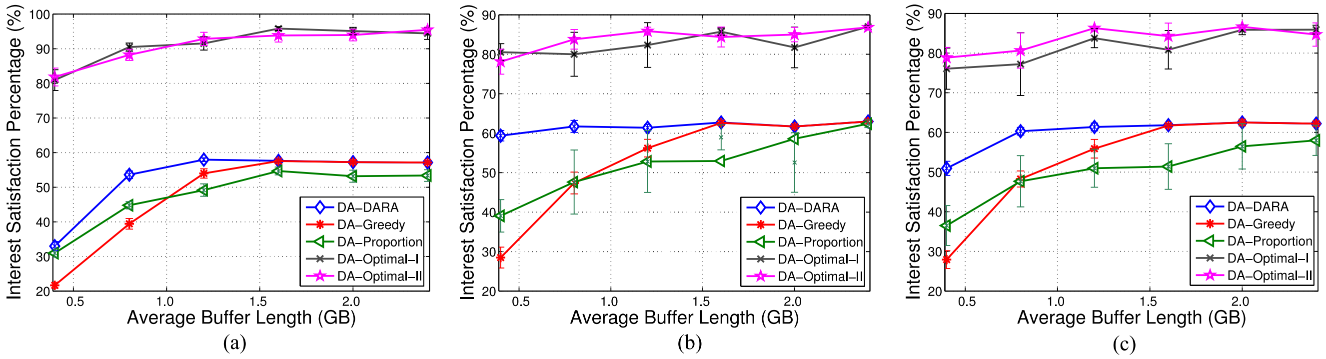


Fig. 6. Performance comparison of our two DA optimal schemes with three existing DA schemes for *Shanghai* trace with the offloading time of 6×10^3 s, where the results are averaged over 30 runs, and the vertical bars indicate the related standard deviations. (a) Exponential interests. (b) Zipf interests. (c) Uniform interests.

average buffer sizes, confirming again the effectiveness of our approach in the environment of heterogeneous helpers' buffer resources.

VI. Conclusion

We have investigated the problem of contact-duration aware optimal mobile data offloading for the cellular network using the vehicular opportunistic network. We have studied this problem in a realistic environment where vehicles' contact rates are heterogeneous and the contact duration time is very limited, as well as the helpers' buffers used in the mobile data offloading are limited and the subscribers have different interests on different data items. By formulating this challenging problem as a utility function maximisation problem, we have designed an optimal and efficient scheme to allocate the network resources in terms of opportunistic contacts and offloading helpers' buffers. Extensive simulation results obtained based on two real large-scale urban vehicular traces have demonstrated that our designs significantly outperform many existing algorithms that are typically used to solve this type of challenging problems, including several existing contact-duration aware schemes and several schemes which do not take into account the contact duration time.

In this work, we replace the individual contact rate of a helper and a subscriber by the average contact rate of a helper to all the subscribers, in order to make analysis and design tractable. Our future work will investigate whether it is feasible to relax this assumption in the formulation of the system expected interest satisfaction optimisation and in the derivation of the optimal solution. Currently, we have only used realistic vehicular mobility traces to simulate the offloading system. Our future work will investigate how to deploy our efficient mobile data offloading schemes in real vehicular network systems in order to evaluate the true system's achievable performance. In our work, we now have only considered the case that all the requests from users can be obtained by the central controller. Our future work will investigate how to implement the offloading scheme when there exist both public and private data. Besides, as stated before, the process of discovering all neighbors in a device's communication range, is of great importance and particularly challenging when devices have directional antennas instead of omni-directional ones. We will investigate on this problem in our future work. Also, it might happen that a helper has a very high contact rate but a very small buffer size, and we will improve the approaches of selecting helpers and allocating buffer in our future work to avoid the extreme case. We now have only considered the allocation of undivided contents. In some cases, it may contribute to the data offloading process if we consider the segmentation of some large video contents. We will take the segmentation of large video contents into account in our future work and improve our offloading scheme. Moreover, a simplified model that all subscribes have the same contact rate with the helper is considered in this paper. In future work, a more general model of different contact rates will be investigated.

Appendix

A. Proof of Lemma 1

Lemma 1: The normal distribution with the expectation $\frac{d_c}{\lambda}$ and the variance $2d_c\lambda^2$, denoted as $N\left(\frac{d_c}{\lambda}, 2d_c\lambda^2\right)$, is a reasonable approximation to the distribution of $\sum_{i=1}^{N(d_c)} r_i$.

Proof: Consider the Probability Density Function (PDF) of the exponential distributed contact duration r_i with parameter λ

$$f_{r_i}(t) = \lambda e^{-\lambda t},\ t > 0.$$

Its characteristic function, denoted by $F_{r_i}(\omega)$, is given by

$$F_{r_i}(\omega) = \frac{\lambda}{\lambda - \mathrm{j}\omega}$$

where $\mathrm{j} = \sqrt{-1}$. Now consider the characteristic function of the random process $\sum_{i=1}^{N(d_c)} r_i$, denoted by $F(\omega)$, which can be expressed as

$$F(\omega) = e^{d_c\left(\frac{\lambda}{\lambda-\mathrm{j}\omega}-1\right)} = e^{d_c\left(\frac{\mathrm{j}\omega}{\lambda-\mathrm{j}\omega}\right)} = e^{\mathrm{j}d_c\omega\left(\frac{\lambda+\mathrm{j}\omega}{\lambda^2+\omega^2}\right)} = e^{\frac{\mathrm{j}\lambda d_c\omega - d_c\omega^2}{\lambda^2+\omega^2}}.$$

Recall that in our problem, the contact duration time is very limited and, consequently, the exponential parameter λ is relatively large. Given a sufficiently large λ, we can approximately have $\frac{1}{\lambda^2+\omega^2} \approx \frac{1}{\lambda^2}$. Therefore, we have

$$F(\omega) \approx e^{\frac{\mathrm{j}\lambda d_c\omega - d_c\omega^2}{\lambda^2}}$$

which is the characteristic function of the normal distribution $N\left(\frac{d_c}{\lambda}, 2d_c\lambda^2\right)$. ■

B. Proof of Lemma 2

Lemma 2: $\widehat{U}(\mathbf{d})$ is a monotonically increasing concave function of $\mathbf{d}$ when $d_c \geq \lambda t_c, \forall c \in \mathcal{C}$.

Proof: Since $\widehat{U}(\mathbf{d}) = \sum_{c\in\mathcal{C}} W_c \widehat{G}(d_c)$ and $W_c \geq 0$, we only need to prove that $\widehat{G}(d_c)$ is a monotonically increasing concave function of d_c. The derivative of $\widehat{G}(d_c)$ is given as

$$\widehat{G}^{'}(d_c) = \frac{\partial \widehat{G}}{\partial d_c} = \frac{\partial \widehat{G}}{\partial \sigma}\frac{\partial \sigma}{\partial d_c} = \frac{\partial \widehat{G}}{\partial \sigma}\left(\frac{1}{\lambda\sqrt{2d_c}}\right)$$
$$= \left(\frac{\lambda}{2} + \frac{t_c}{\sigma^2}\right)\left(\frac{\lambda}{2\sqrt{\pi d_c}}\right)e^{-\frac{\left(\frac{t_c}{\sigma} - \frac{\sigma\lambda}{2}\right)^2}{2}}.$$

By defining $s = \frac{d_c}{\lambda t_c}$, $a = \frac{t_c\lambda}{4}$ and $C = \frac{1}{4\sqrt{\pi\lambda t_c}}e^{\frac{t_c\lambda}{2}}$, we have

$$\frac{\partial \widehat{G}}{\partial d_c} = Ce^{-a\left(s+\frac{1}{s}\right)}\left(1 + \frac{1}{s}\right)\frac{1}{\sqrt{s}} > 0,\ \forall d_c \geq 0. \quad \text{(B1)}$$

Next, we have

$$\frac{\partial^2 \widehat{G}}{\partial {d_c}^2} = Ce^{-a\left(s+\frac{1}{s}\right)}(-a\left(s^{-0.5} + s^{-1.5}\right)\left(1 - s^{-2}\right)$$
$$- 0.5s^{-1.5} - 1.5s^{-2.5})\frac{1}{\lambda t_c}. \quad \text{(B2)}$$

Therefore, when $s \geq 1$ or $d_c \geq \lambda t_c$, $\frac{\partial^2 \widehat{G}}{\partial {d_c}^2} \leq 0$, which together with (A1) prove the Lemma. ■

C. Proof of Theorem 1

*Theorem 1:*The solution obtained by Algorithm 1 is the optimal solution for the optimisation problem (C8).

Proof: According to Lemma 2, the function $\widehat{G}(d_c)$ in Algorithm 1 is a monotonically increasing concave function, and $\widehat{G}^{'}(d_c) > 0$, $\forall d_c \geq 0$. It is clear that the solution obtained by Algorithm 1 for $\mathbf{d}$ meets the conditions

$$W_c \cdot \widehat{G}^{'}(d_c) = C_{\mathrm{A}1},\ \forall d_c \neq 0 \quad \text{(C1)}$$

$$W_c \cdot \widehat{G}'(d_c) < C_{\mathrm{A}1},\ \forall d_c = 0. \quad \text{(C2)}$$

Here the constant $C_{\mathrm{A}1}$ is equal to Φ_{mid} obtained in Algorithm 1. Therefore, we need to prove that the objective function

$$\widehat{U}(\mathbf{d}) = \sum_{c\in\mathcal{C}} W_c \widehat{G}(d_c) \quad \text{(C3)}$$

achieves the maximum value if and only if the conditions (C3) and (C4) are met. We offer the proof in three steps.

First, the range of $\mathbf{d}$ satisfies the condition

$$d_c \geq 0, \forall c \in \mathcal{C},\ \text{and} \sum_{c\in\mathcal{C}} d_c = d_{\mathrm{all}}. \quad \text{(C4)}$$

This defines a closed set. Since a continuous function has at least one maximum on a closed set, the maximum value of $\widehat{U}(\mathbf{d})$ does exist.

Second, assume that $\mathbf{d}$ achieves the maximum value of the objective function (C5). Consider $d_{c1} \leq d_{c2}$. Let us make the following perturbation:

$$d_{c1}^{pert} = d_{c1} + \sigma,\ d_{c2}^{pert} = d_{c2} - \sigma \quad \text{(C5)}$$

where σ is infinitesimal. We have

$$\widehat{U}^{pert} - \widehat{U}$$
$$= W_{c1}\left(\widehat{G}(d_{c1}^{pert}) - \widehat{G}(d_{c1})\right) + W_{c2}\left(\widehat{G}(d_{c2}^{pert}) - \widehat{G}(d_{c2})\right)$$
$$= \sigma\left(W_{c1} \cdot \widehat{G}^{'}(d_{c1}) - W_{c2} \cdot \widehat{G}^{'}(d_{c2})\right). \quad \text{(C6)}$$

Because $\mathbf{d}$ maximises $\widehat{U}(\mathbf{d})$, $\widehat{U}^{pert} - \widehat{U} \leq 0$. Note that $d_c \geq 0$, $\forall c \in \mathcal{C}$, and the perturbation (C7) must be legitimate (in the range of $\mathbf{d}$). This rules out $d_{c1} = d_{c2} = 0$, since σ can be positive or negative. The two legitimate cases are: i) $d_{c1} > 0$, and σ is either positive or negative; and ii) $d_{c1} = 0$, σ is positive, and $d_{c2} > 0$. In the 1st case, $\widehat{U}^{pert} - \widehat{U} \leq 0$ leads to

$$W_{c1} \cdot \widehat{G}^{'}(d_{c1}) = W_{c2} \cdot \widehat{G}^{'}(d_{c2}). \quad \text{(C7)}$$

Thus, we must have $\forall d_c \neq 0$, $W_c \cdot \widehat{G}_c^{'}(d_c) = \text{Constant}$. In the 2nd case, we must have

$$W_{c1} \cdot \widehat{G}^{'}(d_{c1}) < W_{c2} \cdot \widehat{G}^{'}(d_{c2}) = \text{Constant} \quad \text{(C8)}$$

that is, we must have $W_c \cdot \widehat{G}^{'}(d_c) < \text{Constant}$, $\forall d_c = 0$. Combining the both cases proves that the optimal solution of the problem (C8) meets the conditions (C3) and (C4).

Third, we observe that there exists only one point generated by Algorithm 1 that satisfies the conditions (C3) and (C4). This completes the proof. ■

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] "Cisco visual networking index: Global mobile data traffic forecast update, 2015–2020," Feb. 1, 2016. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[3] D. Câmara, N. Frangiadakis, F. Filali, and C. Bonnet, "Vehicular delay tolerant networks," in *Handbook of Research on Mobility and Computing: Evolving Technologies and Ubiquitous Impacts*, M. M. Cruz-Cunha and F. Moreira Eds. Hershey, PA, USA: IGI Global, 2011, pp. 356–367.

[4] B. Han, P. Hui, V. S. Ali Kumar, M. V. Marathe, G. Pei, and A. Srinivasan, "Cellular traffic offloading through opportunistic communications: A case study," in *Proc. 5th ACM Workshop Challenged Netw.*, Chicago, IL, USA, Sep. 20–24, 2010, pp. 31–38.

[5] J. Burgess, B. Gallagher, D. Jensen, and B. Levine, "Maxprop: Routing for vehicle-based disruption-tolerant networks," in *Proc. 25th IEEE Int. Conf. Comput. Commun.*, Barcelona, Spain, Apr. 23–29, 2006, pp. 1–11.

[6] V. N. G. J. Soares, F. Farahmand, and J. Rodrigues, "A layered architecture for vehicular delay-tolerant networks," in *Proc. IEEE Symp. Comput. Commun.*, Sousse, Tunisia, Jul. 5–8, 2009, pp. 122–127.

[7] J. Zhao and G. Cao, "VADD: Vehicle-assisted data delivery in vehicular ad hoc networks," in *Proc. 25th IEEE Int. Conf. Comput. Commun.*, Barcelona, Spain, Apr. 23–29, 2006, pp. 1–12.

[8] K. Lee, I. Rhee, J. Lee, Y. Yi, and S. Chong, "Mobile data offloading: How much can WiFi deliver?" *ACM SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, pp. 425–426, Apr. 2010.

[9] 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*, Vancouver, BC, Canada, Oct. 17–20, 2011, pp. 236–245.

[10] R. Bhatia, G. Narlikar, I. Rimac, and A. Beck, "UNAP: User-centric network-aware push for mobile content delivery," in *Proc. IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 19–25, 2009, pp. 1–9.

[11] J. Whitbeck, M. Amorim, Y. Lopez, J. Leguay, and V. Conan, “Relieving the wireless infrastructure: When opportunistic networks meet guaranteed delays,” in *Proc. IEEE Int. Symp. World Wireless, Mobile, Multimedia Networks*, 2011, pp. 1–10.

[12] A. Balasubramanian, R. Mahajan, and A. Venkataramani, “Augmenting mobile 3G using WiFi,” in *Proc. 8th Int. Conf. Mobile Syst., Appl., Serv.*, San Francisco, CA, USA, Jun. 15–18, 2010, pp. 209–222.

[13] B. Higgins *et al.*, “Intentional networking: Opportunistic exploitation of mobile network diversity,” in *Proc. 16th Annu. Int. Conf. Mobile Comput. Netw.*, Chicago, IL, USA, Sep. 20–24, 2010, pp. 73–84.

[14] U. Shevade *et al.*, “Enabling high-bandwidth vehicular content distribution,” in *Proc. 6th Int. Conf. Emerging Netw. Exp. Technol.*, Philadelphia, PA, USA, Nov. 30–Dec. 3, 2010, pp. 12–23.

[15] P. Costa, C. Mascolo, M. Musolesi, and G. P. Picco, “Socially-aware routing for publish-subscribe in delay-tolerant mobile ad hoc networks,” *IEEE J. Sel. Areas Commun.*, vol. 26, no. 5, pp. 748–760, Jun. 2008.

[16] S. Ioannidis, A. Chaintreau, and L. Massoulié, “Optimal and scalable distribution of content updates over a mobile social network,” in *Proc. IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 19–25, 2009, pp. 1422–1430.

[17] L. Ilias, P. Costa, and C. Mascolo, “A hybrid approach for content-based publish/subscribe in vehicular networks,” *Pervasive Mobile Comput.*, vol. 5, no. 6, pp. 697–713, 2009.

[18] C. Paolo, C. Mascolo, M. Musolesi, and G. P. Picco, “Socially-aware routing for publish-subscribe in delay-tolerant mobile ad hoc networks,” IEEE *J. Sel. Areas Commun.*, vol. 26, no. 5, pp. 748–760, Jun. 2008.

[19] L. Qiu, Y. Li, P. Hui, D. Jin, L. SU, and L. Zeng, “Edge-Markovian dynamic graph based performance evaluation for delay tolerant networks,” in *Proc. Wireless Commun. Netw. Conf.*, Paris, France, Apr. 1–4, 2012, pp. 2129–2133.

[20] H. Luo, X. Meng, R. Ramjee, P. Sinha, and L. Li, “The design and evaluation of unified cellular and ad hoc networks,” *IEEE Trans. Mobile Comput.*, vol. 6, no. 9, pp. 1060–1074, Sep. 2007.

[21] B. Han, P. Hui, and A. Srinivasan, “Mobile data offloading in metropolitan area networks,” *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 14, no. 4, pp. 28–30, Oct. 2011.

[22] X. Zhuo, W. Gao, G. Cao, and Y. Dai, “Win-coupon: An incentive framework for 3G traffic offloading,” in *Proc. 19th IEEE Int. Conf. Netw. Protocols*, Vancouver, BC, Canada, Oct. 17–20, 2011, pp. 206–215.

[23] K. Lee, Y. Yi, J. Jeong, H. Won, I. Rhee, and S. Chong, “Max-contribution: On optimal resource allocation in delay tolerant networks,” in *Proc. 29th IEEE INFOCOM*, San Diego, CA, USA, Mar. 15–19, 2010, pp. 1–9.

[24] H. Zhu, L. Fu, G. Xue, Y. Zhu, M. Li, and L. Ni, “Recognizing exponential inter-contact time in VANETs,” in *Proc. 29th IEEE INFOCOM*, San Diego, CA, USA, Mar. 14–19, 2010, pp. 1–5.

[25] W. Wang, V. Srinivasan, and M. Motani, “Adaptive contact probing mechanisms for delay tolerant applications,” in *Proc. 13th ACM Int. Conf. Mobile Comput. Netw.*, Montrèal, QC, Canada, Sep. 9–14, 2007, pp. 230–241.

[26] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, “Pocket switched networks: Real-world mobility and its consequences for opportunistic forwarding,” Univ. Cambridge Computer Lab, Cambridge, U.K., Tech. Rep. UCAM-CL-TR-617, Feb. 2005.

[27] H. Zhu, M. Li, L. Fu, G. Xue, Y. Zhu, and L. 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.

[28] Y. Li, D. Jin, L. Zeng, and S. Chen, “Revealing patterns of opportunistic contact durations and intervals for large scale urban vehicular mobility,” in *Proc. 2013 IEEE Int. Conf. Commun.*, Budapest, Hungary, Jun. 9–13, 2013, pp. 1646–1650.

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

[30] A. Kulik, H. Shachnai, and T. Tamir, “Maximizing submodular set functions subject to multiple linear constraints,” in *Proc. 20th Annu. ACM, SIAM Symp. Discr. Algorithms*, New York, NY, USA, Jan. 4–6, 2009, pp. 545–554.

[31] Y. Li, G. Su, P. Hui, D. Jin, L. Su, and L. Zeng, “Multiple mobile data offloading through delay tolerant networks,” in *Proc. 6th ACM Workshop Challenged Netw.*, Las Vegas, NV, USA, Sep. 19–23, 2011, pp. 43–48.

[32] M. Ikawa, Y. Goto, Y. Igarashi, H. Kumazawa, K. Koizumi, and K. Oka, “DSRC local communication platform and its application to information push service,” in *Proc. IEEE Intell. Veh. Symp.*, Parma, Italy, Jun. 14–17, 2004, pp. 105–110.

[33] L. Kevin, S. Lee, R. Cheung, U. Lee, and M. Gerla, “First experience with cartorrent in a real vehicular ad hoc network testbed,” in *Prof. IEEE Mobile Netw. Veh. Environ.*, 2007, pp. 109–114.

[34] Z. Thomas, G. O’Shea, and A. Rowstron, “Feasibility of content dissemination between devices in moving vehicles,” in *Proc. ACM 5th Int. Conf. Emerging Netw. Exp. Technol.*, 2009, pp. 97–108.

[35] L. Ming, Z. Yang, and W. Lou, “Codeon: Cooperative popular content distribution for vehicular networks using symbol level network coding,” *IEEE J. Sel. Areas Commun.*, vol. 29, no. 1 pp. 223–235, Jan. 2011.

[36] Z. Xuejun, Q. Li, G. Cao, Y. Dai, B. Szymanski, and T. Porta, “Social-based cooperative caching in DTNs: A contact duration aware approach,” in *Proc. IEEE 8th Int. Conf. Mobile Adhoc Sensor Syst.*, 2011, pp. 92–101.

[37] L. Yong, L. Qiu, D. Jin, L. Su, P. Hui, and L. Zeng, “Contact duration aware evaluation for content dissemination delay in mobile social network,” *Wireless Commun. Mobile Comput.*, vol. 15, pp. 527–537, 2013.

[38] Y. Li, Z. Wang, D. Jin, L. Zeng, and S. Chen, “Collaborative vehicular content dissemination with directional antennas,” *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1301–1306, Apr. 2012.

[39] U. Lee, R. Cheung, and M. Gerla, “Emerging Vehicular Applications,” in *Vehicular Networks: From Theory to Practice.* Boca Raton, FL, USA: CRC, 2009.

[40] S. Tachi, M. Inami, and Y. Uema, “Augmented reality helps drivers see around blind spots,” *IEEE Spectr.*, Oct. 31, 2014.

[41] 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.

[42] *Family of Standards for Wireless Access in Vehicular Environments (WAVE)*, U.S. Dept. Transp., Washington, DC, USA, IEEE Stand. 1609, Apr. 13, 2013.

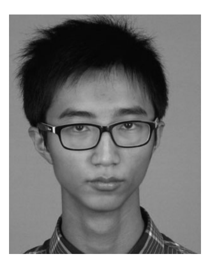
[43] A. Balasubramanian, B. N. Levine, A. Venkataramani, “Replication routing in DTNs: A resource allocation approach,” *IEEE/ACM Trans. Netw.*, vol. 18, no. 2, pp. 596–609, Apr. 2010.

[44] H. Cai *et al.*, “Neighbor discovery algorithms in wireless networks using directional antennas,” in *Proc. IEEE Int. Conf. Commun.*, 2012, pp. 767–772.

[45] C. B. Moler, “MATLAB—An interactive matrix laboratory,” Dept. Math. Stat., Tech. Rep. 369, Univ. New Mexico, Albuquerque, NM, USA, 1980.

[46] J. Whitbeck, Y. Lopeza, J. Leguay, V. Conan, and M. D. de, “Push-and-track: Saving infrastructure bandwidth through opportunistic forwarding,” *Pervasive Mobile Comput.,* vol. 8, no. 5, pp. 682–697, 2012.

[47] H. Yao, D. Zeng, H. Huang, S. Guo, A. Barnawi, and I. Stojmenovic, “Opportunistic offloading of deadline-constrained bulk cellular traffic in vehicular DTNs,” *IEEE Trans. Comput.*, vol. 64, no. 12, pp. 3515–3527, Dec. 2015.



Xiangming Zhu received the B.S. degree from Tsinghua University, Beijing, China, in 2010, where he is currently working toward the Ph.D. degree with the Department of Electronic Engineering. His major research interest focuses on integrated terrestrial-satellite communications.

Yong Li (M’09–SM’16) received the B.S. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2007 and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012.

He is currently a faculty member with the Department of Electronic Engineering, Tsinghua University. He has been the General Chair, TPC Chair, and TPC Member for several international workshops and conferences, and he is on the editorial boards of three international journals. His papers have total citations of more than 2000 (six papers exceed 100 citations, Google Scholar). Among them, eight are ESI highly cited papers in computer science, and four have received conference Best Paper (run-up) Awards. He received the IEEE 2016 ComSoc Asia-Pacific Outstanding Young Researchers award.

Depeng Jin (M'08) received the B.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 1995 and 1999, respectively, both in electronics engineering.

He is an Associate Professor at Tsinghua University and Vice Chair of the Department of Electronic Engineering. His research interests include telecommunications, high-speed networks, ASIC design, and future Internet architecture. He received the National Scientific and Technological Innovation Prize (Second Class) in 2002.

Jianhua Lu (M'98–SM'07–F'15) received the B.S.E.E. and M.S.E.E. degrees from Tsinghua University, Beijing, China, in 1986 and 1989, respectively, and the Ph.D. degree in electrical and electronic engineering from Hong Kong University of Science and Technology, Hong Kong, China, in 1998. Since 1989, he has been with the Department of Electronic Engineering, Tsinghua University, where he is currently a Professor. He is also a Chief Scientist of the National Basic Research Program (973) of China. He has authored more than 200 technical papers in international journals and conference proceedings. His research interests include broadband wireless communications, multimedia signal processing, and satellite communications. He has been an active member of professional societies. He has served at numerous IEEE conferences as a member of Technical Program Committees. He served as the Lead Chair of the General Symposium of the IEEE ICC in 2008 and the Program Committee Co-Chair of the 9th IEEE International Conference on Cognitive Informatics in 2010. He received the Best Paper Award at the IEEE International Conference on Communications, Circuits, and Systems in 2002, ChinaCom in 2006, the IEEE Embedded-Com in 2012, and the National Distinguished Young Scholar Fund by the NSF Committee of China in 2005.