Optimizing Content Dissemination in Vehicular Networks with Radio Heterogeneity

Joon Ahn†, Maheswaran Sathiamoorthy†, Bhaskar Krishnamachari†, Fan Bai‡, Lin Zhang§
† University of Southern California, Los Angeles, CA, USA
‡ General Motors Corporation, ECI Lab, Warren, MI, USA
§Tsinghua University, Beijing, China
{joonahn, msathiam, bkrishna}@usc.edu, fan.bai@gm.com, linzhang@tsinghua.edu.cn

Abstract—Disseminating shared information to many vehicles could incur significant access fees if it relies only on unicast cellular communications. We consider the problem of efficient content dissemination over a vehicular network, in which vehicles are equipped with two kinds of radios: a high-cost low-bandwidth, long-range cellular radio, and a free high-bandwidth short-range radio. We formulate and solve an optimization problem to maximize content dissemination from the infrastructure to vehicles within a predetermined deadline while minimizing the cost associated with communicating over the cellular connection.

We examine numerically the tradeoffs between cost, delay and system utility in the optimum regime. We find that, in the optimum regime, (a) system utility is more sensitive to the cost budget when the allowed delay for the dissemination is not large, (b) the system requires relatively smaller cost budget as more vehicles participate and more delay is allowed, (c) when the cost is very important, it is better not to spread the content if it needs small delay. We also develop a polynomial-time algorithm to obtain the optimal discrete solution needed in practice. Finally, we verify our analysis using real GPS traces of 632 taxis in Beijing, China.

1 Introduction

Although they are widely deployed, and becoming more spectrum-efficient with the ongoing transition to 4G systems, cellular infrastructure networks are feeling the strain of rapidly increasing data traffic due to new mobile platforms and applications. It is widely predicted that the volume of mobile data consumed by users will grow exponentially in the next decade. For example, [1] estimated that the global mobile data traffic will increase from 90,000 Terabytes per month in 2009 to 3,600,000 Terabytes per month in 2015, resulting in a dramatic increase of 39 times in 6 years; similarly, AT&T reported that its wireless data usage jumped almost 5,000% from 2006 to 2009 [2]. At the same time, it is also estimated that the growth of cellular infrastructures might fail to keep up with the pace of mobile data growth [3]. The outcome of exponentially growing mobile data significantly surpassing the limited supply of cellular data pipe is being vividly termed as *Mobile Data Tsunami*. As an early sign of mobile data tsunami, the recent event that newly introduced smartphones overloaded cellular systems in major cities was well documented [2].

The practical bandwidth constraints in cellular systems are not only due to limited wireless spectrum but also because of limited capacity of backhaul; while increasing the cellular capacity through additional spectrum or backhaul infrastructure is possible, it will incur significant capital and operational expenditure, further increasing the cost of cellular access charged to customers [4], [5]. Therefore, we contend that, *as the cellular bandwidth becomes increasingly crowded and more expensive, hybrid protocols that synergistically combine direct cellular access along with store-carry-forward routing through peer-to-peer communication will be proved as a bandwidth-efficient and cost-effective way for offloading the often congested cellular infrastructure*.

In this paper, as a first step towards understanding the potential benefits of using delay-tolerant networks to offload the cellular networks, we scope our research interest only to highly mobile vehicular networks. Particularly, we consider the problem of efficient dissemination of some delay-tolerant content to a group of vehicles that share an interest in this content. The delay-tolerant contents can support a variety of services, ranging from traditional traffic information and weather forecast to futuristic mobile advertisement and music (mp3) sharing. Such applications/services have been envisioned by industry as a key driving force for future vehicular networks [6].

We assume that vehicles in the future could be equipped with two different types of wireless radios – a high-usage-cost low-bandwidth, long-range cellular radio, and a free high-bandwidth, short-range WiFi-like radio[1]. In light of the fact that cellular radios in cars would allow only for unicast communication, and therefore incur a significant unit per-vehicle charge for content download, the use of the free short-range radio to assist in such a broad dissemination

Research was sponsored in part by a contract from General Motors, in part by the U.S. Army Research Laboratory under the Network Science Collaborative Technology Alliance, Agreement Number W911NF-09-2-0053, and in part by the National Science Foundation, under award 1217260.

1. Recent trends in the automotive industry point to an emerging age of vehicular communication networks consisting of cars equipped with both cellular radio devices as well as short range inter-vehicular radios such as those based on IEEE 802.11p/WAVE (wireless access for vehicular environments).

process becomes economically compelling. We formulate in this work an optimization problem from the perspective of a content provider[2], with the goal of maximizing the number of vehicles that obtain the content within a given deadline while minimizing the expense of using the cellular infrastructure. Our contribution in this work is as follows:

1) We analyze mathematically the content dissemination process, formulating it as an optimization problem, and derive the optimum solution for the problem in closed-form using the method of Lagrange multipliers for constrained convex optimization.
2) We then investigate the behaviors of the system in terms under various parameter settings to understand the key tradeoffs. We also develop a polynomial-time algorithm to obtain the practical optimum solution to overcome the non-integrality limitations of our closed-form solution.
3) Finally, we use GPS traces of 632 taxis in Beijing to verify our analysis. We conclude that content can be spread effectively to most vehicles across a city in a reasonably timely manner with very low-cost use of the cellular infrastructure.

2 Problem Formulation

As introduced in Section 1, we consider a heterogeneous vehicular network consisting of cars with both short-range and cellular radios, over which m-types of content need to be disseminated to m-groups of vehicles. The i-th group of vehicles are interested in the i-th type of content. The goal is to efficiently disseminate these contents to their corresponding groups of nodes from the infrastructure exploiting both long-range and short-range communication methods.

One extreme way of the dissemination is to send the contents to each one of vehicles in interest through the long-range radio only. This method incurs significant access fees proportional to the number of the interested vehicles although the associated delay would be small. The other extreme is to send the message to only one vehicle in each interested group through the long-range radio, and let it spread to other vehicles through encounters via the short-range radio. In contrast to the first approach, this incurs the minimum access fees, but the delay for reaching a large number of nodes would be substantial. In between, the delay would decrease as the number of vehicles that obtain the messages directly through the long-range radio (we call them *seed nodes*) is increased, with a corresponding increase in access cost. Thus the number of seed nodes tunes a fundamental tradeoff between delay and cost.

Our goal in this problem is, then, to maximize the expected number of vehicles obtaining the contents in their interest such that the access cost is as low as possible, subject to the long-range radio access cost constraint and tolerable delay constraint. For more specific presentation, let us suppose m types of messages to disseminate from the infrastructure. Let n denote the total number of nodes in the network, and p_i is the proportion of the nodes that are interested in the i-th type of messages. We use interchangeably the terms node and vehicle, and messages and contents, respectively, in this paper. We assume (without loss of generality) that each long-range radio access incurs a unit cost, while k_i is the number of seeds for the i-th type of message. Hence, the total cost $c(\vec{k})$ is the sum of all k_i-s, where $\mathbf{k} = (k_1, k_2, ..., k_m)$ is called the *seed vector*. Let $s_i(k_i, t)$ denote the expected number of satisfied nodes for i-th type content at time t when the number of seed node is k_i. We assume that the seeds are deployed at time 0.

Then the problem formulation is as follows:

$$\begin{aligned} \mathit{PF1} : \text{Maximize}_{\mathbf{k}} \quad & f(\mathbf{k}) = \textstyle\sum_{i=1}^{m} s_i(k_i, d) - w \cdot c(\mathbf{k}) \\ s.t \quad & c(\mathbf{k}) = \textstyle\sum_{i=1}^{m} k_i \leq C \\ & 0 \leq k_i \leq n_i = p_i n, \ \forall i \in M \\ & \mathbf{k} \in \mathbb{N}^m \end{aligned}$$

where $M = \{1, 2, \ldots, m\}$, the cost budget is C, the tolerable delay is $d > 0$ units of time, and $w > 0$ is the total cost weight. The total cost weight is a parameter that helps indicate the relative weight between a) dollars spent on downloading content to the seed nodes $c(\mathbf{k})$, and c) the number of satisfied nodes that have received the content by the deadline. In particular, if the cost of each seed is normalized to one, w is chosen to have the intuitive meaning that deploying one more seed should bring at least w satisfied nodes on average.

The objective function $f(\mathbf{k})$, which is referred to as *system utility* in this paper, is essentially the extra benefits induced by the short-range radio. It is easy to see that the system utility is the expected number of satisfied vehicles through the short-range radio alone, when the total cost weight $w = 1$.

3 Modeling Dissemination

In this section, we derive the expected number $s_i(k_i, t)$ of satisfied nodes obtaining i-th type of content at time t when only k_i seeds are initially deployed at time 0.

3.1 Terminology and Assumption

We first define the symbols used in our analysis as well as state the assumptions.

1) We assume that a node may encounter α proportion of all nodes on average for the time interval in interest;
2) For any pair of nodes, we assume that the inter-encounter time follows the Exponential distribution with rate β;
3) We also assume that the inter-encounter times of pairs of nodes are jointly independent and identical;

The assumption (1) is self-explainable. The assumptions (2) and (3) make our analysis mathematically tractable, and they have been found reasonable when vehicles follow conventional mobility models such as random waypoint model ([7]). At the same time, we acknowledge that these two assumptions might not be always realistic so that we relax both of them in our trace-driven simulation; though neither assumption is perfectly

2. Purely as a motivating example for focusing on cost-optimization from the perspective of a content provider, consider that on purchasing an Amazon Kindle, the customer does not pay directly for the 3G access that may be needed to get content such as books; thus there is an incentive for Amazon to reduce its 3G data access costs as much as possible. We show in this work that if the content being provided in a vehicular network is not too delay-sensitive, the content distribution cost can be driven down by the use of V2V communication.

honored in the empirical traces of Beijing taxis, our simulation results still reasonably agree with our theoretical results.

In our study, we take an *interest-only* caching policy: A node sends the previously obtained messages only to the nodes that are interested in the same type of messages. There does exist some prior work considering non-interest-only caching policies (e.g., [8]). In this paper, we focus on this interest-only caching policy because it can avoid the non-ignorable storage costs for keeping uninterested data incurred otherwise. A different solution is to allow vehicles to carry contents which the vehicle users might not be interested in. It is obvious that the latter could provide an even better performance than the interest-only solution at the cost of extra storage space. The analysis of the latter cache policy is out of our scope, which merits an independent study.

Note that atomic contact among vehicles is assumed, implying that the message exchange between a pair of vehicles could be completed during their encounter process. As shown in [9], [10], it is reported that 30-70 MB data could be transfered as vehicle encounters (with normal driving speeds). Thus, we believe that most types of light-weight content (weather forecast, traffic information, mobile advertisement) could be successfully transmitted during short encounters between vehicles. For larger-size content, if there is no storage and resumption of partial transfers, the approach in this work can be extended in a straightforward manner by considering only those contacts that are long enough. This can be done by reducing the contact rates in the analysis or removing low-duration contacts from the traces for simulation. Taking into account the resumption of partial transfers exactly in simulations requires a more careful accounting of the states of individual transfers depending on the contact duration. Accounting for them exactly in the analysis is more challenging, but the performance in this case can be upper and lower bounded by two different contact rates: with the higher one corresponding to considering all possible contacts, and the lower one corresponding to considering only contacts whose duration is sufficient for complete transfer.

3.2 ODE model

We observe that the expected number of satisfied nodes behaves like the number of infected nodes in epidemic routing ([11]). The previous work has introduced largely two methods to analyze the number of infected nodes; one is using Markov chains and the other is using ordinary differential equations (ODE) ([12], [13]). We use the ODE method, which is quite well-known from its biological applications to modeling the spread of epidemics [14], [15].

First, consider the expected number of newly satisfied nodes $\triangle S$ between time t and $t+dt$, where dt is infinitesimal. There are two groups of nodes at time t; a group of satisfied nodes and a group of unsatisfied nodes. The number of nodes in the former group is $s_i(k_i, t)$ as defined, and that of the latter is $n_i - s_i(k_i, t)$, where $n_i(= p_i n)$ is the number of nodes that are interested in the type i message.

Let us define the *inter-encounter time between the two groups* as the time elapsed until any node in one group meets any node in the other group after such encounter of inter-group nodes happens. Then, the inter-encounter time between the satisfied and the unsatisfied follows the Exponential distribution with rate $\beta\times$*(# of pairs of ever-encounter inter-group nodes)*, because the inter-encounter time of each pair of nodes that ever meets is IID Exponential (Assumptions (2) and (3)) and each node meets a fraction of other nodes (Assumption (1)).

Therefore, the expected number of newly satisfied nodes $\triangle S = s_i(k_i, t+dt) - s_i(k_i, t) = \alpha\beta s_i(k_i, t)\,(n_i - s_i(k_i, t)) \cdot dt$. Note that the expected number of ever-meeting pairs of inter-group nodes is approximately[3] $\alpha s_i(k_i, t)\,(n_i - s_i(k_i, t))$.

From the above equation and the fact that the number of seeds is k_i, we have the following ODE system:

$$\frac{\partial s_i(k_i,t)}{\partial t} = \alpha\beta s_i(k_i,t)(n - s_i(k_i,t)) \tag{1}$$

$$s_i(k_i, 0) = k_i \tag{2}$$

It turns out that this ODE system has the closed-form solution as follows:

$$s_i(k_i, t) = \frac{n_i}{1 + (n_i/k_i - 1)\exp(-n_i\alpha\beta t)} \tag{3}$$

4 Optimization

In this section we derive theoretically the solution of the optimization problem proposed in Section 2. In order to gain better intuition about the system behavior, we relax the optimization problem ignoring the integral constraint on the numbers of seeds k_i. Therefore, we focus on the following optimization problem $PF2$ in this section:

$$PF2 : \text{Maximize}_{\mathbf{k}} \;\; f(\mathbf{k}) = \textstyle\sum_{i=1}^{m} s_i(k_i, d) - w \cdot c(\mathbf{k}) \tag{4}$$

$$s.t \;\; c(\mathbf{k}) = \textstyle\sum_{i=1}^{m} k_i \le C \tag{5}$$

$$0 \le k_i \le n_i = p_i n, \; \forall i \in M \tag{6}$$

We first show that the problem is a convex optimization problem, then, solve the problem using the method of Lagrange multipliers. In the process, we further relax some constraints for easier derivation, and then, provide the condition under which the solution derived with the relaxation is valid for the original problem $PF2$.

4.1 Convexity of the Problem

The expected number s_i of the satisfied nodes is concave with respect to the number of seeds k_i because its first derivative is non-negative and its second derivative is non-positive as follows:

$$\frac{\partial s_i(k_i,d)}{\partial k_i} = \frac{n_i^2 z_i}{k_i^2(1 + (n_i/k_i - 1)z_i)^2} \ge 0, \;\; \forall k_i \in (0, n_i]$$

$$\frac{\partial^2 s_i(k_i,d)}{\partial k_i^2} = -\frac{2n_i^2 z_i(1 - z_i)}{k_i^3(1 + (n_i/k_i - 1)z_i)^3} \le 0, \;\; \forall k_i \in (0, n_i]$$

3. This is because we approximate the expectation of the square of the number of satisfied nodes at time t to the square of the expectation of the number of satisfied, which is not rigorously true with the finite number of nodes. However, it becomes more accurate and eventually exact as $n \to \infty$ because the variance goes to zero. We shall also see when we validate with the real traces, this is still a useful approximation.

where we use the following for concise presentation:

$$z_i = e^{-n_i \alpha \beta d} \tag{7}$$

Therefore, the objective function $f(\mathbf{k})$ is a linear combination of concave functions, which implies that the function itself is concave. From the concavity of the objective function and the fact that all constraints are linear, we can see that the problem is a convex optimization problem.

4.2 Optimum Number of Seeds

We use the Lagrange dual of the convex optimization problem to obtain the optimum solution. We further ignore the constraints in (6) for now for the concise presentation of the derivation. But, we shall provide the conditions under which the obtained solution in this section is valid for *PF2*.

The Lagrangian of the problem is as follows:

$$L(\mathbf{k}, \lambda) = f(\mathbf{k}) - \lambda \left(c(\mathbf{k}) - C\right) \tag{8}$$

where λ is the Lagrange multiplier and $\lambda \geq 0$.

Since the primal problem is concave, it is well-known that the parameter set $(\widehat{\mathbf{k}}, \widehat{\lambda})$ that minimize $\sup_{\mathbf{k}} L(\mathbf{k}, \lambda)$ maximizes the primal. Because the Lagrangian is also concave with respect to $\mathbf{k}$, we have the following conditions for such $(\widehat{\mathbf{k}}, \widehat{\lambda})$;

$$\frac{\partial L(\mathbf{k}, \lambda)}{\partial k_i} = \frac{n_i^2 z_i}{(k_i + n_i z_i - k_i z_i)^2} - \lambda - w = 0 \quad \forall i \tag{9}$$

$$\frac{\partial L(\mathbf{k}, \lambda)}{\partial \lambda} = \lambda(\textstyle\sum_{i=1}^m k_i - C) = 0 \tag{10}$$

As can be seen from (10), we have two cases: one for $\lambda = 0$ (*i.e.* $\sum k_i < C$) and the other for $\sum k_i = C$. When $\sum k_i < C$, the constraint (5) is inactive meaning that the solution of the constrained optimization problem is indeed that of its unconstrained version. Suppose $\widetilde{\mathbf{k}}$ is the unconstrained optimum solution, and let $\widetilde{C}$ be the unconstrained optimum total cost, given by;

$$\widetilde{C} \doteq c(\widetilde{\mathbf{k}}) = \textstyle\sum_{i=1}^m \widetilde{k}_i \tag{11}$$

Then, $\widetilde{C} = c(\widetilde{\mathbf{k}}) < C$, and so, the optimum solution $\widetilde{\mathbf{k}}$ automatically satisfies the constraint (5) in this case.

On the other hand, the constraint (5) is active in the case where $\sum k_i = C$. This means that the unconstrained solution requires more cost than allowed in general, that is, $C \leq \widetilde{C}$. In other words, the system does not afford the unconstrained optimum seed vector, resulting in fewer numbers of seeds to meet the constraint. Therefore, the system utility $f(\mathbf{k})$ would be smaller than its maximum possible.

Now we provide the solution of the constrained optimization problem as follows:

$$\widehat{k}_i = \begin{cases} \widetilde{k}_i = \frac{n_i \sqrt{z_i}}{1 - z_i}\left(1/\sqrt{w} - \sqrt{z_i}\right), & \text{if } \widetilde{C} < C \quad (12a) \\ \breve{k}_i = \frac{n_i \sqrt{z_i}}{1 - z_i}\left(\frac{C + A}{B} - \sqrt{z_i}\right), & \text{if } \widetilde{C} \geq C \quad (12b) \end{cases}$$

where

$$A = \textstyle\sum_{i=1}^m \frac{n_i z_i}{1 - z_i}, \qquad B = \sum_{i=1}^m \frac{n_i \sqrt{z_i}}{1 - z_i} \tag{13}$$

And, $\widetilde{C}$ can be obtained from (11) and (12a). The derivation for the solution is not terribly difficult, and so, we omit it in this paper for more concise presentation. We note that the solution in (12) still ignores the constraint (6). However, we show that the solution is indeed the solution of *PF2* under the conditions in Theorems 1 and 2.

Theorem 1. *Suppose $\widetilde{k}_i$ and $\widetilde{C}$ are defined as in (12a) and (11), respectively. Also, suppose $z_i = \exp(-n_i \alpha \beta d)$. Then, under any one of the following conditions,*

$$\begin{aligned} \mathbb{C}_1 : &\quad \{0 < w < 1, 0 < z_i \leq w\} \\ \mathbb{C}_2 : &\quad \{w = 1, 0 < z_i < 1\} \\ \mathbb{C}_3 : &\quad \{w > 1, 0 < z_i \leq 1/w\} \end{aligned}$$

the optimum numbers of seeds, k_i^*, of the optimization problem PF2 are, if $\widetilde{C} < C$,

$$k_i^* = \widetilde{k}_i \tag{14}$$

Proof: We note that $\widetilde{k}_i$ is the solution of *PF2* when $\widetilde{C} < C$ if we ignore the constraint (6). Hence, what we need to show is that $\widetilde{k}_i$ is in the interval $[0, n_i]$ under any of the conditions $\mathbb{C}_1, \mathbb{C}_2$, or $\mathbb{C}_3$ so that the constraint is satisfied.

We can represent $\widetilde{k}_i$ as follows:

$$\widetilde{k}_i = \tfrac{n_i \sqrt{z_i}}{1 - z_i}\left(1/\sqrt{w} - \sqrt{z_i}\right) = n_i \cdot y(z_i)$$

where

$$y(z_i) \doteq \tfrac{\sqrt{z_i/w} - z_i}{1 - z_i} \tag{15}$$

Then, we only need to show $0 \leq y(z_i) \leq 1$ under any of the three conditions.

When $0 < w \leq 1$, we can see that $y(z)$ is monotonically non-decreasing in $(0, 1)$ because its first derivative is non-negative in that interval as follows:

$$\frac{dy(z)}{dz} = \frac{1 - w + (\sqrt{z} - \sqrt{w})^2}{2\sqrt{wz}(1 - z)^2} \tag{16}$$

Hence, we can easily see $0 = y(0) < y(z) \leq y(w) = 1$ under the condition $\mathbb{C}_1$. Under the condition $\mathbb{C}_2$, we can see $0 \leq y(z_i) \leq 1$ from the following:

$$0 = y(0) < y(z) < \lim_{z \to 1} y(z) = 1/2 < 1 \tag{17}$$

Note that we cannot use $y(1)$ because it is not defined at $z = 1$.

Now consider the last condition $\mathbb{C}_3$. When $w > 1$, we can easily see that $y(z) < 1$ for $0 < z < 1$ from (15). And it is not difficult to see that $y(z) > 0$ for $0 < z < 1/w$. And, these imply that $0 \leq \widetilde{k}_i \leq n_i$ under $\mathbb{C}_3$. □

Theorem 2. *Suppose $\breve{k}_i$ and $\widetilde{C}$ are defined as in (12b) and (11), respectively. Also, suppose $z_i = \exp(-n_i \alpha \beta d)$.*

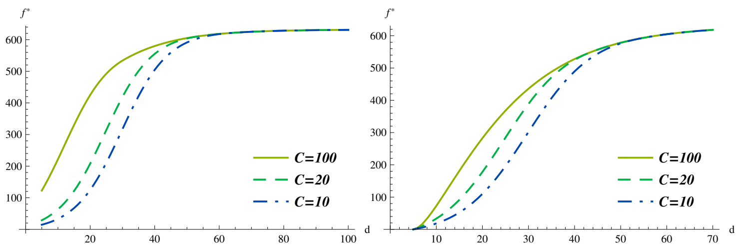
Then, if any of the conditions $\mathbb{C}_1, \mathbb{C}_2$ and $\mathbb{C}_3$ holds, and also if

$$\textstyle\sum_{j=1}^m \frac{n_j \sqrt{z_j}}{1 - z_j} \leq C \leq \widetilde{C}$$

, the optimum numbers of seeds, k_i^*, of the optimization problem PF2 are,

$$k_i^* = \breve{k}_i$$

Proof: $\breve{k}_i$ is the solution of *PF2* when $C \leq \widetilde{C}$ if we ignore the constraint (6). Hence, we only need to show $\breve{k}_i \in [0, n_i]$ under the conditions.



(a) $w = 0.5$ (b) $w = 2$

Fig. 1: Optimum utility vs. delay budget

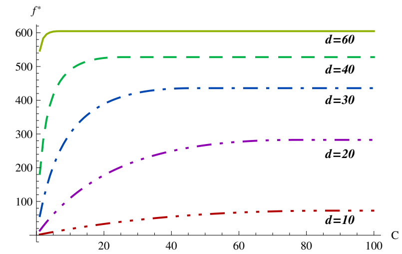


Fig. 2: Optimum utility vs. cost budget ($w = 2$)

First, we will show that $\breve{k}_i \leq n_i$. Since $\widetilde{C} \geq C$,

$$\sum_{i=1}^{m} \frac{n_i\sqrt{z_i}}{1-z_i}\left(1/\sqrt{w} - \sqrt{z_i}\right) = B/\sqrt{w} - A \geq C$$
$$\Rightarrow\ 1/\sqrt{w} \geq (C+A)/B$$

where A and B are defined in (13).

This implies, together with (12) and the proof of Theorem 1, that $\breve{k}_i \leq \widetilde{k}_i \leq n_i$.

Now let us show that $\breve{k}_i \geq 0$. Since $C \geq \sum_j \frac{n_j\sqrt{z_j}}{1-z_j}$,

$$\begin{aligned}\breve{k}_i &\geq \frac{n_i\sqrt{z_i}}{1-z_i}\left[\frac{1}{B}\left(\sum_j \frac{n_j\sqrt{z_j}}{1-z_j} + A\right) - \sqrt{z_i}\right] \qquad (18)\\ &= \frac{n_i\sqrt{z_i}}{B(1-z_i)}\sum_j^m \frac{n_j\sqrt{z_j}}{1-z_j}\left(1 + \sqrt{z_j} - \sqrt{z_i}\right) \geq 0\end{aligned}$$

where (18) follows since $\sqrt{z_j} - \sqrt{z_i} \geq -1$ for all j and i. □

4.3 Optimum System Utility

In this section we investigate the system behavior when the seed vector is optimum $\mathbf{k}^*$. We first derive the optimum expected number of satisfied nodes and the optimum system utility, and look into how they depend on the system parameters, such as the cost budget C, delay budget d, etc., through numerical evaluations.

The optimum number s_i^* of satisfied nodes can be derived from (3) and (12), given by

$$s_i^*(d,C) = \begin{cases} \sum_{i=1}^{m} \frac{n_i}{1-z_i}\left(1 - \sqrt{wz_i}\right), & \text{if } \widetilde{C} < C \quad (19a)\\ \sum_{i=1}^{m} \frac{n_i}{1-z_i}\left(1 - \frac{B}{C+A}\sqrt{z_i}\right), & \text{otherwise} \quad (19b)\end{cases}$$

where z_i, A, and B are given in (7) and (13) respectively.

The optimum system utility is from (4), (12), and (19), as follows:

$$f^*(d,C) = \begin{cases} \sum_{i=1}^{m}(1 - \sqrt{wz_i})s_i^*(d,C), & \text{if } \widetilde{C} < C \quad (20a)\\ \sum_{i=1}^{m}\left(1 - \frac{C+A}{B}w\sqrt{z_i}\right)s_i^*(d,C), & \text{otherwise} \quad (20b)\end{cases}$$

Because of the complexity of the above equations, it is hard to obtain a good intuition on the optimum system behavior from the equations themselves. So, we resort to the numerical evaluations of the equations for better intuition. When it comes to numerical evaluation, the equations are very simple and easy to calculate. However, we need proper parameter values for evaluations in order to have relevant results.

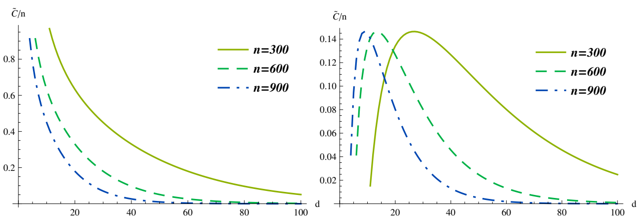
We use the values we obtain from the real traces of vehicles in Section 5; the number of nodes $n = 632$, the inter-encounter rate $\beta = 19.178 \times 10^{-6}$ per second, and $\alpha = 0.191$. And we primarily focus on a single type of content for ease of exposition [4]. From the proof of Theorem 1, we can see that some system property may be different when $w < 1$ than when $w > 1$. So, we compare the system behaviors for $w = 0.5$ and $w = 2$ when appropriate.

Figure 1 shows the optimum utility with respect to the delay d when the allowed cost C is small, medium, and large. When d is large, we can see that the system utility has negligible sensitivity with respect to the value of C. This is because when there is sufficient time to propagate the information, it is possible to make do with a small number of seed nodes regardless of the budget. When w is larger (e.g., in Fig 1), there is also little sensitivity on the value of C when d is small. This is because for w larger than 1, the system is more sensitive to the cost of seeds, and for small values of d there is not sufficient benefit from adding seed nodes (because there is not enough time for encounters to propagate the information) and in fact it costs more to recruit a seed than the resulting benefit. As a result, regardless of the cost budget there is little incentive to use many seed nodes for small d and the utility is insensitive in this case. But, when d is in between, the difference can be significant. As for the influence of w, the utility shows similar tendency regardless of w although the utility is more sensitive to C when $w = 0.5$.

Now we look into the optimum utility with respect to the allowed cost C in more detail through Figure 2. From the figure, we can see that the utility increases up to some point and stays there afterwards as C increases, for each d value. From the analysis, we know that the C value from which the utility is constant is actually $\widetilde{C}$. When d is small, the optimum utility increase for a large range of C, but the slope is very small, which means the sensitivity of the utility to C is small. As d increases, $\widetilde{C}$ decreases while the sensitivity increases. However, when d is large enough, only a small number of seeds is needed to satisfy most of the nodes, and so the cost constraint become less important. Note that we omit the plots for $w = 0.5$ because they look similar to those of $w = 2$ (Figure 2).

Figure 3 shows more directly how the unconstrained opti-

4. The nature of our optimization solution is that it yields for each type of content the optimal number of seeds. Once this is done, the problem decomposes into independent sub-problems, one for each item of interest. There is thus not a significant loss of generality in considering each type of content in isolation.



(a) $w = 0.5$ (b) $w = 2$

Fig. 3: Unconstrained opt. total cost vs. delay budget

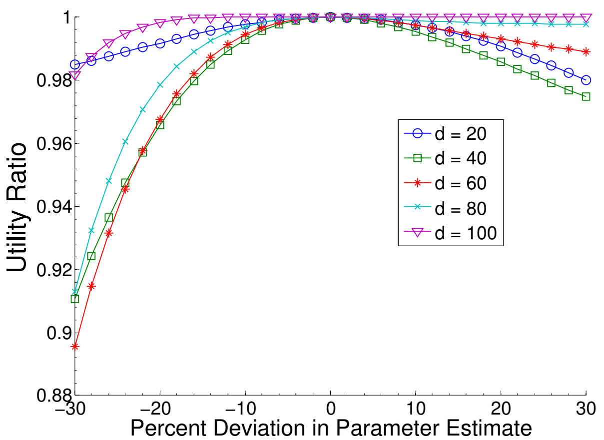


Fig. 4: Sensitivity of Utility to Errors in Parameter Estimation

mum total cost $\widetilde{C}$ changes as the allowed delay d changes. While the cost monotonically decreases as d increases when $w = 0.5$, *remarkably*, the cost initially *increases* and then decreases when $w = 2$. In fact, more generally, our model implies that the optimum cost monotonically decreases when $w \leq 1$, and shows a unimodal increase-then-decrease behavior when $w > 1$. Although perhaps counter-intuitive (since one may expect to see lower cost when the delay constraint is relaxed), this occurs because when $w > 1$, deploying one more seed requires more satisfied nodes besides itself. But when there isn't enough time to satisfy sufficiently many demanding nodes, the addition of seeds may not be productive. When the weight given to seed-cost is very important (high w) and the allowed delay is very small, our model suggests that it is sometimes better to not attempt to disseminate the content at all (zero seeds), depending on other system parameters like encounter rates.

We can also see that smaller portion of total nodes are needed to obtain the seeds for the optimum performance as the number of nodes increases. As for the influence of parameters α and β, we can see they only appear in z_i with d from (20), and d only appears with α and β. Therefore, α and β have effect of shrinking or stretching the performance plot in the direction of d as they increase or decrease, respectively.

Figure 4 shows how the Utility deteriorates for different amounts of percentage error in the product of the two parameters $\alpha\beta$ (note that the two parameters always show up in this product form in the analytical model), for $n = 632, w = 2$. This figure shows that for percentage deviations between -30 to 30 percent, the ratio of the Utility obtained with respect to that for perfect estimation (0% error) is generally above 0.9, suggesting that the model is quite robust to reasonable errors in the estimation. We hypothesize that the reason for this is that when the rate parameter is over-estimated, then the optimization yields that the number of satisfied nodes will be higher than it will be, and consequently chooses a smaller number of seed nodes than optimally needed. On the contrary, if it is under-estimated, then for the symmetric reasons, it chooses a larger number of seed nodes than optimally needed. Particularly when the utility function is cost-sensitive ($w > 1$), the latter has a greater negative impact on the utility. Thus, in this case, there could be a practical benefit to slightly over-estimating the beta parameter, for instance by adding an extra 5 to 10 percent to the estimate obtained statistically.

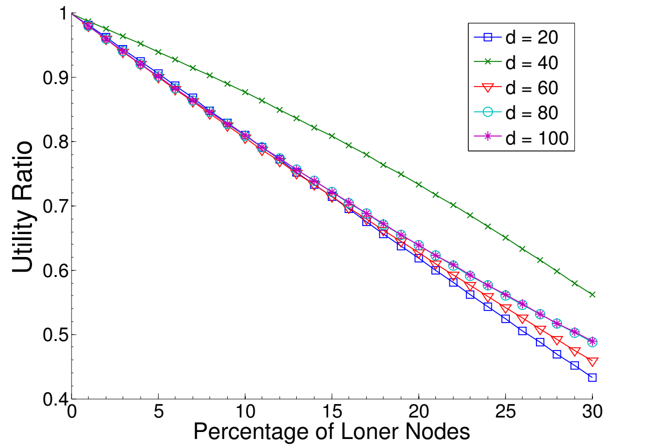


Fig. 5: Performance under Non-homogeneous mobility

4.4 Impact of Loner Nodes

Our analytical model for the utility assumes a homogeneous contact process where each node has the same encounter rates. In networks where there is greater heterogeneity, we can generally expect that the the number of satisfied nodes could be much lower, because of the presence of low-degree nodes that don't connect to many other nodes. In order to understand this a little more rigorously, we consider what happens in a worst-case heterogeneous network in which there is a core of homogeneous, well-connected nodes, and a set of nodes (say a fraction $x\%$ of them) that are completely disconnected. Intuitively, what happens in this case is that from the set of randomly chosen initial seeds a certain random number of the seed nodes will end up being loners and hence get wasted in the sense that they will not contribute to propagating the content to other nodes; the rest of the seed nodes that happen to be in the core will be useful for propagation. We can model the number of useful seed nodes as a Binomial random variable and condition upon it to determine the expected number of satisfied nodes, given that there are n nodes in total, $x\%$ of which are loner nodes. This in turn can be used to determine the expected utility for a given number of seed nodes and total network size.

Using this approach, Figure 5 shows how the expected

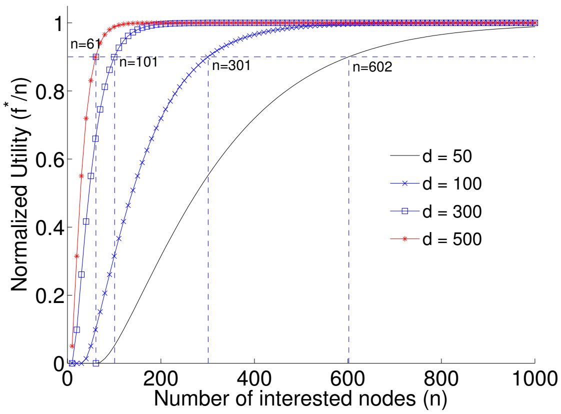


Fig. 6: Normalized utility vs the number of interested nodes

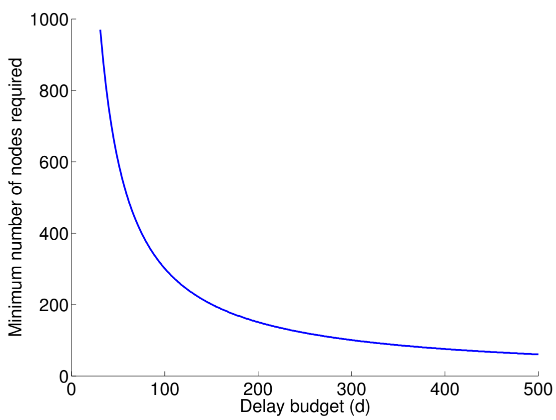


Fig. 7: Minimum number of nodes required for a given delay budget to achieve a normalized utility of at least 0.9

utility deteriorates as the percentage of loner nodes is varied, for an initial network size of $n = 1000$, assuming for content with $w = 2$, and the same α, β parameter values as before, for different delays (normalized to be 1 in the case when there are no loner vehicles). This figure shows that the performance would severely degraded in the presence of such loners, showing why it is important to identify such nodes and exclude them from the dissemination system if seeds are to be chosen at random. For this reason we will filter out "loner nodes" when we later present simulation results based on real vehicle traces.

4.5 Impact of Interest Popularity

Thus far we have focused on different interest groups separately and our scheme restricts only interested nodes to propagate the information. However, for items that are unpopular, i.e. if there are a small number of interested users, this scheme may perform poorly. To analyze this issue, we numerically examine how the utility depends upon the interest popularity (i.e., the number of interested users in a given item). Figure 6 shows how the normalized unconstrained utility varies with the interest popularity, for different values of delay constraint. As we can see the utility improves with interest popularity, and if the delay constraint is strict, then for items with a small number of interested vehicles, the utility is low. Figure 7 plots the minimum number of interested users needed to obtain a high normalized utility as a function of the delay constraint. Again, we see that the minimum number of users to obtain a desired performance level (in this case 90% of maximum possible utility) decreases as the delay constraint is relaxed. In practice, this means that for items with very small number of interested users, the proposed interested-node-only caching policy may need to be modified to seek the aid of other, helper nodes as well.

4.6 Impact of Churn

The analysis thus far has assumed that the number of nodes in the network is a constant and nodes do not join or leave the vehicular network during the process of dissemination. While this is a reasonable assumption for some kinds of vehicles whose presence is relatively stable (such as buses and taxis), it may not hold for other kinds of vehicles (such as personal cars). In the latter kind of system, some nodes may leave the network, while others join it, even as the content dissemination process is taking place. As a first-order modeling of this effect, we introduce a churn rate parameter ρ. For each item, we assume that the total number of interested users in the system N remains the same over time. However, there is a rate ρ at which users leave the network, and at the same rate new users join. This can be modeled by the differential equation $\frac{\partial s_i(k_i,t)}{\partial t} = \alpha\beta s_i(k_i,t)(n - s_i(k_i,t))(1-\rho)$ (this is effectively what in the medical epidemiology literature is referred to as the SI model with equal birth and death rate). It can be seen at a glance that the impact of the churn parameter is to reduce the effective encounter rate. This in turn implies that as the churn rate increases, the number of satisfied nodes by a given time will be smaller for a fixed number of seeds. We can therefore expect the utility of the system to decrease with increase in the churn rate. This is numerically shown in Fig 8. We see that when the delay constraint is tight, there is a marked deterioration in the utility as the churn rate increases. However, when the delay constraint is loose, then the utility remains nearly unaffected till the churn rate gets very large. Thus the sensitivity of the system to level of churn in the network turns out to be significantly dependent on the delay constraint.

4.7 Practical Solutions

In the previous subsections we explored the optimum behavior of the system theoretically. While the theoretical analysis brings better intuition of the system, it is also true that the solution is not either exact nor ready to use in practical systems because it is a continuous solution derived from the relaxed version of the problem (ignoring the integral constraint). The practical systems require integer values for the seed numbers. Hence, in this section, we develop a polynomial algorithm to obtain the exact discrete solution for $PF1$.

Algorithm 1 gives the optimum seed vector, each i-th element of which is integer-valued and in the range $[0, n_i]$. In a nutshell, the algorithm starts with zero seeds for all types, then increments the seed number of the type that gives the

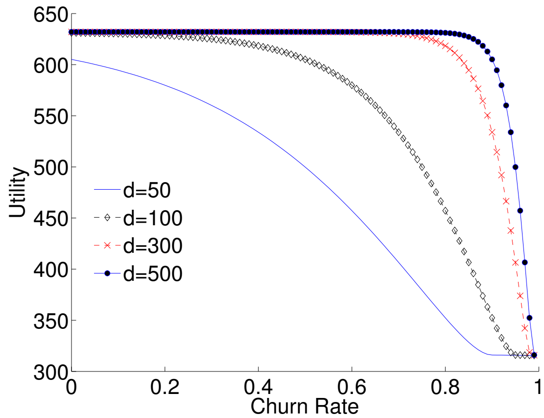


Fig. 8: Effect of the churn rate parameter on the utility for various delays

Algorithm 1 Optimizer(C, m)

```
k := m-sized array initialized to be all zero.
for (c = 0; c < C; c+ = 1) do
    i* := 0 ; δ_max := −∞
    for (i = 1; i ≤ m; + + i) do
        δ := f({k[1], . . . , k[i] + 1, . . . , k[m]}) − f(k)
        if (δ > δ_max) then
            i* := i ; δ_max := δ
        end if
    end for
    if (δ_max ≤ 0) then
        break
    end if
    k[i*]+ = 1
end for
return k
```

maximum increase in the system utility, as long as the total access cost is not over budget and the increase in the system utility is positive. Its correctness is proven in [16]. We resort to omitting the proof due to the short of spaces in this paper. It is easy to see that its time complexity is $O(m^2C)$, where m is the number of types of content, and C is the allowed cost.

In figure 9, the continuous solution from the closed-form analysis is compared numerically with the discrete optimum solution obtained with Algorithm 1, for a system with two content types: 158 nodes interested in content 1 and 474 interested in content 2, a total cost constraint of $C = 100$, and $w = 2$. As this figure shows, in practice, there is a negligible gap between the two. This figure also shows that for small delays it turns out to be optimal to allocate no seeds at all to type 1.

5 Simulation Based on Taxi Traces

In this section we present how the contents dissemination behaves in a more realistic setting. We consider a single type of content in this section because the process of the dissemination does not depend on other contents as shown in Section 3.

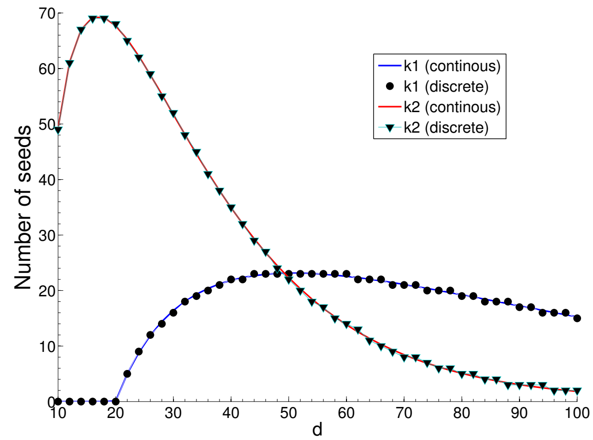


Fig. 9: Numerical comparison of discrete optimum (Alg. 1) with continuous optimum (obtained analytically)

5.1 Beijing Taxi Traces

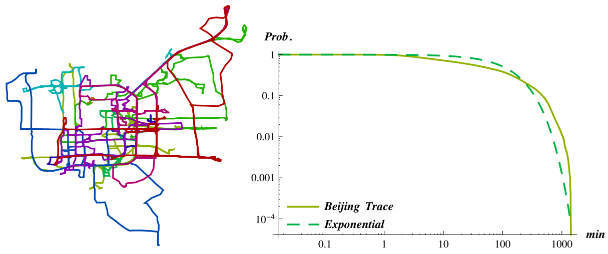
We use the GPS traces of taxis in Beijing gathered from 12:00am to 11:59pm on Jan. 05, 2009 in the local time. The number of subject taxis is 2,927. The number of the GPS points in the trace is 4,227,795, typically one per minute per vehicle. The GPS points span from $32.1223°$ N to $42.7413°$ N in latitude, and from $111.6586°$ to $126.1551°$ in longitude. Figure 10a shows the GPS traces of randomly chosen 10 taxis as an example.

5.2 Encounter Processes

In order to perform a simulation for the contents dissemination through the short-range radio, we need traces of encounters of all pairs of nodes; that is, when which vehicle can communicate with which other vehicle. We can extract these traces from the GPS traces by assuming a radio model. In this paper we assume the circular radio model to decide if two given vehicles encounter each other so that they can communicate directly. The circular radio model has the radio range r so that any two vehicles of distance within r can directly communicate with each other successfully. We use $r = 300$ meters as the literature ([17]) suggests[5].

Suppose a set of error-free time-ordered GPS traces of a pair of vehicles is given. In order to obtain the time-ordered traces of encounters for the pair, we have compared their geodesic distances in some sequence of times. Instead of employing a time sequence of identical intervals, we have checked the distance after the minimum time τ_{min} (its expression is given below) that the pair can encounter each other next, if the

5. We chose a deterministic nominal range primarily for simplicity. The impact of a more realistic radio model would be to alter the number of pairwise contacts between vehicles. In the proposed analytical model, this would translate effectively to a change in the encounter rate parameters α, β. Thus, for instance, if a more realistic communication model results in a lower rate of encounters in expectation, it would result in lower values of these parameters, and in the analytical formulation, this would decrease the expected number of satisfied nodes for a given number of seed nodes, potentially shifting the optimum number of seeds to a higher number (depending on other parameters such as w).



(a) (b)

Fig. 10: Properties of Beijing Taxi Traces

current distance is large enough, for the faster processing and more accurate results. When the current distance is small, we have checked their new distance after a predetermined small time step.

Since the different vehicles do not log their GPS locations at exactly the same times, we cannot simply take the locations of the pair from the logs at a pre-specified time. So, we have interpolated the locations of each vehicle assuming that the GPS traces are dense enough so that a vehicle can be approximated to move in a straight line between a consecutive pair of GPS locations in the traces.

The minimum time τ_{min} for the next encounter is $\tau_{min} = \frac{1}{2s_m}(\textsc{GeoDist}(pos(P_1,t), pos(P_2,t)) - r)$ where GeoDist gives the geodesic distance between the given pair of GPS positions, $pos(P_i,t)$ calculates the estimated position of vehicle i at time t from the set of its GPS traces P_i by interpolating the positions, and s_m is the maximum speed of vehicles in the traces.

We then obtain the time-ordered set of encounters of all pairs by executing the aforementioned algorithm for each pair and sorting their combined result.

We note that the input sets of GPS traces to the algorithm are required to be error-free. However, we have found, as expected, that some GPS units of vehicles experienced errors in some time intervals, so either some erroneous log was reported or there was no data at all in the interval. After removing those erroneous GPS points, we have checked if this removal incurs some side effects. We have found that the removal makes some vehicles untraceable in some non-ignorable time intervals. In other words, some vehicles have no valid GPS points reported for long intervals. And it is difficult to approximate their positions for the duration by interpolating the valid positions. Hence, we resort to excluding those vehicles from the simulation.

Therefore, we have selected 632 vehicles, each of which satisfies the following criteria:

- The GPS points indicating the speed of 80 mph or more are considered erroneous and removed. It is because the speed of more than 80 mph is hard to reach and rarely exercised in the Beijing area.
- The valid GPS points of each vehicle are logged somewhat regularly in time when it is moving so that any two consecutive GPS points of the vehicle do not have a distance of more than 400 meters if their time difference is more than 3 minutes.
- The encounter graph of vehicles forms a well connected graph so that the number of neighbors of a node is at least 2. The encounter graph is defined in Definition 1.

The second condition makes sure that the vehicle has not moved actively when it skipped two consecutive regular GPS reports. We set the distance of 400m so that we can have a better understanding on the timing of encounters (with some tolerance) in the interval of the reports, when the radio range is 300m. The last condition is to remove loner vehicles (which contacted either 0 or only 1 other vehicle during the entire trace). We note that these loner vehicles have almost no interaction with others at all, which means they are in the very different activity region. But, we are interested in the dissemination over the nodes of similar activity region. We showed earlier the severe deterioration in utility that results when loner vehicles are included in the system.

Definition 1 (Encounter Graph)**.** *An* ***encounter graph*** $G(V,E)$ *of vehicles is a graph such that each vehicle is represented by a node* $v \in V$*, and any two nodes* $v_1, v_2 \in V$ *have a link* $e(v_1,v_2) \in E$ *between them if and only if they can communicate with each other (*i.e. *encounter) at any point in the interested time interval.*

The encounter graph of the 632 nodes has 38,139 links; the minimum number of neighbors of a node is 2, the maximum is 261, and the median is 120. Their average number is 120.693. This value is used in later sections for evaluating our model for the number of satisfied nodes.

5.3 Inter-Encounter Time

In this section we analyze on the inter-encounter time of a pair of nodes in order to verify the Exponential assumption of the inter-encounter time and to obtain its rate for evaluating our model.

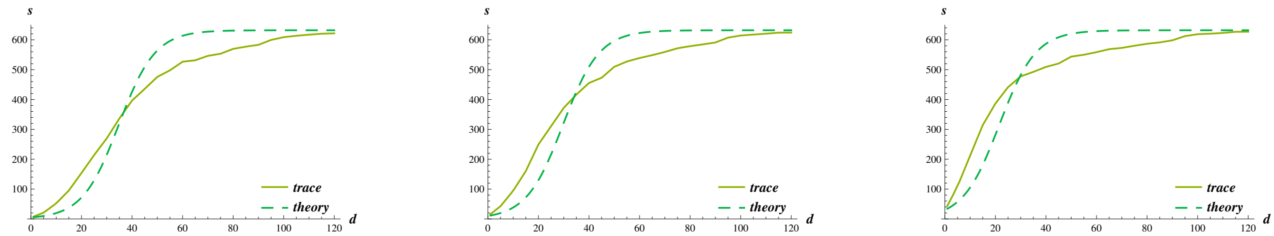
Although the trace data is fine-grained and covers 24 hours of a day, many pairs of nodes have only a few encounters, which is too small to have a good statistical meaning if we focus on the per-pair distribution. So, we hypothesize that the inter-encounter time of every pair follows the identical and independent distribution, particularly, the Exponential distribution as we assume in the analysis in Section 3.

We first examine the aggregate inter-encounter time collecting the available inter-encounter times of every consecutive encounters of all pairs of nodes. The number of samples is 24,205, and their sample mean is 150.005 minutes. Figure 10b shows the tail distribution of the samples and the Exponential distribution with mean 150.005 minutes. As can be seen, they do not show big disparities[6].

Because we assume IID Exponential distributions for per-pair inter-encounter times, their aggregate inter-encounter time has the identical distribution to the per-pair ones, which can be proved easily.

However, the above sample mean for the inter-encounter time is actually an underestimate of the true mean because we

6. This shows that the exponential inter-encounter distribution is a reasonable assumption when considering a vehicular network where the encounter graph is well connected.



(a) k=5 (b) k=10 (c) k=30

Fig. 11: Average number of satisfied nodes vs. tolerable delay

ignore many incomplete samples, which is the time from the beginning of the trace to the first encounter and the time from the last encounter to the end of the trace for each pair of nodes. For each of those samples of time duration, we know that its associated realization of the inter-encounter time is larger than the time duration, but we do not know the exact value. That is why we exclude them from the above estimation. But, now that we have the reason (*i.e.* Figure 10b) to believe that it is fine to assume the Exponential distribution for the inter-encounter time, we can use the incomplete information to obtain a more accurate estimate.

We use the fact that the number of encounters in a time interval T follows the Poisson distribution with mean βT, when the inter-encounter time is Exponential with rate β. Suppose N_i and T_i are the number of encounters and the whole time duration of the trace, respectively, for i-th pair of nodes, and η the number of the pairs that have at least one encounter in the trace. Then, the following equation gives the maximum likelihood estimate β^* of β.

$$\beta^* = \arg\max_\beta \Pr(N_1, N_2, ..., N_\eta | \beta, T_1, ..., T_\eta) \tag{21}$$

where

$$\begin{aligned} &\Pr(N_1, N_2, ..., N_\eta | \beta, T_1, ..., T_\eta) \\ &= \prod_{i=1}^{\eta} \Pr(N_i | \beta T_i) = \prod_{i=1}^{\eta} \frac{(\beta T_i)^{N_i} e^{-\beta T_i}}{N_i!} \\ &= (\prod_{i=1}^{\eta} \frac{T_i^{N_i}}{N_i!}) e^{-\beta \sum_{i=1}^{\eta} T_i} \beta^{\sum_{i=1}^{\eta} N_i} \end{aligned} \tag{22}$$

Note that (22) holds because the inter-encounter times of every pair are assumed to be jointly independent.

After some calculations, we can obtain the maximum likelihood estimate of the rate of the inter-encounter time of a pair of nodes that ever encounter, as follows:

$$\beta^* = \sum_{i=1}^{\eta} N_i \,/\, \sum_{i=1}^{\eta} T_i \tag{23}$$

We shall use this quantity as a parameter value to evaluate our analytical model and compare with the real-trace-based simulation results.

5.4 Simulation Methodology

From the time-ordered traces of the encounters of the Beijing traces, produced by the method in Section 5.2, we have performed the simulations by running Algorithm 2 multiple times until the sample mean of the number of returned satisfied nodes has its error no more than 5% of its value with 97% confidence. Algorithm 2 takes several input arguments; EOL (which stands for Encounter Ordered List) is a time-ordered list of encounters, N is the set of vehicles, $S \subset N$ is the set of seed nodes, t_s is the time when S are deployed, and d is the delay budget. We have performed the simulations for various choices for the number of seeds k and the tolerable delay d, letting the seeds be deployed at time t_s = 9AM. For particular k and d, we have chosen the seed nodes S uniformly at random at each round.

Algorithm 2 SATISFIEDNODES(EOL, N, S, t_s, d)

```
Mark every v ∈ S as satisfied.
for all e ∈ EOL in order of t = time(e), s.t. t_s ≤ t ≤ t_s+d do
    Let v_1 and v_2 be the pair of vehicles for e.
    if only one of v_1 and v_2 is marked satisfied then
        Mark the other node as satisfied.
    end if
end for
return the set of all marked nodes
```

5.5 Number of Satisfied Nodes

Figure 12 shows the average number of satisfied nodes with respect to the number of seeds when the delay constraints are 10, 30, and 60 minutes. When the delay is small (*i.e.* 10 minutes), the real traces suggest more nodes are expected to be satisfied than the theory predicts. When the delay is medium (*i.e.* 30 minutes), the real traces and the theory suggest similar behavior of the dissemination, while the theory overestimates the number of satisfied nodes when the delay is 60 minutes. But, the figure shows qualitatively similar behavior of the average number of satisfied nodes as the number of seeds increases.

Figure 11 shows in more detail how the gap between the theory and the trace suggest changes as the delay constraint increases. The numbers of seeds considered are 5, 10, and 30. And all the cases indicate similar trends of the content dissemination; the real traces suggest that the dissemination is faster than the theory predicts in the early phase, but loses its momentum as more portion of nodes are infected. This difference may be because of the movement dependencies between groups of vehicles in reality. Suppose there is some dependency among the pair-wise encounter processes that is caused by the movement dependency. It is easy to see that the content spread faster to the other nodes of positive correlation than the average, and slower to the nodes of negative correlation. Hence, in the early phase of the dissemination, the content spreads fast to positively correlated nodes, and after consuming most of them, it spreads slowly to the nodes of negative correlation. This can partly address the gap in Figure 11. But, more accurate analysis calls for further investigation, which is out of scope of this paper and the subject of our future research.

Nevertheless, the system behavior with respect to the number of seeds is more important for our problem because it is the parameter to optimize on. And, Figure 12 suggests comparable numbers of seeds for the knees of plots from the theory and the real traces.

5.6 Optimal Number of Seeds

Now we look into the system utility f with respect to the number of seeds. We have compared the system utilities[7] that our model predicts and the Beijing traces suggest, with various delay constraints and cost weights. It turns out they show similar behaviors as in Figure 12; the real traces suggest larger utility values than what the theory predicts when the delay is small. Their difference decreases as the delay budget increases up to some point, after which the difference increases again. In this case the real traces suggest smaller utility values than that of theory. They however share similarities in the shape and trends in the similar manner as in Figure 12.

We also examine how good our analytic solution of the optimal number of seeds, k^*_{thr}, would be in the realistic setting induced from the Beijing traces. Figure 13 shows the optimal number of seeds and the corresponding empirical system utility with respect to the delay budget. Figure 13a compares the empirical optimal number of seeds k^*_{sim} and its analytical counterpart k^*_{thr}. We can see from the figure that k^*_{sim} and k^*_{thr} are getting closer to each other as the delay budget d increases. Although k^*_{sim} and k^*_{thr} have big differences when the delay budget is small, we note that the utility function has a very gentle slope near its optimum in this small delay regime (see Figure 12). This is why our analytical solution provides near-optimal performance even in the small delay regime as can be seen in Figure 13b.

Figure 13b compares the best possible system utility values f^*_{sim} of the trace-based simulations and the empirical utility values $\widetilde{f}_{sim}$ when our solution k^*_{thr} is used. In other words, the figure shows how close the system utility of the real system would be to the system's best possible utility if the system uses our analytic solution. As can be seen, the system utilities in the real world would be within 95% of their real maximums over the entire delay regime if our theoretical optimizers are used. Therefore, these results support the usefulness of our model.

7. We omit the corresponding figure because it looks similar to Figure 12 and due to the lack of pages.

6 Protocol Implementation Sketch

The basic idea of seeding content in vehicles through an infrastructure-based always-on radio and disseminating it further through epidemic contacts certainly seems feasible in light of the substantial existing literature on protocols for delay tolerant and vehicular networks, which have included several alternative proposals for neighbor discovery, interest discovery, distributed data management, and epidemic routing protocols (see [18] for a comprehensive survey). Beyond conceptual proposals evaluated through simulations, there are now several practical implementations of protocols and systems for neighbor discovery [19], vehicle to vehicle link establishment and data transfer [20], [21], [22], deployments on multi-car V2V testbeds [23], [24], [25], [26] and integration of cellular and vehicular radios [27].

As our focus in this paper has been on presenting an optimization framework for this problem, we do not propose or evaluate in this work a detailed protocol-level specification to instantiate the proposed system. However, a rough sketch of how the concept proposed in this paper may be implemented is as follows. The basic idea is to have a two-tier architecture for the vehicular networks, with a cellular-based centralized control plane and a vehicle-to-vehicle distributed data plane. The cellular-based centralized control plane is used to monitor aggregate interest levels in various kinds of content and to estimate relevant statistical parameters (such as the inter-vehicular interaction rate). When a content needs to be disseminated, based on the estimates of these parameters, and the desired application-specific deadline for dissemination, the number of seeds required is calculated, and the content is first downloaded directly through the cellular radio to this number of cars. For further optimization, if finer-grained information about vehicular contact patterns is available (such as encounter degree), the initial seed nodes may be chosen more carefully rather than uniformly at random. In the vehicular data-plane, cars periodically, or on observation of nearby cars, first exchange information to identify the content they have available to transmit and the content they are interested in. If and when a car has a content the other is interested in, this content is transferred (with possibly some form of prioritization if multiple contents need to be exchanged). Over time, the empirically obtained utility may be statistically tracked and used to further adapt the number of seeds so that it is not based purely on theoretical calculations, but rather optimized in a data-driven manner.

It would be of great interest to see in future work a fleshed-out protocol-level design, implementation, and empirical evaluation of a mechanism along these lines on a real vehicular network with heterogeneous radios.

7 Related Work

In the past decade, extensive research has been done to study the technical feasibility of heterogeneous integrated wireless networks. Some of this has focused on integrating wireless local area networks and cellular networks to allow for vertical handoffs [28]. There has also been work on integrating mobile ad hoc networks (MANET) and cellular systems to improve

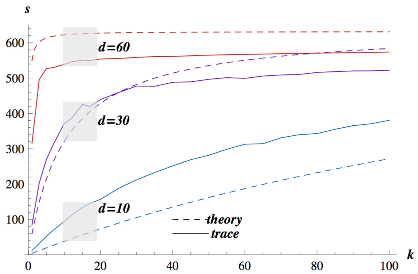
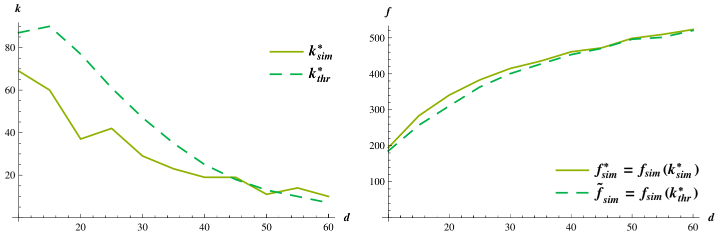


Fig. 12: Avg. #satisfied vs. #seeds



(a) Optimal Seed Number (b) System utility

Fig. 13: System behaviors in optimal regime w.r.t the delay budget

throughput and increase coverage [29], [30], and there has been theoretical analysis of the capacity of such heterogeneous networks [31], [32].

In common with these works, we too propose the integration of the cellular network with another mobile network, however in our context the other mobile network is a delay-tolerant network (DTN) that uses "store-carry-forward" approach for content dissemination. Also, unlike much of the prior focus on capacity improvements, our focus is primarily on maximizing content dissemination within a delay deadline while minimizing the cost of cellular access, though certainly our approach will also free up scarce cellular bandwidth.

Delay-tolerant networking (DTN) is a new network architecture that provides meaningful data service to challenged networks in which continuous network connectivity is not guaranteed [33], such as sparse vehicular networks when such networks are deployed at the first few years [34]. The initial effort for tackling Delay Tolerant Networks was placed on designing reliable and efficient routing protocols under a variety of assumptions on mobility [35], [36]. Encouraged by the above promising results, researchers have explored using opportunistic connections between vehicular nodes to implement delay-tolerant network protocols and applications in empirical testbeds [37], [9]. Our work on vehicular heterogeneous networks is complementary to the above studies on "pure" DTNs.

In sparse DTNs, mobile node encounters are utilized for opportunistic data transfer, and thus the underlying mobility model has a great impact on their performance. The conventional Random Walk model and Random Waypoint model are normally used to evaluate DTN protocols [36], [38]. In order to validate our analysis in a more credible setting, we have used a real large-scale vehicular mobility trace from a large metropolitan area (Beijing) in our study, one of the first studies to do so (a methodology adopted in another recent study [39]).

In our study, we use differential equations to model content replication and dissemination. This is similar to [39], where differential equations are used to model the age of content updates and are found to be a good approximation for large networks. There have been several other prior studies on content dissemination and replication in vehicular networks. In [40], the authors explore the latency performance of different frequency-based replication policies in the context of vehicular networks with limited storage. CarTorrent [41] and AdTorrent [42], present content dissemination mechanisms to distribute files and advertisements, respectively, in vehicular networks. In [8], the authors study how user impatience affects content dissemination. Different from these studies, our focus in this work is on a novel cost optimization problem for disseminating content to the maximum number of vehicles within a given deadline, that leverages both the cellular infrastructure and peer-to-peer vehicular communication.

There has been some research work on cellular multicast, to improve the efficiency of cellular network utilization for multicast applications [43], [44]. However, these works are primarily aimed at improving efficiency in dense settings where the demanding nodes are all or significantly localized within each cell where the multicast takes place. While these techniques can be complementary to the solution proposed in this work, further improving the utility and delay, they are not sufficient in themselves; when propagating content to vehicles city-wide, there may not be sufficient density in individual cells to benefit substantially from cellular multicast.

8 Conclusion

We have investigated the optimum content dissemination in the heterogeneous vehicular network in this work. In this network, each vehicle is equipped with one costly, long-range, low-bandwidth cellular radio and another low-cost, short-range, high-bandwidth radio. We have considered the problem of how to spread relevant content to more vehicles with smaller cost. We have developed the relevant optimization formulation and derived their analytical solutions with some relaxation. One interesting takeaway point is that the contents can be disseminated to a large number of vehicles with a few costly access to the infrastructure, if some delay can be tolerated. We have also developed a polynomial algorithm to calculate the exact optimum seed vector with no relaxation. To verify our analysis and see to what extent the assumptions and approximations made in match reality, we have performed simulations based on the real GPS traces of 632 taxis gathered in Beijing, China.

We believe the modeling presented in this paper makes an important advance in understanding how to mathematically formulate and optimize performance for such problems in vehicular networks. Nevertheless, there is significant room for improvement. Real vehicular systems (including those from our traces) show significant spatio-temporal variation in the encounter rates. For future work, it would also be of interest to conduct simulations with traces obtained from personal vehicles rather than taxis, which are likely to show much greater variation in encounter activity over time. We believe that dividing the day into hour-long time slots and estimating the parameter for each hour separately would result in a better match between the analysis and simulations. Enhancements to the model to consider more non-homogeneous, non-independent encounters, would also be desirable, but in our experience, are likely to be difficult to obtain without sacrificing the tractability and insight provided by the closed-form analysis presented here.

We have not proposed or evaluated a full protocol implementation of the proposed idea in this paper other than the brief sketch outlined in section 6. We hope to see an implementation on a real vehicular network testbed in the future.

References

[1] CISCO, "Cisco visual networking index forecast predicts continued mobile data traffic surge," Feb. 2010. available online at http://www.cisco.com/web/MT/news/10/news_220210.html.

[2] NYTimes, "Customers angered as iphones overload at&t," september 2009.

[3] C. Chan and G. Wu, "Pivotal Role of Heterogeneous Networks in 4G Deployment," in *ZTE Technologies, China*, January 2010.

[4] A. Rath, S. Hua, and S. Panwar, "Femtohaul: Using femtocells with relays to increase macrocell backhaul bandwidth," in *IEEE INFOCOM*, 2010.

[5] O. Tipmongkolsilp, S. Zaghloul, and A. Jukan, "The evolution of cellular backhaul technologies: Current issues and future trends," in *IEEE Communications Surveys & Tutorials*, 2010.

[6] F. Bai and B. Krishnamachari, "Exploiting the wisdom of the crowd: Localized, distributed information-centric vanets," *IEEE Communications Magazine*, vol. 48, no. 5, 2010.

[7] G. Sharma and R. Mazumdar, "Scaling laws for capacity and delay in wireless ad hoc networks with random mobility," in *IEEE ICC*, 2004.

[8] J. Reich and A. Chaintreau, "The age of impatience: optimal replication schemes for opportunistic networks," in *ACM CoNEXT*, 2009.

[9] J. Eriksson, H. Balakrishnan, and S. Madden, "Cabernet: vehicular content delivery using wifi," in *ACM MobiCom*, 2008.

[10] T. Zahn, G. O'Shea, and A. Rowstron, "Feasibility of content dissemination between devices in moving vehicles," CoNEXT '09, 2009.

[11] A. Vahdat and D. Becker, "Epidemic routing for partially-connected ad hoc networks," Tech. Rep. CS-2000-06, UCSD, 2000.

[12] R. Groenevelt, P. Nain, and G. Koole, "The message delay in mobile ad hoc networks," *Elsevier Journal of Performance Evaluation*, 2005.

[13] Z. Haas and T. Small, "A new networking model for biological applications of ad hoc sensor networks," *IEEE/ACM Trans. Netw.*, 2006.

[14] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

[15] M. J. J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2007.

[16] J. Ahn, B. Krishnamachari, F. Bai, and L. Zhang, "Optimizing Content Dissemination in Heterogeneous Vehicular Networks," Tech. Rep. CENG-2010-2, Univ. of Southern California, 2010.

[17] F. Bai and H. Krishnan, "Reliability analysis of dsrc wireless communication for vehicle safety applications," in *IEEE ITSC*, 2006.

[18] S. Basagni, M. Conti, S. Giordano, and I. Stojmenovic, "A taxonomy of data communication protocols for vehicular ad hoc networks," *Mobile Ad Hoc Networking:The Cutting Edge Directions*, 2013.

[19] A. Vinel, D. Staehle, and A. Turlikov, "Study of beaconing for car-to-car communication in vehicular ad-hoc networks," in *Communications Workshops, 2009. ICC Workshops 2009. IEEE International Conference on*, pp. 1–5, IEEE, 2009.

[20] S. Iqbal, S. R. Chowdhury, C. S. Hyder, A. V. Vasilakos, and C.-X. Wang, "Vehicular communication: protocol design, testbed implementation and performance analysis," in *Proceedings of the 2009 International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly*, pp. 410–415, ACM, 2009.

[21] T. Zahn, G. O'Shea, and A. Rowstron, "Feasibility of content dissemination between devices in moving vehicles," in *Proceedings of the 5th international conference on Emerging networking experiments and technologies*, pp. 97–108, ACM, 2009.

[22] B. Yu and F. Bai, "Etp: Encounter transfer protocol for opportunistic vehicle communication," in *INFOCOM, 2011 Proceedings IEEE*, pp. 2201–2209, IEEE, 2011.

[23] M. Cesana, L. Fratta, M. Gerla, E. Giordano, and G. Pau, "C-vet the ucla campus vehicular testbed: Integration of vanet and mesh networks," in *Wireless Conference (EW), 2010 European*, pp. 689–695, IEEE, 2010.

[24] M. Gerla, J.-T. Weng, E. Giordano, and G. Pau, "Vehicular testbeds-model validation before large scale deployment," *Journal of Communications*, vol. 7, no. 6, pp. 451–457, 2012.

[25] J. Ahn, Y. Wang, B. Yu, F. Bai, and B. Krishnamachari, "Risa: Distributed road information sharing architecture," in *INFOCOM, 2012 Proceedings IEEE*, pp. 1494–1502, IEEE, 2012.

[26] D. Raychaudhuri and M. Gerla, *Emerging Wireless Technologies and the Future Mobile Internet*. Cambridge University Press, 2011.

[27] B. Roodell and M. I. Hayee, "Development of a low-cost interface between cell phone and dsrc-based vehicle unit for efficient use of intellidrivesm infrastructure," *Intelligent Transportation Systems Institute Report*, Novemeber 2010.

[28] A. Salkintzis, C. Fors, and R. Pazhyannur, "Wlan-gprs integration for next-generation mobile data networks," *IEEE Comm. Mag.*, 2002.

[29] H. Wu, C. Qiao, S. De, and O. Tonguz, "Integrated cellular and ad hoc relaying systems: icar," in *IEEE JSAC*, 2001.

[30] B. Bhargava, X. Wu, Y. Lu, and W. Wang, "Integrating heterogeneous wireless technologies: a cellular aided mobile ad hoc network (cama)," *Mob. Netw. Appl.*, vol. 9, no. 4, 2004.

[31] B. Liu, Z. Liu, and D. Towsley, "On the capacity of hybrid wireless networks," in *IEEE INFOCOM*, 2003.

[32] L. K. Law, S. V. Krishnamurthy, and M. Faloutsos, "Capacity of hybrid cellular-ad hoc data networks," in *IEEE INFOCOM*, 2009.

[33] K. Fall, "A delay-tolerant network architecture for challenged internets," in *ACM SIGCOMM*, 2003.

[34] F. Bai and B. Krishnamachari, "Spatio-temporal variations of vehicle traffic in vanets: facts and implications," in *ACM VANET*, 2009.

[35] Q. Yuan, I. Cardei, and J. Wu, "Predict and relay: an efficient routing in disruption-tolerant networks," in *ACM MobiHoc*, 2009.

[36] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Efficient routing in intermittently connected mobile networks: the single-copy case," *IEEE/ACM Trans. Netw.*, vol. 16, no. 1, 2008.

[37] X. Zhang, J. Kurose, B. N. Levine, D. Towsley, and H. Zhang, "Study of a bus-based disruption-tolerant network: mobility modeling and impact on routing," in *ACM MobiCom*, 2007.

[38] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Efficient routing in intermittently connected mobile networks: the multiple-copy case," *IEEE/ACM Trans. Netw.*, vol. 16, no. 1, 2008.

[39] A. Chaintreau, J.-Y. Le Boudec, and N. Ristanovic, "The age of gossip: spatial mean field regime," in *ACM SIGMETRICS*, 2009.

[40] S. Ghandeharizadeh, S. Kapadia, and B. Krishnamachari, "Comparison of replication strategies for content availability in C2P2 networks," in *MDM*, 2005.

[41] K. Lee, S.-H. Lee, R. Cheung, U. Lee, and M. Gerla, "First experience with cartorrent in a real vehicular ad hoc network testbed," in *IEEE MOVE*, 2007.

[42] A. Nandan, S. Das, B. Zhou, G. Pau, and M. Gerla, "Adtorrent: Digital billboards for vehicular networks," in *IEEE/ACM V2VCOM*, 2005.

[43] H. W. et al., "Multicast scheduling in cellular data networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 9, 2009.

[44] T.-P. L. et al., "Optimized opportunistic multicast scheduling (oms) over wireless cellular networks," *IEEE Transactions on Wireless Communications*, vol. 9, no. 2, Feb. 2010.

Joon Ahn received his B.S. degree in Electrical Engineering from Seoul National University, Seoul, Korea, in 2000. He received his Ph.D. degree in 2011 in the Department of Electrical Engineering at the University of Southern California. He received the Best Student Paper Award from the Electrical Engineering-Systems Department at the University of Southern California in 2006. He is currently with Ericsson, Inc. as a system architect. His research interests are in the areas of distributed system, mobile networks, and ad-hoc networks.

Maheswaran Sathiamoorthy is a PhD Candidate at the University of Southern California. He received his bachelors from the Indian Institute of Technology Kharagpur, India in 2008. He was an Annenberg Fellow from 2008-2012. His research interests are in distributed storage for vehicular network clouds and data center clouds.

Bhaskar Krishnamachari received his B.E. in Electrical Engineering at The Cooper Union, New York, in 1998, and his M.S. and Ph.D. degrees from Cornell University in 1999 and 2002 respectively. He is currently an Associate Professor and a Ming Hsieh Faculty Fellow in the Department of Electrical Engineering at the University of Southern California's Viterbi School of Engineering. His primary research interest is in the design and analysis of algorithms and protocols for next-generation wireless networks.

Fan Bai (General Motors Global R&D) is a Senior Researcher in the Electrical & Control Integration Lab., Research & Development and Planning, General Motors Corporation, since Sep., 2005. Before joining General Motors research lab, he received the B.S. degree in automation engineering from Tsinghua University, Beijing, China, in 1999, and the M.S.E.E. and Ph.D. degrees in electrical engineering, from University of Southern California, Los Angeles, in 2005.

His current research is focused on the discovery of fundamental principles and the analysis and design of protocols/systems for next-generation Vehicular Ad hoc Networks(VANET), for safety, telematics and infotainment applications. Dr. Bai has published about 50 book chapters, conference and journal papers, including Mobicom, INFOCOM, MobiHoc, SECON, ICC, Globecom, WCNC, JSAC, IEEE Transaction on Vehicular Technology, IEEE Wireless Communication Magazine, IEEE Communication Magazine and Elsevier AdHoc Networks Journal. He received Charles L. McCuen Special Achievement Award from General Motors Corporation ”in recognition of extraordinary accomplishment in area of vehicle-to-vehicle communications for drive assistance & safety”. He serves as Technical Program Co-Chairs for IEEE WiVec 2007, IEEE MoVeNet 2008, ACM VANET 2011 and ACM VANET 2012. He is an Associate Editor of IEEE Transaction on Vehicular Technology and IEEE Transaction on Mobile Computing, and he also serves as guest editors for IEEE Wireless Communication Magazine, IEEE Vehicular Technology Magazine and Elsevier AdHoc Networks Journal. He is also serving as a Ph.D. supervisory committee member at Carnegie Mellon University and University of Illinois - Urban Champaign.

Lin Zhang (B.Sc. '98, M.Sc. '01, Ph.D. '06, all from Tsinghua University, Beijing, China) is currently an associate professor at Tsinghua University. His current research focuses on sensor networks, data and knowledge mining, and information theory. He is a co-author of more than 40 peer-reviewed technical papers and five U.S. or Chinese patents applications. Lin and his team were also the winner of IEEE/ACM SenSys 2010 Best Demo Awards. In 2010, he received Excellent Teaching Awards from Tsinghua University.