Optimizing Content Dissemination in Heterogeneous Vehicular Networks*

Joon Ahn†, 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, 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 heterogeneous 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 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 mathematically analyze the dissemination process using differential equations and convex optimization and derive a closed-form optimal solution. We then examine numerically the tradeoffs between cost, delay and system utility in the optimum regime. We have found that, in the optimum regime, (a) system utility, which is essentially the extra benefits induced by the short-range radio, 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. Since our closed form analysis provides only a continuous solution, we also develop a polynomial-time algorithm to obtain the optimal discrete solution needed in practice. We verify our analysis using real GPS traces of 632 taxis in Beijing, China. The key finding of this work is that content can be spread to a large number of vehicles with minimal use of the cellular infrastructure at low cost, if some delay is allowed.

1. INTRODUCTION

Recent trends in the automotive industry point to an emerging age of heterogeneous 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). As the cellular bandwidth becomes increasingly crowded and more expensive, we contend that hybrid protocols that synergistically combine direct cellular access along with store and forward routing will prove efficient and cost-effective.

We consider the problem of efficient dissemination of some delay-tolerant content (*e.g.* advertisement, traffic information, weather forecast, video/audio clips) to a group of vehicles that share an interest in this content. In light of the fact that cellular radios in cars would allow only for unicast communication, and therefore incur a unit per-vehicle charge for content download, the use of the free short-range radio to assist in such a broad dissemination process becomes economically compelling. We formulate in this work an optimization problem 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: we analyze mathematically the content dissemination process using differential equations, and derive the optimum solution for the problem in closed-form by solving a convex optimization problem. We then investigate the behaviors of the system in terms under various optimal 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. 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 under an hour with very low-cost use of the cellular infrastructure.

This paper is organized as follows. We first formalize the optimization problem in Section 2. In Section 3, we derive one of our key measures, the expected number of satisfied vehicles by the dissemination, using ordinary differential equation (ODE) based modeling. The core optimization problem and its solution is then investigated in Section 4. Then, we develop an algorithm to calculate the practical optimum solution overcoming the limitation of the analytical solution in Section 5. We introduce the Beijing taxi traces and use them to verify our analysis in Section 6. We introduce related work and discuss the differences and novelty of our work in Section 7. Finally, we present concluding comments in Section 8.

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.

*Research was sponsored in part by a contract from General Motors and in part by the U.S. Army Research Laboratory under the Network Science Collaborative Technology Alliance, Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

This method incurs significant access fees proportional to the number of the interested vehicles although the associated delay would be small. On the other extreme is to send the message to one vehicle only 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. Each long-range radio access incurs a unit cost which is assumed one in the paper 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-es, where $\mathbf{k} = (k_1, k_2, ..., k_m)$ is called *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:

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

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 reflects the importance of the cost in the sense that deploying one more seed should bring at least w number of 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 when considering $w = 1$; then, the system utility is the expected number of satisfied vehicles through the short-range radio alone.

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 k_i seeds are deployed at time 0. We note that the number of satisfied nodes depends on the inter-encounter time of nodes, among other things. The *inter-encounter time* of a given pair of nodes is defined as the time duration from the time that the given pair of nodes encounter to the next consecutive time that the pair encounter again. And, we say the two vehicles *encounter* each other when they can communicate with each other directly through the short-range radio.

Note that the inter-encounter time includes the time duration of the former encounter. We include the duration because we assume the high bandwidth for the short-range radio so that the message exchange between a pair of nodes is completed for a very short period time once they encounter.

In order to derive s_i, we assume the following. (1) A node may encounter α proportion of all nodes on average for the time interval in interest. (2) For a pair of nodes that may encounter each other in the time interval, the inter-encounter time follows the Exponential distribution with rate β. (3) The inter-encounter times of pairs of nodes are jointly independent and identical. (4) The seed nodes are selected uniformly at random. (5) A node sends the previously obtained message *only* to the nodes that are interested in the same type of the message.

The assumptions (2) and (3) have been found reasonable when the movement of nodes follows the random waypoint model, random direction model, *etc.* ([1, 2, 3]). Moreover, we have found that the real movement traces of Beijing taxis in Section 6.3 agree with the assumption (2). Assumption (5) implies that a node that has no interest in a particular type of content never acquires nor relays the message of the type, which is often referred to as the *interest-only* caching policy. We focus on this caching policy in this paper because it can avoid the non-ignorable storage costs for keeping uninterested data incurred otherwise. The interest-only policy results in no interaction among the different interest groups. Therefore, we can focus on a particular group of nodes alone to analyze the spreading of the particular type of content.

We note that the expected number of satisfied nodes behaves like the number of infected nodes in epidemic routing ([4]). The differences are that the initial number of sources (that is, the number of seeds in this paper) is more than one, and that the other nodes that a node may ever encounter are not all nodes but a fraction of them (Assumption (1)). The previous work has introduced largely two methods to analyze the number of infected nodes; one is using the Markov chains and the other is using the ordinary differential equations (ODE) ([5, 6, 7, 8]). We use the ODE method with some modification for our analysis.

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 i.i.d. Exponential (Assumptions (2) and (3)) and each node meets a fraction of other nodes (Assumption (1)).

Therefore, the expected number of newly satisfied nodes $\triangle S$ is as follows:

$$\begin{aligned} \triangle S &= s_i(k_i, t + dt) - s_i(k_i, t) \\ &= \alpha\beta s_i(k_i, t) \left(n_i - s_i(k_i, t)\right) \cdot dt \end{aligned} \quad (1)$$

Note that the expected number of ever-meeting pairs of inter-group nodes is approximately[1] $\alpha s_i(k_i, t) \left(n_i - s_i(k_i, t)\right)$.

From Equation (1) and the fact that the number of seeds is k_i,

[1]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.

we have the following ODE system;

$$\frac{\partial s_i(k_i,t)}{\partial t} = \alpha\beta s_i(k_i,t)(n - s_i(k_i,t)) \quad (2)$$
$$s_i(k_i,0) = k_i \quad (3)$$

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

$$s_i(k_i,t) = \frac{n_i}{1+(n_i/k_i - 1)\exp(-n_i\alpha\beta t)} \quad (4)$$

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:

$$PF2: \underset{\mathbf{k}}{\text{Maximize}} \quad f(\mathbf{k}) = \textstyle\sum_{i=1}^m s_i(k_i,d) - w\cdot c(\mathbf{k}) \quad (5)$$
$$s.t \quad c(\mathbf{k}) = \textstyle\sum_{i=1}^m k_i \le C \quad (6)$$
$$0 \le k_i \le n_i = p_i n,\ \forall i \in M \quad (7)$$

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 nonnegative and its second derivative is non-positive as follows:

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

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

where we use

$$z_i = e^{-n_i\alpha\beta d} \quad (8)$$

for concise presentation.

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 Equation (7) 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 the problem $PF2$.

The Lagrangian of the problem is as follows:

$$L(\mathbf{k},\lambda) = f(\mathbf{k}) - \lambda\,(c(\mathbf{k}) - C) \quad (9)$$

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

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 $(\widehat{\mathbf{k}},\widehat{\lambda})$;

$$\frac{\partial L(\mathbf{k},\lambda)}{\partial k_i} = \frac{n_i^2 z_i}{(k_i + n_i z_i - k_i z_i)^2} - \lambda - w = 0 \quad \forall i \quad (10)$$

$$\frac{\partial L(\mathbf{k},\lambda)}{\partial \lambda} = \lambda(\sum_{i=1}^m k_i - C) = 0 \quad (11)$$

As can be seen from Equation (11), 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 (6) 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;

$$\widetilde{C} \doteq c(\widetilde{\mathbf{k}}) = \textstyle\sum_{i=1}^m \widetilde{k}_i \quad (12)$$

Then, $\widetilde{C} = c(\widetilde{\mathbf{k}}) < C$, and so, the optimum solution $\widetilde{\mathbf{k}}$ automatically satisfies the constraint (6) in this case.

On the other hand, the constraint (6) is active in the case where $\sum k_i = C$. It means that the unconstrained solution of the optimization problem requires more cost than allowed in general, that is, $C \le \widetilde{C}$. 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:

$$\widehat{k}_i = \begin{cases} \widetilde{k}_i = \dfrac{n_i\sqrt{z_i}}{1-z_i}\left(\dfrac{1}{\sqrt{w}} - \sqrt{z_i}\right), & \text{if } \widetilde{C} < C \quad (13a) \\ \breve{k}_i = \dfrac{n_i\sqrt{z_i}}{1-z_i}\left(\dfrac{C+A}{B} - \sqrt{z_i}\right), & \text{if } \widetilde{C} \ge C \quad (13b) \end{cases}$$

where

$$A = \sum_{i=1}^m \frac{n_i z_i}{1-z_i}, \qquad B = \sum_{i=1}^m \frac{n_i\sqrt{z_i}}{1-z_i} \quad (14)$$

And, $\widetilde{C}$ can be obtained from Equations (12) and (13a). 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 Equation (13) still ignores the constraint (7). However, we show that the solution is indeed the solution of $PF2$ under the conditions in Theorems 1 and 2.

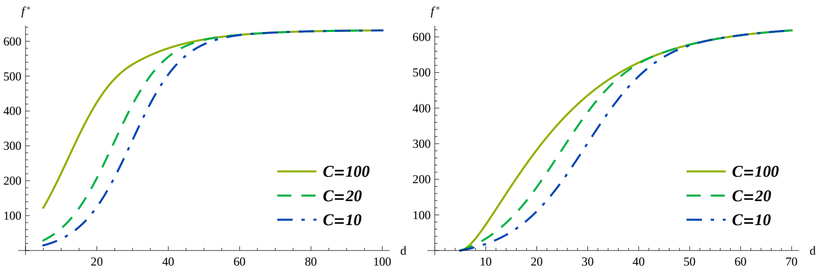
Theorem 1. *Suppose $\widetilde{k}_i$ and $\widetilde{C}$ are defined as in Equations (13a) and (12), respectively. Also, suppose $z_i = \exp(-n_i\alpha\beta d)$. Then, under any one of the following conditions,*

$$\mathbb{C}_1: \quad \{0 < w < 1, 0 < z_i \le w\}$$
$$\mathbb{C}_2: \quad \{w = 1, 0 < z_i < 1\}$$
$$\mathbb{C}_3: \quad \{w > 1, 0 < z_i \le 1/w\}$$

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

$$k_i^* = \widetilde{k}_i \quad (15)$$

Proof. We note that $\widetilde{k}_i$ is solutions of $PF2$ when $\widetilde{C} < C$ if we ignore the constraint (7). 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.



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

Figure 1: Optimum utility vs. delay budget

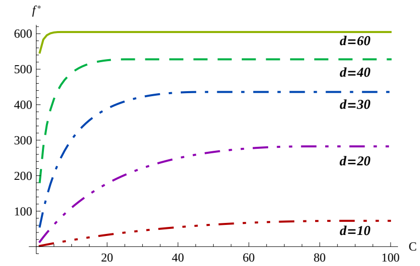


Figure 2: Optimum utility vs. cost budget $(w = 2)$

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

$$\widetilde{k}_i = \frac{n_i\sqrt{z_i}}{1-z_i}\left(\frac{1}{\sqrt{w}} - \sqrt{z_i}\right) = n_i \cdot y(z_i) \tag{16}$$

where

$$y(z_i) \doteq \frac{\sqrt{z_i/w} - z_i}{1-z_i} \tag{17}$$

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

When $0 < w \le 1$, 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:

$$\frac{dy(z)}{dz} = \frac{1 - w + (\sqrt{z} - \sqrt{w})^2}{2\sqrt{wz}(1-z)^2} \tag{18}$$

Hence, we can easily see $0 = y(0) < y(z) \le y(w) = 1$ under the condition $\mathbb{C}_1$.

Under the condition $\mathbb{C}_2$, we can see $0 \le y(z_i) \le 1$ from the following:

$$0 = y(0) < y(z) < \lim_{z\to 1} y(z) = 1/2 < 1 \tag{19}$$

Note that we cannot use $y(z = 1)$ 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 Equation (17). And it is not difficult to see that $y(z) > 0$ for $0 < z < 1/w$. And, these imply that $0 \le \widetilde{k}_i \le n_i$ under $\mathbb{C}_3$. □

Theorem 2. *Suppose $\breve{k}_i$ and $\widetilde{C}$ are defined as in Equations (13b) and (12), 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

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

, 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 $C \le \widetilde{C}$ if we ignore the constraint (7). Hence, we only need to show $\breve{k}_i \in [0, n_i]$ under the conditions.

First, we will show that $\breve{k}_i \le n_i$.

$$\widetilde{C} \ge C \tag{20}$$

$$\Rightarrow \sum_{i=1}^{m} \frac{n_i\sqrt{z_i}}{1-z_i}\left(\frac{1}{\sqrt{w}} - \sqrt{z_i}\right) = \frac{B}{\sqrt{w}} - A \ge C \tag{21}$$

$$\Rightarrow \frac{1}{\sqrt{w}} \ge \frac{C+A}{B} \tag{22}$$

where A and B are defined in Equation (14).

This implies, together with Equation (13) and the proof of Theorem 1, that $\breve{k}_i \le \widetilde{k}_i \le n_i$.

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

$$\breve{k}_i \ge \frac{n_i\sqrt{z_i}}{1-z_i}\left(\frac{1}{B}\left(\sum_j \frac{n_j\sqrt{z_j}}{1-z_j} + A\right) - \sqrt{z_i}\right) \tag{23}$$

$$= \frac{n_i\sqrt{z_i}}{B(1-z_i)} \sum_j^m \frac{n_j\sqrt{z_j}}{1-z_j}\left(1 + \sqrt{z_j} - \sqrt{z_i}\right) \ge 0 \tag{24}$$

where the last inequality follows since $\sqrt{z_j} - \sqrt{z_i} \ge -1$ 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 Equations (4) and (13), given by

$$s_i^*(d, C) = \begin{cases} \displaystyle\sum_{i=1}^{m} \frac{n_i}{1-z_i}\left(1 - \sqrt{wz_i}\right), & \text{if } \widetilde{C} < C \quad (25a) \\ \displaystyle\sum_{i=1}^{m} \frac{n_i}{1-z_i}\left(1 - \frac{B}{C+A}\sqrt{z_i}\right), & \text{otherwise} \quad (25b) \end{cases}$$

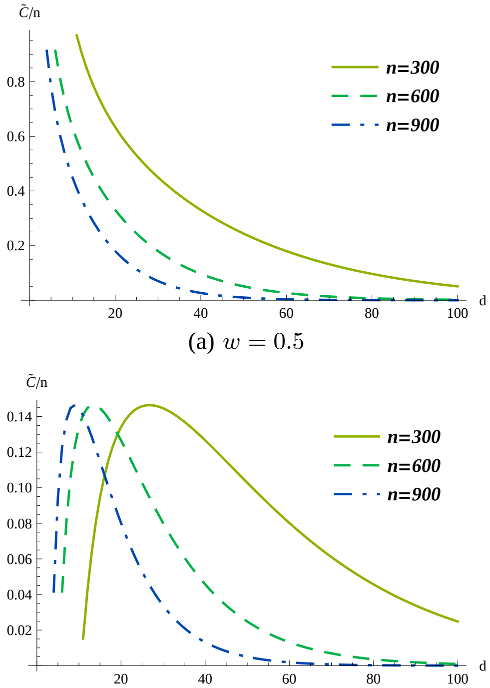
where z_i, A, and B are given in Equations (8) and (14) respectively.

The optimum system utility is from Equations (5), (13), and (25), as follows:

$$f^*(d, C) = \begin{cases} \displaystyle\sum_{i=1}^{m} (1 - \sqrt{wz_i}) s_i^*(d, C), & \text{if } \widetilde{C} < C \quad (26a) \\ \displaystyle\sum_{i=1}^{m} \left(1 - \frac{C+A}{B} w\sqrt{z_i}\right) s_i^*(d, C), & \text{otherwise} \quad (26b) \end{cases}$$

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 6; the number of nodes $n = 632$, the inter-encounter rate



(a) $w = 0.5$

(b) $w = 2$

Figure 3: Unconstrained optimum total cost vs. delay budget

$\beta = 3.663 \times 10^{-6}$ per second, and $\alpha = 0.191$. And we focus on a single type of content in this section. 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 are small, medium, and large. When d is small or large, we can see that the system utility has ignorable sensitivity on the value of C. But, when d is in between, the difference can be quite huge. 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 values. 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 optimum total cost $\widetilde{C}$ changes as the allowed delay d changes. While the cost monotonically decreases as d increase when $w = 0.5$, the cost reaches its maximum and decreases when $w = 2$. In fact, the optimum cost monotonically decreases when $w \leq 1$. The system behavior changes significantly at $w = 1$ because one more seed does not require more than one more satisfied node when $w \leq 1$, and so the system utility never decreases as the seed number increases. However, when $w > 1$, deploying one more seed requires more satisfied nodes besides itself, which may make the utility decreases especially when the delay budget is very small or very large. When the cost is very important (high w) and the allowed delay is very small, our model suggests it is sometimes better not to disseminate the content at all depending on other system parameters like the inter-encounter time.

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 Equation (26), and d only appears with α and β. Therefore, α and β act like shrinking or stretching the performance plot in the direction of d as they increase or decrease, respectively.

5. PRACTICAL SOLUTIONS

In the previous section 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 $[0, n_i]$. Its correctness is proven in Theorem 3. 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.

Algorithm 1 OPTIMIZER(C, m)

$\mathbf{k} := m$-sized array initialized to be all zero.
for $(c = 0; c < C; c+ = 1)$ **do**
  $i^* := 0$
  $\delta_{max} := -\infty$
  for $(i = 1; i \leq m; ++i)$ **do**
    $\delta := f(\{\mathbf{k}[1], \ldots, \mathbf{k}[i] + 1, \ldots, \mathbf{k}[m]\}) - f(\mathbf{k})$
    if $(\delta > \delta_{max})$ **then**
      $i^* := i$;
      $\delta_{max} := \delta$
  if $(\delta_{max} \leq 0)$ **then**
    break
  $\mathbf{k}[i^*] + = 1$
return $\mathbf{k}$

Theorem 3. *Algorithm 1 returns the optimum solution of $PF1$.*

Proof. We first note that the system utility function can be represented w.r.t $\mathbf{k}$ as follows:

$$f(\mathbf{k}) = \sum_{i=1}^{m} s_i(k_i, d) - w \sum_{i=1}^{m} k_i \quad (27)$$

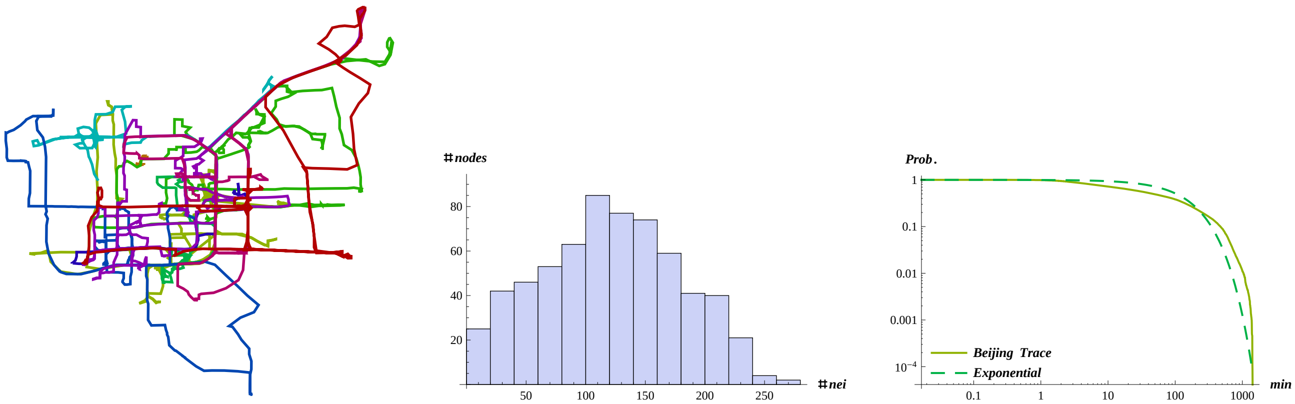
$$= \sum_i \underbrace{(s_i(k_i, d) - wk_i)}_{\doteq f_i(k_i)} = \sum_i f_i(k_i) \quad (28)$$

From Section 4.1, we know that f_i is concave w.r.t the number of seeds, which implies

$$\triangle f_j(k+1) \leq \triangle f_j(k), \quad \forall k \geq 0, k \in \mathbb{N}; \forall j \in M \quad (29)$$

where $M = \{1, 2, \ldots, m\}$, and

$$\triangle f_j(k) \doteq f_j(k) - f_j(k-1) \quad (30)$$



(a) (b) (c)

Figure 4: Properties of Beijing Taxi Traces: (a) geographical movements of 10 sample taxis, (b) Histogram of # of neighbors of a node, (c) tail distribution of the inter-encounter times

Now suppose $\mathbf{k}^*$ is the outcome of Algorithm 1, and $\widetilde{\mathbf{k}}$ is an arbitrary legitimate vector of number of seeds for *PF1* (*i.e.* $\widetilde{C} = \sum_i \widetilde{k}_i \leq C$). It is easy to see that Algorithm 1 ensures that $C^* = \sum_i k_i^* \leq \lfloor C \rfloor \leq C$, and so $\mathbf{k}^*$ is a legitimate seed vector. We shall show that $\mathbf{k}^*$ gives the system utility at least as high as that of $\widetilde{\mathbf{k}}$ so that $f(\mathbf{k}^*) \geq f(\widetilde{\mathbf{k}})$.

i) If $\mathbf{k}^* = \widetilde{\mathbf{k}}$, we have nothing to prove.

ii) When $\widetilde{\mathbf{k}} \preceq \mathbf{k}^*$, $\widetilde{k}_i \leq k_i^*$ for all $i \in M$. Then,

$$f(\mathbf{k}^*) - f(\widetilde{\mathbf{k}}) = \sum_{j \in J} \sum_{i=1}^{k_j^* - \widetilde{k}_j} \triangle f_j(\widetilde{k}_j + i) \geq 0 \tag{31}$$

where $J = \{j \in M | \widetilde{k}_j < k_j^*\}$.

It is non-negative because $\triangle f_j(k_j^*) \geq 0, \forall j$ due to the line 11 of Algorithm 1, and $\widetilde{k}_j + i \leq k_j^*$, which implies from Equation (29)

$$\triangle f_j(\widetilde{k}_j + i) \geq \triangle f_j(k_j^*) \geq 0 \tag{32}$$

Hence, $f(\mathbf{k}^*) \geq f(\widetilde{\mathbf{k}})$.

iii) When $\widetilde{\mathbf{k}} \succeq \mathbf{k}^*$ (*i.e.* $\widetilde{k}_i \geq k_i^*, \forall i$) but $\widetilde{\mathbf{k}} \neq \mathbf{k}^*$, it is easy to see the following:

$$C^* < \widetilde{C} \leq \lfloor C \rfloor \leq C \tag{33}$$

This implies the algorithm has executed the line 11, which again implies with Equation (29),

$$\triangle f_l(k_l^* + c) \leq 0, \quad \forall l \in M, \forall c \geq 1 \tag{34}$$

Hence, letting $J = \{j \in M | \widetilde{k}_j > k_j^*\}$,

$$f(\widetilde{\mathbf{k}}) - f(\mathbf{k}^*) = \sum_{j \in J} \sum_{i=1}^{\widetilde{k}_j - k_j^*} \triangle f_j(k_j^* + i) \leq 0 \tag{35}$$

iv) Consider the remaining cases. For all these cases, we have at least a pair of $(i,j) \in M^2$ such that $\widetilde{k}_i < k_i^*$ and $k_j^* < \widetilde{k}_j$.

Before proceeding, we show a couple of useful inequalities for this proof.

$$\triangle f_x(k_x^* + 1) \leq \triangle f_y(k_y^*), \quad \forall y \neq x \tag{36}$$

If this is not true, Algorithm 1 would have incremented k_x to be $k_x^* + 1$ instead of incrementing k_y when $k_x = k_x^*$ and $k_y = k_y^* - 1$.

From Equations (29) and (36),

$$\triangle f_x(k_x^* + 1) \leq \triangle f_y(k), \quad \forall k \leq k_y^*; \, \forall y \in M \tag{37}$$

Let us define δ as follows:

$$\delta = \min\{|\widetilde{k}_i - k_i^*|, |\widetilde{k}_j - k_j^*|\} \tag{38}$$

And Let $\mathbf{k}^{(1)}$ such that $k_i^{(1)} = \widetilde{k}_i + \delta$, $k_j^{(1)} = \widetilde{k}_j - \delta$, and $k_l^{(1)} = \widetilde{k}_l, \forall l \neq i, j$, which implies $\mathbf{k}^{(1)}$ is also legitimate from $C^{(1)} = \sum_i k_i^{(1)} = \widetilde{C} \leq C$. Now, consider

$$\begin{aligned} f(\mathbf{k}^{(1)}) - f(\widetilde{\mathbf{k}}) &= \sum_{l=1}^{\delta} \triangle f_i(\widetilde{k}_i + l) - \sum_{l=1}^{\delta} \triangle f_j(\widetilde{k}_j - l + 1) \\ &= \sum_{l=1}^{\delta} \Big(\triangle f_i(\widetilde{k}_i + l) - \triangle f_j(\widetilde{k}_j - l + 1) \Big) \end{aligned} \tag{39}$$

Because $\widetilde{k}_i + l \leq k_i^*$ and $k_j^* \leq \widetilde{k}_j - l + 1$ for $\forall l \in [1, \delta]$, we have the following inequality from Equation (37):

$$\triangle f_j(\widetilde{k}_j - l + 1) \leq \triangle f_i(\widetilde{k}_i + l) \tag{40}$$

This implies that the RHS of Equation (39) is non-negative. Hence, $f(\mathbf{k}^{(1)}) \geq f(\widetilde{\mathbf{k}})$, and $k_i^{(1)} = k_i^*$ or $k_j^{(1)} = k_j^*$, which means at least one more element in $\mathbf{k}^{(1)}$ is same as that of $\mathbf{k}^*$ than $\widetilde{\mathbf{k}}$, augmenting the system utility.

Now, we keep doing this augmentation process from the resultant seed vector of each process until there is no such pair (i, j). We need no more than m rounds of this process to reach this state. Then, letting $\mathbf{k}^{(f)}$ denote the final resultant seed vector, we have one of the following exhaustive cases; (a) $\mathbf{k}^{(f)} = \mathbf{k}^*$, (b) $\mathbf{k}^{(f)} \preceq \mathbf{k}^*$, and (c) $\mathbf{k}^{(f)} \succeq \mathbf{k}^*$. In each of the cases, from i), ii) and iii),

$$f(\mathbf{k}^*) \geq f(\mathbf{k}^{(f)}) \geq f(\widetilde{\mathbf{k}}) \tag{41}$$

$\square$

6. SIMULATION BASED ON TAXI TRACES

In this section we present how the contents dissemination behaves in the 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.

6.1 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 4(a) shows the GPS traces of randomly chosen 10 taxis as an example.

6.2 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 ([9, 10]) suggests.

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} (Equation (42)) that the pair can encounter each other next, if the 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 logs of GPS locations are not synchronized, we cannot simply take the locations of the pair from the logs at a given 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 given by

$$\tau_{min} = \frac{1}{2s_m}(\textsc{GeoDist}(pos(P_1, t), pos(P_2, t)) - r) \qquad (42)$$

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 vehicle 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.

After all, 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 distance 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 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. We note that the 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.

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 $v_1, v_2 \in V$ has 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. Figure 4(b) shows the histogram of the number of neighbors of a node.

6.3 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 4(c) 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. Because we assume IID Exponential distribution for per-pair inter-encounter time, it should have the identical distribution to that of the aggregate inter-encounter time.

However, the above sample mean for the inter-encounter time is actually an underestimate of the true mean because we 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 4(c)) 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.

$$\beta^* = \arg\max_\beta \Pr(N_1, N_2, ..., N_\eta | \beta, T_1, ..., T_\eta) \qquad (43)$$

where

$$\begin{aligned} &\Pr(N_1, N_2, ..., N_\eta | \beta, T_1, ..., T_\eta) \\ &= \prod_{i=1}^{\eta} \Pr(N_i | \beta T_i) = \prod_{i=1}^{\eta} \frac{(\beta T_i)^{N_i} e^{-\beta T_i}}{N_i!} \qquad (44) \\ &= \left(\prod_{i=1}^{\eta} \frac{T_i^{N_i}}{N_i!} \right) e^{-\beta \sum_{i=1}^{\eta} T_i} \beta^{\sum_{i=1}^{\eta} N_i} \end{aligned}$$

Note that Equation (44) 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:

$$\beta^* = \sum_{i=1}^{\eta} N_i \,/\, \sum_{i=1}^{\eta} T_i \qquad (45)$$

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

6.4 Simulation Methodology

From the time-ordered traces of the encounters of the Beijing traces, produced by the method in Section 6.2, 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; E 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 are 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(E, N, S, t_s, d)

```
Mark every v ∈ S as satisfied.
for all e ∈ E in order s.t. t_s ≤ time(e) ≤ 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.
return the set of all marked nodes
```

6.5 Number of Satisfied Nodes

Figure 5 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

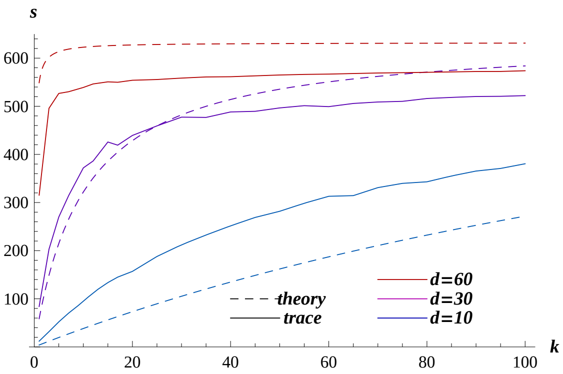


Figure 5: Avg. # of satisfied nodes vs. # of seeds

behavior of the average number of satisfied nodes as the number of seeds increases.

Figure 6 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. And, the results from the real trace do not show the kind of threshold behavior of the theory results. This indicates the real-world encounter process between two groups of vehicles is not perfectly represented in our model. This may be because the pair-wise encounter processes are not IID or do not follow the Poisson distribution. Suppose there is some dependency among the processes even when they are identical. 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 6. But, more accurate analysis calls for further investigation, which is out of scope of this paper and the subject of our future research.

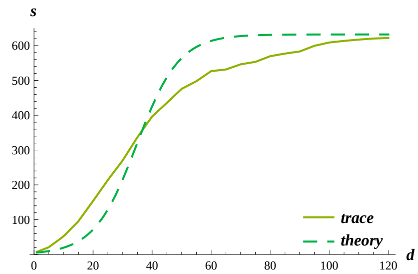
Nevertheless, the system behavior with respect to the number of seeds is more important for our problem because it is the parameter to optimize on. And, Figure 5 suggests comparable numbers of seeds for the knees of plots from the theory and the real traces.

6.6 Optimal Number of Seeds

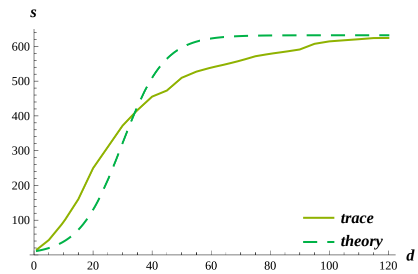
Now we look into the system utility f with respect to the number of seeds. We have compared the system utilities[2] that our model predicts and the Beijing traces suggest, with various delay constraints and cost weights. It turns out they show similar behaviors as in Figure 5; the real traces suggest 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 5.

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 7 shows the optimal number of seeds and the corresponding empirical system utility with respect

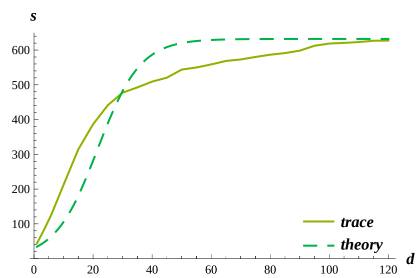
[2]We omit the corresponding figure because it looks similar to Figure 5 and due to the lack of pages.



(a) k=5



(b) k=10



(c) k=30

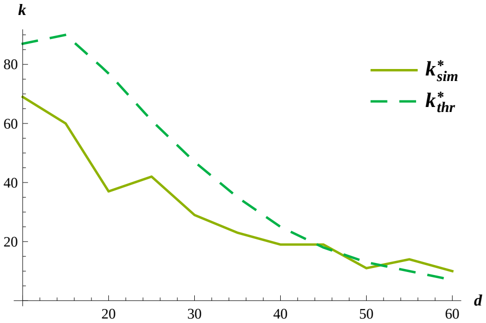
Figure 6: Average number of satisfied nodes vs. tolerable delay

to the delay budget. Figure 7(a) 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 5). This is why our analytical solution provides near-optimal performance even in the small delay regime as can be seen in Figure 7(b).

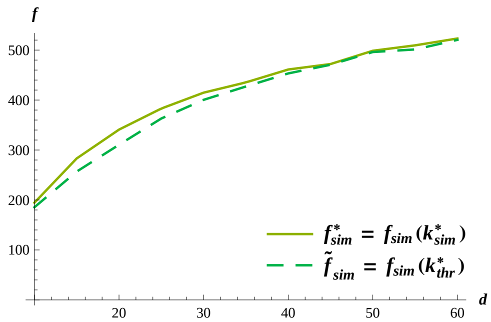
Figure 7(b) compares the best possible system utility values f^*_{sim} of the trace-based simulations and the empirical utility values $\widetilde{f}_{sim}$ when our solution k^*_{thr} is used. In other words, the figure shows how close the system utility of the real system would be to the system's best possible 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. RELATED WORK

In the past decade, extensive research has been done to study the technical feasibility of heterogeneous intergrated wireless networks. Some of this has focused on integrating wireless local area networks and cellular networks to allow for vertical handoffs [11]. There has also been work on integrating mobile ad hoc networks



(a) Optimal Seed Number



(b) System utility

Figure 7: System behaviors in optimal regime with respect to the delay budget

(MANET) and cellular systems to improve throughput and increase coverage [12] [13], and there has been theoretical analysis of the capacity of such heterogeneous networks [14] [15].

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 [16], such as sparse vehicular networks when such networks are deployed at the first few years [17]. The initial effort for tackling Delay Tolerant networks was placed on designing reliable and efficient routing protocols under a variety of assumptions on mobility [18] [19] [20] [21]. 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 [22, 23, 24, 25]. 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 [20] [21]. 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 [26]).

In our study, we use differential equations to model content replication and dissemination. This is similar to [26], 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 [27], the authors explore the latency performance of different frequency-based replication policies in the context of vehicular networks with limited storage. CarTorrent [28] and AdTorrent [29], present content dissemination mechanisms to distribute files and advertisements, respectively, in vehicular networks. In [30], 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.

8. CONCLUSION

We have investigated the optimum content dissemination in the heterogenous vehicular network in this work. In this network, each vehicle is equipped with two radios; one is the costly long-range low-bandwidth radio for direct communication with the infrastructure, and the other is the low-cost short-range high-bandwidth radio for communication with peer vehicles. We have considered the problem of how to spread relevant content to more vehicles with smaller cost. We have developed the relevant optimization formulation, derived their analytical solutions with some relaxation, and examined the behaviors of the system under the optimum regime. 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, on the order of an hour, can be tolerated. We have also developed a polynomial algorithm to calculate the exact optimum seed vector with no relaxation.

In order to verify our analysis and justify our assumptions and approximations, we have performed simulations based on the real GPS traces of 632 taxis gathered in Beijing, China. We have found that the real traces show the aggregate inter-encounter time of vehicles is close to the Exponential distribution agreeing with our assumption, and that their performance of contents dissemination exhibits similarities to what our model predicts.

In this work, we have assumed a low density of subject vehicles to avoid further level of complexity to the problem, for example, radio interference and packet collisions. Although the low density assumption is not so unrealistic, especially for the early phase of vehicular networks, we plan to investigate this issue further in our future work. We have also assumed i.i.d. pair-wise inter-encounter times, and we have pointed out in Section 6.5 that the mismatched increase rate of the number of satisfied nodes may be attributed to this assumption. Relaxing this assumption is also a topic of our future research.

9. REFERENCES

[1] D. Johnson and D. Maltz, "Dynamic source routing in ad hoc wireless networks," *Mobile Computing*, vol. 353, 1996.

[2] A. Einstein, *Investigations on the Theory of the Brownian Movement*. Dover Publications, 1906.

[3] G. Sharma and R. Mazumdar, "Scaling laws for capacity and delay in wireless ad hoc networks with random mobility," in *IEEE International Conference on Communications*, vol. 7, pp. 3869–3873, June 2004.

[4] A. Vahdat and D. Becker, "Epidemic routing for partially-connected ad hoc networks," Tech. Rep. CS-2000-06, University of California San Diego, July 2000.

[5] T. Small and Z. Haas, "The shared wireless infostation model – a new ad hoc networking paradigm," in *Mobihoc*, 2003.

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

[7] Z. Haas and T. Small, "A new networking model for biological applications of ad hoc sensor networks," *IEEE/ACM Transactions on Networking*, 2006.

[8] D. Daley and J. Gani, *Epidemic Modelling*. Cambridge University Press, 1999.

[9] F. Bai and H. Krishnan, "Reliability analysis of dsrc wireless communication for vehicle safety applications," in *IEEE Intelligent Transportation Systems Conference*, pp. 355–362, September 2006.

[10] K. Hong, D. Xing, V. Rai, and J. Kenney, "Characterization of dsrc performance as a function of transmission power," in *ACM VANET*, 2009.

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

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

[13] 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, pp. 393–408, 2004.

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

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

[16] K. Fall, "A delay-tolerant network architecture for challenged internets," in *SIGCOMM '03: Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, pp. 27–34, ACM, 2003.

[17] F. Bai and B. Krishnamachari, "Spatio-temporal variations of vehicle traffic in vanets: facts and implications," in *VANET '09: Proceedings of the sixth ACM international workshop on VehicularAr InterNETworking*, pp. 43–52, ACM, 2009.

[18] Q. Yuan, I. Cardei, and J. Wu, "Predict and relay: an efficient routing in disruption-tolerant networks," in *MobiHoc '09: Proceedings of the tenth ACM international symposium on Mobile ad hoc networking and computing*, pp. 95–104, ACM, 2009.

[19] M. M. Bin Tariq, M. Ammar, and E. Zegura, "Message ferry route design for sparse ad hoc networks with mobile nodes," in *MobiHoc '06: Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing*, pp. 37–48, ACM, 2006.

[20] 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, pp. 77–90, 2008.

[21] 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, pp. 63–76, 2008.

[22] 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 *MobiCom '07: Proceedings of the 13th annual ACM international conference on Mobile computing and networking*, pp. 195–206, ACM, 2007.

[23] A. Balasubramanian, B. N. Levine, and A. Venkataramani, "Enhancing interactive web applications in hybrid networks," in *MobiCom '08: Proceedings of the 14th ACM international conference on Mobile computing and networking*, pp. 70–80, ACM, 2008.

[24] B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A. Miu, E. Shih, H. Balakrishnan, and S. Madden, "Cartel: a distributed mobile sensor computing system," in *SenSys '06: Proceedings of the 4th international conference on Embedded networked sensor systems*, pp. 125–138, ACM, 2006.

[25] J. Eriksson, H. Balakrishnan, and S. Madden, "Cabernet: vehicular content delivery using wifi," in *MobiCom '08: Proceedings of the 14th ACM international conference on Mobile computing and networking*, pp. 199–210, ACM, 2008.

[26] A. Chaintreau, J.-Y. Le Boudec, and N. Ristanovic, "The age of

gossip: spatial mean field regime," in *SIGMETRICS '09: Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems*, pp. 109–120, ACM, 2009.

[27] S. Ghandeharizadeh, S. Kapadia, and B. Krishnamachari, "Comparison of replication strategies for content availability in C2P2 networks," in *6th International Conference on Mobile Data Management (MDM'05)*, May 2005.

[28] 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 Mobile Networking for Vehicular Environments (MOVE)*, May 2007.

[29] A. Nandan, S. Das, B. Zhou, G. Pau, and M. Gerla, "Adtorrent: Digital billboards for vehicular networks," in *IEEE/ACM International Workshop on Vehicle-to-Vehicle Communications (V2VCOM)*, 2005.

[30] J. Reich and A. Chaintreau, "The age of impatience: optimal replication schemes for opportunistic networks," in *CoNEXT '09: Proceedings of the 5th international conference on Emerging networking experiments and technologies*, ACM, 2009.