Geography-Aware Active Data Dissemination in Mobile Social Networks

Jialu Fan*†   Yuan Du*   Wei Gao†   Jiming Chen*   Youxian Sun*

*State Key Lab. of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China
†Pennsylvania State University, University Park, PA 16802, USA
fanjialu@gmail.com, duyuan@zju.edu.cn, wxg139@cse.psu.edu, jmchen@ieee.org, yxsun@iipc.zju.edu.cn

Abstract—In mobile social networks (MSNets), data dissemination is an important topic, which has not been widely investigated yet. Active data dissemination is a networking paradigm where a superuser intentionally facilitates the connectivity in the network. One of the key challenges under this paradigm is how to design the most efficient superuser route to achieve certain properties of end-to-end connectivity. Most existing solutions only focus on the network with stationary users or strongly constrained node mobility, and assume the superuser always moves with a fixed route. In this paper, we propose a flexible approach to design the superuser routes, considering the realistic user movements in MSNets. To the best of our knowledge, this work is the first to study active data dissemination from the social network perspective. We explore the geographic regularity of human mobility in the network, employ a semi-Markov analytical model to describe such mobility pattern, and hence formulate the superuser route design as a combinational optimization problem of Convex Optimization and Traveling Salesman Problem by exploiting social network concepts including communities and centrality. Extensive trace-driven simulations show that our approach consistently outperforms other existing superuser route design algorithms in terms of delivery ratio and energy efficiency.

I. INTRODUCTION

Mobile Social Networks (MSNets) are networks in which wireless mobile users of similar interests or commonalities cooperate to establish network connectivity and communicate with each other in the absence of network infrastructure [1]. Since these networks do not require existing network infrastructure, they can be deployed in a number of critical areas, including large-scale disaster recovery, battlefields and wide-area sensor networks [2][3]. In these environments, the network connectivity in MSNets are usually intermittent due to the unpredictable node mobility, limited radio range, physical obstructions, or malicious attacks. Such intermittent network connectivity and uncertainty make data dissemination in MSNets a challenging problem.

Since users are only connected intermittently in MSNets, users mobility should be exploited to bridge network partitions and to disseminate data. Currently, existing data dissemination approaches for such intermittently connected networks are commonly "store-carry-and-forward" schemes, which exploit the physical user movements to carry data around the network and overcome path disconnection. In some approaches, the data dissemination is *passive*, i.e., the mobility of users carrying data cannot be controlled in favor of data dissemination [4][5][6][7]. The passive scheme relies on the inherent

movement of users, such that the existing users in the network relay the data from the source to the destination in one or more hops. In other approaches, the dissemination is *active*, which means that one or more special mobile users (either existing or extra users) are intentionally deployed in the network. The actions of special users are controllable to facilitate connectivity among other users [8][9][10][11][12][13]. Such special users take the burden of data dissemination away from the regular users, and subsequently save the limited energy and storage resources of the regular users. Our research focuses on the active data dissemination approaches with single special user. In the rest of this paper, the special mobile user is called *superuser*, and the other regular users are called *users* for short.

Active data dissemination depends solely on mobility trajectory of superusers with the knowledge of regular users' movements. Therefore, the design of superuser route has significant impact on network performance. Although there is some initial work on special user route design, such as Message Ferry [8][9], or Data MULEs [13], they only consider the network with stationary users [9][13], or assume that the special users move with fixed routes [8]. Tariq et al. [10] proposed a customized ferry route for mobile networks, but the node mobility is strongly constrained. Besides, none of them considers active data dissemination from the social network perspective.

The basic purpose of this paper is to design flexible superuser routes in MSNets, without any constraints on the movements of regular users. Hence, the main challenge is how to characterize and represent user mobility in MSNets. In this paper, we exploit two key concepts from social network analysis to investigate the natural regularity of user mobility. One is *community*, which is a cluster of users that are "closely" linked to each other according to users' social relationships. In fact, as shown in [14], such social relationships often strongly relates to the geographic locations of mobile users. Therefore, we propose *geo-community* in this paper. Another is *centrality*, which indicates that the contact probability of some users or clusters in the network to other users are higher. Specifically, we focus on exploring the centrality metric of geo-communities, which is named *geo-centrality*, by employing a semi-Markov process to model user mobility, where the communities in the network are represented as Markovian states.

In more details, our objective is to design superuser routes to accelerate the superuser's data dissemination, i.e., minimize the total route duration with a given dissemination probability or maximize the dissemination probability within a given time constraint in MSNets. We determine superuser route that comprises an ordered set of geo-communities and waiting times at these geo-communities. From a probabilistic perspective, the set of selected geo-communities is expected to cover as many users as possible. The cumulative contact probability for a geo-community to users therefore needs to be calculated, and such calculation may require the aforementioned user mobility model. Our detailed contributions are as follows:

- We exploit the concepts of *geo-community* and *geo-centrality* of MSNets, which provides deep insights into the research on both *active* and *passive* data dissemination.
- In terms of the superuser route design, we formulate a Static Optimal Disseminating Algorithm (*SODA*) from the statistic perspective, and further propose a Greedy Adaptive Disseminating Algorithm (*GADA*) excluding the overlap of contact user sets among the geo-communities.

The rest of this paper is organized as follows. Section II provides an overview about problem definition in MSNets and the optimization objectives. Then we describe the MSNets embodying community and centrality in Section III. Section IV presents the semi-Markov model on users' mobility in the network. Based on such modeling, two active data dissemination schemes are studied in Sections V. Section VI evaluates the performance of our approach via trace-driven simulation, Section VII reviews existing work, and Section VIII concludes the paper.

II. OVERVIEW

A. Two Observations in MSNets

There are already two pervasive observations from real social networks [15]. First, mobile users in a social network usually move around several well-visited locations instead of moving randomly. The real situation behind such observation is that users usually belong to several communities, and contact others with similar hobbies, occupations or social functions. For example, the graduate students working in the same office interact more frequently with each other. In fact, such contact preference is also usually correlated to geography information, such that the contacts among officemates mostly happen in the office. We define such geography-related community as geo-community, which will be illustrated later in Section III-A. Second, the user's dwell time at each geo-community is fairly regular, because their social behavior patterns usually remain stable in a relatively long interval.

On the other hand, geo-community affiliations among mobile users can be highly diverse, and hence some geo-communities may have higher dynamic density of users than others. As a result, spatial user distributions in the network are very heterogeneous and possess popular geo-communities of high user density, and where the superuser have therefore a
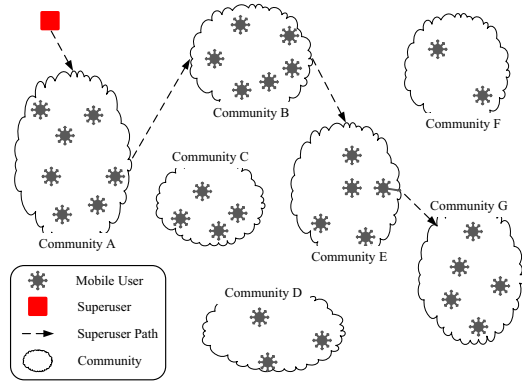

Fig. 1. Diagram demonstrating our active data dissemination scheme in Mobile Social Networks (MSNets)

much better chance of being connected to regular users than on average. Examples of popular geo-communities include public transportation and shopping centers in urban environments, conference rooms and cafeterias in office buildings, or bases and camps in military, etc.

We also propose geo-centrality, which is a geography-related centrality metric in MSNets and will be described in details in Section III-B, to measure the dynamic density of geo-communities. Such metric will be further used in the superuser route design.

B. Big Picture

We consider a following scenario: A salesman is about to advertise his products to on-campus customers (e.g. faculty, staffs, and students). He has to physically move around the campus to transmit advertisements via his smartphone to users'. He is trying to decide his route aiming to disseminate the ads to mobile users as soon as possible.

In our active data dissemination scheme, the users in MSNets are classified into two categories. (a) Regular users, or simply the users, that move based on their social lives. These users are potential data receivers from the superuser. The movements of the users are not controllable. (b) A single special user called superuser (e.g. the salesman) that aims to disseminate data to regular users in the network. In this paper, we consider that the superuser broadcasts data, i.e., all the regular users in the network are data destinations. Since most users are supposed to be selfish [16] and do not like to give up their own resources to help others unless some incentives be responded. We only focus on the data dissemination from superuser to regular users, and do not consider the opportunistic transmission between regular users. The problem is therefore how to design superuser route to facilitate data dissemination most effectively. To solve such a problem, we should focus on answering the following questions:

- What are the appropriate social-based metrics for measuring the dynamic density of geo-communities?
- Given the dynamic density of each geo-community, how should the superuser decide which geo-communities to stay and for how long, respectively?

Generally speaking, exploitation of the regularity of human mobility in social networks will definitely facilitate the calculation of the dynamic density of geo-communities in the network. As stated in Section II-A, users in MSNets move around several well-visited geo-communities, and their dwell times at each geo-community remain stable over time. Hence, we employ a Markov process to model user mobility in the network, where the communities are represented as Markovian states.

Through the analysis of Markov model, we can compute each user's steady-state probability distribution over geo-communities, and further propose geo-centrality, the cumulative contact probability to users, for each geo-community to measure its dynamic density. Suppose the whole network is composed of a number of geo-communities, the problem of superuser route design is then how to choose geo-communities and allocate time accordingly. Fig. 1 gives an illustration of our active data dissemination scheme.

C. Optimization Objective

The specific objective of our work is to design the superuser route to achieve the delivery ratio p with the minimum total duration of route. For example, the practical goal of the salesman is to disseminate the advertisement to a certain percent of users in the network within the minimum time. We define the required delivery ratio p as the average ratio of users have received the data item from the superuser, and such ratio is defined from a statistic perspective.

Our superuser route design scheme follows a utility-based approach. The superuser route comprises some geo-communities and the according dwell times. Suppose that the whole network is composed of a certain number of geo-communities, and the superuser can deliver data to users who affiliate to a geo-community where the superuser is stopping. The utility u_i of geo-Community i describes its potential contribution to the superuser's data dissemination. Obviously, the number of users to whom the superuser delivers data at geo-Community i will not decrease along with t_i, which indicates the superuser's dwell time at geo-Community i. In other words, $u_i(t_i)$ is a non-decreasing function of t_i. However, the key point is that the increase range of $u_i(t_i)$ with t_i is different among geo-communities, we should then allocate the limited time to geo-communities with higher gradient of $u_i(t_i)$. It will be shown later the associated utility of a geo-community is related to its geo-centrality.

The optimization objective is therefore to minimize the total duration of route, and at the same time guarantee the required delivery ratio p.

$$\min\ T$$
$$s.t. \sum_{i=1}^{J} u_i(t_i) \geq p \tag{1}$$
$$t_i \geq 0, 1 \leq i \leq J$$

where J indicates the total number of geo-communities in the network. Note that T represents the total duration of route, which includes waiting time T_w and journey time T_j, i.e., the total route time $T = T_w + T_j$. The corresponding optimization problem can be solved with the methods we outline in this paper, as discussed in Section V.

III. MOBILE SOCIAL NETWORKS (MSNETS)

MSNets are the graphs of interactions between mobile users, which play an important role in the dissemination of information, innovations, and diseases. We explore the geography-related characteristics of MSNets in this section, and further propose geo-community and geo-centrality.

A. geo-Community

A community is defined as a clustering of users that are "tightly" linked to each other, either by direct linkage or by some "easily accessible" users that can act as intermediates. Members of a community usually share interesting properties, such as common hobbies, social functions, and occupations [17]. On campus, graduate students working in the same office interact more frequently with each other; members affiliating with the same club such as football or swimming, have such heavy interactions. In an academic conference, scholars having mutual research interests also make up a community.

Interests often highly relates to geography in human society [14]: Officemates contact each other in the office; volleyball lovers play volleyball together in gyms; scholars share and communicate their research interests in conferences. Some social networks do not relate to geography such as Internet social networks, which are discussed in some literature. However, we do not consider them in this paper, we focus instead on geography-related MSNets and define the community in such networks as *geo-community*.[1] The social relationships among people form stable network structures, and the geo-community shows its efficiency in the design of active data dissemination schemes.

A single user can be a member of several different communities (people have varying roles in society). However, we consider that a user only shows one membership at any single time in this paper. In other words, the whole network is composed of a certain number of geo-communities, where a user contacts exactly one geo-community at any one time. While the user can change geo-community affiliation over time, and we assume he/she does not spend any time on the transition between geo-communities. For example, consider a father in family, who loves chess and swimming, works as a leader in a doll factory. He therefore affiliates to the above four geo-communities: Family, Chess Center, Natatorium and Factory. However, he affiliates to only one of those four geo-communities at any one time. Note that this does not preclude a user from belonging to multiple geo-communities.

---

[1]We will use the terms, *community* and *geo-community* interchangeably in subsequent sections.

B. geo-Centrality

'Betweenness' centrality, who measures the extent to which a node lies on the paths linking other nodes, is proposed in graph theory and network analysis at first and currently a widely used Freeman's measure in social-based data forwarding [5][6]. In this paper, we use 'betweenness' to measure the dynamic density of a geo-community instead, and a geo-community with higher betweenness has better capability of contacting mobile users in social networks.

Certainly, betweenness is defined and calculated based on the topology of network contact graph, and is not sufficient to analytically represent the probabilities for a geo-community to contact mobile users in the network. Inspired by [18], we propose geo-centrality, a centrality metric of geo-communities, in mobile social networks:

Definition 1. *Suppose there are totally N users in the network, and the steady-state contact probability per unit time between geo-Community i to $User_k$ is ϕ_i^k. The **geo-centrality** of geo-Community i is defined as*

$$C_i(t_i) = 1 - \frac{1}{N}\sum_{k=1}^{N}(1-\phi_i^k)^{t_i} \tag{2}$$

where $C_i(t_i)$ indicates the average probability that a randomly chosen user in the network is contacted by geo-Community i within time t_i, and the computation of ϕ_i^k will be described in Section IV-B. Steady state is a situation in which all state variables (i.e., the transition probability matrix P and sojourn time distribution $D_i(t)$ in semi-Markov model described in Section IV) are constants in spite of ongoing processes that strive to change them. The unit time means we focus on the discrete time system in our work. A user is contacted by a geo-community indicates that the user affiliats with that community, such as, the father is contacted by geo-community Factory when he's working.

IV. USER'S MOBILITY MODEL

As users in MSNets always belong to several communities, they usually move around these well-visited locations (i.e., geo-communities). Therefore, we can model user mobility as Markov process, where the states space can be represented by the set of geo-communities. Instead of continuous-time Markov chains, we use semi-Markov processes because the sojourn time during which a user is associated with a community can take multiple forms, besides the exponential or geometric distributions [19][20]. Semi-Markov process was first introduced by Lee et al. [21] to model user mobility in DTNs. They focused on analyzing the load balance among APs, and explored the characteristics of user mobility using the trace data collected from laptops but not cellphones, which weakened the effectiveness of modeling real-time user mobility. Furthermore, they did not elaborately explain the computation methods of the critical parameters in the model. Similar model has also been used in [15] to present a passive routing protocol in DTNs, where they employed a synthetic model to analyze the performance of the presented

protocol. Moreover, [21] and [15] laid particular attention to the transient behaviors of the model and did not consider any social network concept. In this section, we describe the semi-Markov processes we employ to model user mobility in MSNets, and further propose the computation methods of several important parameters in the model. For simplicity, we assume the renewal process is time homogeneous during the period in which the mobility model is built.

A. Time Homogeneous Semi-Markov Model

We consider user's mobility as a Markov renewal process $\{(X_n, T_n) : n \geq 0\}$, where T_n is the time instant of the n-th transition ($T_0 = 0$) and $X_n \in \mathbb{S}$ is the state at the n-th transition. The states space is represented by the set of geo-communities $\mathbb{S} = \{1, 2, \dots, J\}$. A user that moves between two geo-communities transfers in the markov process between the corresponding states. Random variable $T_{n+1} - T_n$ describes the geo-communities sojourn time. Then, the associated time homogeneous semi-Markov kernel Q is defined by:

$$\begin{aligned} Q_{ij}(t) &= Pr(X_{n+1} = j, T_{n+1} - T_n \leq t | X_n = i) \\ &= p_{ij}H_{ij}(t), i, j \in \mathbb{S} \end{aligned} \tag{3}$$

Suppose $P = [p_{ij}]$ is the transition probability matrix of the embedded Markov chain, where the transition probability from state i to state j is

$$p_{ij} = \lim_{t \to \infty} Q_{ij}(t) = Pr(X_{n+1} = j | X_n = i)$$

We also derive the sojourn time probability distribution in state i regardless of the next state.

$$D_i(t) \triangleq Pr(T_{n+1} - T_n \leq t | X_n = i)$$

Note that the distribution of the sojourn time, $D_i(t)$, during which the user is associated with geo-Community i before his/her next transition takes place can be expressed as

$$D_i(t) = \sum_{j=1}^{J} Q_{ij}(t).$$

With the transition probability matrix P and the sojourn time distribution $D_i(t)$ of the above semi-Markov process, we can characterize user mobility in MSNets. Section IV-B describes how to derive these probabilities and further shows the computation of steady-state probability distribution ϕ_i^k (proposed in *Definition 1.*) over geo-communities for users.

B. Steady-state Probability Distribution over geo-communities

Each user has his/her own spatial distribution that reflects his/her own trajectory. We therefore model semi-Markov process on every user separatively, in order to compute the steady-state probability distribution ϕ_i of each user. Without loss of generality, we illustrate how to compute $User_k$'s steady-state probability distribution $[\phi_i^k], i = 1, 2, \dots, J$, a $1 \times J$ vector. To determine the steady-state probability distribution $[\phi_i^k]$ of $User_k$, we need to compute two parameters, the

transition probability matrix P^k and the *sojourn time probability distribution matrix $D_i^k(t)$*. In this section we describe a method to determine these two parameters using user mobility history, and then propose how to compute ϕ_i^k with these two parameters.

1) Transition Probability Matrix: Suppose the transition probability matrix of the embedded Markov chain for $User_k$ is P^k. We take the father mentioned in Section III-A as example, he visits four geo-communities: Family, Chess Center, Natatorium and Factory. At any one of those geo-communities, he could pick to reside for a while or move to another geo-community according to his preferred probability. Those mobility probabilities constitute the transition probability matrix P^k. We now define the P^k probabilities as follows.

Definition 2. *The probability p_{ij}^k that $User_k$ moves from geo-Community i to geo-Community j is defined as the observed transition frequency:*

$$For\ each\ i \neq j, p_{ij}^k = Pr(X_{t+1}^k = j | X_{t+1}^k \neq i, X_t^k = i),$$
$$with\ \sum_{j \neq i} p_{ij} = 1, and\ p_{ii} = 0$$

When we compute P^k, sojourn times are not taken into consideration. We only consider the transition between different states ($p_{ii} = 0$).

2) Sojourn Time Probability Distribution: Let $D_i^k(t)$ be the probability of the sojourn time at geo-community i for $User_k$ regardless of the next geo-community, which can be calculated as follows:

$$D_i^k(u) = Pr(X_{t+u+1}^k \neq i, X_{t+u-v}^k = i, v = 0, \ldots, u-2|$$
$$X_{t+1}^k = i, X_t^k \neq i), u = 1, \ldots M_i$$

where M_i represents the upper bound to the time spent in geo-Community i. We assume that when the network reaches steady state, the mobility history provides a representative sample from which the sojourn time distribution can be drawn.

In Markov processes, the sojourn time is usually considered to have an exponential distribution. The use of a semi-Markov model in this paper eliminates such constraint and can reflect the real world processes even more.

3) Computation of ϕ_i^k: Given the transition probability matrix P^k, we can derive the *steady-state transition probability* $\pi^k = [\pi_1^k, ..., \pi_J^k]$ by solving the following equations:

$$\pi^k = \pi^k P^k, \sum_{i=1}^{J} \pi_i^k = 1.$$

In fact, π_i^k denotes the probability of $User_k$ being in geo-Community i at some transition instants.

Then, with the sojourn time probability distribution $D_i^k(t)$, we define the $J \times 1$ mean residence time vector $\bar{D}^k = [\bar{d}_i^k]$, where $\bar{d}_i^k$ is the mean value for $D_i^k(t)$. We can characterize user mobility by calculating the steady-state user distribution $\phi^k = [\phi_i^k]$ as follows:

$$\phi_i^k = \frac{\bar{d}_i^k \pi_i^k}{\sum_{i=1}^{J} \bar{d}_i^k \pi_i^k}$$

The steady-state distribution $[\phi_i^k]$ is the probability distribution of $User_k$ over geo-communities at any instant, and is hence the long-term average distribution of $User_k$ over geo-communities. We use ϕ_i^k to represent the contact probability per unit time between geo-Community i and $User_k$.

V. DESIGNING ALGORITHMS FOR THE SUPERUSER ROUTE

In this section, we discuss how the superuser controls its trajectory to meet mobile users in the network with the goal of minimizing the total duration, at the same time guaranteeing the given data dissemination probability. We start with describing the key idea behind this process. Later we describe the route design algorithms in detail.

The main difficulty in designing superuser route for MSNets is that we cannot correctly predict the location of the users (a user may affiliates to several geo-communities), and hence it may not be possible to deterministically position the superuser to contact certain regular user. However, the steady-state probability ϕ_i^k of presence of $User_k$ in geo-Community i (he/she belongs to that geo-community), is non-zero, then we can contact $User_k$ with certainty if we wait in geo-Community i for long enough. This probability approaches 1 only as the waiting time approaches infinity. Obviously, we cannot afford to wait for infinitely long, and we hence cannot afford to contact the mobile users with certainty. However, it is possible to meet the users with a desired probability by waiting a finite amount time at a geo-community, as long as the steady state probability of user presence in that geo-community is substantial such that the desired meeting probability is modest (i.e., large but not approaching 1).

We have the knowledge of the geo-centrality $C_i(t_i)$ for every geo-Community i in the network, then our next step is to choose waiting times $t_i(\geq 0)$ at each geo-Community i, and ordering these geo-communities together to form a tour. In this section, we propose two algorithms for solving the optimization problem.

A. Static Optimal Disseminating Algorithm (SODA)

T in Eq. (1) represents the total duration of the superuser route, which has two components (a) Waiting time in the Route T_w: The sum of waiting times at the chosen geo-communities. (b) Journey Time in the Route T_j: The total time that the superuser spends traveling between geo-communities. The total route time $T = T_w + T_j$. We assume that only T_w contributes to the superuser's data dissemination, so we can divide Eq. (1) into two parts: (i) finding a good set of geo-communities and their corresponding waiting times, and (ii) ordering these geo-communities together to form a tour. We then look at the two steps independently.

1) Choosing appropriate geo-communities: We have the knowledge of the geo-centrality function of geo-Community $i(1 \leq i \leq J)$, then our next step is to choose waiting times t_i corresponding to each geo-Community i, so that the total data dissemination probability for the superuser approaches p. Obviously, the geo-communities with $t_i \neq 0$ are selected as the stopping sites of the superuser. Clearly, we want to minimize the total waiting time. The corresponding optimization problem is as follows:

$$\min \sum_{i=1}^{J} t_i$$
$$s.t. \sum_{i=1}^{J} C_i(t_i) \geq p \quad (4)$$
$$t_i \geq 0, 1 \leq i \leq J$$

From Eq. (2), $C_i(t_i)$ is the sum of logarithmic functions, then Eq. (4) becomes a convex optimization problem, which can be solved by *interior-point methods*.

Interior-point methods are always used for solving the following optimization problems that include inequality constraints.

$$\min f_0(x)$$
$$s.t. f_i(x) \leq 0, 1 \leq i \leq m$$

where $f_0, \ldots, f_m : \mathbf{R}^n \to \mathbf{R}$ are convex and twice continuously differentiable. The convex optimization problem is solvable, i.e., an optimal $x^\star$ exists [22]. Obviously, our optimization satisfies the required condition. Since there is only one inequality constraint in Eq. (4), we can transform the inequality into the objective function, then the optimization problem becomes an unconstrained problem as follows:

$$\min f_0(x) + \sum_{i=1}^{m} -(1/t) \log(-f_i(x))$$

To solve the above unconstrained problem, *barrier method* can be employed, which is based on solving a sequence of unconstrained minimization problems, using the last point found as the starting point for the next unconstrained minimization problem. In other words, we compute $x^\star(t)$ for a sequence of increasing values of t, until $t \geq m/\epsilon$, which guarantees that we have an ϵ-suboptimal solution of the original problem.

The sketch of the algorithm [22] is summarized as Algorithm 1, where ϵ is the approximation ratio to the optimal objective value. In the process of *step 1*, the Newton's method is adopted to solve the unconstrained problem $tf_0 + \phi$, where $\phi(x) = -\sum_{i=1}^{m} \log(-f_i(x))$ is the *logarithmic barrier* function.

Algorithm 1 Barrier Method in Static Optimal Disseminating Algorithm (*SODA*)

---

given strictly feasible $x, t := t^{(0)} > 0, \mu > 1$, tolerance $\epsilon > 0$.
repeat
1. *Centering Step.* Compute $x^\star(t)$ by minimizing $tf_0 + \phi$, starting at x
2. *Update.* $x := x^*(t)$
3. *Stopping criterion.* **quit** if $m/t < \epsilon$
4. *Increase t.* $t := \mu t$

---

2) Constructing a path from chosen geo-communities: Once we have determined the geo-communities, we order them so as to minimize the length of the route. This amounts to the traveling salesman problem (TSP) whose exact solution is NP-hard. TSP solvers like Concorde [23] can solve the problem exactly for few hundred points within minutes. If the number of points is large, then we can choose any of the available approximation algorithms [24] that exist for TSP.

However, a user can belong to several communities, which introduces potential *overlap* among contact user sets of geo-communities in the network.

Definition 3. *Suppose there are N mobile users in the network, and the steady-state contact probability per unit time between geo-Community i to $User_k$ is ϕ_i^k. The **contact user set** of geo-Community i is defined as*

$$\mathbb{W}_i = \{v_k | v_k \in \bigcup_{1 \leq k \leq N} (\phi_i^k \neq 0)\}$$

Consider the example of the father described in Section III-A, who is the mutual member of four communities, which means his steady-state distribution contributes to the geo-centrality of all these four geo-communities. However, the superuser route is comprised of some *ordered* geo-communities, i.e., in the form of geo-community scheduling. It is possible that the superuser has already delivered the data to the father in one of those geo-communities, such as Factory. Then the contribution of father to the other three communities' geo-centrality should not be considered anymore, because the superuser does not need to disseminate data to the same users more than once, i.e., we only count delivered users, which are the users that have received the data, but not the number of delivery times. Hence, we define the above algorithm for the superuser route design as Static Optimal Disseminating Algorithm (*SODA*), and further propose Greedy Adaptive Disseminating Algorithm (*GADA*), which introduces scheme of updating geo-centrality for communities each step, in terms of all the non-contacted users in the network.

B. Greedy Adaptive Disseminating Algorithm (GADA)

In this algorithm, we also choose geo-centrality as community's utility, but it instead computes geo-centrality of *non-contacted users* for each community repeatedly. In other words, *GADA* overcomes the overlap among $\mathbb{W}_i (1 \leq i \leq J)$ by updating the current geo-centrality of each community dynamically.

Algorithm 2 Greedy Adaptive Disseminating Algorithm (*GADA*)

1: $\mathbb{G} \leftarrow \varnothing; \mathbb{U} \leftarrow \mathbb{S}; T$
2: Compute $C'_i(0)$ for every $i \in \mathbb{S}$
3: Stop at the geo-community with maximal $C'_i(0)$
4: **for** $(t = 1; t < T; t++)$ **do**
5:     **if** $User_k \in Community_{cur}$ **then**
6:         $\mathbb{G} \leftarrow \mathbb{G} \bigcup User_k$
7:     **end if**
8:     $\mathbf{a}[i] = \bar{C}'_i(0), 1 \le i \le J, i \ne cur$
9:     $temp = \mathbf{a}[j] = \max \mathbf{a}$
10:     **if** $(\bar{C}'_{cur}(t_{soj}) \le temp) \wedge (\bar{C}_j(T - t - t_{cur,j}) \ge (\bar{C}_{cur}(T) - \bar{C}_{cur}(t_{soj})))$ **then**
11:         Move to Community j
12:     **else**
13:         Stay at the current community
14:     **end if**
15: **end for**

Throughout the rest of this section we use the following notation. Given a collection of geo-communities $\mathbb{S} = \{1, 2, \ldots, J\}$ over a domain of users $\mathbb{M} = \{1, 2, \ldots, N\}$. Let $\mathbb{G}$ be a collection of *contacted* users (i.e., the users who have already received the data from the superuser). Given $C_i(t_i)$ as the geo-centrality function of geo-Community i during waiting time t_i, we further propose $\bar{C}_i(t_i)$ to denote such centrality of *non-contacted* users covered by geo-Community i (i.e., facing users not covered by set $\mathbb{G}$).

Algorithm 2 shows the details of *GADA*, where T represents the time constraint for the superuser route, the subscript *cur* indicates the current community where the superuser stays. t_{soj} is the waiting time at the current community, and $t_{cur,j}$ indicates the travel time from the current community to Community j, which is a constant and known by the superuser as described before. $\bar{C}_i{}'(0)$ stands for the gradient of $\bar{C}_i(t_i)$ at $t_i = 0$.

We elaborately illustrate *Step* $7. - 10.$ in Algorithm 2. Intuitively, *GADA* aims to maximize the centrality around the network each step. Obviously, the superuser will choose the geo-community with maximal $C'_i(t)|_{t=0}$ as the first stop. What matters is *if* and *when* the superuser should move to other geo-communities. Without loss of generality, we consider the condition of two geo-communities in the network. Suppose there are two geo-Communities i and j with $C'_i(t)|_{t=0} > C'_j(t)|_{t=0}$, the travel time $t_{i,j}$ between two geo-communities is a constant.

Assumptions on *when* and *if* the superuser should move to other geo-communities are:

$C1$: The sojourn time t_i for superuser stays at geo-Community i before leaving for geo-Community j is

$$C'_i(t)|_{t=t_i} = C'_j(t)|_{t=0} \tag{5}$$

$C2$:

$$C_j(t_j) \ge C_i(T) - C_i(t_i) \tag{6}$$

Theorem 1: Suppose assumptions C1-C2 hold, then the total centrality will achieve maximum within time constraint T.

where Eq. (6) is obvious, since if the travel cost of moving to another geo-community is less than the total utility gain, the superuser should move; otherwise, the superuser would better stay at the current geo-community. However, we prove the optimal transition time instant (Eq. (5)) as follows:

Proof: The time constraint for the superuser is $T = t_i + t_j + t_{i,j}$, let $T' = T - t_{i,j}$, then $t_i + t_j = T'$. Since T and $t_{i,j}$ are constants, T' is a constant.

It is easy to show that $\frac{\partial C_i(t_i)}{t_i} = \frac{\partial C_j(T'-t_i)}{t_i}$ since $C'_i(t)|_{t=t_i} = C'_j(t)|_{t=0}$, then $\frac{\partial C_i(t_i)}{t_i} + \frac{\partial C_j(t_j)}{t_j} = 0$, and further $\frac{\mathbf{d}(C_i(t_i)+C_j(t_j))}{\mathbf{d}t_i} = 0$. The total centrality function $C = C_i(t_i) + C_j(t_j)$, which is a concave function and will achieve maximum at $\frac{\mathbf{d}C}{\mathbf{d}t_i} = 0$. ∎

Note that the prerequisite of *Theorem* 1 is that the two geo-communities have unchanged centrality function, *GADA* instead aims at dynamic centrality of geo-communities. However, the algorithm can guarantee the maximal total utility for the whole system at the transition time instant (i.e., t_i).

In contrast to *SODA*, *GADA* can overcome the overlap among geo-communities in the network by facing *non-contacted users* each step, but the trade-off is introducing more computational overhead.

VI. EMPIRICAL STUDY

A. Simulation Setup

We use two experimental traces collected from realistic MSNets to validate our time-homogeneous semi-Markov modeling, and to evaluate the performance of our active data dissemination schemes. We believe that the chosen traces cover a large diversity of environments, from university campuses (*MIT Reality*) [25] to conference sites (*Infocom 06*) [26], with experimental periods from a few days (*Infocom 06*) to several months (*MIT Reality*).

The reason we choose datasets containing static APs is that APs can help identify geo-communities thanks to their geography-related property. Concretely, we use the *syslog* data for mobile users' association patterns to APs. Each *syslog* message contains a timestamp in seconds, the clients MAC address, the AP name, and the event type. From these *syslog* messages, the mobility of each user is extracted in the form of a series of two tuples (AP name, the timestamp when the association with this AP occurs). In our work, we use the neighborhood of an AP to represent a geo-community, and consider the sojourn time of a participant spending at a geo-community as the time interval of his/her two consecutive contacts with different APs, the former of which is the corresponding geo-community.

B. Semi-Markov Model Evaluation

1) User X's mean residence time distribution: We use *MIT Reality* with larger network scale and longer experiment period to validate our semi-Markov modeling. Fig. 2 gives the mean residence time at different regions of user X on the map, and we have also analyzed the mean residence time at each geo-community($\bar{D}$) of user X. It shows that user X spends
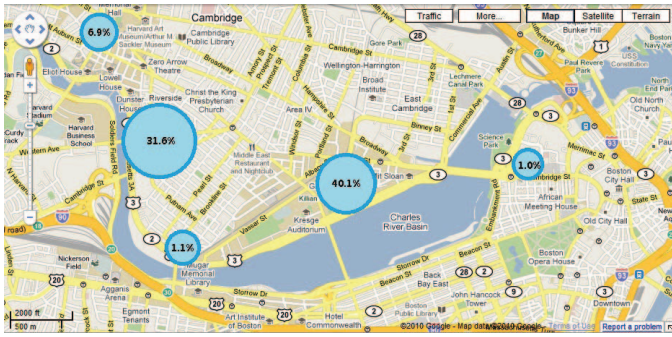
Fig. 2. Mean residence time distribution over geo-communities on the map for user X

	m1	m2	m3	m4	m5	m6	m7
m1	1	0.9693	0.9405	0.3183	0.8366	0.9924	0.9270
m2	-	1	0.9940	0.3210	0.9373	0.9584	0.9900
m3	-	-	1	0.3183	0.9601	0.9292	0.9984
m4	-	-	-	1	0.3001	0.3131	0.3152
m5	-	-	-	-	1	0.8100	0.9643
m6	-	-	-	-	-	1	0.9175
m7	-	-	-	-	-	-	1

	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Sun	1	0.7077	0.7347	0.6937	0.6700	0.2898	0.5709
Mon	-	1	0.9517	0.9422	0.8667	0.4025	0.7478
Tue	-	-	1	0.8865	0.9109	0.2649	0.6760
Wed	-	-	-	1	0.8097	0.6199	0.8870
Thu	-	-	-	-	1	0.1230	0.5250
Fri	-	-	-	-	-	1	0.8758
Sat	-	-	-	-	-	-	1

71.7% of her overall time at home or work, where 40.1% of overall time elapsed at 6 different geo-communities. An even more interesting phenomenon is found when we consider the distribution of the remaining 9% for user X, these remaining time is spent in geo-communities that each appear less than 1% of time. In other words the spatiotemporal distribution for user X shows a heavy tail, which enhances the using of semi-Markov process to model user's mobility in MSNets.

2) Similarity of user mobility in time intervals of different scales: We calculate the steady-state distributions of user X over geo-communities based on traces collected in time intervals of different scales, in order to investigate whether or not, and to what extent, mobility behaviors correlate in time. We choose *cosine distance* as the similarity measure. In the current problem setting, the cosine distance $sim(\vec{p}, \vec{q})$ is defined as

$$sim(\vec{p}, \vec{q}) = \frac{\sum_{i=1}^{J} p_i q_i}{\left(\sqrt{\sum_{i=1}^{J} p_i^2}\right)\left(\sqrt{\sum_{i=1}^{J} q_i^2}\right)} = \frac{\vec{p} \cdot \vec{q}}{|\vec{p}||\vec{q}|}$$

where $\vec{p} = [p_i]$ and $\vec{q} = [q_i]$ are the steady-state distributions of two mobility models under comparison, and p_i and q_i are the probabilities that a user contact geo-Community i under the two models, respectively. Note that $sim(\vec{p}, \vec{q})$ ranges in $[0, 1]$, with $sim(\vec{p}, \vec{q}) = 1$ indicating that the two mobility behaviors are identical as far as the long-term user distribution is concerned.

Table I gives the above similarity measure between steady-state distributions derived based on monthly traces collected in a period of 7 months (28-week period) between 19 September, 2004 and 4 March, 2005. It can be observed that the closer in time the two monthly traces, the more similar the corresponding two steady-state distributions, though there exist some exceptions. For example, the cosine distance becomes much smaller when *m4* is involved. This is because *m4* corresponds to the period of *winter break* (from 12 December, 2004 to 8 January, 2005). This administrative event on campus affects the user mobility dramatically.

Table II gives the cosine distance between steady-state distributions derived based on daily traces. It is again confirmed that the close similarity of user mobility between each day of

a week. An interesting phenomenon is that *Friday* has a lower similarity with other days, since people always have fun at Friday night.

The results verify the two assumptions we made in Sec. II-A on MSNets. First, mobile users in a social network usually move around several well-visited geo-communities instead of moving randomly. Second, the user's dwell time at each geo-community is fairly regular, since their social behavior patterns always remain stable according to their social lives.

C. Performance Evaluation of The Superuser Route Algorithms

We use *Infocom 06* trace with AP locations on the map to evaluate the performance of our active data dissemination schemes. We extracted the distance between any two APs from the map of conference site[2], and treat it as the moving distance of the superuser between the two corresponding geo-communities. We assume that there is a single superuser in the system, and compare our schemes (*SODA* and *GADA*)[3] with the following two Message-Ferry based routing schemes [8] [10]:

1) **Message Ferry moves with Restricted Random Way-point model (MF-RRWP)**: The ferry moves according to the random way-point mobility model, with the restriction that the way-points are only chosen from the center of each geo-community. At each way-point, the ferry pauses for exponentially distributed time with mean of 15 minutes. Note that this ferry model can

---

[2]The performance is supposed to be evaluated in a larger mobile social environment, but to the best of our knowledge, *Infocom 06* dataset is the only one who both provides the geography information of all the APs and also represents the real movements of the users (i.e., it is collected using mobile phones but not laptops). Therefore, we magnify the map of the conference site on a scale of 1:10.

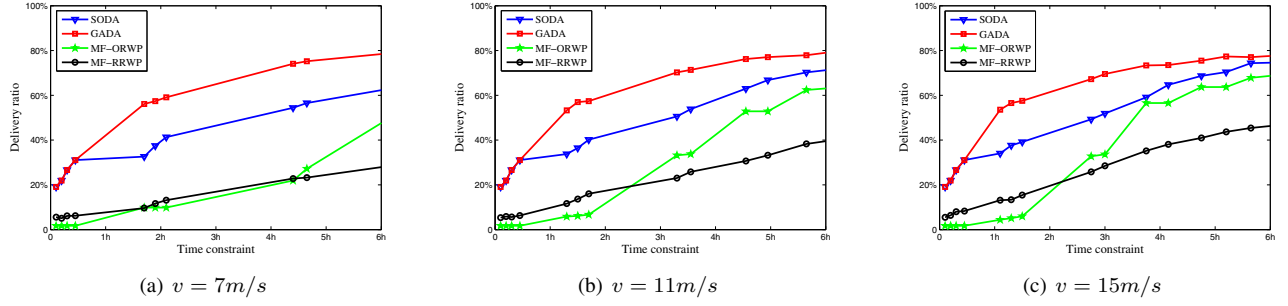[3]We set 3 minutes as time unit in *SODA* and *GADA*.

Fig. 3. Delivery ratio under different superuser routing schemes, and under different superuser speed
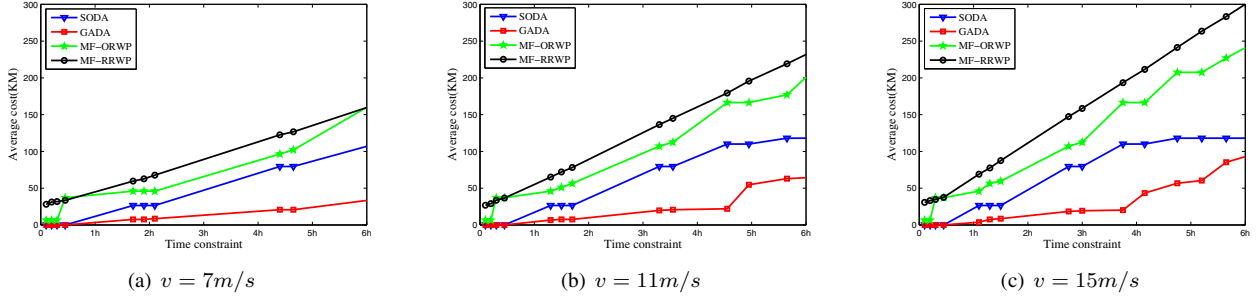

Fig. 4. Average cost under different superuser routing schemes, and under different superuser speed

also thought of as one where the ferry visits one geo-community after the other, at random.

2) **Message Ferry moves along ordered set of way-points (MF-ORWP)**: In this model, we also pick the center of each geo-community as a way-point, and order these way-points so as to form a shortest possible tour using the Concorde Traveling Salesman (TSP) solver [23]. The message ferry route is in the form of traversing the ordered set of way-points repeatedly.

In our simulation, we focus on the following two metrics, which are key characteristics in data dissemination of MSNets.

1) **Delivery ratio**, the ratio of the number of delivered users to the total number of users in the network.
2) **Average cost**, the traveling distance of the superuser. Note that although the superuser is not limited in power supply, we still aim to maximize the energy efficiency.

One might think that since *MF-ORWP* scheme covers the entire region, they would perform well. However, there are good reasons for its poor performance. For dissemination probability $p = 60\%$ (in *SODA*) and the superuser speed of $15m/s$, the length of the route (tour) for *MF-ORWP* is $207.5km$, compare this to the route length of $8.5km$ that we observed for *GADA*. Fig. 4 shows the high cost of *MF-ORWP*, almost equal to *MF-RRWP*. With the same speed, the longer route length means that the superuser takes a long time on the journey, and a significant fraction of this time is spent traveling in the parts of the region that have zero or negligible probability of user presence. Note that even though the superuser covers the entire region, it does not cover the entire region at once, especially when the region area is very

large, so the users and superuser can keep missing each other. We can also observe from fig. 3 that the delivery ratio of *MF-ORWP* rises as the superuser accelerates.

For *MF-RRWP*, the superuser may choose random geo-communities having no mobile user, thus time spent traveling to and staying at such communities is completely futile, except for when the path to these communities intersects the region of some mobile users.

Overall, *SODA* and *GADA* both perform significantly better than the other two MF-based schemes, with higher delivery ratio and lower cost. The main reason is that we balance the traveling time and waiting time, and moreover invest waiting time at geo-communities that are most advantageous in terms of increasing the contact probability with the users. It can be observed from fig. 3 that *SODA* performs as good as *GADA* when the time constraint T is small (less than 30 minutes in the experiment), since the superuser in *SODA* doesn't have enough time to visit a certain user more than once in such a little time. As T gets larger, *GADA* takes advantage of updating geo-centrality metrics in terms of non-contacted users, certainly with more computational overhead.

VII. RELATED WORK

In the context of Delay-Tolerant Networks (DTNs), Pocket Switched Networks (PSNs) and Opportunistic Networks, *passive* schemes taking advantage of the social behaviors of mobile users have been proposed for data forwarding [5][6] and content dissemination [4]. SimBet routing [5] uses ego-centric betweenness metric as utility and only forwards data to nodes with higher ones. Hui et al. consider node centrality as well as social community knowledge in [6]. Researches on

identification of social communities have also emerged [6][17]. However, all of these schemes use the intrinsic mobility of the nodes in the network.

Another set of work considers the possibility of controlled mobility for network routing [8][9][11][12][13]. They proposed the communication models where a special mobile node (such as Message Ferry [8][9] or Data MULEs [13] *etc.*) facilitates the network connectivity. However, those models always consider the network with stationary nodes, or need to disturb nodes' motion trajectory. Tariq et al. [10] aims at designing the ferry route not interrupt nodes' motion in mobile DTNs, but they didn't consider the social nature of the network. On the contrary, our data dissemination scheme exploits the social characteristics of mobile network without any online collaboration between the superuser and regular users in the network. Though we focus on a different application (data dissemination from the superuser to regular users), our superuser also can extend to work as "data carrier" between regular users. As such, it strengthens research of both active routing schemes and further the foundations in the area of MSNets.

VIII. Conclusions and discussion

In this paper, we studied active data dissemination in MSNets, the main idea behind which is to exploit social properties of regular user's mobility to facilitate data delivery on purpose. Specifically, we introduced a semi-Markov process relating geography and social-network concepts for modelling regular user's mobility. The superuser route comprises some geo-communities and the according dwell times that both are calculated carefully based on such model. Extensive trace-driven simulation results show that the dissemination schemes presented by this paper perform significantly better than other existing active routing schemes. We believe that this paper presents the first step in exploiting social network methods for efficient active data dissemination in MSNets.

We focus on the active data-dissemination scheme, i.e., we only consider one-step transmission directly from the superuser to mobile users in the network. Many existing research in related networks, such as in DTNs and PSNs, consider the opportunistic forwarding between peer nodes (i.e., passive scheme), where however still exists one important problem to be solved: How to design the efficient incentive mechanism for participants in the network, since the data forwarding will definitely drain mobile devices' batteries. It must be expected that many users will be selfish and try to exploit the system to gain performance, without giving up their own resources to help others. Certainly it is possible that some users will forward the data to friends who are interested in that. Therefore, our dissemination schemes can be viewed as a lower bound on the delivery ratio in MSNets.

IX. Acknowledgement

References

[1] S. Ioannidis, A. Chaintreau, and L. Massoulie. Optimal and scalable distribution of content updates over a mobile social network. In *INFOCOM*, Rio de Janeiro, Brazil, Apr. 2009.

[2] J. Zhao and G. Cao. Vadd: Vehicle-assisted data delivery in vehicular ad hoc networks. *IEEE Transactions on Vehicular Technology*, 57(3), May 2008.

[3] S. He, J. Chen, Y. Sun, D. K. Y. Yau, and N. K. Yip. On optimal information capture by energy-constrained mobile sensors. *IEEE Transactions on Vehicular Technology*, 59(5), Jun. 2010.

[4] C. Boldrini, M. Conti, and A. Passarella. Modelling data dissemination in opportunistic networks. In *CHANTS*, San Francisco, California, USA, Sep. 2008.

[5] E. Daly and M. Haahr. Social network analysis for routing in disconnected delay-tolerant manets. In *MobiHoc*, Montreal, Quebec, Canada, Sep. 2007.

[6] P. Hui, J. Crowcroft, and E. Yoneki. Bubble rap: social-based forwarding in delay tolerant networks. In *MobiHoc*, Hong Kong SAR, China, May 2008.

[7] A. Balasubramanian, B. Levine, and A. Lindgren. Dtn routing as a resource allocation problem. In *SIGCOMM*, Kyoto, Japan, Aug. 2007.

[8] W. Zhao, M. Ammar, and E. Zegura. A message ferrying approach for data delivery in sparse mobile ad hoc networks. In *MobiHoc*, Roppongi, Japan, May 2004.

[9] W. Zhao, M. Ammar, and E. Zegura. Controlling the mobility of multiple data transport ferries in a delay-tolerant network. In *INFOCOM*, Miami, USA, March 2005.

[10] M. Tariq, M. Ammar, and E. Zegura. Message ferry route design for sparse ad hoc networks with mobile nodes. In *MobiHoc*, Florence, Italy, May 2006.

[11] B. Burns, O. Brock, and B. Levine. Mv routing and capacity building in disruption tolerant networks. In *INFOCOM*, Miami, USA, March 2005.

[12] S. Guo and S. Keshav. Fair and efficient scheduling in data ferrying networks. In *CoNEXT*, New York, NY, U.S.A, Dec. 2007.

[13] R. Shah, S. Roy, S. Jain, and W. Brunette. Data mules: Modeling a three-tier architecture for sparse sensor networks. *Elsevier Ad Hoc Networks Journal*, 1:215–233, Sep. 2003.

[14] L. David, N. Jasmine, K. Ravi, R. Prabhakar, and T. Andrew. Geographic routing in social networks. In *Proceedings of the National Academy of Sciences (PNAS)*, volume 102(33), pages 11623–11628, 2005.

[15] Q. Yuan, I. Cardei, and J. Wu. Predict and relay: an efficient routing in disruption-tolerant networks. In *MobiHoc*, New Orleans, Louisiana, USA, May 2009.

[16] Q. Li, S. Zhu, and G. Cao. Routing in socially selfish delay tolerant networks. In *INFOCOM*, San Diego, California, USA, Mar. 2010.

[17] P. Hui, E. Yoneki, S. Chan, and J. Crowcroft. Distributed community detection in delay tolerant networks. Kyoto, Japan, Aug. 2007.

[18] W. Gao, Q. Li, B. Zhao, and G. Cao. Multicasting in delay tolerant networks: A social network perspective. In *MobiHoc*, New Orleans, Louisiana, USA, May 2009.

[19] H. Cai and D. Eun. Crossing over the bounded domain: From exponential to power-law inter-meeting time in manet. In *MobiCom*, Montreal, Quebec, Canada, Sep. 2007.

[20] T. Karagiannis, J. Boudec, and M. Vojnovic. Power law and exponential decay of inter contact times between mobile devices. In *MobiCom*, Montreal, Quebec, Canada, Sep. 2007.

[21] J. Lee and J. Hou. Modeling steady-state and transient behaviors of user mobility: Formulation, analysis, and application. In *MobiHoc*, Florence, Italy, May 2006.

[22] S. Boyd and L. Vandenberghe. Convex optimization. *Cambridge University Press*, 2004.

[23] W. Cook, D. Applegate, R. Bixby, and V. Chvatal. Concorde: A code for solution of travelling salesman problem. http://www.tsp.gatech.edu/.

[24] V. Vazirani. Approximation algorithms. *Springer*, Aug. 2001.

[25] N. Eagle, A. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. In *Proceedings of the National Academy of Sciences (PNAS)*, volume 106(36), pages 15274–15278, 2009.

[26] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau. CRAWDAD data set cambridge/haggle (v. 2009-05-29). Downloaded from http://crawdad.cs.dartmouth.edu/cambridge/haggle, May 2009.