SOCIAL AND LOCATION BASED ROUTING IN DELAY TOLERANT NETWORKS

by

Ying Zhu

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2013

Approved by:

_____
Dr. Yu Wang

_____
Dr. Jamie Payton

_____
Dr. Weichao Wang

_____
Dr. Jiang Xie

_____
Dr. Yang Cao

ABSTRACT

YING ZHU.Social and location based routing in delay tolerant networks. (Under the direction of DR. YU WANG)

Delay tolerant networks (DTNs) are a special type of wireless mobile networks which may lack continuous network connectivity. Routing in DTNs is very challenging as it must handle network partitions, long delays, and dynamic topology in such networks. Recently, the consideration of social characteristics of mobile nodes provides a new angle of view in the design of DTNs routing protocols. In many DTNs, a multitude of mobile devices are used and carried by people (e.g. pocket switched networks and vehicular networks), whose behaviors are better described by social models. This opens the new possibilities of social-based routing, in which the knowledge of social characteristics is used for making better forwarding decision. However, the social relations do not necessarily reflect the true device communication opportunities in a dynamic DTN. On the other hand, the increasing availability of location technologies (GPS, GSM networks, etc.) enables mobile devices to obtain their locations easily. Consider that an individual's location history in the real world implies his/her social interests and behaviors to some extent, in this dissertation, we study new social based DTN routing protocols, which utilize location and/or social features to achieve efficient and stable routing for delay tolerant networks. We first incorporate the location features into the social-based DTN routing methods to improve their performance by treating location similarity among nodes as possible social relationship. Then, we discuss the possibility and methods to further improve routing performance by adding limited amount of throw-boxes into the networks to aid the DTN relay. Several throw-boxes based routing protocols and location selection methods for throw-boxes are proposed. All proposed routing methods are evaluated via extensive simulations with real life trace data (such as MIT reality, Nokia MDC, and Orange D4D).

ACKNOWLEDGMENTS

There are lots of people I would like to thank for their contribution or support on this thesis.

First, I would like to thank my advisor, Professor Yu Wang for his wonderful instruct and understanding during the past three years. This dissertation could not have been written without his insights and invaluable help and guidance. His visions in the area set the beginning of this thesis, and his involvement in working through vague ideas produced building blocks to the work.

I am grateful to other members of my thesis committee, Professor Jamie Payton, Professor Weichao Wang, Professor Jiang Xie, and Professor Yang Cao, for their service, interests and help.

To past and present members of our research group, Siyuan Chen, Minsu Huang, Chao Zhang, Yong Sun, Yingjian Liu, and Fei Gao, I thank them all for useful meetings and discussions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Delay or disruption tolerant networks (DTNs) [46, 88, 97] is a type of wireless mobile network that does not guarantee continuous network connectivity. In DTNs data delivery often occurs through physical node movement. Figure 1.1 illustrates the process: if node A wants to send a message to node C, the path may lead across node B, which moves at a given time towards the range of node C. The DTN is a network which aims to cope with:

- Network partitioning: Eventually no end-to-end connectivity between sender and destination is possible due to frequent or constant network partition. These partitions may occur because of geographical distance, lacking radio signal strength or other limiting factors.
- Network interruption: DTNs are often deployed in rough and adverse surroundings and nodes may be subject to numerous operation failures. These failures may cause network interruptions and disconnect linked nodes. Another assumption is that partitioned networks can be of heterogeneous structures, which means not all local networks are using the same underlying protocols and applications.
- High error rates: Presumably short connectivity and high mobility in combination with weak signal strength and/or other aggravating circumstances is leading to a high link error rate that makes end-to-end reliability difficult.
- Long delay: The intermittent connectivity causes long and hard to estimate delays. Data often has to be buffered or queued if there is no direct path to the destination node. Delays may last up to hours or days, depending on the mobility and connectivity within the network.
- Asymmetric data rates: Due to the intermittent characteristics of a DTN most com-

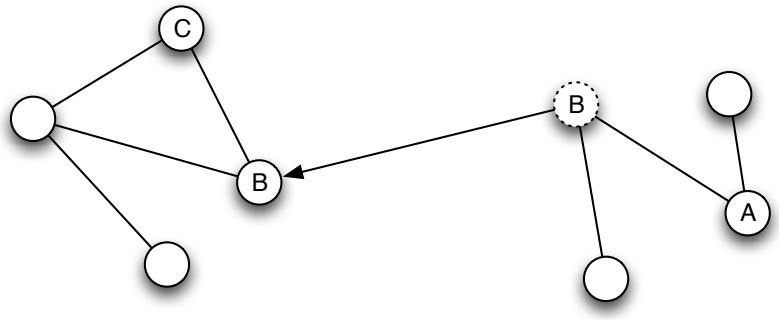


Figure 1.1: DTNs data delivery occurs often through node movement.

munication between two nodes will be mainly asymmetric. Elapsing time between request and answer will be rather hours than milliseconds.

- Energy, bandwidth, buffer and cost restrictions: As within ad-hoc networks, energy is a limited and valuable resource. Nodes may be turned off frequently to save energy and therefore weaken the already low connectivity. Other limiting factors are bandwidth bottlenecks and several cost restrictions. The available buffer-space restriction is far more limiting than in an ad-hoc network because most data can not be delivered immediately and directly to the destination and must be buffered at intermediate nodes.

DTNs have recently drawn much attention from networking researchers due to the wide applications of these networks in challenging environments, such as space communications, military operations, and mobile sensor networks. Intermittent connectivity in DTNs results in the lack of instantaneous end-to-end paths, large transmission delay and unstable network topology. These characteristics make the classical ad hoc routing protocols [45, 73, 78] not being applicable for DTNs, since these protocols rely on establishment of a complete end-to-end route from the source to the destination.

Many routing schemes [8, 10, 54–58, 81, 82, 87, 93, 95, 99, 100] have been proposed for DTNs. Most of these DTN routing protocols belong to three categories: *message-ferry-based*, *opportunity-based* and *prediction-based*. In message-ferry-based methods [10, 81, 99, 100], systems usually employ extra mobile nodes as ferries for message deliv-

ery. The trajectory of these ferries is controlled to improve delivery performance. However, controlling these nodes leads to extra cost and overhead. In opportunity-based schemes [46, 82, 87], nodes forward messages randomly hop by hop with the expectation of eventual delivery, but with no guarantees. Generally, messages are exchanged only when two nodes meet at the same place, and multiple copies of the same message are flooded in the network to increase the chance of delivery. Some DTN routing protocols [8, 54–57, 95] make relay selection by estimating metrics relative to successful delivery, such as delivery probability or expected delay based on a history of observations. Most of these protocols focus on whether and when two nodes will contact. Liu and Wu [58] also proposed a forwarding method based on a probabilistic forwarding metric, which is derived by modeling each forwarding as an optimal stopping rule problem.

All of the current DTN routing methods share a similar paradigm "store and forward". If there is no connection available at a particular time, a DTN node can store and carry the data until it encounters other nodes. When the node has such a forwarding opportunity, all encountered nodes could be the candidates to relay the data. Thus, relaying selection and forwarding decision need to be made by the current node based on certain routing strategy. Various DTN routing approaches adopt different strategies based on different metrics. Example of such metrics include estimated delivery probability to the destination node, network resources available (including bandwidth, storage, and energy), estimated delay and current network congestion level. However, the capricious mobility and restricted resource in DTNs significantly obstruct us from designing an ideal forwarding mechanism.

1.1 Social Characteristics of DTNs

Recently, the consideration of social characteristics provides a new angle for the design of DTN routing protocols. In most of the DTN applications (e.g. vehicular networks [77, 94], mobile social networks [4, 12, 14, 17, 64], disease epidemic spread monitoring and pocket switched networks (PSNs) [38]), a multitude of mobile devices are used and carried by people, whose mobility is better described by social characteristics such as the carrier's

social relations and behaviors. This opens new possibilities of successfully applying social network analysis (SNA) methods on DTN routing, in which the knowledge of social characteristics are used to make better forwarding decision. The social characteristics usually long term and less volatile than node mobility, they usually cost less to maintain and more reliable. In DTN environment, using them to make forwarding decision may significantly reduce the control overhead and improve the routing performance. Common social characteristics includes: community [34, 61, 68, 72, 74], centrality [23, 24, 59, 69], similarity [15, 53, 63] and friendship [7, 62, 96]. More details refer to Chapter 2.

1.2 Social Network Analysis

Social network analysis (SNA)[79, 89], which studies relationships among social entities, the patterns and implications of these relationships, has been proved a powerful tool in many research areas such as anthropology, biology, communication studies, economics, information science, computer science and engineering. The study of information propagation in social networks, which is very relevant to data dissemination and data forwarding in communication networks (such as routing packages in delay tolerant networks), is one of the key topic in SNA. Information propagation has been used in epidemiology to help understand how patterns of human contact aid or inhibit the spread of diseases such as HIV in a population. There are many success examples on information propagation in general(online) social networks.

Lind *et al.* [75] studied a simple model of information propagation in social networks, by introducing the concepts of spread factor (the average maximal fraction of neighbors of a given node that interchange information among each other) and spreading time (i.e. the time needed for the information to reach such a fraction of nodes). They applied this model to real empirical networks and compared spreading dynamics with different types of networks. They found that the number of neighboring connections strongly influences the probability of being gossiped. Yildiz *et al.* [91] considered the problem of asymmetric information diffusion with gossiping protocols in both static and dynamic networks. They

derived conditions under which the network converges to the desired result within limit, and provided policies that offers a trade-off between accuracy and increased mixing speed for the dynamic asymmetric diffusion problem. In [84, 85], Tang *et al.* proposed new temporal distance metrics to quantify and compare the speed of information diffusion with the consideration of the evolution of a network from a local and global view. Lee *et al.* [52] proposed a method to find influentials by considering link structure and the temporal order of information adoption in Twitter. In [98], Zhao *et al.* used communication motifs and maximum-flow communication motifs as the tools to characterize the patterns of information propagation in two real-life social networks (networks from cellular call record and Facebook wall-post history). They concluded that the patterns of information propagation within both social networks are stable overtime, but these patterns are different and sensitive to the cost of communication in synchronous and asynchronous social networks. The speed and the amount of information propagated through a network are correlated and dependent on individual profiles. In [1], Bakshy *et al.* studied the content propagation via user-to-user content transfer history in a time-evolving social network (Second Life). They found that the social network plays a significant role in the propagation of content. Additionally, adoption rate increases as the number of adopting friends increases, but this effect varies with the connectivity of a particular user. They also found that sharing among friends occurs faster than sharing among strangers and some individuals play a more active role in distributing content than others. Kuhlman *et al.* [50] studied the problem of inhibiting diffusion of complex contagions such as rumors, undesirable fads and mob behavior in social networks by removing a small number of critical nodes from the network. They showed that finding minimum number of such nodes is NP-hard, and proposed efficient heuristics for such tasks.

All of the studies above confirm that social structures and properties indeed strongly influence information propagation. These observations inspire the development of social-aware routing protocols for different communication networks as well. Based on the ob-

servation of DTN's social characteristics and taking the recent advances in social network analysis, several social-based DTN routing methods [7, 15, 27, 37, 40, 96] have been proposed recently to exploit various social characteristics in DTNs (such as community and centrality) to assist the relay selections. We will introduce details on exiting social-based DTN routing method in later chapters.

1.3 Location Characteristics of DTNs

Although social characteristics are already proved effective in DTN routing, in real world for many reasons (privacy), we cannot get information also may not reflect the truly device communication opportunities. For example, a mother and a daughter live in two cities. Their mobile device seldom have chance to exchange data directly. On the other hand, the appearance of mobile device equip with sensors (especially GPS) and contact/event logs enables pervasive monitoring of mobile user/mobile device behaviors and mobility. There are several cellular datasets recently collected via smartphone based testbeds: Nokia Data Collection Campaign [11], MIT reality project [51], Nodobo [71], and Context project [70]. These real-life tracing data provide abundant resources to study social, spatial, and temporal characteristics of mobile users in different environments. An individuals location history in the real world implies, to some extent, his/her interests and behaviors. People who share similar location histories are likely to have common interests, behaviors and some kind of relations. Thus, it is possible to analyze the enriched location information and extract location/social characteristics among users. The social characteristics, which extracted from location information, will more accurately represent the physic contact opportunities among users. By seeking such kind of location and social characteristics, DTN routing protocols are expected to have better performance.

1.4 Content and Organization

This thesis focus on study of new routing protocols for delay tolerant networks formed by mobile users, which utilize location and/or social features to achieve efficient and stable routing. We first present a location-based DTN routing protocol, which uses the new metric

geo-similarity to make the routing decision. Some simplified location-based DTN routing protocols are then proposed. After showing the effectiveness of location-based methods, We explore methods to predict a location's semantic meaning (the location's social feature) using Nokia Data Collection Campaign Dataset [11]. Then propose several location-social based routing protocols, which incorporate the location characteristics into the social-based DTN routing methods. At last, We discuss the possibility and methods to further improve routing performance by adding limited amount of throw-boxes into the networks to aid the DTN relay. Several throwboxes based routing protocols and location selection methods for throw-boxes have been proposed.

The rest of this thesis is organized as follows. We first introduce some social analysis methods, social properties related to social-based DTN routing and available real word tracing data in Chapter 2. Then, in Chapter 3, We review and analysis the current social-based routing protocols in DTNs. In Chapter 4, We propose the location-social based routing methods which use location and/or social features to achieve efficient and stable routing, and report a study on Nokia Data Collection Campaign Dataset [11]. In Chapter 5, We propose several throw-boxes based routing protocols and location selection methods for throw-boxes based DTNs. We summarize the thesis and discuss some possible future works in Chapter 6.

CHAPTER 2: PRELIMINARIES

In this chapter, we will introduce some social analysis methods and social properties related to social-based DTN routing. And introduce the real-world wireless tracing data we use in our study.

2.1 SNA Methods and Social Properties

In this section, we will introduce some social analysis methods and social properties related to social-based DTN routing. Many of these social properties have been studied in social network analysis.

2.1.1 Social Graph and Contact Graph

The most popular way, to study the social relations among people and extract their social properties, is building a *social graph* (also called social network). A social graph is a global mapping of everybody and how they are related. Such a graph is an abstract graph where vertices represent individual people and edges describe social ties between individual people. Social ties can be expressed in many forms. For example, different types of social ties may describe different social relationships among people such as friends, family members, and co-workers. Social graphs have been widely used in many applications, such as analysis of online social networks [65] or terrorist networks [49]. With a social graph, a variety of social metrics (e.g., communality, centrality, and similarity) can be easily calculated or estimated, and these metrics can be then used by social-based approaches. Therefore, it is crucial to obtain social graphs for social-based approaches.

A social graph is an intuitive source for many social metrics such as community and friendship. Unfortunately it is not always available (due to either privacy or security reasons) or hard to be obtained via disclosed social data. However, with new networking

technology, we can study relationships among people by observing their interactions and interests over wireless networks. Building a *contact graph* is a common way to study the interactions among people in a network and thus analyze their relationships and estimate the social metrics among them. In DTNs, each possible packet forwarding happens when two mobile nodes are in contact (i.e., within transmission range of each other). By recording contacts seen in the past, a contact graph can be generated where each vertex denotes a mobile node (device or person who carries the device) and each edge represents one or more past meetings between two nodes. An edge in this contact graph conveys the information that two nodes encountered each other in the past. Thus the existence of an edge intends to have predictive capacity for future contacts. A contact graph can be constructed separately for each single time slot in the past, or it can be constructed to record the encounters in a specific period of time by assigning a set of parameters to each edge to record the time, the frequency and the duration of these encounters. From the observation that people with close relationships such as friends, family members, etc. tend to meet more often, more regular and with longer duration, we can extract DTN nodes' relationships from the recorded contact graph, estimate their social metrics, and use such information to choose relays with higher probabilities of successful forwarding.

How to detect people's relationships and create the relative social graph from the recorded contact graph may affect estimation accuracy and the efficiency of social-based approaches. Most of the current social-based DTN routing algorithms [15, 38, 40] directly treat the aggregated contact graph (merging the contact graphs of several time slots into one graph) as the social graph of all entities in the network, and uses this graph to generate social metrics for forwarding selection. This strategy is based on the observation that although the contact graph reflects the encounter history while the social graph reflects the social relations among people, the aggregated contact graph (the sum of contact graph over time) and the social graph are statistically similar. However, Hossmann *et al.* [35, 36] showed that the performance of these algorithms heavily depends on the way the graph is constructed out

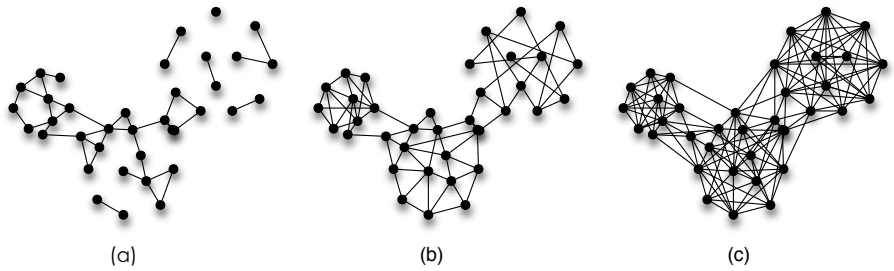
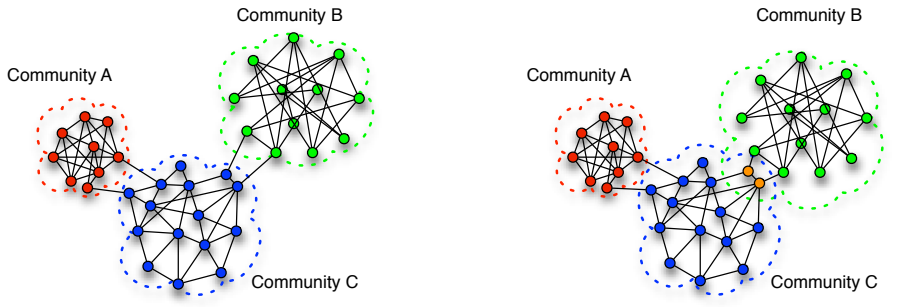


Figure 2.1: Illustration of three different contact graphs using different aggregation periods.

of observed contacts (i.e., contact aggregation) and proposed a method to select an appropriate aggregation period for contact aggregation. For example, in Figure2.1, three different contact graphs are generated from the observed contact history. Figure2.1(a) shows a disconnected contact graph which maybe too sparse to detect any useful social structure. While Figure2.1(c) shows an almost complete graph which is too dense and thus useless. Figure2.1(b) is an appropriate aggregated contact graph. After building the aggregated contact graph, different social metrics can be obtained. For example, Hui, *et al.*[13, 39, 41, 92] proposed several community detection approaches (simple, k-clique, modularity, etc.) with great potential to detect both static and temporal communities. Bulut *et al.* [7] introduced a method of detecting the quality of friendship by calculating the social pressure metric (SPM) from contact graphs.

2.2 Community

Community is an important concept in ecology and sociology [34, 61, 74]. In ecology, a community is an assemblage of two or more populations of different species occupying the same geographical area. In sociology, community is usually defined as a group of interacting people living in a common location. Community ecologists and sociologists study the interactions between species/people in communities at many spatial and temporal scales [34, 61, 68, 72, 74]. It has been shown that a member of a given community is more likely to interact with another member of the same community than with a randomly chosen member of the population [72]. Therefore, communities naturally reflect social relationship among people. Figure2.2 illustrate examples of community structures in Social graphs.



(a) Community without overlapping (b) Community B & C are overlapped

Figure 2.2: Illustration of community structures in contact/social graphs

Since wireless devices are usually carried by people, it is natural to extend the concept of social community into DTNs to explore interactions among wireless devices. It is believed that devices within the same community have higher chances to encounter each other. Therefore, the knowledge of community structures could help a routing protocol to choose better forwarding relays for particular destinations, and hence improve the chance of delivery. Many proposed community detection algorithms [13, 39, 41, 92] are available for identifying social communities from the contact graph of DTNs.

Some of the most common detection methods for communities are summarized as follows.

Minimum-cut method [66]: The minimum-cut method divides the social graph or the network into a predetermined number of components such that the number of edges between components is minimized.

Hierarchical clustering: The hierarchical clustering method uses the concept of similarity to detect communities. The similarity metric aims to measure the degree of similarity (usually topological) between node pairs. Common calculation methods include the cosine similarity, the Jaccard similarity coefficient, and the Hamming distance between rows of the adjacency matrix of the network. A common community detection strategy is: all nodes within a community have similarity greater than a given threshold.

Girvan-Newman algorithm [30]: Girvan and Newman proposed a community detection

method using a graph-theoretic metric, betweenness (we will introduce later), to identify the bridge edges among communities in the network. By removing these bridge edges, the communities can be easily detected.

Modularity maximization [16, 67, 68]: The modularity maximization method detects communities by searching over all possible network divisions to find the one with particularly high modularity. Modularity is defined as a benefit function, which measures the quality of a particular division. Optimization methods (such as greedy algorithms or simulated annealing) are often used because exhaustive search over all possible division is usually too expensive.

The Louvain method [86]: The Louvain method is a greedy optimization method with two phases. It first looks for "small" communities by locally optimizing modularity, it then aggregates nodes in the same community and builds a new network whose nodes are the communities. These two steps are repeated iteratively until a maximum modularity is achieved.

Clique based methods [21]: In a graph, a clique is a subgraph in which every node is connected to every other node in the subgraph. Since a clique is the most tightly connected structure, there are many community detection approaches based on the detection of cliques in a graph. As a node can belong to multiple cliques, these methods may lead to overlapping community structures.

2.2.1 Centrality

In graph theory and network analysis, *centrality* is a quantitative measure of the topological importance of a vertex within the graph. A central node, typically, has a stronger capability of connecting other nodes in the graph. In a social graph, the centrality of a node describes the social importance of its represented person in the social network. In DTNs, the sociological centrality metrics [59] can also be used for relay selections (nodes with high centralities are always good candidates of relay nodes).

There are several ways to define centrality in a graph. Three common centrality mea-

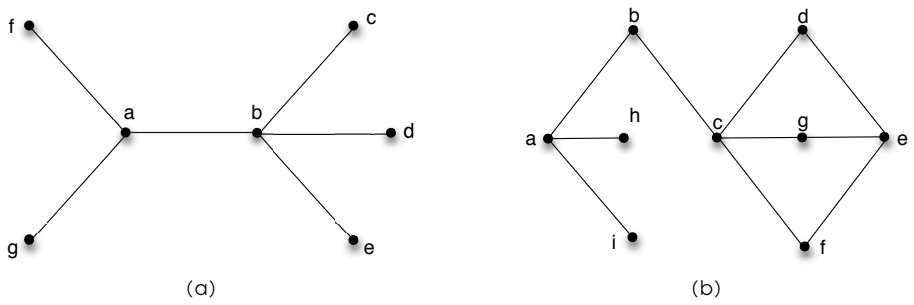


Figure 2.3: Illustration of centrality and similarity measurements over simple social graphs

sures are *degree centrality*, *betweenness centrality*, and *closeness centrality* [23, 24, 69]. Degree centrality is the simplest centrality measure which is defined as the number of links (i.e., direct contacts) incident upon a given node. For example, in Figure 2.3(a), the degree centrality of node a and node b are 3 and 4 respectively while those of the other nodes are 1. A node with a high degree centrality is a popular node with a large number of possible contacts, and thus it is a good candidate of a message forwarder for others (i.e., a hub for information exchange among its neighborhood). Betweenness centrality measures the number of shortest paths passing via certain given node. For example, the betweenness centrality of node a and b in Figure 2.3(a) are 18 and 24, respectively. But for the other nodes, their betweenness centralities are 0 since they are not on any shortest paths. Nodes that occur on many shortest paths between other nodes have higher betweenness than those that do not. A node with high betweenness centrality can control or facilitate many connections between other nodes, thus it is ideal for a bridge node during message exchange. The closeness centrality of a node is defined as a the inverse of its average shortest distance to all other nodes in the graph. If a node is near to the center of the graph, it has higher closeness centrality and is good for quickly spreading messages over the network. For the example in Figure 2.3(a), the closeness centrality of node a is $\frac{2}{3}$ since its average shortest distance to all others is 1.5. The closeness centralities of b, c/d/e, and f/g and $\frac{3}{4}$,$\frac{6}{13}$,and $\frac{3}{7}$, respectively.

2.2.2 Similarity

Similarity [15] is a measurement of the degree of separation. It can be measured by the number of common neighbors between individuals in social networks. For example, in Figure 2.3(b), the similarity betwwen a and c is 1, while that between c and e is 3. Sociologists have long known that there is a higher probability of two people being acquainted if they have one or more other acquaintances in common. In a network, the probability of two nodes being connected by a link is higher when they have a common neighbor. When the neighbors of nodes are unlikely to be in contact with each other, diffusion can be expected to take longer than when the similarity is high (with more common neighbors). In addition, there are other ways to define the similarity beyond common neighbors, such as similarity on user interests [63] and similarity on user locations [53].

2.2.3 Friendship

Friendship is another concept in sociology which describes close personal relationships. In DTNs, friendship can be defined between a pair of nodes. On the one hand, to be considered as friends of each other, two nodes need to have long-lasting and regular contacts. On the other hand, friends usually share more common interests as in real world. In sociology, it has been shown that individuals often befriend others who have similar interests, perform similar actions and frequently meet with each other [62]. This observation is called *homophily phenomenon*. Therefore, the friendship in DTNs can be roughly determined by using either contact history between two nodes [7] or common interests/contents claimed by two nodes [96].

2.3 Real-World Wireless Tracing Data

To understand the social, spatial and temporal dynamics in DTNs is an essential step of protocol design for DTN routing. The theoretical methods do not always lead to accurate observation due to simplified assumptions made for the ease of analysis. Also, none of the theoretical methods can imitate the DTNs environment exactly the same as the realistic one. So, it is more convincing to study social from real word tracing.

Fortunately, with the advance of new wireless devices (smart phones), and social-media websites, there are tremendous amounts of enriched public real-life wireless tracing data available, which provide a possibility of study the social relationships among the participants. For example, the Community Resource for Archiving Wireless Data At Dartmouth (CRAWDAD) [90], a wireless network data resource for the research community, archives different wireless trace data (such as contact tracing from a student social network [83] or a campus bus-based DTN [9] or a Bluetooth network among conference attendees [80] and mobility tracing of taxi cabs in San Francisco [76]). In addition, there are several public datasets recently collected via smartphone based testbeds: Nokia Data Collection Campaign [11], MIT reality project [51], Nodobo [71], and Context project [70], which include multiple social relationships among their users (such as call record, Bluetooth contact, WiFi access, GPS log, and social network information) These real-life tracing data provide abundant resources to study social, spatial, and temporal characteristics of different DTNs.

In this thesis we mainly study three datasets: the MIT reality project data [51], the Nokia Data Collection Campaign data [11] and the D4D datasets [5]. The MIT reality data is a small-scaled dense campus dataset with rich variety of data source, which is very suitable for early stage experiment. It has been used in experiment of a lot of research works. The drawback of this dataset is the social roles of users in this dataset are too flat, they or either students or professors. So do their relationships. Compare with MIT dataset the users in Nokia Campaign data have more colorful social roles and relationships since they are from a citywide region. And its location access information is collected by GPS, which is more accurate than cell tower logs. The drawback of this dataset is it does not have mapping of its users ID and relative Phone Number and Mac Address. So even we have both call logs and Bluetooth trace, we cannot use it for routing algorithm evaluation. We only use it on Nokia Data Collection Campaign to make semantic place prediction. The D4D Challenge Dataset is a rare large-scaled dataset. Although it does not have Bluetooth tracing data, it is the most coincident with our target DTN environment, which is sparse network with rich

Table 2.1: Characteristics of the datasets

Date set	# of users	region	duration	call logs	bluetooth	tower logs	gps
MIT	100	campus	1 year	yes	yes	yes	no
Nokia	200	Lake Geneva	1 year	yes	yes	no	yes
D4D	46,254	Ivory Coast	150days	no	no	yes	no

social relationships between users but without straightforward social relation information. So in this thesis, we use the MIT reality data on our early stage experiments and D4D Challenge Dataset for all stage experiments. Table 2.1 summarizes the characteristics of these three dataset.

2.3.1 MIT Reality Mining Dataset

The MIT Reality Mining Dataset consists of one hundred Nokia 6600 smart phones which pre-installed with several pieces of software. Seventy-five users are either students or faculty in the MIT Media Laboratory, while the remaining twenty-five are incoming students at the MIT Sloan business school adjacent to the laboratory. Of the seventy-five users at the lab, twenty are incoming masters students and five are incoming MIT freshman. The information they are collecting includes call logs, Bluetooth devices in proximity, cell tower IDs, application usage, and phone status (such as charging and idle), which comes primarily from the Context application. The dataset was collected over nine months period on users' location, communication and device usage behavior.

2.3.2 Nokia Data Collection Campaign Dataset

Nokia Research Center Lausanne and its academic partners (EPFL and Idiap) have recently completed a data collection campaign ($http$://research.nokia.com/page/11367) in the Lake Geneva region. Data from smartphones of almost 200 participants were collected in the course of more than one year. This dataset provides a comprehensive and relatively unexplored data set including rich social and geographic informations. Detailed content of the dataset includes:

- Phone usage (complete call message log, application start and close events, audio being played);

Table 2.2: Numbers of users, towers, and contacts in four different settings in D4D

Setting	# of users	# of towers	# of encounters
A) subset users within full region	13,436	1,095	617,136
B) subset users within limited region	6,318	496	327,717
C) all users within full region	46,254	1097	6,787,594
D) all users within limited region	21,768	497	3,736,173

- Personal data (list of pictures and videos taken with the camera, full contact list, full calendar content and update events);
- Environmental data (accelerometer samples, list of available WiFi access points, list of visible Bluetooth devices);
- Phone status data (current attached GSM cell, GPS readings, battery level, alert mode, other system status);
- Information provided by the participants through questionnaires(sex, age, occupation,*et al.*).

We will introduce more details in Chapter 4.

2.3.3 D4D Challenge Dataset

(a) traffic distribution (b) limited region (c) tower distribution

Figure 2.4: Illustration of the limited region in Settings B and D of D4D.

The released D4D datasets [5] are based on anonymized Call Detail Records (CDR) of phone calls and SMS exchanges between 50,000 Orange mobile users in Ivory Coast between December 1, 2011 and April 28, 2012 (150 days). Among the released four datasets,

we mainly use the second one (**SET2**): individual trajectories with high spatial resolution. This dataset contains the access records of antenna (cellular tower) of each mobile user over two-week periods. Such information provides high resolution trajectories for all mobile users. We will use the sequences of visited cellular towers of all users to generate contact encounters among mobile users and location profiles of each mobile user. In the results present in this thesis we only use the first two weeks (December 1 to 14, 2011) data for our simulations.

Since D4D datasets do not have direct encounter information between phones via short range communications (such as Bluetooth or WiFi), to support opportunistic communications we assume that two phones can direct communicate to each other if they share the same cellular tower at particular time. Though this assumption may not be true in reality, it gives us an approximated environment for opportunistic communications. All of our experiments are based on the generated encounter databases from SET2.

We will consider four different settings (A-D) for our experiments. Table 2.2 summarizes some statistics of these settings. In term of number of nodes (mobile users), we either use all 50,000 users or a subset of users (around 15,000) in the original SET2. When we pick up the subset of users, we just simply choose the first 15,000 users in our encounter database. Notice that the number of users in our generated encounter database is less than the number of users in original SET2 (such as $46,254 < 50,000$). This shows that there are many mobile users who do not share any cellular towers with other users. The smaller size of user set could accelerate the execution time of our simulations. Notice that the number of encounters is significantly reduced after picking the subset users, though the cellular towers stay the same level. We also have settings where we limit the physical locations of encounters to a small region. As shown in Figure 2.4(a), the traffic load distribution within Ivory Coast is unbalanced. This figure shows the number of calls (both incoming and outgoing calls) during the first two weeks. Darker color indicates heavier traffic loads. Therefore, when picking up the small region, we choose the region with the heavies traffic

load. The longitude and latitude ranges of the region (shown as a tiny blue rectangle in Figure 2.4(b) around Abidjan) are $[-8.49, -2.69]$ and $[4.41, 10.47]$, respectively. Abidjan is the economic and former official capital of Ivory Coast and the largest city in the nation. From Table 2.2 we can see that this region holds a large number of cellular towers and mobile users. Figure 2.4(c) shows the detailed tower distribution in this region.

CHAPTER 3: EXISTING SOCIAL-BASED ROUTING

In this chapter, we review several social-based DTN routing methods that take advantage of positive social characteristics in DTN networks.

3.1 Label Routing

Hui and Crowcroft [37] introduced a routing method (called as label routing hereafter) based on community labels in Pocket Switched Networks (PSNs). A PSN [38] is a type of DTN where mobile devices are carried by people and communicate with each other when people meet. To reduce the amount of traffic created by forwarding messages in PSNs, the proposed routing method uses a labeling strategy to select forwarding relay. Since people in the same community are likely to meet regularly, they are appropriate forwarders for messages destined to the members of their community. In their solution, Hui and Crowcroft assumed that each node has a small label telling others about its affiliation/group (i.e., its social community), just like name badges used in a conference. Based on the labels, label routing chooses to forward messages to destinations directly or to next-hop nodes which belong to the same group (label) with the destinations.

Label routing takes the advantage of the knowledge of social community. It assumes that people from the same community tends to meet more often than people from different communities and hence can be good forwarders to relay messages destined to the other members in the same community (with the same label). Label routing requires very little information about each individual (only its group/affiliation). This is easy to implement in PSN applications, by tapping a mobile device and writing down the affiliation of the owner. In other words, the community (or group) information relies on user inputs in label routing. However, user-defined communities may not always reflect the position/contact

relationship among nodes. For example, two DTN nodes in the same community may be physically far away and could never meet with each other. In this scenario, using one node to be the forwarder for the other may not be a good choice. In addition, in label routing, the message forwarding from the source to the destination is purely via the members within the same community of the destination. This may significantly increase the delay or even fail to deliver the message. For instance, message delivery will fail when the source does not meet any member from the destination's community, even though there are possible relay nodes from other communities.

3.2 SimBet Routing

Daly and Haahr [15] proposed a social-based routing protocol (called SimBet routing hereafter) which uses *betweenness centrality* and *similarity* metrics to identify some "bridge" nodes (with high values of these metrics) in networks. To avoid exchanging information of the entire network topology, they only estimated the betweenness centrality Bet_n for each node n in its local neighborhood. For similarity metric, they considered the similarity $Sim_n(d)$, the number of common neighbors, of the current node n with the destination node d. Both of the social metrics are maintained and updated dynamically in DTNs. Therefore, the proposed SimBet routing makes forwarding decision by considering not only the pre-estimated betweenness centrality metric but also the locally determined social similarity. Nodes with high betweenness centralities are those nodes who can act as bridges in their neighborhood, while nodes with high similarities with the destination are more likely to find a common neighbor with the destination which can act as the forwarder.

In SimBet routing, when a DTN node n meets another DTN node m and holds a message with destination d, n calculates its relative betweenness utility and similarity utility to node m:

$$\begin{aligned} SimUtil_n &= \frac{Sim_n(d)}{Sim_n(d) + Sim_m(d)} \\ BetUtil_n &= \frac{Bet_n}{Bet_n + Bet_m} \end{aligned}$$

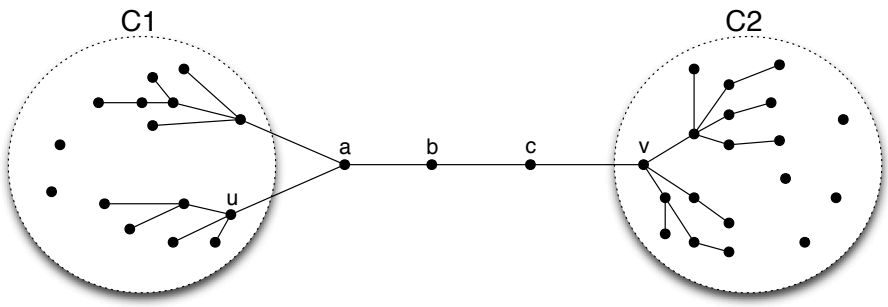


Figure 3.1: Illustration of problems of the SimBet routing.

Then node n can compute its SimBet utility, which is a weighted combination of betweenness utility and similarity utility:

$$SimBetUtil_n(d) = \alpha SimUtil_n(d) + (1-\alpha)BetUtil_n.$$

Here, α is an tunable parameter which can adjust the relative importance of the two utilities. For the message with d as its destination, if $SimBetUtil_m(d) > SimBetUtil_n(d)$, node n forwards the message to node m. Otherwise, it continues to hold the message. Via possible multi-hop relays, the message may eventually reach d.

In summary, SimBet routing uses two social metrics (centrality and similarity) to estimate or predict the probability that potential relay nodes may meet the destination. It is obvious that both metrics are effective at identifying suitable relays in different scenarios respectively. Take an example graph, as shown in Figure 3.1, where a few low-degree bridges (i.e., a, b and c) connect two well-connected components C_1 and C_2. Assume that node u wants to send a message to node v. When node u encounters node a, it compares its SimBet utility with that of node a's. Both u and a have zero similarity to v, but u's global betweenness centrality is less than a's since a sits on more of the shortest paths. Thus, u will transfer the message to a based on SimBet routing. In this case, centrality metric helps to pick the better relay node. On the contrary, if node a wants to send a message to v and it encounters node b, similarity metric will play a role since the global betweenness centralities of a and b are the same. Therefore, a has a smaller similarity (zero common neighbor)

to v than b has (one common neighbor with v). Therefore, combining multiple social metrics may make the social-based protocol more effective in broad situations. However, due to the uncertainty of future encounters and underlying social graph, it is still possible that the node with high SimBet utility fails to delivery the message to the destination.

To avoid global information exchanges, SimBet routing provides a distributed method to calculate social metrics locally, which is desirable in a DTN environment. However, estimating centrality based solely on local information may lead to inaccurate "bridge" identification. For instance, in the example shown in Figure 3.1, it is assumed that u wants to send a message to v. When u encounters node a, based on the two-hop information, u's local betweenness Bet_u is much larger than a's Bet_a. Since both u and a have zero similarities to v, the overall $SimBetUtil_u(v) > SimBetUtil_a(v)$. Then, node u will not pass this message to node a, and thus miss the opportunity to delivery the message. Nonetheless, considering global betweenness, each of the nodes of a, b and c has highest betweenness in the entire network (since they form the only path connecting components C_1 and C_2), and can then be correctly identified. A possible way to increase the chance of correct "bridge" identification is using larger neighborhood information, although this may increase communication cost. Similarly, to increase the chance of delivery, multiple relay nodes could be used. The trade-off is always between delivery performance and communication cost.

3.3 Bubble Rap Forwarding

The forwarding strategy, *Bubble Rap Forwarding*, proposed by Hui *et al.* [40] also relied on two social characteristics (community and centrality). They assumed that each node belongs to at least one community and its node centrality (either betweenness or degree centrality) in the community describes the popularity of the node within this community. Each node has a global centrality across the whole network (or called global community), and a local centrality within its local community. A node may also belong to multiple communities and hence have multiple local centralities. Taking advantages of these social

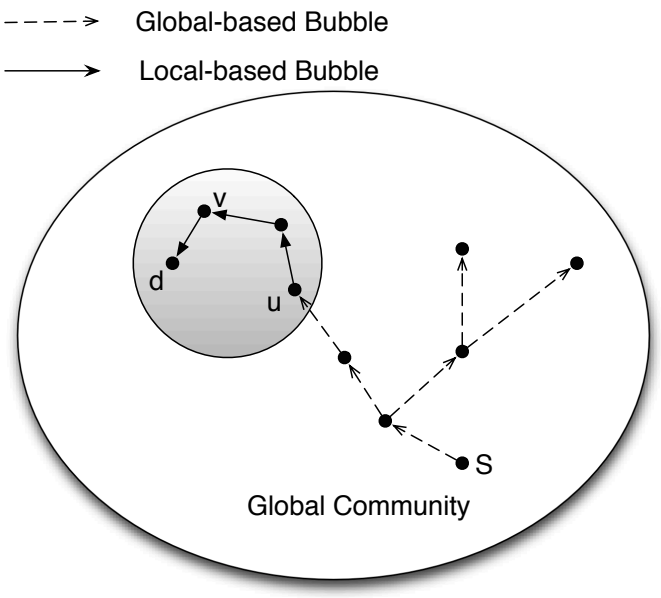


Figure 3.2: An illustration of the Bubble Rap forwarding from source s to destination d

characteristics, Bubble Rap Forwarding basically includes two phases: a bubble-up phase based on global centrality and a bubble-up phase based on local centrality. In both phases, the bubble-up forwarding strategy is utilized to forward messages to nodes which are more popular than the current node (i.e., with higher centrality). When a node s has a message with destination of d, it first bubbles the message up based on the global centrality, until the message reaches a node which is in the same local community C_d as the destination d. This procedure is shown as blue arrows in Figure 3.2. : The blue and red arrows show the bubble-up operations based on global centrality in global community and local centrality in d's community C_d, respectively. After the message reaches d's community at node u, Bubble Rap Forwarding switches to the second phase which uses members of C_d as relays. This forwarding strategy continues to bubble up the message through the local community based on local centrality until the destination is reached. This later procedure is shown as red arrows in Figure 3.2. In order to reduce cost, it is also required that whenever a message is delivered to the community, the original carrier delete this message from its buffer to prevent further dissemination.

Bubble Rap Forwarding uses the concept of community in addition to node centrality to help with the forwarding decision. The introduction of local centrality inside a community is more beneficial than local centrality around local neighborhood (i.e., k-hop) [15]. The bubble-up operations allow fast transfer of a message towards the destination or its community. However, such a strategy may fail when the destination belongs only to the communities whose members are all with low global centrality values. In this case, the bubble-up process in the first phase of Bubble Rap Forwarding cannot find the relay node which is in the same local community as the destination node. A possible solution for this problem to have a timeout timer for bubble-up process and exchange to other backup strategy for data delivery after timeout. In [40], the authors used a flat community (not hierarchical) to demonstrate the efficiency of Bubble Rap Forwarding. However, they did not provide details about how to handle hierarchical communities where the destination d may belong to multiple overlapping communities. In that scenario, they may face problems in the second phase of Bubble Rap Forwarding. For example, if the current encountering node u shares multiple communities with d, a problem arises regarding which one of d's local communities should be chosen to bubble-up. A simple solution to this problem is picking the local community with which d have highest centrality. This solution also matches the spirit of Bubble Rap Forwarding which keeps looking for nodes with high centralities.

3.4 Homophily Based Data Diffusion

Zhang *et al.* [96] proposed a data diffusion scheme based on the "homophily" phenomenon in social networks. Here, data diffusion aims to deliver data to all nodes in DTNs. In DTNs, data may not be completely delivered from one node to another during a contact between them, since the contact time is too short to transmit the data or the buffer available at the receiving node is insufficient to hold the data. Therefore, in the design of data diffusion protocol, not only the contact probability between nodes but also the data propagation orders (which data should be propagated first) affect the diffusion speed and data access delay.

To choose an appropriate relay node to diffuse and an appropriate data item to buffer, Zhang *et al.* introduced a method using the *friendship* among nodes and the "homophily" phenomenon. The "homophily" phenomenon describes the trend in real word that friends usually share more common interests than strangers. By applying the same idea from "homophily" phenomenon, their proposed data diffusion strategy diffuses the most similar data items between friends, and diffuses the most different data items between strangers. If a node meets a new contact who is a friend, it first diffuses the most similar data items of their common interests to its friend first until the contact time is over. If the new contact is a stranger, it starts from the data item most different from their common interest. By theoretical analysis, Zhang *et al.* showed that this data diffusion scheme achieves better diffusion speed and data access delay than the other three possible schemes (including diffusing the most similar data to any encounter, diffusing the most different data to any encounter, and diffusing the most different data between friends and the most similar data between strangers).

This proposed method provides a new angle to social-based approaches. It considers the need of managing data propagation orders which is an important aspect of design issues in DTN routing. With the same amount of communication opportunity and duration, more useful information can be transmitted under this proposed method. In addition, the proposed method is not conflicted with other DTN routing protocols. It can be used together with other DTN routing protocols to make better relay decisions with efficient data propagation orders. In the proposed method, social friendship is the only metric used to predict the encounter's needs of information and friendship is defined by users. However, user defined friendship is not always available in DTNs. Therefore, it is another challenging direction need to be further explored regarding how to efficiently detect the friendship in dynamic DTNs .

3.5 Friendship Based Routing

Bulut *et al.* [7] also used friendship to aid the delivery of packets in DTNs. They introduced a new metric, social pressures metric (SPM), to accurately detect the quality of friendship. Different from [96], where friendship is defined by users based on their social relationships, this approach considered friends as nodes which contact to each other frequently and have long-lasting and regular contacts. Therefore, the social pressures metric between nodes i and j can be estimated from the encounter histories of these nodes (recorded by the nodes) as: $SPM_{i,j} = \frac{\int_{t=0}^{T} f(t)dt}{T}$, where $f(t)$ denotes the remaining time to the first encounter of these nodes after time t and T is the total time period. SPM describes the average forwarding delay if node i has a message destined to j at each time unit. Then, the link quality $w_{i,j}$ between each pair of nodes, (i, j), is defined as $w_{i,j} = \frac{1}{SPM_{i,j}}$. The authors assumed that the bigger value of $w_{i,j}$ represents the closer friendship between i and j. Using the value of $w_{i,j}$, each node can construct its friendship community for each period T as a set of nodes whose link quality with itself is larger than a threshold. When a node i, having a message destined to d, meets with node j, it forwards the message to j if and only if (1) j and d are in the same friendship community (in the current period) and (2) j is a stronger friend of d than i.

In summary, this friendship based routing method uses the node contact information in each period to calculate the friendship metric (i.e., SPM), and constructs the friendship community. These social metrics can indeed help with making smarter forwarding decisions. However, the calculation of these metrics needs the whole contact information during each period, which may not be realistic in most DTNs. To obtain $f(t)$ in the current period, node i needs to know the time of its first encounter to node j after time t in this period, which is an event in future. Therefore, either the values in contact history from previous periods are used for this calculation at the current period or the estimated future contacts in this period are available for this calculation. This is clearly a drawback of this proposed method. In addition, this friendship based routing uses a similar forwarding scheme to la-

bel routing [37], which may lead to the same problem. If the source node fails to meet with any node in the same friendship community with the destination node, the delivery fails. Therefore, more felicitous forwarding strategies should be studied for this friendship based routing.

Although the friendship based method [7] and homophily based method [96] both use friendship metrics for delivery data in DTNs, they are designed for different purposes. In [96] the friendship measurement is used to select which data items to diffuse, while in [7] the friendship metric is used to detect communities and select which relay nodes to forward. Therefore, different social metrics or various calculation methods need to be designed for specific design purposes. There is no universal solution for all applications.

3.6 Other Social-Based Routings

Besides the social-based DTN routing strategies reviewed above, there are also a few recent social-based approaches which define their own social-related metrics to improve either the scalability or accuracy of routing. We briefly review them in this section.

In [63] Mei *et al.* took advantage of the observation, that people with similar interests tend to meet more often, to propose a social-aware and stateless routing (SANE) for pocket switched networks. This routing strategy represents the interest profile of an individual u as an k-dimensional vector I_u. To express the interest similarity between two individuals u and v, the *cosine similarity* is defined as, $\Theta(I_u, I_v) = \cos(\angle I_u I_v) = \frac{I_u I_v}{\|I_u\|\|I_v\|}$. In SANE, a message should be forwarded to individuals whose interest profiles closely resemble that of the destination. They assume that the interest profile of a message m is the interest profile of its destination. Thus, a message m will be relayed to a node u only if the *cosine similarity* of the interest profile between message m and node u is higher than a given threshold ρ. One of the advantages of this method is that each node only needs to maintain the interest profile without extra storage. The cost of maintaining and updating this social metric is also relevantly easy. These advantages improve the scalability of this routing method.

Gao and Cao [26] proposed a user-centric data dissemination approach which consid-

ers both social centrality and user interests simultaneously. Different from the concept of centrality used in [15, 27, 40], this approach creates its own concept of centrality, which indicates the expected number of interesters (nodes interested in the data item held by i) that node i can encounter during the remaining time $T_k - t$ of data dissemination. Here, T_k is the time constraint of the data item and t is the current time. Then their relay selection makes sure that a new relay always has better capability of disseminating data to interesters than the existing relays based on this newly defined time-varying centrality. They consider both local centrality (centrality defined over one-hop neighborhood) and multi-hop centrality (which takes multi-hop opportunistic connection into consideration). With multi-hop centrality, more forwarding chances are considered, and this strategy may thus lead to more accurate estimation of forwarding probability.

In [22], Fabbri and Verdone proposed a sociability-based DTN routing, which is based on the idea that nodes with high degrees of sociability (frequently encountering many different nodes) are good forwarding candidates. They defined the *sociability indicator* metric to evaluate the forwarding ability of a node. This metric quantifies the social behavior of a node by counting its encounters with all the other nodes in the network over a period T and is therefore a time-varying parameter. The routing strategy forwards packets to the most sociable nodes only. It is worth to notice that the strategy also considers both *first hop-based sociability* and *kth hop-based sociability*. For kth hop-based sociability, the highly sociable neighbors are considered during the calculation of a user's sociability.

From these new social-aware approaches, we can see that the design of social-based DTN routing tends to be more sophisticated. It not only directly uses social concepts from social networks but also considers its own reality in a DTN environment.

CHAPTER 4: LOCATION-SOCIAL BASED ROUTING

Most of the social-based routing methods introduced above are implemented by exploring social properties from social or contact graph (e.g. the community in [37, 40], the centrality in [15, 40], the friendship in [7], *et al.*). However in real word for many reasons we cannot get them. For example, in Nokia Data Collection Campaign Dataset, although we can get the list of visible Bluetooth devices (the hashed MAC address of devices), we cannot infer the contacts information between users since the relations between users and their relative device's hashed MAC address is protected.

Even the social and contact relations of users are available, they not necessarily reflect the truly device communication opportunities. For instance, brothers and sisters are socially very close, but they may not have chance to see each other for longtime since they are busy or living in two different countries, etc. On the other side, only look at social relation may also lose many communication opportunites such as communication between familiar strangers (e.g. Stangers take the same bus everyday). So, although pure social-based methods has been proved better performance than traditional opportunity-based routing protocols, their own limitation of accurately represent communication opportunities makes them hard to have further progress.

In order to design more efficient and stable DTN routing algorithm, we need to explore new realistic and effective features, which are easy to get, and simple to maintain. The increasing availability of location technologies (GPS, GSM networks, etc.) enables mobile devices to obtain their locations easily. An individual's location history in the real world implies, to some extent, his/her interests and behaviors. People who share similar location histories are likely to have common interests behaviors and some kind of relations. For

example, family members are living at the same place (home) and colleagues stay in the same office during the day, classmates who take same classes may have the same schedule and members of the table tennis club may show off at the same time in the gym for an athletic event. Thus, it is possible to analyze the enriched location information and extract social features among users, which are stable and easy to maintain. More than this, the social features, which extracted from location information, will more accurately represent the physic contact opportunities among users. Hereafter, we name these kinds of features as location-based social features. By seeking such kind of location-based social features, DTN routing protocols are expected to have better performance.

4.1 Related Works

Location information has been used for communication protocols in different mobile networks. The most notable result is position based routing [6, 48, 60] in mobile ad hoc networks or wireless sensor networks. In position based routing, routing decision is made based on the position information of neighboring nodes and the destination. No routing table is needed at each node, which reduces routing overheads and improves its scalability. The changing topology is reflected as position updates from neighbors, thus routing protocol can handle topology changes without further procedures. However, position-based routing suffer a lot from routing loop or dead ends in mobile networks and DTNs.

Recently, location-aware approaches have been applied in DTNs. There are two main ways to use location information in aid of DTN routing. One is using the current location information to pick the next-hop relay node as the traditional position-based routing. For example, GeoDTN+Nav method [29] combines the GPSR method [47] with DTN routing for a vehicular DTN. Packets are routed to the neighbor who has the smallest distance to the destination. If all neighbors are further away from the destination than the current node is, it switches to perimeter mode and routes the packet based on face routing. If the current node does not have any neighbor at this moment, the protocol switches to DTN mode by carrying the packet. Until a new neighbor node closer to the destination is found,

the routing method switches back to position-based forwarding. Another is performing the mobility prediction via the study of mobility pattern, when historical location information of all nodes is available. Leguay *et al.* [53] built a high-dimensional Euclidean space based on node location patterns. For each node, its coordinates are correspond to its probability of being found in each possible location. By defining different types of distance between two nodes to represent their location similarity, they proposed several location-aware DTN routings. Fan *et al.* [28] explored the geographic regularity of human mobility in the network and employed a semi-Markov analytical model to describe such mobility pattern. By modeling regular users mobility, they further studied how to schedule a superuser to facilitate data delivery. Gao and Cao [25] studied how to characterize the steady-state and transient-state user mobility behaviors at a fine-grained level, based on the Hidden Markov Model (HMM) formulation of user mobility. They showed that their approach is effective in characterizing user mobility pattern and making accurate mobility prediction.

With the availability of enriched location information, location-aware approaches have become an emerging topic in DTNs. However, all existing approaches consider the location information isolated from the social properties. In the following sections, we will explore possible ways to extract social features from location information and use them on DTN routing.

4.2 Location-Based Routing

In this section, we will propose our location-based routing protocol, which mines the similarity between users based on their geographic location history (cell tower ID scan records). Instead of only taking into account of the geographic regions they accessed alone, it also considers relative visited duration and frequency of these regions. By maintain a location profile for each user, when a source/relay node encounters another node, their *geo-similarities* with the destination node could be caculated respectively. The higher *geo-similarities* they have with the destination, the more related they might be with it. With the help of this *geo-similarity* metric, we can easily choose the appropriate relays. We will

first define the *geo-similarity* metric, then we will propose the details of the single-copy location-based routing protocol and extent it into multi-copy scenario.

4.2.1 Geo-Similarity Metric

A user's visit frequency and duration of a place in the past may imply the possibility of this user to visit this place in the future. Therefore, by recording the historical visiting frequency and duration of each location, we can build a *location profile* of each user, which reflects how likely this user will visit a particular place. In this paper, we use the location of each cellular tower as one location. However, our proposed method can work with other definition of locations. Assume that there are n mobile users $(v_1, \cdots, v_n)$ and m cellular towers $(t_1, \cdots, t_m)$. Then we can define the location profile of a user v_i as follows:

Definition 4.1 (Location Profile): The location profile of user v_i is defined as a m-dimensional vector,

$$L(v_i) = \{p(v_i, t_1), \cdots, p(v_i, t_m)\},$$

where $p(v_i, t_j) = \frac{d_{ij}}{\sum_{j=1}^{m} d_{ij}} \cdot \frac{f_{ij}}{\sum_{j=1}^{m} f_{ij}}$. Here d_{ij} and f_{ij} are the total visiting duration and frequency of user v_i to tower t_j, repsectively. Thus, $p(v_i, t_j)$ basically shows the product of the portion of duration/frequency of user v_i to tower t_j compared with the total duration/frequency to all towers. Larger value of $p(v_i, t_j)$ generally indicates v_i visiting t_j more often and staying longer.

Based on location profiles, we can calculate the *geo-similarity* of two users as follows:

Definition 4.2 (Geo-Similarity): The Geo-Similarity of two users v_i and v_j is defined as the inner product of their location profiles, i.e,

$$Sim(i, j) = L(v_i) \cdot L(v_j) = \sum_{x=1}^{m} p(v_i, t_x) p(v_j, t_x).$$

The definition of the *geo-similarity* implies the similarity between location visiting patterns of these two users, which hopefully reflects the probability of their meeting at a cell tower in the future. With the definition of location similarity, the location based routing

```
<cellspan starttime = '2004-09-04 22:51:44' endtime ='2004-09-04 22:52:01' celltower_oid = '8816'/>
<cellspan starttime = '2004-09-04 22:52:01' endtime ='2004-09-04 22:52:26' celltower_oid = '14998'/>
<cellspan starttime = '2004-09-04 22:52:26' endtime ='2004-09-04 22:54:31' celltower_oid = '8816'/>
<cellspan starttime = '2004-09-04 22:54:31' endtime ='2004-09-04 22:59:31' celltower_oid = '8918'/>
<cellspan starttime = '2004-09-04 22:59:31' endtime ='2004-09-04 23:34:37' celltower_oid = '8816'/>
<cellspan starttime = '2004-09-04 23:34:37' endtime ='2004-09-05 03:10:25' celltower_oid = '8918'/>
<cellspan starttime = '2004-09-05 03:10:25' endtime ='2004-09-05 03:10:34' celltower_oid = '8923'/>
<cellspan starttime = '2004-09-05 03:10:34' endtime ='2004-09-05 10:40:08' celltower_oid = '8918'/>
<cellspan starttime = '2004-09-05 10:40:08' endtime ='2004-09-05 10:40:25' celltower_oid = '8816'/>
```

Figure 4.1: Sample of cell tower scan records from MIT Reality Mining Dataset.

```
2    2011-12-18 13:02:00 926
2    2011-12-18 16:42:00 926
2    2011-12-18 17:01:00 926
3    2011-12-06 16:56:00 1080
3    2011-12-06 17:01:00 1080
3    2011-12-06 17:48:00 1080
```

Figure 4.2: Sample of cell tower scan records from D4D Challenge Dataset.

method is straightforward.

4.2.2 Single-Copy Location-Based Routing

In single-copy location-based routing, when a source/relay node encounters another node, if the encountered node has higher *geo-similarity* with the destination, the source/relay node will choose this encountered node as the new selected relay and delete the message from itself. The whole network only has one copy of the message. The details of the steps are as follows.

Step 1 (Each User Setup a Location Profile): Figure 4.1 and Figure 4.2 illustrates examples of the scan records on MIT reality and D4D challenge dataset respectively. The record includes the scanned tower ID and the time of this scan. Initially each user v_i caculate and save their *location profile* following the definition.

Step 2 (Calculate the Geo-similarity): When a source/relay node v_i encounters another node v_j, v_j gives v_i his/her *location profile*, then v_i calculate their *geo-similarities* Sim_{id} and Sim_{jd} with the destination node v_d respectively following the definition of *geo-similarity* (the *location profile* of v_d is included in the message).

Step 3 (Make Routing Decision): After getting the *geo-similarities*, v_i compare them. If

v_j has higher *geo-similarity* with the destination, v_i choose v_j as the new selected relay and delete the message from itself. Otherwise, the v_i will hold the message until it encounters another node and repeat the step 2 or finish routing if it meet the destination.

We now compare the performance of our single-copy scheme with other DTN routing algorithms listed below.

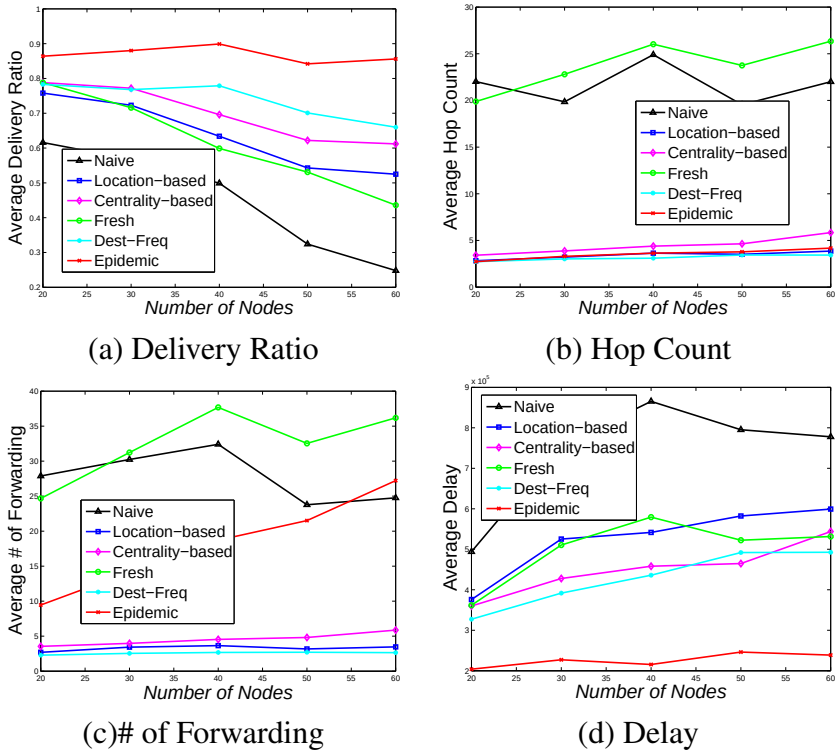
- Epidemic[87]: A broadcast method, during any encounter, a copy of the message is forwarded to all encountered nodes and the current node still hold a copy of the message. The epidemic Forwarding algorithm conducts the upper bound of the successful delivery ratio.
- Naive: during any encounter, the message is always forwarded to the encountered node and the current node will not hold the message after forwarding. If there are multiple nodes during the same encounter, the next hop is randomly picked. It can be treated as a single-copy version of Spray and Wait [82]. The naive algorithm conducts the lower bound of the successful delivery ratio.
- Fresh[18]: the message is only forwarded from the current node v_i to the encountered node v_j if v_j has met the destination more recently than v_i does. If there are multiple nodes satisfying such a condition during the same encounter, v_i forwards the message to the one who has met the destination most recently.
- Destination Frequency[20]: the message is only forwarded from v_i to v_j if v_j has met the destination more often than v_i does. If there are multiple nodes satisfying such a condition during the encounters, v_i forwards the message to the one who has met the destination most often.
- Centrality-Based: the message is only forwarded from v_i to v_j if v_j has higher centrality than v_i does. Here, we simply consider the degree centrality of each node, i.e., how many nodes it has encountered. A node with higher degree centrality is more popular in the network. If there are multiple nodes satisfying the condition during the encounter, v_i forwards the message to the one who has the highest centrality. Similar

idea has been used in Greedy-Total [19], SimBet [15] and Bubble Rap [40].

Our evaluations are conducted on the MIT Reality Mining Dataset. The reason we choose this datasets is it contains enriched cell tower scan trace, which record contacts among users and cellphone towers. The user carried mobile devices record the nearby cellphone towers by periodically scan. This gives us the location information we need to implement our location-based routing protocol. Also it includes various types of data, which may useful for our further continuous research, such as call logs, Bluetooth devices in proximity, application usage, and phone status. The Various experiment periods provide a lot of convenience too. The MIT Reality Mining Dataset is a relatively dense dataset compare with D4D Challenge dataset as it was collected from students and faculties in the same lab, their distributions are concentrated in a small area. On the other hand, users in D4D Challenge dataset are distributed nation wide. Thus compared with D4D Challenge dataset, MIT Reality Mining Dataset is more suitable for evaluate the performance of single-copy routing protocols. Because in a sparse network such as D4D Challenge dataset network, the success ratio of all single-copy routing algorithms are too low to have perceptible difference between them.

We implement all these six algorithms in a simulator developed by our group. We randomly choose 20, 30, 40, 50 and 60 users respectively from the MIT Reality Mining Dataset to build the different experiment environment. For each environment, we randomly generated the data source and destination as the routing assignments, and then apply the six algorithms. For all the simulations, we repeat the experiment for multiple times and compare each algorithm using the following four metrics:

- Average Successful Delivery Ratio: the average percentage of successfully delivered message from the sources to the destinations under the same environment (the same number of users).
- Average Hop count: the average number of hops during each successful delivery from the sources to the destinations.



(a) Delivery Ratio (b) Hop Count

(c)# of Forwarding (d) Delay

Figure 4.3: Performance comparison for single-copy routing on MIT Reality Dataset.

- Average Number of forwarding: the average number of messages forwarding in the network during the whole period.
- Average Delay: the average time duration of successfully delivered message from the source node to the destination node.

Figure 4.3 shows the performance comparison of the six algorithms on the four metrics. From the evaluation of average successfully delivery ratio in figure 4.3(a), we found that our location-based routing algorithm has much better average successfully delivery ratio than Naive algorithm, and similar with the Fresh, Destination Frequency and Centrality-Based algorithm. This proved that our location-based algorithm could achieve the acceptable delivery ratio. By observing the average hops and average copies in Figure 4.3(b) and (c) we can find that our algorithm has the smaller average hops and average copies than the Centrality-Based and Fresh algorithm. This proves our algorithm uses less network resources. So the slightly better successful delivery ratio of Centrality-Based and Fresh

algorithm is obtained with the cost of more communication cost. Comparing the average delay of our algorithm with the other algorithms we found the transmit delay of our algorithm is acceptable. From all these results above, we can conclude that our location-based algorithm is an effective routing method in DTNs environment. It is worth for us to explore location characteristics to improve DTN routing performance.

4.2.3 Multi-Copy Location-Based Routing

Although the single-copy location based algorithm has been proved has high successfully delivery ratio in MIT reality environment, there are still a gap between it and the upper bound (Epidemic Forwarding). Beside, MIT reality dataset is collected from a special campus environment, there are many kinds of DTNs (e.g. D4D Challenge Dataset, vehicular networks, etc.) which are not as dense as it. In such kind of networks, the single-copy routing paradigm may lead to very poor delivery performance. For example, in D4D most of algorithms only have less than 20% successful delivery ratio under the single-copy model. This motivated us to extend our algorithm into multi-copy scenario. In this scenario we allow limited number of message copies in the whole network. We expect that with the multiply copies of messages in the network, we could capture better opportunities to reach the destination.The detailed steps are as follows:

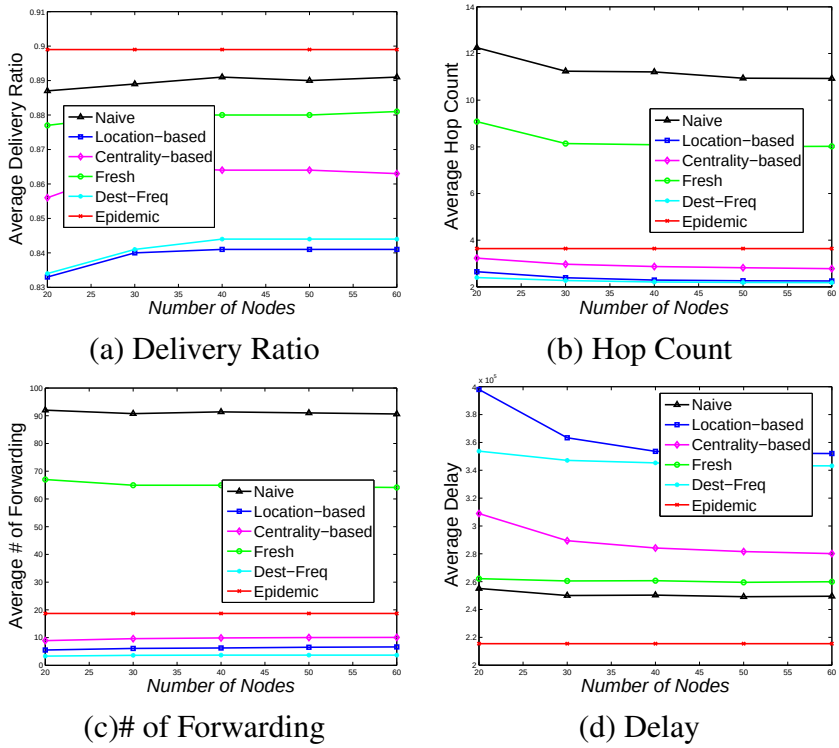
Step 1 (Each Node Setup a Location Profile): Same as in the single-copy location-based algorithm each user built a location profile.

Step 2 (Calculate the Geo-similarity): When a source/relay node v_i encounters another node v_j, v_i calculate their *geo-similarity* with the destination node respectively.

Step 3 (Make Routing Decision): After getting the *geo-similarities* of v_i and v_j with the destination v_d respectively, v_i compare Sim_{id} and Sim_{jd}. If $Sim_{jd} < Sim_{id}$, nothing changes. Else if $Sim_{jd} > Sim_{id}$, v_i choose v_j as the new selected relay. At the same time v_i check the number of copies N_c in the network(N_c could be get by broadcasting). If N_c is less than the copy number limitation N_{max}, both v_i and v_j keep a copy, else v_j keep the copy and delete the copy on v_i. All of the selected relay nodes will hold the message

until they encounter another node, and then they repeat the step 2. The routing process will finish if one of these selected relays meets the destination.

Performance Evaluation for Multi-Copy Location-Based Routing:



(a) Delivery Ratio (b) Hop Count

(c)# of Forwarding (d) Delay

Figure 4.4: Performance comparison for multi-copy routing on MIT Reality Dataset

We evaluate the performance of multi-copy location-based algorithm on both MIT Reality Dataset and D4D Challenge Dataset. The MIT Reality Dataset is a representation of small-scale dense delay tolerant network with limited number of mobile users. Meanwhile, the D4D Challenge Dataset is a representation of large-scale sparse DTN. It provides anonymized call patterns and mobility data of 5000 to 50000 mobile phone users in Ivory Coast. Besides, the relationships of users in D4D Dataset are full of variety compare with MIT Dataset. In MIT Dataset all of the users are students or faculty in the school and most of them are in the same lab. In D4D Dataset, users' relationships are much more colorful. They could be friends, classmates, relatives, colleagues, club members, familar strangers, etc. We believe that the MIT Reality Dataset is the first step of study on data delivery via

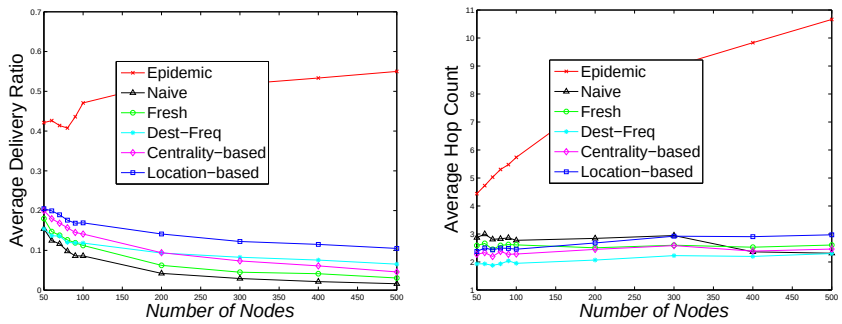
opportunistic communications. And the D4D Dataset represents the more general DTN scenario.

The evaluation on MIT Reality Dataset uses the same ways as it in single-copy location-based algorithm. We obtain the results in Figure 4.4. Compare with the results from single-copy location-based algorithm, we find that the routing performance could be slightly improved by using multi-copy model but not too much. Compare with the increased communication cost using multi-copy way, its may not worth to use multi-copy model in MIT Reality Dataset scenario. After allowing multiple copies in the network, the successfully delivery ratio of all routing algorithms has been very close to the upper bound. Notice that our location-based algorithm is not outstanding among the evaluated algorithms. This is mainly because the users' social relations in MIT Reality Dataset are strong, and the users' activities are always restricted in a small area and very regularly, which means in this kind of network, people's social relations do reflect real communication opportunities. This is not the kind of network scenario, whose routing performance we arms to use location characteristic to improve, as we described at the beginning of this section. So here after, we mainly use the D4D Challenge Dataset as our experiment environment.

In D4D Dataset environment, we assume that two phones can direct communicate to each other if they share the same cellular tower at particular time. Though this assumption may not be true in reality, it gives us an approximated environment for opportunistic communications. All of our experiments are based on the generated encounter databases from SET2. We will considers four different settings (A-D) in Table 2.2 for our experiments.

- Setting A: The first 15,000 users in the encounter database.
- Setting B: The users within the first 15,000 users whose physical locations of encounters limited to Abidjan city.
- Setting C: All 50,000 users in the encounter database.
- Setting D: The users within all 50,000 users whose physical locations of encounters limited to Abidjan city.

We implement seven algorithms (Epidemic, Naive, Fresh, Destination Frequency, Centrality-Based and Location-Based) in these four Setting respectively and compare each algorithm using the following four metrics: average successful delivery ratio, average hop count, average number of forwarding, average delay. For all experiments, we perform $5,000$ random routing tasks among the selected participators. All results reported here are the average over these tasks. For each experiment, we pick different number of nodes to participate the opportunistic communications, ranging from 50 to 500. Here we always pick the most active nodes (based on overall centrality) in the user set, since they are better candidates for opportunistic forwarding. For all opportunistic routing methods except for Epidemic, we allow multiple copies of the same message but limit the number of copies by a small constant. In the default setting, we use 10 as the constant bound.



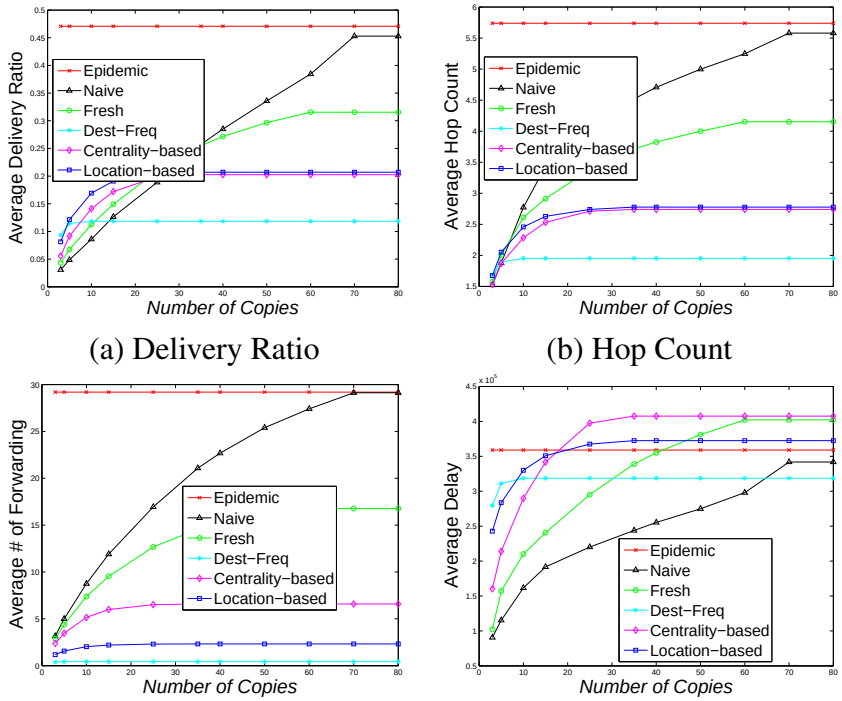
(a) Delivery Ratio (b) Hop Count

(c)# of Forwarding (d) Delay

Figure 4.5: Performance results over Setting A (the number of copies is fixed at 10).

In the first set of simulations, we use Setting A (with around $15,000$ selected users and within the full region). Figure 4.5 illustrate the results. From Figure 4.5(a), we can

see that our location-based methods can achieves better delivery ratio than other methods except the Epidemic algorithm which is the upper bound of the routing performance. This confirms that the understanding and usage of location relationships among mobile users is beneficial for making smarter forwarding decision. Notice that that even though Epidemic routing has the best delivery ratio, it costs extremely large amount of forwarding as shown in Figure 4.5(d). It is also noticeable that the delivery ratio is decreasing as the number of nodes increases. This is reasonable since we always choose the most active nodes as the participators. With more nodes included, more routing tasks are among less active nodes. In terms of hop count and number of forwarding, all opportunistic routing methods are at the similar level except for Epidemic. Notice that for delay since we only consider the successful routes, thus Epidemic usually has the largest delay.



(a) Delivery Ratio (b) Hop Count

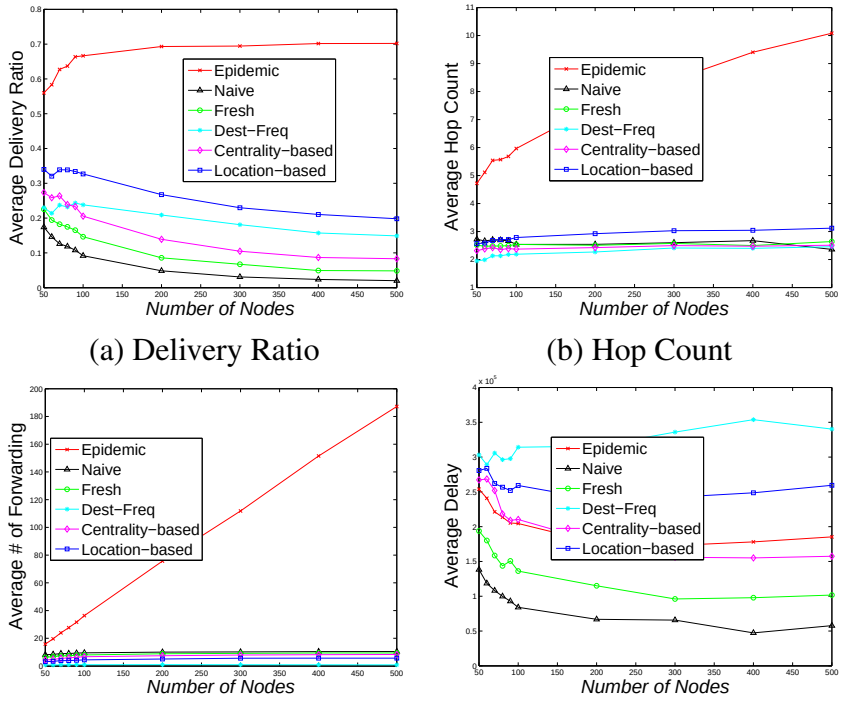
(c)# of Forwarding (d) Delay

Figure 4.6: Performance results over Setting A (the number of nodes is fixed at 100).

For the same Setting A, we then test the effect of the number of copies in multi-copy opportunistic routing. We fixed with 100 nodes and change the number of copies from 3 to

80. Figure 4.6 shows the results. It is obvious that with more message copies all methods can achieve higher delivery ratio but increase the number of forwarding too. There is clearly a trade-off between number of copies and forwarding overhead. When the number of copies reaches certain value, the delivery ratio will be stable. Further adding more copies does not help. For different methods, such critical value of copy number may vary.

To test the performance of all methods in a small and dense region, we then test our methods on Setting B, which limits the region around a rectangle region near Abidjan. Compared with the results in the full region (Setting A), all methods can achieve better performances in this setting. This is reasonable since a limited dense network provides more close opportunities for message delivery among mobile users than a larger and sparser network does.



(a) Delivery Ratio (b) Hop Count

(c)# of Forwarding (d) Delay

Figure 4.7: Performance results over Setting B (the number of copies is fixed at 10).

Last, we also perform simulation over the full population of D4D dataset (Settings C and D). Figure 4.8 and Figure 4.9 show the results, respectively. Compared with previous

results, all methods can achieve better performances too. The reason is still the same that within larger population the selected participants are more active thus lead to better chances for mobile delivery. Once again, better performance can also be achieved in a smaller and denser area.

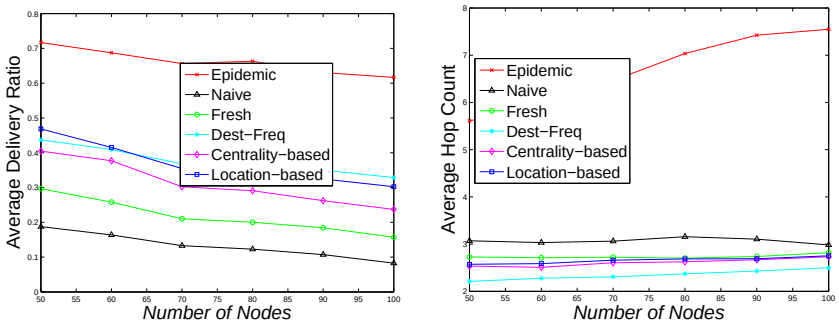
In summary, via the above simulations over the D4D dataset, we can have the following overall conclusions.

- Epidemic can achieve the highest delivery ratio since it takes every forwarding opportunities and does not have limitation on the number of copies. However, it suffers from the large number of forwarding, especially when the number of nodes is large. It could be used as the upper bound of the multi-copy algorithm's routing performance.
- Location-based, Centrality-based, and Destination Frequency can achieve relevant high delivery ratios while still use reasonable number of forwarding. In D4D Challenge Dataset environment, our Location-based algorithm can achieve better routing performance than other social or traditional DTN routing algorithms.
- Compared with different settings, all opportunistic routing can achieve better performance when the participants are active users and the physical region is small and dense. This can be shown in Figure 4.10 which summarizes the average delivery ratios over four different settings under the same parameters.

Since the number of encounters is significantly reduced after picking the subset users, the smaller size of user set could accelerate the execution time of our simulations. More important, the cellular towers stay the same level. So, here after we only evaluate our algorithms on Set B (among the first 15000 users who limits the region around Abidjan).

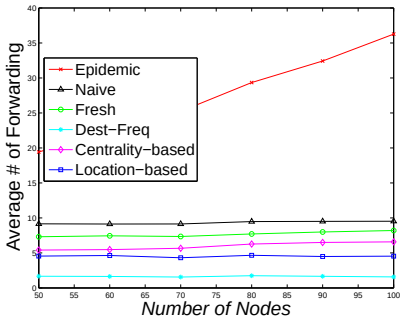
4.2.4 Simplified Location-Based Routing

Our location-based routing protocol has been showed effective on MIT Reality Dataset and excellent performance on D4D Challenge Dataset. However, this method needs to keep a *location profile* for each user and include the *location profile* of the destination node into the message. It works when the method is used for small-scale networks, however

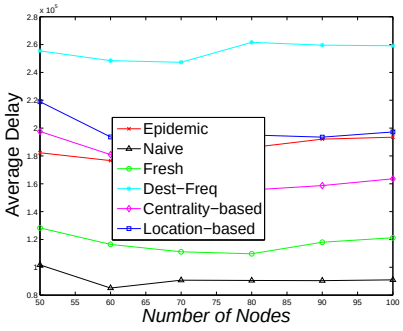


(a) Delivery Ratio

(b) Hop Count

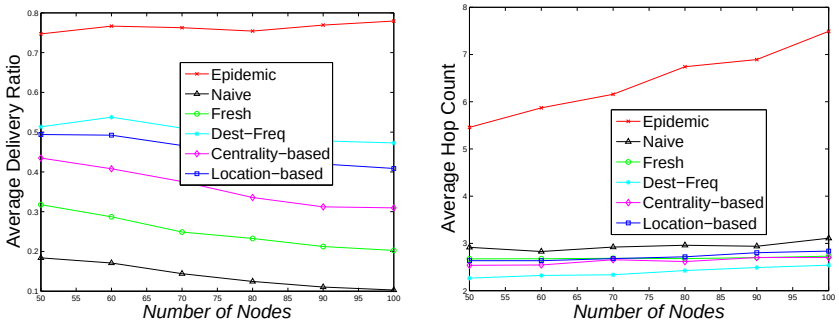


(c)# of Forwarding



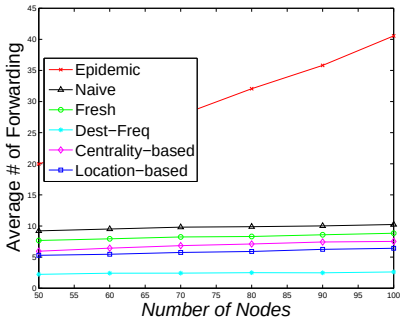
(d) Delay

Figure 4.8: Performance results over Setting C (the number of copies is fixed at 10).

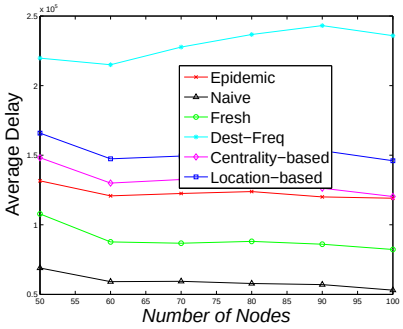


(a) Delivery Ratio

(b) Hop Count



(c)# of Forwarding



(d) Delay

Figure 4.9: Performance results over Setting D (the number of copies is fixed at 10).

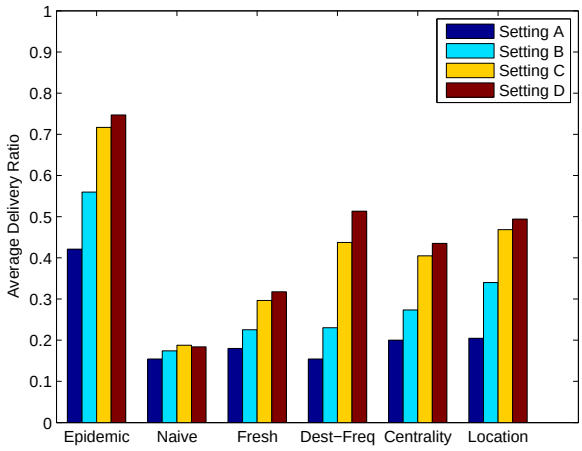


Figure 4.10: Average deliver ratios over Settings A to D (the number of nodes and the number of copies are 50 and 10, respectively).

for large-scale networks such as D4D Challenge Dataset network the huge size of *location profile* could result in considerable communication overhead. In the worst case the *location profile* of a D4D node could include around 1000 towers which could dramatically enlarge the message size. Motived by these consideration, we expect to design some simplified location-based algorithm.

Table 4.1: A D4D user's visited frequency and duration on cell towers.

Cell Tower ID	750	1129	953	898	303	163	1022	404	···
Total Visit Duration	33720	25354	15723	4268	210	113	36	29	···
Total Visit Frequency	216	187	135	52	8	5	2	1	···

From the observation of users *location profile*, we find that the towers which the user visited with large frequency and long duration are mainly limited among a small number of towers. Table 4.1 gives an example of one user's visited frequency and duration on cell towers in D4D Challenge Dataset. We can also observe that for most of the towers their visited duration has direct proportion with visited frequency. Considers short term opportunity encounters are already long enough to finish message transmission, we assume that the most frequently visited towers of a user could be enough to represent the users

location characteristics. Therefore instead of using *location profile*, we use the user's top 10 frequent visited towers to describe his/her location characteristic. We name these towers as *top 10 towers* in short here after, and represent the *top 10 towers* of node v_i as a vector $(t_{i1}, t_{i2}, \cdot, t_{i10})$.

Then we can define a new metric to approximatively measure the *geo-simularity* of two node:

Definition 4.3 (Number of Common Top Towers): Assume the top 10 towers of user v_i and user v_j are $(t_{i1}, t_{i2}, \cdots, t_{i10})$ and $(t_{j1}, t_{j2}, \cdots, t_{j10})$ respectively.

$$I_{kk'} = \begin{cases} 0, \text{when } t_{ik} \neq t_{jk'} \\ 1, \text{when } t_{ik} = t_{jk'} \end{cases}$$

The number of common top towers for node v_i and node v_j is $COMT_{ij} = \sum_{k=1}^{10} \sum_{k'=1}^{10} I_{kk'}$.

By using this new location-based metric, we modify our multi-copy location-based algorithm as follows:

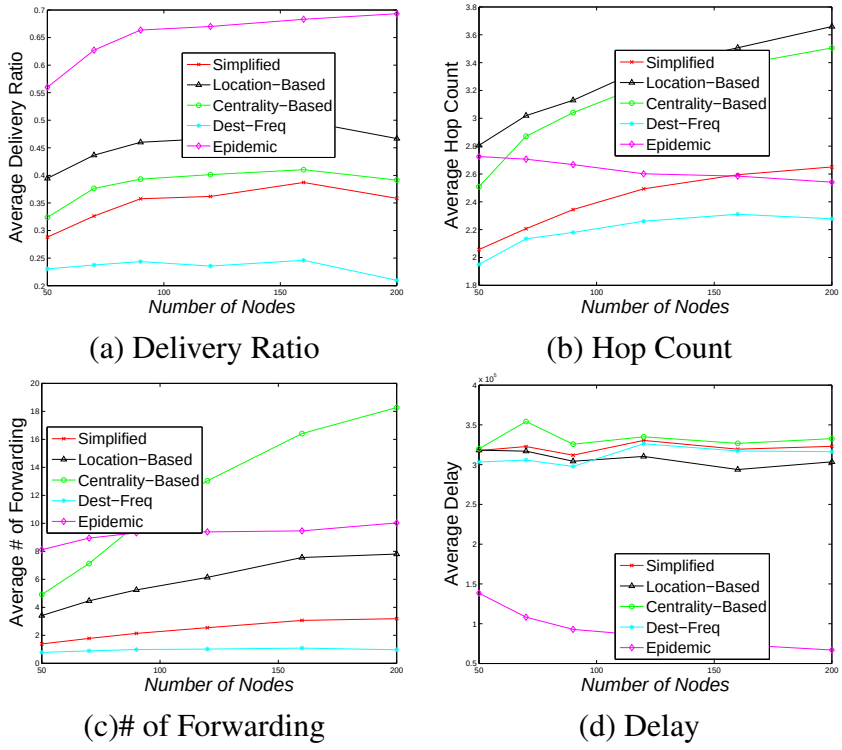
Step 1 (Get the top 10 towers for Each Node): Based on the user's cell tower scan records, each user saves the top 10 towers, which they visited the most frequently as their *top 10 towers*.

Step 2 (Calculate the number of common top towers): When a source/relay node v_i encounters another node v_j, v_i calculate their *number of common top towers* with the destination node v_d respectively.

Step 3 (Make Routing Decision): After getting the *number of common top towers* of v_i and v_j with the destination v_d respectively, v_i compare $COMT_{id}$ and $COMT_{jd}$. If $COMT_{jd} < COMT_{id}$, nothing changes. Else if $COMT_{jd} > COMT_{id}$, v_i choose v_j as the new selected relay. At the same time v_i check the number of copies N_c in the network. If N_c is less than the copy number limitation, both v_i and v_j keep a copy, else v_j keep the copy and delete the copy on v_i. All of the selected relay nodes will hold the message until they encounter another node, and then they repeat the step 2. The routing process will finish

if one of these selected relays meets the destination.

We only introduce the multi-copy form of this algorithm, because we will only evaluate its performance on the D4D Challenge Dataset, which is a sparse network and unsuitable for single-copy routing model. Hear after, we only introduce and evaluate our designed algorithms in multi-copy model, and we believe its easy to extend them into single-copy model also. Using the *number of common top towers* instead of the *location profile* we only need to add 10 tower Ids onto each message and into the user's storage. This way will significant reduce the communication and network management overhead.



(a) Delivery Ratio (b) Hop Count

(c)# of Forwarding (d) Delay

Figure 4.11: Performance results of simplified location-based method over Setting B of D4D.

We now evaluate the performance of simplified location-based algorithm on Setting B of D4D Challenge Dataset. We compare the simplified location-based algorithm with five algorithms(Epidemic, Destination Frequency, Centrality-Based and Location-Based) using the following four metrics: average successful delivery ratio, average hop count, average number of forwarding, average delay. For all experiments, we perform $5,000$ random

routing tasks among the selected participators. All results reported here are the average over these tasks. For each experiment, we pick different number of nodes to participate the opportunistic communications, ranging from 50 to 200. Here we always pick the most active nodes (based on overall centrality) in the user set, since they are better candidates for opportunistic forwarding. For all opportunistic routing methods except for Epidemic, we allow multiple copies of the same message but limit the number of copies by a small constant. In the default setting, we use 20 as the constant bound.

Figure 4.11 illustrate the results. From Figure 4.11(a), we can see that our simplified location-based method can achieves better delivery ratio than Destination Frequency algorithm, and very close to Centrality-Based and Location-Based algorithm. This confirms that the simplified metric do capture the main characteristics of users location features.

4.3 Nokia Mobile Data Challenge: Location Analysis

Location information has been proved very useful on DTN routing. It is interesting to study on the real-world tracing data and answer the following questions:

1. What kind of features we can get from these location information?
2. How could we get them?
3. Are they helpful on DTN routing?

We participated in the Nokia Mobile Data Challenge (MDC)[11], which gives us an opportunity to work with a unique and relatively unexplored rich mobile dataset. One of the task of Nokia Mobile Data Challenge is semantic place prediction. The available data for the Challenge was collected by the Nokia Data Collection Campaign with 200 participants carried the smartphones in the course of more than one year. We name this dataset as MDC dataset for short hearafter. MDC dataset include rich data related to location. For competition purpose, only a subset of 80 users' dataset is released as the training data. For each user in the MDC data, the raw location data (based on GPS and WLAN) was first transformed into a symbolic space which captures most of the mobility information but excludes actual geographic coordinates. This was done by first detect the visited (checked-

in) places and then mapping the sequence of coordinates into the sequence of visits to checked-in places (represented by a place ID). Places are user-specific and they are ordered by the time of the first visit (the visit sequence starts with place $ID = 1$). Each place corresponds to a circle with a radius of 100 meters.

The users' location information we used is in the form of *visitsequence20min.csv* record which is the sequence of place visits which are longer than 20 minutes. It includes:

- userid: id of the user
- unixtimestart: unix time of the phone when the visit started.
- tzstart: time zone of the phone when the visit ended.
- unixtimeend: unix time of the phone when the visit end.
- tzend: time zone of the phone when the visit ended.
- trustedstart: the start time is trusted if there are location data points in the period of 10 minutes before the arrival time ($0 = false$, $1 = true$).
- trustedend: the end time is trusted if there are location data points in the period of 10 minutes before the arrival time ($0 = false$, $1 = true$).
- trustedtransition: the transition between the current visit and the next visit is trusted if there are location data points every 10 minutes between the leaving time of the current visit and the starting time of the next visit. If the transition is trusted then it is not possible to have missing more than 20min visit in the transition.

While people travel further and faster than ever before, it is still the case that they spend much of their time at a few important places. Identifying these key locations is thus central to understanding human mobility and social patterns. In this task , by analyzing anonymized cellular network data, we arms to identify the generally important locations (extracted from location data) and discern the semantic meaning of these places, such as "work place", "restaurant", etc. The list of places to be annotated is belong to one of the 10 categories:

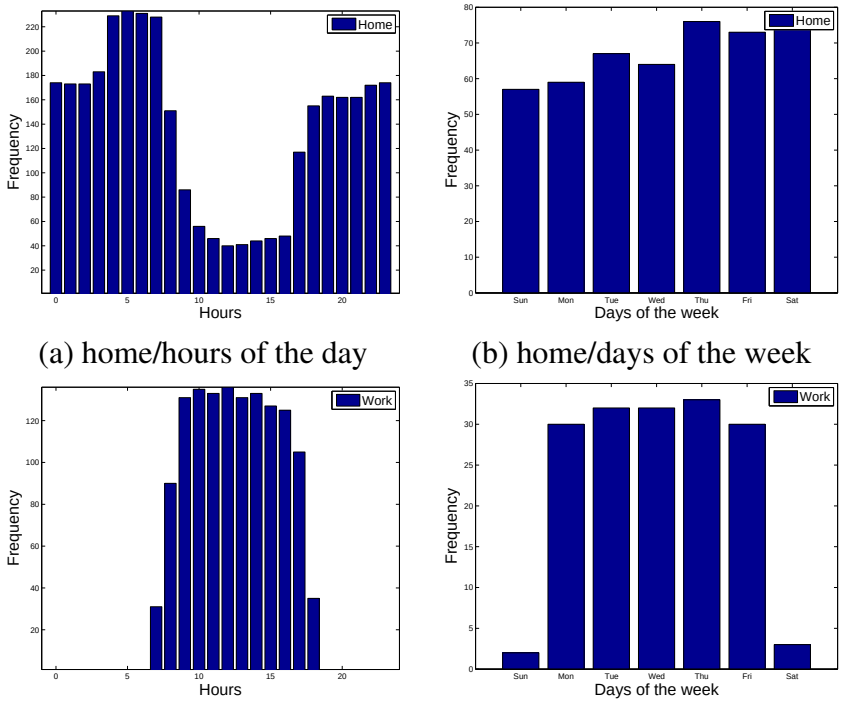
1. Home

2. Home of a friend, relative or colleague
3. My workplace/school
4. Location related to transportation (bus stop, metro stop, train station, parking lot, airport)
5. The workplace/school of a friend, relative or colleague
6. Place for outdoor sports (e.g. walking, hiking, skiing)
7. Place for indoor sports (e.g. gym)
8. Restaurant or bar
9. Shop or shopping center
10. Holiday resort or vacation spot

Beside the collected mobile phone data, we also get some ground truth data: semantic labels of several places that were visited during the data collection period. There are totally 331 instances. This data could be use to verify the classification accuracy of our semantic prediction method.

Semantic Place Prediction Method:

People's access of some particular place may follow some regulations. For example, during weekdays, Bob will get up at 6:00 am and send his daughter to the kindergarten at 6:30 am, he then will came back. After having breakfast, he will go to work at 8:00 am, he will after work at 4:00 pm, when to gym for one hour and pick up his daughter at 5:30 pm, then he will back home and stay. He repeats this routine everyday. These regulations very related to the time. For example, from 8:00 am to 4:00 pm Bob probably not at home because it is his work time. We may distinguish different places by study people's access time on them. Figure 4.12 illustrate a MDC user's access frequency distribution of his home and work place during a day and a week respectively. It's easy to find that the user's access distributions of his home and work place have significant difference. His home has very high access frequency at night and equal access frequency on each day during a week, meanwhile his workplace has very high access frequency at daytime and during weekdays



(a) home/hours of the day (b) home/days of the week

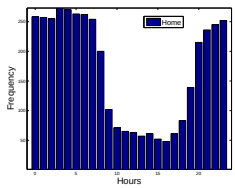
(c)work place/hours of the day (d) work place/days of the week

Figure 4.12: User A's access frequency distribution of home and work place.

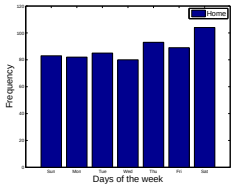
(Monday to Friday). This implies the features of people's access on a place could be used to predict its semantic meaning. Even better, different people may have some similar access features on a particular type of place. Figure 4.13 illustrate the two different MDC users' access frequency distribution of their home and workplace respectively. Both users' home have very high access frequency at night and equal access frequency on each day during a week. Their workplaces also both have very high access frequency at daytime and during weekdays. So we believe, for some specific places there exist common rules to predict them. Like most of people go to restaurant during 12:00am-2:00pm and 5:00pm-8:00pm, most of people sleep at night. By reveal these observable factors, we may predict the semantic meaning of a place.

Studying our ground truth data, we reveal the following observable factors that capture characteristics, which are useful on distinguishing the places semantic meanings:

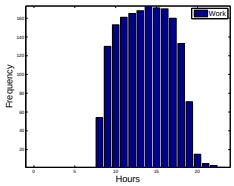
- Days in a month: The number of days, the user accessed the place in one mouth.



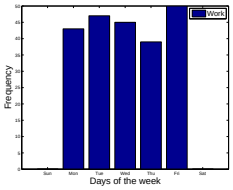
(a) User B's home



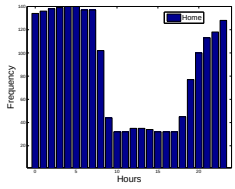
(b) User B' home



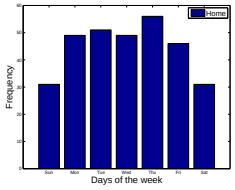
(c) User B workplace



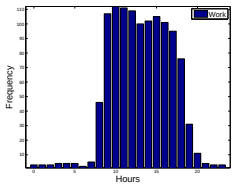
(d) User C's workplace



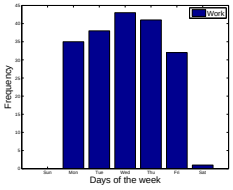
(e) User C's home



(f) User C's home



(g) User C's workplace



(h) User C's workplace

Figure 4.13: User B and C's access frequency distribution of their home and workplace

If the user accessed the place many times on the same day, the day is counted only once.

- Total access frequency: The sum number of times the place was visited during the whole data collection period of the MDC data.
- Total access frequency of visits shorter than 2 hours: The sum number of times the place was visited shorter than 2 hours during the whole data collection period of the MDC data.
- Total access frequency of visits longer than 2 hours and shorter than 4 hours: The sum number of times the place was visited longer than 2 hours and shorter than 4 hours during the whole data collection period of the MDC data.
- Total access frequency of visits longer than 4 hours: The sum number of times the place was visited longer than 4 hours during the whole data collection period of the MDC data.
- The average access frequency in the weekdays: The sum number of times the place was visited in the weekdays during the whole data collection period of the MDC data divide by 5.
- The average access frequency in the weekend: The sum number of times the place was visited in the weekend during the whole data collection period of the MDC data

divides by 2.

- Total access frequency in the daytime: The sum number of times the place was visited between 7:00am to 7:00pm during the whole data collection period of the MDC data.
- Total access frequency at the nighttime: The sum number of times the place was visited between 7:00pm to 7:00am during the whole data collection period of the MDC data.
- Total access frequency in the sleeping time (12:00am-6:00am): The sum number of times the place was visited between 12:00am to 6:00am during the whole data collection period of the MDC data.
- Total access duration: The sum of duration the place was visited during the whole data collection period of the MDC data.
- Total access duration of visits shorter than 2 hours: The sum of duration the place was visited shorter than 2 hours during the whole data collection period of the MDC data.
- Total access duration of visits longer than 2 hours and shorter than 4 hours: The sum of duration the place was visited longer than 2 hours and shorter than 4 hours during the whole data collection period of the MDC data.
- Total access duration of visits longer than 4 hours: The sum of duration the place was visited longer than 4 hours during the whole data collection period of the MDC data.
- The average access duration in the weekdays: The sum of durations the place was visited in the weekdays during the whole data collection period of the MDC data divide by 5.
- The average access duration in the weekend: The sum of durations the place was visited in the weekend during the whole data collection period of the MDC data divides by 2.
- Total access duration in the daytime: The sum of durations the place was visited

between 7:00am to 7:00pm during the whole data collection period of the MDC data.

- Total access duration at the nighttime: The sum of duration the place was visited between 7:00pm to 7:00am during the whole data collection period of the MDC data.
- Total access duration in the sleeping time: The sum of duration the place was visited between 12:00am to 6:00am during the whole data collection period of the MDC data.

4.3.1 Rule Based Prediction: Predict Home and Work Place.

Comparing with other place categories, home and work place are two of the most important places in people's life. They normally have the largest "total access frequency" and they are easier to be detected. An obvious feature of home is: most people sleep at home. So just use the single factor of the "total access duration in the sleeping time" we can detect people's home. The workplace detection is a little bit complicated compare with home detection. As we observed already, most of people's work places should have high "average access frequency in the weekdays" and high "total access frequency in the daytime". They also should have low "average access frequency in the weekend" and low "total access frequency at the nighttime". Following these rules we make our home and work place detection strategies:

- Home: Comparing the "total access duration in the sleeping time" of each place ID for a user, we set the one with highest "total access duration in the sleeping time" as the user's home.
- Workplace: Check the places of each user with the top five "total access frequency". For the places, whose "average access frequency in the weekdays" is larger than its "average access frequency in the weekend" and its "total access frequency in the daytime" is larger than its "total access frequency in the nighttime", we identify the place with the largest "total access frequency" as the user's work place.

We applied our home and workplace detection algorithm on the MDC data and validate the results with the ground truth data. We found our home and workplace detection algorithms have very high classification accuracy. Table 4.2 summarized the detailed classification accuracy of home. Here the class "Yes" represent the place labels classified as home, the class "No" represent the place labels classified as not home. The true positive rate (TP rate) of class "Yes" represent the ratio the real home been classified as home, the true positive rate of class "No" represent the ratio the place which is not home been classified as not home, the false positive rate (FP rate) of class "Yes" represent the ratio the place which is not home been classified as home, the false positive rate of class "No" represent the ratio the home been classified as not home. Similarly, Table 4.3 summarized the detailed classification accuracy of work place. Considering not everyone will have distinct home or work locations: some people work at home, others have no fixed work site, and still others may not use their cell phones at home, our algorithms produce good approximation of true home and work locations.

Table 4.2: The detailed classification accuracy of home

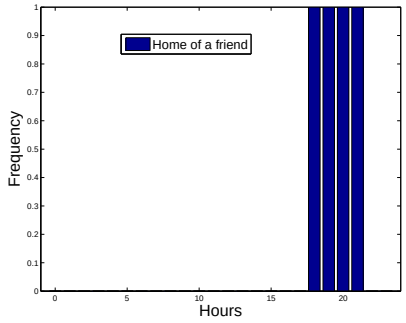
Class	TP Rate	FP Rate
Yes	0.762	0.081
No	0.919	0.238

Table 4.3: The detailed classification accuracy of work place

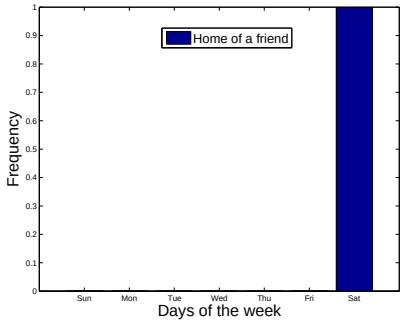
Class	TP Rate	FP Rate
Yes	0.765	0.105
No	0.895	0.235

4.3.2 Semantic Place Predict with Machine Learning :

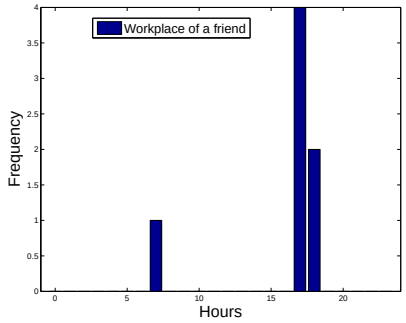
We manually formulated the rules to distinguish home and work place successfully, however not all categories of the places have such obvious regulation. Figure 4.14 illustrate the access frequency of friend's home and friend's work place of user A, it is hard for us to find any characteristics. That means we need to explore more smart ways to learn their regulations.



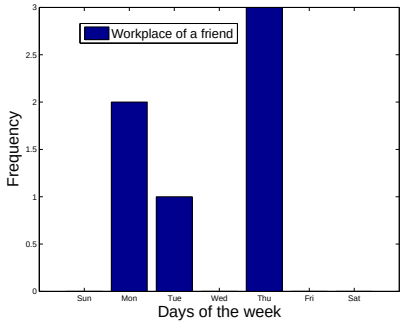
(a) friend's home/hours of the day



(b) friend's home/days of the week



(c)friend's work place/hours of the day



(d)friend's work place/days of the week

Figure 4.14: User A's access frequency distribution of friend's home and friend's work place

Currently the machine leaning methods are widely used on evolve behaviors based on empirical data. A learner can take advantage of examples (data) to capture characteristics of interest of their unknown underlying probability distribution. Data can be seen as examples that illustrate relations between observed variables. The machine learning methods could automatically learn to recognize complex patterns and make intelligent decisions based on data. So we use WEKA [33], a comprehensive tool for machine learning and data mining to explore the user's mobility pattern on different categories of places.

Two basic concepts in WEKA are dateset and classifier. A dataset is a collection of the classified examples/instances. Each instance consists of a number of attributes, any of which can be nominal (one of a predefined list of values), numeric (a real or integer number) or a string (an arbitrary long list of characters, enclosed in "double quotes"). The external representation of an instances class is an ARFF file, which consists of a header

describing the attribute types and the data as comma-separated list. The Classifiers are the machine learning algorithms. The Classifier use the examples/instances in the dataset to train a pattern, then using the obtained pattern we can classify the unclassified data.

```
% Here we state the internal name of the dataset
@relation placelabe1_1
% These lines define six numeric attributes (represent our six observable
factors)
@attribute Duration numeric
@attribute Frequency numeric
@attribute Day numeric
@attribute Night numeric
@attribute Weekday numeric
@attribute Weekend numeric
%The last attribute is the default target or class variable used for
prediction. In our case it is a nominal attribute with two values, “Yes”
represents the place is home, “No” represent it’s not home
@attribute Label {Yes,No}
% The rest of the dataset consists of the token @data, followed by comma-
separated values for the attributes -- one line per example. In our case,
the first six values of attribute describe the factors of the place, the
last represent the ground truth whether this place is home or not.
@data
%107   3 (user and place ID)
7239873,218,74,144,38.8,12.0,Yes
%177   2
3474346,187,175,12,31.0,16.0,No
%081   1
11571846,243,63,180,35.2,33.5,Yes
%146   2
4664400,374,372,2,57.2,44.0,No
%043   11
4641364,275,271,4,40.6,36.0,No
%059   2
2412795,178,167,11,31.0,11.5,No
```

Figure 4.15: Sample of WEKA dataset.

Here is an example of how we use weka to extract the pattern of a specific category of place. Figure 4.15 is an commented example of dataset, which is used to extract the pattern of place category: 1(home). We name this dataset as training set hereafter. This dataset tells the classifier places with what kind of attribute is home or not home. For example, the first instance tells the place with total duration 7239873 second, total frequency 218, 74 total access frequency during the day, 144 total access frequency at night, 38.8 average access frequency in weekday, 12 average access frequency in the weekend is home. The classifier

will then automatically train a rule. We save this rule, and apply it on other instances (we name these instances as testing set hereafter), it will tell us which class the instance belongs to.

To test the performance of this machine learning method we divide our ground truth data into two parts. We randomly pick 1/4 of the instances as the testing set and remaining 3/4 instances as training set. Table 4.4 shows the distribution of our instances on different place categories. We use all the factor we explored("days in a month","total access frequency", etc.) as the instance attributes, and we use five different classifiers: NaiveBayes, BayesNet, IBK, J48 and AdaBoostM1. Table 4.5 list the true positive rate of class "Yes" for each place category and classifier respectively. Table 4.6 list the true positive rate of the class "No" for each place category and classifier respectively. Table 4.7 list the false positive rate of the class "Yes" for each place category and classifier respectively. Table 4.8 list the false positive rate of the class "No" for each place category and classifier respectively. We found that different classifier has different predict accuracy especially for place labels which has less instances. Comparing all the classifier we use, NaiveBayes has the most stable performance on different place categories.

Table 4.4: The distribution of instances

Place Label(Category)	1	2	3	4	5	6	7	8	9	10
ground truth	84	46	102	23	9	25	14	11	17	5
training set	64	35	70	15	6	20	10	8	12	4
testing set	20	11	32	8	3	5	4	3	5	1

The results show the machine learning method have less accuracy of detecting home and work place than our manually formulated detection algorithms. We want to find out if the attributes selection effect the accuracy. Then we again use the single factor "the total access duration in the sleeping time" as the instance attributes and use the classifier IBK to detect home. We get the detailed classification accuracy in Table 4.9. The result shows the detection accuracy is different with the multiply attributes detection. So we found that, the attributes selection effect the predict accuracy too. The key of improving the machine

Table 4.5: The true positive rate of class “Yes”

Place Label(Category)	NaiveBayes	BayesNet	IBK	J48	AdaBoostM1
1	0.5	0.563	0.688	0.625	0.563
2	1	0	0.556	0.222	0.444
3	0.762	0.429	0.524	0.476	0.333
4	1	1	0.5	0.5	0
5	1	0	0	0	0
6	0.9	0.6	0	0	0
7	0.857	0	0	0.143	0
8	0.8	0	0	0	0
9	1	1	0	0	0
10	1	0	0	0	0

Table 4.6: The true positive rate of class “No”

Place Label(Category)	NaiveBayes	BayesNet	IBK	J48	AdaBoostM1
1	0.964	0.909	0.945	0.982	0.982
2	0.21	1	0.774	0.903	0.823
3	0.62	0.96	0.9	0.88	0.9
4	0.701	0.806	0.925	0.925	1
5	0.6	0.729	0.986	1	1
6	0.525	0.639	0.934	1	1
7	0.484	1	1	1	1
8	0.606	1	1	1	1
9	0.464	0.71	0.971	0.957	1
10	0.814	1	0.986	1	0.986

Table 4.7: The false positive rate of class “Yes”

Place Label(Category)	NaiveBayes	BayesNet	IBK	J48	AdaBoostM1
1	0.036	0.091	0.055	0.018	0.018
2	0.79	0	0.226	0.097	0.177
3	0.38	0.04	0.1	0.12	0.1
4	0.299	0.194	0.075	0.075	0
5	0.4	0.271	0.014	0	0
6	0.475	0.361	0.066	0	0
7	0.516	0	0	0	0
8	0.394	0	0	0	0
9	0.536	0.29	0.029	0.043	0
10	0.186	0	0.014	0	0.014

Table 4.8: The true positive rate of class "No"

Place Label(Category)	NaiveBayes	BayesNet	IBK	J48	AdaBoostM1
1	0.5	0.438	0.313	0.375	0.438
2	0	1	0.444	0.778	0.556
3	0.238	0.571	0.476	0.524	0.667
4	0	0	0.5	0.5	1
5	0	1	1	1	1
6	0.1	0.4	1	1	1
7	0.143	1	1	0.857	1
8	0.2	1	1	1	1
9	0	0	1	1	1
10	0	1	1	1	1

learning's accuracy is to set the appropriate attributes and classifier.

Table 4.9: The detailed classification accuracy of home by machine learning

Class	TP Rate	FP Rate
Yes	0.625	0
No	1	0.375

We also observed that one category of place may have multiply patterns, such as the work place. Many people have more than one work places, one of them is their main work place, another, for example just access once a week. So even both of them are work places they have some different features. We split the work place instances into two groups. Group one is the instances for the main work place, group two includes the instances of others. If a user has only one work place we put it into group one. If a user has multiply work places we put the one with largest "total access duration" into group one, others into group two. We train the two groups separately with classifier IBK. Then we use the obtained patterns to classify our test set. Table 4.10 and 4.11 list the detailed accuracy of these two patterns respectively. We combines the identified workplace from these two patterns together, Comparing this result with our manually formulated method, the combined result has better accuracy Table 4.12 list the detailed accuracy of it.

Table 4.10: The detailed classification accuracy of work place by machine learning with pattern obtain from group one

Class	TP Rate	FP Rate
Yes	0.614	0.082
No	0.918	0.386

Table 4.11: The detailed classification accuracy of work place by machine learning with pattern obtain from group two

Class	TP Rate	FP Rate
Yes	0.597	0.041
No	0.959	0.403

4.3.3 Other Tasks in MDC

Besides the task of semantic place prediction, we also participate in the nexplace prediction taks of MDC. In this task, we explore methods of predicting the user's next destination using the user's current context. We also solved this problem using machine learning method and get a good correct classify ratio. Among all the teams competitor in Nokia Mobile Data Challenge, the prediction accuracy of our methods rank in the fourth and fifth for the two tasks.

4.4 Location-Social Based Routing

From the review of previous works on social-based DTN routing and our experimental results on location-based DTN routing, we found both methods are effective. We are going to explore the novel joint social- and location-aware approaches to design more efficient routing protocols for the DTNs' environment.

A feasible idea is extract user's social properties from their location information. By study the Nokia Data Collection Campaign dataset (MDC Dataset), we found some ap-

Table 4.12: The detailed classification accuracy of work place by machine learning with combining two patterns

Class	TP Rate	FP Rate
Yes	0.867	0.122
No	0.878	0.133

proaches to identify important places of user. These analysis results can also help us to predict user's social relations. For example if two users' home are very close, they are neighbors. If they live at the same place, they are probably family members. If two users work at the same place they might be colleagues. If the two user's friends lives or works at the same place they may also friends. If we found a user visit another user's home very often they may be friends. If a user only visits another user's work place, they may be business partner but they probably are not friends. In this section we will focus on how to extract these social relations, which observed from the location information, as useful social properties for DTN routing. And, we will explore how to apply them on DTN routing.

4.4.1 Location-Social Based Metrics

In Section 4.3, we explored two ways to identify important places of users. One is manually establish some rules; the other is use machine learning method. Using the machine learning method can get more accurate results and increase the possibility of discovering some inconspicuous characteristics of certain specific place. However these benefits brings large computation cost, which is both unaffordable and unworthy for DTN routing. Especially considers that, most people spend a huge percentage of their time at their home and work places, (for example, a normal person who sleep eight hours a day and works eight hours during weekdays, spend at least $8 * 7 + 8 * 5 = 96$ hours per week at his/her home and work place which is more than 57% of time in his/her life), and our manually rules to detect home and workplace already have very good accuracy, we will explore how to use these rules to extract user's social features from location information.

Our simplified location-based method proved that using cell towers, which a user frequently visited, is a good way to represent the user's location characteristics. Similarly, we use the towers a user frequently visit during "sleeping time to distinguish a user's home feature. This feature describes that the user lives around the location of these towers. Also, we distinguish a user's work place feature as the towers the user frequently visit during daytime of weekdays (Monday to Friday), which describes the user works around the physic

location of these towers. So we give the definition of the user's *top 5 home-towers* and *top 5 work-towers* as follows:

Definition 4.4 (A User's Top 5 Home-Towers): A user v_i's top 5 home-towers is five of the user v_i's visited towers appears in its visited history (visit scan record), which user v_i visited with the longest sum of duration, during 11pm$\sim$7am every day of the test period.

Definition 4.5 (A User's Top 5 Work-Towers): A user v_i's top 5 work-towers is five of the user v_i's visited towers appears in its visited history (visit scan record), which user v_i visited with the longest sum of duration, during 8am$\sim$7pm every weekday of the test period.

With the two metrics (top 5 home/work-towers) to describe a individual's live and work feature, we can define two new metrics to approximatively measure how likely two nodes will be colleagues or neighbors:

Definition 4.6 (Number of Common Home-Towers): Assume the top 5 home-towers of user v_i and user v_j are $(t_{i1}, t_{i2}, \cdot, t_{i5})$ and $(t_{j1}, t_{j2}, \cdot, t_{j5})$ respectively.

$$I_{kk'} = \begin{cases} 0, \text{when } t_{ik} \neq t_{jk'} \\ 1, \text{when } t_{ik} = t_{jk'} \end{cases}$$

The number of common home-towers for node v_i and node v_j is $COMH_{ij} = \sum_{k=1}^{5} \sum_{k'=1}^{5} I_{kk'}$.

Definition 4.7 (Number of Common Work-Towers): Assume the top 5 work-towers of user v_i and user v_j are $(t_{i1}, t_{i2}, \cdot, t_{i5})$ and $(t_{j1}, t_{j2}, \cdot, t_{j5})$ respectively.

$$I_{kk'} = \begin{cases} 0, \text{when } t_{ik} \neq t_{jk'} \\ 1, \text{when } t_{ik} = t_{jk'} \end{cases}$$

The number of common work-towers for node v_i and node v_j is $COMW_{ij} = \sum_{k=1}^{5} \sum_{k'=1}^{5} I_{kk'}$.

The more number of common home/work-towers two users have, the more likely they will live/work in the same area, which means they will probably have more communication opportunities. Thus, we can roughly construct the users' home/work contact graph:

Definition 4.8 (Users' Home Contact Graph): We model the users' home contact graph as a 2-dimensional undirected graph $G_h = (V, E)$. where $V = (v_1, \cdots, v_n)$ denotes the set of users and E denotes a set of links. A link $v_i v_j \in E$ denotes the number of common home-towers for user v_i and v_j is larger than 2, which means user v_i and v_j live within the same area.

Definition 4.9 (Users' Work Contact Graph): We model the users' work contact graph as a 2-dimensional undirected graph $G_w = (V, E)$. where $V = (v_1, \cdots, v_n)$ denotes the set of users and E denotes a set of links. A link $v_i v_j \in E$ denotes the number of common work-towers for user v_i and v_j is larger than 2, which means user v_i and v_j work within the same area.

A link in home/work contact graph indicates the two users may have strong connection(communication opportunity).

With the users' home and work contact graph, we can have the users' degree and betweenness centrality on them:

Definition 4.10 (The User's Home/Work Degree Centrality): The user v_i's degree centrality is the number of links incident upon v_i on the home/work contact graph.

Definition 4.11 (The User's Home/Work Betweenness Centrality): The user v_i's betweenness centrality is the number of shortest paths passing via node v_i on the home/work contact graph.

A user with a high degree centrality on home/work contact graph is a popular user with a large number of possible contacts. He/she may live/work in the area a large number of users in the network live in/work at. A user with high betweenness centrality on home/work contact graph can control or facilitate many connections between other users.

4.4.2 Location-Social Based Routing Protocols

The metrics we explore in the last section represent users social/location characteristic and their relationships, which could be helpful on DTN routing. Now the question is how

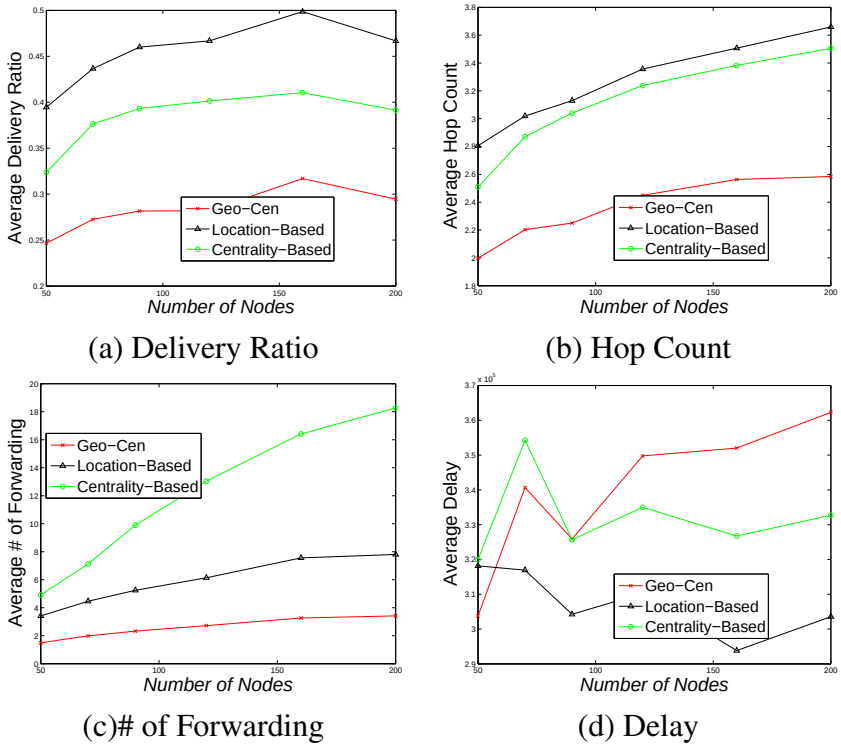
to smartly use them to make routing decision.

Using multiply metrics as relay selection criterion is a straight forward method to take advantage of varies information resource. The use of multiply social-based metrics has been proved effective. In [15], Daly and Haahr uses *betweenness centrality* to identify those nodes who can act as bridges in their neighborhood, and uses *similarity* metrics to identify nodes are more likely to find a common neighbor with the destination which can act as the forwarder. Hui *et al.* [40] also relied on two social characteristics (community and centrality). They assumed that each node belongs to at least one community and its node centrality (either betweenness or degree centrality) in the community describes the popularity of the node within this community. We explore the combination of using social- and location-based metrics.

We first designed two group of experiments, both of which demonstrated that simple combination of multiply location and/or social metric may lead to poor routing performance.

Our first experiment test the combination of one social and one location metric: the *geo-similarity* and *degree centrality*. We name our new method as Geo-Cen. This method only forward the data to the node whose geo-similarity to the destination and degree centrality are both larger than that of the current node. We compare the method with the method which only use geo-similarity and centrality as routing metric (Location-Based and Centrality-Based). The evaluation is in multi-copy model on the Setting B of D4D Challenge Dataset. The result in Figure 4.16 shows that such simple combination approach does not improve the performance. The successful delivery ratio of the Geo-Cen algorithm is even lower than both the location-based algorithm and the centrality-based algorithm. The reason is this method only choose the node, whose geo-similarity to the destination and degree centrality are both larger than that of the current node, as relay node. Surely such kind of node should be ideal to relay the message, however using such a strict rule, it is hard for the source node to find a eligible relay. If the source node does not encountered

such kind of node, it just wait and lose the successful communication opportunities.



(a) Delivery Ratio (b) Hop Count

(c)# of Forwarding (d) Delay

Figure 4.16: Simply combining geo-similarity with degree centrality: poor performance.

Our second experiment test the combination of two location metrics: the *number of common home-towers* and *number of common work-towers*. We compare the performance of three routing strategies: Geo-Home, Geo-Work, Home-Work. The details of these three algorithms are listed as follows:

Geo-Home : when a node v_i encountered another node v_j. If the *number of common home-towers* between the encountered node v_j and the destination node v_d is larger than the *number of common home-towers* between the node v_i and v_d, forward the message.

Geo-Work : when a node v_i encountered another node v_j. If the *number of common work-towers* between the encountered node v_j and the destination node v_d is larger than the *number of common work-towers* between the node v_i and v_d, forward the message.

Home-Work : when a node v_i encountered another node v_j. If both the *number of common home-towers* and *number of common work-towers* between the encountered node

v_j and the destination node v_d is larger than 1, forward the message.

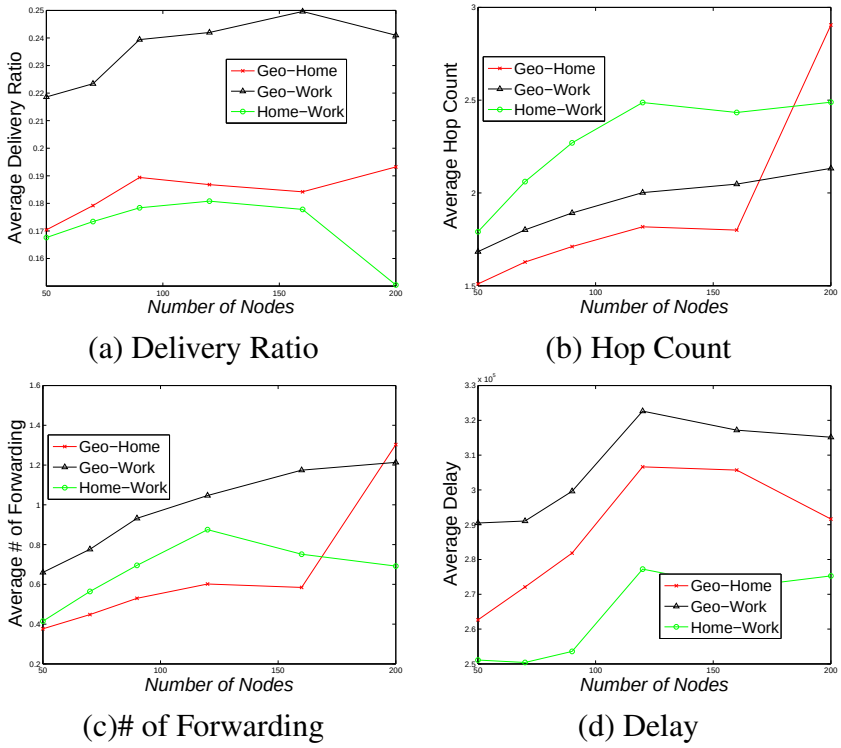
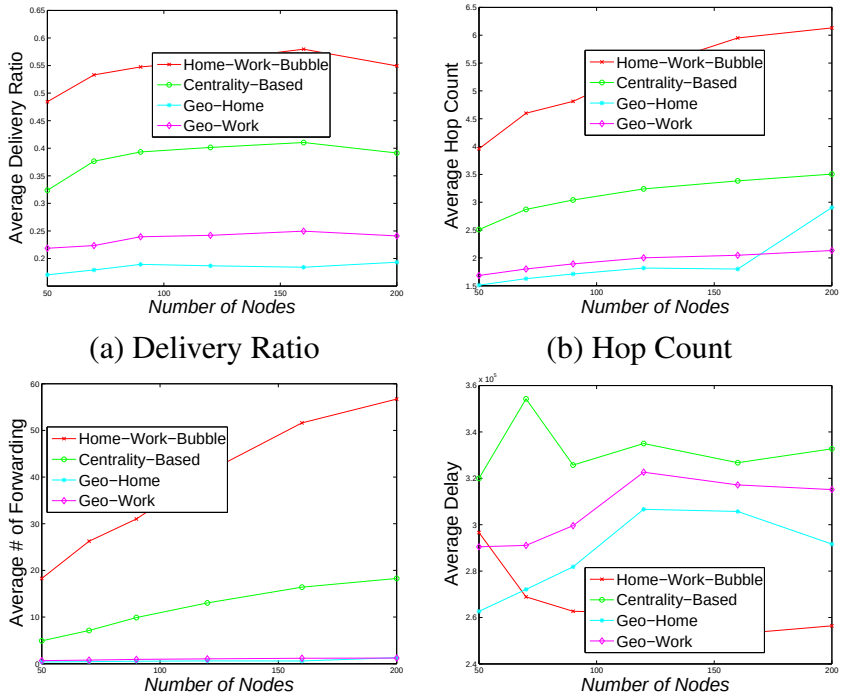


Figure 4.17: Simply combining # of common home&work-towers:poor performance.

We also evaluated these protocols in multi-copy model on the Setting B of D4D Challenge Dataset. The result in Figure 4.17(a) shows that the successful delivery ratio of the Home-Work algorithm is worse than Geo-Home and Geo-Work algorithm. The reason is similar with the one in our first experiment. Although the node which both has common top 5 home tower and top 5 work towers with the destination node could be a good relay, the number of such kind of nodes is limited. There is very low possibility that a user both work and live at the same place with the destination user. So making too strict requirement on relay node selection is not a good way.

Therefore, again, it is an interesting research challenge regarding how to smartly make use of the advantage from both location-based and social-based metrics to improve the routing performance.

The good way to take advantage of multiply metrics (various information source) is not



(a) Delivery Ratio (b) Hop Count

(c)# of Forwarding (d) Delay

Figure 4.18: Smartly use multiply metrics significantly improve the routing performance.

only use them to look for a perfect relay node. It is also important to use appropriate metrics in appropriate situation. Bubble rap forwarding in [40] is a good example. In its bubble-up phase, it first bubbles the message up based on the global centrality to increase the relay node's opportunity to encounter with the more ideal relay node (the node which is in the same local community as the destination). After the first phase, since the current relay node already have relatively high probability to meet with the destination node, its relay selection rule became more rigid (it only choose the node with higher local centrality). Inspired by this observation, we proposed our new location-social based routing algorithm which jointly use *number of common home-towers*, *number of common work-towers*, *home degree centrality* and *work degree centrality*, We call it Home-Work-Bubble.

The basic idea of this algorithm is when the current node doesn't have any common *top 5 home/work- towers* with the destination, which means the current relay node does not have close home or work place with the destination, the source relay will choose the node

with higher *home/work degree centrality* or higher total *number of common home and work towers* with the destination as the relay. Until the current relay meet with the node, which has common *top 5 home/work towers* with the destination, it begins the second stage. Since then, it only relays message to the node, which has higher total *number of common home and work towers* with the destination. It's detailed steps are as follows.

Step 1 (Calculate the Top 5 Home/Work Towers): Using the user's cell tower scan records, find the *top 5 home/ work- towers* .

Step 2 (Construct the Home/Work Contact Graph): Create matrix D_H to represent the users' home contact graph, D_W to represent the users' work contact graph. For each $d_{ij} \in D_H$, $d_{ij} = 1$ if *the number of common home-towers* between user v_i and v_j is larger than 2, otherwise $d_{ij} = 0$. Similarly, for each $d'_{ij} \in D_W$, $d'_{ij} = 1$ if *the number of common work-towers* between user v_i and v_j is larger than 2, otherwise $d'_{ij} = 0$.

Step 3 (Calculate the Home/Work Degree Centrality): Calculate the *home/work degree Centrality* for each user.

Step 4 (Make Routing Decision): When a user v_i encounters another user v_j, if user v_i doesn't have any common *top 5 home/work-towers* with the destination node, v_i choose v_j as the relay node if v_j have common *top 5 home/work-towers* with the destination or the home/work degree centrality of node v_j is larger than node v_i. If user v_i already have common *top 5 home/work- towers* with the destination node, v_i choose v_j as the relay node only if v_j have larger total *number of common home/work- towers* with the destination node than v_i have .

We evaluate the performance of Home-Work-Bubble algorithm on setting B of D4D Challenge Dataset in multi-copy model and limit the number of copies by 20. We compare the simplified location-based algorithm with three algorithms(Geo-Home, Geo-Work and Centrality-Based) using the following four metrics: average successful delivery ratio, average hop count, average number of forwarding, average delay. The result in Figure 4.18 shows that Home-Work-Bubble algorithm has much better successfully delivery ratio(about

two times) than the other three algorithms.

4.5 Time Aware Location-Social Based Routing

By analysis user's access frequency on cell phone towers in different timeslot, we found that people has different preferences in different timeslot. Considers people's natural activity features, we divide a day into four timeslot: 12am∼8am, 8am∼1pm, 1pm∼6pm, 7pm∼12am. Table 4.13 lists a D4D user's top 5 frequently accessed towers in four timeslots respectively. We can see that in each timeslot, his/her access preferences have big difference.

Table 4.13: User 3061's top 5 towers in four time slots

TimeSlot	Top1	Top2	Top3	Top4	Top5
12am∼8am	628	455	311	727	707
8am∼1pm	86	925	772	455	38
1pm∼6pm	960	1020	628	707	727
6pm∼12am	38	1020	568	707	960

Enlightened by this observation, we believe accurately establish specific routing strategies for different timeslot may improve the routing performance. We first extend all of the metrics we use for location-social based DTN routing into timeslot aware version. The definitions of these metrics are exactly the same as them in the whole time period, the only difference is the metrics for each timeslot only considers the contact or cell tower scan records happen in that timeslot. Then we compare two pair of DTN routing methods, with and without consideration of timeslots on setting B of D4D Challenge Dataset in multi-copy model and limit the number of copies by 20.

We first compares the betweenness centrality based algorithm with and without the timeslot consideration:

Betweenness Centrality Based: Calculation the betweenness centrality use the contact graph get from the *top 10 towers* of users. (If the two users have more than 3 common top towers, they have a link in the contact graph). When a node v_i encountered another node v_j. If the node v_j has larger betweenness centrality than node v_i, forward the message.

Timeslot Aware Betweenness Centrality Based : Calculation the betweenness centrality use the contact graph get from the *top 5 towers* of users in each timeslot. (If the two users have more than 2 common top towers during a timeslot, they have a link in the contact graph of that timeslot). Then each node has four different betweenness centralities pair with the four timeslots. When a node v_i encountered another node v_j. v_i first look at the the timeslot encounter happens in which timeslot. If the node v_j has larger betweenness centrality than node v_i in that timeslot, forward the message.

We also compares the simplified location-based algorithm with and without the timeslot consideration:

Simplified Location-Based: When a node v_i encountered another node v_j. If the node v_j has larger *number of common top towers* with the destination node than node v_i, forward the message.

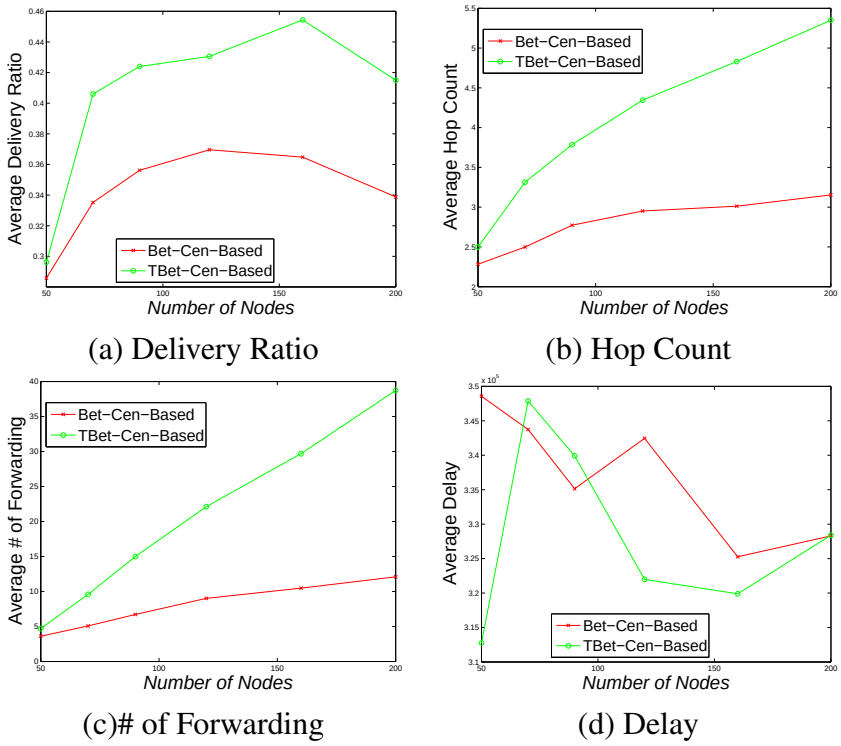
Timeslot Aware Simplified Location-Based : When a node v_i encountered another node v_j. Check the encounter happens in which timeslot. If the node v_j has larger *number of common top towers* with the destination node than node v_i in that timeslot, forward the message.

The result in Figure 4.19 and Figure 4.20 shows that timeslot aware algorithms has much better successfully delivery ratio than their original ones.

Since the timeslot aware methods achieve higher successful delivery ratio, we extent our Home-Work-Bubble algorithm into the timeslot aware version. For this algorithm we only have two timeslots: day time (8am 7pm) and night time (7pm 12am & 12am 8am). It's detailed steps are as follows:

Step 1 (Calculate the Top 5 Home/Work Towers): Using the user's cell tower scan records, find the *top 5 home/ work- towers* for each user.

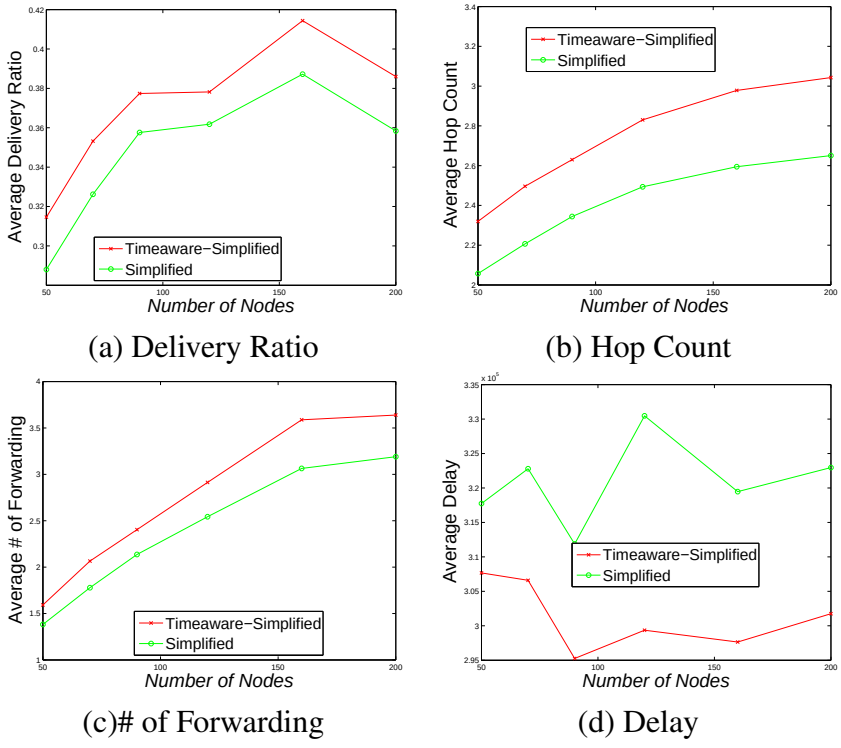
Step 2 (Construct the Home/Work Contact Graph): Create matrix D_H to represent the users's home contact graph, D_W to represent the users' work contact graph. For each $d_{ij} \in D_H$, $d_{ij} = 1$ if *the number of common home-towers* between user v_i and v_j is larger



(a) Delivery Ratio (b) Hop Count

(c)# of Forwarding (d) Delay

Figure 4.19: The timeslot aware bet-centrality algorithm achieves higher delivery ratio.



(a) Delivery Ratio (b) Hop Count

(c)# of Forwarding (d) Delay

Figure 4.20: Timeslot aware simp. location-based algorithm achieves higher delivery ratio.

than 2, otherwise $d_{ij} = 0$. Similarly, for each $d'_{ij} \in D_W$, $d'_{ij} = 1$ if *the number of common work-towers* between user v_i and v_j is larger than 2, otherwise $d'_{ij} = 0$.

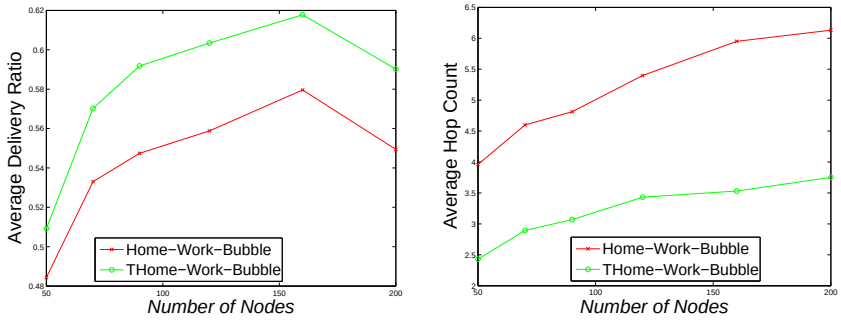
Step 3 (Calculate the Home/Work Degree Centrality): Calculate the *home/work degree Centrality* for each user.

Step 4 (Make Routing Decision): When a user v_i encounters another user v_j, v_i first check the encounter happens in which timeslot. If the encounter happens during day time, v_i aims to looking for relay node close to destination's work place. If v_i doesn't have any common *top 5 work- towers* with the destination node, v_i choose v_j as the relay node when v_j has common *top 5 work- towers* with the destination node or v_j has higher work degree centrality. If user v_i already have common *top 5 work- towers* with the destination node, v_i choose v_j as the relay node only if v_j have larger *number of common work-towers* with the destination node than v_i have . If the encounter happens during night time, on the contrary v_i aims to looking for relay node close to destination's home. If v_i doesn't have any common *top 5 home- towers* with the destination node, v_i choose v_j as the relay node when v_j has common *top 5 home- towers* with the destination node or v_j has higher home degree centrality. If user v_i already have common *top 5 home-towers* with the destination node, v_i choose v_j as the relay node only if v_j have larger *number of common home-towers* with the destination node than v_i have .

We evaluate the performance of Home-Work-Bubble algorithm with and without the timeslot consideration on setting B of D4D Challenge Dataset in multi-copy model and limit the number of copies by 20. The result in Figure 4.21 shows that timeslot aware algorithms has better successfully delivery ratio.

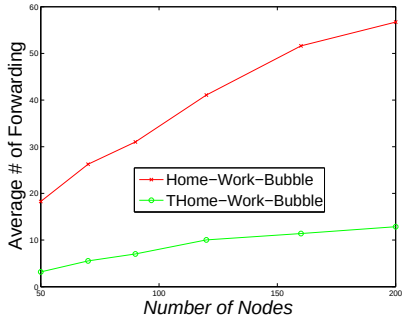
4.6 Summary

In this chapter, we study the problem of how to design the DTN routing algorithms using social and location based methods. We first prove the location features are useful on DTN routing. Then we explore methods to predict the location's semantic meaning. Extend the experimental results into location social based routing, we proposed several location-social

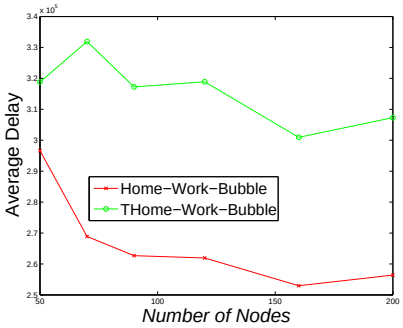


(a) Delivery Ratio

(b) Hop Count



(c)# of Forwarding



(d) Delay

Figure 4.21: The timeslot aware Home-Work-Bubble algorithm has higher delivery ratio.

based metrics and proposed good ways to design DTN routing algorithms with them. And we evaluated all proposed routing methods via extensive simulations with real life trace data (such as MIT reality, Nokia MDC, and Orange D4D).

CHAPTER 5: DTN ROUTING WITH THROW-BOXES

In chapter 4, we proposed location and social based DTN routing algorithms to improve the DTN routing performance. However due to the intermittent connectivity characteristics of the DTNs, especially for some sparse networks like D4D Challenge Dataset, the successful package delivery ratio is still low. To further increase the successful delivery ratio and reduce the communication latency in DTN, we use throw-boxes to increase the contact opportunities in DTN and improve the network performances.

Throw-boxes are small and inexpensive devices equipped with wireless interfaces and storage. They are stationary, and they relay data between mobile nodes in a store-and-forward way, thus they can operate without communication with other throw-boxes. They are especially helpful when the network users move within a region of network field (e.g. In D4D Challenge Dataset, most of users move around several towers.).

In this chapter, we will first explore different network models using throw-boxes, and discuss how the number of throw-boxes could effect the routing performance. Then, we will propose different methods to choose the locations of throw-boxes.

5.1 Related Works

A few research works have already considered routing in DTNs equipped with throw-boxes. To the best of my knowledge, [101] is the first work in this area. The authors investigated the gain on the network throughput from deploying throwboxes. They investigated three different deploy scenarios, and for each scenario, evaluated three different relay schemes: an epidemic, a single path and a multi-path relay scheme. They showed that to maximize the throw-boxes' effectiveness, their placement should be considered simultaneously with the routing algorithm. They provided guidelines on the design and deployment

of throwboxes in DTNs.

In [42], Ibrahim *et al.* studied the impact of adding throwboxes for two common relay protocols: the epidemic protocol and a multi-copy two-hop relay protocol. Results have shown that, under the epidemic protocol, the network performance increases greatly in the presence of throwboxes at the cost of increasing the number of copies in the network. On the other hand, only marginal improvements can be expected when the use of throwboxes is restricted to relay data to their corresponding destinations and not to intermediary nodes. In [43], they proposed and evaluated several relay strategies to minimize the resources consumption. They introduced a framework to calculate the main performance of each relay scheme under Markovian assumptions.

In [2], Banerjee *et al.* concluded that, due to possible energy constraints, deploying throw-boxes need to consider the energy efficiency. They proposed an energy-efficient hardware and software architecture for throwboxes. And gave an approximate heuristic for solving the NP-hard problem of meeting an average power constraint while maximizing the number of bytes forwarded by it. In [3], they studied infrastructure-enhanced (base stations, relays, mesh nodes) mobile networks in the context of vehicular networks. They observed that deploying x base stations can reduce the average packet delivery by a factor of two and that the same reduction requires 2x mesh nodes and 5x relays. Another conclusion is that deploying small infrastructures is superior to using ferry nodes. They also complemented their experimental work with an analytical model of large-scale networks in the presence of infrastructure and for different spreading (relay) protocols.

In [44], Joel *et al.* studied the performance impact of deploying stationary relay nodes on two different application scenarios for vehicular opportunistic networks.

In [31] and [32], Bo Gu *et al.* introduced a capacity-aware routing protocol which aims to search the shortest path with the consideration of time-varying delay and capacity of virtual link. Markov Chain is used to model the evolution of the link delay and capacity.

5.2 Network Model Using Throw-boxes

We now describe the characteristics of throw-boxes and present the network models we consider. Throw-boxes can be used in variety of scenarios. Here, we focus on the use of throw-boxes in mobile DTNs, where the throw-boxes and nodes could communicate with each other when they are in the transmission range of each other. The throw-boxes are stationary, and we only consider the locations of the cellphone towers as their candidate displacement locations. The reason we have this assumption is, these locations (the locations of the cellphone towers) could be the hot spots of the network nodes (From Table 4.1, we can find a network node mainly moves around several cellphone towers.), and since we can get cell tower access record from real tracing data, it's easy for us to get the network location and social characteristics of these places for our analysis.

For a normal mobile user in DTN, we extract its location and social characteristics from its contact information (device to device Bluetooth scan record or synthesized device to device contact record) and cellphone tower access record. In throw-box scenario, the location and social characteristics of normal mobile users are extracted from the same resource and we synthesize the similar contact information and the cellphone tower access record of throw-boxes. We treat the network's cellphone tower access record as the contact information of the throw-boxes, which means throw-boxes only have contact record with the mobile users (which throw-box contact to which user at what time). We use the network's cellphone tower access record and the mobile user's *top 10 towers* to synthesize the cellphone tower access record, which reflect a two hop relation: throw-box to mobile user and then to the user's top 10 towers. For example, if we have a cellphone tower record (v_i, t_j, T_{enc}), which describes that node v_i contacted to cellphone tower t_j at T_{enc}, and the *top 10 towers* of node v_i are $(t_{i1}, t_{i2}, \cdots, t_{i10})$, we will synthesize ten cellphone tower access record for throw-box located at tower t_j. They are $(t_j, t_{i1}, T_{enc}), (t_j, t_{i2}, T_{enc}), \cdots, (t_j, t_{i10}, T_{enc})$. This synthesized record indicate that a at particular time, the messages on a throw box could be carried by mobile users who has high probability to appears around some partic-

ular cellphone towers.

Now we can extend our definition of location and social characteristic for network nodes (normal mobile users) into the throw-box scenario. With contact information and cellphone tower access record available for both network mobile nodes and throw-boxes, most of the definitions don't need to change. We list the modified ones as follows:

The mobile user's centrality: The mobile user v_i's centrality is the total number of throw-boxes and users, which v_i contact to.

The throw-box's centrality: The throw-box t_k's centrality is the number of users it contact to.

5.2.1 Different Communication Models

With our modified location and social characteristic in throw-boxes scenario and our proposed location and social based DTN routing algorithms. We compare three different mobile nodes and throw-boxes communication models:

Model A: We treat the throw-boxes exactly the same as the mobile users in our routing algorithms. Which means neither the throw-boxes nor the mobile users could hold the message permanently. When an encounter happens, between two mobile users or user and throw-box, the message always will forward to the one with higher evaluated probability to meet the destination node in our proposed routing algorithms. And for a message, it's total number of copies in the whole network (include ones on both throw-boxes and mobile users) should not be more than N_{max}. If the total number of copies for a message is larger than N_{max}, after forwarding, the original one will be delete.

Model B: We treat the throw-boxes and the mobile users differently. The mobile users in the network could hold no more than N_{max} copies of a message. However there is no constraint on copies on throw-boxes. When a mobile user encounters a throw-box, it will give a copy to the throw-box, the throw-box will keep the copy permanently (We assume it can hold long enough until the routing is finished), and the mobile user will still keep the copy on itself. The throw-boxes here are not allowed to forward message copy to mobile

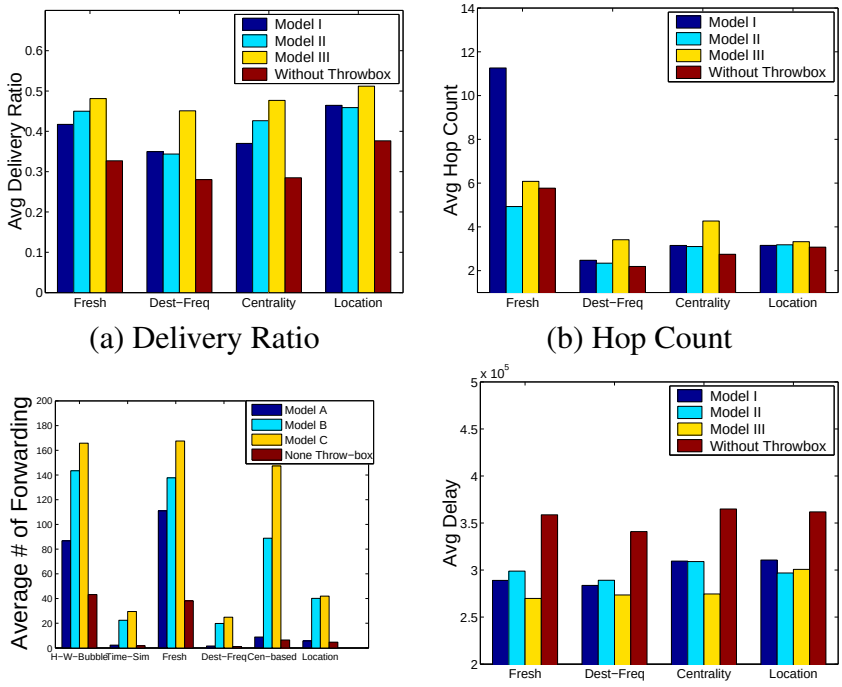
users except the destination node.

Model C: The same as in model B, there is no constraint on copies on throw-boxes. The difference is the throw-boxes are allowed to forward message copy to mobile users. If the total copies on all the mobile users is less than N_{max}, and the throw-box evaluated that the mobile user has higher probability to meet the destination node than itself, it will give a copy to this mobile user.

We evaluate these three different communication models on setting B of the D4D Challenge dataset. We applied these three models on four routing algorithms: Fresh, Destination Frequency, Centrality-based and Location-based. And compare their routing performance with the one without throw-boxes using the following four metrics: average successful delivery ratio, average hop count, average number of forwarding, average delay. For each model, we pick 100 nodes and 20 throw-boxes to participate the opportunistic communications and we allow total 10 message copies on mobile users. The 20 throw-boxes location are selected from total 268 candidate cellphone tower locations using method A in the next section. Figure 5.1 illustrates the results. We can see that all of our proposed models have higher average successful delivery ratio than the one without throw-boxes. Model C achieves the highest successful delivery ratio, the smallest delay with the cost of largest number of forwarding. So there is always a tread off between the routing performance and the communication cost.

5.2.2 Number of Throw-boxes

We also study the effect of the number of throw-boxes using in the networks by evaluation on setting B of the D4D Challenge dataset. We applied Model C on Four routing algorithms: Fresh, Destination Frequency, Centrality-based and Location-based. For each model, we pick 100 nodes and 5 to 70 throw-boxes to participate the opportunistic communications and we allow total 10 message copies on mobile users. The throw-boxes location are selected from total 268 candidate cellphone tower locations. We compare their routing performance using the following four metrics: average successful delivery ratio, average



(a) Delivery Ratio (b) Hop Count

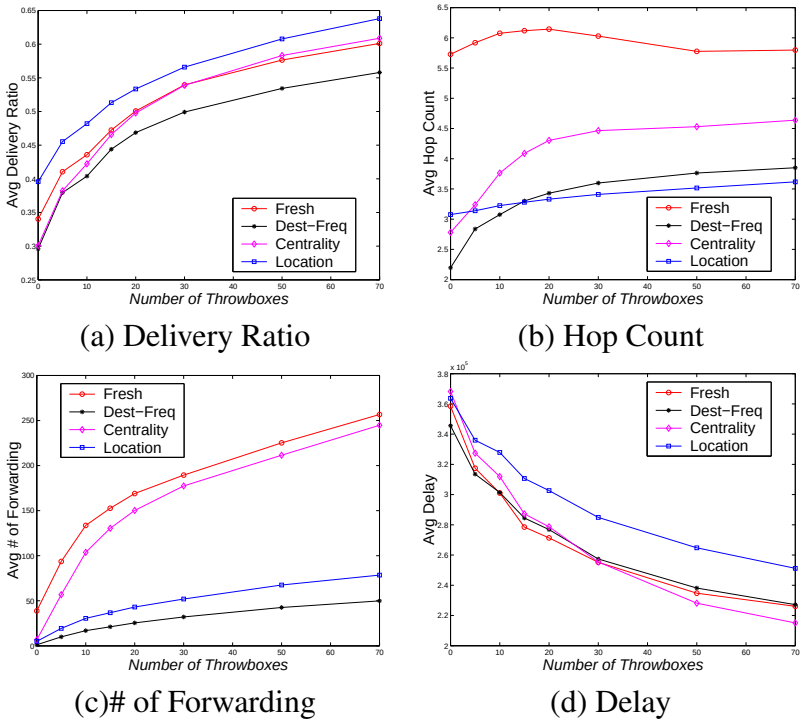
(c)# of Forwarding (d) Delay

Figure 5.1: Different network models comparison.

hop count, average number of forwarding, average delay. Figure 5.2 illustrate the number of throw-boxes comparison with throw-box selection Method A in the next section. and Figure 5.3 illustrate the number of throw-boxes comparison with random throw-box selection. We can see that in both scenarios, the successful delivery ratio increases as the number of throw-boxes increase. With smart throw-box location selection (Method A), the successful delivery ratio increase faster than randomly choose the throw-box location, especially at the beginning of throw-box amount increasing.

5.3 Throw-boxes Location Selection

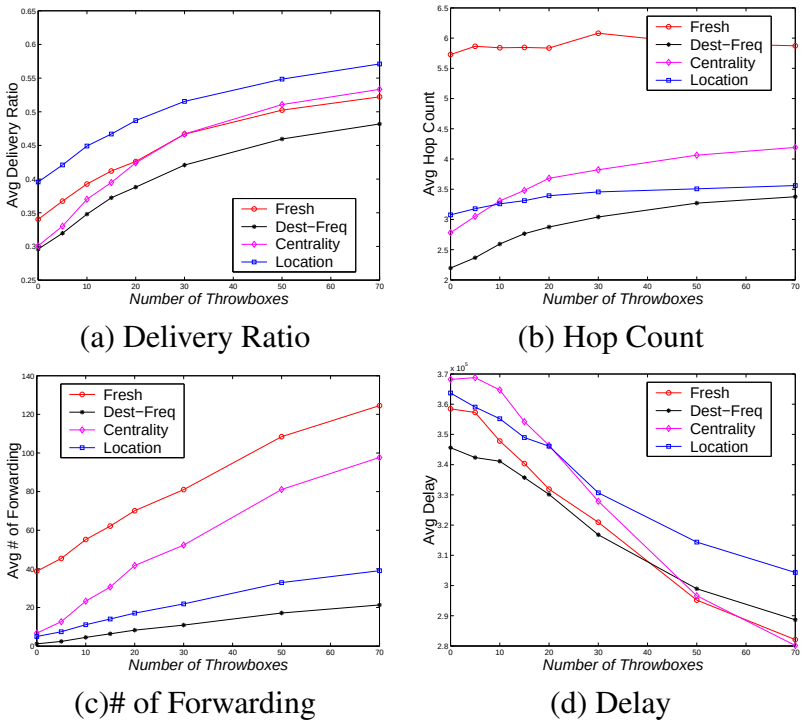
In last section, we already demonstrated that, the successful delivery ratio increase faster with carefully throw-box location selection than randomly choose the throw-box location, especially at the beginning of throw-box amount increasing. So, it is meaningful to consider how to choose the appropriate throw-box locations to maximum the benefits of throw-box, especially when there is only small amount of throw-boxes available. In this



(a) Delivery Ratio (b) Hop Count

(c)# of Forwarding (d) Delay

Figure 5.2: Number of throw-boxes comparison on Model C with throw-box selection Method D.



(a) Delivery Ratio (b) Hop Count

(c)# of Forwarding (d) Delay

Figure 5.3: Number of throw-boxes comparison on Model C with random throw-box selection.

section we study the throw-box location selection problem. For a DTN network, which has N mobile users $V = \{v_1, v_2, \cdots, v_N\}$, we set all the *top 10 towers* of these nodes $T = \bigcap_{i=1}^{N} \{t_{i1}, t_{i2}, \cdots, t_{i10}\}$ as the candidate throw-box locations. We aim to select M throw-box locations from T, such that the network's routing performance(average successful delivery ratio) is maximized.

Here, we propose throw-box selection methods by considering the user and the candidate throw-box location's popularity (degree centrality) and importance on routing (betweenness centrality). We believe that the cellphone tower locations which have contact with a lot of users or which occur on many shortest paths between mobile nodes could be good throw-box location candidate. We also believe that the important towers (*the top 10 tower*) of important users(user with high popularity or betweenness centrality) may have large contribution on routing performance. From these observation, we have the following five throw-box selection methods:

Method A: In this method, we only considers the cellphone tower locations' degree centrality. We define the candidate cellphone tower location's degree centrality as the number of mobile users the cellphone tower act as their *top 10 towers*. And we choose the cellphone tower locations with the top M degree centrality as the throw-box locations.

Method B: In this method, we only considers the cellphone tower locations' betweenness centrality. We virtually add a throw-box on all the candidate cellphone tower locations, and calculate their betweenness centrality in the network. Then we choose the cellphone tower locations with the top M betweenness centrality as the throw-box locations.

Method C: In this method, we only considers the users' betweenness centrality. We calculate the users' betweenness centrality in the network. Then we choose the M throw-box locations by first choose all the *top 10 towers* of the mobile user which has the highest betweenness, then choose all the *top 10 towers* of the mobile user which has the second highest betweenness . keep going, until we choose enough throw-boxes.

Method D: In this method, we considers both the cellphone tower locations' degree

centrality and the users' degree centrality. We first normalized the user v_i's degree centrality as c_i between 0 and 1. Then we define a metric to describe the candidate cellphone tower location's popularity among the mobile users with high degree centrality. We name it Centre-Centre. The Centre-Centre of the candidate cellphone tower location t_j is $\sum_{i \in U} c_i$, where U is the set of mobile users whose *top 10 towers* include t_j.

Method E: In this method, we considers both the cellphone tower locations' degree centrality and the users' betweenness centrality. We first normalized the user v_i's betweenness centrality as b_i between 0 and 1. Then we define a metric to describe the candidate cellphone tower location's popularity among the mobile users with high betweenness centrality. We name it Centre-Between. The Centre-Between of the candidate cellphone tower location t_j is $\sum_{i \in U} b_i$, where U is the set of mobile users whose *top 10 towers* include t_j.

We compares the five throw-boxes location selection methods using by evaluation on setting B of the D4D Challenge dataset. We applied Model C on four routing algorithms: Fresh, Destination Frequency, Centrality-based and Location-based. For each model, we pick 100 nodes and 5 throw-boxes to participate the opportunistic communications and we allow total 10 message copies on mobile users. We selected the throw-box location from total 268 candidate cellphone tower locations using method A to E respectively. We compare their routing performance with the random selection using the following four metrics: average successful delivery ratio, average hop count, average number of forwarding, average delay. Here we considers two different scenarios:

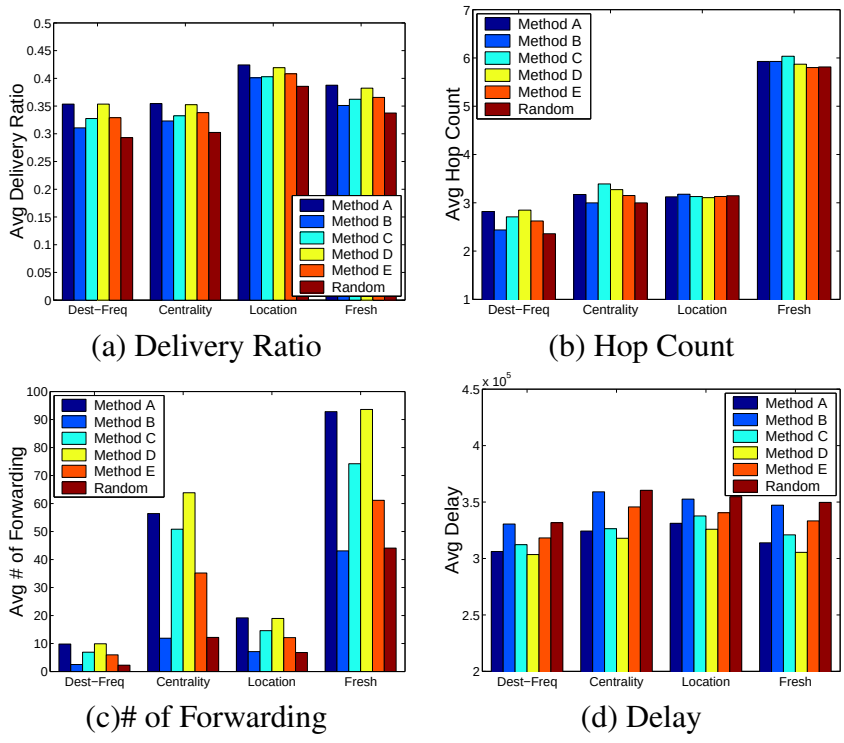
Scenario1: The whole network is well connected in one component;

Scenario2: The network has several separate components, which are not strongly connected to each other.

Figure 5.4 and Figure 5.5 illustrate the results on Scenario 1 and Scenario 2, respectively. We can see that in Scenario1, all of our social-based methods have similar successful delivery ratios which are higher than that of random deployment. Among the five methods, Method A and Method D have the slightly better delivery ratios in most of routing meth-

ods. Method B, which considers the locations' betweenness centrality, has significantly less number of forwarding than other methods. This may due to that putting throwboxes at "bridge" locations (locations with high betweenness centrality in the social graph G) reduces unnecessary forwardings among throwboxes and mobile users. In Scenario2, the performances are much poorer than those in the previous simulations since the connectivity between two components are very loose. Now Method B and Method E, which considers betweenness centrality, have better successful delivery ratio than others. This is mainly because the locations selected by these two methods can act as "bridge" nodes to connect the separate components.

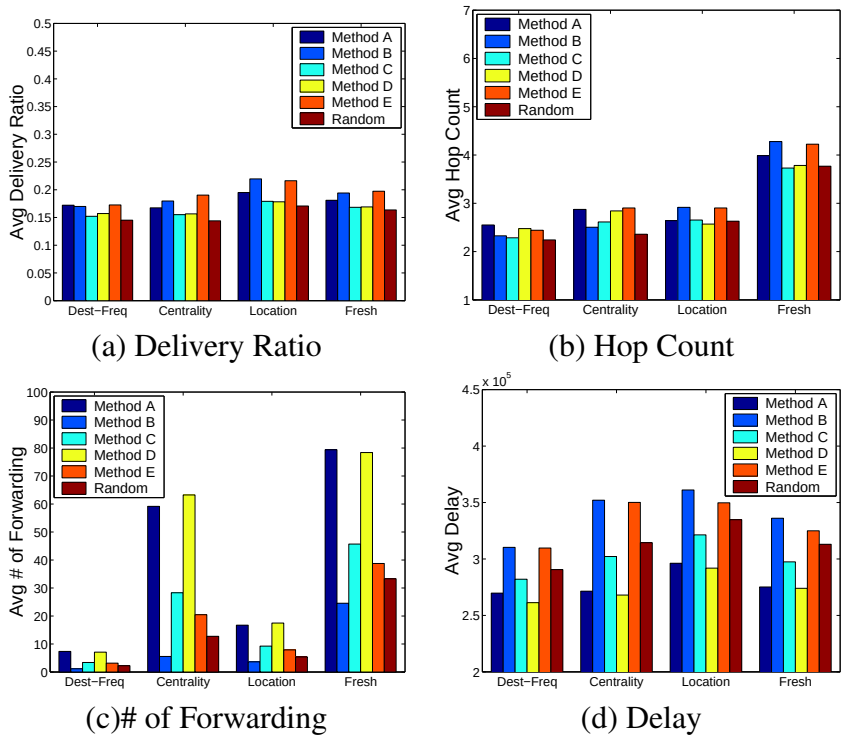
Overall, our proposed social-based methods can indeed improve the performances for all routing methods by smartly pick the locations of deployed throwboxes.



(a) Delivery Ratio (b) Hop Count

(c)# of Forwarding (d) Delay

Figure 5.4: Performance comparison of different throwbox placement schemes, Scenario1.



(a) Delivery Ratio (b) Hop Count

(c)# of Forwarding (d) Delay

Figure 5.5: Performance comparison of different throwbox placement schemes, Scenario2.

5.4 Summary

In this chapter, we study the usage of throw-boxes in DTN routing. We first explore different network models using throw-boxes, and discuss how the number of throw-boxes could effect the routing performance and then propose different methods to chose the locations of throw-boxes. Our experiment results proved that the successful delivery ratio increases as the number of throw-boxes increase. With smart throw-box location selection, the successful delivery ratio increase faster than randomly choose the throw-box location, especially at the beginning of throw-box amount increasing. For the throw-box location selection, our proposed social-based methods can indeed improve the performances for all routing methods by smartly pick the locations of deployed throwboxes.

CHAPTER 6: CONCLUSION

We briefly summarized our completed work and future work.

6.1 Summary

We make the following contributions:

- We proposed the geo-similarity metric to measure how similar the two users' access regulations are, based on their historic tracing data. According to *geo-similarity*, we proposed a location-based routing algorithm (both singlecopy and multicopy version) and showed our location-based algorithms could achieve the acceptable delivery ratio in both MIT and D4D datasets.
- We proposed a simplified location-based routing algorithm, which uses the new metric the *number of common top towers* instead of *geo-similarity* to make routing decisions. We demonstrated that the simplified metric do capture the main characteristics of users location features while reducing the communication overhead.
- We studied how to predict semantic meaning of the important places using Nokia MDC dataset. Both rule-based and machine learning based methods are proposed. Both methods can produce good accuracy for home and work location.
- We proposed several location-social based metrics to extract users' social features from location information. We showed that, simple combination of using multiply metrics may result in poor performance. The good way to take advantages of multiply metrics is not only use them to look for a perfect relay node. It is more important to use appropriate metrics in appropriate situation.
- We proposed the Home-Work-Bubble routing algorithm which jointly use *number of common home-towers*, *number of common work-towers*, *home degree centrality*

and *work degree centrality*. We showed that it achieve much better performance than purely use one of these metrics.

- We proposed three time aware location-social based routing protocols and we demonstrated that accurately establish specific routing strategies for different timeslot could improve the routing performance.
- We considered DTN routing with throw-boxes available scenario. We proposed three network models and proved both of them can achieve better performance than the one without throw-boxes.
- We discussed how the number of throw-boxes usage will affect the routing performance. We found that the successful delivery ratio increases as the number of throw-boxes increase. With smart throw-box location selection, the successful delivery ratio increases faster than randomly choose the throw-box location, especially at the beginning of throw-box amount increasing
- We proposed five throw-box selection algorithms, and we showed that, our proposed social-based methods can indeed improve the performances for all routing methods by smartly pick the locations of deployed throwboxes.

6.2 Future Works

From our research presented in this thesis we found that multiple social, spatial, and temporal characteristics of both individual components and network structure can affect the protocol performance in DTNs. We already successfully explored the message delivery opportunities by joint considering social, location and temporal metrics, and demonstrated in this way we have better chance to find the appropriate message transmission routine.

There are still several open problems left as our future works. We'd like to explore some more comprehensive way to design the hybrid social- and location- based routing, such as constructing a multi-level social- and location-graph to model DTNs. The relationships between individual devices (i.e., humans) could be very complex: social attributes, such as age groups, interests, membership of organizations and working relationships; Location

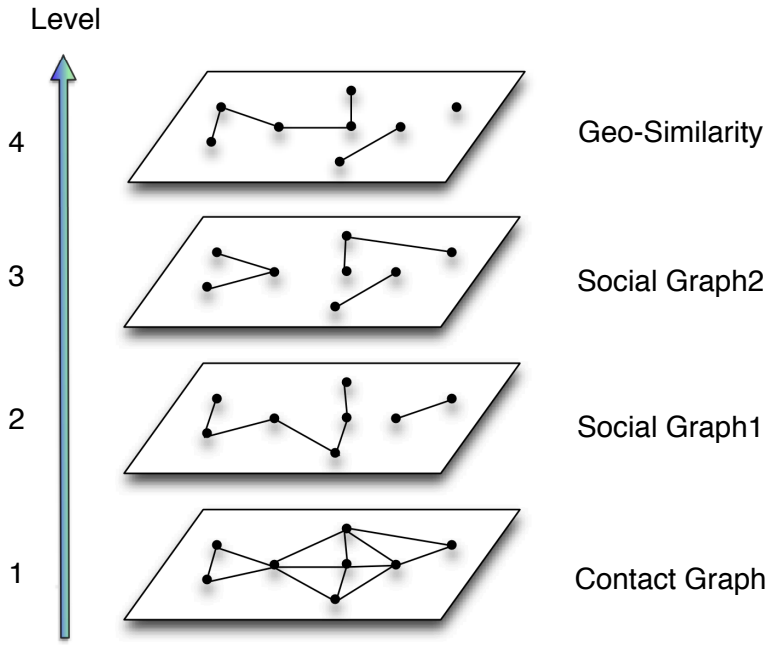


Figure 6.1: Multi-level graphs for modeling different social/location characteristics in DTNs.

attributes, such as geo-similarity, common live area, common workplace. We are going to model this relationships with a multi-level social- location-graph as shown in Figure 6.1. Different levels of graph represent different social/location relationships among user in the network, which could abstracted from different data resources. For instance, the first level may be the contact graph constructed from Buletooth contact records; the second level could be the call history graph from call and message logs, the third level could be a social graph constructed from a social network website, the fourth level could be a geo-similarity graph constructed from the cell phone tower access records, and so on.

We plan to use two kinds of routing strategy with this multi-level graph.

- When a relay node encounters another node, evaluate the successful forwarding probability of both the current node and the encountered node on each levels, compare their aggregated successful forwarding probability to make routing decision.
- When a relay node encounters another node, evaluate the successful forwarding probability of both the current node and the encountered node on each levels, if the en-

countered node has better successful forwarding probability than the current node on more than one level, we choose the encountered node as the forwarder. In this way, we may not find the “best” relay, but we catch the opportunities more quickly and frequently.

REFERENCES

[1] E. Bakshy, B. Karrer, and L. Adamic. Social influence and the diffusion of user-created content. In *Proceeding of the tenth ACM conference on Electronic commerce*, 2009.

[2] N. Banerjee, M. D. Corner, and B. N. Levine. An energy-efficient architecture for dtn throwboxes. In *Proc. IEEE Infocom*, 2007.

[3] N. Banerjee, M. D. Corner, D. Towsley, and B. N. Levine. Relays, base stations, and meshes: Enhancing mobile networks with infrastructure. In *Proceedings of ACM Mobicom*, 2008.

[4] A. Beach, M. Gartrell, S. Akkala, J. Elston, J. Kelley, K. Nishimoto, B. Ray, S. Razgulin, K. Sundaresan, B. Surendar, M. Terada, and R. Han. WhozThat? evolving an ecosystem for context-aware mobile social networks. *IEEE Network*, 22(4): 50–55, 2008.

[5] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. *Data for development: The D4D challenge on mobile phone data*. 2013.

[6] P. Bose, P. Morin, I. Stojmenovic, and J. Urrutia. Routing with guaranteed delivery in ad hoc wireless networks. In *3rd int. Workshop on Discrete Algorithms and methods for mobile computing and communications*, 1999.

[7] E. Bulut and B. K. Szymanski. Friendship based routing in delay tolerant mobile social networks. In *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM)*, 2010.

[8] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine. Maxprop: Routing for vehicle-based disruption-tolerant networks. In *Proceedings of the 25th IEEE International Conference on Computer Communications (INFOCOM)*, 2006.

[9] J. Burgess, B. N. Levine, R. Mahajan, J. Zahorjan, A. Balasubramanian, A. Venkataramani, Y. Zhou, B. Croft, N. Banerjee, M. Corner, and D. Towsley. Crawdad data set umass/diesel (v. 2008-09-14). *Downloaded from* $http$ *://crawdad.cs.dartmouth.edu/umass/diesel*, 09 2008.

[10] B. Burns, O. Brock, and B. N. Levine. Mv routing and capacity building in disruption tolerant networks. In *Proceedings of the 24th IEEE International Conference on Computer Communications (INFOCOM)*, pages 398–408, 2005.

[11] N. R. Center. Mobile data challenge (mdc). In $http$ *://research.nokia.com/page/12000*. 2012.

[12] A. Chaintreau, P. Fraigniaud, and E. Lebhar. Opportunistic spatial gossip over mobile social networks. In *WOSN '08: Proceedings of the first ACM Workshop on Online Social Networks*, pages 73–78, 2008.

[13] S. Y. Chan, P. Hui, and K. Xu. Community detection of time-varying mobile social networks. In *Proc. of the 1st International Conference on Complex Sciences: Theory and Applications (Complex 2009)*, February 2009.

[14] F. Chierichetti, S. Lattanzi, and A. Panconesi. Gossiping (via mobile?) in social networks. In *DIAL M-POMC '08: Proceedings of the 5th ACM international workshop on Foundations of mobile computing*, pages 27–28, 2008.

[15] E. M. Daly and M. Haahr. Social network analysis for routing in disconnected delay-tolerant manets. In *MobiHoc '07 Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*, 2007.

[16] L. Danon, J. Duch, A. Arenas, and A. Daz-guilera. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 9008: 09008, 2005.

[17] Z.-B. Dong, G.-J. Song, K.-Q. Xie, and J.-Y. Wang. An experimental study of large-scale mobile social network. In *Proc. of the 18th International World Wide Web Conference (WWW2009)*, 2009.

[18] H. Dubois-Ferriere, M. Grossglauser, and M. Vetterli. Age matters: efficient route discovery in mobile ad hoc networks using encounter ages. In *Proc. of ACM MobiHoc*, 2003.

[19] V. Erramilli and M. Crovella. Diversity of forwarding paths in pocket switched networks. In *Proc. of ACM IMC*, 2007.

[20] V. Erranmilli, M. Crovella, A. Chaintreau, and C. Diot. Delegation forwarding. In *Proc. of ACM MobiHoc*, 2008.

[21] M. G. Everett and S. P. Borgatti. Analyzing clique overlap connection. *Connections*, 21(1):49–61, 1998.

[22] F. Fabbri and R. Verdone. A sociability-based routing scheme for delay-tolerant networks. *EURASIP Journal on Wireless Communications and Networking*, January 2011.

[23] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[24] L. C. Freeman. Centrality in social networks: conceptual clarification. *Social networks*, 1(3):215–239, 1979.

[25] W. Gao and G. Cao. Fine-grained mobility characterization: Steady and transient state behaviors. In *Proceedings of the 11th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, 2010.

[26] W. Gao and G. Cao. User-centric data dissemination in disruption tolerant networks. In *Proceeding s of the 30th IEEE Conference on Computer Communications (INFOCOM)*, 2011.

[27] W. Gao, Q. Li, B. Zhao, and G. Cao. Multicasting in delay tolerant networks: a social network perspective networks. In *MobiHoc '09: Proceedings of the 10th ACM international symposium on Mobile ad hoc networking and computing*, 2009.

[28] W. Gao, J. Chen, J. Fan, Y. Du, and Y. Sun. Geography-aware active data dissemination in mobile social networks. In *Proceeding of Mobile Ad Hoc and Sensor Systems (MASS)*, 2010.

[29] M. Gerla, P.-C. Cheng, K.-C. Lee, and J. Harri. Geodtn+nav: Geographic dtn routing with navigator prediction for urban vehicular environments. *Mobile Networks and Applications*, 15(1):61–82, 2010.

[30] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci*, 99(12):7821–7826, 2002.

[31] B. Gu and X. Hong. Capacity-aware routing using throw-boxes. In *Global Telecommunications Conference (GLOBECOM 2011)*, pages 1–5, 2011.

[32] B. Gu and X. Hong. Optimal routing strategy in throw-box based delay tolerant network. In *Communications and Networking in China (CHINACOM), 2011 6th International ICST Conference on*, pages 501–506, 2011.

[33] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11.

[34] G. A. Hillery. Definitions of community: areas of agreement. *Rural Sociology*, 20 (4):111, 1955.

[35] T. Hossmann, F. Legendre, and T. Spyropoulos. From contacts to graphs: pitfalls in using complex network analysis for dtn routing. In *INFOCOM'09: Proceedings of the 28th IEEE International conference on Computer Communications Workshops*, pages 260–265, 2009.

[36] T. Hossmann, T. Spyropoulos, and F. Legendre. Know thy neighbor: Towards optimal mapping of contacts to social graphs for dtn routing. In *INFOCOM'10: Proceedings of the 29th IEEE International conference on Computer Communications*, 2010.

[37] P. Hui and J. Crowcroft. How small labels create big improvements. In *International Workshop on Intermittently Connected Mobile Ad hoc Networks in conjunction with IEEE PerCom 2007*, pages 19–23. MarchMarch, 2007.

[38] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and the consequences of human mobility in conference environments. In *WDTN '05: Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, 2005.

[39] P. Hui, E. Yoneki, S. Y. Chan, and J. Crowcroft. Distributed community detection in delay tolerant networks. In *Proc. of ACM SIGCOMM Workshop, MobiArch'07*, 2007.

[40] P. Hui, J. Crowcroft, and E. Yonek. Bubble rap: Social-based forwarding in delay tolerant networks. In *Proc. of the 9th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, May 2008.

[41] P. Hui, E. Yoneki, and J. Crowcroft. Identifying social communities in complex communications for network efficiency. In *Proc. of the 1st International Conference on Complex Sciences: Theory and Applications (Complex 2009)*, February 2009.

[42] M. Ibrahim, A. Al Hanbali, and P. Nain. Delay and resource analysis in manets in presence of throwboxes. *Perform. Eval.*, 64(9-12):933–947, Oct. 2007.

[43] M. Ibrahim, P. Nain, and I. Carreras. Analysis of relay protocols for throwbox-equipped dtns. In *Proceedings of the 7th international conference on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, WiOPT'09, pages 222–230, Piscataway, NJ, USA, 2009. IEEE Press.

[44] V. N. G. J. S. Joel J. P. C. Rodrigues and F. Farahmand. *Stationary Relay Nodes Deployment on Vehicular Opportunistic Networks*. Auerbach Publications, 2010.

[45] D. B. Johnson and D. A. Maltz. *Dynamic source routing in ad hoc wireless networks, in Mobile Computing*. Kluwer Academic Publishers, p.153–181, T. Imielinski and H. F. Korth (Eds.), Norwood, MA, 1996.

[46] P. Juang, H. Oki, Y. Wang, M. Martonosi, L. S. Peh, and D. Rubenstein. Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences with zebranet. *SIGOPS Oper. Syst. Rev*, 36(5):96–107, 2002.

[47] B. Karp and H. T. Kung. Gpsr: Greedy perimeter stateless routing for wireless networks. In *Proc. of the ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom)*, 2000.

[48] Y.-B. Ko and N. H. Vaidya. Location-aided routing (lar) in mobile ad hoc networks. In *MobiCom '98: Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking*, pages 66–75, NY, USA, 1998.

[49] V. E. Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, 2002.

[50] C. J. Kuhlman, V. S. A. Kumar, M. V. Marathe, S. S. Ravi, and D. J. Rosenkrantz. Finding critial nodes for inhibiting diffusion of complex contagions in social networks. In *Proceeding of the 2010 European conference on Machine learning and knowledge discovery in databases: Part 2*, 2010.

[51] M. M. Lab. Reality mining project. In $http$ *://reality.media.mit.edu/*. 2005.

[52] C. Lee, H. Kwak, H. Park, and S. Moon. Finding influentials based on the temporal order of information adoption in twitter. In *Proceeding of the 19th international conference on World wide web*, 2010.

[53] J. Leguay, T. Friedman, and V. Conan. Dtn routing in a mobility pattern space. In *Proceeding of ACM WDTN*, 2005.

[54] J. Leguay, T. Friedman, and V. Conan. Evaluating mobility pattern space routing for DTNs. In *Proceedings of the 25th IEEE International Conference on Computer Communications (INFOCOM)*, SpainBarcelona, Spain, April 2006. Barcelona.

[55] A. Lindgren, A. Doria, and O. Schelén. Probabilistic routing in intermittently connected networks. *SIGMOBILE Mob. Comput. Commun. Rev*, 7(3):19–20, 2003.

[56] C. Liu and J. Wu. Scalable routing in delay tolerant networks. In *MobiHoc '07: Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*, pages 51–60, 2007.

[57] C. Liu and J. Wu. Routing in a cyclic mobispace. In *MobiHoc '08: Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing*, pages 351–360, 2008.

[58] C. Liu and J. Wu. An optimal probabilistic forwarding protocolin delay tolerant networks. In *MobiHoc '09: Proceedings of the 10th ACM international symposium on Mobile ad hoc networking and computing*, pages 105–114, 2009.

[59] P. V. Marsden. Egocentric and sociocentric measures of network centrality. *Social Networks*, 24(4):407–422, October 2002.

[60] M. Mauve, A. Widmer, and H. Hartenstein. A survey on position-based routing in mobile ad hoc networks. *in Network, IEEE*, 15(6):30–39, 2001.

[61] D. W. McMillan and D. M. Chavis. Sense of community: A definition and theory. *Journal of Community Psychology*, 14(1):6–23, 1986.

[62] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[63] A. Mei, G. Morabito, P. Santi, and J. Stefa. Social-aware stateless forwarding in pocket switched networks. In *Proceeding of the 30th IEEE Conference on Computer Communications(INFOCOM) mini-conference*, 2011.

[64] E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell. Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application. In *SenSys '08: Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 337–350, 2008.

[65] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (IMC '07)*, 2007.

[66] M. E. J. Newman. Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):321–330, 2004.

[67] M. E. J. Newman. Fast algorithem for detecting community structure in networks. *PHYS.REV.E*, 69:066133, 2004.

[68] M. E. J. Newman. Modularity and community structure in networks. In *Proceedings of the National Academy of Sciences*, no. 23, pp. 8577, vol. 103, 2006.

[69] M. J. Newman. A measure of betweenness centrality based on random walks. *Socail networks*, 27(1):39–54, 2005.

[70] U. of Helsinki. The context project. In *http ://www.cs.helsinki.fi/group/context/data*. 2005.

[71] U. of Strathclyde. Nodobo data release. In *http ://nodobo.com/release.html*. 2011.

[72] S. Okasha. Altruism, group selection and correlated interaction. *British Journal for the Philosophy of Science*, 56(4):730–725, 2005.

[73] C. Perkins and E. Royer. Ad-hoc on-demand distance vector routing. In *Proc. of the 2nd IEEE Workshop on Mobile Computing Systems and Applications*, February 1999.

[74] D. D. Perkins, P. Florin, R. C. Rich, A. Wandersman, and D. M. Chavis. Participation and the social and physical environment of residential blocks: Crime and community context. *American Journal of Community Psychology*, 18:83–115, 1990.

[75] J. A. P.G. Lind, L.R. da Silva and H. Herrmann. Spreading gossip in social networks. *Physical Review E*, 76(3), 2007.

[76] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser. Crawdad data set epfl/mobility (v.2009-02-24). *Downloaded from http ://crawdad.cs.dartmouth.edu/epfl/mobility*, February 2009.

[77] M. Piorkowski, N. Sarafijanovoc-Djukic, and M. Grossglauser. A parsimonious model of mobile partitioned networks with clustering. In *Proc. of the 1st International Conference on COMmunication Systems and NETworkS (COMSNETS)*, pages 1–10, 2009.

[78] E. Royer and C. Toh. A review of current routing protocols for ad-hoc mobile wireless networks. *IEEE Personal Communications*, 6(2):46–55, April 1999.

[79] J. Scott. *Social Network Analysis: A Handbook*. Sage Publications, 2000.

[80] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau. Crawdad data set cambridge/haggle (v. 2006-09-15). *Downloaded from http ://crawdad.cs.dartmouth.edu/cambridge/haggle*, Sept 2006.

[81] R. C. Shah, S. Roy, S. Jain, and W. Brunette. Data mules: modeling a three-tier architecture for sparse sensor networks. In *Proc. of the 1st IEEE International Workshop on Sensor Network Protocols and Applications*, pages 30–41, 2003.

[82] T. Spyropoulos, K. Psounis, and C. S. Raghavendra. Spray and wait: an efficient routing scheme for intermittently connected mobile networks. In *WDTN '05: Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, pages 252–259, 2005.

[83] V. Srinivasan, M. Motani, and W. T. Ooi. Crawdad data set nus/contact(v.2006-08-01). *Downloaded from http ://crawdad.cs.dartmouth.edu/nus/contact*, Aug 2006.

[84] J. Tang, M. Musolesi, C. Mascolo, and V. Latora. Temporal distance metrics for social network analysis. In *Proceeding of the 2nd ACM workshop on Online social networks*, 2009.

[85] J. Tang, C. M. M. Musolesi, and V. Latora. Characterising temporal distance and reachability in mobile and online social networks. *ACM SIGCOMM Computer Communication Review*, 40(1), 2010.

[86] R. L. V. D. Blondel, J. L. Guillaume and E. Lefebvre. Fast unfolding of community hierarchies in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, page 10008, 2008.

[87] A. Vahdat and D. Becker. Epidemic routing for partially connected ad hoc networks. Technical report cs-200006, Duke University, April 2000.

[88] Y. Wang, H. Dang, and H. Wu. A survey on analytic studies of delay-tolerant mobile sensor networks. *Wirel. Commun. Mob. Comput*, 7(10):1197–1208, 2007.

[89] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[90] J. Yeo, D. Kotz, and T. Henderson. Crawdad: a community resource for archiving wireless data at dartmouth. In *SIGCOMM Comput. Commun. Rev.*, pages 21–22, vol. 36, no. 2, 2006.

[91] M. E. Yildiz, A. Scaglione, and A. Ozdaglar. Asymmetric information diffusion via gossiping on static and dynamic networks. In *Proceedings of the 49th IEEE Conference on Decision and Control (CDC 2010)*, 2010.

[92] E. Yoneki, P. Hui, and J. Crowcroft. Visualizing community detection in opportunistic networks. In *Proc. of ACM MobiCom Workshop on Challenged Networks (CHANTS)*. SeptemberSeptember, 2007.

[93] Q. Yuan, I. Cardei, and J. Wu. Predict and relay: an efficient routing in disruption-tolerant networks. In *MobiHoc '09: Proceedings of the 10th ACM international symposium on Mobile ad hoc networking and computing*, pages 95–104, 2009.

[94] 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*, pages 195–206, 2007.

[95] X. Zhang, G. Neglia, J. Kurose, and D. Towsley. Performance modeling of epidemic routing. *Comput. Netw*, 51(10):2867–2891, 2007.

[96] Y. Zhang and J. Zhao. Social network analysis on data diffusion in delay tolerant networks. In *MobiHoc '09: Proceedings of the tenth ACM international symposium on Mobile ad hoc networking and computing*, 2009.

[97] Z. Zhang. Routing in intermittently connected mobile ad hoc networks and delay tolerant networks: overview and challenges. *IEEE Communications Surveys and Tutorials*, 8(1):24–37, 2006.

[98] Q. Zhao, Y. Tuan, Q. He, N. Oliver, R. Jin, and W. Lee. Communication motifs: a tool to characterize social communications. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010.

[99] W. Zhao, M. Ammar, and E. Zegura. A message ferrying approach for data delivery in sparse mobile ad hoc networks. In *MobiHoc '04: Proceedings of the 5th ACM international symposium on Mobile ad hoc networking and computing*, pages 187–198, 2004.

[100] W. Zhao, M. Ammar, and E. Zegura. Controlling the mobility of multiple data transport ferries in a delay-tolerant network. In *Proceedings of the 24th IEEE International Conference on Computer Communications (INFOCOM)*, 2005.

[101] W. Zhao, Y. Chen, M. Ammar, M. Corner, B. Levine, and E. Zegura. Capacity Enhancement using Throwboxes in DTNs. 2006.