Human Dynamic Networks in Opportunistic Routing and Epidemiology

A Thesis Submitted to the

College of Graduate Studies and Research

in Partial Fulfillment of the Requirements

for the degree of Master of Science

in the Department of Computer Science

University of Saskatchewan

Saskatoon

By

Mohammad Seyed Hashemian

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science

176 Thorvaldson Building

110 Science Place

University of Saskatchewan

Saskatoon, Saskatchewan, Canada

S7N 5C9

Abstract

Measuring human behavioral patterns has broad application across different sciences. An individual's social, proximal and geographical contact patterns can have significant importance in Delay Tolerant Networking (DTN) and epidemiological modeling. Recent advances in computer science have not only provided the opportunity to record these behaviors with considerably higher temporal resolution and phenomenological accuracy, but also made it possible to record specific aspects of the behaviors which have been previously difficult to measure.

This thesis presents a data collection system using tiny sensors which is capable of recording humans' proximal contacts and their visiting pattern to a set of geographical locations. The system also collects information on participants' health status using weekly surveys. The system is tested on a population of 36 participants and 11 high-traffic public places. The resulting dataset offers rich information on human proximal and geographic contact patterns cross-linked with their health information.

In addition to the basic analysis of the dataset, the collected data is applied to two different applications. In DTNs the dataset is used to study the importance of public places as relay nodes, and described an algorithm that takes advantage of stationary nodes to improve routing performance and load balancing in the network. In epidemiological modeling, the collected dataset is combined with data on H1N1 infection spread over the same time period and designed a model on H1N1 pathogen transmission based on these data. Using the collected high-resolution contact data as the model's contact patterns, this work represents the importance of contact density in addition to contact diversity in infection transmission rate. It also shows that the network measurements which are tied to contact duration are more representative of the relation between centrality of a person and their chance of contracting the infection.

Acknowledgments

I take this opportunity to specially acknowledge and extend my gratitude to the people who made the successful completion of this thesis possible.

First and foremost I want to express my sincere appreciation to my supervisor, Dr. Kevin Stanley, for his incredible support throughout my thesis. Dr. Stanley encouraged me to work in Human Contact Pattern and Delay Tolerant Networks as there was plenty of scope for investigation in this area, which I found it to be fascinating. I also want to thank Dr. Nathaniel Osgood, who closely collaborated in epidemiological aspects of the work. His insights and advices during different stages of the project were very valuable and effective. Dr. Stanley and Dr. Osgood steered me in the right path by providing me the necessary guidance and sharing their valuable knowledge during the meetings and discussions. They have taught me different ways to approach a problem which in turn helped me to improve my critical thinking and reasoning skills. I also appreciate the amount of effort they put in correcting my thesis.

I gratefully acknowledge my committee members, Dr. Derek Eager and Dr. Cheryl Waldner (external), for their invaluable feedback on my thesis. I am grateful to the office and technical staff members in the computer science department for assisting me in many different ways. In particular, I want to thank out graduate secretary Janice Thompson for providing timely helpful whenever needed.

I wish to thank each and every friend of mine for being on my side during the good and bad times. I am also very much indebted to the University of Saskatchewan Persian Student Association for providing me assistance during my initial days in Saskatoon.

Lastly, and most importantly, I wish to sincerely thank my Parents, Ali Hashemian and Zahra Yasemi, and my sisters, Fatemeh, Maryam, and Fahimeh, for their unconditional support and love which has been greatest strength all through my life. I dedicate this thesis to my loving parents and sisters.

Contents

List of Tables

List of Figures

List of Acronyms

DPV	Delivery Probability Value
DTN	Delay Tolerant Network
ER	Epidemic Routing
ESD	Electro Static Discharge
FOI	Force of Infection
ILI	Influenza-Like Illnesses
LBR	Location Based Routing
LBR-PS	Location Based Routing Popular Stationary
LBR-RL	Location Based Routing Reduced Load
PSN	Pocket Switched Network
RSSI	Received Signal Strength Indicator
TOS	Tiny OS
TTL	Time To Live

CHAPTER 1: INTRODUCTION

Understanding human behaviors and the dynamics of human networks, particularly in terms of proximal and geographic contacts, have diverse applications in different sciences. In Delay Tolerant Networks (DTNs, networks where delivery us more important than latency), which use humans as mobile agents, this knowledge can be used to create an infrastructure-independent network, based on dynamic and opportunistic connectivity. In public health and epidemiology, this information helps explain infection transmission and can improve the accuracy of epidemiological models of contagious diseases. Generally this information helps reveal existing patterns in human behavior and aids is designing systems which are capable of predicting these behaviors and applying the acquired knowledge on the specific domain. For example using this information in DTN allows designing systems which can predict the future connectivities and use this knowledge to optimize the routing decisions. Currently, different approaches are used to collect and analyze human contact patterns, such as self-report by study participants or observations by experimenters. This work focuses on the design and implementation of a novel approach which uses small wireless devices to record proximal and geographic human contacts cross-linked with their health status. The resulting dataset is applied to two seemingly disparate areas: first, to improve the routing performance in DTNs, and second, to examine the importance of detailed contact data in epidemiological modeling.

1.1 Human Dynamic Networks

Human contact and movement patterns underlie many fields in science. In urban planning and city development, movement patterns help determining the high traffic paths based on city structure and aid in designing the pathways and streets accordingly [1]. In Human-Computer Interaction, knowledge about human proximal and social contact patterns, how these patterns change based on humans' geographical contact patterns, and location-oriented services can help design context-aware applications [2]. This knowledge also forms the fundamental part of P3 [3] (People-to-People-to-Geographical-Places) systems which offer richer content to the user by including contact- and location-related data in the applications [4, 5].

One of the key prerequisites of characterizing human dynamic networks is having substantial knowledge of both proximal and geographical contacts, which can be acquired by studying datasets on human contact patterns. *Proximal contacts* primarily focus on people's physical proximity with each other, but in a broader view other aspects also can be placed in this category, such as electronic contacts (e.g. phone calls, texting, emails, and online social networks). *Geographical contacts* focuses on people's location, such as the individual's movement pattern (partly recordable by GPS) or their visiting pattern to certain named locations.

To record human contact pattern datasets, traditional methods use observations by experimenters and self-report by individuals. These datasets are usually collected and analyzed in the context of a specific application. For example, many studies in epidemiology collect data on human contact patterns using self-report methods, and focus on extracting the paths of infection transmission by analyzing past contacts between members of the infected population [60].

Although the data collected using traditional methods has offered important insights, these methods are error-prone, time-consuming, and have limited ability to record aspects of behavior, such as contact duration or temperature of the surrounding environment. These limitations have direct effects on the collected datasets, and lessen their applicability and reliability for many applications.

Recent advances in computer technology have made it possible to record datasets on human contact patterns and environmental data, using the small sensor devices or wireless-enabled handheld smart phones. These devices are usually capable of recording different environmental variables from the surroundings and their proximity to other people and locations for long periods. Collecting high resolution datasets on human contact patterns using these smart devices not only improves the knowledge for characterizing human dynamic networks, but also helps to investigate problems difficult to study using datasets based on participants self-reports.

1.2 Applications in Delay Tolerant Networks

Characterization of human dynamic networks has been widely considered in Delay Tolerant Networks, where the packet delivery is more important than latency. Particularly in Pocket Switched Networks (PSNs) which focuses on human as underlying DTN agents, this knowledge helps to achieve the main goal in this field, i.e. designing infrastructure-less communication protocols primarily based on human dynamic contact networks. The main focus in DTN routing has been recording and analyzing social aspects of human contact patterns. Datasets like Reality Mining [6], the Cambridge/Haggle datasets [7], and others publicly available via the CRAWDAD repository [8] are examples of recorded human contact patterns in this area. The majority of these datasets have population selection bias toward university students and staff, and cover a range of durations (from a few days to a year), study population size (from dozens to hundreds), locations and demographics. Considering that the essence of this field of research is related to computer science, prior work in this area has typically employed smart devices for data collection and therefore certain high-resolution contact granularity.

In addition to the wide attention to human proximal contact patterns seen in DTN datasets, a few works have focused on geographical aspects of human dynamic networks and recorded the pattern of visiting different locations by people, usually using devices capable of recording GPS coordinates to collect the information on participants' locations. The number of collected datasets which have included human geographical contact patterns are substantially smaller than those focused on proximal contact patterns.

The knowledge acquired on human contact patterns from these datasets are generally used to design opportunistic routing algorithms, which help create communication and data transmission protocols for human dynamic networks, using wireless-enabled handheld devices carried by people. These routing algorithms have to be capable of dealing with intrinsic properties of DTNs such as lack of end-to-end path between the source and destination at any given time, opportunistic and semi-predictable connectivity between nodes, and constraints on available resources [9].

Oblivious approaches like Epidemic Routing [10], history based approaches such as PRoPHET [11], and social-structure based approaches such as BUBBLE [12] are examples of efforts in designing DTN routing algorithms. Although this research has significantly improved understandings on human social contact patterns and has demonstrated considerable progress in routing, the achieved performance is not yet practical and there are still many other areas remaining uninvestigated.

As one of the applications of human dynamic networks, this work focuses on using geographical aspects of human contact patterns to improve DTN routing performance. It particularly investigates the potential impact of stationary nodes, acting as relay stations for mobile agents in DTNs. It also demonstrates that appropriate utilization of place as a resource can increase network performance and decrease the load on certain nodes which are over-utilized because of the popularity of their location.

1.3 Applications in Epidemiology and Public Health

Similar to the situation with DTNs, knowledge regarding human contact patterns has important applications in epidemiology and public health. Contact recording in public health, and the disease transmission networks created from it, can be based on intimate ties or distant contacts. Many behavior-based disorders such as obesity or suicidality can spread through the network of intimate social ties; a network of distant proximal contacts also has direct biological importance with respect to the spread of infections and pathogen transmissions [13], and can also influence related risk behaviors.

Currently, tracing both distant-contacts and intimate ties in the health domain are accomplished manually using questionnaires which are filled out by study participants or by health care workers (e.g. public health nurses) in collaboration with those participants. These participants are mostly selected from the population of infected cases (either laboratory-confirmed or self-reported cases) and their contacts. Although the disease-specific parameters collected via these questionnaires and the definition of contacts can differ based on the goal of the investigation, it frequently includes participants' contacts with others, and their habitual meeting places.

Data collected by contact tracing can be used to forecast outbreaks and understand the consequence of different interventions [14], by using either egocentric (focusing on the individual nodes) or sociocentric (focusing on the population under study) analyses. For infections with direct transmission (i.e. via droplet, indirect, and direct contact) having additional data on people's contacts with different locations can help identify environments with high risk of transmission. For example, it has been shown that 80% of HIV infections occur through contaminated injection equipment during drug use in public places, which generally happen in specific locations across a city [15]. In the case of air-borne infections such as tuberculosis, knowledge about individual's contact patterns with public locations plays a critical role in completing the network of infection spread and planning for potential outbreaks [14, 16].

Unlike DTNs, the vast majority of the contact pattern datasets in epidemiology and public health are collected using self-report or observation-based methods, filled by or together with the participants subject to contact tracing. These participants are sometimes reluctant to report contacts for reasons of confidentiality; even given a willingness to offer contact histories, they can have limited ability to recall the occurrence and timing of a contact, particularly for those of shorter duration or with familiar-strangers, for example on a bus or commuter train [67]. These issues not only limit the generality and completeness of the collected data, but also cause the self-report to be unable to record important attributes such as the heterogeneity of the reported contacts.

These limitations strongly motivate the use of smart-devices to automate the contact recording process. The data collected through smart-devices not only offers minute-resolution contact records, it also can record different attributes for each contact (e.g. different environmental variables such as temperature) to provide deeper insights on a population's contact patterns and subsequently improve the accuracy of the resulting network.

This work designs and implements the proposed automated contact tracing system such that the resulting dataset can be applied to public health research. The contact data collected via smart-devices in this system are cross-linked with participants' health status during the same time period, and are used to model the pathogen transmission of airborne infections such as H1N1. The high-resolution contacts provided by this system are used to study the importance of contact duration in infection transmission and the impact of duration on network centrality measurements.

1.4 Thesis Contribution

This thesis offers three primary contributions. The first contribution is the creation and development of an automated system designed to collect proximal and geographical contact data which form the essential components for characterizing human dynamic networks. Using smart devices such as sensors or cell phones to collect human behavior is not completely new as several researches have presented similar systems capable of recording human contact patterns. The previously collected datasets have enhanced our understanding of human dynamic networks, but they still do not capture the breadth of human endear, or include health data which would make the contact records more applicable to epidemiological modeling. The dataset presented in this thesis can provide deeper insights into human behavior, particularly in terms of their proximal and geographical contact patterns, and health status. This contribution has been published in WiOpt 2010 [53].

The second contribution of this thesis is the application of the collected dataset on the design of DTN routing protocols. Previous work in this area has focused on using the human proximal contact patterns to improve the routing performance. The work presented in this thesis is one of the first which considers the importance of human geographical contact patterns in DTN routing and is the first to utilize the resources available at stationary locations as relay nodes and to assess the results using an empirical dataset.

Combining the high-resolution collected data with airborne infection transmission models is the third contribution of this thesis. The resulting model leverages the understanding of the impact of contact duration (in addition to the more traditional measures of contact frequency) in contagious pathogen transmission, particularly for airborne infections such as H1N1.

1.5 Thesis Organization

The remainder of the thesis is organized as follows. Chapter 2 describes related research for each aspect of this work, including characterizing human dynamic networks, datasets available on human proximal and geographical contacts, routing protocols in Delay Tolerant Networks, and contact tracing in epidemiology and public health. Chapter 3 describes the design and implementation of the contact tracing system and provides analysis of the characteristics of the collected dataset. Chapter 4 proposes a new algorithm for DTN routing focusing on utilizing stationary resources in specific public places, and applies the collected dataset to evaluate the performance of the proposed protocol. Chapter 5 describes an agent-based H1N1 transmission model which uses the collected high-resolution data as the underlying contact pattern of the agents. Chapter 6 summarizes the thesis and outlines the possible areas for future work.

CHAPTER 2: RELATED WORK

The work presented in this thesis focuses on automated contact tracing systems and their application to DTN routing and epidemiological modeling. Previous researchers have collected valuable datasets on human contact patterns, either using smart devices such as sensors (mainly in DTN research) or via individual self-report and diaries (mainly in epidemiology and public health). Each of these datasets has been used to improve the understandings of human dynamic networks and subsequently have provided insights to either routing in DTN, or models of infection transmission. This chapter starts by reviewing previous automated contact tracing systems and the datasets collected by these systems. This is followed by the history of using human contact pattern in designing routing protocols for DTNs. The last part of this chapter describes works focusing on contact tracing in epidemiology, and researchers who have tried to use automated contact tracing in modeling infection transmission.

2.1 Human Contact Patterns

Recording and analyzing human contact patterns, or human behavioral patterns in a more general context, underlies many scientific disciplines. With advances in the design of wireless sensor modules such as MicaZ [17] or TelosB [18], which can be easily carried by people, researchers were able to automatically collect aspects of human behavior. Reality Mining [6] is one of the first to use smart phones to collect and analyze human contacts. Their dataset included the contacts between 100 university students and staff and their proximity to cell phone towers over 9 months.

After Reality Mining, other researchers obtained different datasets using automated contact tracing. These systems either used Bluetooth [6], Wi-Fi [19], or Zigbee [7] devices to record the relative proximity of participants. Example datasets include university students and staff [20], conference attendees [21], university wireless usage [22], and rollerblade tours [23]. Researchers have also attempted to use secondary measures to estimate contact patterns by inferring them from published or recorded schedules. Examples include student attendance in classes based on anonymized schedules [24], vehicular networks such as from bus schedules [25], or subway transit records [26].

These datasets are mostly available through public repositories such as CRAWDAD [8], covering a wide range of participant selection (university students [6], randomly selected citizens of a city [27], people in a theme park [27]), population size (from a few people [21] to more than 300 individuals [19]), and experiment duration (from a few days [21] to approximately one year [6]). However, the main focus in most of these studies has been on contacts between participants. In fact, most systems only recorded contacts between participants, and either don't provide any data on contacts between participants and different locations, or the provided data are very coarse and unreliable. The datasets collected by NCSU [27] are one of the few which provide information on participants' locations by recording the GPS coordinates of participants during 5 different experiments.

The dataset presented in this thesis not only improves understanding regarding human proximal contact patterns, but also is one of the first datasets which represents the human geographical contact pattern with respect to public locations. This provides the opportunity to study the patterns of people visiting locations and its relation to human proximal contacts.

2.2 DTN Routing

Since the introduction of DTNs, many authors have focused on overcoming the intrinsic challenges with respect to routing, such as lack of an end-to-end path between source and destination, opportunistic and unreliable connectivity, and constraints in available resources [9]. Early work tried to solve general problems without making any assumptions about the underlying nodes' characteristics, such as Epidemic Routing [10], SWIM [28], Spray-And-Wait [29], or other similar works [30, 31]. The main idea behind these algorithms is that a packet can be delivered to the destination by sending a copy of it to many of the available nodes in the network, hoping one of the packet holders will eventually meet the destination; the algorithms usually differ in terms of the number of packets allowed. These oblivious algorithms only offer an acceptable performance in extreme cases such as wildlife monitoring [32], because their focus on generality limited the ability of the algorithms to utilize the regularities in behavior of the mobile agents [33].

As one of the basic characteristics of mobile agents, algorithms such as PRoPHET [11, 34] used knowledge of past encounters to improve routing performance. The basic assumption in these algorithms is that nodes encountered more frequently in the past are more likely to be encountered in the future. Employing this assumption considerably increased the performance of routing algorithms.

Subsequently algorithms focused on more detailed characteristics of mobile agents. In particular, PSNs [35], a subset of DTN which only considers human as mobile agents, tries to reveal and utilize human contact patterns to optimize routing protocols. Social patterns [36, 37] or characteristics of contacts' intermeeting time [38] are examples of the characteristics employed by previous authors to better model human contact patterns and therefore improve network latency, efficiency and reliability. These characteristics are used by social based routing protocols such as Bubble [12] and LocalCom [39] to improve routing performance. These algorithms divide people into different labeled communities and suppose that each individual has certain popularity in each community, in addition to certain popularity in the whole network. The algorithm uses local and global popularity to select the next relay node. Nodes with higher popularity have a greater chance to be selected as relay nodes. The stronger assumptions used by these algorithms more precisely reflect human behavioral patterns leading to considerably improved routing performance, but unfairly utilize popular nodes along path routes, potentially comprising their device performance. In other words, the load distribution in these algorithms follows a similar trend to nodes' centrality distribution, and a node with higher centrality will receive proportionally higher load [40]. Other researchers have used contact history to shift the burden away from high-centrality nodes, and distribute it evenly between other nodes [41].

These social-based algorithms focus on the underlying social structure of human contact patterns, which were extracted from empirical datasets, such as Reality Mining [6] or the Cambridge/Haggle datasets [7]. In fact, these datasets play an important role in providing and validating the fundamental and basic assumptions, such as the difference between peoples' popularity, or the way they form communities [42].

More recently, some authors have started to study the geographic aspects of human contact patterns and employ patterns in locations as well as proximal contact for routing purposes. Yuan et. al. [43] used a semi-deterministic mobility model to simulate the mobility pattern of people moving between places and employed a similar semi-deterministic model to exploit the patterns for routing purposes. Tian and Li [44] analyzed the available datasets on human geographical contact patterns in PSN routing, while other work divided the system environment into different regions to study the movement of the nodes between each region [45]. These algorithms use the history of visiting regions by each node as the primary factor to select the subsequent relay node.

A major obstacle in understanding geographical aspects of human contacts is the paucity of empirical datasets. To the best of our knowledge, the datasets from NCSU [27], which use GPS, and part of data collected in the INFOCOM06 dataset [7], which includes fixed and mobile nodes, are the only publicly available datasets which include contact patterns with locations and mobile nodes. The work by MIT [46] is a recently collected but not publicly available dataset which also includes location information.

This thesis will use the collected dataset to analyze the importance of potential resources available at stationary locations, their applicability as relay nodes, and their effect on reducing load on mobile nodes. It presents a routing algorithm which takes advantage of the available stationary nodes to improve the routing performance in DTN.

2.3 Epidemiological Modeling

Human behavioral patterns, and in particular human contact patterns, form an important part of epidemiology and public health, creating networks of human proximal contacts [47, 13]. Without having knowledge of human contact networks, the social structure through which infectious disease, social influence, norms, information, or any other socially transmittable constructs must flow cannot be mapped and studied [13].

Social networks of a community can either be created based on intimate ties of the members or their distant contacts [13]. While a network based on intimate ties plays an important role in behavioral epidemiology, studying the transmission of many infectious pathogens requires the information on networks of proximal contacts [13].

A network of proximally distant contacts is important for studying the spread of respiratory pathogens, such as airborne infections. The transmission of these pathogens can occur either through direct contact, indirect contact, droplets, or through the air [16]. Although some of these infections require close contact between infected and susceptible individuals for transmission, many pathogens can also be transmitted solely through the air and without any direct contact [48, 49, 16]. Therefore, a contact network useful for studying the spread of airborne infections not only needs to represent the contact pattern between individuals subject to the study, but also it has to include information about contacts between individuals and geographical locations. The information on contact with public places also plays an important role in spread of blood borne infections [50, 51, 52, 15], albeit for more concrete reasons than simple proximity to an infectious agent.

The majority of the contact networks created in epidemiology and public health are based on data collected through manual methods, such as self-report or diaries filled by individual participants. In addition to their intrinsic limitations of being error-prone and time consuming, the collected data are also unable to represent attributes proven to be important in modeling infectious diseases, such as contact diversity [54] and concurrency (multiple contacts during the same time) [55]. However recent studies have considered individual [56, 57] and aggregate [58] transmission models using data collected by self-report that distinguish multiple contact intensities.

Limitations that exist in manually collecting required data to create human contact network strongly motivate a move toward automated contact tracing systems, similar to the ones which already are used for DTN research. Recently, some researches have focused on using contact tracing automation in public health and epidemiology. For example, LogSensor by the Mosar Project seeks to control AMR bacteria in hospitals through analysis of staff and patient interaction. FluPhone by researchers at Cambridge and the work by MIT researchers [46] are other projects focused on the effect of proximal contacts on infection spread. There have been other works which have attempted to model pathogen transmission using sensor-based data collection [78], but the researchers employ a highly stylized model that is not tied to any specific pathogen or real-world epidemiologic context.

As the Flunet dataset offers minute-resolution contact pattern in addition to health status of participants, this thesis combines the provided contact data with an H1N1 pathogen transmission model to study the importance of contact density in comparison with contact diversity and its relation to infection risk. It also demonstrates that centrality measurement metrics which considers contact density exhibit higher correlation with infection rate than traditional centrality measures.

CHAPTER 3: DATA COLLECTION AND ANALYSIS

Empirical datasets which represent human behavioral patterns provide a basis for research into human dynamic networks, particularly in terms of proximal and geographical contacts. Although other works have described systems to collect datasets on these behaviors, these datasets cannot describe all the existing patterns and regularities. There is a dearth of data on aspects including human geographical contact patterns and the relation of personal contacts with people's health conditions. Here a system is presented which is capable of recording human contact patterns cross-linked with their health information. The dataset provided in this chapter can offer valuable insights on human contact patterns and its relation to their health status.

This chapter focuses on experimental design and implementation. The experimental work is divided into two separate sections: contact data collection and simulation. Contact data collection explains the design, implementation and deployment of the system which is used to record the contact patterns of the participants during the data collection period. The simulation section focuses on the basic design of the system used for simulating different scenarios. In addition, different characteristics of the dataset and a preliminary statistical analysis are presented.

3.1 Contact Data Collection

The goal of the experiment was to collect participants' proximal and geographical contact patterns in addition to their health information, such that the resulting dataset can be used to improve understanding of human contact patterns in DTN networks and epidemiological modeling. As the collected data from participants represented certain patterns in their life, ethical issues potentially existed in the project. To make sure the study follows the standard ethical regulations, an application submitted to Behavioral Research Ethics Board in University of Saskatchewan which represented the study and its goals. The study followed by the approval of the board.

The participants' contacts were recorded using pocket-size wireless devices, and resulting data included the exact time and duration of each contact between individuals, and the amount of time they spent in different public places. The health information and self-reported contact data were recorded on a weekly basis using online surveys, mainly for epidemiology and health modeling purposes. This section describes strategies for participant and location selection, the hardware and software platform used to design the system, online surveys, and challenges and difficulties encountered.

3.1.1 Participant and Location Selection

During data collection, the contact patterns of 36 participants and the time they spent in proximity to 11 different public places was recorded over 3 months, starting from Nov. 9th 2009 and ending at Feb. 9th 2010. The participants included graduate students from 7 labs in the Computer Science Department, departmental staff, and undergraduate students. In this document, participants are referred to as *Mobile Nodes*. Also 11 stationary nodes was placed in public places throughout the core areas of campus to measure the amount of time each person spent in specific public areas. Out of 11 stationary nodes, 3 were connected to a networked PC and also served as data sinks, while the other 8 nodes solely recorded contacts with participants. In the following text, the stationary nodes are collectively referred to as *Stationary Nodes*, the 3 data sink nodes are referred to as *Server Nodes*, and the 8 offline stationary nodes are referred to as *Fixed Nodes*.

Selected locations are shown in Figure 3-1. In this figure, server nodes are shown with circles and fixed nodes are marked with diamonds. Locations for fixed nodes were chosen by experimenters in high traffic, public locations, or in or near the participating graduate laboratories, while locations for server nodes were picked mainly to facilitate collecting buffered data from mobile nodes.

Some mobile nodes had their spent a significant portion of their time at location in close proximity to one of the stationary nodes (either a server node or a fixed node). Also the mobile nodes could have a similar contact pattern to the associated stationary node. For example, if a mobile node spends a considerable time at its desk, the set of people it visits is similar to the set of people that visited the stationary node.

3.1.2 Setup Overview

To record the contact data, one MicaZ module (Crossbow) was assigned to each node, both stationary and mobile. Participants were asked to carry the nodes with them during the experiment, while stationary nodes were fixed at a specific location. In addition, each server node was connected to a networked PC in order to transfer the received data from mobile nodes to a central database.

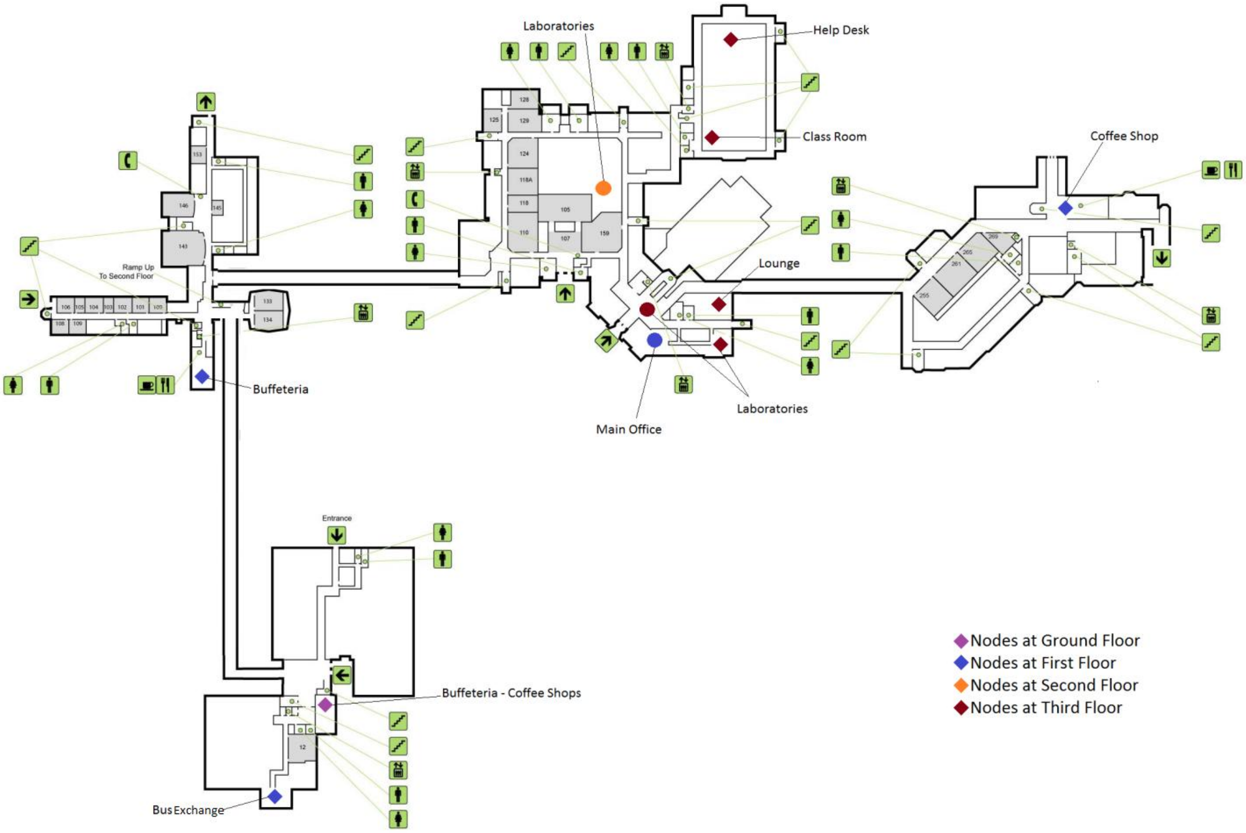


Figure 3.1: Map of stationary nodes deployed throughout the campus.

The mobile nodes buffered the recorded contact information while they were not in contact with any server node. After each connection between a mobile node and any of the server nodes, the mobile node offloaded the buffered data accumulated since the last offload time. To collect all the data at a central location, server nodes were connected to a central database which periodically uploaded collected data from the network to the main database.

Contact with a server node triggered clock synchronization by setting the mobile node's internal clock to the value received from the server node. When two mobile nodes established a new contact, they synchronized their clocks to the node which had more recently visited a server. This algorithm caused the synchronized time to propagate through the network, mitigating clock drift in mobile nodes with infrequent server contact. It bears emphasizing that while a peer-to-peer protocol was used for clock synchronization, no peer-to-peer transmission protocols were used to transfer node data back to the server.

To probe adjacent nodes, each node broadcast a "HELLO" message every 30 seconds (4-second interval for stationary nodes) with random drift of +/-2 s to prevent packet collision. If a node currently in contact failed to receive 4 consecutive HELLO messages from its partner, it labeled the partner as departed and added a new contact record to its internal buffer. Each contact record included: control flags, adjacent node ID, contact start time, contact end time, distance (discretized to CLOSE, MEDIUM, and FAR based on RSSI value), and temperature (if available). New contact records were also started if the binned RSSI value between two nodes changed.

3.1.3 Hardware Platform

Contact patterns were recorded for each node using a MicaZ module carried by the participant (in case of the stationary nodes, fixed at their specific location). Each module recorded the contacts within its Serial Flash non-volatile memory, allowing the node to retain contact information even after a hardware reset due to power failure or other problems. Nodes used the CC2420 RF-Transceiver built-in to MicaZ modules with RF power set to maximum as the communication interface.

With the exception of server nodes, all the nodes in the network were powered by two AA rechargeable batteries. Each participant was asked to replace the batteries at least every three days from one of 3 battery pools located next to server nodes around the department, as pilot studies demonstrated that each module could lasted at most 4 consecutive days. Batteries for the fixed nodes were replaced at the same rate by the experimenters. Server nodes were powered by the serial interface board to which they were connected.

To determine the environment's temperature, half of the mobile nodes were equipped with MTS310 sensor boards to periodically measure the ambient temperature. MIB510CA and MIB520CA serial interface boards were used to program the modules with the application image. These boards were also used to connected server nodes to their associated networked PC through a UART interface.

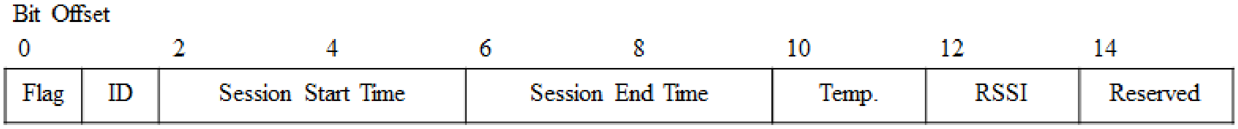
3.1.4 Software Platform

To implement the software of the system, MoteWorks (by Crossbow) was used as the primary SDK. MoteWorks is a development kit for MicaZ modules which uses TinyOS 1.1 and employs Cygwin to run required POSIX-based tools. Two separate programs were implemented in programing language nesC to be deployed on the nodes: one for mobile nodes and fixed nodes, and another for server nodes. In addition to nodes, networked PCs attached to server nodes used a custom C# program to transfer received data from server nodes to the central database. In the following sections, each of these components will be discussed in detail.

3.1.4.1 Mobile and Fixed Node Program

The MicaZ node has a non-volatile flash memory which is capable of holding 32,768 lines of 16-byte data. Considering that each node was powered with an unreliable source, and the sensitivity of the mote to electrostatic discharge (ESD), there was a chance that the node might restart at any time. To overcome the problem of data loss after reset, the program recorded collected contact data and status of the node periodically in different parts of the memory, called the Contact Block and Configuration Block, respectively.

The contact block, which covered the majority of the flash memory, started from line 1000 and ended at line 32,500. Each recorded contact occupied one line of the contact block, and included control flags, visited node ID, start and end time of the session, last recorded temperature, and RSSI level of the contact. Figure 3.2 shows the bitmap of a contact record.



Bit Offset

0		2	4	6	8	10	12	14
Flag	ID	Session Start Time		Session End Time		Temp.	RSSI	Reserved

Figure 3.2: Contact record bitmap to store in Contact Block

The configuration block held the data required to recover following reset to the latest state prior to reset, and occupied lines 32,500 to 32,600 of flash memory. These data included the number of contacts recorded in the contact block, the number of offloaded contacts from the contact block, and the current time. Each set of configuration data occupied one line of the configuration block, and the program added a new line in a circular fashion. Therefore, at any given time, one line of the block had the most recent configuration information. Lines 0 to 1000 were used to record the program images and lines 32,600 to 32,768 were left unused.

Time was recorded in a 32-bit value as the number of 8 milliseconds periods elapsed from starting the experiment. The time accuracy was decreased from millisecond-accuracy due to space limitations in memory and packets. A factor of 8 was employed for rapid binary division. Each time value consisted of two sections: base and offset. Base is the time between start of the experiment and last clock reset operation in the node. A node resets the clock to synchronize the value of its internal clock with the system in one of the following conditions:

1) **Visiting a server node**: In this case the node receives the correct time from the server and therefore resets the internal clock and updates the base time with the received time. It also records the current time as the latest time (as given by the server) at which it updated the internal clock.

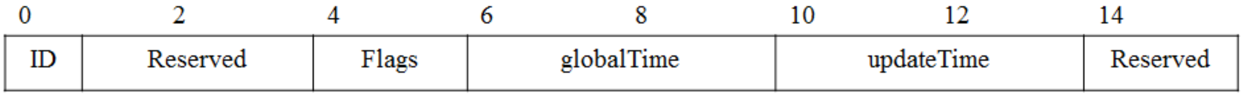
2) **Visit an adjacent node which had more recently visited a server node**: The node resets its internal clock and uses the time received from the adjacent node as the base time. It also records the clock update time from the adjacent node to show how long ago the current time has been updated with a server node.

3) **Read from configuration block as the last recorded time before reset**: During the startup routine, the node finds the latest recorded time in the configuration block and uses it as the base time.

Offset showed the number of milliseconds passed since last base update (i.e. the last internal clock reset) due to any of the reasons described above. Adding base time and offset time always resulted the most accurate time in the node. Note that the time accuracy of a node was proportionally related to the frequency of visiting server nodes or other updated nodes in the network.

The Time module used here had 1024 ticks per second which caused a drift in both base and offset time. As this was a uniform and constant drift, it was corrected by post-processing the received data.

Each node periodically broadcasted a HELLO packet to the environment. The broadcast interval for mobile nodes was every 30 seconds, while for stationary nodes was every 4 seconds. The HELLO packet format is shown in Figure 3.3. In addition to node ID and a control flag, each node also included its internal time. In this packet, globalTime is a 32-bit value which shows number of 8 milliseconds passed since the experiment start time, and updateTime shows the previous time that globalTime updated with a server node.



0	2	4	6 8	10 12	14
ID	Reserved	Flags	globalTime	updateTime	Reserved

Figure 3.3: HELLO packet format

Each node holds a session record for every other known node in the network. This record can be active or inactive, depending on if the respective nodes are present. A session record includes the node ID, if the session is active, session start time, session RSSI, and a counter to count the number of steps without receiving a HELLO packet. When a node receives the first HELLO packet from an adjacent node, it activates that node's session (if it is the first time the nodes meet, it first creates a new session record for the adjacent node), and records the session start time. Based on the RSSI of the received packet, it marks the session as CLOSE, MEDIUM, FAR, or FAINT.

For each active session the counter is incremented every 30 seconds and reset upon receipt of a HELLO packet. If the counter for an active session reaches to 4, the node marks the node as 'Departed' (deactivates the session) and adds a contact record to the contact block of the memory.

In addition to 4 consecutive HELLO packet failures, changes to the RSSI value can also cause the previous session record to be committed to the contact block and a new session record started. For example, if a session started with a CLOSE RSSI and subsequently the RSSI dropped to MEDIUM, the node finalizes the current session for the adjacent node, records it in the contact block with CLOSE RSSI, and starts a new session with MEDIUM RSSI. This gives an estimate of the distance between two nodes in each recorded session, and ensures that changes to relative distance in a single continuous contact are accurately recorded.

The thresholds for each distance label were chosen based on the results of pilot experiments. The CLOSE group is related to line of sight contacts with a distance of up to five meters. The MEDIUM group represents contacts with a line of sight distance of up to 15 meters. The FAR group is related to all line of sight contacts farther than 15 meters. The 'Faint' group shows very low RSSI values (less than 7%). This often happens when two nodes are very close to each other (approximately less than 20 centimeters) and can be explained by the "Near Field and Far Field" behavior of electromagnetic radiation from an antenna [59]. While the distances were calibrated using line of sight RSSI values, values with intervening obstacles were inevitably recorded during the experiment. It is worth noting that obstacles generally make the RSSI appear farther away, suggesting the CLOSE category as a reasonable approximation of an epidemiologically relevant contact. Hysteresis was applied to the RSSI bounds to prevent small fluctuations at the edge of an RSSI band creating a proliferation of unnecessary records. The thresholds and their associated hysteresis bands are shown in table 3.1.

Table 3-1: RSSI thresholds for different proximities

	Low bound	**High bound**
CLOSE	210 dB	265 dB
MEDIUM	195 dB	220 dB
FAR	30 dB	205 dB
FAINT	0 dB	30 dB

Each node could receive a set of commands from the server nodes and react accordingly. The following commands were implemented:

- Erasing data in configuration block
- Erasing all the data in the memory
- Offload the configuration block data to the server: used to fetch the recorded values in the configuration block
- Offload the contact block data to the server: used to manually restore the packets which delivered from the device but dropped on the server side.
- Offload the whole memory to the server
- Pause the operation
- Resume the operation
- Change the threshold for RSSI proximity levels: in cases where the pre-defined boundaries were not suitable (for example due to different materials in the environment), the command could be used to define new boundaries.

3.1.4.2 Server Node's Code

To implement basic operations of the server node, the MoteWorks pre-system was employed. These operations include receiving packets from adjacent nodes and recording them in a local MySQL database, and recording the amount of time each node stays in proximity to the server node. This calculated by number of HEALTH packets which broadcasted by nodes periodically, received by the server node from any adjacent node every two minutes. HEALTH packets, which are different than HELLO packets, are part of framework which used for development (i.e. MoteWorks) and are broadcasted from mobile nodes during their adjacency to server nodes.

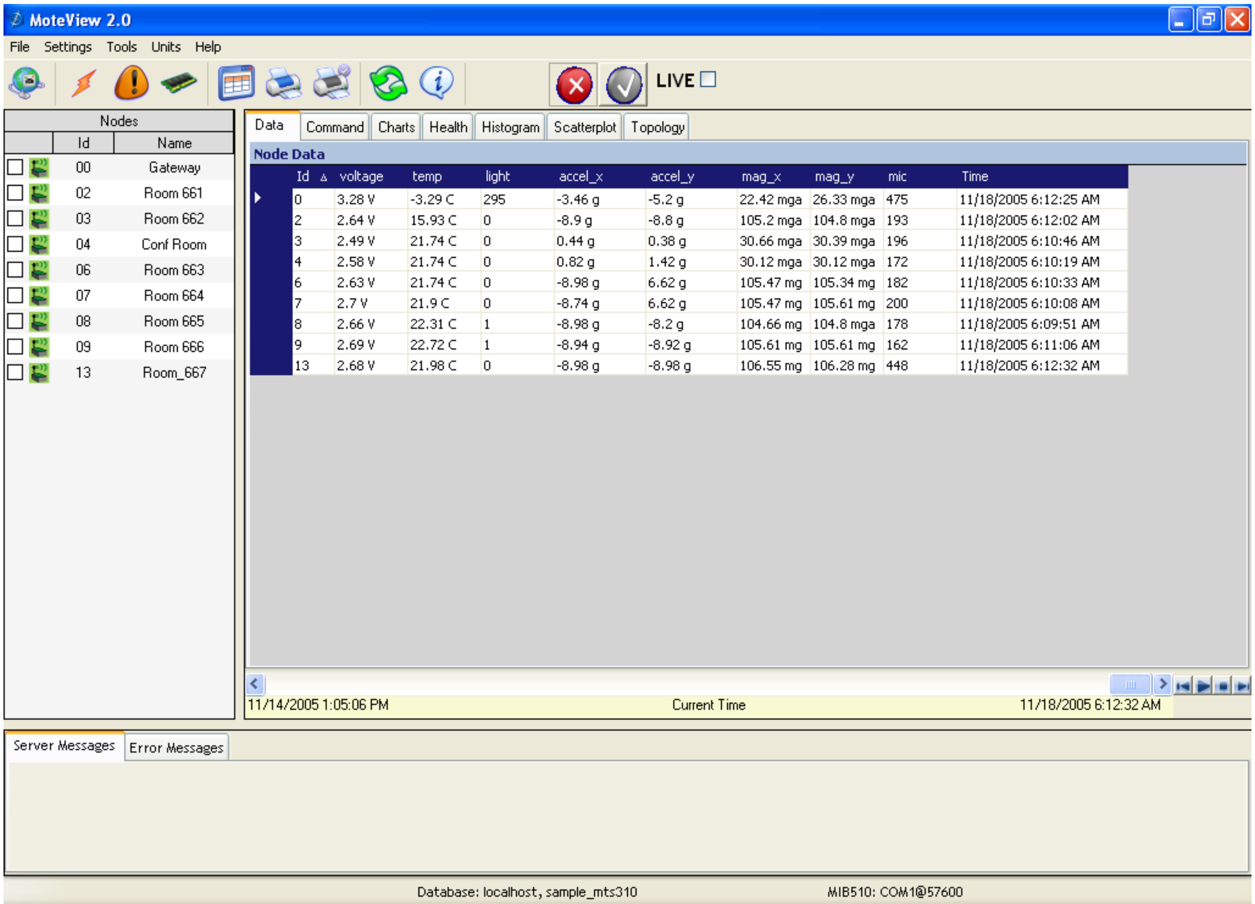
A command mode was also implemented for the server nodes, allowing the experimenter to control the operation of the server node using the connected PC. The implemented commands for server nodes are as follows:

- Pause server node
- Resume server node
- Set server node's internal clock

These commands were primarily used to set the time in the server after restart. Setting the correct time in the server was accomplished manually by the experimenter through pausing the server, setting the correct time, and resuming normal operation.

3.1.4.3 Server PC Configuration

MoteView, one of the tools available in MoteWorks, was used to receive the data from the server node and to record it in the local MySQL database. A snapshot of this tool is shown in Figure 3.4. MoteView connected via a serial port to the server node and received the delivered packets from the server node. After receiving each packet, it converted the packet into a proper SQL INSERT command and executed that command on the local database for storage. A custom application written in C# was used to move the recorded data from the local MySQL database to


MoteView 2.0
File Settings Tools Units Help
LIVE □
Data Command Charts Health Histogram Scatterplot Topology
Server Messages Error Messages

Nodes

	Id	Name
□	00	Gateway
□	02	Room 661
□	03	Room 662
□	04	Conf Room
□	06	Room 663
□	07	Room 664
□	08	Room 665
□	09	Room 666
□	13	Room_667

Node Data

Id	voltage	temp	light	accel_x	accel_y	mag_x	mag_y	mic	Time
0	3.28 V	-3.29 C	295	-3.46 g	-5.2 g	22.42 mga	26.33 mga	475	11/18/2005 6:12:25 AM
2	2.64 V	15.93 C	0	-8.9 g	-8.8 g	105.2 mga	104.8 mga	193	11/18/2005 6:12:02 AM
3	2.49 V	21.74 C	0	0.44 g	0.38 g	30.66 mga	30.39 mga	196	11/18/2005 6:10:46 AM
4	2.58 V	21.74 C	0	0.82 g	1.42 g	30.12 mga	30.12 mga	172	11/18/2005 6:10:19 AM
6	2.63 V	21.74 C	0	-8.98 g	6.62 g	105.47 mg	105.34 mg	182	11/18/2005 6:10:33 AM
7	2.7 V	21.9 C	0	-8.74 g	6.62 g	105.47 mg	105.61 mg	200	11/18/2005 6:10:08 AM
8	2.66 V	22.31 C	1	-8.98 g	-8.2 g	104.66 mg	104.8 mga	178	11/18/2005 6:09:51 AM
9	2.69 V	22.72 C	1	-8.94 g	-8.92 g	105.61 mg	105.61 mg	162	11/18/2005 6:11:06 AM
13	2.68 V	21.98 C	0	-8.98 g	-8.98 g	106.55 mg	106.28 mg	448	11/18/2005 6:12:32 AM

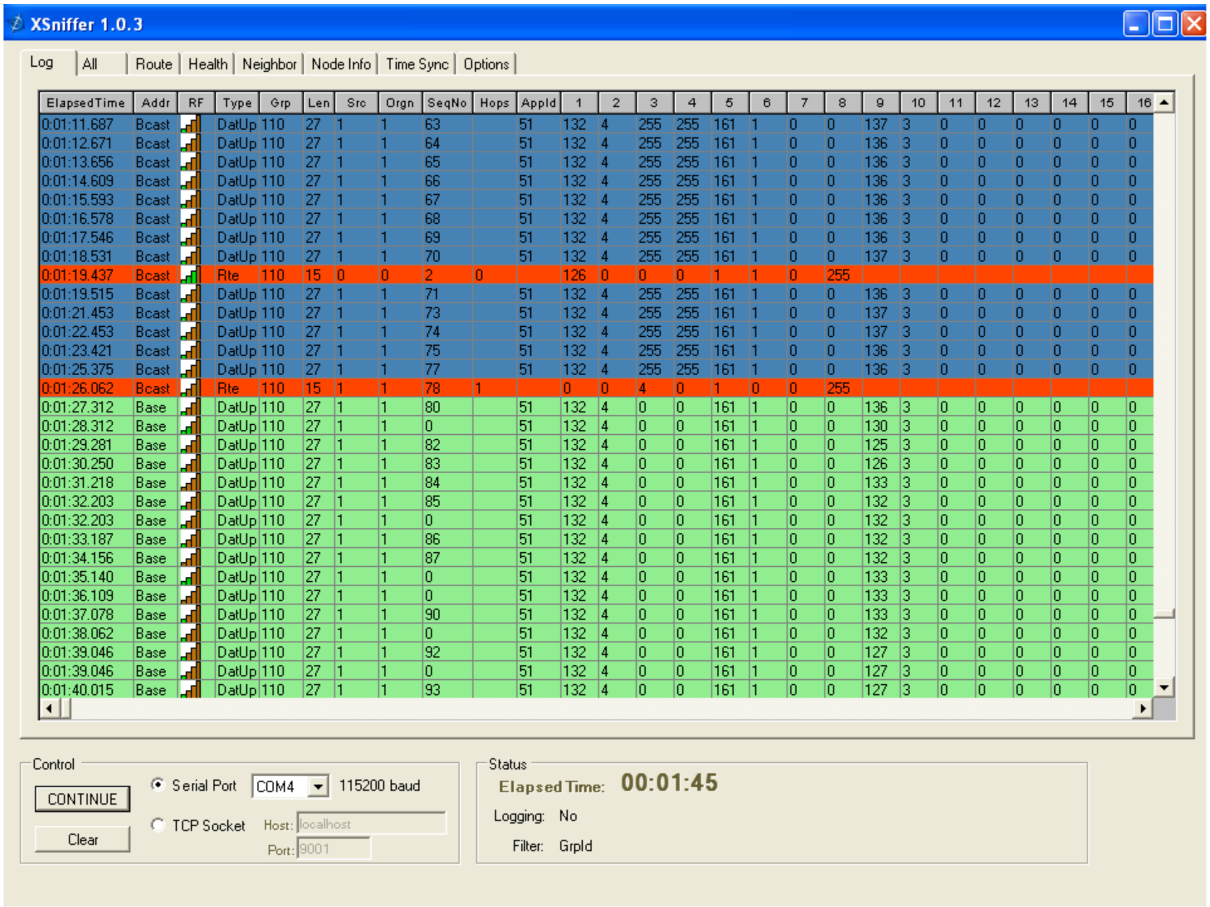
11/14/2005 1:05:06 PM Current Time 11/18/2005 6:12:32 AM

Database: localhost, sample_mts310 MIB510: COM1@57600

Figure 3.4: MoteView screen, used for receiving data from server nodes and storing them in the database.

a central SQL Server database. In addition to recording the raw received data, this application parsed the raw data and stored in it final form in the central database.

Xserve, a tool in TinyOS which is available in Windows through the Cygwin command line, was used to communicate with the nodes (stationary or mobile) in the network via the server PC. The tool used a UART port through the connected serial interface board to pass the received commands and arguments to the specified node. Other tools such as XSniffer (shown in figure 3.5) were used to control and monitor the network during the study.



ElapsedTime	Addr	RF	Type	Grp	Len	Src	Orgn	SeqNo	Hops	AppId	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0:01:11.687	Bcast	▂▄▆	DatUp	110	27	1	1	63		51	132	4	255	255	161	1	0	0	137	3	0	0	0	0	0	0
0:01:12.671	Bcast	▂▄▆	DatUp	110	27	1	1	64		51	132	4	255	255	161	1	0	0	136	3	0	0	0	0	0	0
0:01:13.656	Bcast	▂▄▆	DatUp	110	27	1	1	65		51	132	4	255	255	161	1	0	0	136	3	0	0	0	0	0	0
0:01:14.609	Bcast	▂▄▆	DatUp	110	27	1	1	66		51	132	4	255	255	161	1	0	0	136	3	0	0	0	0	0	0
0:01:15.593	Bcast	▂▄▆	DatUp	110	27	1	1	67		51	132	4	255	255	161	1	0	0	136	3	0	0	0	0	0	0
0:01:16.578	Bcast	▂▄▆	DatUp	110	27	1	1	68		51	132	4	255	255	161	1	0	0	136	3	0	0	0	0	0	0
0:01:17.546	Bcast	▂▄▆	DatUp	110	27	1	1	69		51	132	4	255	255	161	1	0	0	136	3	0	0	0	0	0	0
0:01:18.531	Bcast	▂▄▆	DatUp	110	27	1	1	70		51	132	4	255	255	161	1	0	0	137	3	0	0	0	0	0	0
0:01:19.437	Bcast	▂▄▆	Rte	110	15	0	0	2	0		126	0	0	0	1	1	0	255								
0:01:19.515	Bcast	▂▄▆	DatUp	110	27	1	1	71		51	132	4	255	255	161	1	0	0	136	3	0	0	0	0	0	0
0:01:21.453	Bcast	▂▄▆	DatUp	110	27	1	1	73		51	132	4	255	255	161	1	0	0	137	3	0	0	0	0	0	0
0:01:22.453	Bcast	▂▄▆	DatUp	110	27	1	1	74		51	132	4	255	255	161	1	0	0	137	3	0	0	0	0	0	0
0:01:23.421	Bcast	▂▄▆	DatUp	110	27	1	1	75		51	132	4	255	255	161	1	0	0	136	3	0	0	0	0	0	0
0:01:25.375	Bcast	▂▄▆	DatUp	110	27	1	1	77		51	132	4	255	255	161	1	0	0	136	3	0	0	0	0	0	0
0:01:26.062	Bcast	▂▄▆	Rte	110	15	1	1	78	1		0	0	4	0	1	0	0	255								
0:01:27.312	Base	▂▄▆	DatUp	110	27	1	1	80		51	132	4	0	0	161	1	0	0	136	3	0	0	0	0	0	0
0:01:28.312	Base	▂▄▆	DatUp	110	27	1	1	0		51	132	4	0	0	161	1	0	0	130	3	0	0	0	0	0	0
0:01:29.281	Base	▂▄▆	DatUp	110	27	1	1	82		51	132	4	0	0	161	1	0	0	125	3	0	0	0	0	0	0
0:01:30.250	Base	▂▄▆	DatUp	110	27	1	1	83		51	132	4	0	0	161	1	0	0	126	3	0	0	0	0	0	0
0:01:31.218	Base	▂▄▆	DatUp	110	27	1	1	84		51	132	4	0	0	161	1	0	0	133	3	0	0	0	0	0	0
0:01:32.203	Base	▂▄▆	DatUp	110	27	1	1	85		51	132	4	0	0	161	1	0	0	132	3	0	0	0	0	0	0
0:01:32.203	Base	▂▄▆	DatUp	110	27	1	1	0		51	132	4	0	0	161	1	0	0	132	3	0	0	0	0	0	0
0:01:33.187	Base	▂▄▆	DatUp	110	27	1	1	86		51	132	4	0	0	161	1	0	0	132	3	0	0	0	0	0	0
0:01:34.156	Base	▂▄▆	DatUp	110	27	1	1	87		51	132	4	0	0	161	1	0	0	132	3	0	0	0	0	0	0
0:01:35.140	Base	▂▄▆	DatUp	110	27	1	1	0		51	132	4	0	0	161	1	0	0	133	3	0	0	0	0	0	0
0:01:36.109	Base	▂▄▆	DatUp	110	27	1	1	0		51	132	4	0	0	161	1	0	0	133	3	0	0	0	0	0	0
0:01:37.078	Base	▂▄▆	DatUp	110	27	1	1	90		51	132	4	0	0	161	1	0	0	133	3	0	0	0	0	0	0
0:01:38.062	Base	▂▄▆	DatUp	110	27	1	1	0		51	132	4	0	0	161	1	0	0	132	3	0	0	0	0	0	0
0:01:39.046	Base	▂▄▆	DatUp	110	27	1	1	92		51	132	4	0	0	161	1	0	0	127	3	0	0	0	0	0	0
0:01:39.046	Base	▂▄▆	DatUp	110	27	1	1	0		51	132	4	0	0	161	1	0	0	127	3	0	0	0	0	0	0
0:01:40.015	Base	▂▄▆	DatUp	110	27	1	1	93		51	132	4	0	0	161	1	0	0	127	3	0	0	0	0	0	0

Control

● Serial Port COM4 115200 baud

○ TCP Socket Host: localhost Port: 9001

Status

Elapsed Time: 00:01:45

Logging: No

Filter: GrpId

Figure 3.5: XSniffer, used to monitor the packet transfer in the network during the study.

3.1.4.4 Database Structures

A local MySQL database was responsible for temporarily buffering the data. A copy of this database was created on each server PC which consisted of two tables. The first table, RawData, was used to hold the packets received from nodes in the network. Each row in this table represents one contact record, and consisted of the record's reception time by the server node, and 12 integer columns to hold different parts of the message. These fields matched the fields in the received packet.

Table 3-2: AnalyzedData table format

Column Name	Type	Description
result_time	Date & Time	The time which this record is received at server node
nodeID	Number	Node ID who reported this record
seenNodeID	Number	ID of the adjacent node
msStartContactTime	Number	Start time of the contact, in millisecond format
startContactTime	Date & Time	Start time of the contact in real time format
fixedStartTime	Date & Time	Actual contact start time after correcting the 2.4% drift (refer to ‘Time System’ in 3.1.4.2 for more details)
msEndContactTime	Number	End time of the contact, in millisecond format
endContactTime	Date & Time	End time of the contact in real time format
fixedEndTime	Date & Time	Actual contact end time after correcting the 2.4% drift (refer to ‘Time System’ in 3.1.4.2 for more details)
temperature	Number	Reported temperature during the contact (valid if the node equipped with a sensor board)
Flag	Number	Control flags associated with the record
rssiLevel	Number	RSSI level during the contact
serverID	Number	ID of the server which received the record.

The second table, HealthPackets, was used to hold the health packets sent every two minutes to the server node from adjacent nodes. As mentioned earlier, the data collected in this table can be used to calculate the time each node spends in proximity to each server node.

A central SQL Server database collected all the data from server PCs in the network. In addition to a copy of data in RawData and HealthPackets tables, the central database also had a table to hold the decoded version of data in RawData, called AnalyzedData. The design of AnalyzedData and description of each field are shown in Table 3.2.

3.1.5 Surveys

While it is technically feasible to record contact patterns and health information automatically, collecting health symptom information remains difficult and potentially invasive. In traditional contact tracing based on self-report [60], health symptoms would often be reported by people at the same time as they were asked to record their contacts. Although in this work the contact pattern between participants and places was recorded automatically, participants were still asked to report their health state, any symptoms they might be experiencing, and a recollected estimate of their past week's contact patterns through surveys. Each participant received an electronic invitation for the weekly survey at the end of the week by email. The questions in the survey targeted the last week's health and contact data. Participants had one week to complete the survey; the old survey was closed before sending the invitation for the next survey.

The weekly surveys consisted of two parts: estimation of participants' contact time with other members of the study, and their health symptom information during the past week. In first part, participants received a list of all members of the study and were asked to report the total duration of contacts in minutes with their top five contacts (judged by contact durations) during the past week. They were also asked to specify if the contacts corresponded to 'Friends', 'Colleagues', both, or none. In the second part, participants were asked to report any health-related symptoms they experienced during the preceding week. They were specifically asked if they experienced flu-like symptoms, the start and end time of the illness (if available), and whether they had visited a physician. At the end of the survey, they were asked to describe the condition of their MicaZ module and any failures which had occurred. The second part of the weekly survey is shown in Appendix A.

In addition to weekly surveys, participants were also asked to complete a one-time demographic/background survey at the end of the experiment, which included questions on basic demographic variables, the average amount of time spent with people and on campus, whether and when they had received a flu shot or H1N1 vaccination, and their attitudes towards the wireless module. The list of questions in the demographic survey is shown in Appendix B.

3.1.6 Failure Modes

Wireless module failures can be divided into two main categories: technical problems and human carelessness. The primary technical failure mode during the study was an RF transceiver failure in the MicaZ receiver amplifier. The node could not receive HELLO packets from other nodes, but it could broadcast packets properly. Therefore, other nodes could record the faulty node’s existence, while the faulty node couldn’t find any adjacent nodes.

Based on the diagnostic report by the manufacturer, the problem was due to electrostatic discharge (ESD) damage to the receiving amplifier, due to the prototype design of MicaZ motes and the dry, cold conditions which dominate Saskatchewan winters. To prevent this failure, the nodes were covered in three different types of protection: 1) using bubble wrap packets 2) using MicaZ standard boxes provided by the manufacturer 3) using anti-electrostatic bags. Figure 3.6 shows the node when covered by each of these different covers. None of the applied methods completely prevented subsequent failures. Thirty-two nodes experienced receiver failure during the study. Defective nodes were detected manually based on the reported contacts in the central database and were replaced after failure detection.

Battery failure or displacement, sensor board displacement caused by shaking the module (for the modules equipped with MTS310 sensor board), and storage memory failure were other technical problems which occurred rarely during the study.

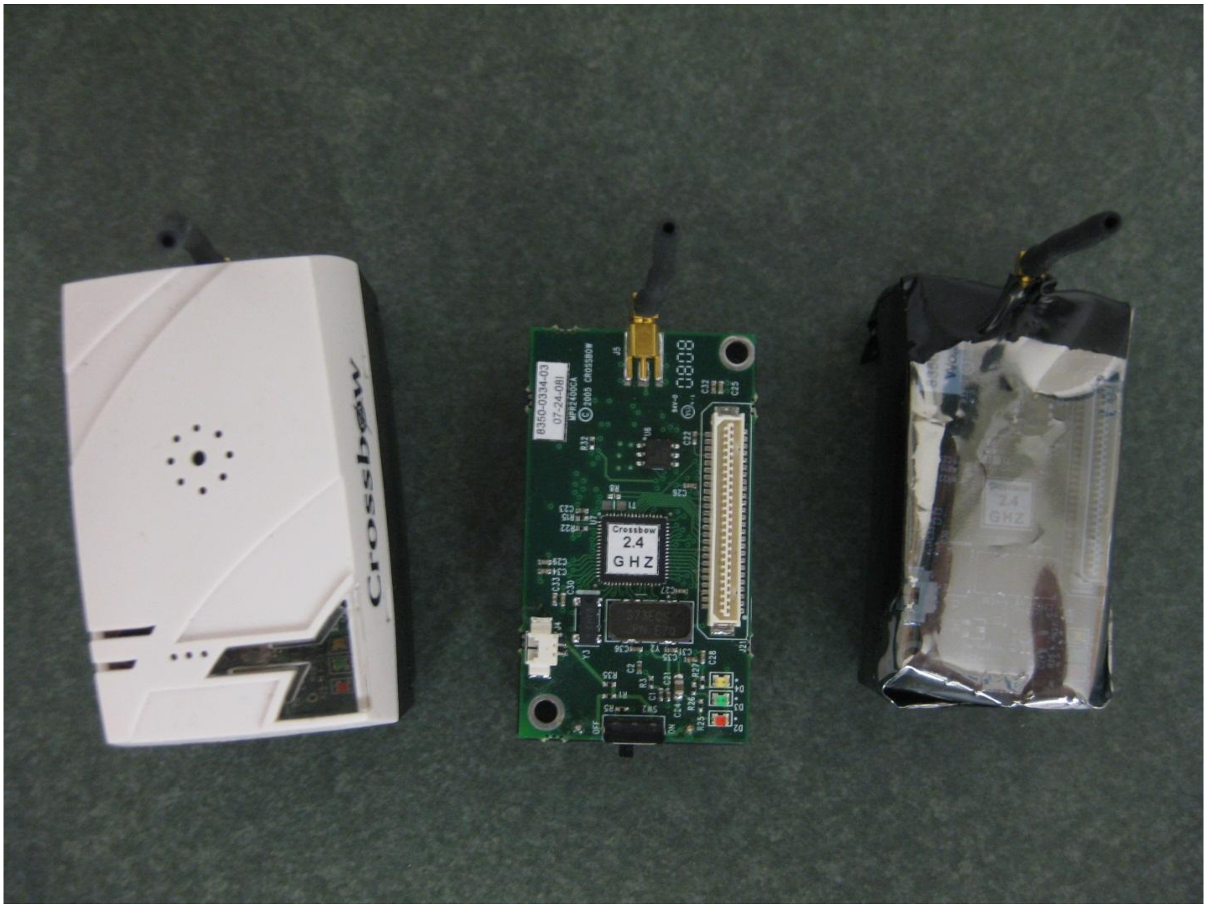


Figure 3.6: Different packaging for MicaZ modules. The left module is placed in recommended Crossbow box; the middle module is bare module without any cover; the right module is covered with anti-static bags to prevent Electro-Static discharge. The modules with the Crossbow boxes were also equipped with MTS310 sensor boards.

Human error was dominated by participants forgetting to carry the module, and forgetting to periodically replace the battery. When the participant forgot to carry the module, either the module was left on or off. In the former case, the module recorded additional erroneous data and in the latter case, potential contact information was not recorded. Compliance issues were addressed by sending reminders, mentioning compliance in the weekly surveys, and directly reminding the more egregious offenders. It was observed that as time progresses, people cared less about the mote, and more often forgot to change the battery or to keep the device on. Voluntarily turning the node off, and forgetting to switch the device on after turning it off, are other examples of human carelessness which appear to heave rarely affected data collection.

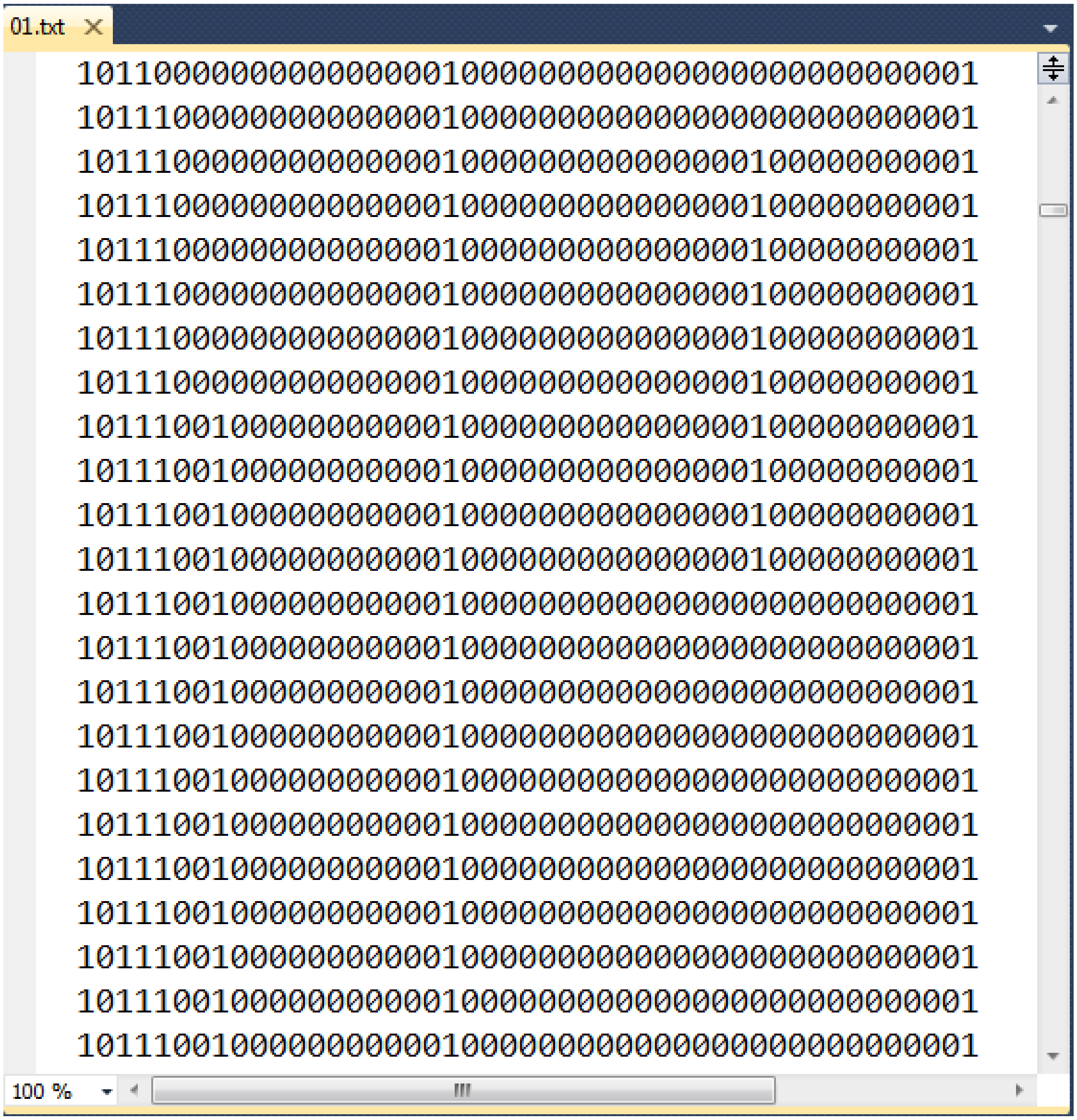
All the described obstacles were adequately controlled during the experiment, and therefore had the minimal impact on the collected dataset. The impact of the node' transmitter failure was mitigated by defining contact between a pair of nodes as the union of their contact records. Therefore, if a defective node were to visit a healthy node, the contact could be recorded by the healthy node as the defective node could broadcast HELLO packets successfully. The only condition where data could potentially be lost was the cases where both nodes were defective, which happened rarely due to closely monitoring the network for defective nodes. In order to diminish the effect of human carelessness, the collected data was observed every day and implausible records were deleted manually at the end of the experiment, as described in section 3.3.

3.2 Simulation and Modeling Preparation

Processes associated with both of the application areas (epidemiological modeling and DTN routing) can be represented as agent-based models where the results from contact data collection determine the relations (contacts) between the agents. To input the recorded contact pattern to the agent-based model, the experimental period was discretized into 30-second time steps and the connectivity pattern for each node (agent) for each time step was extracted from the dataset based on the reported contacts from the related nodes. The contact graph used here is considered undirected; therefore a reported contact between node A and node B at a certain time from either node, implied a contact at that time of both nodes to each other, that is, the contact record between nodes A and B is the union of A's contact record with B, and B's contact record with A.

Nodes in the network were sorted based on their type and numbered from 1 to 47. Nodes 1 to 36 represent the 36 mobile nodes in the network; nodes 37 to 44 represent the 8 fixed nodes; and nodes 45 to 47 represent the 3 server nodes.

To simulate the agent based model, either for a DTN with human agents or for models on infection spread, network simulator 3 (ns-3) was employed [61]. The contact pattern for each node was recorded to a node-specific file and fed to the simulator to determine the connectivity



```
10110000000000000001000000000000000000000000001
10111000000000000001000000000000000000000000001
10111000000000000001000000000000000100000000001
10111000000000000001000000000000000100000000001
10111000000000000001000000000000000100000000001
10111000000000000001000000000000000100000000001
10111000000000000001000000000000000100000000001
10111000000000000001000000000000000100000000001
10111001000000000001000000000000000100000000001
10111001000000000001000000000000000100000000001
10111001000000000001000000000000000100000000001
10111001000000000001000000000000000100000000001
10111001000000000001000000000000000000000000001
10111001000000000001000000000000000000000000001
10111001000000000001000000000000000000000000001
10111001000000000001000000000000000000000000001
10111001000000000001000000000000000000000000001
10111001000000000001000000000000000000000000001
10111001000000000001000000000000000000000000001
10111001000000000001000000000000000000000000001
10111001000000000001000000000000000000000000001
10111001000000000001000000000000000000000000001
10111001000000000001000000000000000000000000001
```

Figure 3.7: Snapshot of a contact pattern file.

between the nodes during the simulation. Each line in the file for a given node consisted of 47 characters, ‘0’ or ‘1’, and each character determined the connectivity of the node associated with the file and the other nodes in the network. The line number mapped to the index of the 30-second time step in the experiment period. For example if node *X* was connected to node *Y* during time step *k*, the *Y*th character of the *k*th line in connectivity file for *X* is equal to ‘1’. Note that the connectivity is considered undirected. Therefore in this example, the *X*th character of *k*th line in the connectivity file for *Y* is also equal to ‘1’. Each node is considered as connected to itself for all the time steps. Figure 3.7 shows a snapshot of one of the contact pattern files.

3.3 Dataset Characteristics and Analysis

The collected dataset consists of two types of data: data collected automatically using sensor modules, and data collected from participants through surveys. The combination of these two represents human behavioral pattern with respect to their contacts and health status. Therefore certain characteristics are expected in the dataset such as the frequency and density of contacts at different hours of the day, clustering of members based on their lab association, and specific contact duration distributions. Analyzing these characteristics allows examining the validity of the dataset regarding expected behaviors and comparing it with other existing datasets. It also reveals basic features of the dataset which can be helpful in more detailed analyses.

This section starts with the results from a demographic survey which show the characteristics of the participant population. Then the basic analysis and descriptive statistics of the collected contact data is presented. At the end, the cumulative health data from weekly surveys are presented and their relation to the contact data is discussed.

3.3.1 Participants Demographics

Participants in this study consisted of 75% males and 25% females, of which 47% were in contact with other participants solely on campus. 14% of the participants received a regular flu shot every year, and 39% of them had received the H1N1 vaccination. 83% of participants never smoked, while 6% smoked occasionally, and 11% smoked every day. As the study occurred during the winter and people were advised about the concerns regarding the risks of catching H1N1 flu or the risks of transmitting flu from themselves to others, 2 people reported they changed their behavior in response to the concerns regarding catching the flu, and 5 people reported they changed their behavior in response to the concerns regarding transmitting the flu (2 people had a positive answer to both of these questions). Figure 3.8 shows the age diversity of

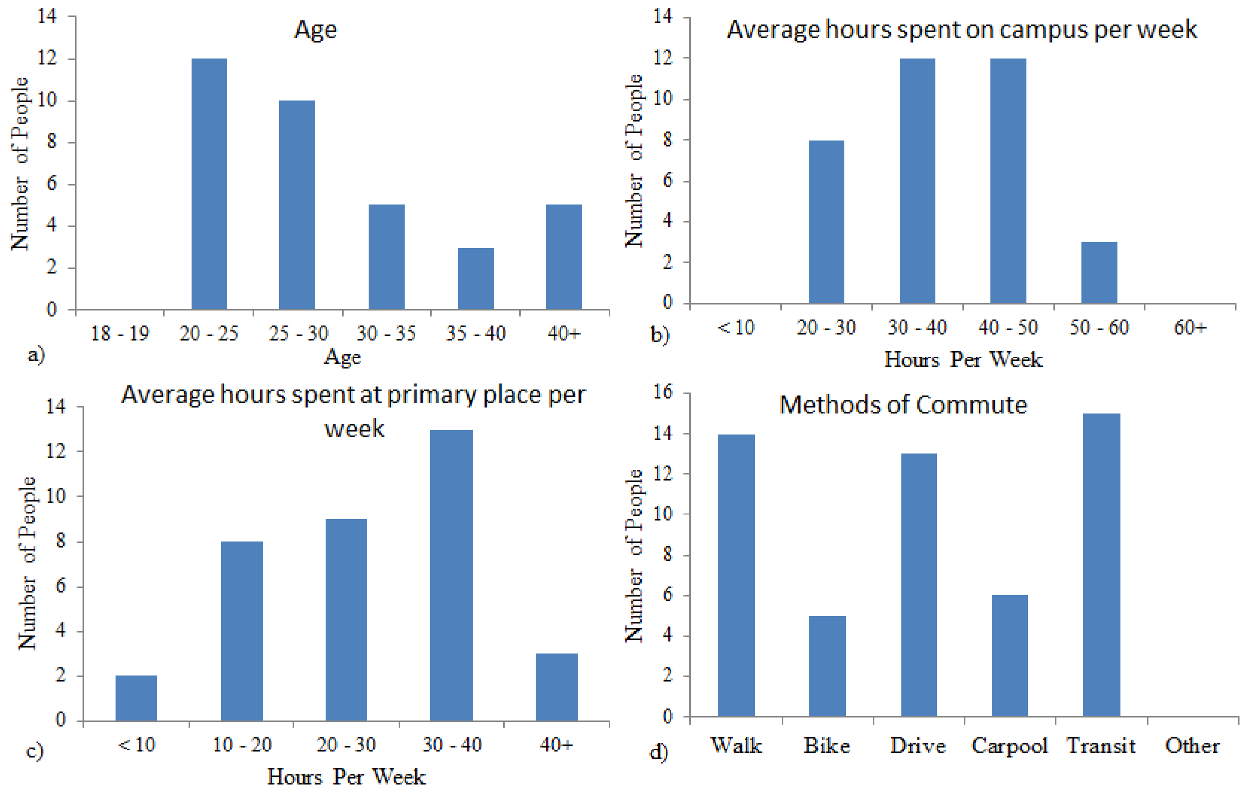


Figure 3.8: Results from different questions of demographic survey.

the participants, the amount of time they spent at the university and at their primary locations (e.g. office, desk, laboratory), and the method employed to commute to and from the university.

3.3.2 Dataset Characteristics

This section focuses on the collected data and its characteristics. The collected contact data can be divided into four different groups based on the RSSI classification: CLOSE, MEDIUM, FAR, and FAINT. Out of total reported contacts, 32.31% occurred with CLOSE proximity, 67.63% happened in MEDIUM proximity, and only 0.04% reported FAINT proximities. No contact was reported as FAR. Lack of FAR contacts can be mainly due to defined thresholds for RSSI values. Unless stated otherwise, the analysis presented at this section takes all proximity groups into account.

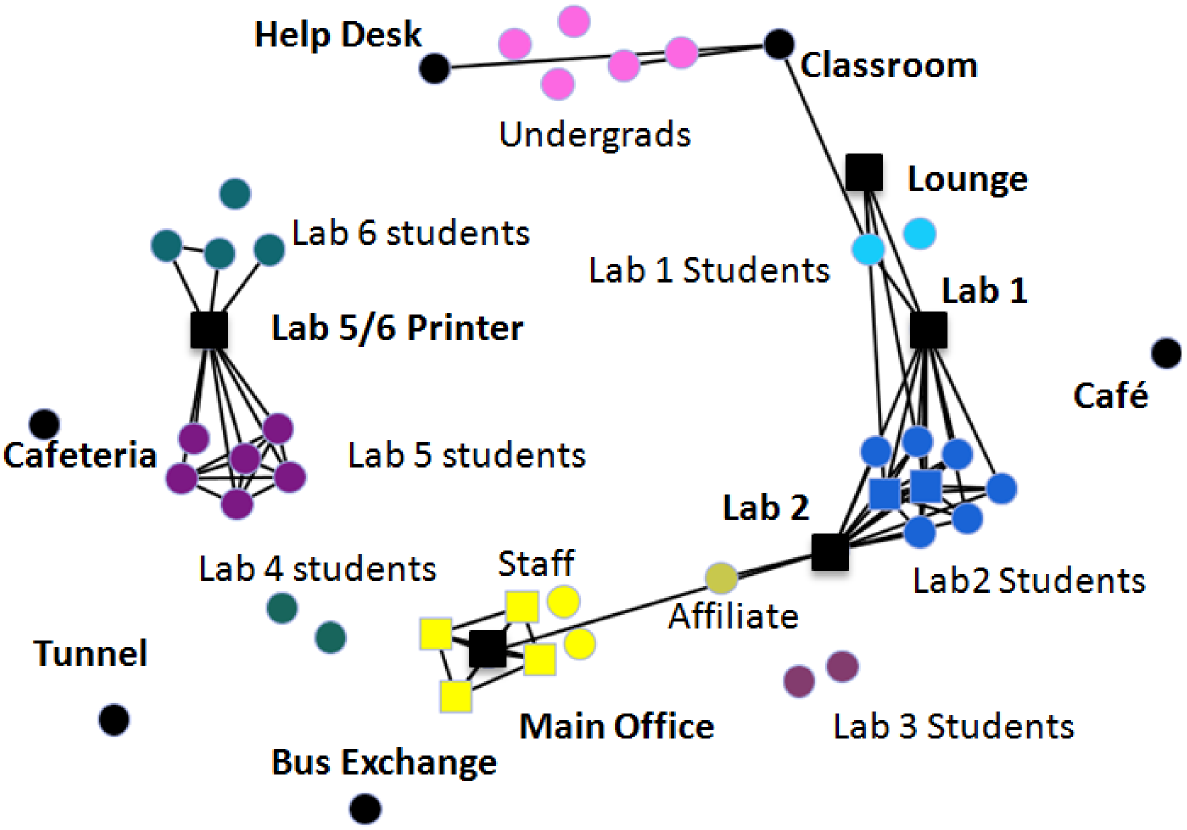


Figure 3.9: Network structure for those with at least 18 minutes contact per day

As mentioned earlier, participants were drawn from 7 different Computer Science Laboratories, undergraduate students, and departmental staff. Mobile nodes were classified into 9 different labels based on their group association. Including two different types of stationary nodes (i.e. fixed nodes and server nodes), the network was divided into 11 different categories. Figure 3.9 plots a realization of the network as a graph. Each vertex represents a node, and the color indicates the node’s association with a group. An edge between two nodes implies a mean of more than 18 minutes of contact per day between two participants during the course of experiment. Square nodes represent popular stationary nodes as described in section 4.5.4.

Because the dataset represents a human environment and contacts occurring between work colleagues, it is suspected that most of the contact occurred during the day. Figure 3.10 validates this hypothesis. As shown in this figure, the total number of reported contacts and hence total

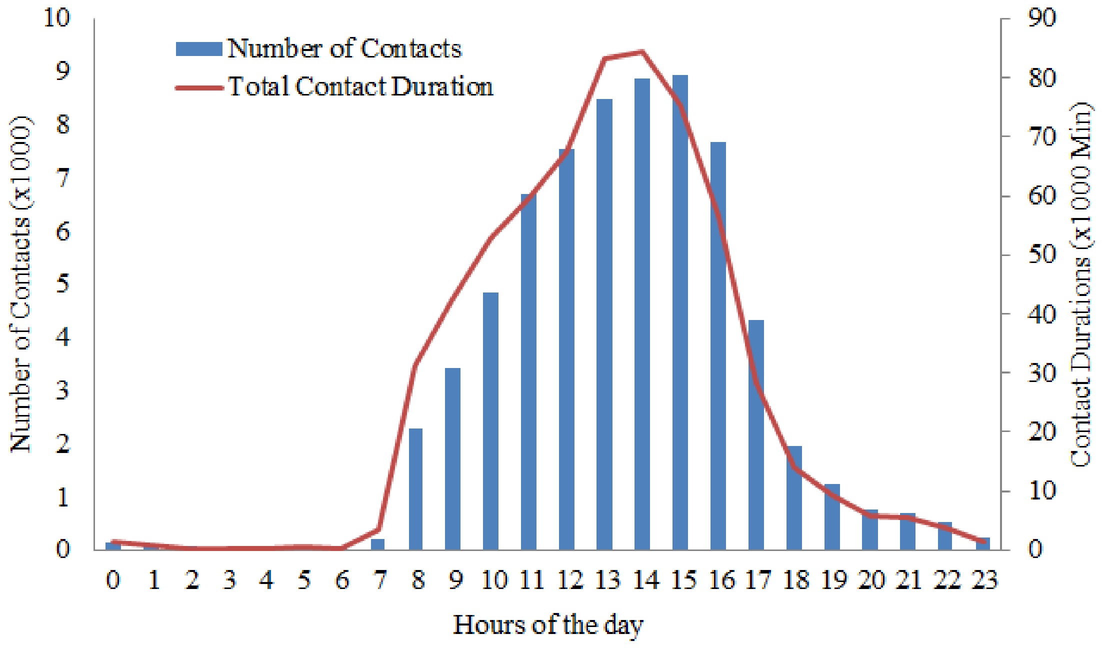


Figure 3.10: Contact duration and number of reported contacts at different hours of a day.

contact durations between Midnight and 6 AM is negligible. The number of contacts increases at 7 AM, reaches its peak at 3 PM, and decreases during afternoon and evening. Contacts reported as changing RSSI values were recorded as unique contacts. This peak is strongly dependent on the participants chosen for the experiment and can change in different human environments. In this experiment, the majority of participants were students with their own work schedules, and the minority were staff with fixed working hours. However, these schedules tended to overlap in the afternoon, increasing the probability of contact.

The average duration of contacts at different hours of the day is interesting for both application areas. As shown in Figure 3.11, contacts which happened in the morning between 7 AM to 9 AM are longer than those occurring during other times of the day. This can be explained by considering the differences between staff contacts and student contacts. Staff contacts primarily happen in working hours, and they present longer contacts as they mostly stay at their primary location in contact with other staff in their proximity. The dominance of these contacts during the working hours results a longer average contact duration. On the other hand,

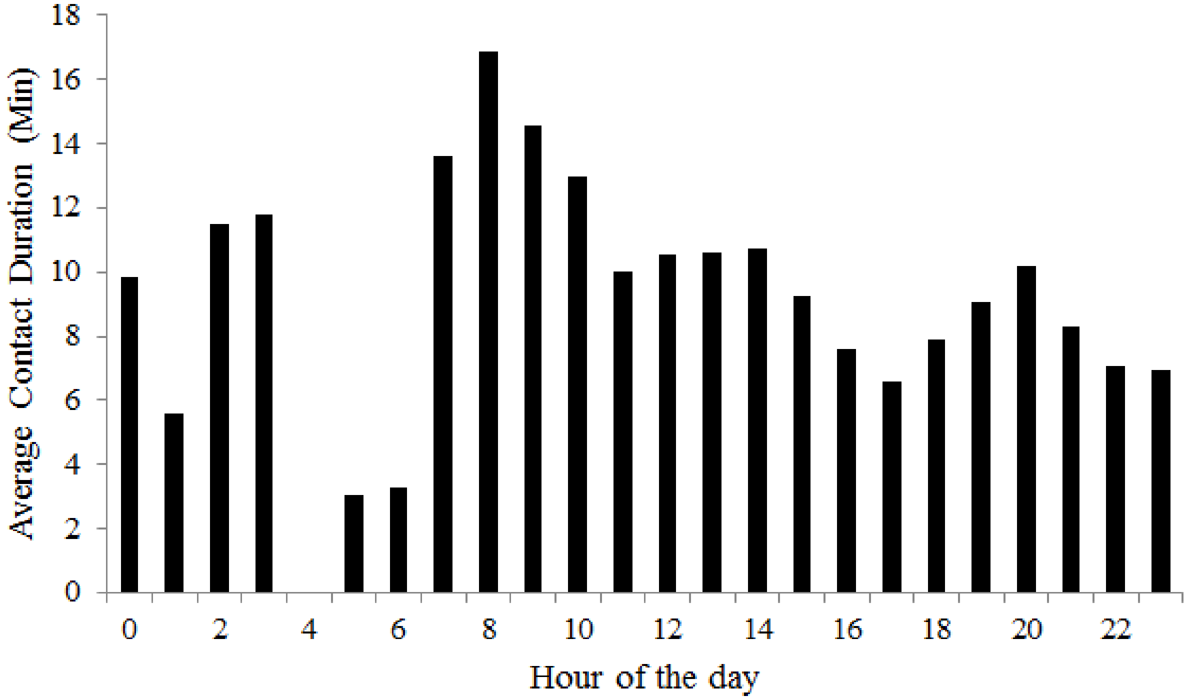


Figure 3.11: Average contact duration at different hours of a day.

contacts in the afternoon and evening are dominated by students in the department, which can be shorter and not tied to a fixed location, due to their freedom in mobility and working location. During midnight to 6 AM, although the contact durations are similar to the contacts in the afternoon, the total number of contacts was unsurprisingly much smaller.

Figure 3.12 shows the complementary cumulative distribution function (CCDF) of contact duration. The CCDF for all the collected data is shown with a dashed line. The figure also shows the CCDF with contact durations of more than 10 hours (0.03% of total reported contacts) removed, since contacts of this duration were likely due to human carelessness in leaving motes abandoned near each other. In this figure, the graph of the pruned data is shown with a solid line. Removing this section of data yields a considerable difference in the resulting plot. Actual data of this duration would have required that participants be so synchronized in their movements they used the washroom together. Although the graph shows a small fluctuation between 300 and 400 minutes, it is difficult to separate the cause of the departure from the curve, because while it

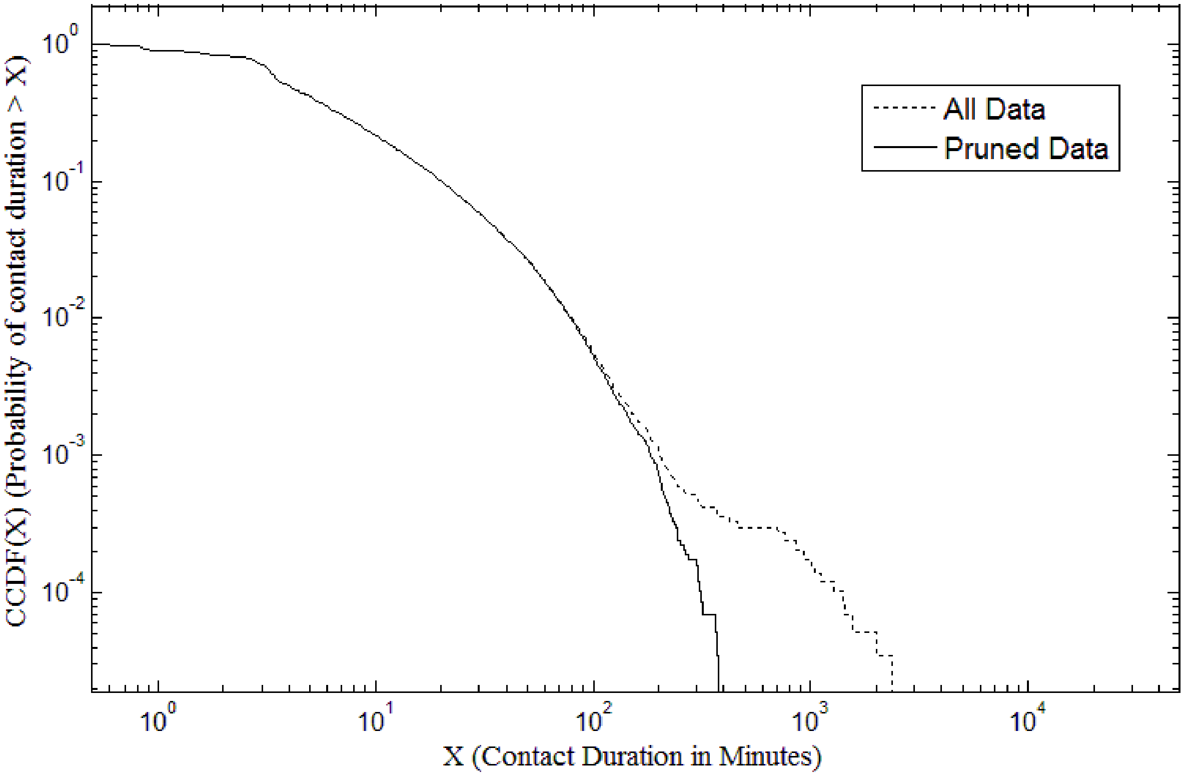


Figure 3.12: CCDF of collected data.

is likely due to abandoned nodes behaving like stationary nodes, it could be caused by people changing their behavior when in continuous proximity.

Figure 3.13 and Figure 3.14 show contact duration and number of reported contacts for all mobile nodes during the experiment, respectively. These graphs are plotted for ‘CLOSE’, ‘MEDIUM’, and ‘Total’ reported contacts. All of the contact proximities show the same trend in data: there are fewer nodes which have long contact times and a large number of reported contacts, while there are many nodes with small to medium contact time and fewer contact records.

Figure 3.15 shows the duration of reported contacts from three selected participants with all other nodes in network. Other nodes in contact with primary nodes are sorted by duration of

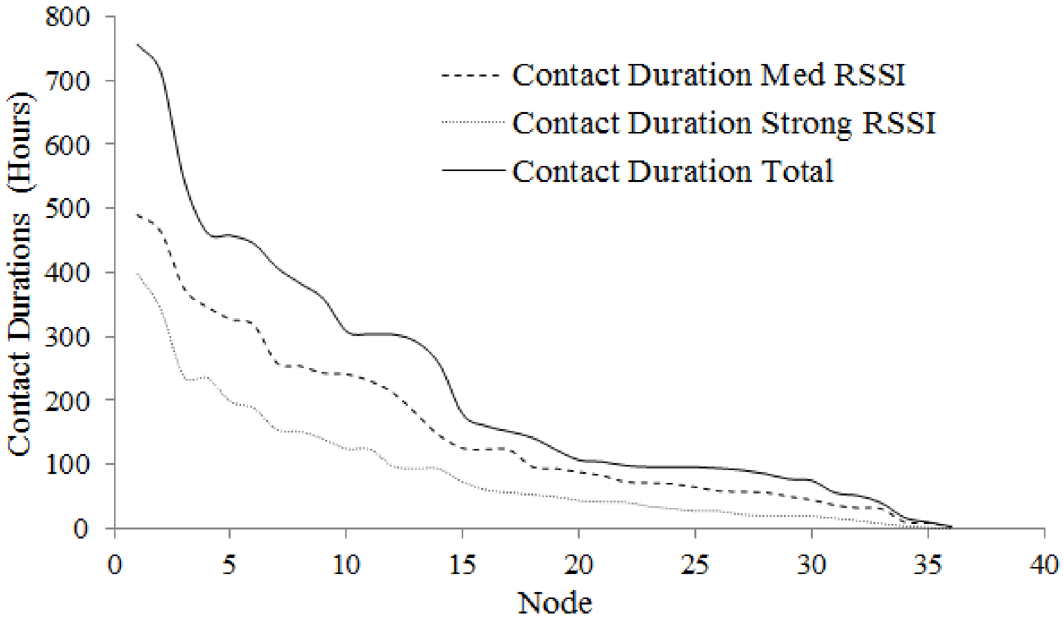


Figure 3.13: Reported contact duration by each node during the experiment. Nodes in X axis are sorted based on contact duration.

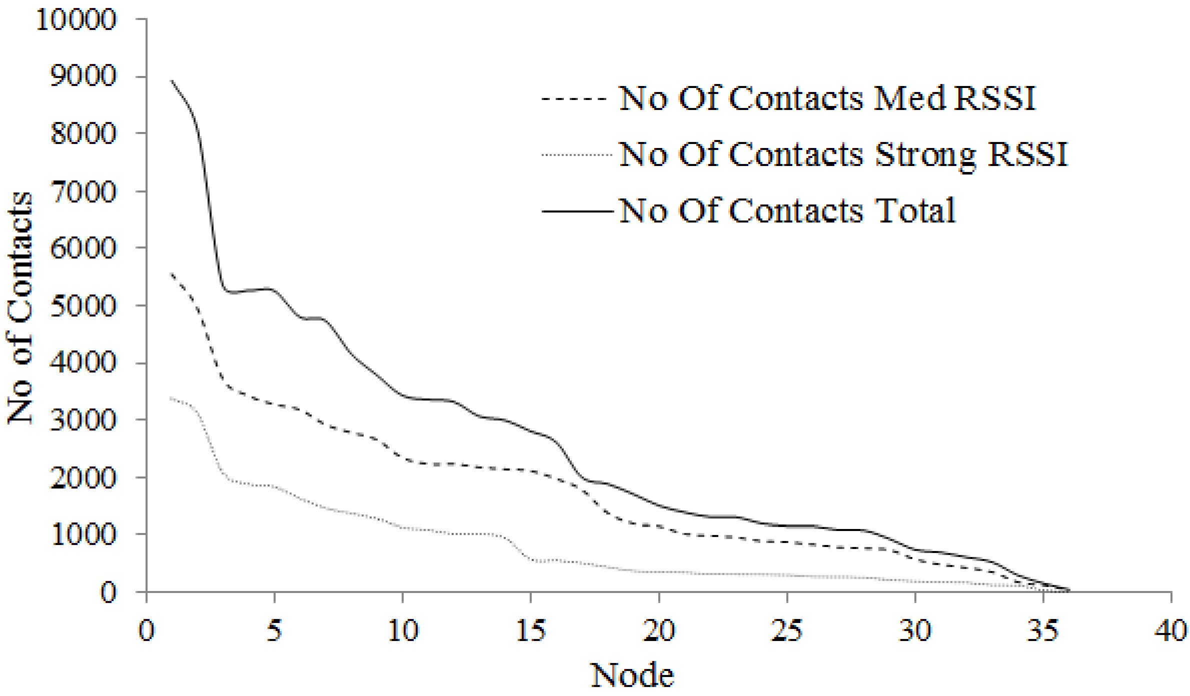


Figure 3.14: Reported number of contacts by each node during the experiment. Nodes in X axis are sorted based on contact duration.

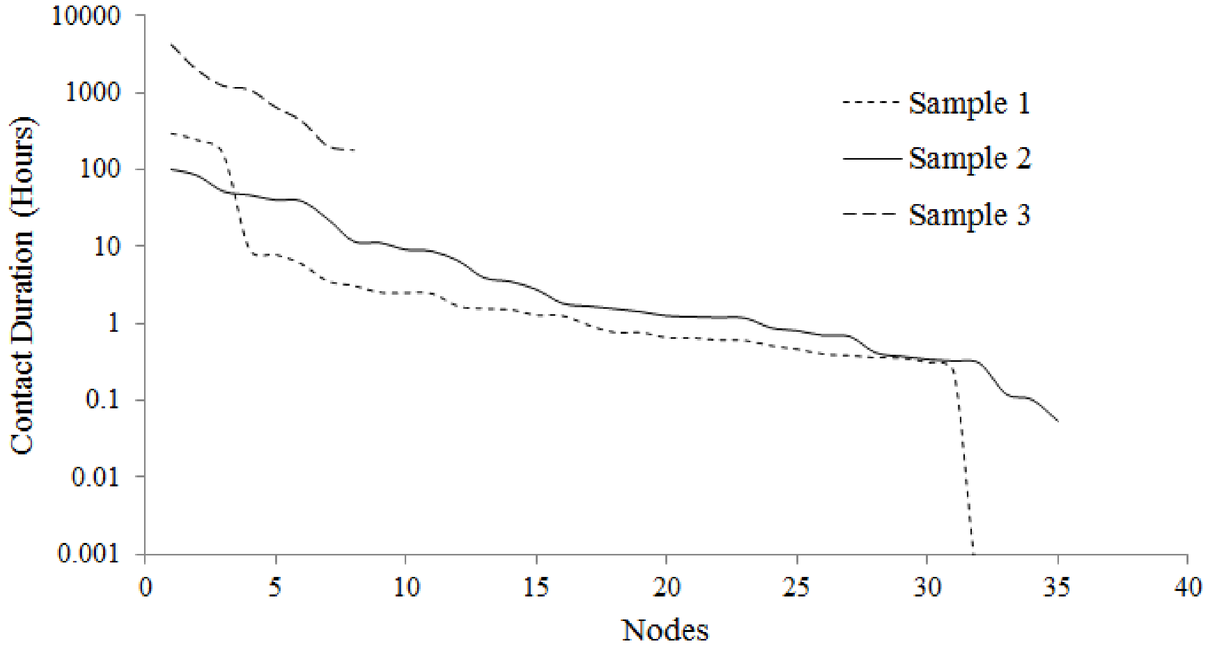


Figure 3.15: Duration of reported contacts between three sample participants and other participants during the experiment.

contact and shown on the X axis. Out of 36 participants, 3 of the plots were similar to ‘Sample 2’, which shows people with high centrality. Even though these people do not have extended contact times with other people, they have seen all other participants at least once during the study. The plot for the other 33 people was similar to either ‘Sample 1’ (representing the participant with the highest number of visited nodes, but not all nodes) or ‘Sample 3’ (representing the participant with the lowest number of visited nodes). All the other 31 graphs fall between Sample 1 and Sample 3. Sample 3 represents the participant with the least number of visited nodes, as the line finishes before 10 nodes, and sample 1 shows the participants with the most number of visited nodes (but not all nodes). This is reasonable for people who worked exclusively in a lab, as they spent a considerable amount of time close to their lab mates, but they did not have sustained or frequent contacts with other people in the study.

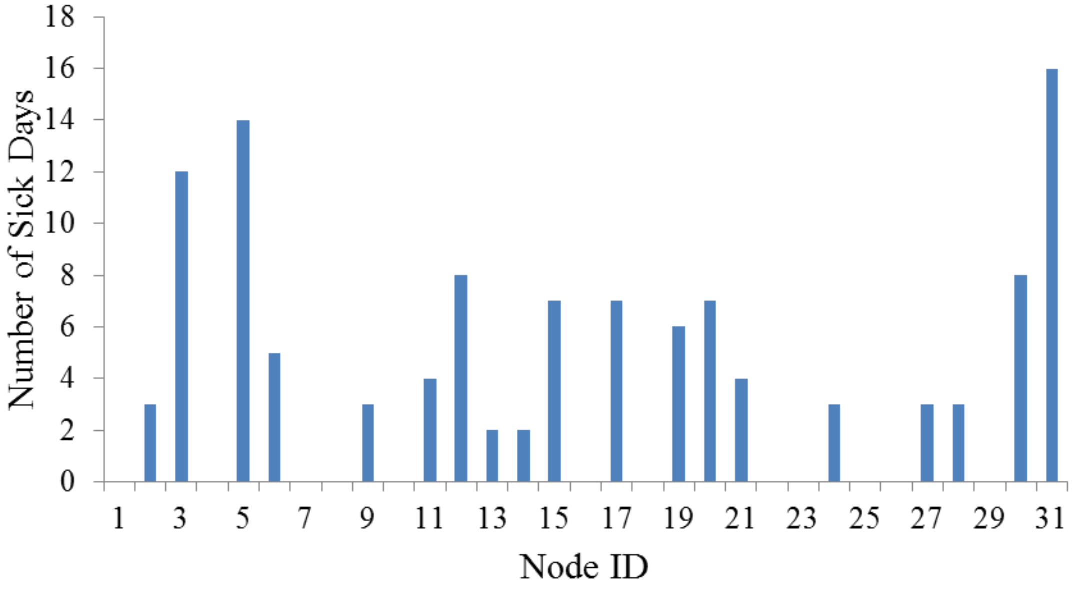


Figure 3.16: Number of reported sick days by each participant.

3.3.3 Health Related Characteristics based on the survey results

Information was recorded on participants' health during the study using weekly surveys. Compliance with the survey was sporadic; out of total 11 posted surveys, 5 participants didn't fill any and just one filled all surveys, while 20 people filled more than half of the posted surveys. For analysis in this section, the 5 participants who completed zero surveys were ignored.

Figure 3.16 shows total number of sick days reported by each participant during the study. These data can be used to analyze the relation between peoples' contact patterns and their likelihood of being ill. Participants were logically grouped into those who did not report a sick day, those that did reported less than 5 sick days, and those with greater than 5 sick days. While it is recognized that processes of viral infection in humans are incompletely captured with aggregate statistics, I sought to determine if there was any correlation between the report of illness and participants' contact patterns in an attempt to validate the automated approach and to inform both future analysis of this data and the collection of new data. Descriptive statistics are shown in Table 3.3.

While there were differences in contact duration for those that reported less than 5 days of illness and those with 5 or more days, there are no apparent differences between those who reported more than 5 sick days and those who reported none. To investigate the relation between number of reported sick days and total contact duration, Pearson correlation applied on these variables which resulted to $\rho = -0.0381$ and p-value = 0.8386. This result rejects any correlation between number of reported sick days and total contact duration in Flunet dataset. While this conclusion cannot be generalized, further research will be required to tease out the underlying causes. It's suspected that some participants who reported no sick days did not fill out the surveys properly, and therefore underreported their sickness, causing a misclassification. However, these human factor issues can be difficult to isolate in data, and may require additional experiments to resolve. The trends noted are also influenced by important confounders, as older participants were more strongly represented in the more than 5 days category, and older participants in the study tended to be office staff, with consistent and prolonged contact patterns rather than graduate students with more sporadic schedules.

Table 3-3: Descriptive statistics about participants with different sick days

	Never	**< 5 Days**	**≥ 5 Days**
No. of People	12	9	10
Mean Duration – Close (Hour)	1.42	1.08	1.42
Mean Duration – Any (Hour)	3.23	2.52	3.19
Mean Contacts	36.37	29.18	32.2
Male/Female	75%/25%	89%/11%	60%/40%
Flu shot	0%	11%	10%
Mean Age	29.8	28.7	31.2
Std Dev Age	7.3	5.6	10.9

3.4 Discussion

This chapter presents the detailed design and implementation of proposed data collection system which resulted in a new dataset incorporating automated contact monitoring using MicaZ motes and health survey data. Our contact pattern measures are similar to previously reported analyses [6], particularly the contact frequency and duration distributions.

In addition to the proximal contacts between participants, the collected dataset also represented their geographical contacts regarding the amount of time they spend at different high-traffic public locations. I believe, similar to proximal contacts, regularities can be seen in participants' geographical contacts. Analyzing and revealing these potential regularities and patterns can be useful in different applications.

In following two chapters, the collected dataset is used for two different purposes. The focus in DTN realm is on geographical contacts and their benefits in routing efficiency when used in combination with proximal contact data. It will be shown how resources at different locations can be used as relay nodes in order to improve the routing performance. In the epidemiology and health modeling realm, the focus will be on minute-resolution proximal contacts and use them as the contact pattern for our designed infection transmission model. This high-resolution contact data can reveal new aspects of the relation between contact density and contact diversity with rate of infection transmission.

CHAPTER 4: DTN ROUTING PROTOCOL FRAMEWORK

Today, massive numbers of devices capable of wireless communication can be found in most parts of the world, such as PCs, or cell phones carried by people. We can imagine each of these devices as a node potentially connected to all other nodes around it via a short-range connection, and could therefore use this opportunistic connectivity as an infrastructure for a low price, large-scale ad-hoc network. Although several authors [35, 12, 11] have focused on different methods and protocols to improve the performance of this type of network, shortcomings in this area have limited the scope and performance of the results, primarily because the network depends on human mobility patterns. In this work, the Flunet dataset is used to explore the impact on network routing performance by focusing on human-location contact, and use this pattern to improve the performance of DTN routing.

4.1 Introduction

With an increasing number of Bluetooth-enabled cell phones and smart devices, I approach a world where people can create dynamic networks as they move. This introduces the potential for a new set of applications that do not require full connectivity and offer location- or situation-related information to their users, at a lower price than WiMax or GPRS. Even in a WiFi-enabled environment, applying this type of network can have a significant effect on network's performance [63]. These opportunistic networks are considered as a subset of Delay Tolerant Networks, where packet delivery is more important than latency.

DTNs, their utility and properties were first described by [9]. Since then, designing efficient routing protocols with an acceptable performance has remained an open issue. Creating a network infrastructure based on human dynamic networks strongly depends on designing a routing protocol which can deal with a DTN's intrinsic characteristics such as a lack of an end-to-end path between source and destination, opportunistic and semi-predictable connectivity, and resource allocation constraints [9]. The early routing algorithms for DTNs were constructed to handle the general case, without using any well-defined assumptions regarding the nodes, or the network’s surrounding environment, and therefore they underutilize the potentially available existing patterns. This focus on generality leads to slower or more resource-intensive routing protocols with insufficient performance for human centered data applications [33].

An early example of the use of human mobility patterns in DTN routing was the Prophet algorithm [11], which was based on the simple assumption that nodes encountered more frequently in the past are more likely to be encountered in the future. While the algorithm achieved substantial savings over the baseline flooding algorithm [10], the overhead incurred by transmitting contact tables made multihop routing prohibitively expensive. In [12], Hui et al. described BUBBLE, a routing algorithm based on properties of social networks. BUBBLE first attempts to label cliques of individuals, then attempts to route packets through highly connected individuals both between and within cliques. BUBBLE provided a performance advantage over Prophet, but also potentially can introduce a considerable overhead due to the need for dynamically distributing the data about cliques’ structures through the network. Other shortcomings of BUBBLE were that the work did not attempt to discern why individuals were highly connected, and the algorithm imposed a disproportionate burden on highly connected nodes. In [44], Tian and Li offered a similar approach to BUBBLE, but clustered people by the time they spent in different locations instead of according to their social contacts. In other words, each node was assigned to a physical location component based on the amount of time they spend there, and the packet owner attempts to pass the packet to the nodes which usually appear at the same location as the destination. Regardless of the importance of the insights provided by this work on humans’ visitation pattern to different locations, the shortcomings of BUBBLE apply to this work as well. Although these algorithms led to significant improvement with regards to specific performance metrics, the complexity of human behavioral patterns leave uninvestigated many potential assumptions for exploitation, particularly the relationship between human dynamic networks and physical locations.

This chapter focuses on human dynamic networks and studies the effect of visiting public places (i.e. high traffic physical locations) on routing performance. It is suspected that people not only have a regular pattern in contacting other people, but they also have a regular pattern in visiting different places around them, and this regularity can be used together with contact pattern regularity to improve the performance of routing in DTNs with human agents. I also believe a subset of highly connected people are highly connected not because they are particularly social, but because they are often stationary in a high-traffic area, for example, a receptionist in a front office area.

To validate these assumptions, Flunet, the dataset described in Chapter 3, will be used. Based on the assumptions and observations from the data set, a new routing protocol is proposed, location based routing (LBR), which takes advantage of regularities by adding stationary nodes to frequently visited locales. These stationary nodes can be either nodes with nearly unlimited resources (e.g. a Bluetooth-enabled PC) or a backbone which acts as an asynchronous bridge between source and destination. Algorithm performance is evaluated against the Flooding and Prophet algorithms using random packet generation over the contact record from the experiment. The overall efficiency is evaluated as well as the impact on mobile nodes which are proximate to the high traffic locations.

4.2 Background

As in common with other researchers [43, 44] the primary interest is in the impact of place on the efficiency and utility of DTNs, and in particular on PSNs. People inhabit spaces and knowledge of which spaces, the probability that specific individuals will inhabit them, and for how long, can lead to better DTN routing algorithms. In the particular case of isolated PSNs, packet transfer can only take place when nodes are collocated. This leads to the interesting possibility of placing static nodes at locations known to be visited frequently, or even using potentially available static nodes in these places to facilitate routing.

4.2.1 Assumptions

The following assumptions are made to facilitate the analysis, beyond the common DTN and PSN assumptions such as delay tolerance and mobile power limitations.

- **Agency:** It is assumed that all mobile nodes in the network have agency, and cannot be influenced to change location externally.
- **Isolated:** It is assumed that the nodes are isolated by choice or circumstance from any form of network separated backbone (such as is posited in [12, 43]).
- **Heterogeneous:** Because the impact of introducing stationary nodes is considered, the strict power and memory requirements on those nodes can be relaxed, as they can be connected to larger or continuous power sources and are not as space constrained for memory as in [64].
- **Sparse:** Finally it is assumed that the network is sparse. If possible, locations are represented as a graph of unknown connectivity with L vertices, and there are N mobile agents traversing this graph, then it is assumed $L >> N$.

4.2.2 Problem Description

If additional resources are allocated to a network, the network designer must determine where to place them, and the policy designer must determine when to use the resources and when to ignore them. These questions are primarily addressed through simulation analysis over an empirical contact data set, but to illustrate the viability of stationary nodes' utility a simple statistical argument is provided.

Consider two nodes, a and b, on a graph of L location vertices, where communication between nodes can only happen at vertices as in [43]. If there is no information about the graph structure or the temporal behavior of a and b, then the safest assumption is that the probability of a and b being in locations n or m are unknown independent distributions. This is a worst-case scenario for algorithms like [12], as they depend on the correlation between contact patterns and measures of time, place and circumstance to improve performance. The probability of a packet being delivered from a to b under the independence assumption is simply the probability that a and b are collocated:

$$P_d(a,b) = \frac{\sum_{i \in L} P(a,i)P(b,i)}{\sum_{n \in L} \sum_{m \in L} P(a,n)P(b,m)} \tag{4.1}$$

where $P_d(a,b)$ is the delivery probability between a and b and $P(z,k)$ is the probability of the z^{th} node being at location k. For uniform distributions, this expression reduces to simply $1/L$. However if a has previously delivered a copy of the packet to a node x which is stationary at location j, then the probability of at least one packet being delivered is the probability of node b being at location j, or the probability of a and b being collocated:

$$P_d(a,b,x) = P(b,j) + P_d(a,b) - P(a,j) \tag{4.2}$$

Obviously $P_d(a,b,x) \geq P_d(a,b)$. Note that this includes the possibility of a and b being collocated at j, and more than one packet transmission occurring. The at-least-one-packet-delivered formalism is employed to make the results more concise, but the analysis that follows holds equally true for the exactly one packet delivered formulation.

By adding x, resources are added to the system, so one would expect the probability to increase. Consider adding an additional mobile node c, which has the same packet as a to deliver to b. In this case, the probability of at least one packet delivery is the probability that a and b or b and c are collocated, or:

$$P_d(a,b,c) = P_d(a,b) + P_d(b,c) - P(a,c) \tag{4.3}$$

I know that $P_d(a,b,c) \geq P_d(a,b)$, but what can be concluded about the relationship between $P_d(a,b,c)$ and $P_d(a,b,x)$? Without knowing the actual distributions it is impossible to conclude anything concrete, because it has to be evaluated whether it is more likely that b will visit j than contact c. In the uniform distribution case, which is the least predictable, $P_d(a,b,x) = P_d(a,b,c)$.

Assuming that all mobile nodes have agency (A1), there is no control over the trajectory of *c*, but a designer can pick *j* to be in locations that maximize the probability of delivery, or for opportunistic use of existing resources, which resources are potentially useful. While it cannot be concluded that adding stationary nodes will always be better than adding mobile nodes, it can be expected to have a greater degree of control over the return on placing or choosing stationary nodes. However, as it is assumed, $L >> N$, so it may be difficult to find places which facilitate routing sufficiently to justify the additional resource. Based on other work [65, 66] it seems logical to conclude that at an individual level, the numbers of locations regularly visited in *L* is actually quite small. If there exists a subset of agents which visit similar subsets in *L*, then it is plausible that placing additional nodes at specific vertices on *L* could be advantageous.

Finally, the impact of adding stationary nodes is considered as it affects the mobile nodes surrounding them. Many algorithms (e.g. [12, 39]) employ estimates of a node's popularity (as measured by frequency of contacts) to attempt to enhance the probability or speed of packet delivery, by selectively routing to more popular nodes based on some measure of utility. However, it is possible that a mobile agent's popularity is due to their residence in a popular place, for example, a receptionist in a busy front office or a waitress in a busy restaurant. Rather than burden a resident's agent with additional packets, it seems sensible to preferentially route through the stationary node due to its larger data store and power reserve (A3).

Given that it is possible to estimate the distributions for both contact patterns and locations using modern monitoring techniques [27, 46] we can do substantially better than the assumptions of uniform independent distributions, as the distributions are never uniform (unless the node can teleport to another node without crossing intervening nodes) and not necessarily independent for nodes with high probabilities of contact. The remainder of this chapter focuses on an examination of the role of stationary relay points in PSN, given one specific dataset which contained mobile and stationary nodes, and the development of policies to select stationary nodes, route through stationary nodes and balance mobile node loads using stationary nodes.

4.3 Algorithm Design

I have implemented two standard algorithms and three variants of a novel algorithm, Location Based Routing (LBR). The standard algorithms are used as a base comparison to understand the tradeoffs and benefits conferred by the new proposed algorithm against both of the standard algorithms and against more novel algorithms such as [12, 43, 44] which also use these algorithms as a basis for comparison. Several variants of these standard algorithms exist, so the specific implementations which are employed for clarity are briefly described.

4.3.1 Reference Algorithms

Epidemical routing (ER) [10] is a standard benchmark algorithm in DTNs and sensor networks. ER passes a copy of all packets in its buffer to each node regardless of the likelihood of seeing the destination. The implementation in this work exchanges the packet list at each node to determine which packets each node requires, then synchronizes the packet list, implying that each packet is passed to each node exactly once. This is in contrast to other implementations that do not pass packet lists, but simply broadcast all available packets whenever nodes meet.

PRoPHET [11] is the simplest history-based routing system, positing only that the likelihood of one node contacting another is equivalent to the likelihood of such an encounter occurring encountering them in the past. PRoPHET nodes exchange changed entries in their routing tables every time they meet, detailing their past contacts with other nodes. More aggressive variants of PRoPHET are possible that record the cost of multihop paths, and exchange these likelihood tables. However, our implementation only transfers the likelihood for single hop paths.

4.3.2 Location Based Routing (LBR)

LBR is a variant of PRoPHET that takes the role of stationary nodes into account. LBR assumes that stationary nodes have superior resources, and can therefore buffer more data for longer. The only drawback of transmitting data to a stationary node is the cost incurred on the mobile node to transmit the packet. Therefore it is sensible to spread packets to as many stationary nodes as possible to increase the chance of transmitting to the target node. This section presents a detailed description of LBR in addition to different variations of it.

4.3.2.1 Packet Transmission

Packets can be generated by a mobile node and are always destined for another mobile node. A history-based routing protocol is used for mobile-mobile packet forwarding. Each mobile node records their contact duration with other nodes in the network. When two mobile nodes meet, they broadcast a *Delivery Probability Value* or *DPV* for each of the available destinations (i.e. nodes with packets buffered for them in these particular nodes), using a packet format similar to Figure 4.1.a. DPV shows the probability of transferring a packet between two nodes and can be a function of a history-based visit probability, the node's resource availability, or other parameters typically considered in DTN routing algorithms. In this implementation, the DPV depends solely on the probability of visiting the destination based on the contact duration history. After receiving the list of DPVs from adjacent nodes, if the receiver has a better DPV for one or more destinations in the list, it returns a packet with all packet IDs for which it has a higher DPV. The format of this packet is shown in Figure 4.1.b. The sending node receives all responses from adjacent nodes (if there are any) and for each available destination sends all the packets for that destination to the node with higher DPV (Packet format is shown in Figure 4.1.c). The sender also marks all passed packets as *Forwarded* (described in 4.3.2.2); therefore in the future the sender delivers these packets only to the final destination.

Stationary nodes are considered stationary buffers; however this algorithm could be extended to incorporate stationary nodes forming a network backbone as in [64]. When a mobile node visits a stationary node, the stationary node looks in the buffer for all the packets available for the recently-arrived mobile node, and delivers them to the destination (using a packet format is shown in Figure 4.1.d). The mobile node also sends packet IDs for all of its packets to the stationary node (packet format in Figure 4.1.e), so the stationary node can request the packets which are not available in its buffer. The packet request at this stage happens by using a packet format similar to that shown in Figure 4.1.e. After the mobile node has received a packet request

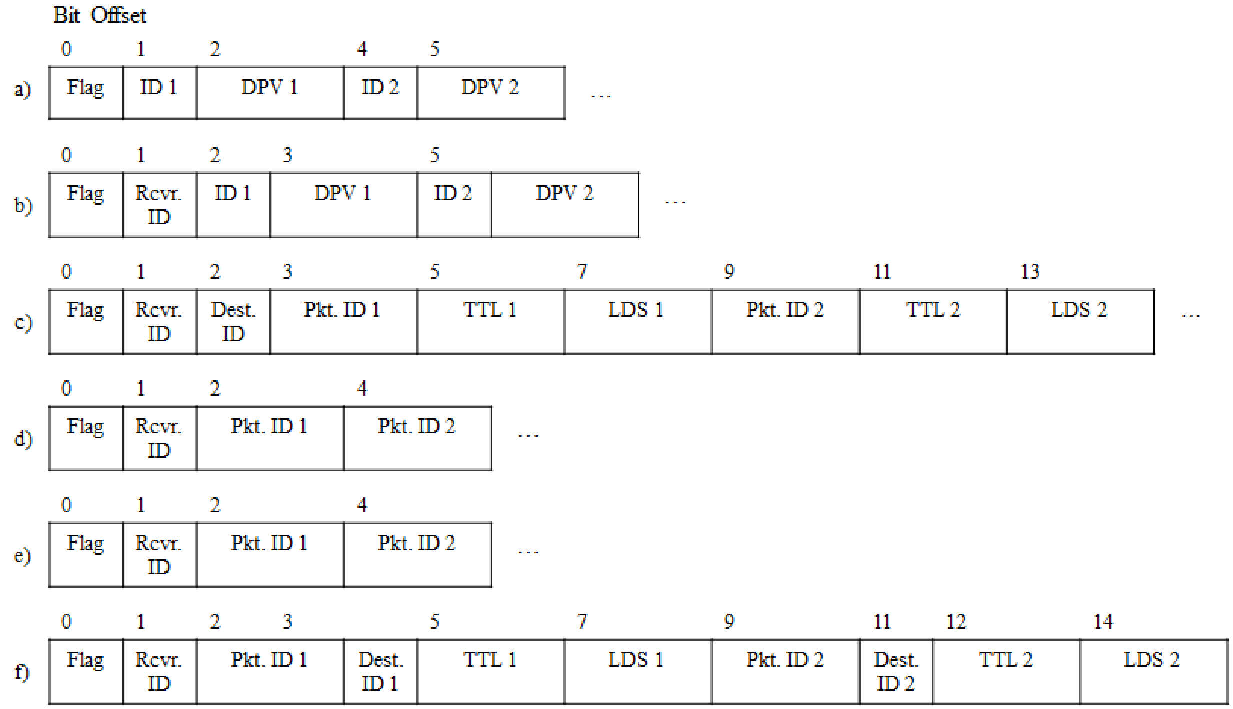


Figure 4.1: Different packet formats used for communication between mobile and stationary nodes in LBR. Description of each packet type is included in the text.

from stationary node, it sends a copy of all requested packets back in a packet similar to Figure 4.1.f.

4.3.2.2 Buffer Management

In order to utilize the available buffer space, each node (mobile and stationary) tries to fill its remaining buffer with lower prioritized packets, including overheard packets, timed-out packets, or forwarded packets. These packets can be delivered directly to the destination, or they can be deleted to prevent buffer overflow, but they can't be passed to any other mobile or stationary nodes. In order to define relative priority of deletion, four types of packets are defined as follow:

- **Main packets**: Primary copies of live packets which the node has received for delivery to the final destination, which have not been passed to any other nodes by the current node.

- **Forwarded packets**: Includes packets which have been passed to other nodes with better DPV for the destination.
- **Overheard packets**: Packets which have been recorded due to overhearing a transaction between two adjacent nodes.
- **Timed-out packets**: Packets for which their time to live (TTL) has expired.

Each node assigns a delete priority to each of the packets in its buffer, which shows how beneficial the packet is. A low delete priority for a packet increases its chance to be deleted. Main packets have the maximum delete priority, as the current node is the primary packet holder. Packets from other types, i.e. forwarded, overheard, and timed-out are assigned a delete priority based on the DPV between the node and the destination.

As long as the node has sufficient buffer space, it keeps all of the available packets. When a buffer overflows, packets are deleted in reverse priority, starting with timed out packets and ending with main packets.

4.3.2.3 Anti-Packet System

The algorithm includes a passive anti-packet system to help remove delivered packets from the network while introducing minimal overhead to the system. Each node has a list consisting of IDs of delivered packets. When the node either delivers a packet or sniffs a final packet delivery, it adds the packet ID to the list. Meanwhile if the node sniffs any communication between two adjacent nodes about a delivered packet (a packet whose ID is in the delivered list), it broadcasts an anti-packet message (Figure 4.1.g) so all the surrounding nodes will delete that particular packet and update their delivered list.

4.3.2.4 Load Reduction (LBR-RL)

I suspected that the popularity of some nodes in the network was due to the popularity of their location, and therefore their contact pattern should be similar to the contact pattern of their associated stationary nodes. To check this hypothesis, two factors are measured: first, the load on each mobile node (through a test simulation run), and second, a simple metric to measure the contact similarity between each pair of mobile and stationary nodes. The load on different mobile nodes is shown in Figure 4.2, where Y axis shows the number of transmitted packet via the node and the X axis shows the node number, sorted by transmitted packets during simulation.

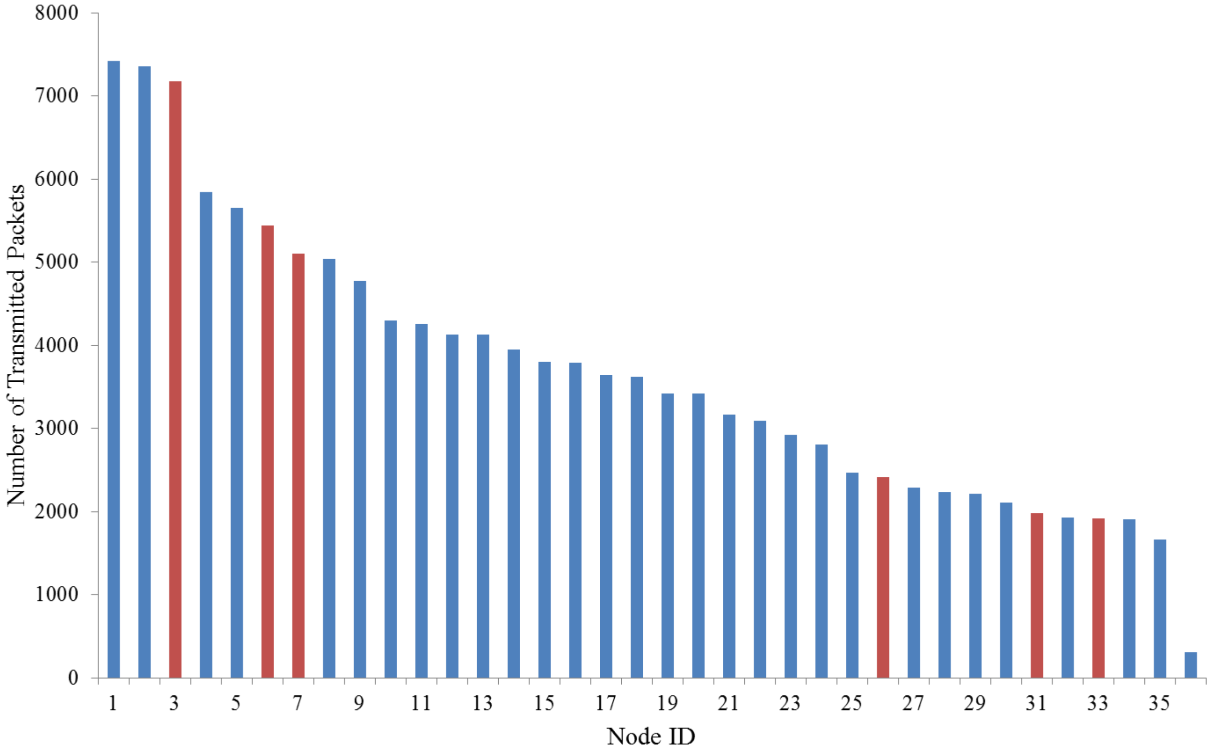


Figure 4.2: Load on each mobile node in the network.

The similarity between each mobile and stationary node pair in the network is defined as total duration which both nodes spent connected to the third node, and is defined as follow:

$$\psi(s,m) = \sum_{i \neq m}^{N} min\big(T_w(s,i), T_w(m,i)\big) \quad (4.4)$$

Where m and s represent mobile and stationary node respectively, i represents all the other mobile nodes in the network, and $T_w(a,b)$ denotes the contact duration between nodes a and b. The cost function measures the size of the intersection between the contact patterns of the stationary and mobile node. Contact intersection is employed as the metric to capture cases where one node's contact list is a subset of the other, which it would be considered as similar. The larger the value of ψ, the larger the intersection between the mobile and stationary routing tables.

The graph for the $\psi(s, m)$ function between each stationary node and all mobile nodes is shown in Figure 4.3, while each bar in each graph shows the similarity between the mobile node with ID specified at X axis and the stationary node mentioned in the graph title. As it can be seen, around 97% of the mobile-stationary node pairs have less than 3 hours similar contact per day with all the other mobile nodes in the network ($T_w(m, s) < (3 * 3600 * 92)$).

I defined a mobile-stationary node pair as similar if their total contact duration per day includes more than 4 hours of similar contact. A similar contact for two nodes is defined as a contact with a third mobile node during the experiment, regardless of time of the contact. Applying this similarity threshold to the Flunet Dataset resulted in 8 similar stationary-mobile node pairs. These pairs consisted of 3 distinct stationary nodes and 6 distinct mobile nodes (2 mobile nodes met the threshold for two different stationary nodes).

The loads on selected mobile nodes are shown with red bars in Figure 4.2. Half of the nodes have a relatively high load (3, 6, and 7), the second half (26, 31 and 33) carry more modest load in the network. Two of the mobile nodes with high load (index 3 and 7) both showed a high contact similarity with stationary nodes located at 'Lab 1' and 'Lab 2' (Figure 3.9), while the other node with high load (index 6) and all of the nodes with modest load showed a high contact similarity with the third stationary node located at the 'Main Office'. The 4 staff participants in the main office created a sub-network in the system, where they each had long contact durations with the other staff and with the stationary node, while only one of them had a desk at a location

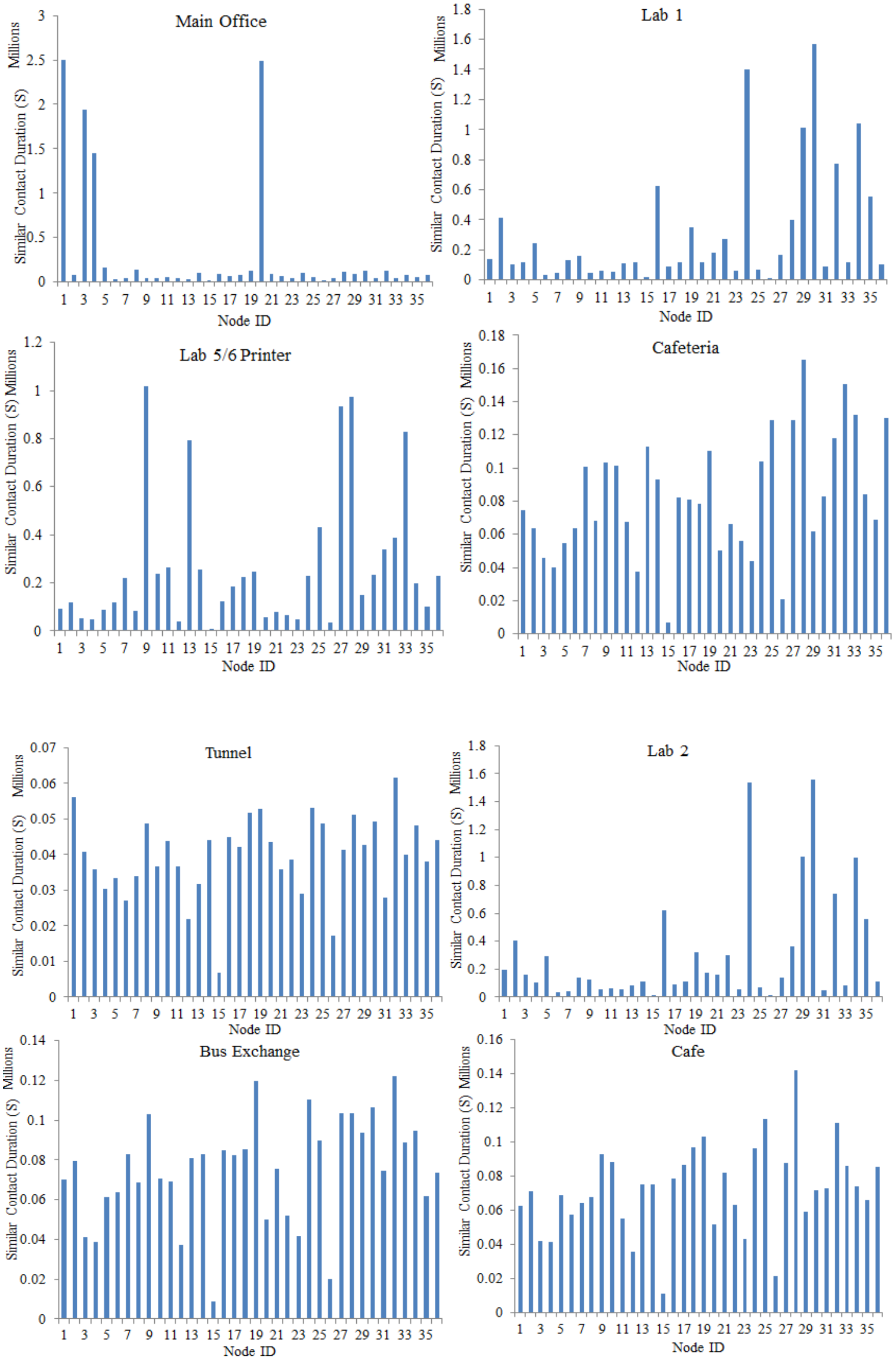


Figure 4.3: The similarity between each stationary node and all mobile nodes in the network.

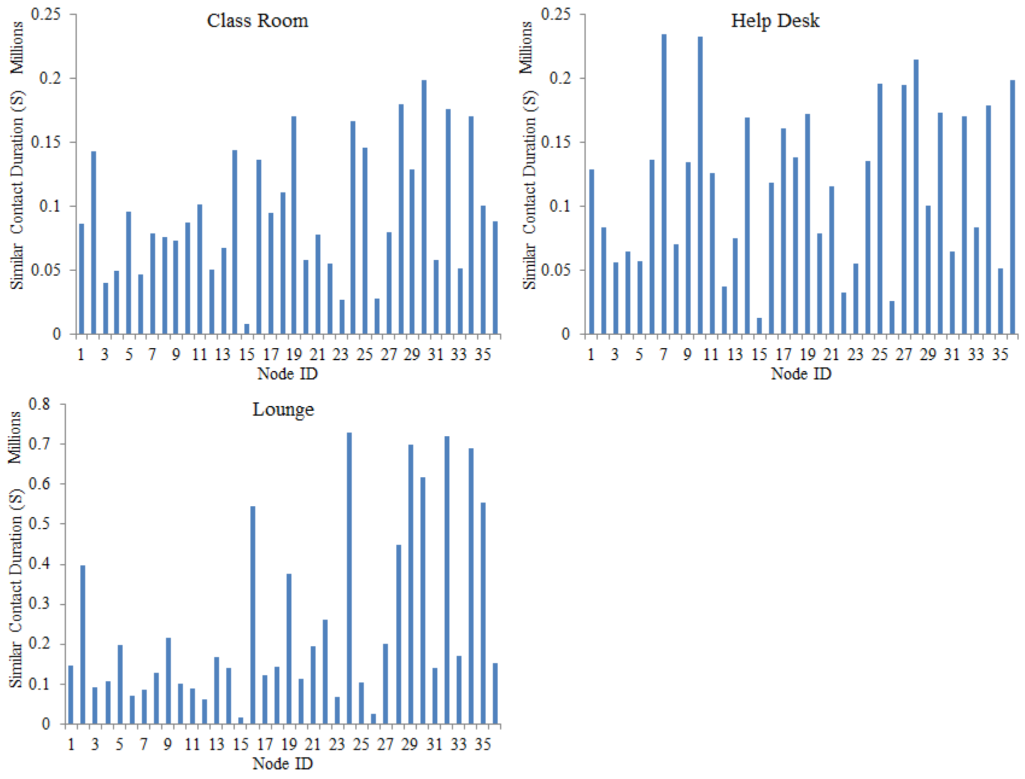


Figure 4.3 (Cont.): The similarity between each stationary node and all mobile nodes in the network.

that allowed visitor contacts as well. Therefore the high contact similarity between the nodes with low load are due to long contact duration with a small set of nodes, and low contact span.

From the graphs in Figure 4.2 and 4.3, it can be concluded that some of the nodes which act as relay nodes have a similar contact pattern to the stationary nodes near their primary location. In other words, the popularity of these nodes could be due to people who visit the location and not the person. The load on these nodes can be diminished by preferentially routing to stationary nodes if both the stationary and the similar mobile node are present.

I designed Location-Based Routing Reduced-Load (LBR-RL) protocol as a variation of LBR to be capable of capturing this effect. In this variation, a mobile node is considered as associated with a stationary node if it has more than 4 hours of similar contacts with the stationary node (based on Equation 4.4) over the experiment. If a mobile node is associated with a stationary node, forwarded packets follow a different policy in two cases: first when the mobile node is in contact with the associated stationary node, and second when the packet has already been buffered at the associated stationary node. For this purpose, when the packet owner determines the associated mobile node as the selected relay node, it waits for two time units (two 30-seconds) and probes the associated stationary node. If the packet owner finds itself connected to the associated stationary node, it sends the packet to the stationary node instead. If the packet owner is not connected to the stationary node, it checks whether the associated stationary node has received the current packet or not. If true, the owner skips transmitting the packet to the mobile node, otherwise it forwards the packet. Hence for each mobile-stationary node pair, LBR-RL limits the communication between the associated mobile node and other mobile nodes in the network to the packets which haven't yet been buffered in the associated stationary nodes.

Applying this policy limits the communication between mobile nodes associated with a stationary node and all other mobile nodes in the network only to the time slots when mobile and stationary nodes are not connected and only to the packets which already have not been buffered in the stationary node. In this way, the majority of the packet forwarding is to the stationary node instead of to the associated mobile node, which, by considering the assumption of contact similarity between mobile and stationary node, should reduce the load on the mobile node without considerable effects on routing performance. Also, the policy introduces the least amount of control-packet transmission in the network, which helps keeping the total overhead at a low level.

4.3.2.5 Popularity of Stationary Nodes (LBR-PS)

Previous studies on human social contact patterns have shown that different people have different popularity and centrality in the network. It is suspected that the difference in popularity can be seen in stationary nodes as well. This hypothesis is reasonable by considering the stationary nodes as agents which reflect the visiting pattern at their location, and knowing that different places have visiting patterns based on the importance of the service they offer to the population. These facts lead to different popularity level for each public location, and subsequently for each stationary node placed at such a location.

Some basic observations of the Flunet dataset also strengthened the hypothesis of different popularity and centrality for different stationary nodes. In the Flunet dataset, two of the stationary nodes were placed close to the main campus bus exchange. However, based on self-reports, only 17% of the participants primarily used public transit to commute. Therefore, for most of the participants in Flunet, the Bus Exchange is not a popular place with frequent and regular visits. Note here it is not claimed that places such as the Bus Exchange are not popular locations, but it argues that the service offered at this location is not of interest for most of the sample population, while it may remain a place with frequent and regular visits for many people outside of our sample.

This popularity difference between different public locations make some of the stationary nodes associated with these locations high centrality nodes in the networks and good candidates as relay nodes, while it gives less centrality to other stationary nodes. This heterogeneity can be used to increase the routing performance in the network, as routing all packets to all stationary nodes might lead to wasted resources with only modest gains in efficiency.

Location-Based-Routing Popular-Stationary (LBR-PS) is a variant of LBR which takes this hypothesis into account by dividing stationary nodes into two groups. In the first group, called *Ordinary Stationary Nodes*, the stationary node receives the packets based on its chance to be visited by the destination compared to the corresponding chance for adjacent nodes. While in the second group, called *Popular Stationary Nodes*, the stationary node receives a copy of all the live packets in the network, regardless of its chance to be visited by their destination.

A stationary node is marked as popular based on the total amount of time it spends in contact with mobile nodes, regardless of the number of nodes. This metric captures both high churn locations (high probability of finding an intermediate node, low probability of finding a specific node), and localized areas where specific nodes can be found most of the time (lower probability of finding an intermediate node, higher probability of finding a specific node). A measure such as betweenness centrality might neglect the second case, as the high-duration-low-numbers node would only be on shortest paths to the small number of mobile nodes close to it.

If the stationary node is popular, it acts as a stationary buffer and keeps all of the available packets. Therefore, when it connects to an adjacent mobile node, it broadcasts a message to indicate itself as a popular node and requests for all the existing packets which are not currently buffered in the stationary node. By receiving the popularity-indicator message in the mobile node, a process similar to the default behavior in LBR (explained in section 4.3.2.1) takes place.

Each mobile node waits for 2 timeslots after establishing a connection with a stationary node to receive a popularity-indicator message. If the mobile node doesn’t receive this message, it considers the recently-connected stationary node as an ordinary node and uses the same routing policy as with another mobile node (described in section 4.3.2.1), even though it is fixed at a stationary location.

4.4 Experimental Setup

The algorithm was evaluated using the Flunet dataset in the ns3 simulation environment, as described in section 3.2. Two different application modules were used in ns3 to simulate the operation of mobile nodes and stationary nodes, with a similar set of input parameters applied for both. Table 4.1 shows the parameters and ranges examined over simulation runs. In total, 232 simulations were run. Each such simulation explored the impact of a particular assignment of values to variables. The delivery ratio and transmission overhead of each run were recorded by the system. The delivery ratio was measured to show the percentage of the packets which can be delivered successfully using LBR and the amount of time required to deliver them, and transmission overhead to measure the load introduced on the network by using LBR. The results of these two factors were compared to the benchmark algorithms.

Variations of Epidemic Routing and PRoPHET were employed as benchmark algorithms. Epidemic routing represents the best achievable delivery ratio regardless of transmission overhead because it implements a parallel exhaustive search, and PRoPHET represents a standard intelligent history-based opportunistic routing algorithm. For each of the benchmark algorithms, the same code was used to simulate the operation of mobile and stationary nodes;

therefore mobile and stationary nodes in these algorithms acted identically. The parameters used for these two algorithms and their ranges are also shown in Table 4.1.

Table 4-1: Simulation parameters used as input for each algorithm and the range of values assigned to them.

Parameter Name	**Description**	**Range in LBR**	**Range in Prophet**	**Range in Epidemic**
Packet TTL	Time-to-live for each packet.	Changed from 0.5 day to 14 days with 0.5 day intervals, plus infinite TTL.	Same as LBR	Same as LBR
Mobile nodes' buffer size	Number of packets each mobile node could buffer.	Set to 100 packets for all simulations.	Either 100 or infinite packets for all simulations	Same as Prophet
Stationary nodes' buffer size	Number of packets each stationary node could buffer.	Either 100 or infinite packets for all simulations		
Reduced load	Determined if the code should use the reduced load model or not.	True or False	N / A	N / A
Popular stationary nodes	Determined if the code should differentiate popular stationary nodes or not.	True or False	N / A	N / A

In Flunet, mobile nodes represent people carrying mobile devices, and stationary nodes represent wireless-enabled devices that can be found in public locations. People send messages to other people through the DTN network, while stationary nodes act as relay nodes. Therefore in all the tested algorithms (Epidemic, PRoPHET, and LBR) only mobile nodes were able to generate packets, and those packets were only destined for other mobile nodes. For each timeslot, each mobile node drew from a uniform random variable with 0.1% probability to generate a packet in a time slot, an average of 3 packets every 2880 time slots (one day). Two different cases were considered for data segment size, based on TinyOS 1.1 limitations: a 0-byte data segment (which represented the lower bound of data segment size), and 129-byte data segment (which is the maximum size in TinyOS).

4.5 Results

This section presents the effect of adding available stationary nodes to the network and shows how existing routing algorithms, like Epidemic routing and PRoPHET, take advantage of them as additional resources. LBR is compared with these results to demonstrate that having knowledge about stationary nodes and knowing the fact that these nodes are stationary helps the algorithm to make better decisions, increasing the routing performance. Subsequently, the performance of the LBR variations (i.e. LBR-RL, LBR-PS, and LBR-RL-PS) are compared with LBR and benchmark algorithms using delivery ratio and transmission overhead, to demonstrate how the described hypotheses can affect the algorithm output.

4.5.1 Effect of Using Stationary Nodes

Figure 4.4 shows the impact of including stationary nodes on the delivery ratio using the two benchmark algorithms: Epidemic Routing and Prophet. For Epidemic routing with a 100-packet buffer (capable of holding 1% of total generated packets), the difference between the delivery ratio with stationary nodes and without stationary nodes decreases as TTL increases, because as TTL increases, the number of live packets in the network increases and therefore many packets will drop due to buffer overflow. In contrast, the difference in delivery ratio for Epidemic Routing with infinite buffer size and the PRoPHET algorithm increases by increasing the TTL, as they don’t suffer from considerable packet drop.

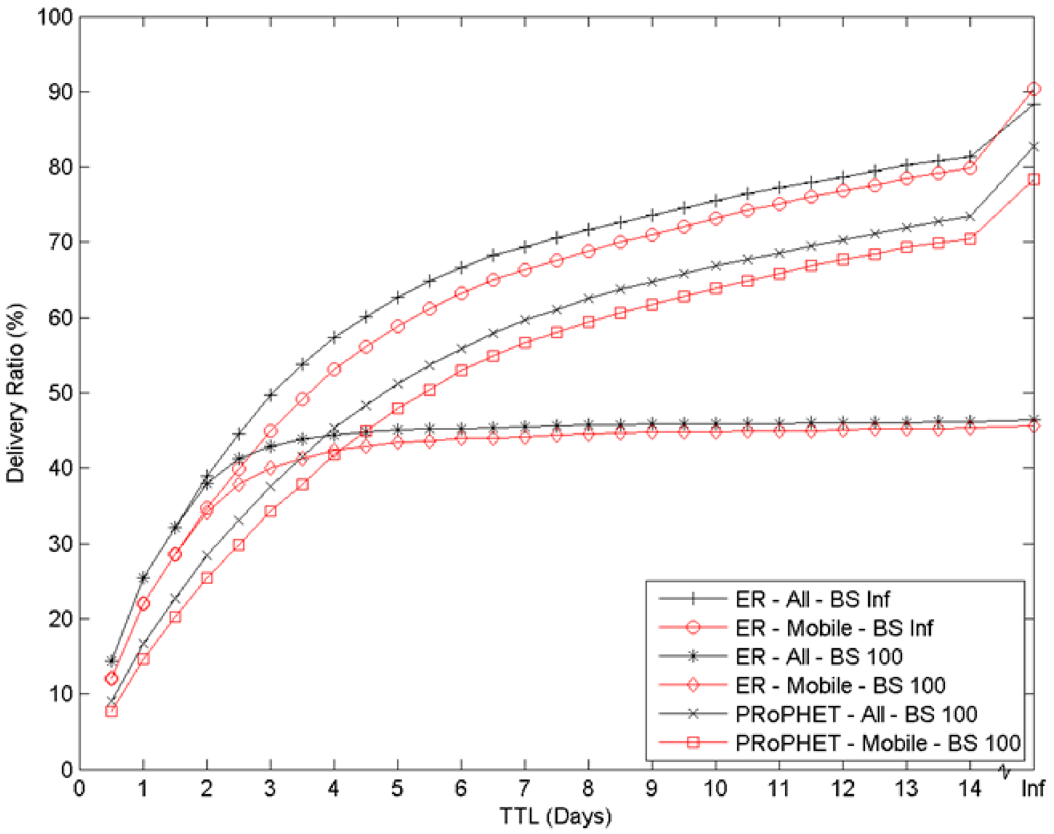


Figure 4.4: Effect of adding stationary nodes on delivery ration of benchmark algorithms.

Figure 4.5 shows the transmission overhead for each of the benchmark algorithms, with and without stationary nodes, with a data segment size of 0 and 129 bytes. Without any data segment, the overhead introduced by Epidemic routing for low TTL values is even better than PRoPHET, due to PRoPHET's control packets. Although having no data segment is an uncommon scenario, Figure 4.5 shows that in this case sending packets to all the nodes is reasonable as it eliminates the need for sending any control packets. By increasing the size of the data segment to 129 bytes, the overhead due to Epidemic routing increases as expected while the PRoPHET algorithm scales more reasonably.

Comparing the overhead between the two different cases of each algorithm (with and without stationary nodes) shows that adding stationary nodes increases the overhead by

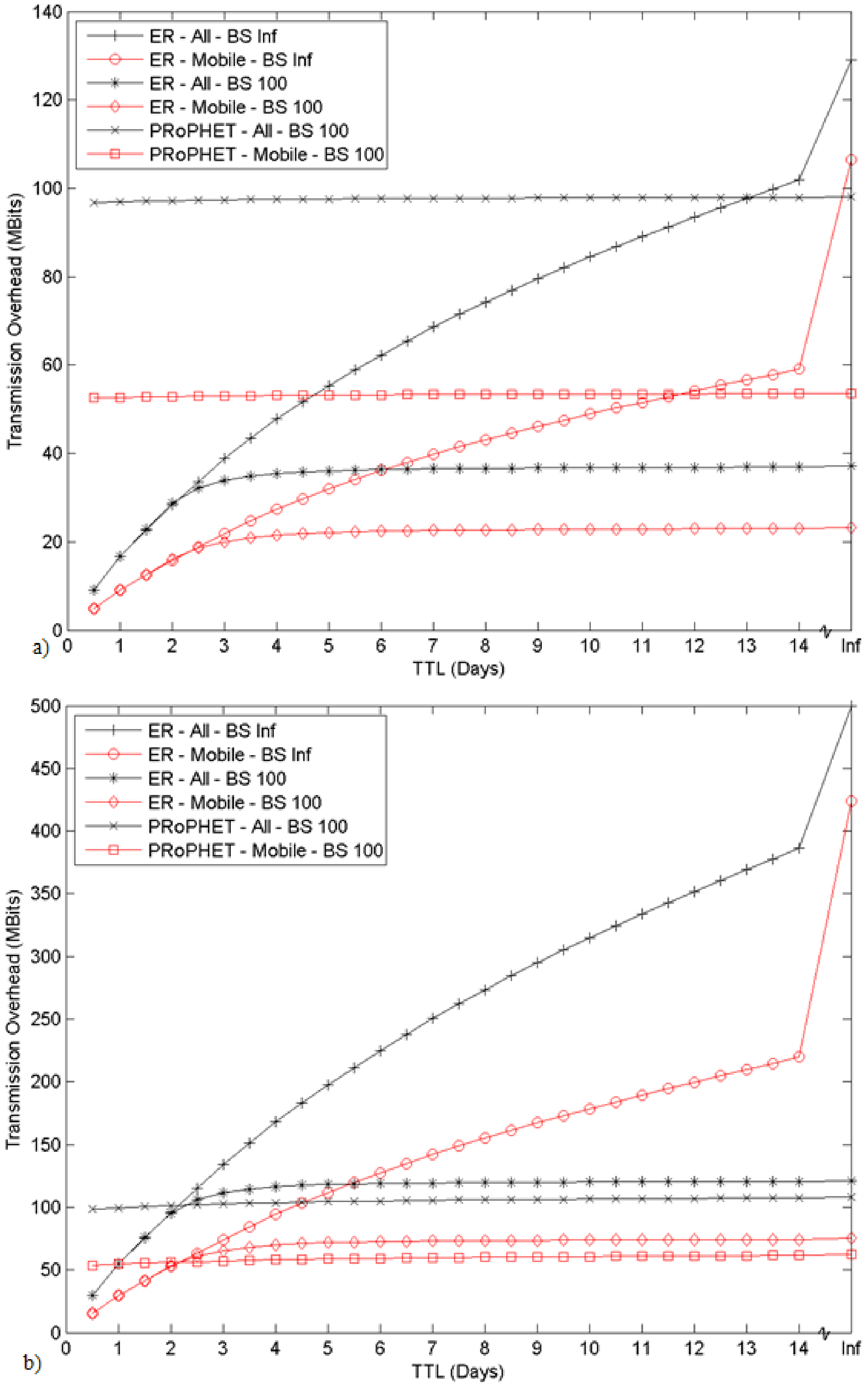


Figure 4.5: Transmission overhead for benchmark algorithms with a) no data segment per packet b) 129-byte data segment per packet (maximum TOS 1.1 data segment size)

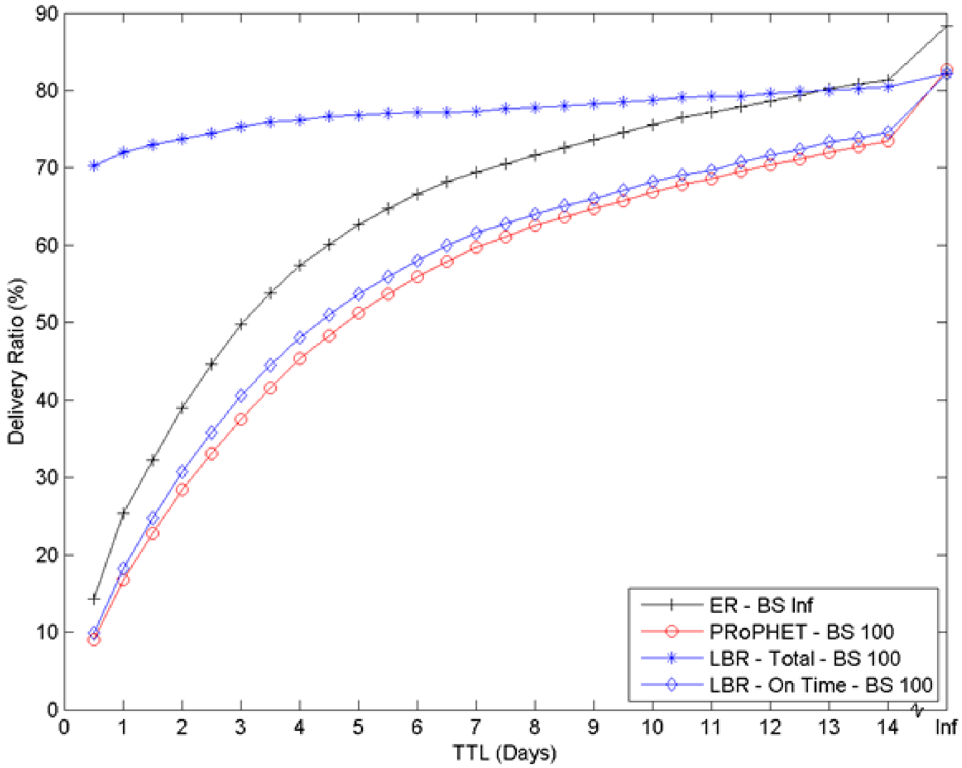


Figure 4.6: Delivery Ratio comparison between LBR, Epidemic Routing and the PRoPHET algorithms.

approximately 50%. This is primarily due to adding more nodes to the network, which causes more and larger control messages and increases the number of forwarded packets.

Looking at the performance improvement gained by adding stationary nodes (of approximately 10%) and the overhead introduced by these nodes (approximately 50%) in benchmark algorithms, reveals the general inefficiency of benchmark algorithms in dealing with stationary nodes and overcoming the differences between additional nodes in the network. This inequality in the ratio between delivery improvement and increase in overhead for each stationary node and the inability of standard algorithms to adjust to this structure leads to the poor performance of these algorithms.

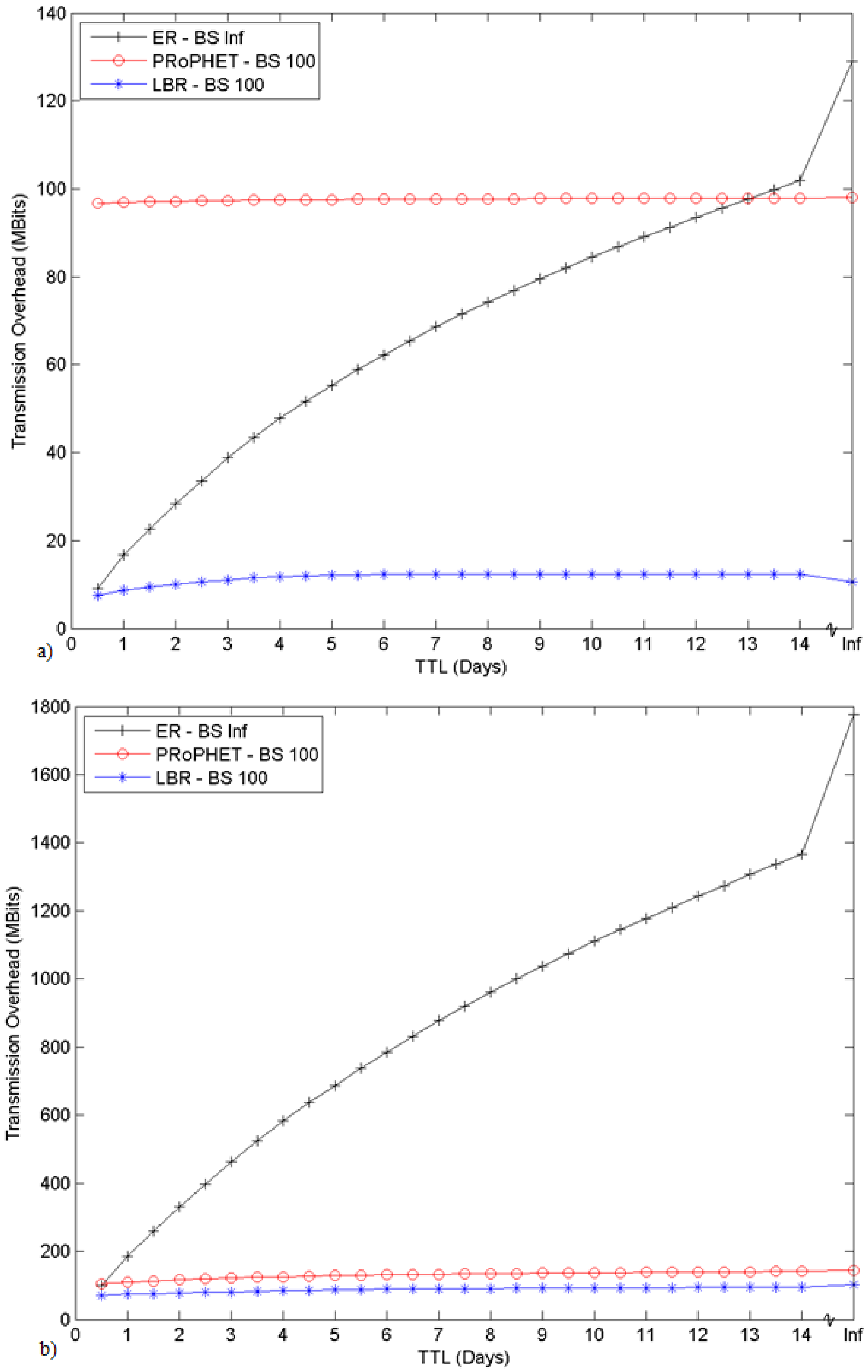


Figure 4.7: Transmission overhead for LBR compared to Epidemic and PRoPHET algorithms with a) no data segment per packet, and b) 129-byte data segment

4.5.2 LBR Routing Performance:

Figure 4.6 compares the delivery ratio of LBR with PRoPHET and Epidemic routing. The buffer size for Epidemic is set to infinite in order to show the best achievable performance in the network, while the buffer size for all nodes (stationary and mobile) in both LBR and PRoPHET is set to 100 packets. The ratio of the packets delivered before TTL time-out by LBR has improved by a maximum of 19% compared to PRoPHET, and on average it has a 7% smaller delivery ratio than Epidemic routing. Taking delivered timed-out packets into account as well, on average LBR improves the delivery ratio 39% more than Epidemic routing with Infinite Buffer Size.

Figure 4.7 shows the transmission overhead of LBR compared with Epidemic routing and PRoPHET, with 0 and 129-byte data segments. As in the results given above, the buffer size in Epidemic routing is set to infinite while for LBR and PRoPHET it is equal to 100 packets, and all of the algorithms used both stationary and mobile nodes. With no data segment, the control packets' size in PRoPHET considerably increases the load and makes it equivalent to Epidemic, while LBR has much lower control packet overhead. With a 129-byte data segment, the load in Epidemic routing dramatically increases.

As mentioned before, it is assumed that the stationary nodes are fixed at specific locations and are therefore have access to more resources compared to mobile nodes, such as more power and greater storage capacity. LBR can substantially benefit from additional resources in stationary nodes. Figure 4.8 shows the effect of having a 1000 packet buffer size in stationary nodes on delivery ratio. As expected, increasing the space available to stationary nodes increases the delivery ratio.

Figure 4.9 shows the effect of increasing buffer size for stationary nodes on transmission overhead. Interestingly, for all two data segments, the scenario with larger buffer space for stationary nodes imposes a lower transmission overhead compared to the case with smaller buffer space. This is primarily due to buffer overflow in stationary nodes offering limited buffer space, which causes retransmission of packets by mobile nodes and leads to a higher

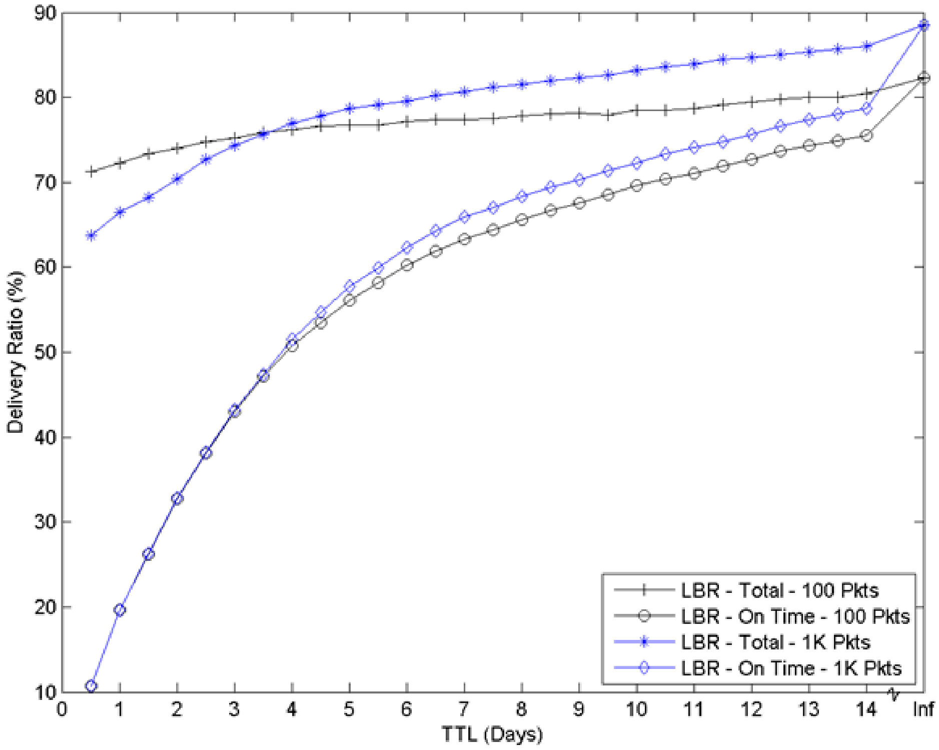


Figure 4.8: Effect of increasing stationary nodes' buffer size in LBR algorithm on delivery ratio

transmission load, while with larger buffer space the stationary nodes face less buffer overflow and fewer packet retransmissions are required.

4.5.3 Load Reduction

As described in section 4.3.2.4, the reduced load variant of LBR (LBR-RL) sought to reduce the load on certain high-centrality mobile nodes which are associated with one or more stationary nodes by determining these associations and moving the load from mobile nodes to the stationary nodes.

Figure 4.10 shows the effect of LBR-RL on selected nodes with contact patterns similar to one or more stationary nodes. For the PRoPHET algorithm, the buffer size was capable of holding 100 packets while for LBR (both normal and reduced) buffer size for mobile nodes was set to 100 packets, while it was set to 1000 packets for stationary nodes. TTL value for all three

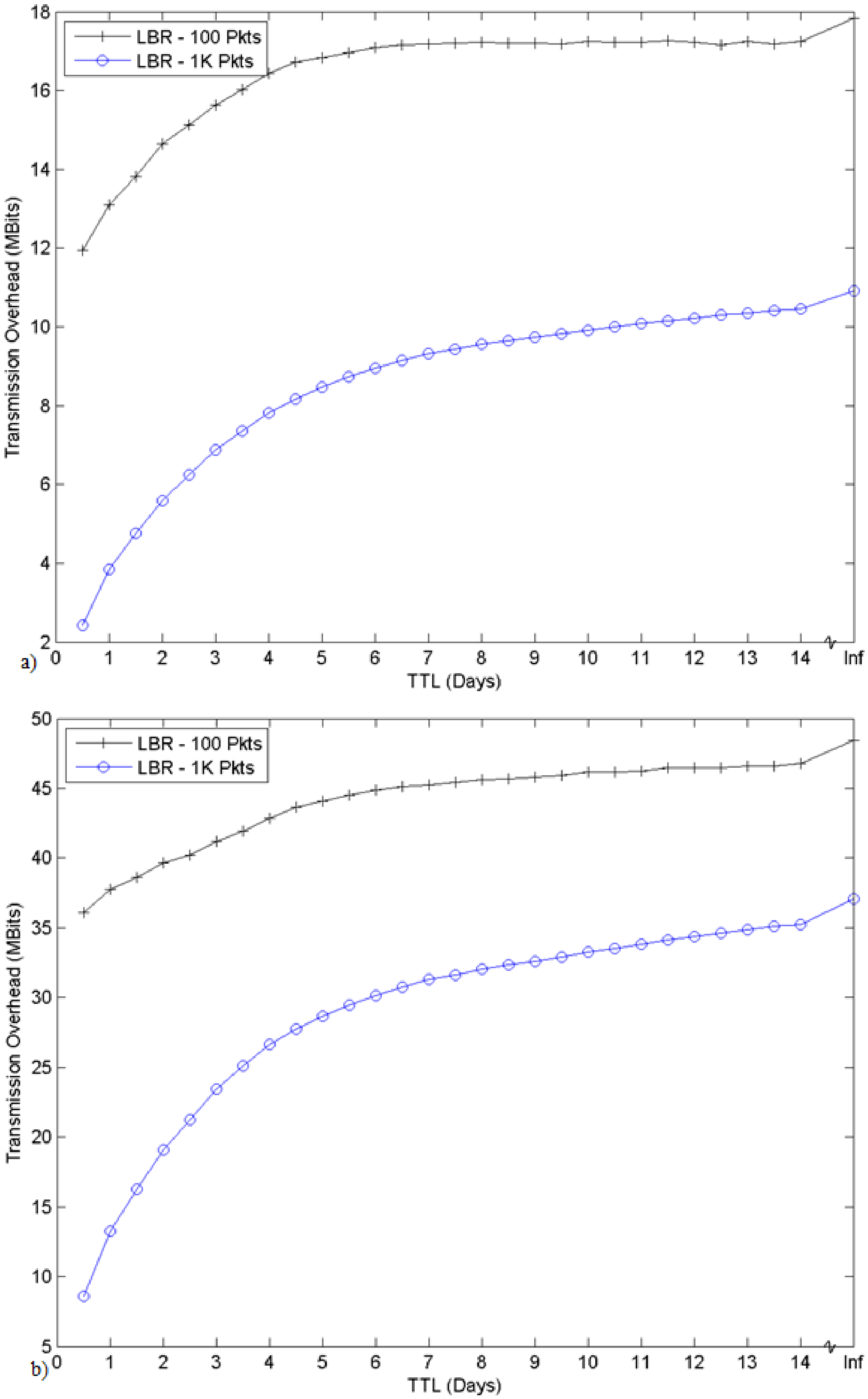


Figure 4.9: Effect of increasing stationary nodes' buffer size in LBR algorithm by order of 10 on transmission overhead with a) 0-byte data segment b) 129-byte data segment

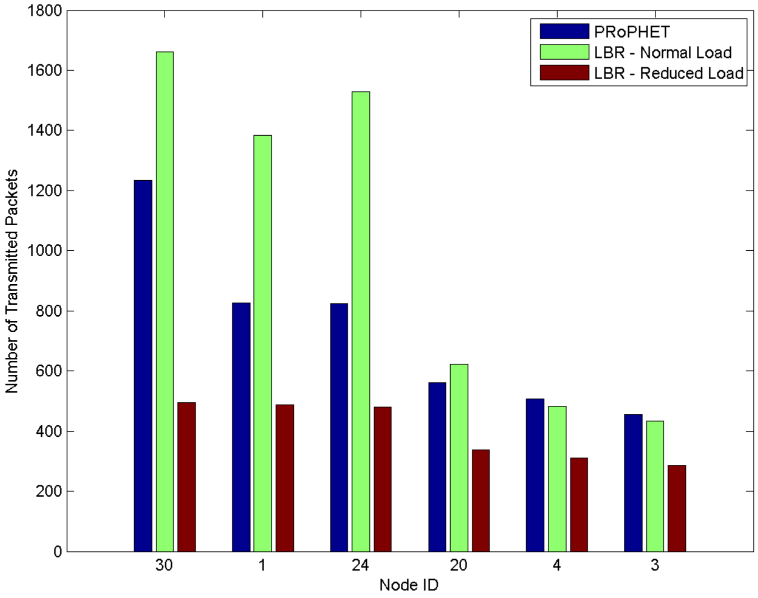


Figure 4.10: The effect of load reduction model on the selected nodes

algorithms was set to infinity. The X axis represents the node ID and the Y axis shows the number of packets transferred via the node using different algorithms.

Figure 4.11 shows the effect of the load reduction model on delivery ratio and plots both on-time and total delivered packets. The graph shows a maximum of 1.7% reduction of the delivery ratio for on-time delivered packets, which is due to a slightly higher latency in the LBR-RL in comparison with LBR, while in terms of total delivered packets (on-time and timed-out) there is no significant difference in delivery ratio, and both lines are almost identical.

4.5.4 Popular Stationary Nodes

A set of 12 simulations were run, where the number of stationary nodes remained constant, but where the number of popular stationary nodes changed from 0 to 11 and the number of ordinary stationary nodes changed from 11 to 0. Stationary nodes were sorted based on their total contact duration with mobile nodes, and at each simulation run the highest node moved from the

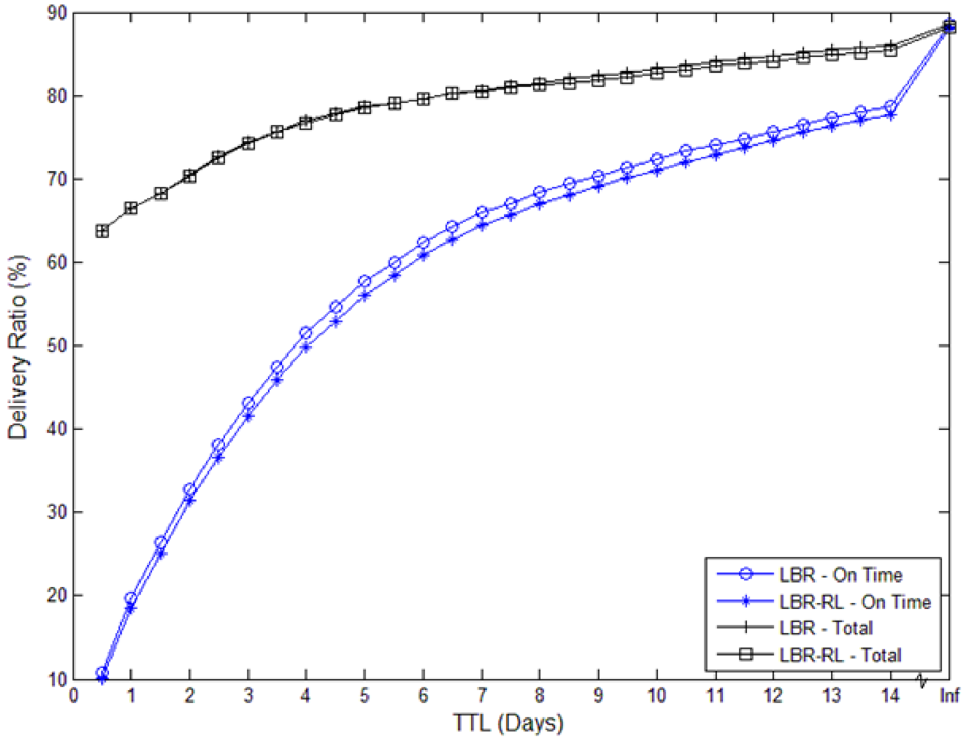


Figure 4.11: Delivery ratio comparison between LBR and LBR-RL

ordinary group to the popular group. The case with all the stationary nodes marked as "popular" is equivalent to the default behavior of LBR and represents the most aggressive policy regarding message forwarding to stationary nodes, while the case with all the stationary nodes marked as "ordinary" represents the most conservative policy, and the rest fall in between. In this simulation set, the packet TTL was infinite, the mobile buffer size was equal to 100 packets, and the stationary nodes buffer size was infinity.

Figure 4.12 and 4.13 show the effect of changing stationary nodes behavior from ordinary to popular on delivery ratio and transmission overhead, respectively, while the X axis on both graphs show the number of popular stationary nodes and the Y axis shows delivery ratio in Figure 4.12 and transmission overhead in Figure 4.13.

As shown in Figure 4.12, the delivery ratio increases by changing the behavior of the top five stationary nodes from ordinary to popular, while the delivery ratio quickly saturates after the

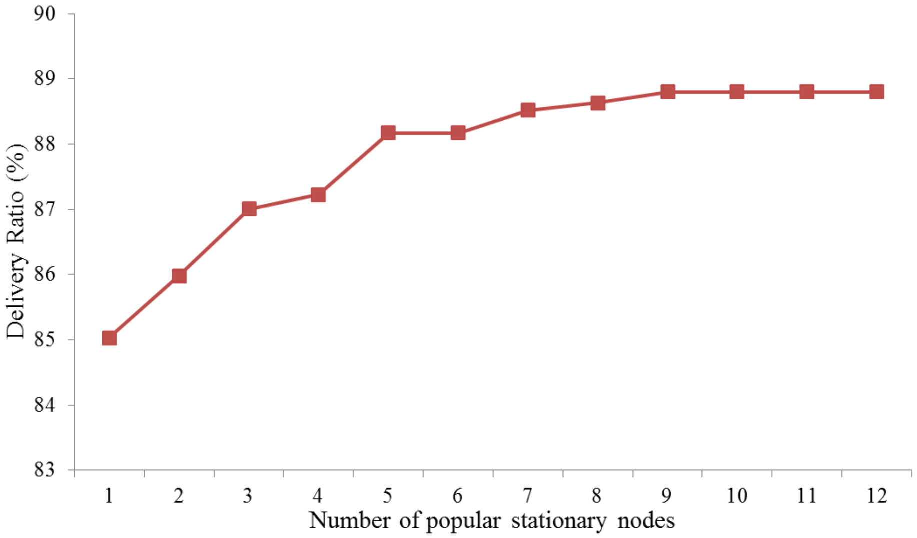


Figure 4.12: Effect of changing stationary nodes' behavior from ordinary to popular in delivery ratio

6th node. Unlike delivery ratio, the transmission overhead increases linearly. In addition, by looking at the total contact duration of each stationary node, it can be seen that all of the first five stationary nodes are connected to mobile nodes more than 10% of the time, while the rest have less than 10% total connectivity with mobile nodes (Figure 4.14).

I selected 10% connectivity between each stationary node and the mobile nodes as the threshold for selecting the node's behavior in terms of buffering packets. Therefore, if the total contact duration between the stationary node and all mobile nodes is more than 10% of the time elapsed since beginning of the experiment, the stationary node behaves as a popular node, otherwise it behaves as an ordinary stationary node.

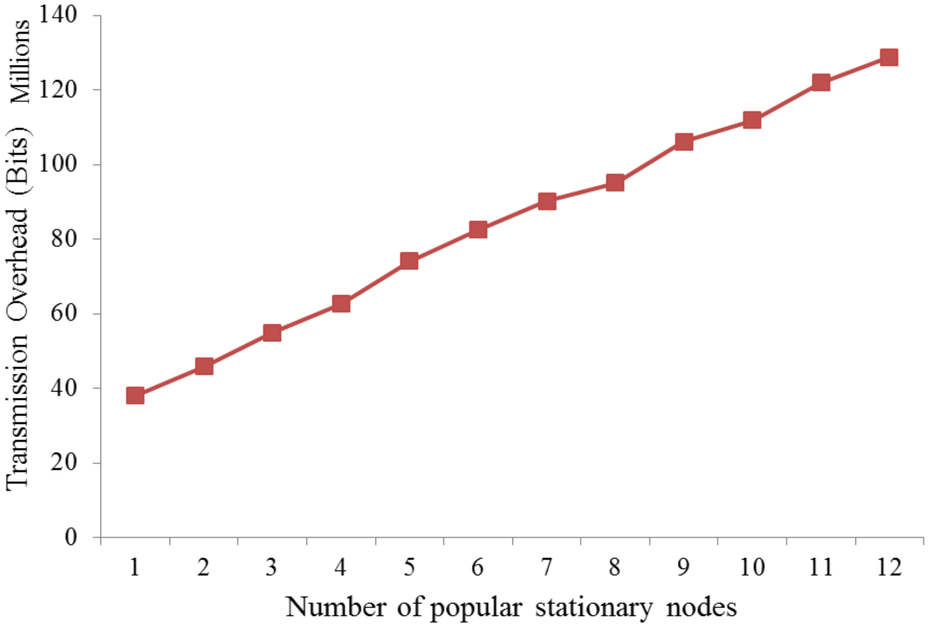


Figure 4.13: Effect of changing stationary nodes' behavior from ordinary to popular on transmission overhead

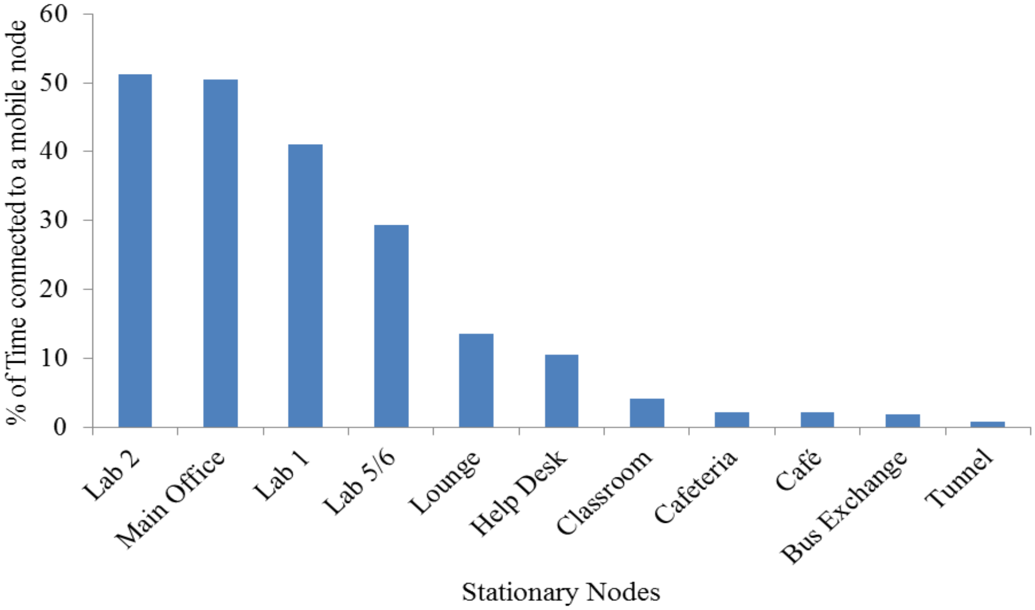


Figure 4.14: Stationary nodes' total connectivity with mobile nodes. Y axis shows the total contact duration between the stationary node and all mobile nodes over the total experiment duration.

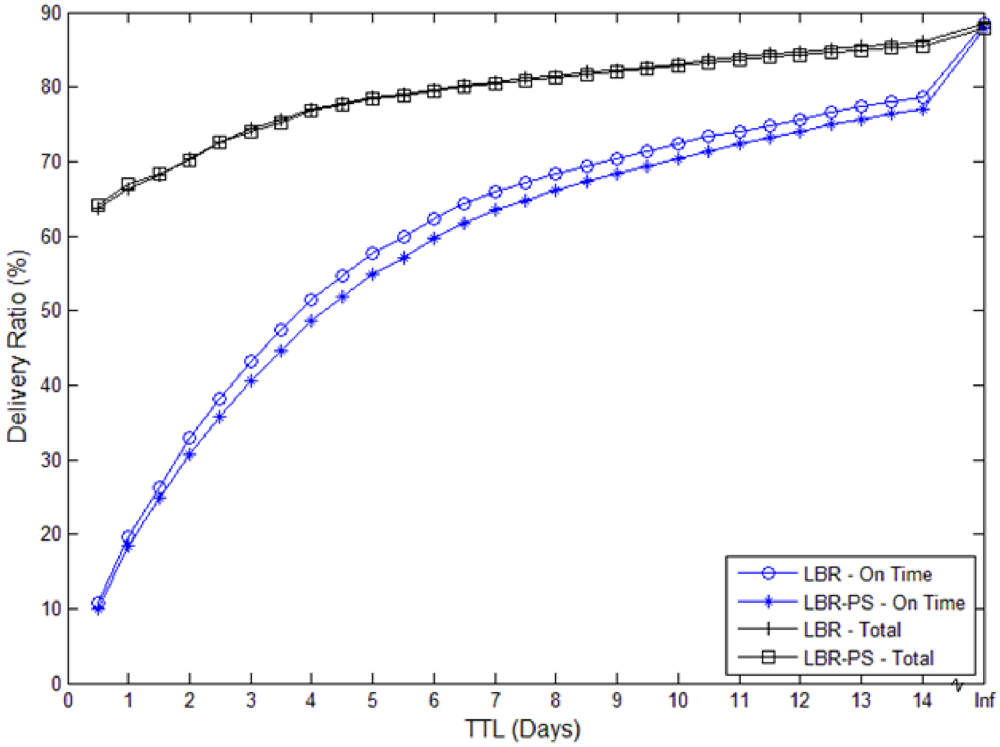


Figure 4.15: Effect of differentiating popular and ordinary stationary nodes on delivery ratio

Figure 4.15 represents the effect of applying LBR-PS on total and on-time delivery ratio. As this figure shows, by applying LBR-PS, the maximum decrease in delivery ratio is 2.9% for on-time delivered packets, and 0.5% for total delivered packets. Similar to the effect of LBR-RL on delivery ratio, introducing popular stationary nodes also slightly affects the latency due to longer paths for a small portion of packets, which leads to the difference in delivery ratio for on-time delivered packets and total delivered packets.

Figure 4.16 shows the effect of using LBR-PS on total transmission overhead, using different sizes of data segment. In the case with no data segment (Figure 4.16.a), the graph shows using popular stationary nodes model decreased overhead 29% on average compared to LBR, while with 129-byte data segment (Figure 4.16.b) the mean reduction is 36%.

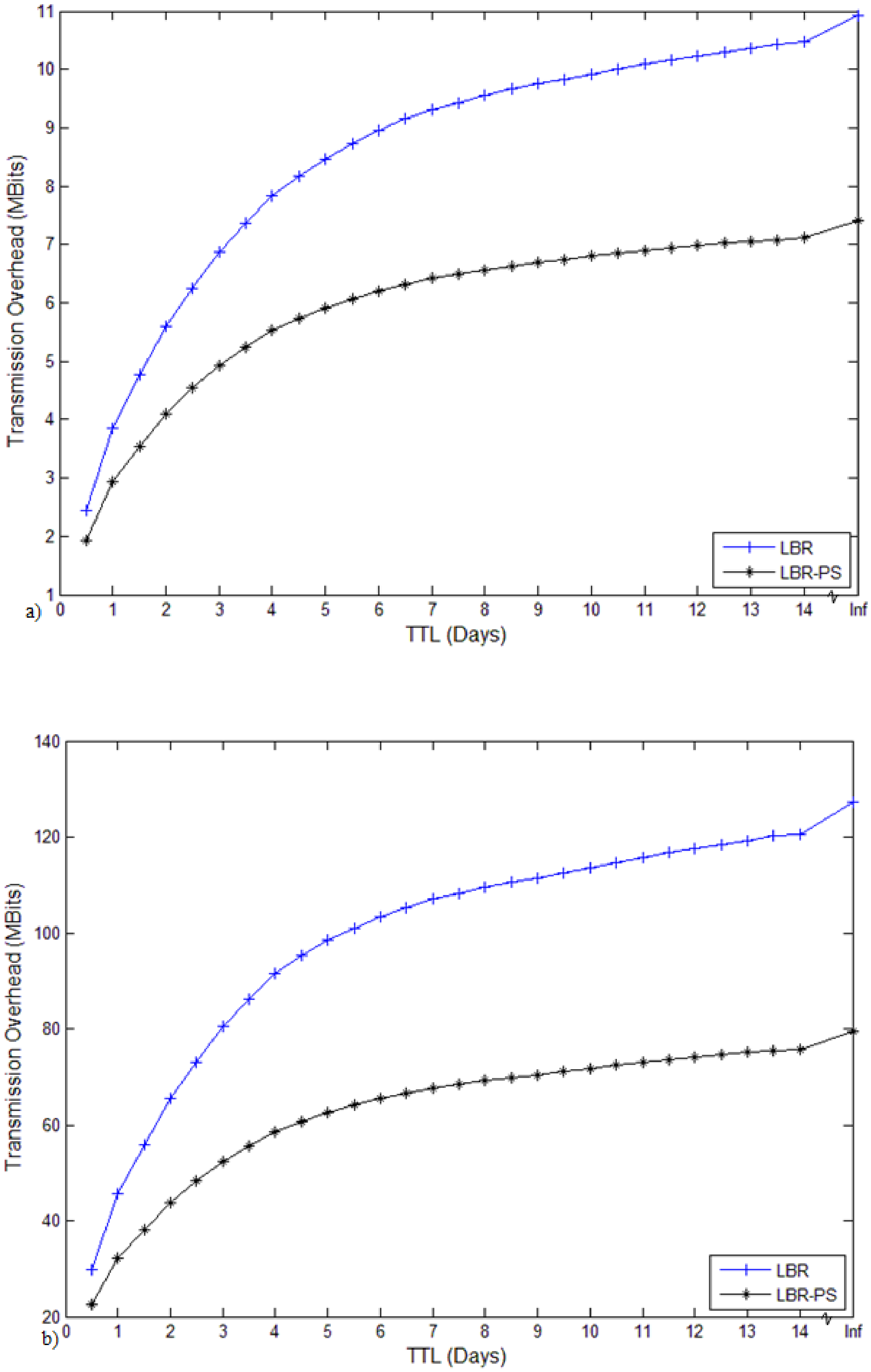


Figure 4.16: Effect of differentiating popular and ordinary stationary nodes on transmission overhead using a) 0-byte data segment, and b) 129-byte data segment.

4.5.5 Effect of Data Segment Size

The graphs presented in previous sections considered two different data segment sizes, i.e. 0-byte, and 129-byte (maximum TOS 1.1), and in certain cases these two values showed a considerable change on total transmission overhead. Figure 4.17 plots the effect on total transmission overhead of increasing data segment size exponentially. The TTL value for all the presented algorithms is set to infinity. Epidemic routing with infinite buffer size expectedly shows the maximum transmission overhead, regardless of data segment size. While for data segments smaller than 32 bytes, Epidemic routing has less overhead compared to PRoPHET. This is mainly because of the overhead introduced by control packets via PRoPHET, which is ameliorated with increasing data segment size. Both LBR and LBR-RL-PS have less overhead compared to Epidemic routing for any data segment size. LBR imposing greater load than

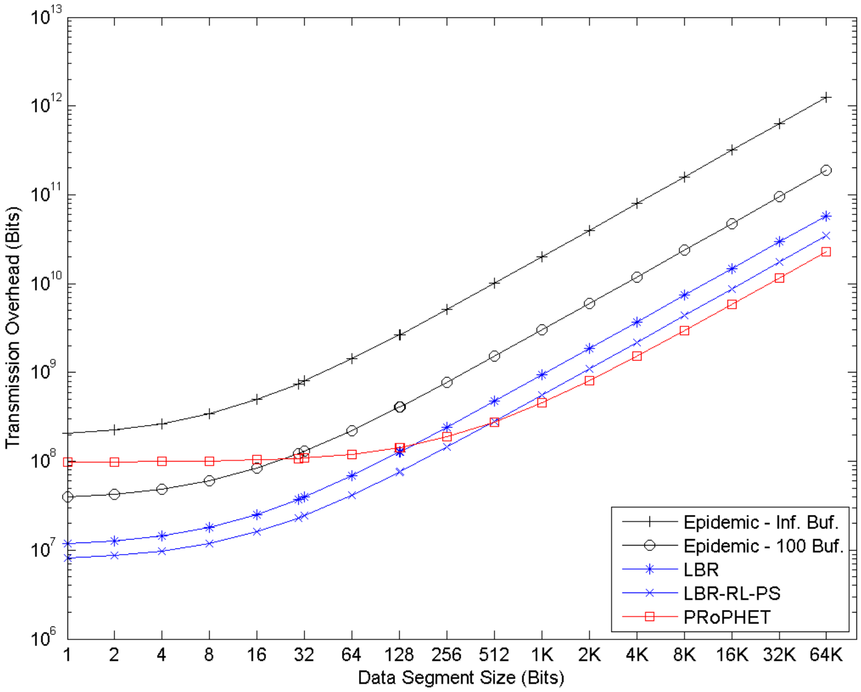


Figure 4.17: Effect of increasing data segment size on transmission overhead

PRoPHET for data segments larger than 128 bytes, while LBR-RL-PS puts considerably less load on the network for data segments smaller than 512 bytes.

4.6 Discussion

The combination of the collected dataset and the algorithms which are presented in this chapter provide an interesting picture of the interaction between people, places, and algorithm design. I have been able to demonstrate that static nodes can be leveraged to significantly improve routing performance, even with relatively simple policies for determining when and where to copy DTN packets. However, the analysis presented here is not without limitations, particularly with respect to the scope and generalizability of the dataset.

4.6.1 Impact of Place

Algorithms for PSN routing have focused on taking advantage of the opportunistic connectivity between human agents. However, stationary resources exist in a human environment, such as Bluetooth enabled personal computers, which are either not connected to a backbone, or are unable to share their backbone connection due to cost or security issues. Our algorithms and analysis can be leveraged to utilize these potential resources in an efficient way. Additionally, for remote or military applications where nodes are deployed away from infrastructure, LBR can be utilized to best understand where to place stationary nodes to enhance the delivery ratio and reduce node load.

By comparing LBR algorithmic variants against PRoPHET, it is demonstrated that knowing the stationary nodes are in fact stationary provides benefits beyond simply increasing the number of nodes. LBR benefits more strongly from the inclusion of stationary nodes than does PRoPHET, which treats all nodes identically.

This chapter demonstrated that using even mathematically simple policies in conjunction with stationary nodes can make a significant difference in the efficiency of PSNs. This efficiency is primarily gained by a conservative packet copy policy, which combines concepts of duration and location to make copies of packets at the right time and place.

4.6.2 Copying

A crucial component of PSN policy design is determining the circumstances under which to copy a packet, in an attempt to reach the destination faster using multiple routes. This question has been addressed to varying degrees by many authors, but most pointedly in [28, 12]. Epidemic Routing and direct pass (where nodes which create a packet only directly deliver to the destination) provide an upper and lower bound on this tradeoff with always copy and never copy policies. Algorithms must attempt to match the delivery ratio of Epidemic Routing, while minimizing the number of transmissions of both data bytes and overhead.

The copying policy in this chapter is based on the assumption that stationary nodes will be better candidates than their centrality would directly indicate, as there is a high probability of finding specific nodes in specific locations. Popular nodes can then be selected to always receive packet copies, in the belief that they will either see the node, or see a mobile node with a substantially better chance of meeting the destination than the current node.

In the small-scale network described by the dataset, an individually labeled system performed sufficiently well. However, for larger systems the number of mobile nodes may make contact tables unwieldy. In this case, pruning and labeling along the lines of [12] could be employed, retaining individual records for close contacts, and only clique labels for more infrequent contacts. Knowledge of a node’s localization pattern is a potentially attractive method of establishing cliques, as nodes which are localized to the same stationary node are likely part of the same subnet.

4.6.3 Dataset Limitations

Despite the significant contributions offered by the work in this chapter, there are still limitations to the scope and generalizability of our results, primarily due to the dataset which is employed. Like most microcontact datasets [6, 7, 27, 46], Flunet has a strong selection bias towards academic life. The contact patterns and localization behavior discussed here are biased towards graduate student mobility patterns, with flexible arrival and work hours, fixed cubicles or offices and intermittent proximal contact. The strength of the Main Office subnet and its isolation from the other networks is partly due to the different workplace habits and expectations of clerical workers and graduate students. This bias would be even more obvious in extreme cases like a retirement home, where the mobility of many residences might be contingent on the proximity of a care provider, or at an automotive plant where strict shift and break timing and workplace location would create temporally alternating linear and clustered subnets. This limitation is primarily due to the youth of microcontact monitoring as a field. Once tools mature enough for use by sociologists, epidemiologists, and automation engineers, data breadth will increase. Results in this chapter must be taken as limited to academic environments until further data can be obtained which either confirms or refutes their generality.

The size of our study also introduces bias into the analysis. While our study covers 13 weeks, a reasonable duration, it contains only 36 mobile and 11 stationary nodes. Because these nodes were primarily graduate students in a single department at the university, reasonable contact data was obtained. However, I strongly suspect that the dataset is missing the scale necessary to capture the churn typical of locations like the bus exchange, cafeteria and tunnel locations. The examined dataset eliminates many potential multihop paths because of the small number of agents. However, these multihop paths are often undesirable in PSNs because they increase the number of transmissions and energy drain on the system.

Finally, only data is available on those locations which were specifically telemetered. Again this is not uncommon in datasets where either person-location [27] or person-person [6] interactions are missing, or highly inaccurate, or limited to a specific set of locations [7]. The lack of specific location information will become less of an issue in future datasets, as new localization techniques including Bluetooth connection to known static devices, better WiFi localization and GPS [27] become more widely adopted and integrated into microcontact and location measurement systems.

CHAPTER 5: APPLICATIONS IN EPIDEMIOLOGICAL MODELING

Simulation models of infection transmission are proven as effective tools to understand and evaluate interventions to control the spread of communicable diseases. Unfortunately, even the best models suffer from inaccuracy due to limited availability and detail of contact data. Accounting for these deficiencies requires novel means of contact data collection. This chapter integrates an agent-based model of H1N1 pathogen transmission with cross-linked minute-resolution contact patterns and disease status self-reports resulted from the Flunet dataset. The results demonstrate that the richer social network information provided by such contact microdata can be readily incorporated into transmission models, and can lend significant insights into transmission patterns. Most significantly, agent contact duration correlated significantly better than traditional centrality measures with risk of infection and with counts of secondary cases caused by an infected agent. These insights suggest that cross-linking of simulation models with automated ambulatory data collection such as Flunet can offer rich insights into the drivers of infectious disease transmission, observed patterns of disease, and inform environmental interventions.

5.1 Background

The threat of emerging infectious diseases has stimulated the search for techniques to prevent and control communicable disease spread [68]. Simulation models have emerged as key tools in examining tradeoffs between multiple health interventions, and in aiding the control of communicable diseases [69]. While properly parameterized and calibrated models can inform decision making, building such models is challenging because critical parameters are difficult to measure precisely, including the structure and dynamics of contact networks among population members, which shape the spread of both pathogens and risk behaviors.

Data collected by contact tracing [70] and self-reporting [56] has provided some important insights into network structure for many notifiable illnesses. Unfortunately, even for the best models, contact data depends heavily on unreliable self-report data collection methodologies [60] which omit detail and place a substantial recording burden on participants [56]. Because of the less tangible character of the contacts involved, determining contact network structure for air-borne pathogen spread requires the collection of additional information on casual contacts [56]. While self-reported measures can provide insight, some leading studies have noted the desirability of employing automated data-collection approaches to capture higher fidelity contact frequency and duration information [56].

As the first influenza pandemic in decades, the H1N1 pandemic served as a catalyst for research into control of emerging infectious diseases. In Saskatoon H1N1 first emerged in Spring 2009, and followed the typical summer quiescence, and autumnal re-emergence. By mid-October, cases of H1N1 began a notable rise [71]. At the same time, vaccination was initiated in a staged fashion. Mass vaccination proceeded aggressively from early November through December 18. Vaccination data suggest that approximately 50% of the city population was immunized [72]. Aided by the staged vaccination process, reported cases of influenza in 2009-2010 peaked unusually early (mid-November). Low numbers of influenza cases were reported in December 2009 and thereafter. Most circulating influenza transmission in Saskatchewan over this period was drawn from the H1N1 strain [71].

This chapter seeks to integrate rich contact microdata from Flunet with an adaptation of a well-grounded individual-level Canadian transmission model [73]. Our study objectives were threefold: to assess the effectiveness of incorporating contact microdata with models of infectious disease, to identify features within empirical contact patterns that exerted disproportionate impact on infection spread, and to validate these findings against self-reported health status information.

5.2 Methods

The key methodological innovation presented by the work in this chapter was the use of dynamic contact data such as Flunet in the agent-based simulation of infection spread. While other contact datasets are available [7], and others with health information have been described

[46], Flunet is unique in providing longitudinal information on contact patterns, occurrence of influenza-like illness (ILI) symptomology, and vaccination. As described in Chapter 3, Flunet contained detailed inter-participant contact patterns but lacked data about the threat of infection from non-participants, which is modeled using reported H1N1 incidence data for the same time and province as the contact data. A system diagram outlining the flow of data and state changes is shown in Figure 5.1.

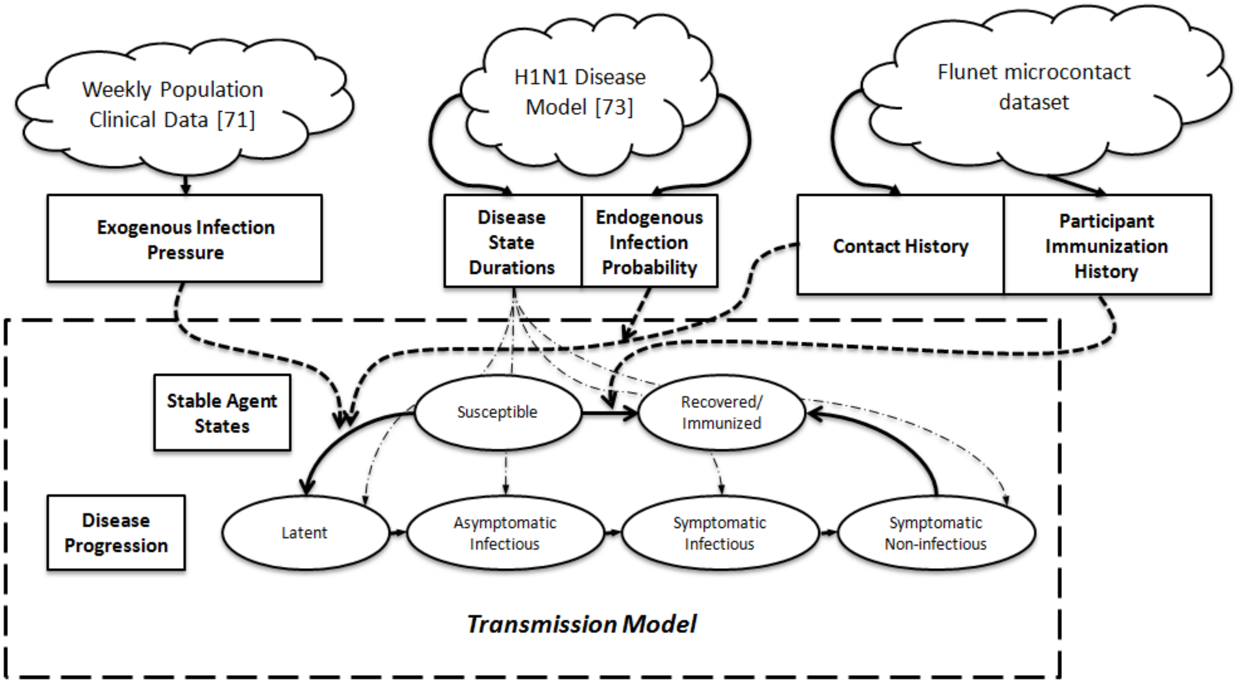


Figure 5.1: Simulation structure and flow

The simulation model is encapsulated in the dashed box. Agents remained in a susceptible state unless acted on by an infection event. Such stochastic events were triggered externally using the exogenous pressure data derived from case reports [71], or internally through contact with an infected agent. The likelihood of endogenous infection was governed by contact and vaccination data from Flunet surveys and infection risk information drawn from [73]. Because it is assumed that H1N1 re-infection risk to be negligible, agents in the recovered or immunized state remained there until the simulation ended. If an agent became infected, the infection ran its course deterministically through the stages of infection. The duration of each infection stage was drawn from the distributions in [73] and derived quantities.

I performed Monte Carlo ensembles of stochastic dynamic simulations operating on the contact data, where the primary variables drawn from distributions were disease stage durations, exogenous infection events and transmission from infected endogenous contacts. The contact record was stepped through like an animation, creating exactly the same sequence of contacts in the course of every Monte Carlo realization. While agents repeatedly relived the same sequence of contacts in different realizations, the stochastic associated with infection progression and transmission gave rise to differences in disease spread.

5.2.2 Transmission Model

5.2.2.1 Model Design

The simulation model classified each individual in the sample population into one of seven states: *Susceptible*, *Latent*, *Asymptomatically Infectious*, *Symptomatic Infectious, Symptomatic Non-Infectious*, *Recovered*, and *Immunized*. All the agents in the model started in the *Susceptible* state. A susceptible individual could contract the infection either from exogenous or endogenous sources. Exogenous sources are defined as the population outside the study who were in contact with Flunet participants and could transmit the infection to the monitored individuals, while endogenous sources are other Flunet participants in an infectious state.

Dynamic transmission models differ in their treatment of contacts. For some epidemiological contexts, the contacts underlying transmission are of defined or bounded duration – for example, needle sharing, sexual encounters, and blood transfusions. For this class of contacts, the frequency rather than the duration of contacts is the primary source of variability in transmission risk. For air-borne infections, however, the likelihood of transmission rises not only with contact frequency, but also with contact duration [57].

For the case of H1N1 influenza transmission, our model assumes that ongoing contact between two discordant individuals provides a conduit for transmission, where the likelihood density of transmission is a constant independent of contact duration. More specifically, an infectious individual gives rise to potentially contagious events (e.g., sneeze or cough) at a fixed rate v. Any susceptible individual in contact with that person as having a likelihood β of contracting the infection for each such infectious event. Similar to the analysis [57], infections are more likely to occur in a longer contact than in a shorter one, as by considering a fixed rate for infectious events, longer contacts include more such events and subsequently cause higher likelihood of infection transmission.

Given this model, the basic reproductive number is as follows:

$$R_0 = \overline{T_\iota}\overline{F_C}v\beta \tag{5.1}$$

where $\overline{T_\iota}$ is the mean duration of infectiousness in days, $\overline{F_c}$ represents the average daily cumulative contact duration of an individual in time slots (summed regardless of concurrency), v is the number of potentially infecting events per time slot of contact between two individuals, and β is the mean likelihood that a susceptible agent will be infected by a given infecting event.

For endogenous infections, it is assumed that the mean of the basic reproductive number R_0 for our study population was identical to that reported in a prominent Canadian H1N1 study [73]. For an infective person, the infection hazard of infecting an adjacent susceptible individual per time step (βv) was thus determined as follow:

$$\beta v = \frac{R_0}{\overline{T_\iota} * \overline{F_c}} \tag{5.2}$$

$\overline{F_c}$ is computed using recorded data in the Flunet dataset according to the following formula:

$$\overline{F_c} = \frac{\sum_{i=1}^{N_p} \sum_{j=1}^{N_p \,\&\, j \neq i} T_c(i,j)}{N_p\, N_d} \tag{5.3}$$

where N_p represents the number of study participants, N_d gives the study duration, and $T_c(i,j)$ indicates the total duration of contacts (in days) between nodes i and j during N_d days. Stationary nodes were not considered in the study because we only considered person to person transmission.

The per-week infection hazard of acquiring the infection from exogenous sources was determined according to the following formula:

$$P_x(k) = \frac{N_k}{N_U - \sum_{i=1}^{k-1} N_i} \tag{5.4}$$

where *k* refers to the week number since the start of the simulation, N_i gives the number of laboratory confirmed H1N1 cases in the whole province during the i^{th} week of the 2009-2010 influenza season, and N_U refers to the total population of Saskatchewan. The denominator represents an estimate of the number of susceptible individuals in the province. The entire denominator thus estimates the number of susceptible individuals who remain at risk in week *k*. It should be noted here that the computation of susceptible individuals in the denominator considers neither the vaccination status of the population nor the infections that took place the previous influenza season. This formulation therefore systematically underestimates the actual infection pressure. This systematic error is compensated by providing a thorough exploration of the space of possible transmission probabilities (Figure 5.3).

A susceptible agent receiving the infection from either exogenous or endogenous sources transitions to the latent state. Before starting the Latent period, the model computed the duration for each of the subsequent four stages of illness (Figure 5.1). In determining these durations, I sought to reproduce the observed variability in H1N1 progression by drawing the duration of incubation (T_{Inc}) and duration of symptoms (T_S) from two log-normal distributions with parameters from [73]. The illness duration T_{Ill} was calculated by adding $T_{Inc}$and T_S. Using these three values, the total duration of infectiousness T_{Inf} was calculated as:

$$T_{Inf} = \frac{T_{Ill} * \overline{T_{Inf}}}{\overline{T_{Inc}} + \overline{T_S}} \tag{5.5}$$

where T_{Ill} gives the computed duration of illness, $\overline{T_{Inf}}$ is the average duration of infectiousness, $\overline{T_{Inc}}$ shows the average incubation period, and $\overline{T_S}$ is average duration of symptoms. Following the computation of the duration of infectiousness, the duration of the *Symptomatic Infectious* state $T_{sinf}$was estimated using:

$$\frac{T_{sinf}}{T_{Inf}} \approx \frac{\overline{T_{s\imath nf}}}{\overline{T_{Inf}}} \tag{5.6}$$

The remainder of the durations were computed using following subtractions:

$$T_{ainf} = T_{Inf} - T_{sinf} \tag{5.7}$$

$$T_{lat} = T_{Inc} - T_{ainf} \tag{5.8}$$

$$T_{nInf} = T_S - T_{sinf} \tag{5.9}$$

where T_{ainf} represents the asymptomatically infectious duration, T_{lat} shows the latent period duration, and T_{nInf} shows the symptomatic non-infectious duration.

Each infected agent experienced the four illness states sequentially with the passage of time. A person in the *Asymptomatically Infectious* or *Symptomatic Infectious* state was considered infective, and could infect other susceptible adjacent individuals with infection hazard βv. At each time step and for each adjacent susceptible, the infective person transmitted the infection with likelihood density βv. A susceptible receiving H1N1 vaccination transitioned to the *Immunized* state, and was thereafter considered immune.

For the sake of the simulation, it is assumed that no H1N1 mortality. The study here lacked sufficient data to predict whether a specific individual would elect to self-quarantine given a symptomatic infection, and did not consider hospitalization outcomes. Given these assumptions, it is chosen that to regard an individual's contact patterns as unaffected by the health status of that individual and those around them. To examine the degree to which these assumptions might shape simulation results, an additional Monte Carlo ensemble is simulated examining the extreme situation in which individuals removed themselves from circulation for the duration of their symptomatic period. Finally, in light of the dominance of the H1N1 strain during the Saskatchewan 2009-2010 influenza season, only one strain of influenza was considered.

5.2.2.2 Simulation Setup

The model described in the previous sections was implemented in network simulator 3, using a network of 36 agents, where each agent represented one individual. The simulation setup was similar to the description in section 3.2.

To estimate P_x, the model required a time series of the laboratory-confirmed H1N1 cases in Saskatchewan. This data was extracted from the Public Health Agency of Canada FluWatch [71] on a weekly basis. As it is shown in Figure 5.2, the number of reported cases for the 13 weeks of the simulation declines monotonically to zero after the 9th week (January 4th 2010).

5.2.2.3 Scenarios

The model was simulated to explore a four-dimensional scenario space that examined the impact on model outputs of four distinct assumptions. The first two assumptions related to the exogenous and endogenous forces of infection (FOI). An exogenous FOI coefficient linearly scaled P_x to values that were 1, 2, 4, 8, 16, and 32 times the baseline. Similarly, the endogenous FOI coefficient scaled βv by 0.5, 1, 1.5, 2, 2.5, and 3 times the baseline value.

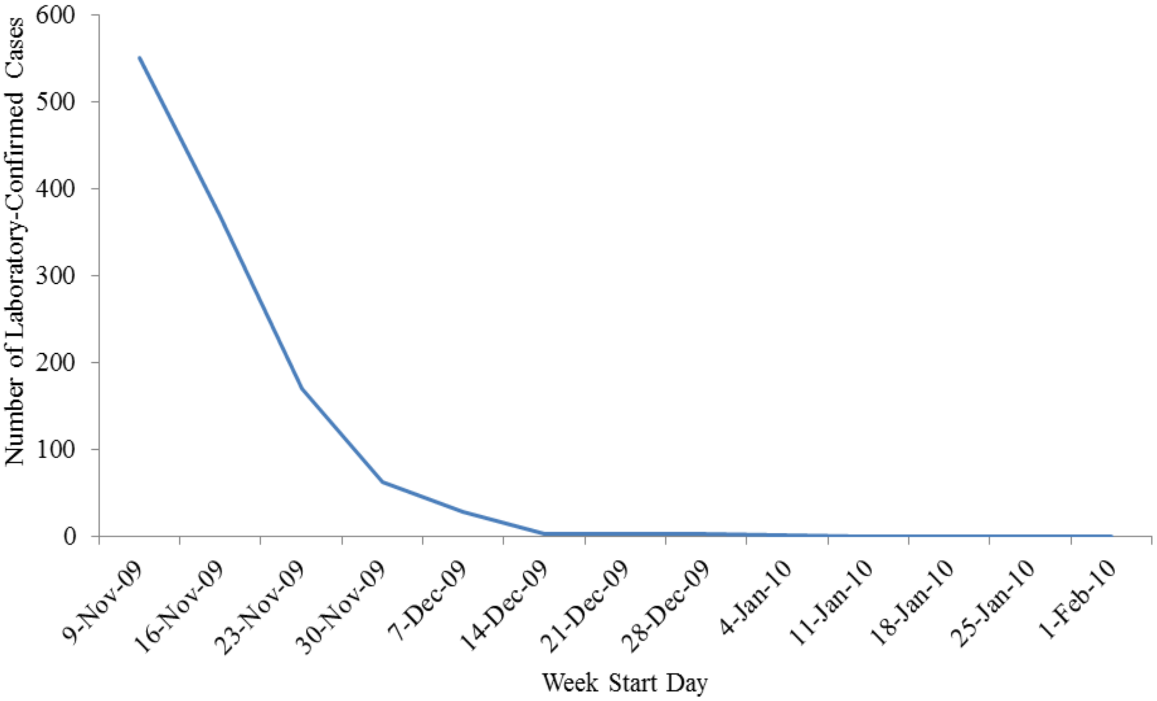


Figure 5.2: Weekly laboratory-confirmed H1N1 cases reported in Saskatchewan.

The third assumption varied was whether the H1N1 vaccination status from the Flunet survey results was considered during simulation. For the case without H1N1 immunization, none of the self-reported H1N1 vaccination data was considered. For the scenarios that account for H1N1 immunization, participants started susceptible but transitioned to the Immunized state according to the time they reported an H1N1 vaccination in Flunet health surveys. Individuals who did not report vaccination in the surveys never entered the Immunized state. It is assumed that negligible benefit of immunization if the agent was infected at the time of vaccination, and allowed the infection to run its course.

The final assumption varied related to the distance required to transmit an infection endogenously. Specifically, the impact of assuming that transmission required close contact (judged via RSSI), or any detectable contact is examined. Based on the contact criteria, T_c represented either total duration of close contacts or total duration of all contacts.

Two supplementary baseline (for close and all contacts) scenarios explored the impact of participants removing themselves from circulation during their symptomatic period. Note that to compute βv for these two scenarios, $\overline{T_\imath}$ was replaced with $\overline{T_{a\imath nf}}$, where $\overline{T_{a\imath nf}}$ was calculated as the average duration of asymptomatic infectiousness of all the previous baseline and alternative scenarios.

In total, the scenario space consisted of six baseline scenarios and 144 additional scenarios. Each baseline scenario was simulated using 100,000 Monte Carlo realizations; the other 144 alternative scenarios were each simulated using 2,500 Monte Carlo realizations. Exploration of the scenario space (including the baselines and alternative scenarios) required running 960,000 different realizations.

5.2.2.4 Metrics for Contact Networks Structure

While static representations of social networks are convenient, popular, and can yield powerful insights [74, 75, 76, 77], the temporal aggregation involved may obscure features of real contact networks that serve important roles in the transmission of infectious disease. Our experimental and simulation design provided us with rich information on study network dynamics. However, because network structure – particularly evolving network structure – is difficult to represent concisely, derivative measures are often employed [76]. To quantify the structure of our network, four centrality measures are employed: Betweenness, degree, time degree, and log time degree. Betweenness centrality is a classic measure of network structure that attempts to capture the importance of the node to the graph's connectivity, by summing the number of times a node lies on the shortest path between two other nodes, calculated using:

$$C_B(v) = \sum_{a\, \in Nodes} \sum_{b\, \in Nodes -\, \{a\}} \frac{\sigma_{ab}(v)}{\sigma_{ab}} \tag{5.10}$$

where v is the vertex in question, σ_{ab} is the number of shortest paths between a and b, and $\sigma_{ab}(v)$ is the number of shortest paths between vertices a and b that pass through v, summed over all pairs of vertices in the graph.

While betweenness captures a global picture of the network by examining shortest paths, degree centrality only considers a node's number of one-hop neighbors. For a static graph, degree centrality is calculated according to:

$$C_D(v) = \frac{deg(v)}{n-1} \tag{5.11}$$

where $deg(v)$ is the number of edges incident on v, and n is the number of nodes in the graph. Degree centrality can capture the local conditions of a node more accurately, but does not take into account the heterogeneous nature of the contact patterns and durations evident in our dataset. As a result, additional centrality measures are defined to capture elements of network dynamics.

Time degree centrality for a node can be defined as the average over all time slots of the fraction of all other agents with whom that node is in contact in a given time slot. This is computed in our case by summing up the count of that node's contacts over all 30-second time slots in the entire study, and then dividing by both the number of time slots in the whole study and by the number of participants minus one. This measure captures the duration of (potentially concurrent) interpersonal contact patterns. People who encounter many others for short periods would have a large degree centrality, and a similar time degree centrality to individuals who spend a great deal of time in a smaller group and have a smaller degree centrality. Because time degree aggregates over contact times, and because contact times are often characterized by power law distributions [6, Section 3.3], log time degree centrality also is introduced as a measure of contact density. Time degree and log time degree centralities are calculated discretely as:

$$C_{TD}(v) = \frac{1}{N_k}\sum_{k} C_D(v,k) \tag{5.12}$$

$$C_{LDT}(v) = ln\big(C_{TD}(v)\big) \tag{5.13}$$

where N_k is the total number of time slots in the period and $C_D(v,k)$ is the degree centrality of the v^{th} vertex at time k. The log time degree is simply the natural logarithm of the time degree. Time degree is normalized and therefore always less than or equal to one, causing log time degree to be always negative, with increasingly negative numbers indicating a lower centrality.

If the heterogeneity of the system is dependent on the network structure, then the likelihood of a participant's infection at some point during the study should be correlated with appropriate network structure metrics. Pearson and Spearman correlations are run using the MATLAB statistical toolbox against the probability of infection in two baseline simulations (using close proximity, with and without immunization) against the four measures of centrality described above. Given that an individual's network location may also shape their likelihood of transmitting a pathogen when infected, for the same scenarios as above correlations of the four centrality measures are run against the average number of secondary endogenous infections directly caused by a node once it was infected. Finally, to better understand the effect of vaccination status on the correlations derived above, Student's t-test is also used to examine the difference in the four measures of centrality between participants who did and who did not report vaccination.

5.3 Results

I analyzed the response of our simulated infections to changing endogenous and exogenous infection pressure and proximity threshold required for transmission to confirm that the simulation did not produce any significant artifacts. This served both as a cross-check on the H1N1 influenza model proposed in [73] using highly detailed contact data, and as a confirmation of the plausibility of our model and approach. With plausibility of approach established, the baseline scenarios is then used to examine the impact of network dynamics on the spread of infectious disease.

5.3.1 Transmission Model

Figure 5.3 shows the attack rate (fraction of the simulated study population infected) for 72 different scenarios, where vaccination effects were considered. Figure 5.3.a shows the attack rate for simulations that only considered close proximity, which resulted in attack rates from 0.007 to 0.37. Figure 5.3.b shows the graph of scenarios where all detectable contacts were allowed for infection transmission, yielding an attack rate ranging from 0.02 to 0.8. This echoes the findings of [57], who used diary entries to demonstrate that contact quality and duration had a significant impact on disease transmission, and both the duration and quality of contact was significantly heterogeneous.

Based on the self-reports of participants' health conditions in the Flunet dataset, one individual was diagnosed with influenza by a physician and two others reported symptoms characteristic of ILI (2.7% and 8.3% of the study population, respectively). Given that the parameters related to exogenous and endogenous pressures in this model are derived based on laboratory-confirmed cases, model results for H1N1 infections are compared to the single physician-diagnosed case. Because of the stochastic elements in the model, the number of H1N1 cases occurring in a model scenario varies from realization to realization. The statistics from baseline simulations limited to contacts judged "close" and incorporating vaccination effects give a simulation mean of 0.39 for H1N1 case counts. 82.14% of realizations yielded no infections within the study population; 10.26% of realizations contained exactly one infection; 7.6% of realizations yielded 2 or more infections. As the observed count of 1 person infected falls readily within the 95% empirical fractile around the mean, it cannot be disproved the null hypothesis that our model is consistent with the underlying epidemiological process.

Figure 5.4 shows the number of exogenous and endogenous infections for two baselines restricted to the assumption that only close contacts can transmit infection, but varying the treatment of vaccination. As anticipated from its definition, the graph of exogenous infections is

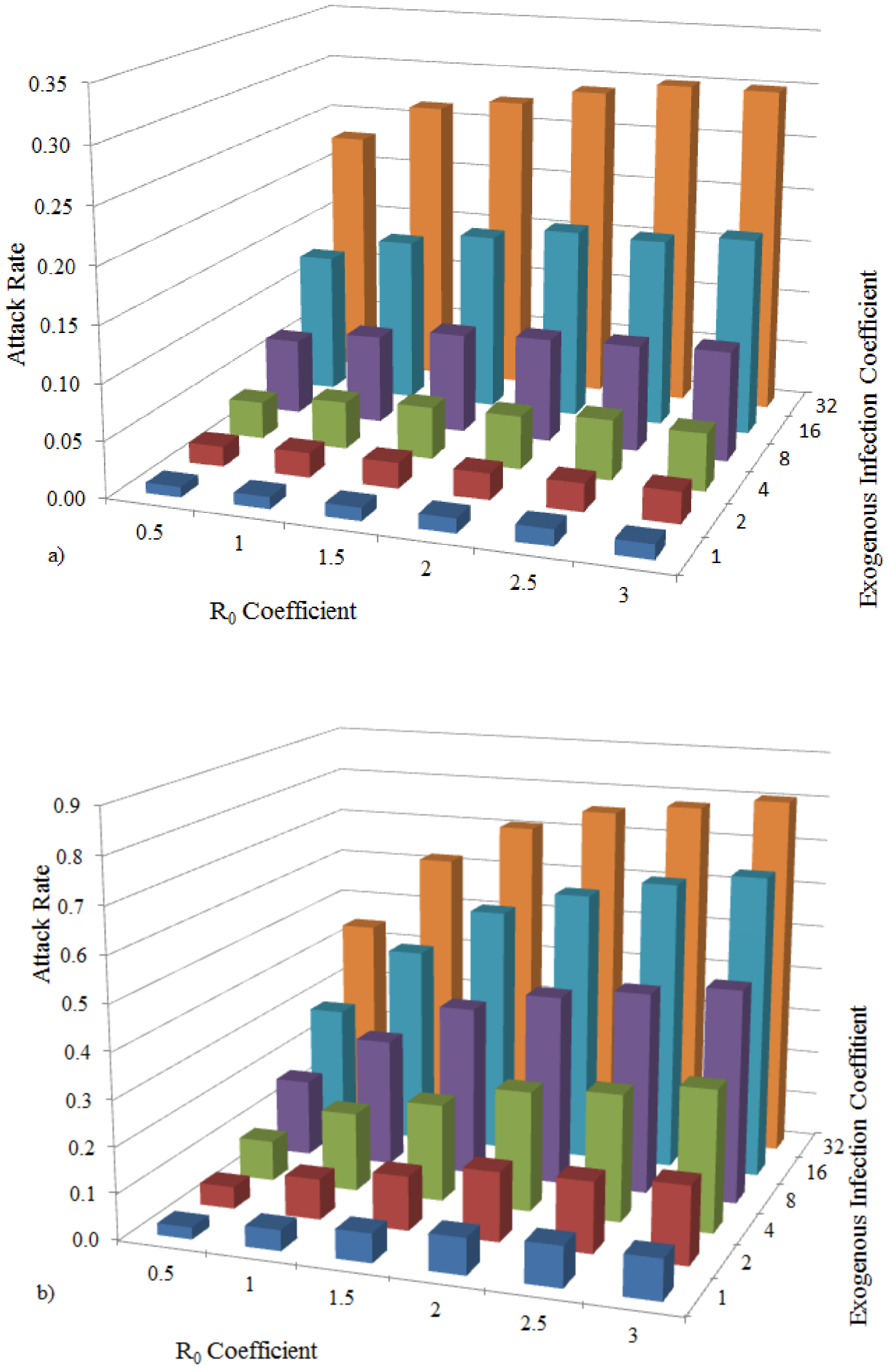


Figure 5.3: Attack rate (fraction of endogenous population infected) according to different assumptions about endogenous and exogenous infection pressure. Both graphs incorporated vaccination effect showing a), close proximity is required for contacts; b) any detectable presence qualified as contact.

independent of scenario vaccination assumptions, and it follows a trend similar to Figure 5.2. The endogenous cases require exogenous infections to begin transmission through the network. The lag in response between the exogenous and endogenous curves is therefore expected. Because the exogenous pressure diminishes to nearly zero by January, the endogenous infections also disappear quickly. As there is no chance of reinfection in the model, and because recorded contacts virtually disappear over the Christmas break, the endogenous infections also diminish to nearly zero by January.

5.3.2 Impact of Network Structure

Unlike previous works in agent-based modeling, this study has recourse to detailed contact records, containing not only high-fidelity temporal data, but also proximity estimates. The "close" category of inter-contact distance fits better with our understanding of the primary infection range for ILI. All subsequent analysis in this section is therefore performed solely on the subset of close contacts. By constraining our inquiry to a single contact criteria and ILI, the strength of our dataset is leveraged to investigate the impact of contact network structure and contact duration on the spread of disease. Because of our large-scale Monte Carlo ensembles, I believe that the variations in the underlying H1N1 model data have been well explored. Therefore, it is expected that heterogeneity in the results of the simulations to be dominated by the impact of network structure and contact duration rather than simulation artifacts.

5.3.2.1 Infection Impacts

Before analyzing impacts, it is necessary to establish appropriate metrics for measuring network connectivity. Table 5.1 shows correlation metrics (ρ) and *p* values for Pearson and Spearman correlations between likelihood of infection (average number of infections received by each node over a specific set of simulation runs) and the four measures of centrality introduced in Section 5.2.2.4, for close-contact baseline simulations with and without vaccination.

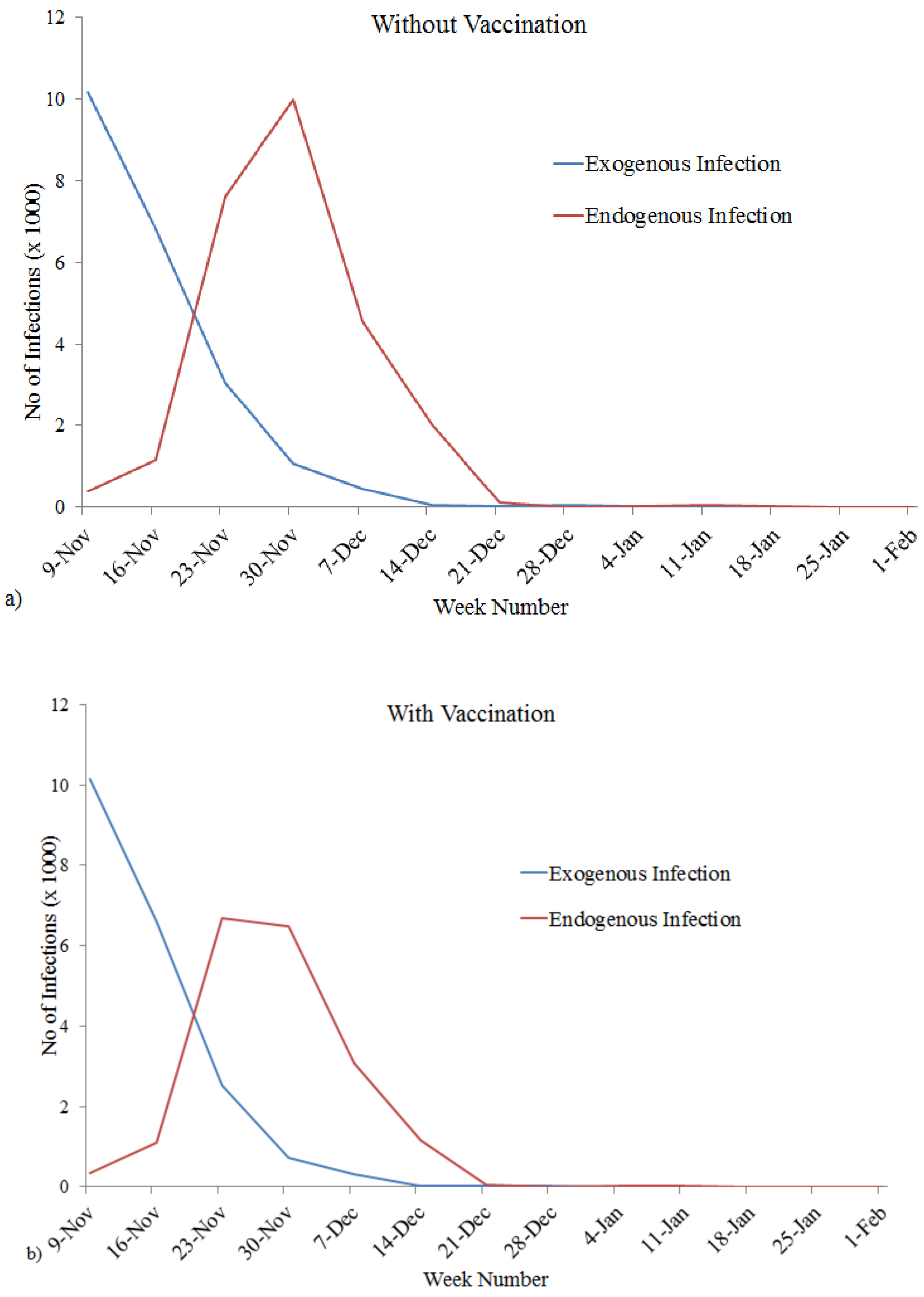


Figure 5.4: The number of exogenous and endogenous infections per week using close contacts without vaccination (a), and with vaccination (b) over the course of 100,000 runs.

Table 5-1: Correlations between centrality measures and probability of infection in baseline simulations

	Ignoring H1N1 Vaccination				*Considering H1N1 Vaccination*			
	Pearson		**Spearman**		**Pearson**		**Spearman**	
	ρ	**P**	ρ	**P**	ρ	**p**	ρ	**p**
Betweenness	0.172	0.316	0.245	0.149	0.110	0.522	0.239	0.160
Degree	0.415	0.012	0.292	0.084	0.296	0.080	0.258	0.128
TD	0.514	0.001	0.744	<0.001	0.344	0.040	0.519	0.001
LTD	0.740	<0.001	0.744	<0.001	0.503	0.002	0.519	0.001

Table 5.1 suggests that, in our experiments, betweenness centrality fails to capture the elements of network structure enhancing infection risk, as both experiments produced weak correlations with non-significant p-values. Degree centrality has similar shortcomings, producing only one significant result with a moderate correlation and p = 0.012 using Pearson's test, which is counterbalanced by the non-significant correlation using Spearman's test. However, both time degree and log time degree centralities produce significant correlation results for the non-vaccinated case for both correlation coefficients, and moderately (time degree) or very (log time degree) significant correlations for the vaccinated case.

Time degree and log time degree centralities correlated with the probability of infection, while degree and betweenness centralities did not. Because these metrics are variants of degree centrality with the addition of contact duration, it can be inferred that duration has a significant impact on infection risk. It is implicitly demonstrated that network heterogeneity drives infection rates due to the existence of a positive correlation with a network metric. If no heterogeneity existed, then no significant correlation would be likely as the variation would be random, or correlated to an independently varying parameter. Because the variation correlates with a network structure metric, it can be inferred that the heterogeneities in simulation results arise from the network structure or related parameters.

While the correlations between time degree and probability of infection remain significant in the simulations that included vaccination information, the degree of correlation diminishes. This result is not surprising, as immunization has a direct impact on the likelihood of infection, which goes to zero in the model regardless of the individual's network connectivity. This observation is interesting for two reasons: first, it demonstrates that independent variation in infection likelihood diminishes the impact of network structure; and second, that even in the face of a highly non-linear, but not universal, disturbance (not all nodes are immunized), the underlying impact of network structure remains significant.

Student's t-test was applied for each centrality measure to test the null hypothesis that those reporting and not reporting vaccination held identical mean centrality values. As it is shown in Table 5.2, p-values returned for the tests were all more than 0.7, indicating that the hypothesis of equal means could not be rejected. This further suggests that participants did not base vaccination decisions on their centrality and that differences between the vaccinated and non-vaccinated correlations in Table 5.1 are not due to a selection bias of high or low centrality participants opting for or against vaccination.

The impact of network structure and immunization on endogenous cumulative infection probability can be illustrated by plotting log time degree against the cumulative likelihood of infection for both scenarios, as shown in Figure 5.5, with a least-squares regression line for the non-immunized, log time centrality data.

Table 5-2: p-values resulting from applying Student's t-test to test the hypothesis of equal means centralities for those reporting and not reporting vaccination.

	Betweenness	**Degree**	**TD**	**LTD**
P-Value	0.93	0.723	0.721	0.911

Log time degree centrality only offers an approximation of the likelihood of infection. The log-linear regression suggests a centrality below which it is impossible to become infected (near -11). Our own simulated data refutes this, as even the least connected individuals had non-zero

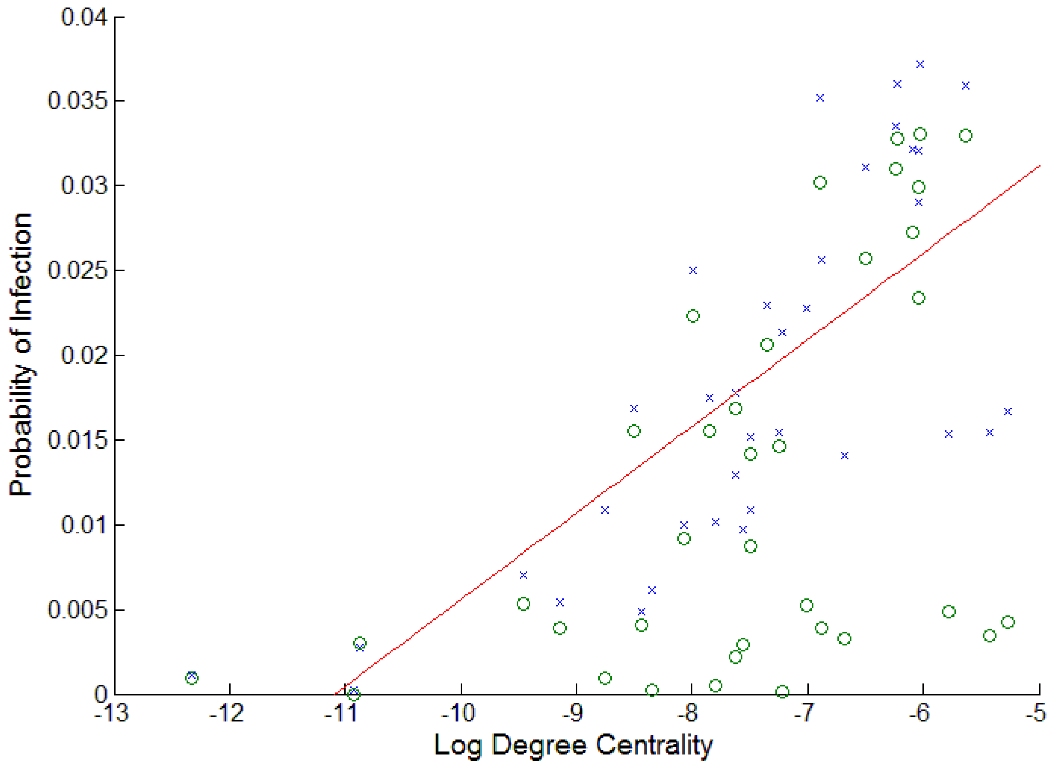


Figure 5.5: Impact of a node's log-transformed TD centrality (LTD) and immunization on endogenous infection probability. Results from two simulation scenarios are shown: One where the vaccination effect is considered (x's), and another where this effect is ignored (o's). The red line indicates the least squares fit for the case without vaccination. Dashed lines represent outliers; solid lines denote a single identified subnet.

endogenous infection counts. However, the x intercept could be usefully interpreted as the centrality below which infection probability is negligible.

People with larger log time degree centrality are linearly more likely to get infected, indicating a degree of predictive power. To fully verify this hypothesis requires a more rigorous statistical treatment and a larger participant population. This relation has limited predictive power because the slope of the line depends not only on the time degree centrality, but the parameters of the disease, which may be difficult to derive in practice. The best can be concluded is that people with logarithmically larger time degree centrality will have a measurably increased risk of infection, with all other factors held equal.

There are two primary sets of outliers in Figure 5.5: those who had very low centralities and did not often get infected in simulation (dashed line), and the main office staff who formed a semi-isolated subnet (solid lines). The office staff had a high vaccination uptake, with 3 of the 4 members receiving H1N1 vaccination, ensuring herd immunity for the entire subnet.

Finally, significant correlations between likelihood of infection and both time degree and log time degree centralities were maintained for those Monte Carlo ensembles in which behavior change was assumed to limit infection transmission to the asymptomatic period. The persistence of the correlations in this extreme case suggests that even in the presence of strong behavioral change on the part of symptomatic infectives themselves, duration-based measures are likely to remain important indicators of infection risk.

5.3.2.2 Transmission Impacts

While network structure impacts the population spread of a pathogen due to its strong effect on infection risk, it also changes risk of transmission given infection. Table 5.3 shows the correlations between an agent’s centrality and the average number of secondary infections caused by that agent per episode of infection of that node. The results are generally similar to the infection risk correlations reported in Table 5.1. Traditional network measures (betweenness and degree centrality) still exhibit weak and non-significant correlations. By contrast, time degree and log time degree centralities exhibit stronger and consistently statistically significant correlations. As with risk of infection, when H1N1 vaccination is considered, the correlations are lowered.

5.4 Discussion

This work used Flunet as high-resolution contact data and combined it with transmission models of airborne infections. Previous work in epidemiology and public health has focused on heterogeneity of contacts in human population, particularly regarding duration and diversity. These works have shown the importance of this heterogeneity in modeling infection transmission [54, 55]. But the traditional data collection methods such as self-reports and diaries were unable

Table 5-3: Correlations between network measures for a node and number of secondary infections caused by that node per each time it is infected.

	Ignoring H1N1 Vaccination				*Considering H1N1 Vaccination*			
	Pearson		**Spearman**		**Pearson**		**Spearman**	
	ρ	**P**	ρ	**P**	ρ	**p**	ρ	**p**
Betweenness	0.184	0.283	0.244	0.151	0.139	0.420	0.253	0.137
Degree	0.399	0.016	0.300	0.076	0.302	0.073	0.296	0.080
Time Degree	0.665	<0.001	0.895	<0.001	0.472	0.004	0.615	<0.001
Log Time Degree	0.802	<0.001	0.895	<0.001	0.590	<0.001	0.615	<0.001

to record and represent this heterogeneity. Therefore the models built on traditionally collected data lack sufficient information on contact duration and have to make assumptions in this regard. These assumptions have first order effect on the model outcome and the accuracy of the predictions from the model. Here Flunet is used as high-resolution contact pattern to reveal the importance of this heterogeneity in epidemiological modeling.

The main contribution of the chapter is combining epidemiological modeling of infection transmission with electronically-collected minute-resolution contact data. This combination has shown the importance of contact duration together with contact frequency in modeling pathogen transmission particularly for air-borne infections.

The study in this chapter has shown two important findings. First the results showed the network measurements which are tied with contact duration such as time degree centrality are more strongly associated with infection risks compared to traditional measures like degree centrality and betweenness. Secondly, the results from Table 5.3 showed the direct relation between contact duration and the infection rate. In other words, having longer contact duration, even with a smaller number of people leads to more infection transmission by the infective individual.

Although even the traditional self-reports also can collect contact duration from individuals and put weights on each collected contact, the survey results in Flunet showed the limited tolerance of people to regularly record this information. The results of the surveys represented that participant compliance with requested self-reporting of their 5 most common weekly contact durations was just above 25%. Even with their willingness for reporting these contacts, remembering all the contacts with a minute-resolution, as collected in this dataset and used in the model, is nearly impossible using self-report and diaries.

In addition to the importance of the finding presented in this chapter, this work has some limitations. First, the Flunet dataset is limited to a small sample size and a selection bias toward university students. Although the sample size of this study is consistent with other similar electronically data collection systems, collected data through self-reports is normally represented by notably larger population sizes. The second limitation of the work is regarding the behavior change of the infected individuals during the symptomatic period. In this study, it is assumed that people don't change their behavior during the symptomatic period of the illness. While I tried to compensate for this by simulating an extreme behavioral change and removed the infected participants from the simulation cycle, this scenario doesn't reflect all potential behavioral outcomes. Thirdly, due to a lack of data on different infection transmission rates at each illness period, it is simply assumed the same transmission rate for both asymptomatic shedding and symptomatic periods.

Regardless of the limitations in this study, here it is been demonstrated significant findings regarding the importance of contact duration in measuring individuals' centrality in infection transmission networks. These findings can be improved in the future by using datasets with larger population sizes, and improving the modeling of behavioral change during symptomatic period.

CHAPTER 6: CONCLUSION

Understanding human behavioral patterns forms an important part of many sciences, ranging from urban planning and city development to epidemiology and public health. Different methods have been used to collect and analyze data on human behaviors, including diary recording or self-report by individuals or observations by experimenters. Recent advances in computer technology have made it possible to record these behaviors together with any other pertinent environmental data using wearable sensors and subsequently to analyze them in order to reveal the existing characteristics. Although the datasets collected by previous researches contributed significant understanding of human behavioral patterns, the complexity of human movement and contact patterns leaves many areas uninvestigated. Here it is focused on human proximal and geographical patterns and tried to use sensor devices to collected these patterns. In addition to basic analyses, the resulting data was used in two applications, delay tolerant networking and epidemiological modeling, in order to leverage the existing knowledge in these fields.

6.1 Discussion

Designing a system capable of recording all aspects of human behavioral patterns is not feasible. One of the preliminary decisions is to determine essential parameters to measure based on the research purpose, and to design the data collection system accordingly. This work targeted the applications in networking and epidemiological modeling, leading to a data collection system focused on human proximal and geographical contact patterns, cross-linked with information on their health status.

Another important design decision is to pick the population from which participants will be selected. This decision depends on the purpose of the research and also highly affects the feasibility of the study. For example, certain studies might require focusing on a specific population, such as study of infection transmission in long-term care facilities or primary school students, as they are highly vulnerable to receiving or transmitting the infection, while other research might not require a special set of participants. Selecting a certain population can both simplify the experiment and introduce new risks. Issues such as controlling the experiments, performing required coordination with participants, and introducing selection bias to the dataset potentially make the dataset inappropriate or incomplete. On the other hand, focusing on a special population can tailor the population to reflect the goals of the study and also provides the opportunity to reveal potential differences between the selected populations and more general populations enumerated in other similar studies. This work focused on students and university staff as the participant population. This decision simplified participant selection and data collection, but also introduces a selection bias to the recorded data towards academic life.

Another important design issue is the confidentiality of participants' information. Research in human behavioral patterns deals with recording and analyzing data related to the everyday life of participants. Depending on the type of the collected data, this information potentially can disclose the identity of participants even after the anonymization process. For example, data on human movement patterns collected through different positioning systems such as GPS can simply reveal the location of a participant's home or work-place and the times of day they are typically in each. Therefore, the participants have to be fully informed about privacy aspects prior to collecting their data. The data collected in this work could be fully anonymized. This study was reviewed by the Research Ethics Board of the University of Saskatchewan and found to be of minimal risk to the participants.

Determining the study duration and subsequently keeping participants motivated during the study period are other factors which need to be considered. Datasets from short studies might not be able to represent or reveal existing patterns, while in long studies it is harder to keep the participants motivated. Experiences from Flunet showed that as the study progressed, participants become indifferent to study goals. This reluctance can have a considerable effect on the quality of the data collected. Selecting an appropriate study length based on the research requirements and to maintaining participants' motivation are especially important for long-term studies. Although there is not a single solution for this issue and a useful strategy depends on many factors, resolving these concerns before starting the study is important. A 3-month period during winter was selected as a study length for two reasons: first, it is long enough to provide useful data on human contact patterns, and second, it covered a flu season, which is important for health modeling purposes and for studying the spread of infections. Participants were motivated to contribute to the study by providing an honorarium to each person, and also assigning prizes after the study based on compliance. The data collection is monitored during the experiment and tried to improve the collection process by reminding apparently forgetful participants either directly or through email. The collected data in this study emphasized the advantages of proposed automatic data collection method using sensors over traditional survey and diary-based methods.

6.2 Future Work

The work presented in this thesis can be continued along different vectors of research. In terms of data collection, further studies can be conducted to record additional datasets on human proximal and geographical contact patterns. Such data could provide additional insights into and understanding of human contact patterns by applying the experience gained through this work and considering gaps which exist in current datasets. Future data collection studies also can focus on different populations in order to reveal the potential differences between a certain population and the academic participants in this study. Using devices other than sensor modules (such as smart phones) to motivate people to participate and to provide wider and more reliable results can improve data collection in future studies.

With respect to routing in Delay Tolerant Networks, the research described here raises several interesting avenues for additional work. As noted previously, the current state of empirical data available for the study of the detailed interaction of people and place is severely limited. Both better tools and additional datasets could help fill this void. Further analysis of the hypotheses regarding the role of stationary nodes, particularly with respect to high churn locations, would lead to a better understanding of not only routing efficiency, but infrastructure utilization. Also, additional analyses on the relation between the centrality of people and centrality of their primary location in order to better recognizing nodes whose popularity is based on their location, and subsequently designing heuristics to move the load from localized nodes to stationary resources could have considerable impact in DTN performance.

There are two possible venues to extend the research presented here on epidemiological modeling: regarding the dataset aspects and model design respectively. From a dataset point of view, using a broader and more diverse population and longer study period would overcome some of the limitations of this work. For example, the characterization of individual's behavioral change during symptomatic period could be investigated with a different design. Having finer contact resolution in the dataset to capture close contacts could provide better data on infection transmission. From a model design perspective, the model used in this work can be refined by designing a more detailed H1N1 transmission model by adding additional agent states, such as hospitalization and death. Using more representative data for contacts between participants and the population outside of the study, and more detailed information on population immunization states could refine the representation of exogenous infection pressure used in Chapter 5, yielding better results.

In addition to the areas discussed in this thesis, the collected dataset or future datasets can be used together with more advanced techniques, such as Social Network Analysis to reveal new characteristics and patterns from human contacts and to apply them into these fields of study.

6.3 Thesis Summary

The primary focus of this thesis is on human behavioral patterns, particularly regarding proximal and geographical contacts, which underlie human dynamic networks. A system capable of recording proximal and geographical contacts of people, together with information on their health status is designed and implemented. The system collected data from a sample population consisting of 36 participants and their proximity to 11 high-traffic public places during 3 months of winter 2009-2010. The dataset analyses demonstrated the consistency of the collected dataset with previous work over different measures such as contact and meeting time distributions. The presented dataset, which combines participant's proximal and geographical contacts with their health status information, is the main contribution of this part of thesis and its results have been published in [53].

The dataset was applied to Delay Tolerant Networks to improve routing performance by studying the interactions between people and places. The focus is on the effect of stationary nodes potentially available in public places as relay nodes and designed an algorithm which uses these nodes to increase the delivery ratio. Unlike previous work which considered a social network representation of human contacts for routing, this work considered both proximal and geographical aspects of human contacts for routing purposes. The results showed using stationary nodes available at different public places not only improves the routing performance, but can also provide load balancing for mobile connectivity to the stationary nodes. The results also suggested that, similar to people, places have different popularity; considering this in the algorithm design, can reduce the total overhead in the network. Studying the impact of stationary nodes on routing performance and load balancing in DTNs together with the presented algorithm which utilizes these nodes is the main contribution of this thesis in DTNs.

In the second application, the dataset was used for modeling the spread of airborne infection pathogens in a population. A new model for H1N1 infection spread is designed by combining an existing individual-level Canadian infection transmission model with collected contact data as the population’s contact pattern. The minute-resolution contact patterns used in the model reflected the importance of considering existing heterogeneity in contact density and diversity on model outcome. It is also shown that the network measurements which take the contact duration into account are more closely associated with risk of infection transmission compared to traditional centrality measurements. The results demonstrated that longer contact duration increases the chance of getting an infection, regardless of the diversity of the contact. Using minute-resolution contact data as the contact pattern of an infection transmission model, which yielded valuable insights with respect to the relation between contact duration and transmission risk, is the main contribution of this thesis in epidemiology and public health.

CHAPTER 7: REFERENCES

1. O'Neill, E., Kostakos, V., Kindberg, T., Schiek, A. F., Penn, A., Fraser, D. S., and Jones T. 2006. Instrumenting the City Developing Methods for Observing and Understanding the Digital Cityscape. *Lecture Notes in Computer Science*, Volume 4206/2006, 315-332.
2. Kjeldskov, J. and Paay, J. 2005. Just-for-us a context-aware mobile information system facilitating sociality. *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*, September 19 – 22, Salzburg, Austria, 23 - 30.
3. Jones, Q., Grandhi, S. A., Terveen, L. and Whittaker S. 2004. People-to-People-to-Geographical-Places: The P3 Framework for Location-Based Community Systems. *Computer Supported Cooperative Work (CSCW)* Volume 13, Numbers 3-4, 249 - 282.
4. Jones, Q. and Grandhi, S. A. 2005. P3 Systems Putting the Places Back into Social Networks. *Internet Computing, IEEE* Sept.-Oct. 2005 Volume 9 Issue 5, 38 - 46.
5. Jones, Q., Borcea, C., Hiltz, S. R., Manikopoulos, C. and Amento, B. 2006. Urban Enclave Location-Aware Social Computing. *Proceedings of Internet Research 7.0: Internet Convergences*, Brisbane, Australia.
6. Eagle, N. and Pentland, A. 2006. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing,* Volume 10 Issue 4, 255 - 268.
7. Chaintreau, A., Mtibaa, A., Massoulie, L. and Diot, C. 2007. The diameter of opportunistic mobile networks, *Proceedings of the 2007 ACM CoNEXT conference*, Article No. 12, December 10 – 13, New York, United States.
8. Crawdad Community Resource for Archiving Wireless Data At Dartmouth, viewed 5 January 2011 <http://www.crawdad.org>.
9. Fall, K. 2003. A delay-tolerant network architecture for challenged internets, *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, August 25-29, 2003, Karlsruhe, Germany, 27 - 34.
10. Vahdat, A. and Becker D. 2000. Epidemic routing for partially connected ad hoc networks, *Technical Report CS-200006*, Duke University.

11. Lindgren, A., Doria, A. and Schelén, O. 2004. Probabilistic Routing in Intermittently Connected Networks, *Lecture Notes in Computer Science*, Volume 3126/2004, 239 - 254.

12. Hui, P., Crowcraft, J. and Yoneki, E. 2008. Bubble rap: social-based forwarding in delay tolerant networks, *In Proceeding MobiHoc 2008 Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing*, May 26 – 30, Hong Kong, China, 241 - 250.

13. Smith, K. P. and Christakis, N. A. 2008. Social Networks and Health, *Annual Review of Sociology*, Vol. 34, 405 - 429.

14. Al-Azem, A. A. 2006. Social Network Analysis in Tuberculosis Control Among the Aboriginal Population of Manitoba, *University of Manitoba Ph.D. Thesis*

15. Tempalski, B. and McQuie, H. 2008. Drugscapes and the role of place and space in injection drug use-related HIV risk environments, *International Journal of Drug Policy*, Volume 20, Issue 1, February 2008, 4 - 13.

16. Brankston, G., Gitterman, L., Hirji, Z., Lemieux, C. and Gardam, M. 2007. Transmission of Influenza A in Human Beings, *The Lancent Infection Diseases*, Volume 7, Issue 4, April 2007, 257 - 265.

17. MicaZ sensor board datasheet, viewed 8 January 2011 <http://www.openautomation.net/uploadsproductos/micaz_datasheet.pdf>.

18. TelosB sensor board datasheet, viewed 8 January 2011 <http://www.willow.co.uk/TelosB_Datasheet.pdf>.

19. Lenczner, M., Grégoire, B. and Proulx, F. 2007. CRAWDAD/data set ilesansfil/wifidog (v. 2007-08-27)

20. Srinivasan, V., Motani, M. and Ooi, W. T. 2006. CRAWDAD/data set nus/contact (v. 2006-08-01)

21. Schulman, A., Levin, D. and Spring, N. 2008. CRAWDAD/data set umd/sigcomm2008 (v. 2009-03-02)

22. Kotz, D. and Essien, K. 2005. Analysis of a Campus-wide Wireless Network.*Wireless Networks*, Volume 11, January 2005, 115 – 155.

23. Tournoux, P. U., Leguay, J., Benbadis, F., Conan, V., Amorim, M. D. and Whitbeck, J. 2009. The accordion phenomenon: Analysis, characterization, and impact on DTN

routing. *Proceedings IEEE INFOCOM*, April 19 – 25, 2009, Rio De Janeiro, Brazil, 1116 - 1124.

24. Srinivasan, V., Motani, M. and Ooi. W. T. 2006. Analysis and Implications of Student Contact Patterns Derived from Campus Schedules. *In Proceedings of ACM MobiCom*, September, 24 – 29, 2006, Los Angeles, USA, 86 - 97.
25. LeBrun, J. and Chuah, C. 2006. Bluetooth Content Distribution Stations on Public Transit. *In MobiShare 2006: Proceedings of the 1st Workshop on Decentralized Resource Sharing in Mobile Computing and Networking*, July 25, 2006, Los Angeles, USA, 63 - 65.
26. Mcnamara, L., Mascolo, C. and Capra, L. 2008. Media sharing based on collocation prediction in urban transport, *In Proceedings of ACM MobiCom*, September 14 – 19, 2008, San Francisco, USA, 58 - 69.
27. Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S. and Chong, S. 2009. CRAWDAD trace ncsu/mobilitymodels/GPS/NC_State_Fair (v. 2009-07-23)
28. Small, T. and Hass, A. 2005. Resource and Performance trade-offs in delay-tolerant wireless networks, *In Proceedings of the 2005 ACM SIGCOMM Workshop on Delay-Tolerant Networking*, August 26, 2005, Philadelphia, USA, 260 - 267.
29. Spyropoulos, T., Psounis, K. and Raghavendra, C. S. 2005. Spray and wait: an efficient routing scheme for intermittently connected mobile networks, *In Proceedings of the 2005 ACM SIGCOMM Workshop on Delay-tolerant Networking*, August 26, 2005, Philadelphia, USA, 252 - 259.
30. Spyropoulos, T., Psounis, K. and Raghavendra, C. S. 2004. Single-copy routing in intermittently connected mobile networks, *First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks (SECON)*, October 4 – 7, Santa Clara, USA, 235 - 244.
31. Spyropoulos, T., Psounis, K. and Raghavendra, C. S. 2008. Efficient routing in intermittently connected mobile networks: the multiple-copy case, *Journal IEEE/ACM Transactions on Networking (TON)* Volume 16 Issue 1, February 2008, 77 - 90,
32. Juang, P., Oki, H., Wang, Y., Martonosi, M., Peh, L. S. and Rubenstein, D. 2002. Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences

with ZebraNet, *In Proceedings of the 10th international conference on Architectural support for programming languages and operating systems,* October 5 – 9, 2002, San Jose, USA, 96 - 107.

33. Balasubramanian, A., Levine, B. and Venkataramani, A. 2007. DTN routing as a resource allocation problem, *In Proceedings of the 2007 conference on Applications, technologies, architectures, and protocols for computer communications*, August 27-31, August 27 – 31, 2007, Kyoto, Japan, 373 - 384.
34. Liu, C. and Wu, J. 2009. An Optimal Probabilistic Forwarding Protocol in Delay Tolerant Networks, *Proceedings of the tenth ACM international symposium on Mobile ad hoc networking and computing*, May 18 – 21, 2009, New Orleans, USA, 105 – 114.
35. Hui, P., Chaintreau, A., Scott, J., Gass, R., Crowcroft, J. and Diot, C. 2005. Pocket switched networks and human mobility in conference environments, *In Proceedings of the 2005 ACM SIGCOMM Workshop on Delay-tolerant Networking*, August 26, 2005, Philadelphia, USA, 244 - 251.
36. Yoneki, E., Hui, P. and Crowcroft, J. 2008. Distinct types of hubs in human dynamic networks, *In Proceedings of the 1st Workshop on Social Network Systems*, April 1, 2008, Glasgow, Scotland, 7 - 12.
37. Yoneki, E. 2008. Visualizing communities and centralities from encounter traces, *In Proceeding CHANTS 2008 of the Third ACM workshop on Challenged Networks*, September 14 – 19, 2008, California, USA, 129 - 132.
38. Yoneki, E., Greenfield, D. and Crowcroft, J. 2009. Dynamics of Inter-Meeting Time in Human Contact Networks, *In Intl. Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, July 20 – 22, 2009, Athens, Greece, 356 - 361.
39. Li, F. and Wu, J. 2009. LocalCom: a community-based epidemic forwarding scheme in disruption-tolerant networks, *In Proceedings of the 6th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communication and Networking*, June 22 – 26, 2009, Rome, Italy, 1 - 9.
40. Grundy, A. and Radenkovic, M. 2010. Decongesting Opportunistic Social-based Forwarding, *Seventh International Conference on Wireless On-demand Network Systems and Services (WONS)*, Feb. 3 – 5, 2010, Kranjska Gora, Slovenia, 82 - 85.

41. Pujol, J., Toledo, A. L. and Rodriguez, P. 2009. Fair routing in delay tolerant networks, *In INFOCOM 2009*, April 19 – 25, 2009, Rio De Janeiro, Brazil, 837 - 845.

42. Yoneki, E. 2009. The importance of Data Collection for modeling Contact Networks, *In Proceedings of Intl. Conference on Computational Science and Engineering*, August 29 – 31, 2009, Vancouver, Canada, 940 - 943.

43. Yuan, Q., Cardei, I. and Wu, J. 2009. Predict and relay: an efficient routing in disruption-tolerant networks, *In Proceedings of the 10th ACM Intl. Symposium on Mobile ad hoc Networking and Computing*, May 18 – 21, 2009, Louisiana, USA, 95 - 104.

44. Tian, Y. and Li, J. 2010. Location-aware routing for delay-tolerant networks, *In Proceedings of ChinaCom'10,* August 25 – 27, 2010, Beijing, China

45. Wen, H., Liu, J., Lin, C., Ren, F., Li, P. and Fang, Y. 2009. RENA region-based routing in intermittently connected mobile network, *In Proceeding MSWiM 2009 the 12th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, October, 26 – 30, 2009, Canary Islands, Spain, 280 - 287.

46. Madan, A., Cebrian, M., Lazer, D. and Pentland, A. 2010. Social Sensing for Epidemiological Behavior Change, *In Proceedings of the 12th ACM Intl. Conference on Ubiquitous Computing*, September, 26 – 29, 2010, Copenhagen, Denmark, 291 - 300.

47. Coleman, J. S. 1990. Foundations of Social Theory, *Cambridge, MA, Harvard University Press*

48. Roy, C. J. and Milton, D. K. 2004. Airborne transmission of communicable infection—the elusive pathway. *The New England Journal of Medicine*, 350: 1710 - 1712.

49. Health Canada. 1999. *Routine practices and additional precautions for preventing the transmission of infection in health care: revision of isolation and precaution techniques*. CCDR supplement. Ottawa: Health Canada, 1999.

50. Jolly, A. M., Muth, S. Q., Wylie, J. L. and Potterat, J. J. 2001. Sexual Networks and Sexually Transmitted Infections: A Tale of Two Cities. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, Volume 78, 433 - 445.

51. Wyliea, J. L., Shahb, L. and Jolly, A. 2006. Incorporating Geographic settings into a social network analysis of injection drug use and blood-borne pathogen prevalence, *Health & Place*, Volume 13, Issue 3, 617 – 628.

52. Darke, S., Kaye, S. and Ross, J., 2001. Geographical injecting locations among injecting drug users in Sydney, Australia. *Addiction 96*, 241 – 246.

53. Hashemian, M. S., Stanley, K. G. and Osgood, N. 2010. Flunet: Automated tracking of contacts during flu season, *In Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, 348-353.

54. Hethcote H. W. and Yorke J. A. 1984. Gonorrhea transmission dynamics and control. *Springer Lecture Notes in Biomathematics*. Berlin, Springer.

55. Kretzschmar, M. and Morris, M. 1996. Measures of Concurrency in Networks and the Spread of Infectious Disease. *Mathematical Biosciences*, Volume 133, 165 – 195.

56. Read, J. M., Eames, K. T. D. and Edmunds, W. J. 2008. Dynamic social networks and the implications for the spread of infectious disease. *Journal of The Royal Society Interface*, Volume 5, 1001 – 1007.

57. Smieszek, T. 2009. A mechanistic model of infection: Why duration and intensity of contacts should be included in models of disease spread. *Theoretical Biology and Medical Modelling*, November 2009, 6 – 25.

58. Wallinga, J., Teunis, P. and Kretzschmar, M. 2006. Using Data on Social Contacts to Estimate Age-specific Transmission Parameters for Respiratory-spread Infectious Agents. *American Journal of Epidemiology,* 936 – 944.

59. Near and Far Field Wikipedia Article, viewed 5 January 2011 <http://en.wikipedia.org/wiki/Near_and_far_field>.

60. Morris, M., 2004. International Union for the Scientific Study of Population. Network epidemiology: a handbook for survey design and data collection. *Oxford; New York: Oxford University Press*.

61. Network Simulator 3 Reference Manual, viewed 5 January 2011 <http://www.nsnam.org/docs/release/manual.pdf>

62. Chaintreau, A., Hui, P., Crowcroft, J., Diot, C., Gass, R. and Scott. J. 2005. Pocket Switched Networks: Real-World Mobility and its Consequences for Opportunistic Forwarding. *Technical Report UCAM-CL-TR-617, University of Cambridge, Computer Laboratory,* February 2005.

63. Hui, P., Lindgren, A. and Crowcroft, J. 2009. Empirical evaluation of hybrid opportunistic networks. *In Proceedings of COMSNETS 2009,* January 5 – 10, 2009, Bangalore, India, 1-10.

64. Hui, P. and Lindgren, A. 2008. Phase transitions of opportunistic communication, *In Proceedings of the third ACM workshop on Challenged networks*, September 15, 2008, 73 – 80.

65. Song, C., Qu, Z., Blumm, N. and Barabási, A. 2010. Limits of Predictability in Human Mobility, *Science 327(5968)*, 1018 – 1021.

66. Lee, K., Hong, S., Kim, S. J., Rhee, I. and Chong, S. 2009. SLAW: A Mobility Model for Human Walks, *In INFOCOM 2009*, April 2009, Rio de Janeiro, Brazil, 855 – 863.

67. Paulos, E. and Goodman, E. 2004. The familiar stranger: anxiety, comfort, and play in public places, *In Proceeding CHI 2004 of the SIGCHI conference on Human factors in computing systems*, April 24 – 29, Vienna, Austria, 223 – 230.

68. Binder, S., Levitt, A. M., Sacks, J. J. and Hughes, J. M. 1999. Emerging Infectious Diseases: Public Health Issues for the 21st Century, *Science 284*, 1311 – 1313.

69. Mabry, P. L., Marcus, S. E., Clark, P. I., Leischow, S. J., and Mendez, D. 2010. Systems science: A revolution in public health policy research. *American Journal of Public Health*, 1161 – 1163.

70. Eames K. T. D. and Keeling M. J. 2003. Contact tracing and disease control. *Proceedings of the Royal Society B: Biological Sciences*, 270, 2565 – 2571.

71. FluWatch, Public Health Agency of Canada, viewed 14 November 2010 <http://origin.phac-aspc.gc.ca/fluwatch/09-10/w34_10/index-eng.php>.

72. H1N1 Update, Saskatoon Health Region, viewed 13 November 2010 <http://regionreporter.wordpress.com/2010/01/08/h1n1-update/#more-439>.

73. Tuite, A. R., Greer, A. L., Whelan, M., Winter, A., Lee, B., Yan, P., Wu, J., Moghadas, S., Buckeridge, D., Pourbohloul, B. and Fisman, D. N. 2010. Estimated epidemiologic parameters and morbidity associated with pandemic H1N1 influenza. *CMAJ*, 182(2), 131 – 136.

74. De, P., Singh, A., Wong, T., Yacoub, W. and Jolly A. 2004 Sexual network analysis of a gonorrhoea outbreak. *Sex Transmission Infections*, 280 – 285.

75. Jolly, A. M., Muth, S. Q., Wylie, J. L. and Potterat, J. J. 2001 Sexual Networks and Sexually Transmitted Infections: A Tale of Two Cities. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, Volume 78, 433 – 445.
76. Valente, T.W. 2010 Social Networks and Health: Models, Methods, and Applications, *Oxford University Press*.
77. Keeling, M. J. and Eames, K. T. D. 2005 Networks and epidemic models. *Journal of The Royal Society Interface*, 2, 295 – 307.
78. Yoneki, E., Hui, P. and Crowcroft, J. Wireless Epidemic Spread in Dynamic Human Networks, *Bio-Inspired Computing and Communication*, Volume 5151, Springer, 116 – 132.

Appendix A: Flunet Weekly Survey Questions

1. Please enter your device ID.
2. What percentage of contact time does your top five contact lists represent out of contact with study participants over the last week?
3. What percentage of contact time does your top five study participant contact list represent out of contact with anyone, anywhere (including at home, work, standing in line for coffee, etc) over the last week?
4. Have you experienced flu like symptoms since your last report?
5. Which symptoms:
   - Cough
   - Runny nose
   - Headache
   - Sore throat
   - Chest pain
   - Muscle pain
   - Diarrhea
   - Abdominal pain
   - Cold shivers
   - Nausea
   - Irritated eyes
6. Did you have sudden fever?
7. How high was the fever? (between 37C and 40C, in steps of 0.5 degree)
8. When did it start?
9. When did it finish? (If symptoms are persisting, leave the end date blank. If symptoms have crossed two reporting periods, report actual start and end dates.)
10. Were you diagnosed by a physician with the flu since your last report?
11. Did your wireless device ever become inactive for any reason (e.g. voluntarily turned off, battery failure, node failure) since your last report?

12. Approximately how many hours was your wireless device inactive?

13. Please enter the cause or reason for the wireless device inactivity.

Appendix B: Flunet Demographic Survey Questions

1. Please enter your device ID:
2. Please select your age:
   - 18 - 19
   - 20 - 25
   - 25 - 30
   - 30 - 35
   - 35 - 40
   - 40+
3. Please select your gender:
   - Female
   - Male
4. Approximately how many hours per week do you spend at the university on average?
   - Less than 20
   - 20 - 30
   - 30 - 40
   - 40 - 50
   - 50 - 60
   - More than 60
5. Approximately how many hours do you spend in your primary location at the university on average?
   - less than 10
   - 10 - 20
   - 20 - 30
   - 30 - 40
   - more than 40

6. How do you travel to and from the university? Check all that apply.
   - Walk
   - Bike
   - Drive
   - Carpool
   - Transit
   - Other
7. Do you see members of the university community socially off campus?
   - Yes
   - No
8. Do you normally get a flu shot?
   - Yes
   - No
9. Did you receive a regular flu vaccination this year?
   - Yes
   - No
10. If you have received a regular flu vaccination, please supply the date (If you cannot recall the exact day, provide your best estimate):
11. Did you receive an H1N1 vaccination this year?
    - Yes
    - No
12. If you have received an H1N1 vaccination, please supply the date (If you cannot recall the exact day, provide your best estimate):
13. During the course of the study, did you smoke tobacco on a daily basis, less than daily, or not at all?
    - Daily for at least a portion of the survey
    - Less than Daily for at least a portion of the survey but not daily
    - Not at all
14. Did you consciously change your behaviour during the survey in response to concerns regarding the risk of catching flu?

- Yes
- No

15. Did you consciously change your behaviour during the survey in response to concerns regarding the risk of transmitting flu from yourself to others?
    - Yes
    - No
16. You answered "yes" with respect to either of the above, did this behaviour change involve limiting interaction with other people?
    - Yes
    - No
17. Please note that the following question will not impact your chance of winning a prize in the draw, or affect your compensation: On a scale of 0 to 10, please rate how careful you were with carrying your mote with you whenever on campus?
18. Please note that the following question will not impact your chance of winning a prize in the draw, or affect your compensation: On a scale of 0 to 10, please estimate how careful you were with refreshing your batteries.
19. Please leave any comments you wish for the study organizers.