Identifying and Comparing Opportunistic and Social Networks

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Mona Abdelbaset Sadek Ali

2013

Cardiff University

School of Computer Science & Informatics

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed . . . . (candidate)

Date . . . .

Statement 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed . . . . (candidate)

Date . . . .

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed . . . . (candidate)

Date . . . .

Statement 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed . . . . (candidate)

Date . . . .

Abstract

Recent developments in computation and communication technologies are making big changes to the method in which people communicate with each other. Online social networks, wireless technologies and smart-phones are very common and enable communication to be maintained while people are mobile. These types of communication are closely related to humans because they carry the devices.

In this research, we detect, analyse and compare opportunistic networks with human social networks. Currently opportunistic networking platforms have not gone beyond the research and development stage where they are challenging to design and implement. Therefore we develop an indoor mobility tracking system to track the participant movement inside buildings and to record the physical interaction between participants using Bluetooth technology. This system has two different ways of abstracting opportunistic networks from the experimental data: mobility (device-to-building) and co-located (device-to-device) interactions.

The mobility detection system has been studied using a volunteer group of students in the School of Computer Science and Informatics. This group also has been studied to understand their social networking structure and characteristics by using electronic survey methodology. Different techniques have been used to investigate the individuals' networks from the survey and mobility movements and a comparison between them.

From a precision and recall technique, we find that 60-80 % of the participants' social network is embedded in the opportunistic network but a small proportion 10-20% of the opportunistic network is embedded in the social network. This shows the presence of many weak links in the opportunistic network that means the opportunistic network connectivity requires a very small number of key-players to disseminate information throughout the network. We also examine both networks from the perspective of information dissemination. We find that device-to-device create many more weak links to disseminate information rather than server detection. Therefore, information quickly floods throughout the co-located network.

Acknowledgements

This thesis would not have been possible without the help and support of ALLAH. I would like to extend my thanks to ALLAH. Also The help and support of all the kind people around me. I would like to extend my thanks to all individuals who stood by me throughout the years of my study, but I wish to particularly mention a number of people who have had a profound impact on my work.

Above all, I am heartily thankful to my supervisor, Prof. Roger Marcus Whitaker, whose encouragement, guidance and support from start to end enabled me to develop my research skills and understanding of the subject. His wealth of knowledge, sincere advice and friendship with his students made this research fruitful as well as enjoyable. Also Dr.Stuart M. Allen who support and encourage me a lot to finish my study on a desirable time.

It is an honour for me to extend my thanks to my employer Minia University, Minia, Egypt, for offering my postgraduate scholarship, and providing all financial support needed to complete my degree.

I would like to acknowledge the academic and technical support of the School of Computer Science & Informatics at Cardiff University, UK and all its staff members. I am also grateful for the kindness, help and support of all my colleagues at Cardiff University (Dr. Manar I. Hosny, Dr Hessa Al-Salamah, Dr Ehab Elgindy, Dr Ahmed A. Alazawii, Haya Almagwashi, Shada Al-Salamah, Asma Abdaljalil, Soha A.A. Mohamed, Hannah A. Aldahawi, Matthew M. Williams, Chris Gwilliams, Will M. Webberley and Wan S Wan Awang) .

Last but by no means least, I owe my deepest gratitude to my parents, my husband (Dr Mohamed Moustafa), and my cute princesses (Haneen and Salma) for their unconditional support and patience from the beginning to the end of my study.

Contents

List of Figures

List of Tables

List of Algorithms

List of Acronyms

AOA Angle of Arrival

TOA Time of Arrival

API Application Programming Interface

BT Bluetooth

FTP File Transfer Protocol

GByte Giga Byte

IR Infrared

J2ME Java to Mobile Enterprise

KB Kilo Byte

MAC Media Access Control

OPNET Opportunistic Network

RSSI Received Signal Strength Indicator

RF Radio Frequency

SNA Social Network Analysis

US Ultrasonic

List of Mathematical Notations

G is a graph that represent a social network.
V is the set of *nodes* (or *vertices*).
E is the set of *links* (or *edges*) connecting a pair of nodes.
$N = |V|$ is the total number of nodes in the network.
A is an adjacency matrix with dimension N-by-N.
$a_{i,j}$ is an element in the adjacency matrix.
$D(k)$ is degree distribution of a network with N nodes.
r is assortativity coefficient between network nodes.
d_{jk} is joint probability distribution of the remain degree, this value is symmetric on the undirected graph.
$q(k)$ is distribution of the remaining degree.
σ_q is the standard deviation of q.
J_{ij} is a Jaccard coefficient .
E_{ij} is the total number of links (in or out) that i and j has a link to any third node (node i and j have a value 1 at the adjacency matrix).
E_i is the total number of links (out) that i has a link to any third node and j has not (node i has a value 1 at the adjacency matrix and node j has value 0).
E_j is the total number of links (out) that j has a link to any third node and i has not (node j has a value 1 at the adjacency matrix and node i has value 0).

$\delta_{ij}(e)$ is the proportion of shortest paths between i and j that pass an edge e.
$Sp_{ij}(e)$ is the number of shortest path that path by edge e.

Sp_{ij} is the total number shortest path between i, j.
$B(e)$ is the betweenness centrality of an edge e.
Q is the modularity that is a measurement for optimization methods for discovering community structure in networks. d_{S_j} is the distance from kp-set S to node j .
D^R is the proportion of all nodes reached by the set S.

P is a set of m participants where $p_i \in P$ and $1 \leq i \leq m$.

L is a set of K Locations where $L_k \in L$ and $1 \leq k \leq K$.

$[t_A^{p_iL_k}, t_B^{p_iL_k})$ A *participant's time interval* is a pair where $t_A^{p_iL_k}$ & $t_B^{p_iL_k}$ are times between which participant p_i is continuously detected at location L_k.

$Du_{t_st_e}(t_A^{p_iL_k}, t_B^{p_iL_k})$ is the *duration* that a participant p_i spends at Location L_k between t_s & t_e.

$f_{L_kp_i}(t_s, t_e)$ is *location function*; the presence (or not) of p_i at L_k between t_s & t_e where $t_s \leq t_e$.

$TV_{p_iL_k}((t_1, t_2), \cdots, (t_{n-1}, t_n))$ is the trajectory vector of participant p_i at location L_k.

$DV_{p_iL_k}((t_1, t_2), \cdots, (t_{n-1}, t_n))$ is the duration vector of participant p_i at location L_k.

$S_{p_ip_j}(t_s, t_e)$ is the trajectory opportunistic vector between participants p_i, p_j.

$\hat{S}_{p_ip_j}(t_s, t_e)$ is the duration opportunistic vector between participants p_i, p_j.

$Fr_{p_ip_j}$ is the frequency of meeting between participants p_i, p_j.

$S_C(Fr_{p_ip_j})$ is the co-located opportunistic vector between participants p_i, p_j.

p_s is the probability an encounter is not detected by a single device.

N_b is the number of encounters observed by both devices.

N_s is the number of encounter observed by single device.

N_m is the number missed (observed by neither).

$Ex[N_b]$ is the expected number of encounters observed by **both devices**.

$Ex[N_s]$ is the expected number of encounters detected by a single device.

$H_B = (V, E)$ is undirected network that presents different extracted opportunistic networks from the experimental study.

$Pr(A, B^*)$ is the precision between network A and network B^*.

$R(A, B^*)$ is the recall between network A and network B^*.

$F_1(A, B^*)$ is the harmonic mean between the precision and recall of networks A and B^*.

Chapter 1

Introduction

1.1 Motivation

Communication technologies have become an important part of our daily life and we use such technologies to help sustain our relationships and social networks. These communication technologies include devices such as mobile phones, and applications such as Facebook, Twitter and SMS. These technologies increase our ability to communicate with more people and can improve our social lives, increase our knowledge and can facilitate business activities. At the same time the use of technology opens up the opportunity to discover more about our social networks, potentially through implicit means, where our behaviour can be locally monitored and potentially used to detect who we are, to infer our relationship with others, to discover our position in the social network or to help deduce a local context. While acknowledging the important issue of privacy in making inferences from human mobility, this behaviour can be used as a fundamental input to pervasive computing technologies, for example to personalise the content on a display [3] or to provide messaging to others through opportunistic means [38].

Opportunistic networking is an important new technology because it uses the location proximity and interactions between people carrying devices to convey data between devices. It is related closely with human social networks because, in many cases, people interact because of the relationships that they hold with each other. Also, people may be willing to sacrifice resources such as storage and battery power to have communication with others because the personal relationship that they hold with other people means there is trust and goodwill. This aspect is often assumed to be the basis for routing algorithms [38] in opportunistic networks. However, the extent to which this is the case for a real world scenario is not well understood - to what extent would the social networks between humans and opportunistic networks coincide and what happens to data dissemination if different assumptions on the use of social networks are made?

This thesis investigates, in a general context, the relationship between a group's physical mobility and proximity and self-reported social network structure. It will allow us to determine the extent to which opportunistic networking can use the relationships that exist between humans who may carry the devices and what differences there are between the social networks that devices see and the networks that exist between humans. In order to achieve this, two main activities are involved:

1. Firstly, this research develops the capability to track individuals' physical proximity in a particular well-defined group, with their consent, in an indoor setting, gaining knowledge of visits to particular locations and their interaction with each other. A bespoke distributed system is developed to achieve this, using Bluetooth "motes" that are suitable to be carried by participants to collect the required data set. This system has been developed to detect the opportunistic network between the participants from their movements indoor as we can not create an opportunistic network.
2. Secondly, the overall social network of a large group is determined, within which the tracked participants are embedded and analysed. This is achieved using a survey methodology that determines the strength of relationships, the types of interaction and methods of communication that are used between peers.

From this data, diverse social network analysis techniques are employed, including the development of specific techniques, that allow us to determine the network structure that the opportunistic devices would see as compared to the human social network structure. In addition, the correlation between the opportunistic network and static social network of the participants is explored. Conclusions may then be drawn on data dissemination property in opportunistic networks could be affected by using different assumptions on social network properties.

1.2 Opportunistic Networks (OPNET)

An opportunistic network is a mobile network that is created by links between mobile and wireless enabled devices. The network is an infrastructure-less peer-to-peer network where there is no end-to-end connectivity, therefore a complete path between source and destination is not required. Instead, transient connected sub-networks exist, which may just be a pair of devices. Because of the temporal nature of links, data may be stored before there is the opportunity to forward it to another node. An important issue in this network is the mobility of its devices which determines the opportunities for forwarding data to other nodes. As a result, the mobility patterns for devices, for example how often pairs of devices see each other, govern the characteristics of the network and how it transfers data. The quality of service guarantee is a very big challenge as the network is dynamic in topology, infrastructure-less, and may have very short connection times.

A fundamental and interesting point about opportunistic networks is that the devices are in an overall network structure where the links can be thought of as familiarity relationships with other devices (for example, how often a pair of devices see each other). At the same time, the people carrying the devices have social and familiar relationships with the people they interact with. It is also convenient to assume that these two types of networks, opportunistic networks and human social networks, may be similar. However, this is not necessarily the case and it is aimed to investigate this similarity or difference further. This is important knowledge because it can help to inform the design of future protocols for information provision in opportunistic networks.

1.3 Thesis Hypothesis

The rate at which information can be disseminated is a key performance measure of opportunistic networks. Information provision (namely, sending data to a recipient) is challenging because of the lack of certainty over network topology. Flooding or epidemic routing offers a high delivery rate but at the expense of high message overheads as a basic mode of communication. More efficient and effective methods for forwarding in opportunistic networks minimise resource utilisation and duplication of data transmission through knowledge and understanding of the underlying characteristics of contact opportunities in human social networks. Many of sources of context can be used in order to forward messages to more useful peers, from raw data concerning mobility and individual interactions, to data which encapsulates the real-world social relationships of users. Our hypothesis in this thesis is:

> *Networks induced by physical interaction and user mobility provide a measurably more robust framework to predict contact opportunities compared to the human social network formed by " real-world relationship", leading to a quantifiable improvement in information dissemination in opportunistic networks.*

Providing evidence for this hypothesis in this thesis requires:

- The ability to detect real-world human social networks through data collection.
- The ability to detect user mobility both indoors/outdoors.
- The ability to identify the opportunistic network from user mobility and physical interaction.

1.4 Research Contributions

The research contributions in this thesis concern the detection of an opportunistic network through local interactions and the comparison of this with the human social network that the owners of the devices have between them. To determine the opportunistic network, a significant pervasive mobility tracking system has been developed to track the participants' movement inside a large building and to log interactions when outside this building (as far as devices will permit). As this is a large case study for this research, the developed system has been applied with first year undergraduate students in 2011/2012 in the School of Computer Science and Informatics. The system supports the functionality of recording the physical interaction of the participants between each other. Different algorithms have been developed to infer the opportunistic network based on automatic data collection between participants. These inferences let us know information about the social structure between opportunistic devices and the chances for opportunistic dissemination.

To complement this, the social network between the human participants was assessed as well as the extent to which it can be used as a basis for information dissemination. A survey methodology has been used on the same large sample and has assessed their relationships using multiple dimensions. This is an interesting study in its own right because it demonstrates dense clustering that binds the student community together, including aspects such as strength of ties, use of relationships to gain information from others and the different types of technologies used. We are then able to compare the different types of networks, opportunistic and social, to understand their different characteristics and similarities. This extends to how they can perform information dissemination between participants. This is seen in Figure 1.1.

1.5 Research Methodology

The approach in this research has two carefully designed main components. Firstly, the social network between participants is explored by using a survey study and secondly, the movement and physical interaction between participants is tracked by developing a bespoke pervasive computing system. The Tracking system allows us to detect the opportunistic network. This is composed of two parts: devices detecting each other and devices being detected by servers in the building. These issues are described below.

Discovering the social relationships: A comprehensive survey study has been conducted with participants in order to state their self-reported points of view concerning sharing information with each other and maintaining relationships. From this information, different social networks

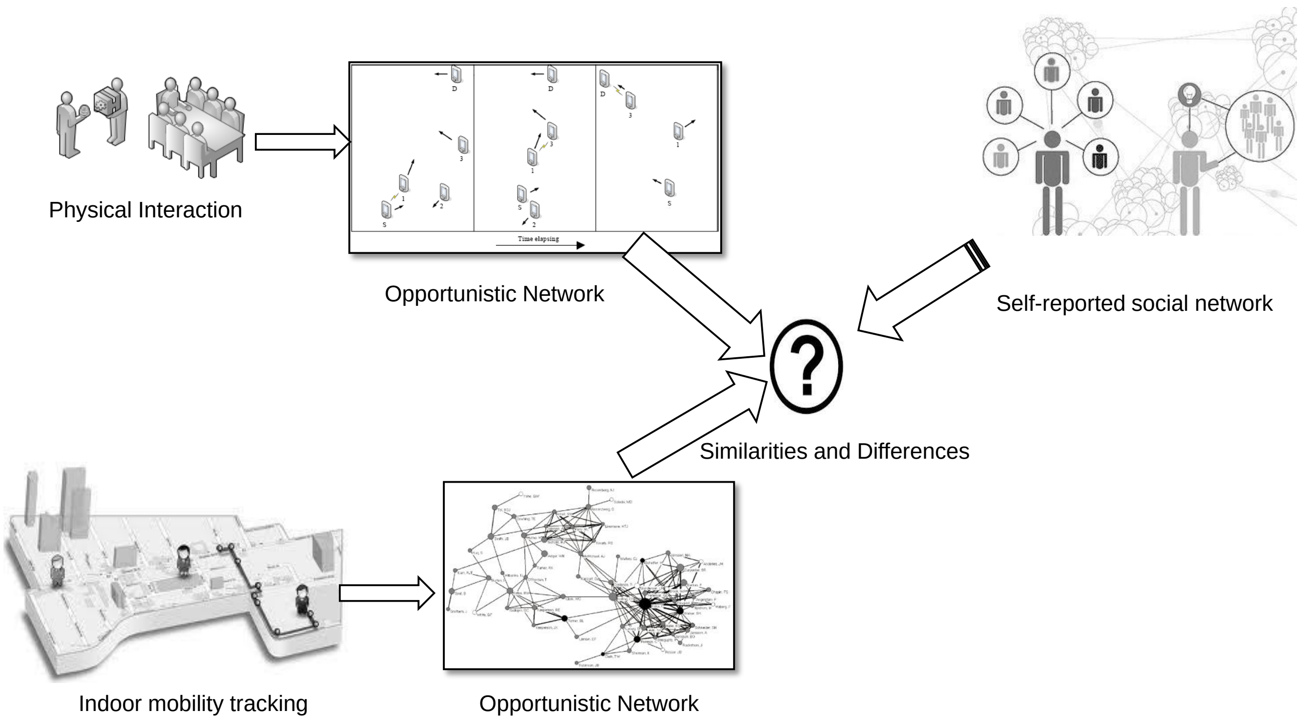


Figure 1.1: Thesis Contribution.

have been built and behaviour and coincidence analysed. These allow us to discover how different people communicate with each other by exploring different communication methodologies, different frequencies of communication and the strength of relationships.

Detecting the opportunistic network through device interactions: Here, we are concerned with the physical interactions between participants and the frequency of interactions. These not only potentially capture people's relationships but also contain noise in the form of random behaviour and coincidence. Different pervasive computing technologies have been employed in indoor positioning and tracking [7, 45, 5]. In this study, Bluetooth technology has been used for participants' indoor positioning and tracking by Bluetooth mote devices.

Detecting the opportunistic network through tracking participants' mobility inside the building: We are interested in the mobility trajectories of the participants that occur inside the building during day time. These trajectories allow us to know different participants' behaviour patterns inside the building. An indoor tracking system has been developed that achieves specific requirements and functionality. This system has been tested thoroughly. Its advantages include robustness, with no commercial product available that performs the same functionality, and it is specifically designed through this research. Figure 1.2 shows the flow of the methodology that has been followed through this research.

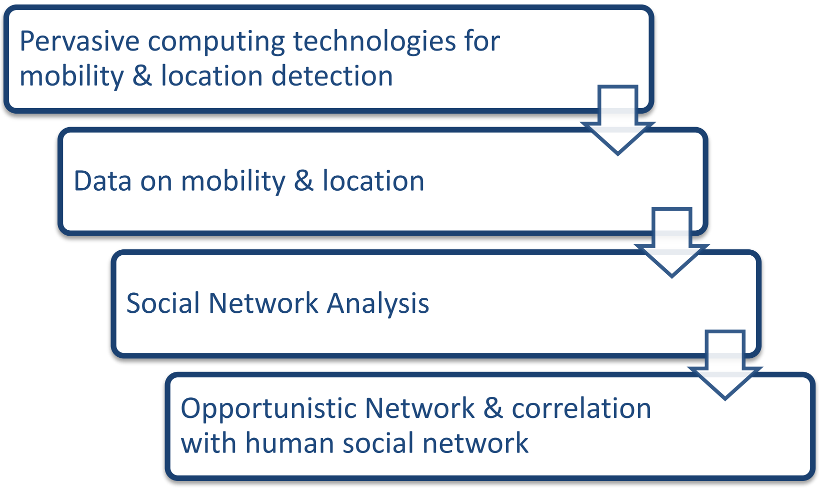


Figure 1.2: Methodology Flow.

1.6 Thesis Overview

The thesis is structured as follows:

Chapter One introduces the research motivation, research contribution and research methodology.

Chapter Two outlines the basic concepts of social networks and the metrics that are used for analysis. It overviews different technologies and techniques for indoor mobility tracking. The opportunistic network information provision and its application is also discussed.

Chapter Three focuses on the acquisition of mobility tracking and physical interaction in order to build up the participants' opportunistic network. It includes all the steps that have been achieved to build a system able to automatically collect information on the mobility tracking and physical interaction between participants. This chapter also emphasises the difficulty faced when convincing the participants to participate in this practical stage of the research.

Chapter Four explores the self-reported data set collection by a survey study. This study focuses on information sharing between participants to understand the hidden relationship and extract the social network between them from their perceptive.

Chapter Five explains different analysis techniques to uncover the characteristics of the extracted opportunistic networks. This chapter shows three different types of extracted opportunistic networks by using different algorithms of extraction. In addition, it analyses the resulted network statistically and concludes with interesting key findings.

Chapter Six overviews the different analytical and visualization methods to profile the self-reported social network. It also focuses on the extraction of the key-user of the social groups. It sums up with the key points of the analysis.

Chapter Seven discusses the evaluation of the extracted social network. This includes the use of precision and recall concept in order to understand and measure the relevance between the social network and the co-located and trajectory opportunistic networks. Additionally, the similarities and differences between the opportunistic and social network are extracted using the dissemination properties in order to show how relationships can affect the flow of opportunistic networks.

Chapter Eight highlights the key features of the research, evaluates the achievements against the aims and concludes with an appraisal of the overall research experience and outcomes. It concludes by suggesting future research which could be carried out based on this research.

Chapter 2

Background & Literature Review

Overview

The development of communication technology motivates different means for information to be transferred to the destination. An opportunistic network is an infrastructure-less network that is established between mobile and wireless devices. Section 2.1 demonstrates a brief summary of information provision techniques in opportunistic networks and its applications. An opportunistic network cannot be easily created, so it is necessary to determine the network that the device can locally see. Due to the advances in pervasive computing technologies, it is feasible now to detect the indoor location of people. Related work is described in Section 2.2 that discusses mobility tracking using different pervasive computing technologies such as WiFi and Bluetooth. Recently, location and mobility data has been used to extract social networks between people as discussed in Section 2.3. From the extracted links between people, relationship inferences have been proposed. Different analysis techniques have been demonstrated to support the analysis of the questionnaire data set in Section 2.4. To conclude, Section 2.5 outlines different examples that have used social network analysis on different types of data.

2.1 Information Provision in Opportunistic Networks

Information provision (namely, sending data to a recipient) is challenging because of a lack of certainty over network topology. As a result, broadcasting is often studied as a basic mode of communication. A node sends information to all its neighbours. This algorithm suffers from the storm problem [74] as the node is working as both host and router at the same time. Multicast is another type of communication where the node sends information to a known group of nodes simultaneously. This method is more organized than broadcasting. Another method is called Unicast, where the information has been sent to a single node in the network.

Different ways of delivering information are proposed that include Flooding, Epidemic and

Context Aware Routing [12]. D.Agrawal [2] introduced the Epidemic in order to maintain the replicated data in a transaction database between servers. The Epidemic selects a neighbour on a probabilistic basis to forward information instead of forwarding information to all neighbours as in broadcasting. Also, this technique has been utilized to reduce the duplication problem in Flooding techniques [95]. In addition, Glance [42] who used this a technique to disseminate information by taking into account people's movements. Another approach in information spreading in the opportunistic is Context Aware Routing which monopolizes the user or node information in information provision. A node that preserves a local database is required in this approach to be able to forward information to the next node.

2.1.1 Flooding Based Routing

Flooding technique is the simplest approach in information dissemination where it broadcasts a received message or information to all known neighbours whenever possible. There is no consideration taken into account before forwarding the message to any node. A broadcast storm [74] happens when a large number of nodes are frequently sending the same information.

This approach is effective when the path between source and destination is unknown as the network topology is always changing. In this approach, nodes continue exchanging information without any consideration at any meeting. This approach suffers from information duplication but it is nevertheless very quick in sending information when the network topology is unknown.

2.1.2 Epidemic Based Routing

Initially this approach was proposed as an amendment of flooding based techniques [95]. This approach reduces the amount of duplicated information in order to improve the flooding technique. A summary vector is created by the sender to be sent to all known neighbours. This vector includes a summary of received information. By exchanging this summary vector when the nodes are connected, each node is able to determine which information has not been received from other nodes. Once unseen information is detected, the node can request this information from another.

In order to implement epidemic routing, a unique code is attached to each message in addition to the source and destination information. This is to enable each node to recognize unseen messages easily. In addition, by avoiding exchanging information to the same node, it helps to decrease the amount of exchanged and updated information between them, especially if they see each other frequently. This approach has been used in sensor networks to disseminate information in [10], and for information collection from sensors through zebra interaction. Juang [55] uses this approach.

2.1.3 Context-Based Routing

Context-based routing utilizes the user information (preferences, location and address) for routing purposes. This approach requires each node to have enough space to store information and exchange it with seen nodes. Additionally, there is the need to keep up-to-date with the availability of its neighbours when transferring information to others.

A history of social interaction and community information for each node is utilized to support information exchange in HiBOp [12]. This information is transferred to other nodes during their interaction and this in turn enables a node to understand its context environment that instructs the node to be a good transmitter. In principle, two tables are required to be updated by the node and these are the History Table (including the current context) and the Identify Table (includes personal information).

Mainly, all information provision techniques that are used in wireless communication are based on two concepts, namely push and query. Push authorizes the node to spread information to other nodes in the network. Query authorizes the node to request information from other nodes in the networks. This thesis does not focus on forming the opportunistic between the participants, but rather it aims to present what it looks like between the participants as well as find the correlation between the detected opportunistic and the social networks between the participants and their perceptive.

2.1.4 Opportunistic Network Applications

Initially, the concept of opportunistic networks was designed to assist military communications and recently it has been successfully deployed in several civil environments where users are interested in gaining information that is relevant to their interests. Opportunistic networks are the alternative method of communication where installing network infrastructure is not feasible (such as in a deep forest, in the ocean or in a rural area).

Animal movement has been monitored by using opportunistic network concepts where SWIM [87] and ZebRaNet [55] is an example of using opportunistic networks to study the behaviour of animal mobility. It has been observed that it is possible to connect people however continuous connection is not available from the DarkNet project [80]. This project enabled internet access to a small village outside New Delhi.

In this thesis, flooding performance is used as the benchmark for addressing the hypothesis of this study. This technique is used to find the similarities and differences between the mobility and self-reported social networks that are demonstrated in Chapter 3 and 4.

2.2 Indoor Mobility Tracking

To create an opportunistic network it is necessary to detect the opportunities that wireless devices have to locally share information. As a result, it is crucial to understand how to track the mobility and co-location of the devices. In addition, no opportunistic network platforms are available other than the prototype developed in the Haggle project [88]. Therefore, a different approach has been taken and this looks at detecting opportunities for peer-to-peer transmissions. This section discusses the meaning of indoor mobility tracking and applications that can utilize tracking inside buildings. Indoor mobility tracking in this study refers to the aggregation of the individual's trajectories inside the building in the same order in which they occur in real life. It has been found that mobility tracking can be successfully used to support different applications, such as monitoring elderly people, detecting the location of people in crisis, smart homes, information dissemination, and monitoring employees inside organizations.

For real-time indoor tracking, a physical sensing mechanism is required to measure the position of humans [69]. Furthermore, it is necessary to repeatedly apply such techniques so that the progression of an individual through the environment can be established. In order to measure human position, various techniques have been developed using different sensors that can be easily carried. These include: Infrared (or IR) [97], Ultrasonic (or US) [45, 82, 7, 51], Radio Frequency Identification (RFID) [57, 44], Bluetooth [103] and WiFi [37, 81].

In this research, it is important to distinguish between position and location [93]. *Position* refers to the exact position of the object that can be provided as latitude and longitude whereas *location* refers to a general area from which specific deductions can be made, such as extracting the location of that object, for example stating "the second room in the north building second floor". *Position* is considered more accurate than *location* but some systems require determining the location of the users to achieve their purposes. In order to achieve an accurate *position*, it requires expensive hardware and software. Knowledge of mobile human/object location is crucial information for monitoring elderly people, the development of geographic routing, and tracking mobile objects through networks systems. There are three fundamental approaches to determine location that are based on physical layer processes [15, 90, 98] that use wireless transmission in different ways.

1. **Received Signal Strength Indicator (RSSI)**: This approach utilizes the inverse propor-

tional relationship between received signal strength and the distance from the receiving station which is linearly represented. Unfortunately, it is impossible to accurately position objects due to reflection and multipath [93]. However, it can still be used to localize objects [98] although it is susceptible to the impediment of radio waves by the local environment.

2. **Angle of Arrival (AOA)**: This approach calculates the angle of arrival of the signal from the mobile object to more than one receiver that operates with antennas. The distance is calculated using simple geometric rules on the angle measurement [93]. This approach has strong limitations for indoor application due to the reflection of the signal on walls, clutter and multipath effects.

3. **Time of Arrival (TOA)**: This approach is based on the time of arrival of a signal, and is transmitted by the mobile object and received by a minimum of three receivers using a triangulation algorithm. This technique is not supported for positioning when using technologies such as Bluetooth as the measure time delay is very small [93, 98] but can be used for cellular mobile communications.

In summary, Table 2.1 presents different systems and related works that have been proposed for mobility tracking by using different wireless technologies such as WiFi, Bluetooth, InfraRed, and WLAN. Some systems strive for highest accuracy whilst others seek only localization within the room level. The previous approaches, such as the physical layer signal processing techniques, are considered the most popular as they are based on detailed information. In many cases, these techniques can be used by engaging specialist equipment. In a limited number of other cases, they are embedded in the function of standardised wireless technologies. In this research, it is necessary to adapt the technology rather than study it. This research aims for for localization within room level accuracy in order to study human behaviour and relationships inside the buildings.

Table 2.1: Positioning and Localization Selective Related Works.

Author/System/year	Sensor	Purpose	Method	Accuracy & Evaluation
Active Badge (1992)	Infra-red	localization	RSS	Low accuracy (room level), affected by fluorescent light and sunlight.
Bahl (RADAR,2000)	WLAN	Localization	Triangulation	Low level of accuracy, used the existing WLAN infrastructure.
Active Bat	Ultrasound	Positioning	Multi-lateration	Very good accuracy within large covered area, require large number of transmitters on the ceiling.
Cricket (2000)	Ultrasound,RF	Positioning	TOA & triangulation	Very good accuracy, address privacy, low cost, more energy consumption.
Bolliger (2008)	GSM, Bluetooth & WiFi	Localization	Fingerprint	Improve the accuracy by using multiple sensors.
Filho (2008)	Bluetooth	Localization	Response of inquiry	Low level of accuracy.
Genco (2005)	Bluetooth	Positioning	Link quality	It requires specific hardware to measure link quality.
Hay & Harle (2009)	Bluetooth	localization/tracking	Proximity connection-base	
WIFE (2009)	WiFi	Positioning	fingerprint	

2.2.1 Indoor Localization Systems

By knowing the technologies' techniques and software that can be used to perform mobility tracking, localization is considered as the first step in mobility tracking in regards to a moving object. As seen in Figure 2.1, this step is required to provide the required data set. In tracking a moving object, firstly the location of this object is determined, then the changes of its location during the time are continuously aggregated. Localization refers to the determination of the location of an object at a single point in time. This study focuses on the determination of the location of an object inside buildings that are known as *Indoor Localization*.

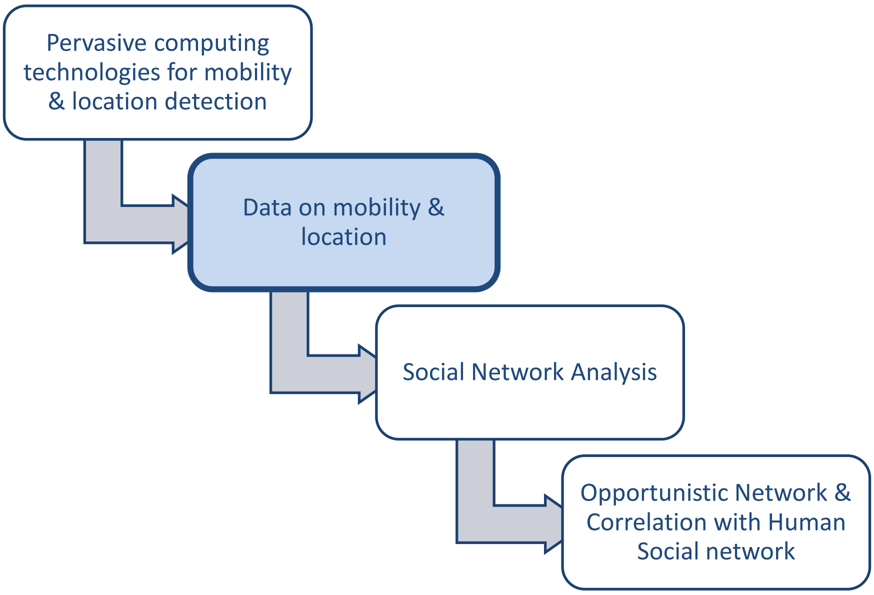


Figure 2.1: The Second Step of Methodology Flow.

As indoor localization has become a hot topic of research, researchers have developed a large number of indoor localization systems that are based on standardised technologies, such as radio communication in the ISM band. The advantages of these systems are that they are easily available and can be readily adapted for a scenario. There are some localization techniques that use a visual means to recognise human faces [60, 84] and the most used approaches involve detection of devices carried by humans using wireless/sensing technologies.

For instance, the Active Badge system [97] uses IR (Infra Red) technology in a badge that emits an IR signal. This signal is received by the IR receiver that has been placed in each room. As a result the system accuracy is a room level (3-5m). RADAR [7] and SpotON [51] use RF (Radio Frequency) signal strength for indoor localization. Intel and Washington University [44] and Utah State University [57] use RFID tags on reference points of the building to enable a user carrying an RFID reader to find a desired route.

Ultrasonic signal has been used in the Active Bat project [45] as an inexpensive alternative. Active Bat is an active ultrasonic transmitter carried by the user during its movement inside the building. A short RF message is sent from the Bat as a pulse that triggers multiple listening devices placed on the ceiling. Each listener utilizes the arrival time of the pulse to compute the distance between the Active Bat and itself. Then the resulting distance is forwarded to a host computer that localizes the Active Bat with an approximation to within 10 centimetres. Another system utilizing ultrasonic communication is called Cricket [82], but Cricket reverses the configuration (namely the transmitter on ceiling and the listeners by the users). This has been done to respect user privacy. Additionally, in the Cricket user's device, it can locally calculate its location information instead of this being sent to a host computer as in Active Bat.

It is important to note that all of the previous systems are single data sources, based on one type of sensor for data collection. Multiple data sources have been used in other systems to improve the accuracy of localization. A system based on RFID, WiFi and Vision is presented in [5]. In [77] a localization system has been proposed based on Bluetooth and WLAN. In [6] indoor location estimation based on Bluetooth and WLAN is presented and evaluated using RSS from WiFi.

Bluetooth technology has become very popular in indoor localization systems [17, 4, 9, 14, 70] and it is likely to be the most common. For example, [9] utilizes the response rate of Bluetooth inquiries to localize people indoors. This is due to numerous properties including low power technology and almost all deployed handsets support it. It is also easy to engage local desktop machines in Bluetooth communication by adding hardware such as a dongle. Bluetooth indoor localisation systems all locate mobile device by continuously inquiring packets from a network of fixed machines often called beacons. This technique has advantages as there is no low level code required or a Bluetooth connection needed as in [46]. It can be referred to as a connection-free method as it does not require any connection between devices to detect each other. This connection free method avoids security issues that arise when a connection is established, however, the mobile device is still under risk of being tracked by anyone when it is in discoverable mode. Additionally, most of handset companies increase security by limiting the time of the handset being on discoverable mode. This increases the challenge of using Bluetooth in indoor localization and tracking.

2.2.2 Location-based Services

Social location-based services are very recent and have been enabled by the combined features of the *smart phone* including GPS, WiFi, Bluetooth, internet connection and mobile software applications. These features can be used to mix the place of the user and his/her social network together and enable the study of users' physical behaviours by continuously tracking their location. Previously, it was conflicting to study and relate the behaviour of physical social network with the behaviour of online social network. Until recently, researchers found conflicting results while comparing the two worlds together. Some studies argue that online social networks contribute in isolating people in physical world [20].

The growing ubiquity of location-enabled *smart phone* reduces the difference between online and offline social networks [20]. *Foursquare* is one of the most mobile applications that supports new means for online interaction based on the physical location of their users. Although applications such as *Foursquare* have unique properties that support different research, it always requires human interaction (the user must check-in for every new location he/she has moved to). This will affect the accuracy of the research as the user may intentionally or unintentionally check-in at certain places.

2.3 Analysing Data through Social Network Structures

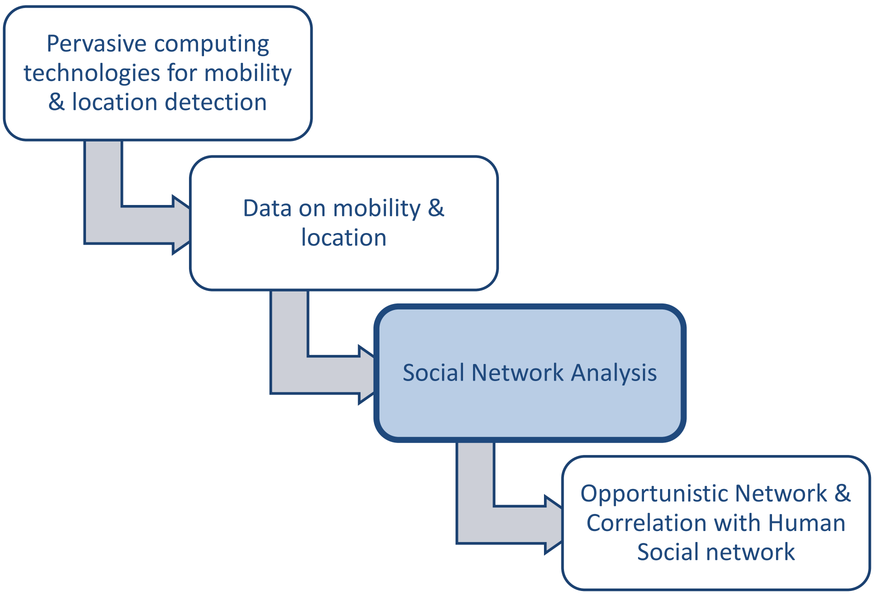


Figure 2.2: The Third Step of Methodology Flow.

2.3.1 Extract Social Network Structure

Aron et al.[22] is the first paper that operates an end-to-end system that integrates information extraction and social network analysis together. The system builds up the social network links by extracting mentioned people from web pages and creating a link between the page owner and the extracted person. A recursive extraction and process builds larger networks that contain "friends of friends of friends". It has been noticed that Bluetooth proximity data can be used for building a social network structure of the users [27].

An analysis of a Bluetooth proximity network has been presented in [19]. Pairwise links between people have been defined in [29] by using Bluetooth and phone call data instead of using questionnaire-based self reported data. In [62] an algorithm proposed group discovery that focused on fully connected sub networks of the Bluetooth proximity network. Some drawbacks emerging from this method are the noise effects and group size that grows gradually. Some researchers focus on the face-to-face aspect in order to overcome Bluetooth drawbacks by using specific mobile devices [41]. Using this method, the actual physical interaction has been sensed but it requires people to carry an additional device.

As every mobile phone is now equipped with Bluetooth, the research in individual/group behaviour mining [102, 43] has received more attention recently. The Reality Mining project utilizes data from the mobile phone to observe and characterize the social behaviour of individual users and organizations [30]. Roue et al [85] extracted a social hierarchy by analysing the flow of emails within organizations. They focused on the users' behaviour patterns to determine the strength of communication links between users. Corresponding to the relation between communication media and relationship strength between users, it was found that strong links (strong friendships or close friends) communicate more frequently using more media than weaker links [48].

To sum up, most of the recent research uses mobile phone data to study users' mobility pattern and internet data such as emails, Facebook to learn about the users' physical social network. However, they do not explore indoor mobility datasets to understand users' mobility pattern which are challenging. These relationships can be used for information dissemination. Another limitation with previous research is that they do not record and utilize the physical interaction to understand the opportunistic networks between users and analyse their structure. These challenges have been overcome through this research.

2.3.2 Infer Relationship Strength

Social network structures are important in order to analyse and produce some inferences from the collected mobility data. As seen in Figure 2.2, it is the third step in the research methodology of this study. Social network analysis can be applied to different types of data, such as email, phone logging and proximity, since the data represents interaction between people. Onnela et al [76] analyse the structure and tie strength of social and communication networks by utilizing the recorded calls of mobile phones. The results from the social network analysis provide social information that represents the frequency of interaction between people and the pattern of communication between them.

Humans are social by nature and social interaction is a common activity in daily lives [33]. Generally, there is always a link between people and each other, and this link is the evidence to show how people communicate in different ways. Recently, research has focused on identifying these kinds of links from peoples' mobility patterns. This is to explore the extent to which new communication technology can maintain, expand or decrease existing relationships. In addition, it explores how these relationships differ or support the physical interaction relationship (face-to-face).

One of the challenging problems in this area is inferring the characteristics of users' social behaviour from their location. Eagle et al [32, 31] and Li et al [58] develop and utilize the measure of users' similarity to infer the social structure between users. This inference is very challenging where co-location users are defined as being in the same place at the same time, but there is no evidence of having a relationship between them. Cranshaw et al [20] introduce a set of features that focus on the social context of the locations that users visit. Then they evaluate these features by predicting both whether two co-located users are friends on Facebook, as well as the number of friends each has in the social network.

Dong et al [28] propose to explore individual friendship patterns from cellular phone call logs. On the other hand, Gho et al [18] study the relation between human geographic movements, temporal dynamics, and social network ties. They analyse the effect of geography on daily routine and social ties on the human mobility pattern. Tang et.al [92] infer social ties from heterogeneous networks and find that 70% of the online social networks (such as Facebook, Twitter, and LinkedIn) have not been well labelled.

Yu [101] used sensing data that was captured by the MIT Reality Mining project to explore the social relationships of evolution. Firstly, they define friendship as a directed link. Secondly, they recognize human friendship from a supervised learning perspective. They adapted an inference model that is able to predict friendship from a variety of features extracted from mobile phone data, including proximity, outgoing calls, outgoing text messages, incoming calls, and incoming text messages. This approach is called Support Vector Machine (SVM). Finally, they use social balance theory to demonstrate the social relation evolution.

In another study focusing on on-line communication methodologies, it was found that strong links use email and instant message more frequently than weak links [47]. The discovered relationship between mobile users can be used for information dissemination in different network environments [52]. [53] proposed a novel paradigm that focuses on behaviour-driven communication, enabling a new class of services in mobile societies.

2.4 Analysis Techniques

In this section, it is important to introduce some selective social network analysis metrics as they have been used throughout the thesis. In general, a social network can be represented by several means. In this study, it is presented in two presentations by a graph and the associated adjacency matrix. A graph G can be defined as (V, E) where V is the set of *nodes* (or *vertices*) and E is the set of *links* (or *edges*) connecting a pair of nodes. A graph be weighted or unweighted - where the weight is a value assigned to the link. A graph can be directed or undirected where the link between two nodes is either symmetric or not symmetric. A social network or a graph can be represented as adjacency matrix A where A is a square matrix N-by-N and $N = |V|$. Each element in matrix $a_{i,j}$ has a value if there exists an edge from node i to node j. What follows is a brief summary of the analysis techniques that have been applied in this thesis. These concern: small-world, degree distribution, shortest path length, assortativity, clustering analysis (hierarchical clustering and Griven and Newman), and key-player analysis.

Small-world Concept

A network theory of "Small Worlds" shows that any network is only stable if there is a logarithmic relationship between the number of nodes in the connection between any two nodes (namely source and destination). The telephone network is considered as a useful example of the small-world concept where two persons calling each other might require a number of steps to complete a call.

The *small-world* concept originated with Stanley Milgram [65] in 1967 with an experimental study that tracked a chain of acquaintances in the United States. The experiment result shows that a packet could be delivered through a path between acquaintances. The average path length was six and this is known as *six degree of separation*.

Degree Distribution

The degree of a node is the number of edges that the node has to other nodes in the network. In a directed network, there are two types of node degree and these are in-degree and out-degree, where in-degree is the number of incoming edges and out-degree is the number of outgoing edges. The degree distribution $D(k)$ of a network with N nodes is the fraction of nodes with the degree k where $D(k) = N_k/N$. N_k = number of node with degree k.

Some networks have a distribution that follows *power law* distribution where these networks are called *scale-free* networks. These networks include a low degree for a majority of their nodes.

Other networks have *random* distribution that follow Gaussian distribution and these are called *random* networks. These networks contain a median degree for a majority of nodes.

Shortest Path

Shortest path aims to find the paths defined with the lowest cost between a node and other nodes in the network. This technique has been used in extracting the characteristics of communication technologies and relationship strength and their effects on social network structure. Dijkstra's algorithm [24] has been used to measure the shortest path length between network pairs.

The shortest path length algorithm can be used for many applications, such as maps [86], robot navigation, texture mapping, urban traffic planning [39], and network routing protocol [68]. In this study, this algorithm has been used to study the connectivity characteristics of different networks between different vertices. In this study, shortest path length represents the number of hops as there is no cost between nodes.

Assortativity Coefficient

Assortativity is a measure of similarity between a network's nodes in terms of the nodes' degree. In social networks, the correlation between nodes of similar degree (namely highly connected nodes connected to other high degree nodes) can be observed in many networks. The assortativity coefficient is essentially the Pearson correlation coefficient of the degree between pairs of linked nodes [71]. The value of the assortative coefficient lies between $[-1, 1]$, where a positive value indicates a correlation between nodes of similar degrees while negative values indicate relationships between nodes of different degrees. Assortativity properties are useful in understanding the spread of disease or cure [66]. In this study, assortativity is used to understand the spread of information inside the network.

$$r = \sum_{jk} jk(d_{jk} - q_{out}(j)q_{in}(k))/\sigma_{q_{in}}\sigma_{q_{out}} \tag{2.1}$$

Where

d_{jk} is joint probability distribution of the remain degree, this value is symmetric on the undirected graph.

$q(k) = \frac{(k+1)p_{k+1}}{\sum_j jp_j}$ is distribution of the remaining degree.

σ_q is the standard deviation of q.

Hierarchical Clustering

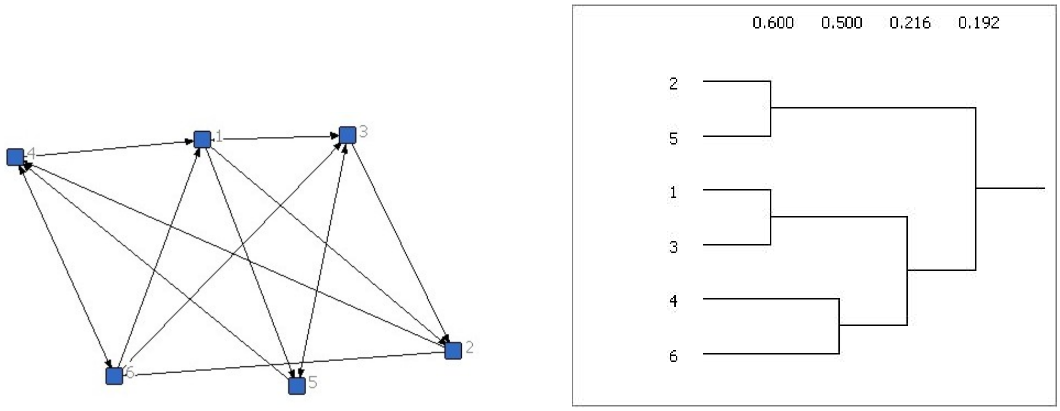
The basic idea behind the hierarchical clustering approach is to calculate a measure of similarity between node i and node j depending on the given network structure as seen in Algorithm 1. Hierarchical clustering is commonly created by agglomerative methods, which are proceeded by a sequence of aggregations of subsets of nodes into clusters. An alternative approach is to use a divisive method, which progressively separates all nodes by successively dividing subsets of nodes into clusters. In this research, the well-used agglomerative method of Johnson is adopted [54].

In implementing the hierarchical clustering approach, the Jaccard coefficient was used to measure the similarities (weights) between nodes [54]. As the adjacency matrix of the network is represented in binary, Jaccard coefficient is a useful measure for determining the similarities of binary data. This coefficient can be calculated by counting the number of times that both node i and node j has a link to node k divided by the total number of links.

$$J_{ij} = E_{ij}/(E_i + E_j + E_{ij}) \tag{2.2}$$

where
E_{ij}= the total number of links (in or out) that i and j has a link to any third node (node i and j have a value 1 at the adjacency matrix)
E_i= the total number of links (out) that i has a link to any third node and j has not (node i has a value 1 at the adjacency matrix and node j has value 0)
E_j= the total number of links (out) that j has a link to any third node and i has not (node j has a value 1 at the adjacency matrix and node i has value 0)

An example of a hierarchical clustering is given via the dendrogram as seen in Figure 2.3. There are six nodes, each of which are their own cluster at the start of the clustering algorithm. Then the Jaccard coefficient is calculated for each pair in the network as a measure of similarity between nodes (this will be discussed below). In Figure 2.3(a) we can see that node 1 and node 3 have three common links (links to node 2 and node 5 and link from node 6). Also, they have two uncommon links (node 1 receives from node 4 and node 3 receives from node 5). The similarity measure can be calculated as follows: $J_{3,1}$ = $3/(3 + 1 + 1)$ = 0.6 (as seen in Figure 2.3(b)) Then the highest ratio is chosen that represents the most similar pair and they are merged together. In this example, we can see that node 1 and node 3 are combined together and node 2 and node 5 are merged together in the same iteration as they have the same similarity ratio (i.e 0.60). After this, the Jaccard coefficient is recalculated between new clusters and other nodes in the network and there is repeat merging until there is one cluster.



(a) Social Network (b) crossponding hierarchical clustering

Figure 2.3: An Example of Simple Hierarchical Clustering.

Algorithm 1 Hierarchical Clustering

Input $Graph$ of the social network (adjacency matrix)

Output A hierarchy of clusters;

Let each node be a cluster;
Compute proximity matrix (the weights between all pairs);
repeat
    Merge two "closest" clusters based on a certain criterion (weights);
    Update proximity matrix;
until only one cluster remains

Girvan and Newman Clustering

As Girvan and Newman [73] state, hierarchical clustering has weakness in some cases. Their approach is also adopted in detecting a community. Girvan and Newman focus on constructing a measure to indicate the edges which are least central to the cluster and they remove them. This divisive technique is repeatedly applied. The algorithm is displayed in Algorithm 2.

Betweenness centrality of an edge assesses the centrality of an edge in a network. It is calculated by summing up the number of shortest paths (geodesics) that pass the observed edge. The edge betweenness of an edge e in a graph $G(V, E)$ is defined as the number of shortest paths along it. The proportion of shortest paths between i and j that pass an edge e is defined as [35]

$$\delta_{ij}(e) = \frac{Sp_{ij}(e)}{Sp_{ij}} \tag{2.3}$$

Where

$Sp_{ij}(e)$ is the number of shortest path that path by edge e.

Algorithm 2 Edge-betweenness Clustering (Girvan and Newman)

Input $Graph$ of the social network (adjacency matrix)

Output Cluster Model (Dendrogram)

```
repeat
    Compute edge betweenness for all edges;
    Remove edge with highest betweenness;
until no more edges in graph
return (Dendrogram);
```

Sp_{ij} is the total number shortest path between i, j.

The betweenness centrality $B(e)$ of an edge e is

$$B(e) = \sum_{i \in V} \sum_{j \in V} \delta_{ij}(e) \tag{2.4}$$

Modularity

To validate the quality of clustering, modularity is used as an external measure [73][72]. Modularity compares the number of edges inside a cluster with the expected number of edges that one would find in the cluster if the network were a random network with the same number of nodes and where each node keeps its degree but edges are randomly connected. It provides no guide as to how many communities a network should split into but is a useful measure on the quality of a division of a network into clusters.

Let us assume a network consists of N nodes connected by $m = |(E)|$ edges and let A_{ij} be the number of directed edges between nodes i and j (adjacency matrix). This means $A_{ij} = 2$ when node i communicates with node j via two different methodologies (Face-to-face and Facebook). Let us also assume the vertices are divided into some number of groups. The modularity of this division is defined to be the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random. In this case, the expected number of edges falling between two vertices i and j following randomization is $k_i k_j / 2m$ where k_i is the degree of vertex i, and hence the actual less expected number of edges between the same two vertices is $A_{ij} - k_i k_j / 2m$. Summing up all pairs of vertices in the same group, the modularity, denoted Q, is then given by

$$Q = 1/2m \sum_{ij} [A_{ij} - k_i k_j / 2m] \delta(c_i, c_j) \tag{2.5}$$

where c_i is the group to which vertex i belongs and $\delta(c_i, c_j)$ is the Kronecker delta symbol. The value of the modularity lies in the range [-1,1]. It is positive if the number of edges within groups exceeds the number expected on the basis of chance. The best clustering configuration occurs when modularity measure is high [73]. The most common use of the modularity is as a measurement for optimization methods for discovering community structure in networks.

Key-players of The Network

There are applicable concepts for the key-players; one of them is the Dominating Sets. Dominating Set is a set of nodes whose members are adjacent to all other nodes in the graph. A new measurement of centrality; m-reach centrality, that counts the number of nodes within distance m of a given node, is used to get the key-players set in a given social network. The advantage of this measure is its face validity. The disadvantage is that it assumes that all paths of different lengths are equally important (where in fact a path of length 1 is more important than a path of length 2).

A more sensitive measure named distance-weighted reach is used to overcome this disadvantage. It can be defined as the sum of the reciprocals of distances from the Key-player set S to all nodes (see equation 1). The Key-players take the minimum distance as appropriate since the fact that the distance to an outside node might be large for a given member of the set will usually be irrelevant.

$$D^R = \sum_j (1/d_{s_j})/n \tag{2.6}$$

In the equation, the distance from kp-set S to node j is indicated by d_{S_j}. In addition, it should be noted that the summation is across all nodes and the distance from the set to a node within the set is defined to be 1. As before, the reciprocal of an infinite distance is defined to be 0. Taking some interpretive license, we can view D^R as the proportion of all nodes reached by the set, where nodes are weighted by their distance from the set and only nodes at distance 1 are given full weight. Hence, D^R achieves a maximum value of 1 when every outside node is adjacent to at least one member of the kp-set (the kp-set is a dominating set). The minimum value of 0 is achieved when every node is an isolate.

The algorithm started by selecting the node with the highest D^R score. Then, for each of the remaining $k - 1$ selections, the node with the highest score is chosen that is not adjacent to any of the nodes already selected. This algorithm, a variant of the first fit decreasing algorithm for the bin-packing problem, is fast and easy, but often yields solutions that are considerably less than optimal. The extent to which the community is susceptible to information flow from informed subgroups was also considered. This is of particular relevance to communities that are exposed to ambient information (such as display screens or notice boards). Using key-player analysis, the difference that the structure has on dissemination was explored. Introduced in [13]

Algorithm 3 Key-players Greedy Optimization

Require: $Graph$ of the social network (adjacency matrix)

Select k nodes at random to populate set S
Set $F =$ fit using appropriate key-player metric
for each node u in S and each node v not in S **do**
    DELTAF = improvement in fit if u and v were swapped
    Select pair with largest DELTAF;
    if DELTAF $<=$ **then**
        terminate
    else
        swap pair with greatest improvement in fit
        Set $F = F + DELTAF$
    end if
end for

key-player-analysis allows for the trade-off between the number of informed participants and minimum dissemination path length for global information spread to be determined [8].

2.5 Examples of SNA from Mobility Tracking and Other Data

This section introduces different examples of how social network analysis has been applied in a number of different studies. Social networks provide a structure for information flow. Consequently, much of the literature in this field is relevant and some of the most interesting techniques have been highlighted in which social network analysis has been adopted and exploited.

Considering physical social networks are based on participant (face to face) interaction, information flow has been of particular interest for business and organisational analysis. In many cases, social networks are used and extended as an analysis tool that allows communication to be discovered and exploited. Among these contributions, Cross et al. [21] propose an approach that is based on creating a sociogram for an information flow network in an organisation. Also related to organisational needs, Mueller-Prothmann & Finke [67] use social network analysis to develop a method for expert localisation and knowledge transfer. They adapt social network analysis to fit organisational practice where it provides a tool to identify knowledgeable communities and to analyse the structure of information flow within and between organizations. Their analysis uses a range of graph-based metrics to assess the social network structure. Dahel & Pedersen [23] use a questionnaire to examine the role of informal contacts in a specific cluster or sub-network. The authors analyse the knowledge flow and determine whether the employees actually acquire valuable knowledge through informal information networks.

Helms & Buijsrogge [49] seek to extend social network analysis and develop a technique called *knowledge network analysis*. In doing so, they add various concepts that are aimed at making social network analysis more suitable for knowledge networks. These concepts include knowledge management roles, expertise levels, knowledge velocity and knowledge viscosity. Helms et al. [50] focus on evaluating the limitations of network analysis for knowledge sharing. The evaluation is carried out through a case study at an international product software developer event. The authors compare and contrast qualitative and quantitative studies and use this to extract limitations of network analysis in the context of their work.

Fischbach et al. [36] also study information flow between workers within an organisation but extend their analysis to a range of different tools and technologies. The study uses dynamic social network analysis and includes face-to-face interaction, email and instant message communication. These studies show how social networks can be exposed and exploited. It is noted that identifying scenarios within which both online and offline social networks co-exist is a challenge.

Beyond the use of social network analysis in an organisational context, social network analysis has been employed to study information flow through diverse technologies. In [63], Martinez et al study collaborative support in computer aided learning. A mixed evaluation method is used that integrates quantitative statistics, qualitative data analysis and social network analysis. They combine different types of data source to support their approach where the data sources include computer log events, face-to-face interaction, questionnaires and focus groups. Email communication is the focus of Kossinets et al. [56] who study the temporal dynamics of communication using emails between the staff within a University, defining a network *back-bone* that maps the quickest information flows in the network.

Tang et al. [91] propose a new approach for information detection and tracking on blogs using both social features and text. They focus on discovering hidden communities within a social network and develop a weighted graph representation showing closeness of relations. They combine these methods to extract the information flow and investigate both temporal and spatial dimensions. Lim [59] looks at patterns of information flow in the chat exchanges of two virtual learning groups using social network analysis. Quantitative statistical network analysis is applied to textual data (tutorial transcripts) containing the exchanges of information within the online collaborative learning context. Information flow is also tackled from a physical perspective in [99]. Statistical calculation is used to analyse the information flow of instant messages between employees in an organization, taking into account the observation that people interested in an item are likely to belong to the same social groups.

Currently, relatively few studies examine both online and offline social communication structures over a single community. The *hybrid* online-offline social network has been defined as a social network in which social links are maintained using both online and offline methods of communication [40]. Some of the few studies tackling the hybrid network include [1] where the authors consider the embedding of social networks in different technologies. Further studies focus on the psychology viewpoint [89] where they establish how networks of *friends* of young adults relate to their online social networks. These studies examine the social network structure using social network analysis approaches including an ego-centric viewpoint. Haythornthwaite et al [48] study the impact of social media on existing relationships and the influence on strength of relationship between parties.

The focus of a number of studies concerning hybrid social networks concerns the issues of trust and identity. This affects the sharing of information and in some cases it is claimed that a mixture of virtual and physical social networks may overcome these difficulties. For the purposes of knowledge sharing, [64] [40] investigate communities where physical interactions are extended in to the virtual world. Subrahmanyam et al [89] attempt to answer the question as to whether having physical interactions combined with virtual interactions reduces problems concerning trust and online knowledge sharing. Conversely, in a reverse study, Xie [100] found that as well as enabling the creation of online social relationships, the Internet can affect offline relationship formation.

2.6 Summary

A summary was provided about different routing and information provision techniques in the opportunistic networks where the flooding technique will be used to evaluate the hypothesis in Chapter seven. Mobility tracking techniques were studied in order to be able to detect the opportunistic network that the device sees locally. From the chapter as a whole, it is important to determine the type of sensor that will be used to support the developed system. It was found that BT is a reasonable approximation for sensing face-to-face social interaction. Bluetooth is embedded into all the new mobile devices such as mobile phones, PDA, laptops. This means that it is not required of the participants to carry other device for experimental purpose. Bluetooth was designed as a short-range communication system with range of room sizes (class 2 Bluetooth device ranges 10m in free space and less indoors) which eases the implementation of proximity-based location. As a result, most Bluetooth tracking systems are based on proximity [16].

Social network analysis plays an important role in understanding the characteristics of any type of data that represents interaction between people, such as phone-calls, email, Facebook, and physical interaction. It has been found that it is challenging to apply social network analysis in indoor mobility tracking to extract the users' social network; this has not been found in other related studies. What is more, other previous research did not apply social network analysis on the recorded physical interaction between users. In addition to, studying the dynamics and mobility patterns of users with the recorded physical interaction detect their opportunistic network. All these challenges have been addressed throughout this research.

All the previous related research compared the nodes in the network together in order to extract the similarities and other characteristics such as tie strength. In this research, different networks are compared with each other and this has not been done before. Information retrieval evaluation measurements were suggested to compare the extracted networks together, such as precision and recall.

Chapter 3

Mobility Tracking Data Collection System

Introduction

In order to investigate the hypothesis, we need to be able to track a group's physical mobility behaviour and interactions between each other. We cannot create an opportunistic network so we need to detect when opportunistic transmission could be made. From our knowledge, no opportunistic network platforms are available other than the prototype developed in the Haggle project [88]. Therefore, we have taken a different approach and looked at developing a system to detect opportunities for peer to peer transmissions. This is a complex tracking problem that cannot be addressed by a single commercial system as far as we can establish.

This chapter details the steps to design and implement a robust mobility tracking system for data collection. The system is complex since it has been developed for mobility tracking (a path of locations over time) and also identifying the interactions between participants' devices. To make this viable, we are examining behaviour at a local level for mobility tracking, where mobility tracking is monitored inside a complex of buildings and we are capturing the participant-to-building interactions. This allows for a well-defined area where we can consider location and use computing resources and infrastructure at our disposal through the School of Computer Science and Informatics.

For the purposes of physical interaction, we judge participants to have a physical interaction if we can establish they are in a 2-3m range of each other for more than one minute. We note that this is an approximate measure but it is sufficiently accurate for us to have a basis for study, analysis and to investigate the validity of the research hypothesis. It is possible to try to deduce participant-to-participant interaction from detection of their independent presence at the same location in the building at the same time. This can be complex to deduce and not always accurate because accurate positioning in the building is challenging to achieve with a limited infrastructure. A better alternative is a user-centric approach where devices detect other participants and directly report these detections back to the system. We adopt the latter approach as our assumption in this thesis for the development of system design.

This chapter demonstrates the approach used for the design and development of our indoor mobility tracking system. It is designed with the overall need to be robust and deployed by experts (researchers) who fundamentally require reliability and accuracy. It is important to note that because the system is being deployed on a test case of real participants, it is vital that the mobility tracking system performs well because repeating the experiments with a large number of users is hard to achieve and is dependent on the goodwill of many people. Importantly, the mobility data needs to be reliably collected with multiple users concurrently operating in the same area. The overall constraints and requirements of the system are discussed in Section 3.1. Section 3.2 includes detailed information about the system design issues and the challenges faced during the design stage. Section 3.4 demonstrates in detail the implementation of the system including pre-deployment testing. The chapter concludes with the experimental configuration to apply the research on test users.

3.1 Indoor Mobility Tracking System Constraints & Requirements

This section identifies the system constraints and necessary requirements. Although the tracking system is for research purposes, it is sufficiently complex to warrant some detailed considerations of requirements. It should be noted that many different components need to be developed and organised to achieve functionality. These span wireless communications and data transfer, data storage, mobile platform processing and databases. These issues span a number of computational layers. This can be achieved with a number of different technologies and a number of design choices. However, the requirements provide a basis to make an appropriate selection of technologies and to design an appropriate system. Therefore, we are concerned with user requirements, functional requirements and data requirements.

User requirements define the needs of the users and, in this research; the user is taken as us, namely the researcher. Functional requirements specify the output we require from a given set of inputs. Non-functional requirements define the characteristics of the system beyond the operations that it should perform. These requirements are derived from the need to capture data that includes the movement of participants and interactions between them. This is needed to address the research hypothesis and it is important to note that, to the best of our knowledge, there is no commercial product that is designed to support this functionality. We address the issue of requirements in the following sections.

Before addressing these requirements, we consider the technological assumptions in Section 3.1.1 and 3.1.2. As described in Chapter Two, there are different ways to create location tracking using wireless technology. Our basic assumption is that participants will have a wireless enabled device which is detectable and enables data on mobility and interaction to be recorded. The device will record interaction with other devices and detection in wireless cells can be used to generate a central record for location detection. This is described in more detail below. We are deploying the system for an indoor application over a large area which means that particular technological choices exist.

3.1.1 Wireless Detection

There are now many different techniques that support indoor mobility tracking. Because of the nature of our study, it is sensible to use a simple discrete approach based on micro-cells provided by a wireless location area network or other ISM-band technology. The detection of a mobile device by the base station over a time period is then noted and recorded as location discovery. The two main technologies that can be used for this approach are WiFi and Bluetooth. WiFi is well-used and there is an infrastructure already available in the test bed area across the School of Computer Science and Informatics. However, unfortunately this network is a closed network run under strict regulation by the Information Services division of the University. We cannot gain permission to extract logs for individual IP addresses and only partial information would be available (not necessarily the duration). This means that the only option is to set up a secondary network specifically for wireless device detection. Also required is the ability to co-locate individuals. As coverage for WLAN technologies is increasingly good, wireless cells are large. Therefore, inferring that two participants are interacting just because they are detected in the same wireless cell is not accurate enough. To solve these issues, we adopt and assume a network of wireless cells operating using the Bluetooth wireless communication protocol. These can be configured using antennae (embedded in dongles) that are hosted by PCs. These are easily transportable and can be located around a building.

Adopting the Bluetooth solution for device detection has other advantages. The Bluetooth protocol has an explicit sniffing process for the detection of devices which occurs before a communication channel is set up between two devices. This is ideal as it means that we do not require connections to be established so detection is quick and efficient. Additionally, it is possible for devices to discover each other using the Bluetooth protocol. Again, detection is efficient and does not require devices to handshake and set up a communication channel. This means that discovery of other nearby devices is quick to establish and easily stored.

3.1.2 Mobile Platform

Bluetooth is available on mobile platforms such as most smartphones, for example, Android and Apple. This makes it possible for us to consider the use of mobile phones as a means to track participants and detect interactions between devices. This is very attractive but a significant problem is that there are many different types and versions of operating systems that control the Bluetooth protocol. Furthermore, there are variations in the version of Bluetooth run on different phones. This makes it very challenging to develop software that can be adopted by a participant on their own phone. Power consumption is another major issue because constant Bluetooth scanning for detection purposes uses a lot of power. This will make running software on a smartphone problematic for users and may affect the level of participation. Because of this, it is necessary to have consistency where all users have the same platform running the same version of the Bluetooth protocol. On the negative side, it means that an additional wireless device has to be carried by the participant but the charge of this can be minimised by adopting small sensor-scale platforms that are easy to include in bags and pockets. Details of the mobile platform adopted are described in Section 3.2.1.

3.1.3 User Requirements

The assumptions on wireless detection and the mobile platform requirements can be used to structure design. To carry out the research, we need to easily deploy the system on a large number of individuals who are involved in the test bed trial. Because many participants are required, it is important that the design allows convenient deployment and relevant data capture. The user requirements below describe this in more detail..

U1. **Ability to track and record participants movements**

The primary purpose of tracking is to establish a complete picture of the participants' movement pattern inside the buildings. A discrete approximation of the participants' continuous movements is the minimum requirement so that a sample of movement is achieved. This represents knowing that a particular participant is at a particular location at a point in time. A time sequence of samples of mobility provides a *trajectory* of the participants inside the building. The system must be able to record the trajectories of all participants for analysis of the movement behaviour offline.

U2. **Ability to track and record physical interaction between participants**

The system must be able to track the physical interaction between participants. This requirement is to detect the presence of a pair or more participants being in the same place at the same time. A detection distance of about 2-3 metres is needed with a co-location of at least one minute. These parameters are set as requirements that make it possible to conduct meaningful analysis and filter out co-location that is not part of an interaction between participants. The system should detect these interactions at least indoors across the deployed area but it is not restricted to this. The system must be able to record the interaction data for analysis of the social network offline.

U3. **Provide privacy and confidentiality for participants**

The system should preserve the identity of participants. Any personal data should not be stored on any component of the system that is not protected. If a participant's device in the system is stolen, it should not be possible to deduce any personal data concerning any individual; the anonymity of individuals must be preserved. This should also extend to the analysis of the data so that no individual identities can be deduced from the findings of the research.

U4. **Real-time access**

The system should support the real-time access to the collected data so that piloting and testing can occur quickly. This will allow verification of detection to occur and allows for reliability and robustness to be checked in real-time. For example, this could be monitoring what is detected when test participants are instructed to behave in particular ways. This requirement means that we can test the accuracy, reliability and robustness of the collected data by following the creation of trajectories and interactions between participants.

U5. **Automatic functionality**

The system should operate automatically meaning that both the participants and the users (researcher) are not required to interact with the system for mobility tracking and interaction detection to occur.

3.1.4 Functional Requirements

Functions need to be carried out by the system so that user requirements can be fulfilled and the system can support the research hypothesis. Functions concern the generation, communication and storage of data. The functional requirements are stated below.

S1. **Recording the presence of a participant at a location.**

Device identity, cell identity, arrival time, date and day of week are required for local storage at the host PC and for transfer to a central database. This central database is updated every time a Bluetooth device has been detected. This is required to keep the collected information in a minimal set of fields.

S2. **Ability to record the presence of the participants at different locations.**

It must be possible to extract from the central database the time sequence of detection at different locations so that, for each participant in the system, a trajectory can be created. Wireless cells that are hosted by PCs are distributed throughout the building where different PCs at different locations are connected with Ethernet LAN network. This connection between PCs is required to enable recording of the detected Bluetooth devices at different locations in the central database.

S3. **Ability to record the presence of multiple participants simultaneously.**

The system must be able to record the location of multiple participants simultaneously. This is because multiple wireless cells are able to record the detected Bluetooth at the same time. A feature of the database is that it is required to manage the recording of different PCs at the same time as each PC is separately performing and while the central database controls the traffic of recording data. The database must have the ability to control the traffic and enable multiple records to be stored at the same time. At least five cells and a maximum of 15 cells will be used for detection, which is sufficient to completely cover the case study building (i.e. School of Computer Science and Informatics).

S4. **Ability to detect Physical interaction between participants.**

The mobile platform must able to detect the physical interaction between participants by detecting their presence in range. The mobile platform should able to sniff for nearby Bluetooth devices. The Bluetooth devices are required to support two functionalities: sniffing for nearby devices and being discoverable for other devices to be detected by them to be able to detect the physical interaction between the participants.

S5. **Ability to record the physical interaction between participants**

The mobile platform is required to be able to record all of the detected physical interaction. The mobile platform must support the facility to save the devices into a hash table to minimize the required space to record the interaction into a system file. As long as the Bluetooth device is a small device with a very small storage, file transfer protocols are required to be established to be able to backup the recorded files into a server machine (PCs).

3.2 Overview of the Mobility Tracking System Architecture

This section explains the design of the mobility tracking system that directly addresses the requirements and constraints stated in Section 3.1. The overall design is presented in Figure 3.1. The hardware components concern the wireless devices, workstations acting as hosts for Bluetooth cells and a central server that records all mobility activity. There are two main aspects to the overall design.

- Mobility data capture

  The wireless devices scan for other wireless devices that come in range to record their presence (physical interaction). The presence of other wireless devices is stored in a file called system file. Once workstations with local storage come within range of the wireless device, FTP connection has been established in order to send the system file (this includes physical interaction data) using FTP protocol. These work stations must have a dongle Bluetooth plugged in them to be able to establish this connection. FTP connection is crucial to be able to dump the physical interaction data as there is not enough space in wireless devices to dump all of the experimental data on its local storage. FTP protocol for data dumping is explained in detail in subsequent sections.

- Data storage and aggregation

  Dongle Bluetooth is plugged into other workstations that are continuously scanned for the presence of any wireless devices in range. These machines are connected with a central database server through the LAN. If a device has been found, the detailed information of its presence is recorded into a central database (mobility tracking data). This central database is able to receive records from multiple workstations where it is able to store multiple data from different workstations simultaneously. It is able to receive records from up to 15 workstations at the same time.

  In this system, two different types of workstation are utilized: a workstation with a local database, and a workstation connected with a central database. Although both workstations have the same specifications, they have different functionalities to perform. The workstation with a local database listens for any wireless devices, such as Bluetooth mote, that come in range to establish FTP connection to dump the system files. However, the workstation that is linked to a central database sniffs for any wireless devices that come in range to record its presence into the central database. It is not applicable for dongle Bluetooth to perform two functionalities (sniffing and establishing FTP connection) at the same time; controlling the traffic between the two functionalities is very challenging.

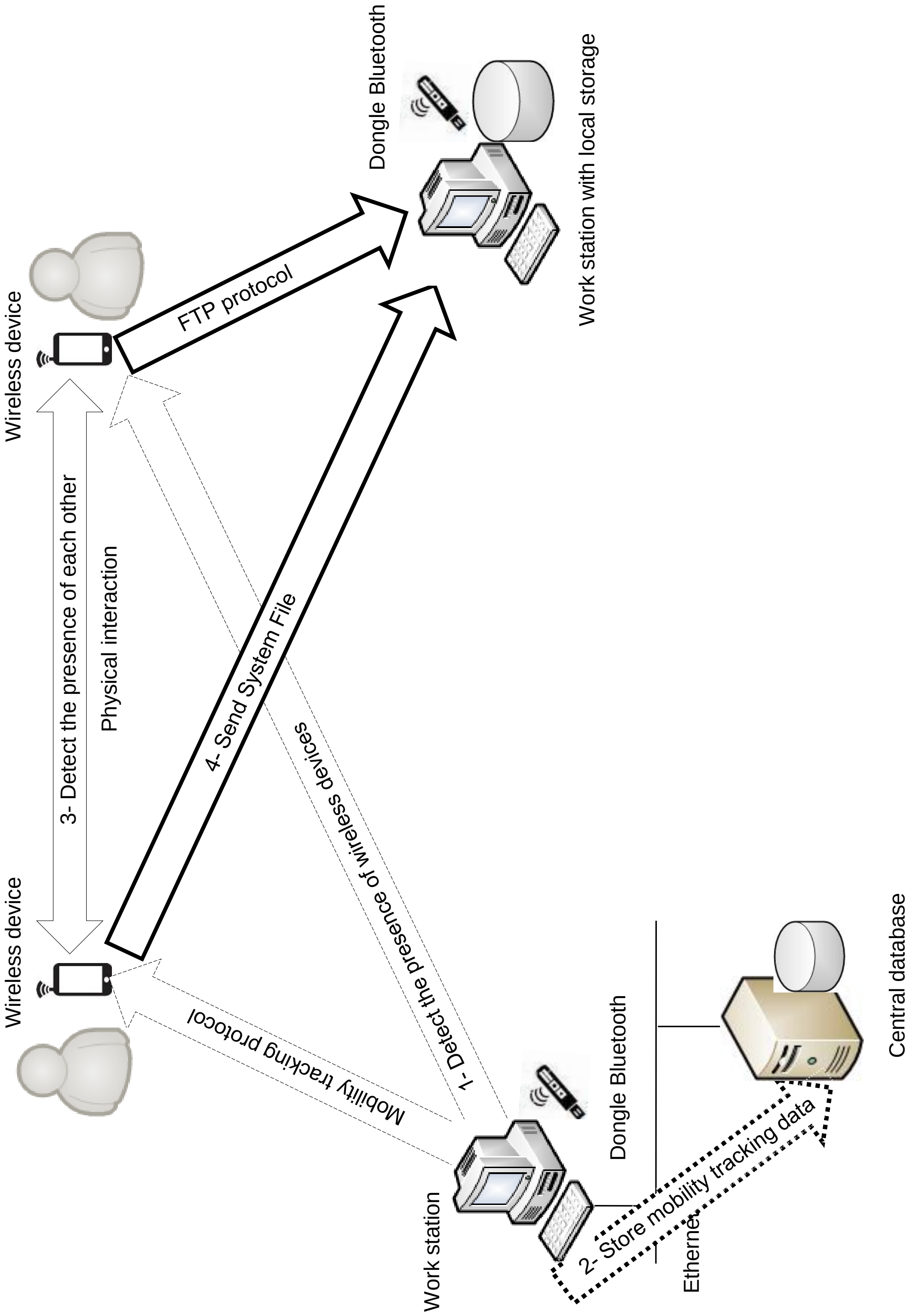


Figure 3.1: System Framework.

3.2.1 System Hardware: Wireless Platform

Bluetooth technology was chosen to be utilized in the tracking system for the reasons described in Section 2.2.1. The design issues concerning Bluetooth also address user requirements $U1$ and system requirements $S1$. Excluding the possibility of using smart-phones as the wireless platform means that we need to adopt bespoke wireless devices. The AIRcable Mini platform was chosen as a simple programmable platform that is Bluetooth enabled and can be controlled through the development of a single application.

The AIRcable Mini is actually a breakout board for the AIRcable SMD. AIRcable SMD is a unit that integrates an enhanced BASIC interpreter, a wirelessly accessible file system for data logging and program code and an industry standard version of Bluetooth. Therefore, the AIRcable Mini is a standalone Bluetooth device that can be programmed to control the specific functionality of the wireless protocol. This allows the system to overcome some important technical challenges that cannot be addressed through the creation of a single application for diverse smartphones. For example, the AIRcable Mini can be programmed to be in discoverable mode without any limitations in contrast to the operating systems in new mobile phones. This allows us to overcome the issues in new mobile phones, such as HTC, iPhone, Android, where restrictions related to controlling Bluetooth have been implemented to increase security. In addition, to support user requirements $U3$, the identity of the Bluetooth MAC address has less importance if discovered by attackers because the AIRcable Mini is not being used for storage of any personal or valuable data.

Figure 3.2: AIRcable Mini.

AIRcable's Mini can be programmed wirelessly for applications written in BASIC with no code compiler required. The on-board file system allows us to run BASIC programs and store information gathered by its data-logging function. For example, data streaming in from a sensor connected to the serial port can be processed and stored. This means that it will support data gathering functions that required for the detection of other devices. In addition, data transfer

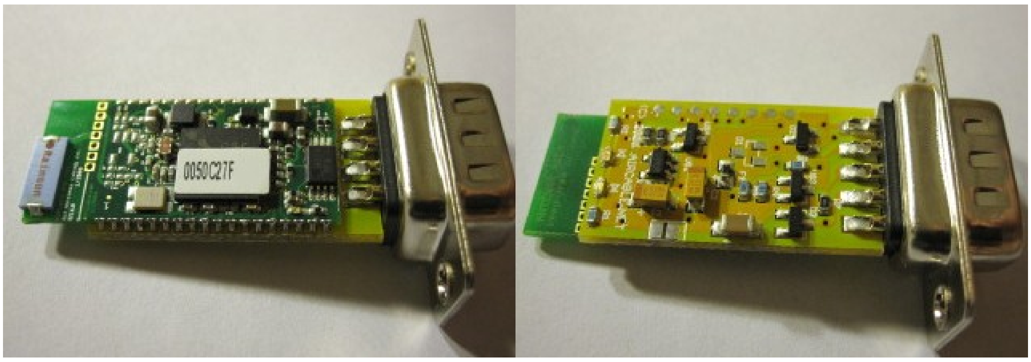


Figure 3.3: Inside AIRcable Mini.

of data in the form of plain files can be instructed through FTP operations over the Bluetooth channel. This is the communication mechanism that we can use to extract data from individual devices to the central database. The AIRcable SMD requires about an 10mA average power. The time for which it is operational can be calculated when the capacity of the battery is reliably known. In discoverable mode, power consumption is only about 1-2mA on average, while the inquiry mode is the most expensive, being 10-20mA. For operational use, the AIRcable mini needs to be cased with its battery supply and there is no need for a physical interface to be included.

3.2.2 System Hardware: Servers and Hosts for Bluetooth Cells

A small number of dedicated workstations are used for two purposes: to operate the Bluetooth cells and detect/communicate with wireless devices and secondly, to store captured data and stream this to a central database. These workstations have the standard Windows 7 operating system with the characteristics of 2.6 GHz processor and 1 GByte of RAM. One of these workstations is used as a server that forms the central database for recording the collected tracking data. The wireless cells are provided using Dongle Bluetooth devices which have been plugged in via USB connection for scanning and data transfer. The class 2 of Bluetooth dongle has been chosen as it supports 10 metre accuracy outdoors and 2-3 metres indoors. This results in relatively small cells which are ideal for the tracking scenario that we are considering here.

3.3 System Communication Protocols & Data Storage

To capture and deliver data through the system, communication protocols and data storage need to be defined. The system has two main functions to perform : detection and recording of co-location of participants and secondly, tracking and recording the presence of individual participants in cells. We discuss these issues in the following sections.

3.3.1 Physical Co-location Tracking

There are three different communication components that need to be addressed. These are shown in Figure 3.1 and involve the following.

- Detection of another mobile device by a mobile device;
- Storage of the event at the mobile device;
- Transfer of the record to the server via wireless connection.

These components need to operate together as follows. The AIRcable Mini detects the wireless devices in range using the Bluetooth scan operation. Wireless devices that are detected have their presences recorded and stored locally as a file where no connection is established with the detected wireless devices. When the AIRcable Mini detects one of the workstation stations, the AIRcable Mini establishes a connection with the server and the file transfer protocol is used to send the file from the AIRcable Mini to the server. After sending the system file to the server, the Mini deletes the file from its local storage as it is too limited. This is the scenario in steps 3 and 4 as seen in Figure 3.1. It is a challenge to program the AIRcable Mini device so that it achieves the specific functional requirements. This is because of the very limited memory and processing capabilities of the device. Therefore, a very simple and robust approach is used where hash codes of devices are used to record their identities. For each AIRcable Mini and workstation, a hash code is assigned. This hash code comes from the hash table that we have created to avoid storing lengthy Bluetooth MAC addresses. The hash table includes 45 devices (40 Bluetooth devices and 5 servers' MAC addresses). As a result, only two digits are stored in the system file instead of 12 digits for each found device. Programming for the AIRcable mini has been undertaken using the BASIC language and has been developed from scratch to enable the Bluetooth device to perform specific functionality. This involves the following steps:

1. Inquiry for nearby filtered Bluetooth devices.
2. Store hash index to the system file.
3. Send the system file to the server machines using FTP service.
4. Be discoverable to be detected by nearby Bluetooth devices

Algorithm 4 Physical Interaction Detection Algorithm

```
L ← 300 {the start of the hash table index}
M ← 45 {the number of the devices in hash table including servers }
The device become discoverable for 5 seconds
for Each detected device D do
  if D in Hash table then
    create file "found.log"
    write the index on "found.log"
    if D is one of servers then
      send "found.log" to the server D
      if "found.log" sent successfully then
        Delete "found.log"
      end if
    end if
  end if
end for
```

Although writing programs in BASIC is normally a straightforward task, it has been complex here because of the limitations of the AIRcable Mini. Many functions have not been well documented or supported and testing is a lengthy process because the AIRcable Mini has no physical interface; programs are installed through wireless connection and the behaviour of the device then has to be assessed. The developed BASIC program is short and robust and is presented in Algorithm 4. In Algorithm 4 particular values have been assigned to different variables to populate them. These are L which is the starting index of the hash table and is set at $L = 300$ so that the hash table starts after the last line of code, and M which represents the maximum number of Bluetooth devices that need to be monitored in the system.

The storage space in the device is 32 KB and so frequent transfer of files is required to the server by Bluetooth. Each AIRcable Mini device is programmed to create a system file that includes the code for all the seen Bluetooth devices indoors and outdoors. A file transfer protocol has been developed where each AIRcable Mini device has the ability to create an FTP connection with the server in-range and send the system file. The channel set-up and FTP protocol is described in Figure 3.4. The system is defined such that the AIRcable Mini only seeks to set up a channel and transfer the system file to one of the identified workstations. The connection and FTP transfer works as follows. The AIRcable Mini device scans for any server in range. If there is a server in range, it responds to the AIRcable Mini, identifying the MAC address of the dongle Bluetooth that is plugged into the server (a local machine with local database). Then, the AIRcable Mini device asks for an FTP connection to be made if this is available. The server asks the AIRcable Mini device for its MAC address for authentication. The AIRcable Mini device sends the MAC address to the server as the authentication credential. The server replies with a confirmation that the FTP connection has been established. The AIRcable Mini device then sends the file.

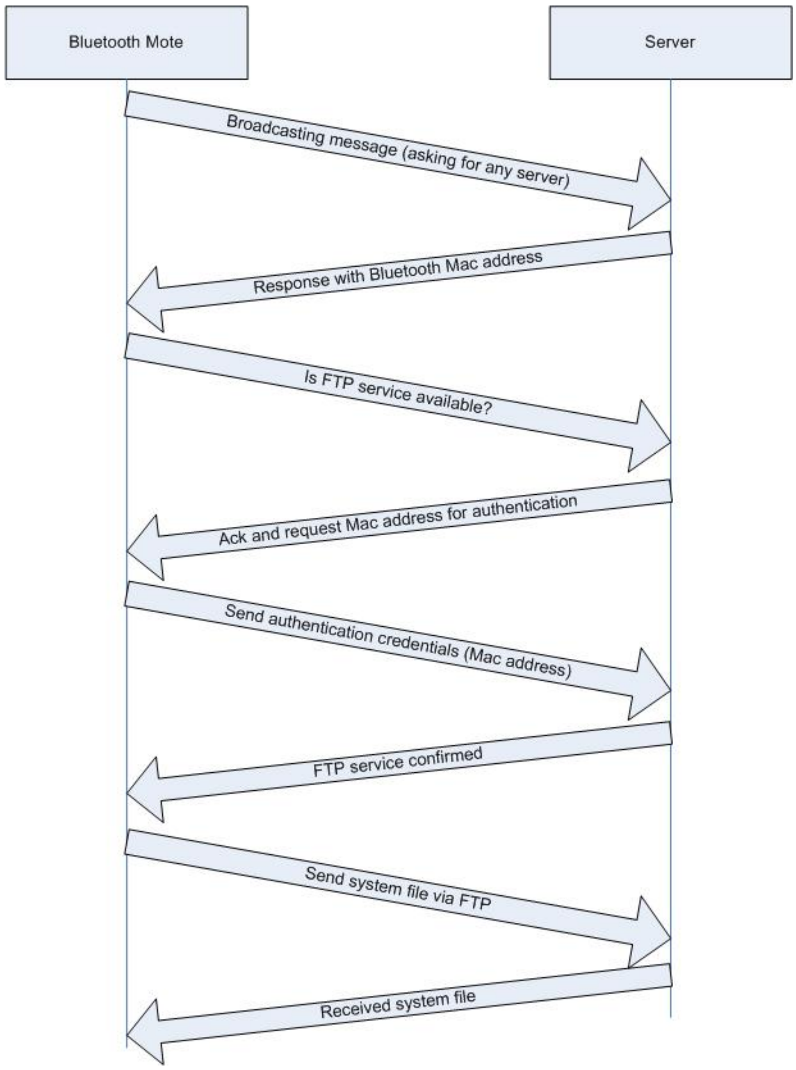


Figure 3.4: FTP Connection Protocol.

3.3.2 Mobility Tracking

As seen in the system framework shown in Figure 3.1, workstations are responsible for detecting the presence of mobile devices. There are essentially two techniques for using Bluetooth to track users indoors: tracking with inquiry and inquiry-free tracking [9]. As regards the first method, the base station sends a discovery packet in each of the 32 radio channels. Bluetooth devices that are in range and discoverable respond to this packet by identifying themselves. This response follows a random delay to avoid collision when multiple devices respond. As a result, the inquiry process is repeated every 10.24s to reliably detect all the devices in range. The second method is a *connection-based* tracking, where the two devices are defined as proximate when one can connect to the other [46]. In this study, the first method has been chosen to (tracking with inquiry) as it is a common method of tracking indoors via Bluetooth and no connection is required to discover the presence of other devices in range.

Using the workstations to perform tracking is preferable because most of the dependency is on the workstations and these do not have resource limitations as compared to the AIRcable Mini devices. Workstations are programmed to scan for detection of AIRcable Mini devices and then the instance of the detected device is sent to a central server where the data is stored in a central database, as seen in steps 1 and 2 in Figure 3.1.

This approach to tracking is used by many systems [79, 96, 9] and is based on continuously scanning (inquiring) from a network of fixed beacons. This technique requests the tracked Bluetooth devices required to be on discoverable mode all the time. This is one of the main reasons why using the AIRcable Mini device is an advantage. It avoids privacy issues which may arise from tracking mobile phone handsets and avoids addressing variable security practices on different mobile handsets. We have found that mobile phone manufacturers are increasing the security by limiting the time that a handset can be in discovered mode. It is worth noting that being discoverable all the time can make a device a target for hackers. Finally, the scanning process can disturb the communication channel [83] which is one aspect that has motivated pilot testing in Section 3.4.

This application has been developed from scratch to support the system and consists of three processes that are sequentially completed. Firstly, a process retrieves the information from a local device. Secondly, a process starts looking for the Bluetooth devices that are in range of the workstation. Finally, a process caches the list of nearby devices. Figure 3.5 shows the flow chart representation of the application. This application represents step 1 in Figure 3.1.

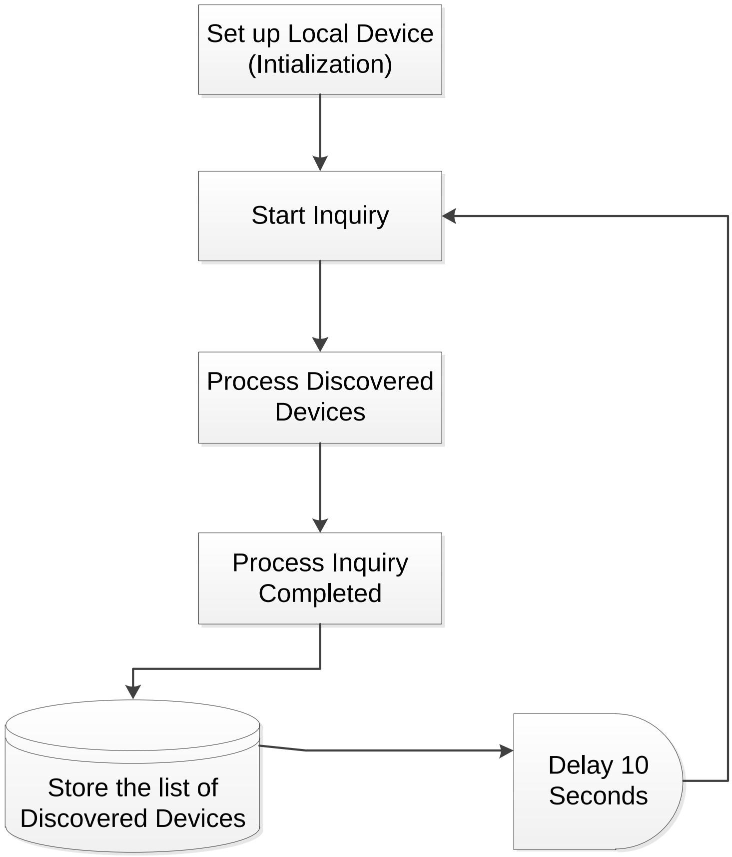


Figure 3.5: Bluetooth Mobility Tracking Protocol.

3.3.3 Java Implementation and Database Schema

Java has been used in developing the system for deployment in workstations as it supports a formal procedure to incorporate a simple concept of Bluetooth into standard Java. This is called JSR-82 but it has the limitation of only being implementable on a J2ME platform, which is designed for embedded systems, such as mobile devices. Consequently, to implement JSR-82 in workstations, we need to adopt a suitable API. In order to solve this, the Bluecove API has been used to fix the problem. It is a free API that supports the ability to program Bluetooth on desktop machines rather than mobile phones or mobile emulators. Bluecove supports different versions of Bluetooth and it can be used on different operating systems.

The mobility tracking data collected by workstations needs to be stored to satisfy the functional system requirement S2. MySQL database has been used to record detection events and there are two simple tables in the database: one table for participant registration and the other for storing tracking information. Participant registration is collected only once before the beginning of the experiment. The second table is for storing the presence data of the user in the building. The storing process of this table is repeated until the end of the experiment. Figure 3.6 shows the schema of the two tables. The relationship between the two tables is one to many; in other words, each user in the registration table has many instances of records in the discovered devices table. Therefore, each user has a unique Bluetooth address (registration table) and he/she can be detected in many places (discovered devices table). This simple schema is consistent with the needs of the devices that can only maintain simple records of events and is a minimal set of fields that addresses all the requirements. The date, time and day of week represent the temporal start of the event. BtAddress is the unique code that identifies each participant based on the wireless device that is carried. Friendly name is the nick name of the Bluetooth device that can be changed by the users where it is not a unique name. In terms of the Bluetooth mote, the friendly name has been initiated by the producer company and it can be changed only by the developer (the researcher). As a result, we use this name and field *Type* to filter the detected devices to the experimental devices only.

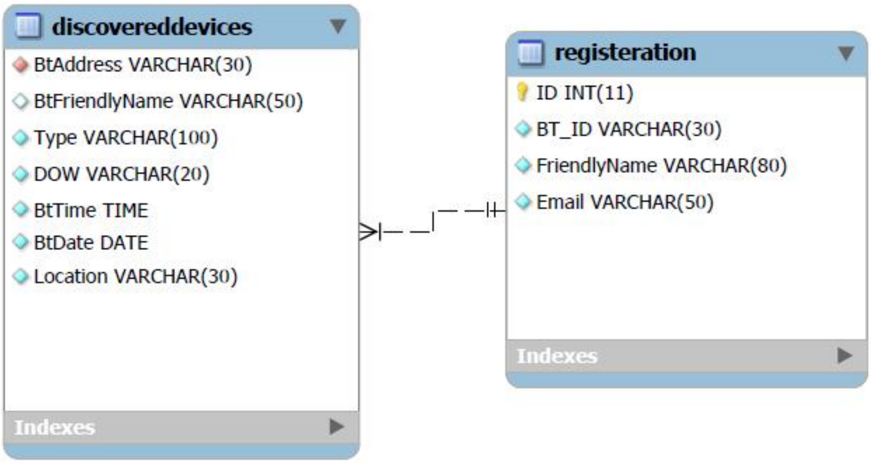


Figure 3.6: Database Tables Schema.

This database has been created in a server machine to perform step 2 in Figure 3.1. This database enables multiple machines to store their mobility tracking data simultaneously. Furthermore, the database is able to support up to 15 machines to store data at the same time where this feature has been added to the database to support the system requirements S3.

In order to record the physical interaction between wireless devices, a system file has been created on the wireless device local storage to store the hash code of the detected wireless devices in range. The system file is sent to the server (workstation with local database) once the server comes within range of the wireless device. The sending process occurs by establishing a FTP connection between the server and the wireless device. These system files are dumped on the server machine by developing an application that changes the file name to a unique name and moves it to a database folder. This application has been developed to overcome the limitation of determining the time of interaction and the difficulty in changing the file name automatically from the wireless device BASIC program.

3.4 Pilot Study Summary

Beyond single device testing, a pilot study was conducted in order to ensure that the system was fully operational. The pilot study was used to test the functionality of the system, its reliability and robustness when more than pairs of devices are considered. Therefore, six devices were distributed to a group of postgraduate students in the School of Computer Science and Informatics for a period of one week. In particular, this was used to determine the power consumption characteristics under realistic usage patterns. The battery life was found to be three days minimum and, for this reason, the experimental study was divided into slots of two days to ensure battery life did not expire. For the mobility tracking perspective, the positioning and location of eight workstations was also tested along with the ability to track different participants. After a week of running the system we found that:

- The building indoor environment is very sensitive to the thickness of walls. The locations of some workstations needed to be changed as they did not detect any participants.
- The battery life lasts for at least three days of continued inquiry. Two types of batteries were used but both gave the same results in terms of battery life.
- The speed of the participants could affect the sending of system files to the servers. The speed governs how long a participant remains in the cell.
- All the system functionalities performed accurately and robustly and were established by case-by-case black box testing.

3.5 Experimental Set-up

In order to set up the experimental study, the workstations were located in tactical locations to provide coverage over the critical areas for the participants in the School of Computer Science and Informatics. These areas were chosen to satisfy the following issues. Firstly, we needed to cover all the school building entrances and there are three main entrances. Secondly, we needed to ensure that workstations could cover the main rooms used by students for their studies. Finally, we needed to cover the main junctions and corridor areas. Coverage characteristics were established by placing workstations in different locations throughout the school buildings and measuring the received Bluetooth signal strength in the locality. After many trials of measuring, the following figures approximate the Bluetooth footprint at the different chosen locations. We used five locations that should be covered, denoted L_1, L_2, L_3, L_4, L_5. Figures 3.7, 3.8(a) and 3.8(b) show the coverage area for four different locations. The coverage is not 100% accurate as there are some obstacles inside the school, such as walls. In addition L_4, which is the library, is a short distance from the mapped area on the diagram. A topology of the locations is described in Figure 3.9, with edges describing adjacency between the workstations (in the sense of a direct walking route). The solid edge means that there is a direct path between the two locations where the dotted edges denote an indirect path between this pair of locations (we have to path by many locations to reach to the destination).

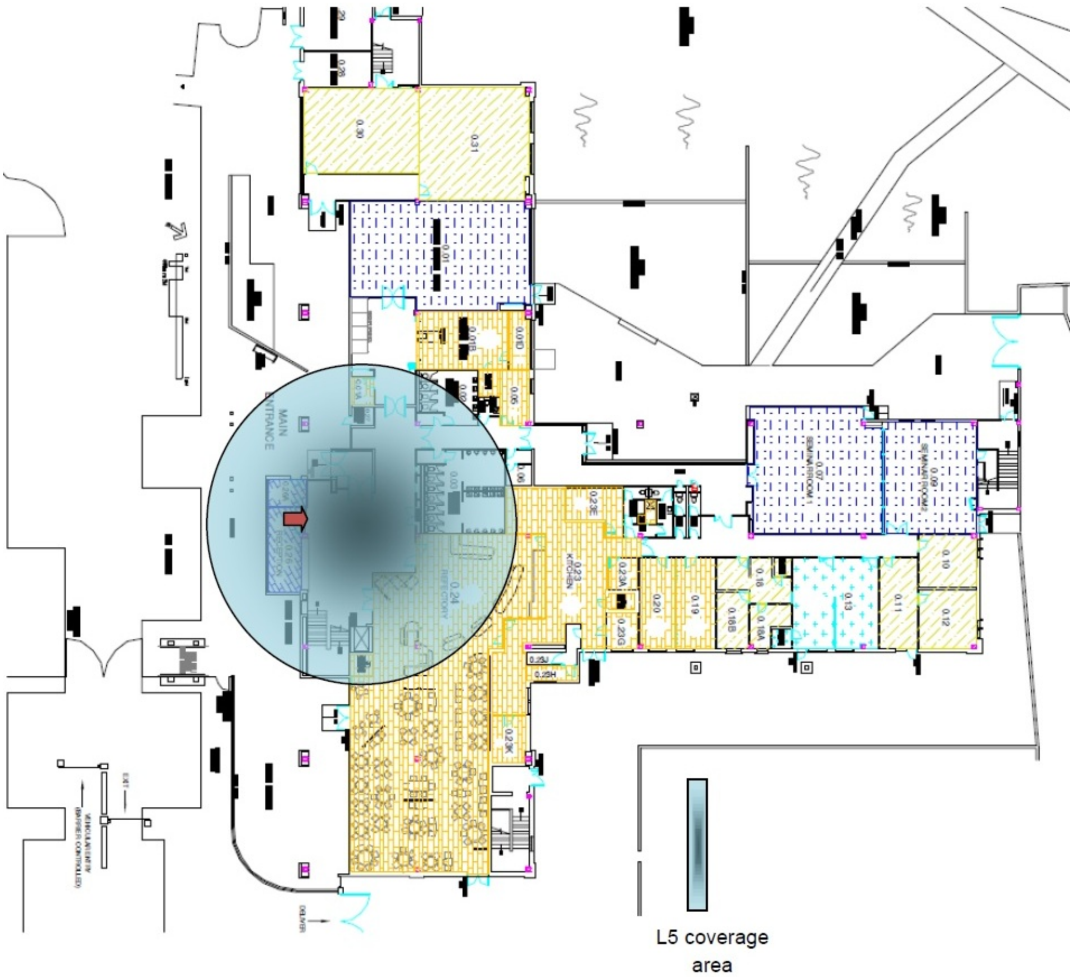
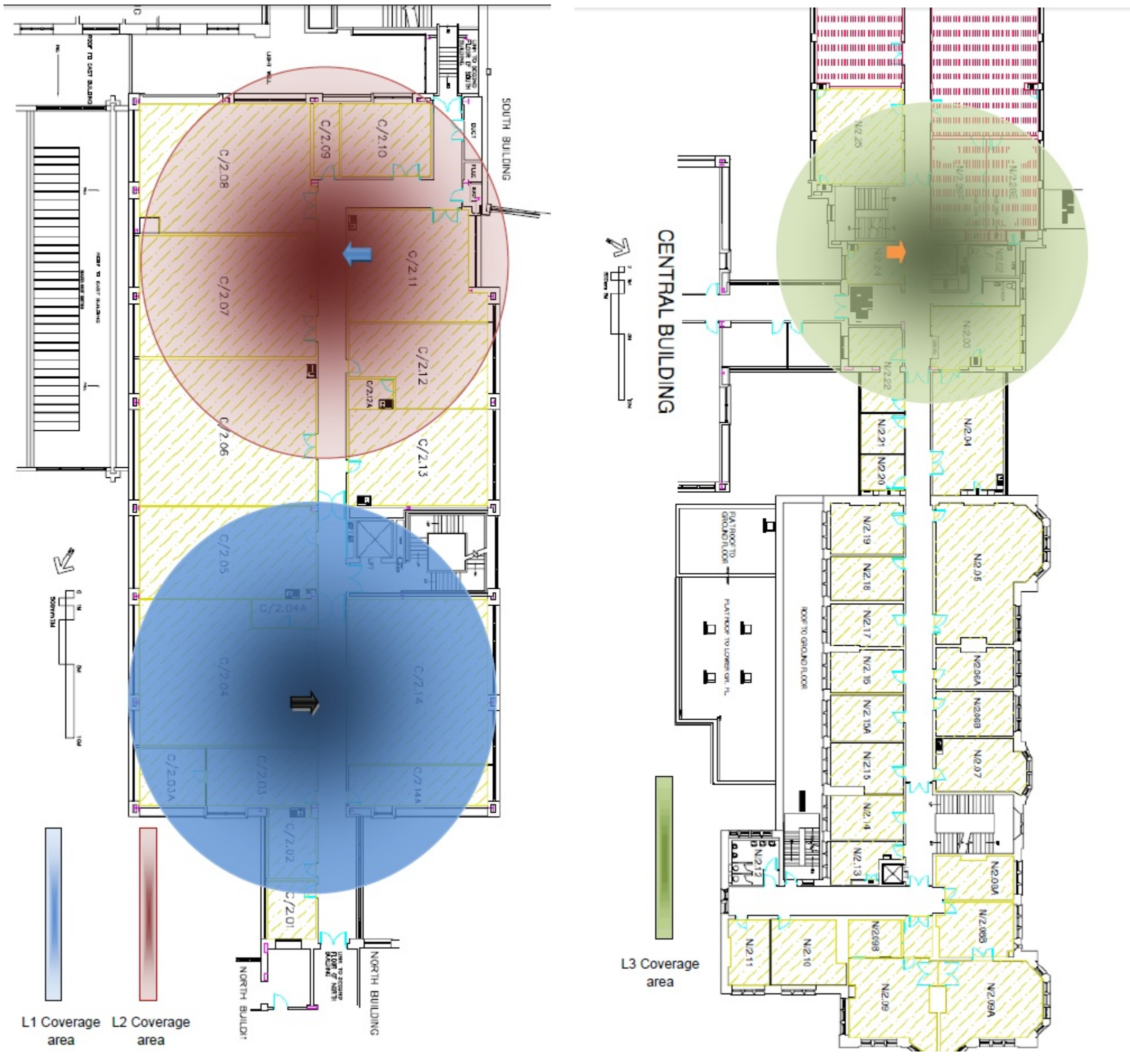


Figure 3.7: Bluetooth Footprint for Trevithich Ground Floor.



(a) Central Building Second Floor (b) North Building Second Floor

Figure 3.8: Bluetooth Footprint for Central Building & North Building Second Floor.

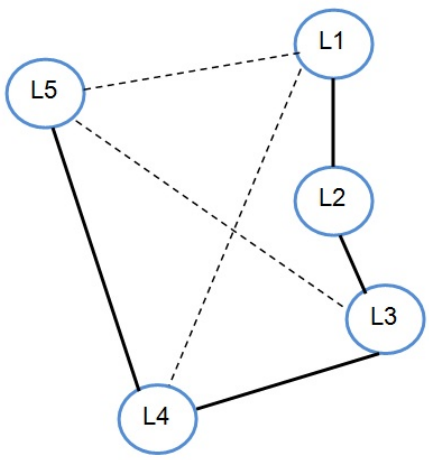


Figure 3.9: Locations Topology.

3.5.1 Experimental Hardware

The hardware used for the experimental set-up can be summarised as follows:

- 10 desktop workstations
  These workstations ran the two applications (presence detection application and physical interaction data collection). These machines were distributed throughout the school building to mostly cover different school buildings (namely Central, North, South and Trevitech buildings). Five of them work as servers (connected with central database) that are distributed throughout the school building to detect the mobility of the participants. The other five work as servers (with local database) for Bluetooth devices to receive the system files of Bluetooth motes by establishing FTP connection. Dongle Bluetooth devices were plugged in all the workstations to be able to perform their functionality (detecting presence of nearby Bluetooth devices, and receiving the file system from the wireless devices).

- 40 AIRcable Mini devices
  Forty Bluetooth motes were prepared in small boxes to make them portable for users. Each box included the batteries plugged in a battery holder to support the power for the Bluetooth device. This is seen in Figures 3.10 and Figure 3.11. Each device has a unique code on the box that identifies different users and this code also renders anonymous the collected data from the system file.

- Central Database Server Machine
  This machine is responsible for receiving all the tracking information that is stored simultaneously by five distributed servers where the other server has its own local database to receive the system file from the wireless device. The central database was used to store real time user mobility during the experiment time.

The experiment deployment was delayed by three months due to a delay in the manufacturing of the AIRCable Mini devices in North America and devices with similar capabilities could not be sourced. This is bespoke research equipment that is not stored in large quantities in a supply chain.

Figure 3.10: Bluetooth-mote in Box.

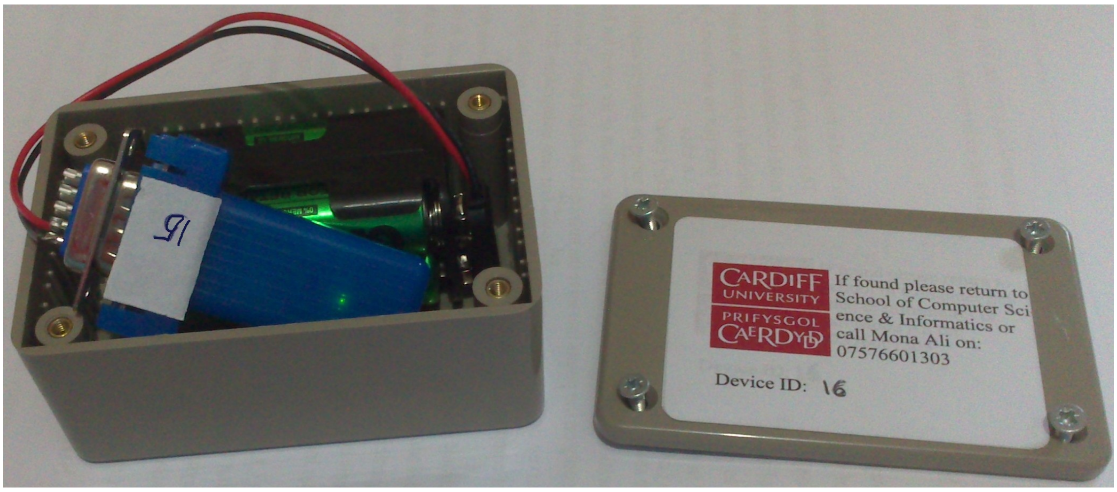


Figure 3.11: Bluetooth-mote in Open Box.

3.6 Summary

This chapter presented, in detail, the design and development of an indoor mobility tracking system to meet all the user and system requirements that are generated for research purposes. It was challenging to create opportunistic networks and, from our knowledge, there are no opportunistic network platforms available other than the prototype developed in the Haggle project. Therefore, we took a different approach and looked at developing a system to detect opportunities for peer to peer transmissions. A detailed framework of the system was proposed. The chapter demonstrated the requirements, design and configuration of the system for deployment and capture of mobility data. Bluetooth technology was chosen for its applicability of localizing people indoor rather than GPS. It was challenging that smart phones limit the discoverability option for their Bluetooth services where the phone Bluetooth can be on discoverable mode for two minutes only and then it becomes undiscoverable.

An external Bluetooth device (AIRcable Mini) was chosen to support specific functionality for its users. In addition, this choice helped us to manage issues, such as using mobile phones that have limited Bluetooth functions and protecting the identity and privacy of individual participants. The device has some useful characteristics, such as the ability to be programmed, it is small in size, and can perform independently with battery support. The device is also able to automatically record interaction with devices carried by other users. Workstations with a local database were used to dump the system files from the Bluetooth devices. Dongle Bluetooth devices were plugged into these workstations to support the ability to establish FTP connection between Bluetooth devices and the servers when they come within range. FTP connection is established to send the system file to be dumped. The system files are dumped in the workstations by means of a new unique name to avoid overwriting any of these files while storing.

Another workstation that is connected with a central database was utilized to identify for any nearby Bluetooth devices. A Java application was developed to track the participant's mobility indoor while carrying the AIRcable Mini. The Java programming language was used to develop the tracking system as it supports the API that enables us to program Bluetooth operations in the device. The Bluecove API was used to develop the system on desktop machines instead of mobile devices. Different hardware components were integrated together to address the two applications for mobility monitoring and interaction between devices. Communication protocols were used to provide data transfer between different components of the system. The reliability of the system was tested before starting the experimental study and the functionality of each component was tested separately. Finally, the performance of the system as a whole was validated through the pilot study.

Chapter 4

Social Network Survey Design

Overview

Two types of data sets have been used to support the aim of this study (self-reported data, mobility tracking data). The aim of this study focuses on detecting the opportunistic network between the participants and comparing the human social network with the opportunistic network which will lead to addressing the study hypothesis. The mobility tracking data collection system (opportunistic network) was represented in Chapter Three, demonstrating the detailed steps in developing a framework for opportunistic network detection. This chapter is concerned with building up an *ego-centric* view of the social network by asking people about their own perceptions. This is therefore *self-reported data* and compared to data in Chapter Five which was collected from observation. This chapter focuses on the methodology for collecting the self-reported social network data in Section 4.1 and the design used for collection in Section 4.2. It includes the challenges faced during the collection stage. The chapter closes with a summary on the collected data.

4.1 Information Collection Methods

Our aim in this chapter is to uncover the social network that the population of participants have. This is hidden because it is based on users' perceptions. Therefore, we require a research methodology that will uncover these perceptions. Generally, in order to collect qualitative information from participants, there are six common methods for collection [11]. In this study, it was appropriate to collect information using an electronic questionnaire. This method enables us to explore social networks from an ego-centric viewpoint. This approach has been used in a number of other studies and is suitable for a large number of participants. For example, Eagle, Pentland, and Lazer have all used self-reported data in social network analysis and inferring the friendship ties [30].

A mixed evaluation method was used that integrates quantitative statistics, qualitative data analysis and social network analysis. They combine different types of data source to support their approach where the data sources include computer log events, face-to-face interaction, questionnaires and focus groups. Email communication is the focus of Kossinets et al. [56] who study the temporal dynamics of communication using emails between the staff within a University, defining a network *back-bone* that maps the quickest information flows in the network. Dahel & Pedersen [23] use a questionnaire to examine the role of informal contacts in a specific cluster or sub-network. The authors analyse the knowledge flow and determine whether the employees actually acquire valuable knowledge through informal information networks.

4.2 Social Networking Survey

This section addresses the steps that were followed to conduct the survey on social networking. To design the survey a number of steps were conducted. Firstly, the purpose of the survey requires a clear definition. From this step, the requirements of the survey study could be defined and these could be used to form the survey questions. Before deploying the survey, it was crucial to pilot it to ensure that all the questions could be interpreted by participants and that the flow of questions was appropriate [25, 26, 11, 61]. The following subsections describe in detail each step involved in designing the survey.

4.2.1 Survey Study Purpose

Our purpose was to determine the physical social network that the participants of the survey have between them. Given the nature of the participants as students, this picture was built while examining the relationships they have with one another, how these are sustained, and how these relationships function as a basis to share information. This provides a concrete way to examine how interactions are sustained between individuals and the extent to which online interactions and electronic media (as well as face-to-face interactions) function and support social relations. Examining the physical and online interactions that support relations would allow us to build up an in-depth picture of the students social network and also how this functions to facilitate communication.

4.2.2 Survey Design

We explore the form and role of relationships in terms of communication. The population that we are exploring is a social network of students in higher education and this is a well-defined and self-contained social network with hybrid online and offline (physical) properties where technology and face to face interaction are both important. The survey was designed to explore the network structure in different ways, specifically:

- *relationship strength and social network structure:* determine the role of weak links and the modes of communication relative to relationship strength;
- *communication technologies and social network structure:* determine the dependency on and role of different types of communication in the social network;
- *frequency of interaction and social network structure:* determine the critical structures that are facilitating most of the communication traffic and the frequency with which different technologies are used.

Currently, relatively few studies examine both online and offline (physical) social communication structures over a single community. The *hybrid* online-offline social network has been defined as a social network in which social links are maintained using both online and offline (physical) methods of communication [40]. One of the few studies tackling the hybrid network include [1] where the authors consider the embedding of social networks in different technologies. Further research from the psychology viewpoint [89] establishes how networks of "friends" of young adults relate to their online social networks. These studies examine the social network structure using social network analysis approaches including an ego-centric viewpoint. Haythornthwaite et al [48] study the impact of social media on existing relationships and the influence on strength of relationship between parties. These aspects will help us to understand a complete picture of the social network and how it functions from the users' own perspectives. These issues have been focussed on after developing a pilot questionnaire as originally structured in Figure 4.1. From implementing the pilot questionnaire, a number of issues were found that led to significant refinements. In particular:

- The number of issues being tackled in the questionnaire were too numerous and led to partial completions;
- There was some potential overlap between issues that could be refined and simplified further;
- There was some participant confusion over the meaning of questions which needed refinement.

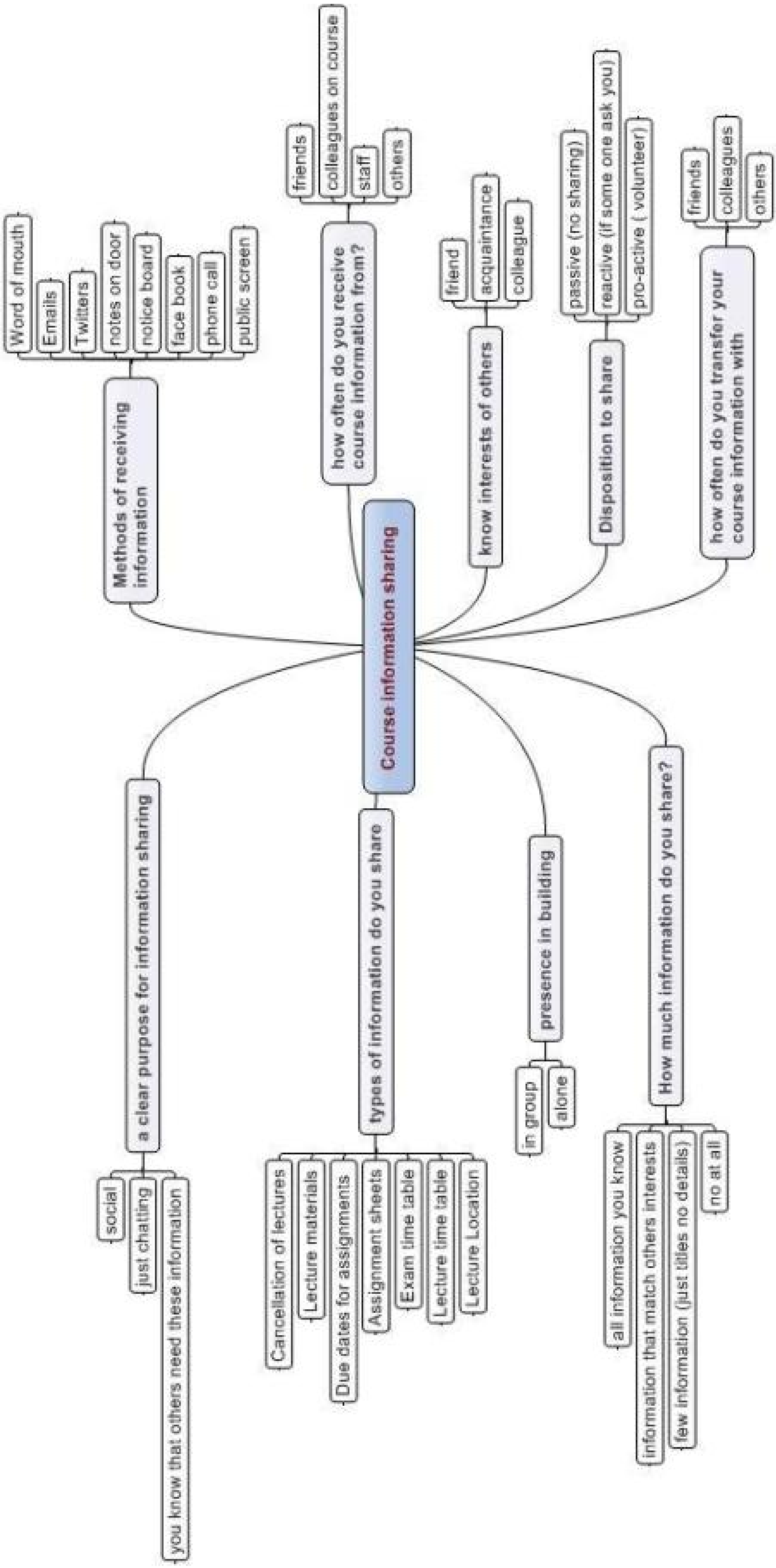


Figure 4.1: Course Information Sharing Issues.

In light of these issues, a significantly refined questionnaire was developed. The questionnaire also allows us to see the mix of different technologies that are used to support relationships. Therefore, from these questions, we were able to significantly increase our understanding of the interplay between online communication and offline social networks and the augmentation effect of online communication technologies.

The questionnaire was designed to establish the characteristics of the social network from each participant's personal view point (an ego-centric perspective). This required participants to express their perceived communication activity with other subjects. This approach allows us to ask questions about a variety of online communication technologies but a disadvantage is the potential for mis-perception of a subject's own activity (in certain circumstances this has been measurable by considering the different perceptions of two parties engaged in a relationship). Anonymity was preserved by recoding personal identifiers in the data prior to analysis.

For each participant, the questionnaire was designed to identify the relationships maintained with others and we investigated how these relationships were maintained. Specifically considered was the intensity of the relationship (relationship strength in three categories), the frequency of interaction and the mode of interaction (offline, online and by which communication mode). Relationship strength was categorised as a "strong friendship" (someone with whom you have significant level of trust and interaction),"friendship" (someone with whom you have empathy or common views and may socialize with) or "course- mate" (someone you know and would acknowledge but with whom you have little other contact). Frequency of interaction was categorised as the most frequent option from daily, at least a few times per week, at least every month, or at least once per semester. Communication was categorised as either face-to-face, mobile phone text messaging, telephone, email, micro-blogging, chatting on the Internet (VoIP), or Facebook, which was a-priori known to be the dominant social networking service used by this group. Thus, each possible relationship could be maintained in $3 \times 4 \times 7 = 78$ different combinations of relationship intensity, frequency of interaction and mode of interaction.

There are many types of questions where each type has its advantages and disadvantages as seen in Figure 4.2. Different question types were used in the survey in order to make it easy for the participants to fill in and simpler for the researcher to analyse as well. There were three main categories of survey questions:

- **Demographic questions**: This set of questions is the open-ended type in order to obtain short answers about their personal information. The collected information from this set will be processed anonymously as the respondents have been informed.
- **Sociality**: This set of questions is partially open-ended to enable asking many questions on the same topic in order to discover relevant information about their social network.

This set includes questions about the sociality of the respondent where he/she has been asked about the names his/her social network member. For each member, he/she must state the relationship strength, frequency of interaction and the most popular methodology of communication. These questions examine the social status of each respondent as social status can affect knowledge sharing [94].

- **Sharing information**: This set of questions is the scaled type in order to obtain a clear picture about the popularity of their preferences. Each respondent states his/her preferences in knowing and sharing specific types of information taking into account the observation that an item relevant to one person is more likely to be of interest to individuals in the same social group [99].

Question Type	Uses	Advantages	Disadvantages	Examples
Open-ended (essay or short-answer)	• Discover relevant issues • Obtain a full range of responses • Explore respondents' views in-depth	• Identifies issues most relevant to respondents • Generates new ideas about topic • Clarifies respondents' positions • Provides detail and depth	• Requires more time, thought, and communication skill to complete • Requires time-consuming data entry • May generate incomplete or irrelevant data • Complicates data summary and analysis	1. Describe the steps you took to prepare for your last exam. 2. What did you enjoy most about this course?
Close-ended (multiple-choice or yes/no)	• Ask many questions in a short time period • Assess learning or attitudes when issues are clear • Measure knowledge or ability	• Fast and easy to complete • Enables automated data entry • Facilitates data analysis and summary of data	• Limits response options • May omit a preferred answer • Requires moderate knowledge of the topic to write appropriate questions and responses • Lacks detail and depth	1. Which aspect of the course do you feel is most effective (mutually exclusive)? a. Lecturing by instructor b. In-class interactive exercises c. Assigned readings d. In-class videos 2. Which aspect of the course is effective (not mutually exclusive)? a. Lecturing by instructor b. In-class interactive exercises c. Assigned readings d. In-class videos
Partial open-ended (multiple-choice with 'other' option)	• Ask many questions in a short time period • Assess learning or attitudes when issues are clear and identifiable • Discover relevant issues	• Enables respondents to create their own response if choices do not represent their preferred response • Generates new ideas about topic • Fast and easy to complete	• Requires moderate knowledge of the topic to write appropriate questions and responses • Lacks detail and depth • Complicates data analysis and summary	Which aspect of the course do you feel is most effective? a. Lecturing by instructor b. In-class interactive exercises c. Assigned readings d. In-class videos e. Other (specify)
Scaled [AT-Planning-Method-AT Survey Overall&Planning-Survey Response Scales]	• Determine the degree of a response, opinion, or position	• Provides a more precise measure than yes/no or true/false items • Fast and easy to complete • Enables automated data entry	• Requires moderate knowledge of the topic to write appropriate questions	Re-reading the text improves my performance on exams. a. Strongly agree b. Agree c. Neutral d. Disagree e. Strongly disagree
Ranking	• Determine the relative importance to respondents of various options • Choose among various options	• Allows respondents to indicate the relative importance of choices • Enables automated data entry	• More difficult to answer • Limits number of response options • May omit a respondent's preferred answer	Rank the following activities in this course by how engaging you found them to be (1 = the most engaging) __Reading the textbook __Listening to the instructor lecture __Watching videotapes in class __Writing the term paper

Figure 4.2: Types of Survey Questions (source: Instructional Assessment Resources [75]).

By the end of this stage, it was required that the flow of survey questions would be reasonable and in the right order. The survey contained understandable questions specifically targeted at undergraduate students. The survey had a clear instructions for participants to help them to understand the required information and easily complete the survey, as well as to ensure that the questionnaire did not omit any information or have any redundancy.

4.2.3 Survey Implementation

In order to fulfil the aim of the survey, three issues were proposed to be asked to the participants, such as:

- **Method of sharing information**: This sub-category combines three sub-categories from Figure 4.1 which are different methodologies of sending/receiving information, and shared information types. This combination minimises the set of survey questions to be a minimal set that serves the survey aims.
- **Presence in the building**: to gain information about participants' preferences of being alone or in groups.
- **Actual social network**:This sub-category combines three different sub-categories, namely the frequency of sharing information, with whom they share (relationship strength), and sharing methodologies. This combination builds a clear picture of the participants' social network members that includes the most popular communication methodology with each member, the frequency of interaction, and the relationship strength.

An electronic survey was conducted on higher education undergraduate first year students of Cardiff School of Computer Science and Informatics 2011/2012. 119 out of 130 participants completed the survey. Eleven participants were dropped from the study as they did not complete the on-line survey at all. In the final sample of 119 participants (16 female and 103 male), two participants did not fill out the information about their networks and one participant gave an incorrect email. Therefore, the actual number of responses included in this study is 116 participants.

Convincing students to participate was a challenge. Following the University regulations, they were informed that all the collected information would be anonymous before applying any analysis. All the participants' personal information was anonymous and confidential and only the anonymous version has been used in this study.

A trial implementation was conducted on first year undergraduates in the academic year 2010/2011 in the Cardiff School of Computer Science. All the collected data was confidentially analysed using different social network analysis metrics. We sought to determine the role of online communication in the physical social network, in particular the role that online communication plays in augmenting face-to-face interactions. These observational results are ready to be published in a journal paper. As it was challenge to convince the participants to take part in the experimental study as part of our research, all the observational results are not included in this thesis.

This trial implementation aided us in producing a new version of the survey as we found that some questions could be combined in one question. As a result, a new version of the survey was conducted focusing only on the points that served the aims of the study. Figure 4.3 shows the new issues of the survey that were conducted with the first year undergraduates 2011/2012 of Cardiff School of Computer Science. The full version of the updated survey is included in Appendix A.1.

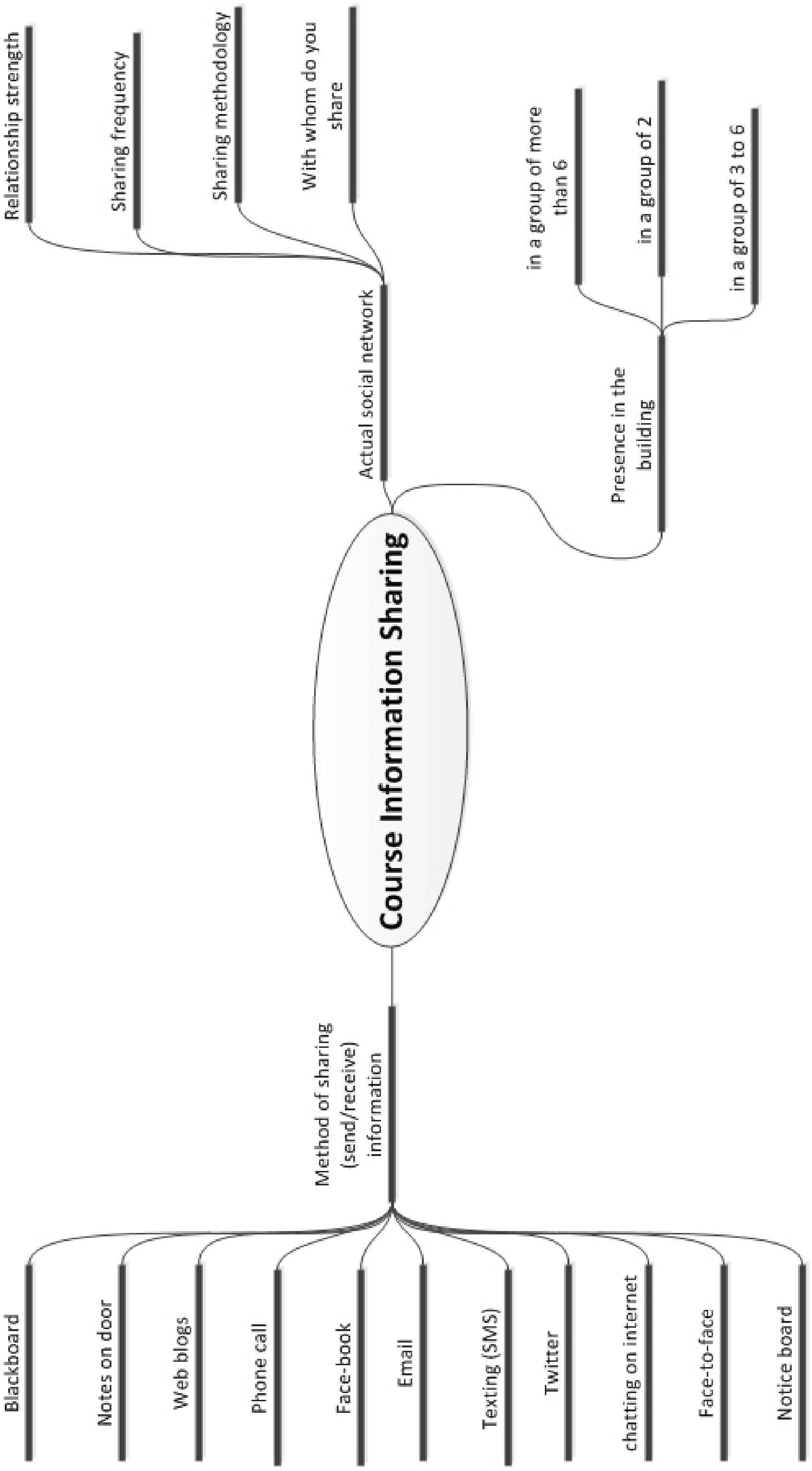


Figure 4.3: Issues of Course Information Sharing.

4.3 Summary

In this chapter, we used the electronic survey as a research methodology to uncover the social network for a group of participants. It was required to understand the participants' social network from their perceptions. This chapter mainly focused on the design and the implementation of the survey in order to collect the ego-centric social information about the participants.

Different stages of the design the survey were followed in order to design a good survey that satisfied its aim. The survey aimed to explore how interactions are sustained between individuals and the extent to which online interactions and electronic media (as well as face-to-face interactions) adapt and support social networks. A concrete picture of students' social networks was built by examining the physical and online interactions that support relations. The survey was piloted on a group of postgraduate students before it was conducted with the participants. Many issues were considered during the survey design, such as question wording, and the order of the questions being rearranged into a meaningful order.

The survey was applied in two case studies: a trial implementation on first year undergraduates 2009/2010 and an implementation on first year undergraduates 2010/2011 in the School of Computer Science and Informatics. The results of the trial implementation study were analysed and integrated into a journal paper where the results of the second case study have been analysed and integrated in this research. A brief summary about the collected data was outlined. The collected data from the survey revealed different types of social networks (physical interaction network, online communication networks, relationship strength networks) that supported the analysis of their social behaviour from participants' own perceptions. These social networks will be compared with the opportunistic ones to address the thesis hypothesis.

Chapter 5

Detecting Opportunistic Networks

Overview

As detailed in Chapter Three, an indoor mobility tracking system was developed to collect mobility and physical interaction data. This allows us to approximate the opportunistic network for our case study of participants. A detailed set-up process for the experimental study was discussed in Chapter Three.

This chapter starts with the experimental schedule in Section 5.1. The mathematical representation of the experimental data set is presented in Section 5.2. In Section 5.3 a statistical description of the collected data is presented. Different algorithms developed to extract the different opportunistic networks (in terms of trajectory, duration, and co-location) are outlined in Section 5.4. Section 5.5 describes the characteristics of the extracted opportunistic networks by using some social network analysis metrics. The chapter ends with a summary of the key findings of the analysis.

5.1 Experimental Schedule

The empirical study was carried out on 23rd April 2012 and lasted for three weeks. This period of time was chosen based on the arrival of the Bluetooth mote devices. In addition, it was the end of the second semester when participants' presence would be regular for exam revision and coursework submission. The experiment timing was divided into six slots where each slot had only two days. The idea of dividing the experiment period into two day time slots was to overcome the limitation of the wireless devices' battery lifetime. It was found that the lifetime of the battery was, on average, two days and five hours. As a result, it was agreed that the time period would be two days before changing the batteries.

On Monday and Thursday of each week, participants collected the devices at 8:30 am in the morning. On Tuesday and Friday, the participants submitted the devices at 6:00 pm in the evening to change the batteries. Wednesday was not included in the experiment as the participants have this day off. In order to encourage participants to participate in this practical study, a prize was proposed for a winner of a draw. In addition, guidance notes Appendix A.3 were provided to identify the rules for participants who were happy to be monitored.

24 participants out of 119 students took part in this study. The participants took part willingly. On completion of the practical study, we had 77 completed slots from different participants as not all students completed all of the experiment's slots. We had 1609 logging events for mobility trajectory over the whole period. 2578 system files were received from different Bluetooth motes by different servers inside the school.

5.2 Data Representation

In order to understand the collected data sets, it is important to include the data representation. This representation provides a clear definition for each key concept in the collected data set.

Definition 1

P is a set of m participants where $p_i \in P$ and $1 \leq i \leq m$ Participants refers to the users who took part in this empirical study.

Definition 2

L is a set of K Locations where $L_k \in L$ and $1 \leq k \leq K$.

Generally, location here denotes the room level range of each beacon of the fixed place beacons network that are distributed throughout the school buildings. For more detailed information about different locations throughout the school, see Chapter Three.

Definition 3

A *participant's time interval* is a pair $[t_A^{p_iL_k}, t_B^{p_iL_k})$ where $t_A^{p_iL_k}$ & $t_B^{p_iL_k}$ are times between which participant p_i is continuously detected at location L_k . By this definition, $t_A^{p_iL_k} \leq t_B^{p_iL_k}$.

Definition 4

A pair of participant's time intervals $[t_A^{p_iL_k}, t_B^{p_iL_k})\cap$ $[t_C^{p_iL_k}, t_D^{p_iL_k}) = 0$ is disjoint if $t_B^{p_iL_k} \leq t_C^{p_iL_k}$ and $t_A^{p_iL_k} < t_C^{p_iL_k}$

Definition 5

For a participant's time interval $[t_A^{p_iL_k}, t_B^{p_iL_k})$ at a location L_k let t_s & t_e denote two points in time where $t_s \leq t_e$ and

$Du_{t_st_e}(t_A^{p_iL_k}, t_B^{p_iL_k}) = |[t_s, t_e) \cap [t_A^{p_iL_k}, t_B^{p_iL_k})|$

where $Du_{t_st_e}(t_A^{p_iL_k}, t_B^{p_iL_k})$ indicates the *duration* that a participant p_i spends at Location L_k between t_s & t_e.

Definition 6

Let t_s, t_e denote two points in time where $t_s \leq t_e$. The *location function*

$$f_{L_kp_i}(t_s, t_e) = \begin{cases} 1 & if Du_{t_st_e}(t_A^{p_iL_k}, t_B^{p_iL_k}) \geq 0 \\ 0 & \text{Otherwise} \end{cases}$$

Thus $f_{L_kp_i}$ indicates the presence (or not) of p_i at L_k between t_s & t_e

Definition 7

Consider time intervals $([t_1, t_2), [t_3, t_4), \cdots, [t_{n-1}, t_n))$ where $t_i \leq t_{i+1}$. Let p_i be a participant at a location L_k. The *Trajectory vector* is defined as follows:

$TV_{p_iL_k}((t_1, t_2), \cdots, (t_{n-1}, t_n)) = (f_{L_kp_i}(t_1, t_2), \cdots, f_{L_kp_i}(t_{n-1}, t_n))$

Definition 8

Consider time intervals $([t_1, t_2), [t_3, t_4), \cdots, [t_{n-1}, t_n))$ where $t_i \leq t_{i+1}$. Let p_i be a participant at a location L_k. The *Duration vector* is defined as follows:

$DV_{p_iL_k}((t_1, t_2), \cdots, (t_{n-1}, t_n)) = (Du_{t_1t_2}(t_1^{p_iL_k}, t_2^{p_iL_k}), \cdots, Du_{t_{n-1}t_n}(t_{n-1}^{p_iL_k}, t_n^{p_iL_k}))$

Definition 9

Consider t_s,t_e denote two points in time, let p_i, p_j be two participants where $i \neq j$ and $1 \leq i, j \leq m$. The *trajectory opportunistic vector* is defined as follows:

$$S_{p_ip_j}(t_s, t_e) = \begin{cases} 1 & if \exists k : TV_{p_iL_k}(t_s, t_e).TV_{p_jL_k}((t_s, t_e) > 0 \\ 0 & \text{Otherwise} \end{cases}$$

Definition 10

Consider t_s,t_e denote two points in time, let p_i, p_j be two participants where $i \neq j$ and $1 \leq i, j \leq m$. The *duration opportunistic vector* is defined as follows:

$$\hat{S}_{p_ip_j}(t_s, t_e) = \begin{cases} Du_{t_st_e}(t_A^{p_iL_k}, t_B^{p_iL_k}) \cap Du_{t_st_e}(t_A^{p_jL_k}, t_B^{p_jL_k})) & if \exists k : Du_{t_st_e}(t_A^{p_iL_k}, t_B^{p_iL_k}) \\ & \& Du_{t_st_e}(t_A^{p_jL_k}, t_B^{p_jL_k}) > 0 \\ 0 & \text{Otherwise} \end{cases}$$

Definition 11

The number of times p_i & p_j meet together is called the *frequency*, denoting: $Fr_{p_ip_j}$.

In terms of the physical interaction between participants, we counted how many times they met (frequency of meeting) to be able to infer the strength of their relationship.

Definition 12 For a pair of participants p_i & p_j *the co-located opportunistic vector*, denote:

$$S_C(Fr_{p_ip_j}) = \begin{cases} Fr_{p_ip_j} & ifFr_{p_ip_j} > 0 \\ 0 & \text{Otherwise} \end{cases}$$

Example: A sample is given in Figure 5.1 to demonstrate the described definitions. As seen, we have three participants p_1, p_2, p_3 who have been detected at four different locations $\{L_1, L_2, L_3, L_5\}$. For each participant, time interval blocks represent his/her route throughout the school buildings. This is shown in Figure 5.1. From *Definition 3* the time that is assigned to each block is called the participant's time interval.

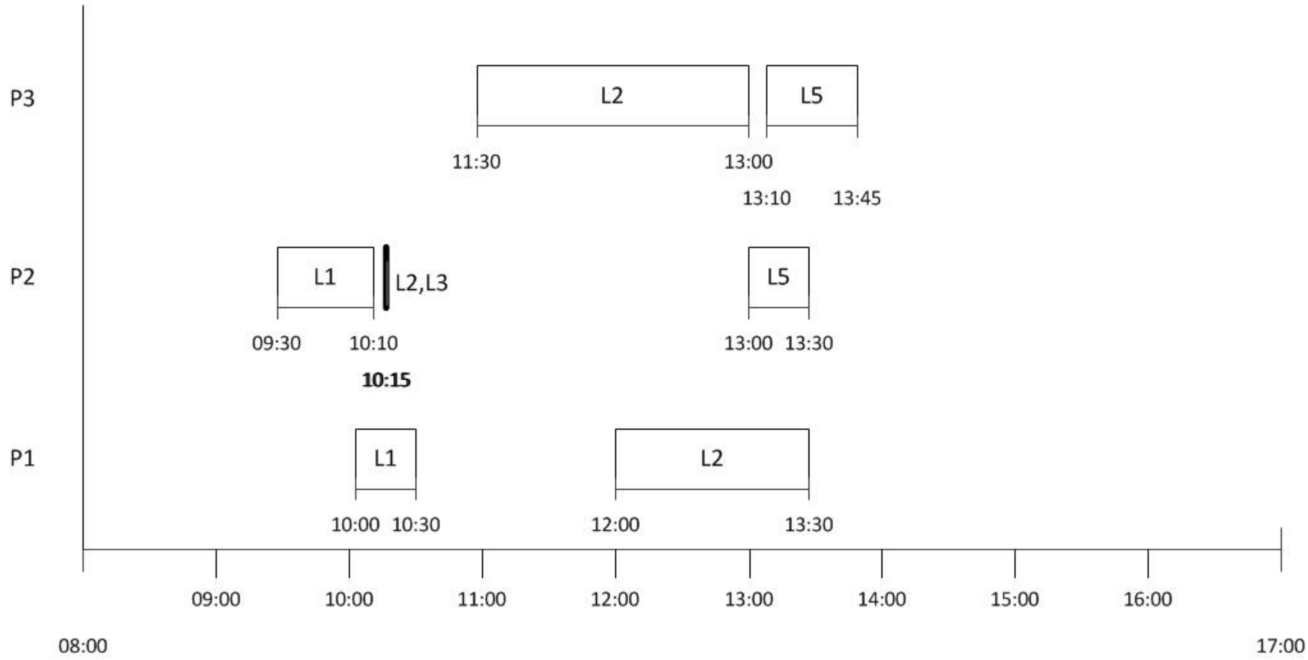


Figure 5.1: Simple Example.

From *Definition 5*, the *duration* for participant p_1 at location L_1 when $[t_s, t_e) = [10:00, 11:00)$ equals 1800 seconds. *From Definition 6*, the *location function* for the same participant equals 1. From *Definition 7*, if we proposed the time intervals to be [8:00,9:00) [9:00,10:00) [10:00,11:00) [11:00,12:00) [12:00,13:00) [13:00,14:00) then the *Trajectory vector* for participant p_1 at location L_1 equals(0,1,1,0,0,0). *From Definition 8*, *Duration vector* equals (0,0,1800,0,3600,1800)

At *Definition 9*, we need to calculate the *trajectory opportunistic vector* between two participants so we need to consider p_1, p_2 from the example in Figure 5.1. For each participant, we need to calculate the trajectory vector (*from definition 7*)

$TV_{p_1L_1}([8:00, 9:00), [9:00, 10:00), [10:00, 11:00), [11:00, 12:00), [12:00, 13:00), [13:00, 14:00)) = (0, 1, 1, 0, 0, 0)$

$TV_{p_2L_1}([8:00, 9:00), [9:00, 10:00), [10:00, 11:00), [11:00, 12:00), [12:00, 13:00), [13:00, 14:00)) = (0, 1, 1, 0, 0, 0)$

then $[S_{p_1p_2}(8:00, 9:00) = 0$

$[S_{p_1p_2}(9:00, 10:00) = 1$

$[S_{p_1p_2}(10:00, 11:00) = 1$

$[S_{p_1p_2}(11:00, 12:00) = 0$

It is straightforward to calculate other calculations on different participants in the example.

5.3 Description of Mobility Tracking Data

This section focuses on the descriptive statistical analysis of the collected data set. The aim of this analysis is to provide a clear picture about the collected data before analysing it. Having a clear picture supports the understanding of the nature of the data and assists the analysis of the extracted opportunistic network.

5.3.1 Mobility Trajectory Profile

This type of data set represents the route of different users throughout the school buildings. For each route, detailed information was provided that includes: date and time of being at a specific location (for more detailed information - see Chapter Three). Generally, this data set was highlighted to be able to extract the opportunistic network between different users. In order to analyse these data sets, they were visualized from different points of view as described in the following subsections. In short, 1609 mobility time intervals were collected during the experiment time slots. The time interval represents the presence of the participants for a continuous time period at the same place. In total, the mobility trajectories were collected for 160 hours through the experiment time slots.

User routine

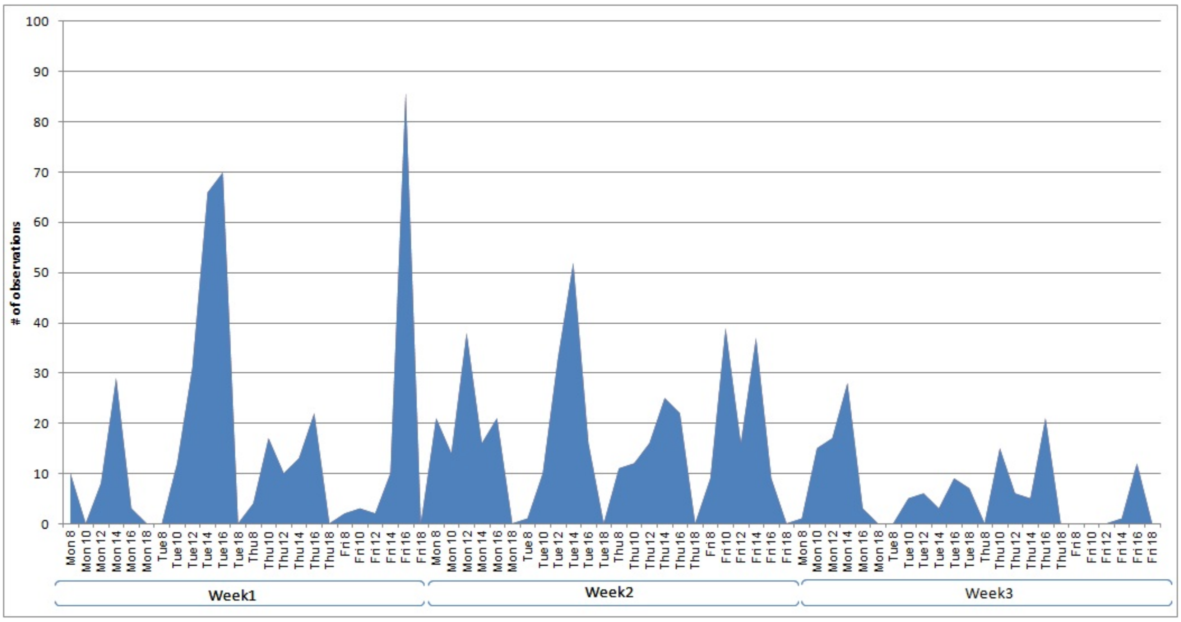


Figure 5.2: Sum of Location Function per Day.

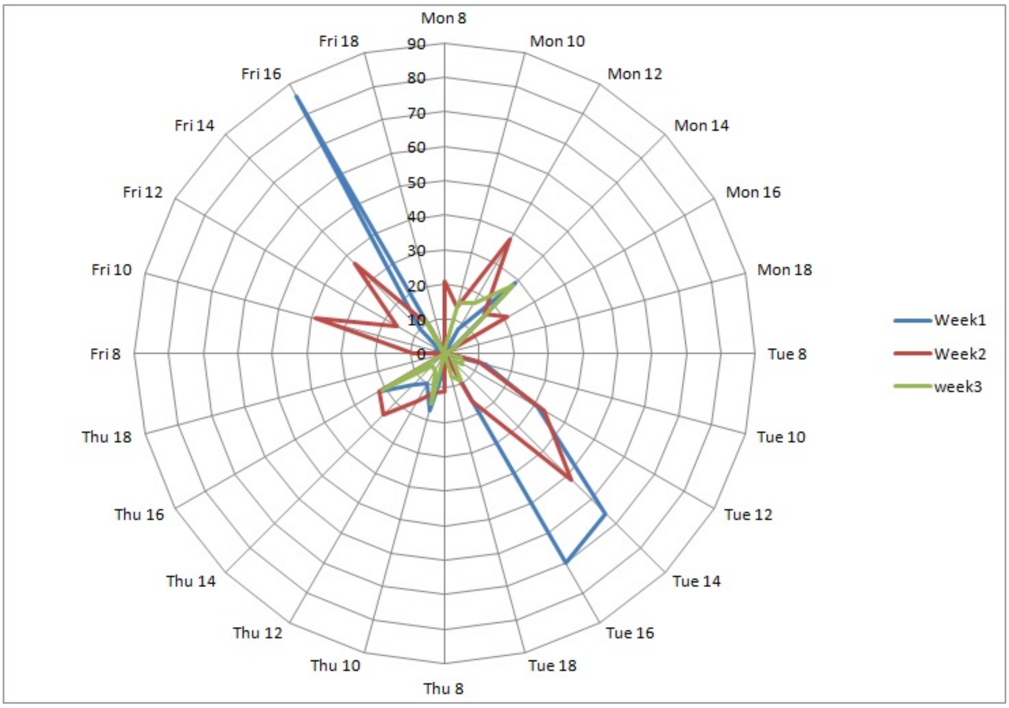


Figure 5.3: Radar Plot of Sum of Location Function per Day.

In this empirical study, the tracking of the participants started at 8:00 am and ended at 6:00 pm. This time periods reflects the existence of the participants in the school building according to the school time table. Figure 5.2 collectively presents the different users' routines by summing all location functions (*Definition 6*) per day (i.e. $\sum_{P,L} f_{L_k p_i}(t_s, t_e)$). Although the first two weeks were normal study weeks, there was no specific routine that the participants followed in their routine during their time at school. It was found that Tuesday and Friday in the first two weeks gained significant amounts of activity rather than week three. Mainly, participants were active from 10:00 am until 4:00 pm.

Figure 5.3 provides a different representation of users' activities during every time slot in the experiment. The radar representation was used to represent the results as it represents periodic information about the users' activities. It is very hard to represent periodic data using a linear plot as there would be a cutting point from the x-axis. This would affect the perception of various trends. However, it would also make the process of establishing any common pattern in their activities easier.

Venue Popularity

In this study, five different venues were strategically chosen as access points to monitor the participants' mobility. Where L_1 covers the one class room, lab and the south entrance. L_2 covers two labs in the central building. L_3 covers the North building the school office, and part of the bridge between North building and Trevitech building. L_4 covers the library at second floor of the Trevitech building. L_5 covers the Trevitech entrance and the refractory. The topology of these venues is declared in Chapter Three. Regarding venue popularity, the participants were able to access different venues during all time slots of the experiments that were aggregated $\sum_{P,T} f_{L_k p_i}(t_s, t_e)$ as seen in Figure 5.4. It is clear that participants' activity mainly focuses on $L1$ and $L2$ as they cover the most important class rooms and labs. In addition, $L4$ did not gain a lot of access except for the last two days as this week was a revision week where some participants preferred to study in the Library.

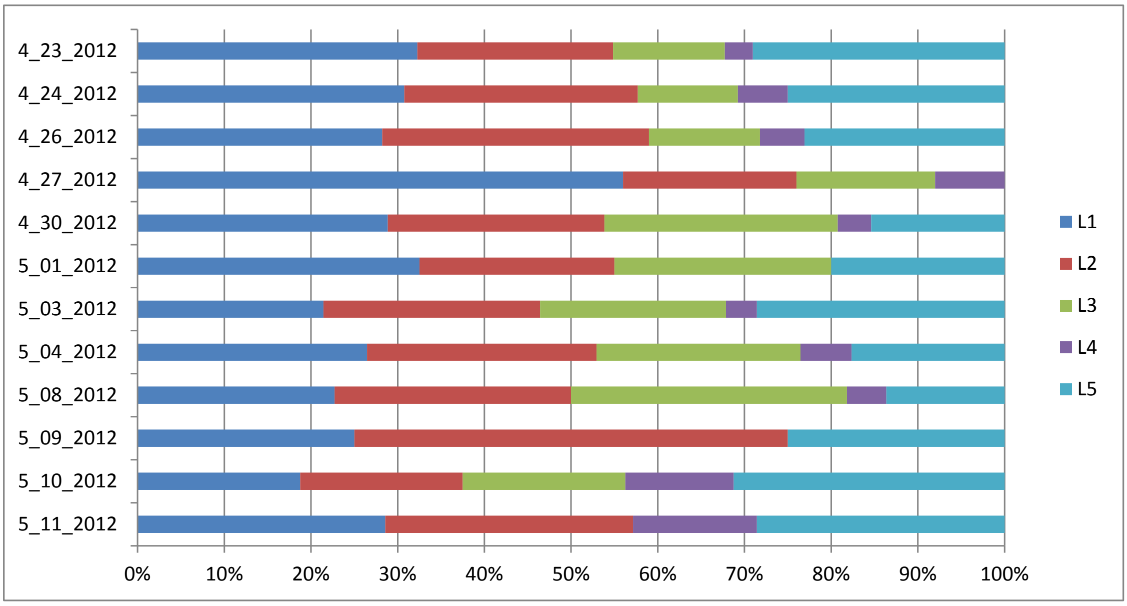


Figure 5.4: Venue Popularity Concerning Sum of Location Function per User.

Another visualization of the venues' popularity in terms of trajectories (i.e *Definition 7*) and durations (*Definition 8*) is seen respectively in Figures 5.5 and 5.6. The former presents the popularity in terms of the number of trajectories for each venue (i.e $\sum_{P,T} TV_{p_i L_k}((t_1, t_2), \cdots, (t_{n-1}, t_n))$) where the latter presents the popularity in terms of the total time intervals have been spent at each venue (therefore $\sum_{P,T} DV_{p_i L_k}((t_1, t_2), \cdots, (t_{n-1}, t_n))$). Although the calculations are different, the same result are produced so that L_1, L_2 are the most important (popular) venues to the participants.

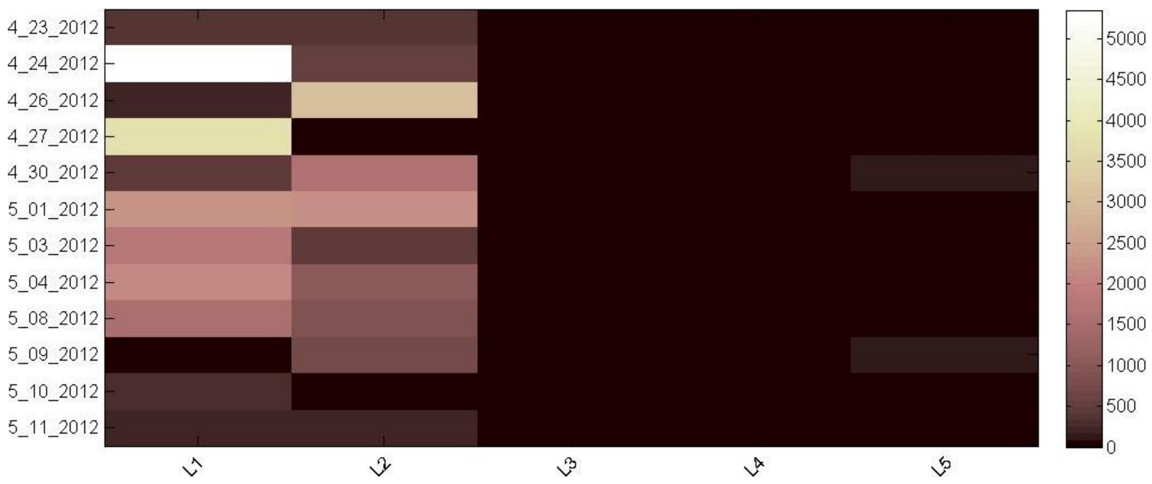


Figure 5.5: Venue Popularity in Terms of Trajectories .

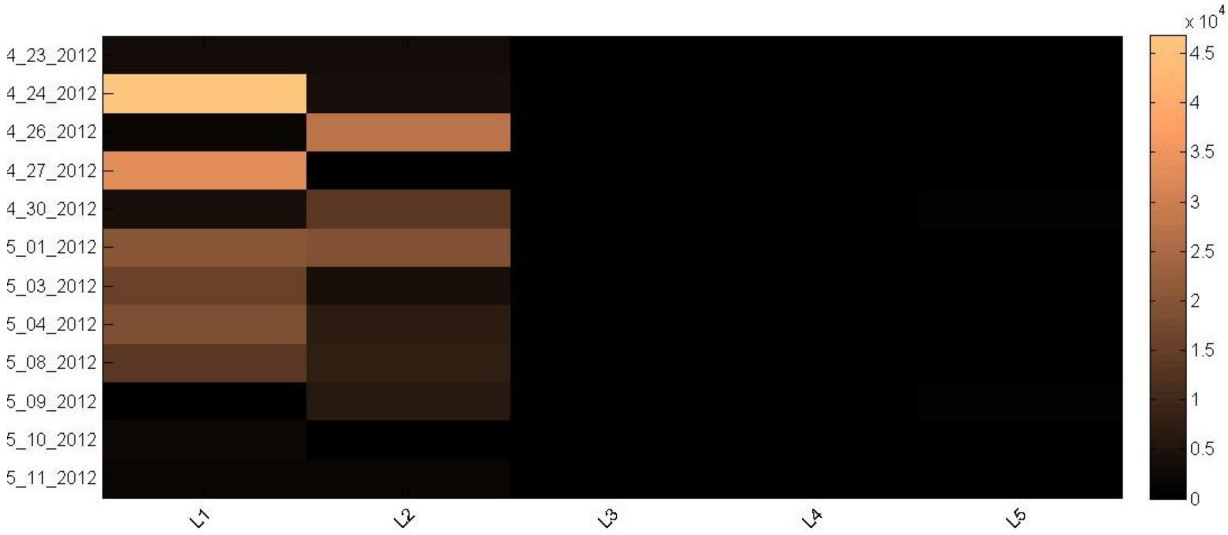


Figure 5.6: Venue Popularity in Terms of Durations .

Additionally, it has been found that the patterns seen in both Figures 5.5 and 5.6 are consistent with the occurrence of events for students.

5.3.2 Physical Interaction Profile

As mentioned in Chapter Three, the physical interaction file (system file) contains a number of hash codes. Each hash code represents the unique identification code for each wireless device (Bluetooth mote). This code was used instead of using the Bluetooth MAC address to overcome the limitation of the system file size of the Bluetooth mote. 2578 files were collected, with each file containing many codes with a maximum of 32 KB. Due to the features of this data set as well as the limitation of Bluetooth mote, it did not support the time of meeting, and so it is hard to visualize the data. As a result, the following subsection discusses the probability of missing encounters of the data set where encounter is the case where the Bluetooth devices become in range of each other (without any interaction) so that it can be detect by each other.

Probability of Missing Encounters

As the Bluetooth devices did not run simultaneously, there had to be missing data in recording some of the encounters. The following discussion calculates and clarifies the probability of missing encounters.

Let:
p_s - the probability an encounter is not detected by a single device.
N_b - number of encounters observed by both devices.
N_s - number of encounter observed by single device.
N_m - number missed (observed by neither).

Hence the probability an encounter is detected by a single device is $(1 - p_s)$. As a result, the expected number of encounters observed by **both devices** $Ex[N_b]$ equals the total number of observed and missed encounters multiplied by **the square** of probability an encounter is detected by a single device

$$Ex[N_b] = (N_b + N_s + N_m).(1 - p_s)(1 - p_s) \tag{5.1}$$

The expected number of encounters detected by a single device $Ex[N_s]$ is the total number of observed and missed multiplied by the probability of an encounter not detected by an single device times the probability of an encounter detected by a single device.

$$Ex[N_s] = (N_b + N_s + N_m).2.p_s.(1 - p_s) \tag{5.2}$$

From the experimental data,

$N_b = 160$, $N_s = 52$ hence (5.1) and (5.2) become

$160 = (160 + 52 + N_m).(1 - p_s)^2$

$52 = (160 + 52 + N_m).2.p_s(1 - p_s)$ using Mathematics to solve for p, N_m, gives

p_s = 0.14 N_m = 4.225

Hence, it is important to remember in the following analysis that around four edges may be missing observed by either. That is quite small for the total number of observed encounters.

5.4 The Extraction of Different Opportunistic Networks

This section focuses on the algorithms that were developed to extract different opportunistic networks from the mobility tracking data set. Firstly, the trajectory and the duration opportunistic networks were extracted from the mobility tracking data followed by the co-located opportunistic network that was extracted from physical interaction data. The trajectory opportunistic network is the opportunistic network that represents the links between the participants in terms of the frequency of being at the same location at the same time. The duration opportunistic network represents the links between participants in terms of the duration of time of being at the same location. The co-located opportunistic network represents the links between participants in terms of the frequency of physical interaction between them. The following three algorithms have been implemented using MATLAB and the resulted network stored on EXCEL sheets as adjacency matrices.

5.4.1 Trajectory and Duration Opportunistic Network Algorithms

This section discusses the developed procedure to extract two different opportunistic networks (trajectory and duration). The basic assumption for this procedure is that any overlapping between the participants' time interval at the same location within a 2-3 metre range on the same day means that they potentially have an interaction. We note that this is an approximate measure but this contains sufficient accuracy for this study.

Based on this assumption, two types of opportunistic networks were extracted, as seen in *Algorithm* 6. One represents a trajectory opportunistic network and the other represents the duration opportunistic network. Both of the matrices are square matrix 24×24. In terms of the duration matrix, each entity represents the joint duration that this pair of participants spent at a specific location on a specific day. As regards the trajectory matrix, each entity represents the frequency of being at the same location between this pair of participants. As a result, an edge has been added between these two participants. The strength of the relationship or the weight of the edge has been updated according to the frequency of participants overlapping each other.

To extract these opportunistic networks *Algorithm* 5 and *Algorithm* 6 were developed. Firstly, *Algorithm* 5 calculates the trajectory vector and duration vector for each participant at each location on a specific day, as defined in *definition 7,8*. The trajectory vector $TV_{p_iL_k}((t_1,t_2),\cdots,(t_{n-1},t_n))$ represents the presence of participants p_i at location L_k during time interval $((t_1,t_2),\cdots,(t_{n-1},t_n))$. As shown in *definition 7*, the trajectory vector is a vector of the location function where the location function represents the existence of the participant p_i in location L_k at time $[t_s,t_e)$ (see definition 6).

The duration vector $DV_{p_iL_k}((t_1,t_2),\cdots,(t_{n-1},t_n))$ represents the time duration for a participant p_i in location L_k during time interval $((t_1,t_2),\cdots,(t_{n-1},t_n))$ (see *definition 8*). Each entity in the duration vector is a duration value that is defined in *definition 5*. In brief, *Algorithm* 5 calculates these two vectors by firstly reading all of the tracking data that includes the trajectory time intervals for each participant. For each given time interval, the overlapped tracking time interval is found for participant p_i at location L_k. The trajectory vector for each given time interval equals one if the participant was found at location L_k during this time interval and zero otherwise. The duration vector for each given time interval is the overlapped duration between the given time interval and the tracking time interval.

Algorithm 5 Trajectory/Duration Vector Function

Input $p_i, L_k, Date, t_S, t_E, delta$
  $[Eparticipants, EDates, ELocations, [t_A, t_B)]$ =read (Experimental data set)
  $j = 0$;
  $counter = 0$;
  $t_s = t_S$;
  $t_e = t_s + \delta$;
  while $t_s < t_e$ **do**
    $j = j + 1$;
    $counter = 0$;
    for $i = 1$ to size (t_A, t_B) **do**
      if (p_i = Eparticipants(i))& (L_k = ELocations(i)) & (Date = EDate(i)) **then**
        if $(t_A(i) > t_e || t_B(i) < t_s$ **then**
          $counter = counter$
        else
          $DV(j) = [t_A(i), t_B(i)) \cap [t_s, t_e)$
          $counter = counter + 1$
        end if
      end if
    end for
    $t_s = t_e + 1$;
    $t_e = t_s + \delta$;
    if $counter > 0$ **then**
      $TV(j) = 1$;
    else
      $TV(j) = 0$;
    end if
  end while
  return TV, DV

In order to extract the duration opportunistic matrix and the trajectory opportunistic matrix, *Algorithm* 6 was applied. For each participant, the duration vector DV and trajectory vector TV were calculated for each location L and on each day. Then, all the participants' trajectory vectors were combined to formulate the trajectory matrix and all participants' duration vectors were combined to formulate duration matrix. For each pair of participants, we compared their trajectory matrices to extract the joint (or common) intervals and summed all the joint intervals together to formulate an entity of the trajectory opportunistic matrix of this pair of participants. This step was repeated for all pairs to formulate the whole trajectory opportunistic matrix. For the duration matrix, the same two steps were applied and repeated to calculate the duration opportunistic matrix.

Algorithm 6 Trajectory/Duration Opportunistic Matrix Algorithm

Input $P, L, Dates, t_S, t_E$
  $DVmatrix = []$; {this matrix contains the duration vector for all participant at specific location and date for all time intervals}
  $TVmatrix = []$; {this matrix contains the trajectory vector for all participant at specific location and date for all time intervals}
  for j=1 : size(Dates) **do**
    for k = 1 : size(L) **do**
      for i = 1 : size(P) **do**
        [TV, DV]= Trajectory/Duration vector($P(i)$,$L(k)$,Dates(j),t_S, t_E) { Call the function in Algorithm 5}
      end for
      $DVmatrix$ = [$DVmatrix$, DV];
      $TVmatrix$ = [$TVmatrix$,TV]
      for ii = 1: size($DVmatrix$) **do**
        for jj = ii+1 : size($DVmatrix$) **do**
          if There is overlapping between participants **then**
            $\hat{S}matrix(ii, jj)$ += overrlapping period {as defined in definition 10}
          end if
        end for
      end for
      for ii = 1: size($TVmatrix$) **do**
        for jj = ii+1 : size($TVmatrix$) **do**
          if There is overlapping between participants **then**
            $Smatrix(ii, jj)$ += overrlapping trajectory {as defined in Definition 9}
          end if
        end for
      end for
    end for
  end for
  return $Smatrix, \hat{S}matrix$

5.4.2 Algorithm The co-located Opportunistic Network

This section focuses on extracting the co-located opportunistic network from the recorded system files (as discussed in Chapter Three). In order to extract the co-located opportunistic network *Algorithm* 7 was developed. For each recorded system file, it finds out which Bluetooth device (BluetoothID) recorded and reads the file content. Then, for each entity in the current system file, it increases by one on the entity of the opportunistic matrix that matches BluetoothID as a column and file entity as a raw. This step is repeated until the end of collected files.

Algorithm 7 Physical Interaction Opportunistic Matrix Algorithm

---

Input All recorded system file ,BlutoothID
  $S_C matrix$ =[];
  for each system file in the directory) **do**
    for i = 1 : size(BluetoothID) **do**
      M= load (current-file)
      if current-file-name = BluetoothID(i) **then**
        for j = 1: size(current-file **do**
          if M(j) = BluettothID(i) **then**
            co-located-matrix(i,M(j)) += 1 {as defined in Definition 12, count the number of frequency of meeting}
          end if
        end for
      end if
    end for
  end for
  return $S_C matrix$

5.4.3 Experimental Networks Adjacency Matrices

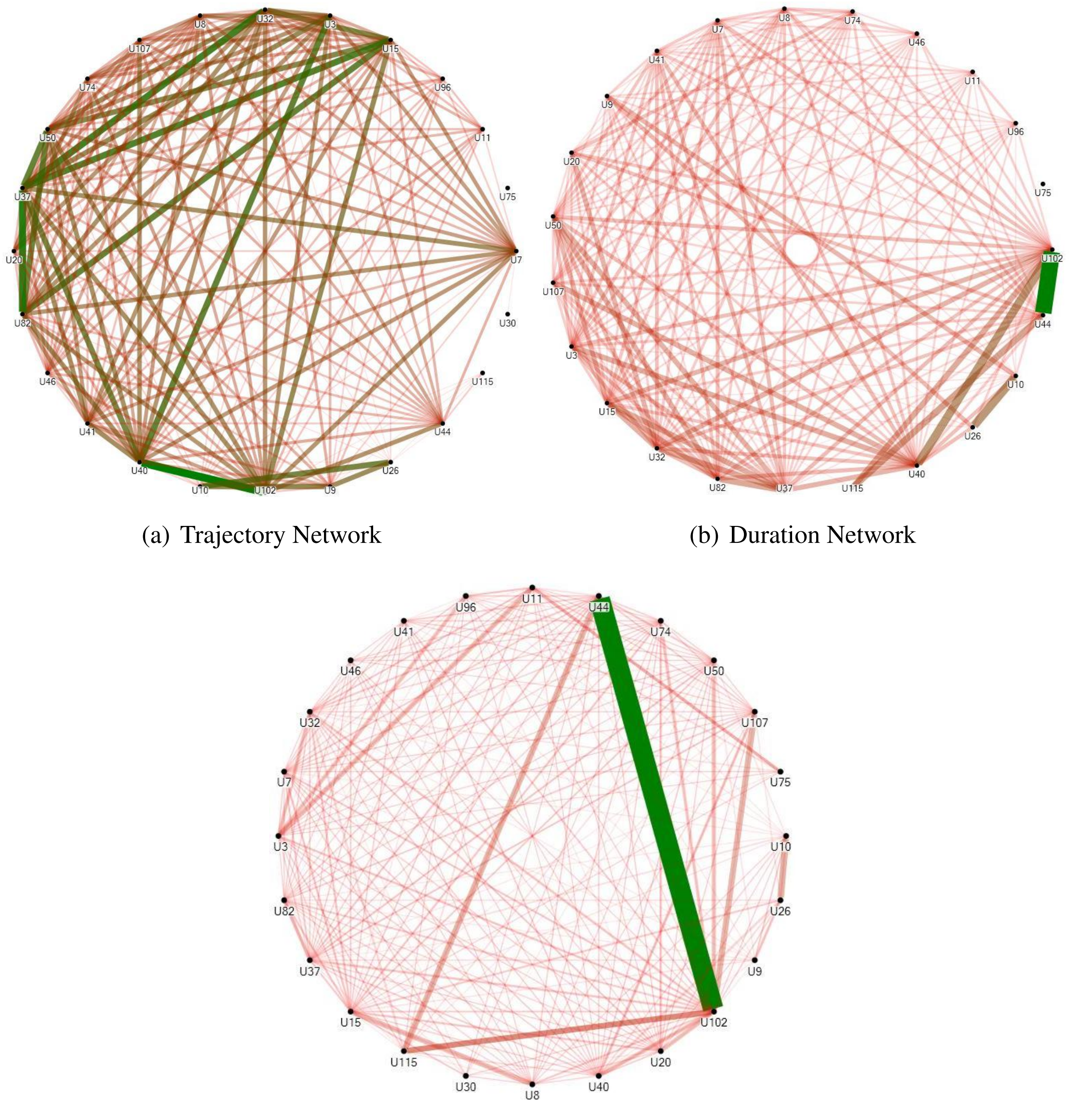
From previous algorithms, we have three opportunistic networks that are extracted from the experimental study. The adjacency matrix for each network was extracted to simplify the analysis of networks. We have cell co-location(trajectory), device co-location (physical interaction) and duration adjacency matrices. Cell co-location represents the joint presence frequency of each pair of participants in the building throughout the experimental study. This matrix is the aggregation of the joint presence frequency of every pair of participants at five different locations in school buildings (see Chapter Three). Each location in the building has its own matrix that represents the joint presence frequency for every pair of participants. These frequencies are calculated from the mobility tracking records in the central database. First, for each participant, the trajectory vector was extracted, as seen in Chapter Five. Secondly, multiple vectors of the participant trajectories were combined together to formulate the trajectory matrix for each participant. Thirdly, for each pair of participants', joint trajectories were extracted and aggregated to formulate the trajectory opportunistic network for each location individually, where an aggregation of all the locations is calculated to form a complete picture about the common trajectories between participants and this is called *cell co-location* (see AppendixB).

The same procedure was followed to extract the duration opportunistic network that represents the joint presence duration at specific locations between every pair of participants. Then, an aggregation was applied to have a clear picture of the duration that every pair of participants spend together (see Appendix B).

An algorithm was developed in order to produce the device co-location adjacency matrix that represents the frequency of physical interaction between every pair of participants. In brief, this matrix was formulated from the system files that were sent from the Bluetooth mote to the server(see Chapter Three). Each file is identified by the MAC address of the producer Bluetooth mote. It includes the hash code for all Bluetooth mote devices that come within range of this device before being in range with the receiving server and sending the system file. For each file, the frequency of the detected devices that come within range is calculated and aggregated with other frequencies on different days between the same pair of Bluetooth motes. A square matrix 24×24 is produced where each entity presents the overall frequency of interaction between this pair for the overall experiment slots.

5.5 Characteristics of Extracted Opportunistic Networks

The undirected network $H_B = (V, E)$ presents different extracted opportunistic networks that result from applying previous algorithms, where V is the set of participants. An edge $(i, j) \in E$ exists if and only if node i at least being in the same location or being recorded by node j, where $B \subset \{b_1, b_2, b_3\}$. The trajectory opportunistic matrix is stored in a graph H_{b1} and the duration opportunistic matrix is stored in a graph H_{b2}. Co-located opportunistic matrix is stored in a graph H_{b3} as seen in Figure 5.7. These matrices can be seen in Appendix B. Edges between the participants were weighted where the edge colours, widths and opacities are based on frequency of meeting values or duration of being together.



(a) Trajectory Network

(b) Duration Network

(c) Co-located Network

Figure 5.7: Visualization of Different Opportunistic Network.

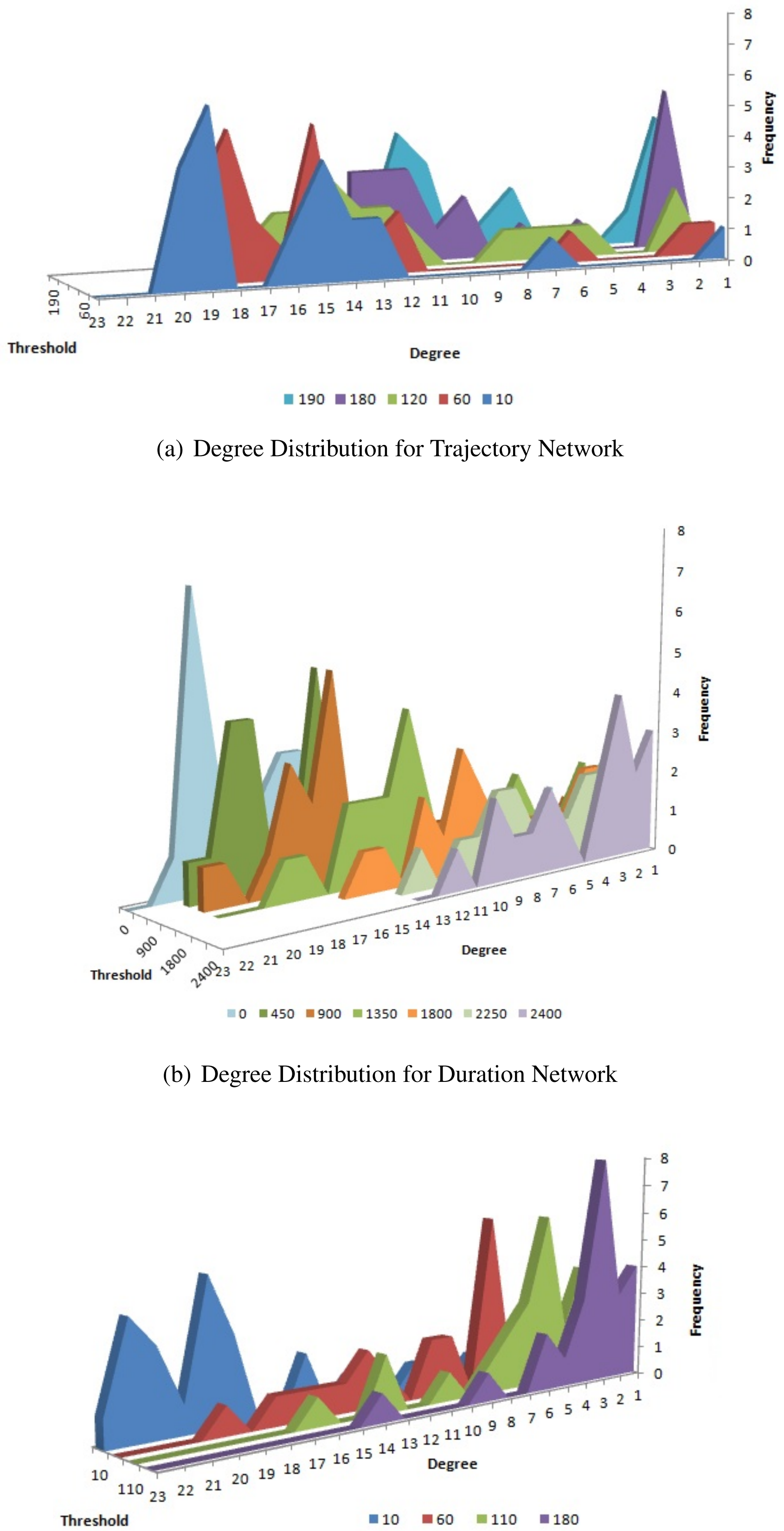
5.5.1 Shortest Path Length and Network Structure

It was found that the resulting opportunistic networks included some noise due to the nature of Bluetooth technology in the room level (i.e. floor) (2-3m). As a result, threshold levels were applied in different networks where each network has its own threshold levels due to the nature of each network. For example, with the co-located network, as it represents the frequency of physical interaction between participants in and out school buildings, the threshold will be applied regarding the frequency. For a trajectory network, as it represents the number of joint time intervals of participants of being at same location, the threshold will be applied regarding the number of joint time intervals. Furthermore, the same applies to duration; the threshold will be applied regarding the time duration.

It is challenging to detect the suitable threshold for different types of data. Different thresholds were tested on the extracted networks. By applying different thresholds on the different networks, a selective threshold degree distribution is displayed in Figure 5.8 to show the effect of removing noise on the data set connectivity. It is found that by removing weak links from the opportunistic network large numbers of participants have small degree values. On the other hand, Figure 5.9 shows the shortest path length for all different threshold levels on different networks. It is found by removing weak links it affects on the degree distribution that affects on the network connectivity particularly in the trajectory and duration opportunistic networks. This figure explores the effects of removing weak links on the connectivity between the participants by increasing the threshold level.

5.5.2 Key-player Identification

In Figure 5.10, it is seen that co-located network is highly connected through different level of frequency threshold than trajectory and duration network. This means that small number of key-players are required to cover the whole network, even at high levels of threshold.

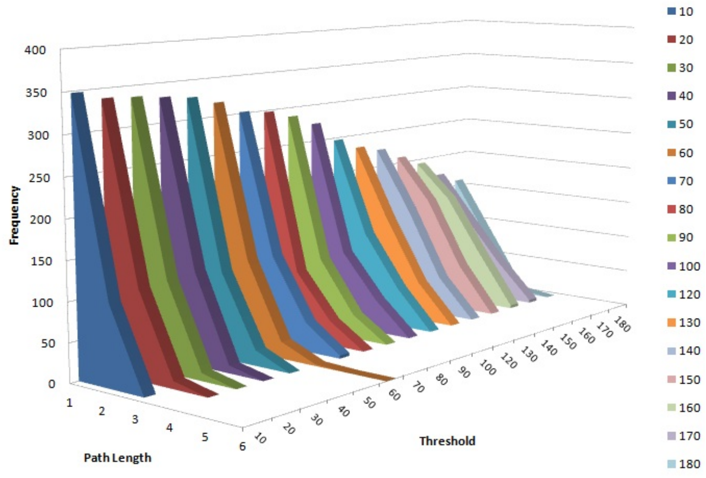


(a) Degree Distribution for Trajectory Network

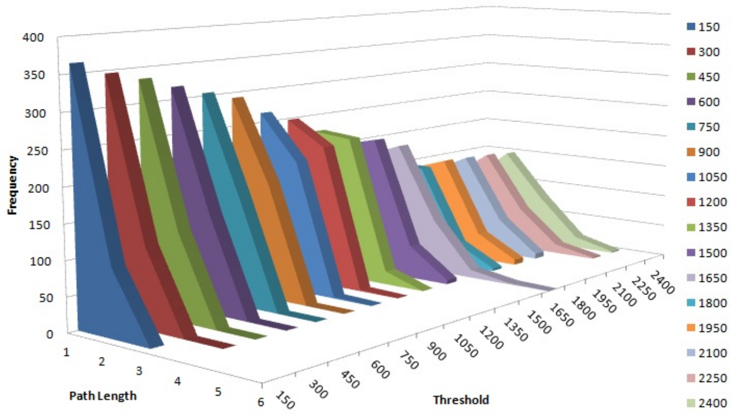
(b) Degree Distribution for Duration Network

(c) Degree Distribution for Co-located network

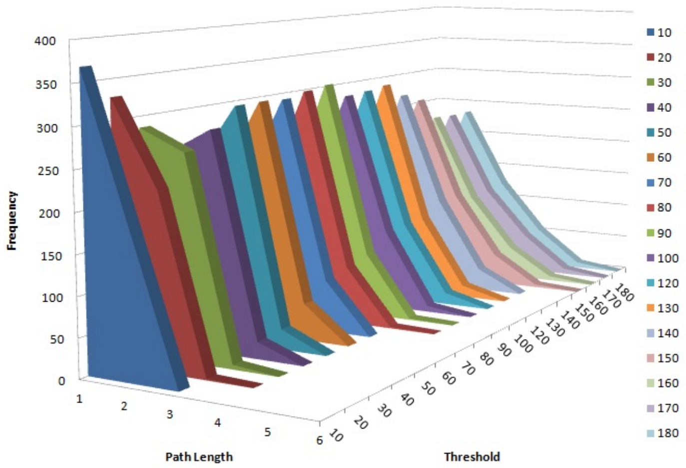
Figure 5.8: Degree Distribution of Different Opportunistic Network at Different Threshold.



(a) ShortestPath for Trajectory Network

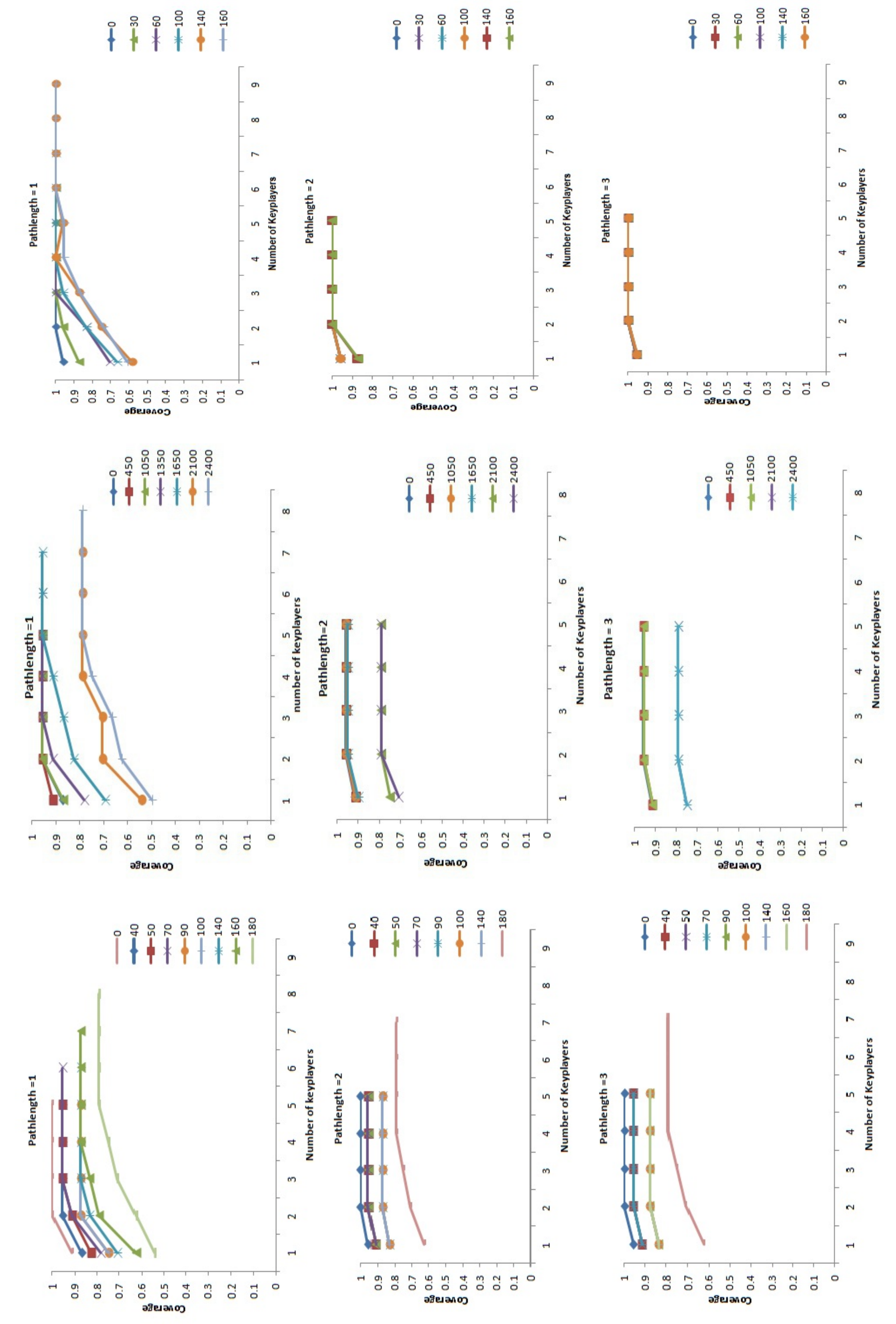


(b) ShortestPath for Duration Network



(c) ShortestPath for Co-located network

Figure 5.9: ShortestPath Length of Different Opportunistic Network at Different Threshold.



(a) Keyplayers of Trajectory Network

(b) Keyplayers of Duration Network

(c) Keyplayers of Co-located network

Figure 5.10: Keyplayers of Different Opportunistic Network at Different Threshold.

5.6 Summary

To sum up, this chapter focused on extracting the characteristics and analysing the collected data from the empirical study. A mathematical notation was provided in order to facilitate the analysis of this data set. Statistical analysis was applied to this data to describe the mobility characteristics, such as users' routine and venue popularity. Different algorithms were proposed to extract the opportunistic network from the collected data.

From the experimental data, we were able to extract different types of networks. Basically, we have two types of collected data, namely device-to-device (physical interaction) and device-to-server (mobility tracking). From the device-to-device data, we extracted the co-located network that represents what devices see. From the device-to-server data, we extracted two types of network; one represents the occurrence of being at the same place at the same time (known as trajectory) and the other represents the duration of being at same place at same time when it happens (known as duration). In order to extract these networks' characteristics, social network analysis was applied where certain key findings were highlighted.

- **Participants routine**

  It is surprising that although the first two weeks of the experiment were normal study weeks, there was no specific routine (the participants did not have a static routine for a specific day of every week for the three weeks of experimental study) that the participants follow during their time at the School, as seen in Figure 5.2 and 5.3. Compared with the participants' study timetable, no specific routine was followed.

- **Linkage between participants and the buildings**

  It was found that the participants were very active at the two access points that cover the most important labs and class rooms to them. It is not surprising that the pattern of the participants' movements were consistent with the occurrence of events for students (namely coursework submission) as seen in Figures 5.4, 5.5 and 5.6.

- **Extracted key networks**

  In this experimental study, there are three useful views of opportunistic networks. H_{b1}, H_{b2} and H_{b3} are undirected graphs that explore the extracted social networks. H_{b1} is the mobility trajectory network that represents the participants' trajectory through the school building. H_{b2} is the duration network that represents the joint duration period of time between participants being at same place. H_{b3} is the co-located social network that represents the physical interactions between participants that have been recorded by Bluetooth mote in system files (experimental data) as seen in Figure 5.7.

- **Extracted opportunistic network connectivity**

  It was found that all the extracted opportunistic networks were highly connected even after removing all weak links by increasing the threshold to remove noise. It is not surprising that these opportunistic networks only require a small number of key-players to cover the whole network due to their connectivity, as seen in Figure 5.9 5.10.This means that they are potentially highly effective for opportunistic communication.

Chapter 6

Characteristics and Analysis of Self-reported Social Networks

Overview

This chapter presents an analysis of the collected data from the survey. In Chapter Four, the questionnaire was designed to address a number of dimensions. From the participants' perception, it addressed who they considered to be within their friendship and acquaintance group and the perceived strength of this relationship. Confidentiality was maintained since, in many cases, reciprocal relationships did not exist. Also explored was *how* relationships were supported by communication, including face-to-face communication and communication technologies (including Email, SMS, Facebook, telephone calling, micro-blogging and Internet messaging). The frequency of usage of these modes of communication was ranked to assess their relative importance. In addition, participants were asked about the frequency with which they meet in groups. Confidentially was preserved and anonymity applied to the full data set prior to analysis. The outcome of the survey is a set of social network graphs that express multiple problem dimensions in this chapter. From this, analysis techniques were applied and developed to determine key structural characteristics of the population's interactions and relationships.

Initially, this chapter focuses on the characteristics of social network structure. In section 6.1 it is focused on the modes of communication that are ordered by popularity and their effect on network connectivity. In section 6.2 the social network is studied in terms of the modes of relationship strength. In addition, each relationship strength will be aggregated on the network structure. In section 6.3 different frequencies of interaction network structures are studied. It considers the way in which different communication technologies are used with different frequencies. It also considers the frequency of interaction between different relationship strengths. In section 6.4 two types of social network clustering are presented. Finally, the key-players of the network are extracted, based on different path length.

6.1 Communication Methodology and Network Structure

Different modes of communication were considered through a directed network $G_C = (V, E)$ where V is the set of participants and an edge $(i, j) \in E$ exists if and only if node i communicates with node j via at least one mode of communication, where $C \subseteq \{c_1, \ldots, c_5\}$, the set of possible modes of communication. The effect of different sets C were examined on shortest path length of all pairs while keeping the population V fixed. Table 6.1 displays different communication modes C ordered by popularity, with the greatest first. The popularity of communication was measured by the number of participants who report that they use a particular mode of communication.

Table 6.1: Popularity of Different Modes of Communication.

Identifier	Communication Mode	% Participants
c_1	Face-to-Face	87.82%
c_2	SMS	46.9%
c_3	Facebook	45.2%
c_4	Phone call	14.78%
c_5	Chat on internet	11.3%
c_6	Email	9.5%
c_7	Twitter	1.7%

The table starts with the most popular and progressively explores the effect of adding less popular modes of communication. This shows a progressively denser network structure. The effect of additional communication modes cause additional connectivity which was examined by considering the profile of all shortest path lengths within the network . Table 6.2 presents the mean path progressively decreasing.

Table 6.2: Statistics on Average Path Lengths in Graph G_C.

Communication set C	Mean Shortest Path Length	Standard Deviation on Path Length	% with no path
$\{c_1\}$	3.96	1.4406	26.5%
$\{c_1, c_2\}$	3.66	1.266	21.5%
$\{c_1, c_2, c_3\}$	3.40	1.118	17.89%
$\{c_1, c_2, c_3, c_4\}$	3.39	1.1149	16.4%
$\{c_1, c_2, c_3, c_4, c_5\}$	3.35	1.097	15.7%
$\{c_1, c_2, c_3, c_4, c_5, c_6\}$	3.33	1.088	15.0%
$\{c_1, c_2, c_3, c_4, c_5, c_6, c_7\}$	3.33	1.08	14.25%

From Figure 6.1 it is seen that two major changes presented in the mode of shortest path length due to combining SMS and Facebook. From Table 6.1 SMS and Facebook are well adopted (46.9%,45.2%) respectively. From Table 6.2 the addition of SMS and Facebook communication to the network marginally increases participant inclusion, with the percentage of participants having no path between them decreasing from 26.5% to 21.5%, 21.5% to 17.89% respectively.

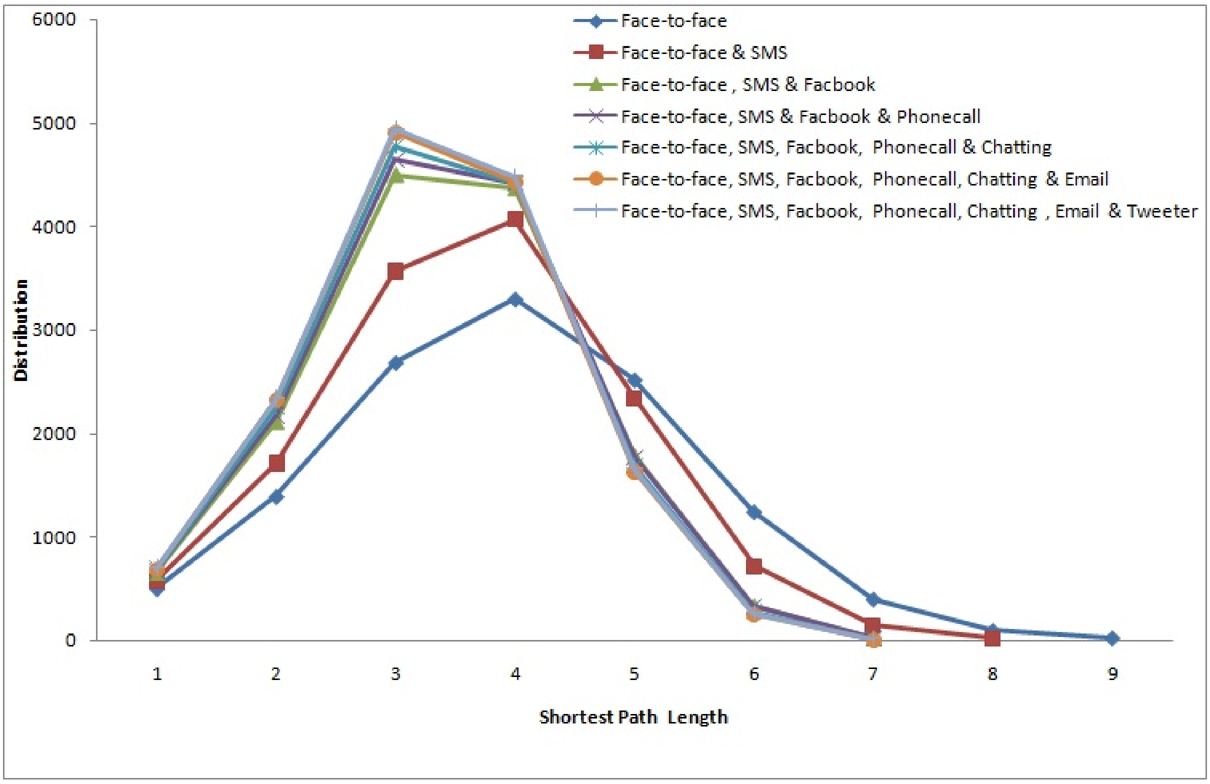


Figure 6.1: The Distribution of Shortest Path Lengths for Different Modes of Communication..

Figure 6.2 presents the degree distribution. Face-to-face and Facebook are a random network distribution as they follow Gaussian distribution. This means most of face-to-face and Facebook nodes have medium degree. However, other communication methodologies are scale free networks as their degree distributions follow the power low distribution. This means that most of nodes in these networks has a small degree.

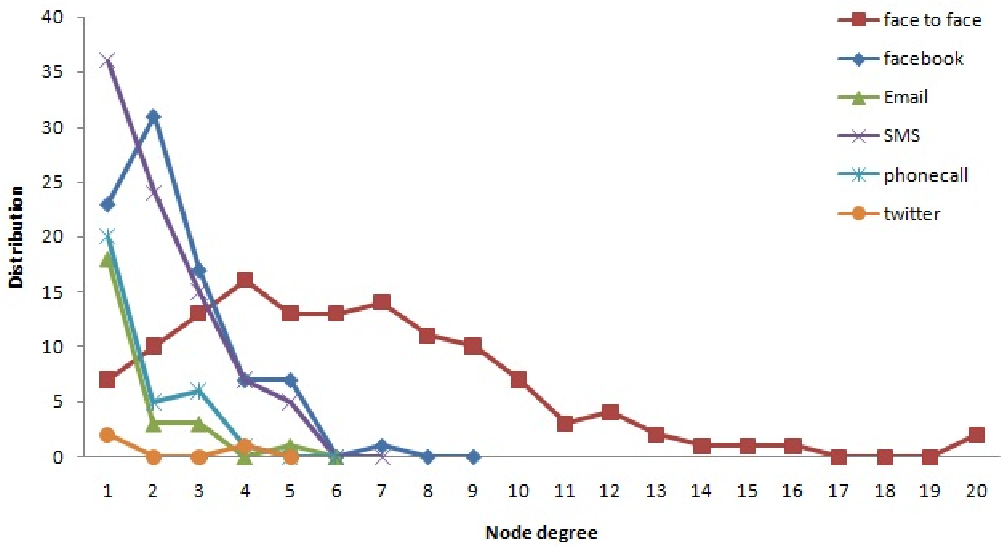


Figure 6.2: The Distribution of Degree for Different Modes of Communication..

According to Table 6.3, most communication methodologies correlate with nodes of similar degree, except Face-to-face, Email and Twitter. This result was separate for each methodology. In Table 6.4 by combining the methodologies together progressively, it was found that all of the nodes in the network correlate with nodes with similar degrees except in the network where Facebook is combined to the network.

Table 6.3: Assortative Coefficient of Different Modes of Communication.

Identifier	Communication Mode	Assortative coefficient
c_1	Face to Face	-0.051045195
c_2	SMS	0.18107925
c_3	Facebook	0.003008933
c_4	Phonecall	0.386138614
c_5	chatting on the internet	0.226873385
c_6	Email	-0.201430274
c_7	Twitter	-0.666666667

Table 6.4: Assortative Coefficient of Combining Different Technologies G_C.

Communication	Assortative coefficient
$\{c_1\}$	-0.051045195
$\{c_1, c_2\}$	0.006119557
$\{c_1, c_2, c_3\}$	-0.003682968
$\{c_1, c_2, c_3, c_4\}$	0.023202953
$\{c_1, c_2, c_3, c_4, c_5\}$	0.012287923
$\{c_1, c_2, c_3, c_4, c_5, c_6\}$	0.009336862
$\{c_1, c_2, c_3, c_4, c_5, c_6, c_7\}$	0.013304971

6.2 Relationship Strength and Network Structure

In order to examine the effect of different levels of relationship strength on the shortest path length, we considered a directed network $G_R = (V, E)$. V is the set of participants and there exists an edge $(i, j) \in E$ if and only if node i perceives it has a relationship with node j with strength from set $R \subseteq \{R_1, R_2, R_3\}$. Where R_1 is the strongest relationship strength, represented as *strong friendship* (someone with whom you have significant level of trust and interaction), R_2 is the medium level of relationship strength, referred to *friendship* (someone with whom you have empathy or common views with and may socialize with them) and R_3 is the weakest relationship strength, denoted as *course-mate* (someone you know and would acknowledge but with whom you have little other contact). The effect of relationship strength was examined by combining different relationship types together, starting with the strongest strength (which is the least popular) and progressively observing the effect of adding weaker links. The effect of this additional connectivity was analysed using shortest paths.

In Figure 6.3, each person has a number of friends and course-mates greater than the number of strong friendships. These numbers of weak links play an important role in increasing the connectivity of the network. Figure 6.4 displays different densities and structures for each type of relationship strength.

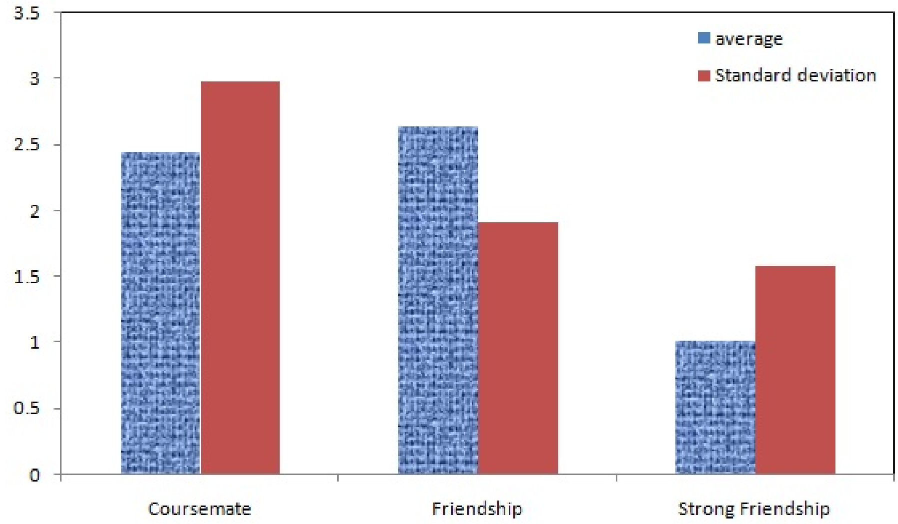
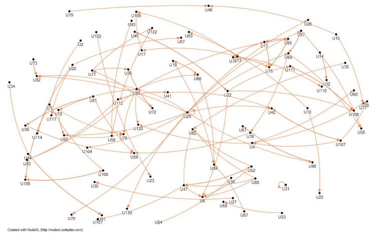
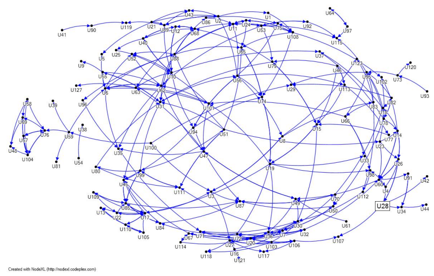


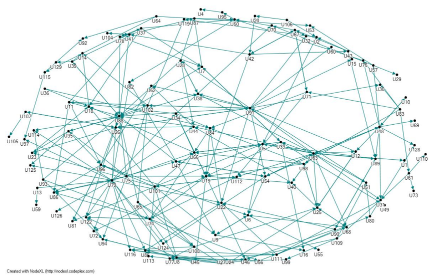
Figure 6.3: Relationship Strengths.



(a) Strong-friendship



(b) Friendship



(c) Course-mate

Figure 6.4: Different Relationships Social Network.

In Table 6.5 there are two types of counted links: the first one is the directed link that counts the links according their directions between network nodes. The second link is the undirected link that counts the links without considering the direction (there is a link from i to j and a link from j to i and these two links are counted as one link in the undirected network). These two types of links were used in order to clarify the percentage of bi-directional links for each strength of the relationship. It was found that 32% and 37% bi-directional links were in friendship and strong friendship strength where only 17% were in course-mate. This means that course-mate strength connects between many different participants together, which increases the connectivity inside the network, as presented in Figure 6.5. The role of the weak relationships (namely

Table 6.5: Popularity of Relationship Strength.

Identifier	Relationship Strength	% Participants	# of Directed Links	# of Undirected Links
R_1	Strong Friendship	48.69%	116	74
R_2	Friendship	84.34%	303	207
R_3	Course-mate	69.56%	281	233

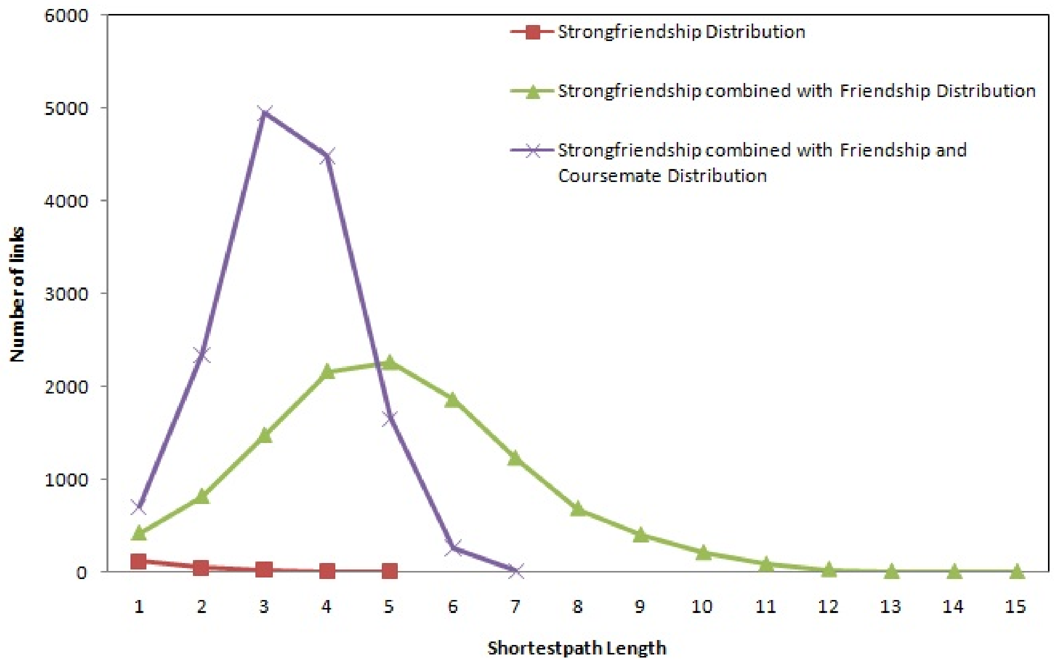


Figure 6.5: The Distribution of Shortest Path Length for Different Modes of Relationship..

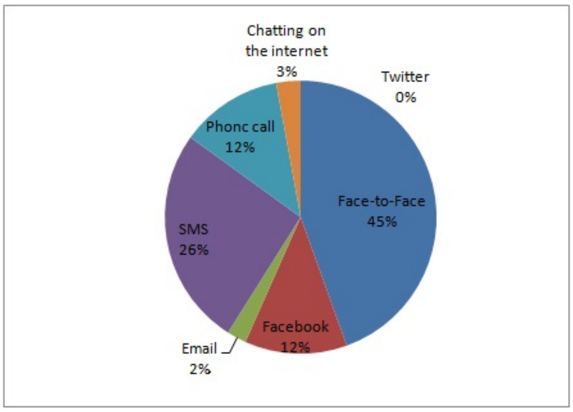
course-mate) is notable. Table 6.6 shows the effect of combining different types of relationships, starting with the strongest relationships and progressively adding the weaker ones. The weakest relationships play a significant role in reducing the mean path length and improving connectivity. This is also shown in Figure 6.5 where the distribution of shortest path lengths between all pairs of nodes (within connected components) is given. The addition of the course-mate relationships to friendships and strong friendships leads to a reduction in mean shortest path length from 5.046 to 3.39. The path length reported for strong friendship (1.547) is low due to the graph having a small connected component.

Table 6.6: Statistics on Average Path Lengths within Connected Components .

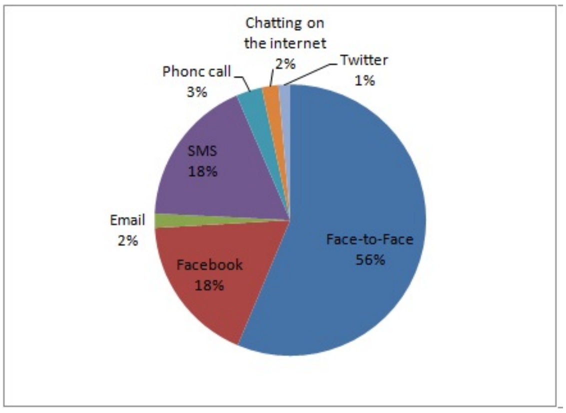
Communication set R	Mean Shortest Path Length	Standard Deviation on Path Length	% with no path
$\{R_1\}$	1.547	0.866	97.6%
$\{R_1, R_2\}$	5.046	2.107	23.7%
$\{R_1, R_2, R_3\}$	3.39	1.089	14.25%

6.2.1 Modes of Communication and Relationship Strength

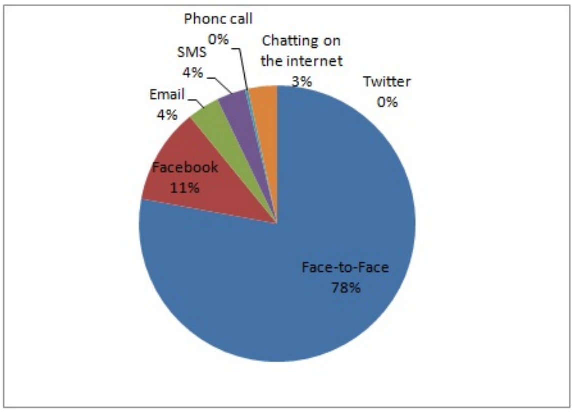
In Figure 6.6, it is notable that, as long as the relationship strength reduces the preferences of physically communicating with other increases,the preferences of using other communication methodologies decline.



(a) Strong Friendship



(b) Friendship

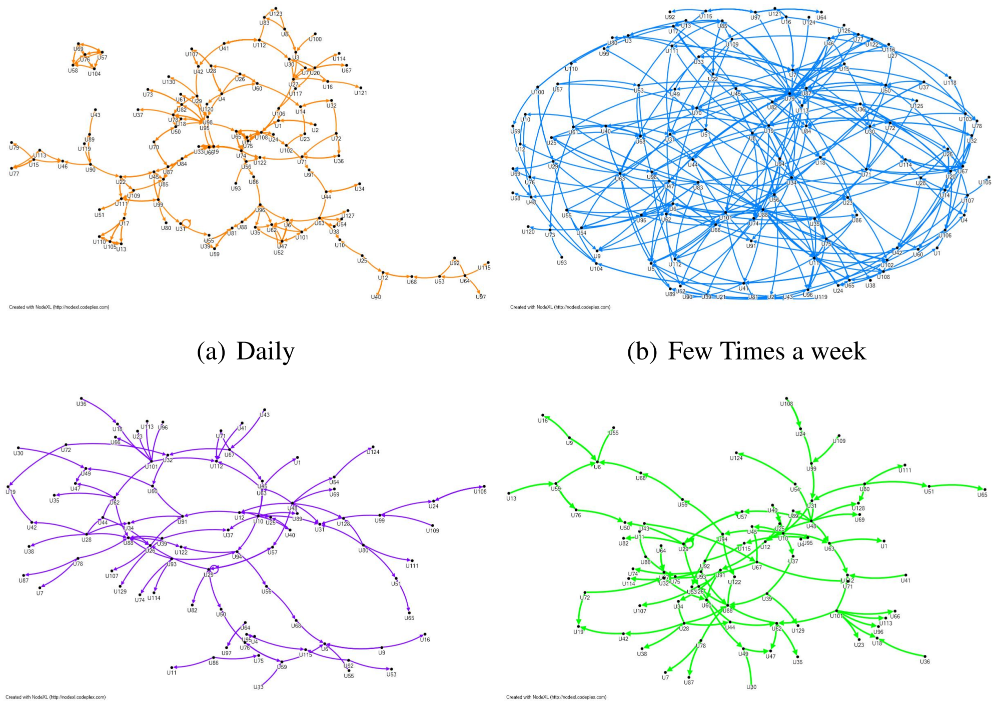


(c) Coursemate

Figure 6.6: Relationship and Methodology Usage.

6.3 Frequency of Interaction in the Social Network

The communication frequencies are classified into four categories, namely: daily, few times a week, once a month and once a semester. Each class has a different network structure as seen in Figure 6.7. It clear that, in terms of daily interaction, the participants interact via a connection of clusters as seen in Figure 6.7(a). Regarding a few times a week, the communication between participants is more regularly distributed and is shown in Figure 6.7(b). As regards the other two frequencies that represent the rare frequencies of interaction, the networks are sparse, as seen in Figure 6.7(c),6.7(d). Figure 6.8 represents the distribution of different categories of communication frequencies.



(a) Daily

(b) Few Times a week

(c) Once a month

(d) Once a semester

Figure 6.7: Figures of Different Communication Frequencies.

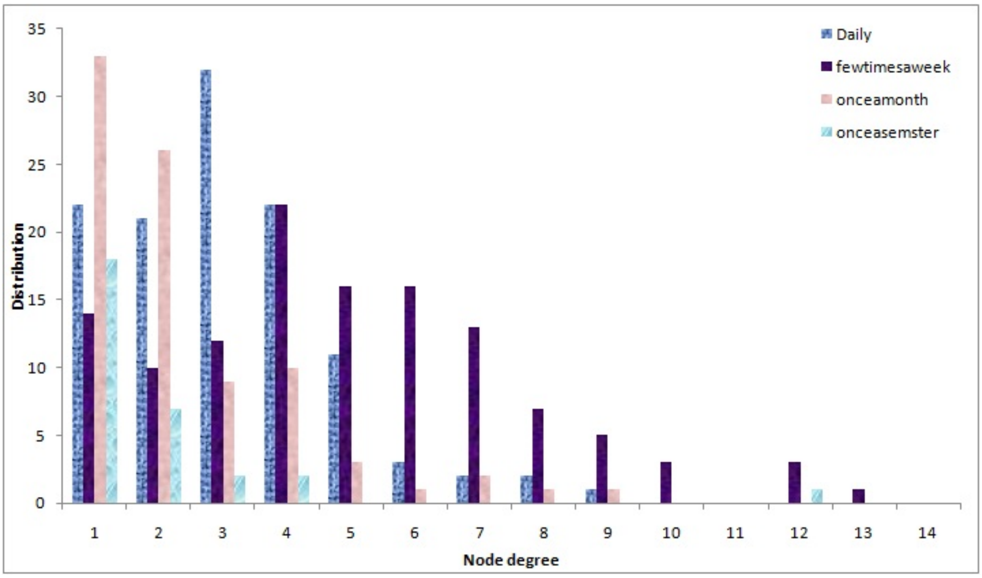
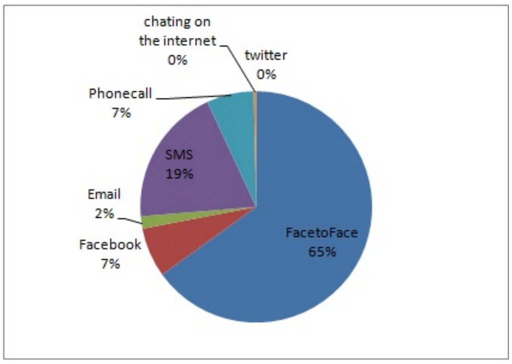


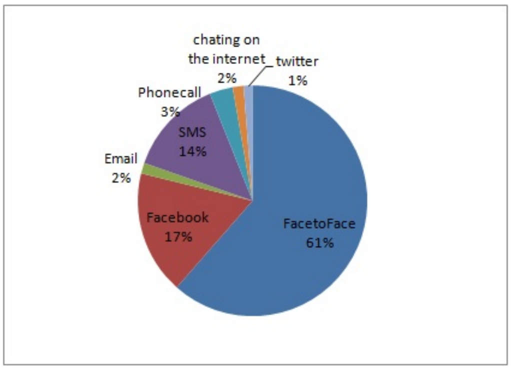
Figure 6.8: Degree of Frequencies.

6.3.1 Frequency of Interaction and Mode of Communication

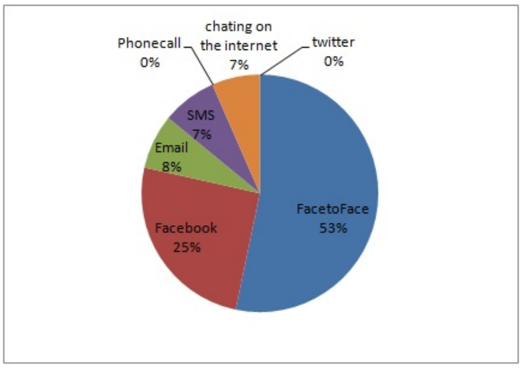
In Figure 6.9 we address the types of communication and the frequency with which they are used. A number of findings are notable. Firstly, highly frequent communication, such as daily, are sustained primarily by Face-to-face interactions. As the frequency of communication in a relationship is reduced, Face-to-face and SMS interactions reduce and Facebook interactions are increasingly used, especially in the case of highly infrequent interactions (once a month). At the same time, it seems that chatting on the internet becomes a substitute for Facebook communication.



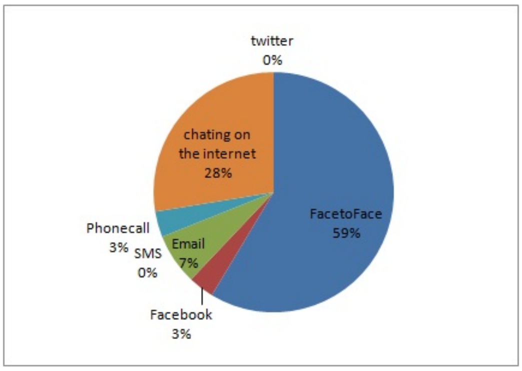
(a) Daily



(b) Few times a week



(c) Once a month

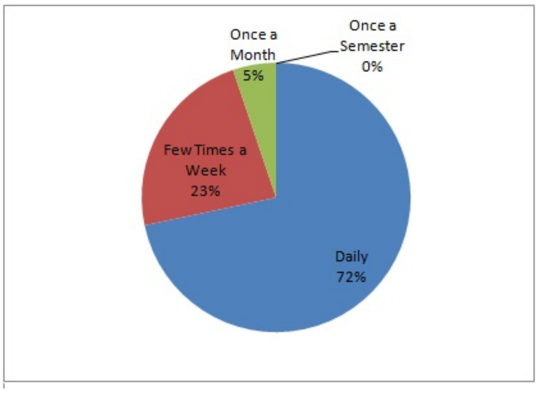


(d) Once a semster

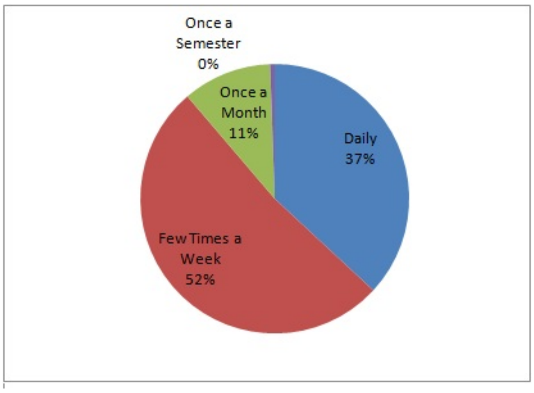
Figure 6.9: Frequency of Usage for Different Communication Technologies.

6.3.2 Frequency of Interaction and Relationship Strength

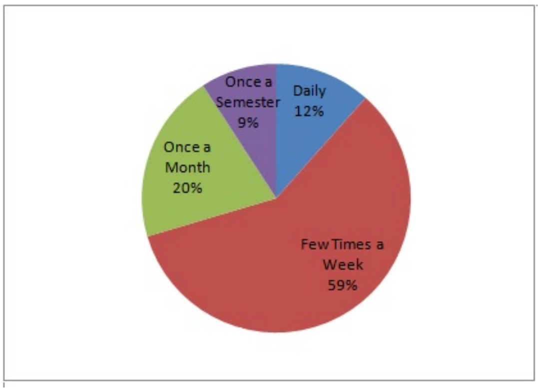
In Figure 6.10 there is a correlation between relationship strengths and the frequency of interaction. Strong relationships prefer to interact with frequent interaction (strong friendship). In contrast, weak strength relationships (course-mate) prefer to interact less frequently.



(a) Strong Friendship



(b) Friendship



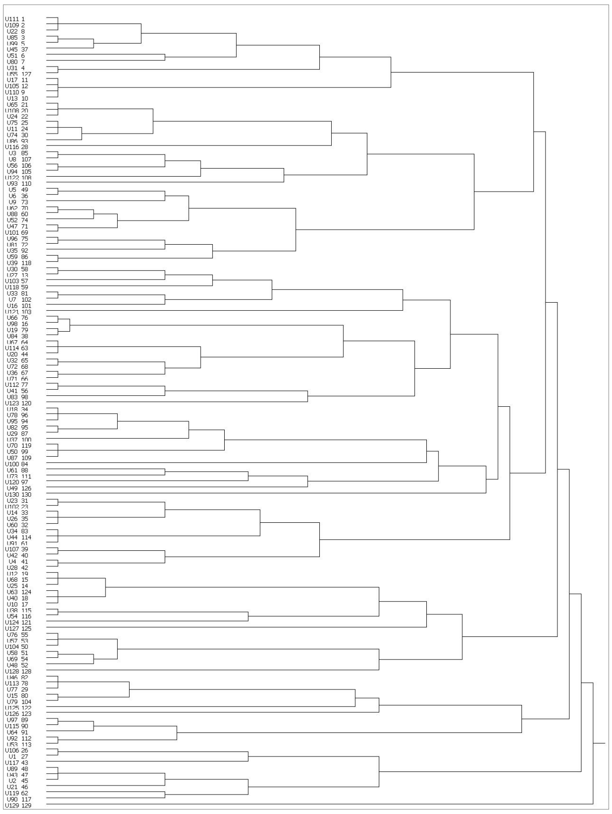
(c) Coursemate

Figure 6.10: The Frequency of Interaction and Relationship Strength.

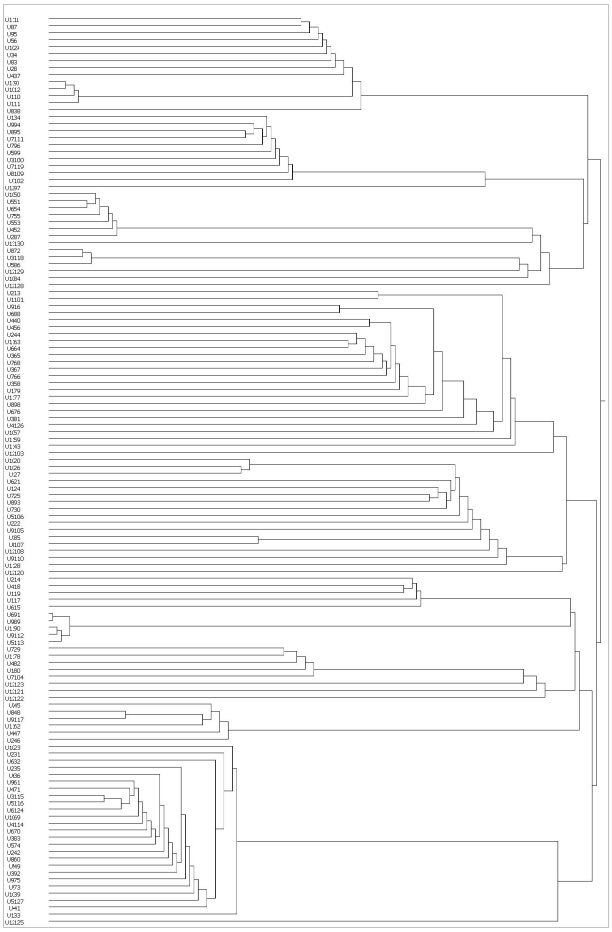
6.4 Community Detection and Clustering

An interesting feature of the hybrid online and physical social network is the extent to which different technologies support dense sub-structures and the role played by individuals in the network. To explore this, we analysed the clustering characteristics induced by different technologies. We applied the Grivan-Newman clustering method [73]. Using betweenness centrality, Grivan and Newman focus on constructing a measure to indicate the edges which are least central to the cluster and they remove them. This divisive technique is repeatedly applied as described in [73].

Figure 6.11 show two different results of community clustering where 6.11(a) presents the community clusters using Hierarchical clustering and 6.11(b) presents the community clusters using Grivan and Newman clustering. As seen in Figure 6.11 using different similarity measure between communities lead to different community clusters.



(a) Hierarchical clustering results

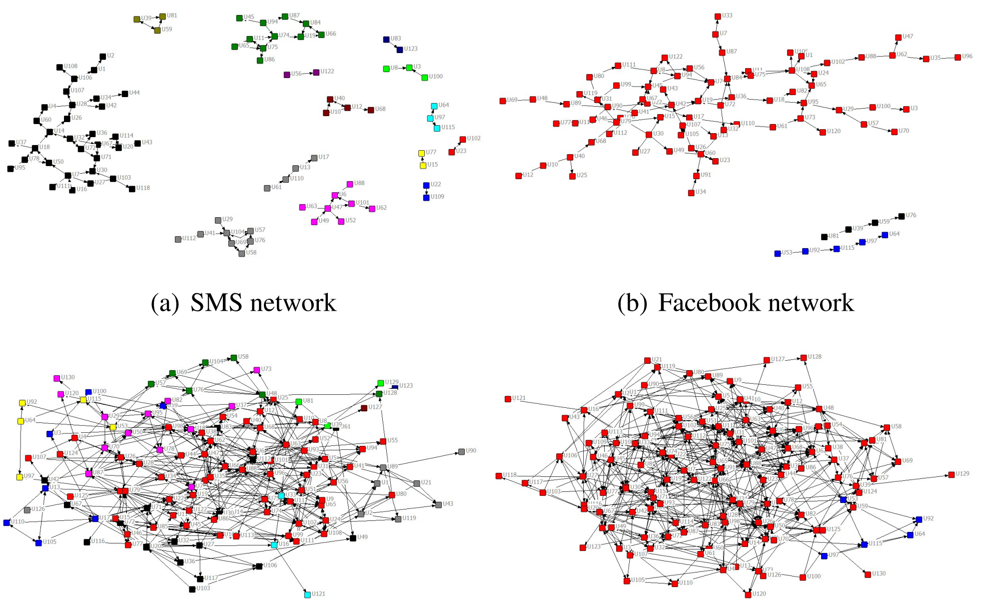


(b) GrivanNewman clustering

Figure 6.11: Clustering Results.

To assess the strength of different clustering levels, we used modularity as an external measure [73, 72]. Modularity compares the number of edges inside a cluster with the expected number of edges that one would find in the cluster if the network were a random network with the same number of nodes and where each node keeps its degree but the edges are randomly connected. It does not provide a guide as to how many clusters a network should ideally be split into but is a useful measure of the quality of a division of a network into clusters, with higher modularity measures indicating increased density within the clustering.

Figure 6.12 shows the clustering groups for SMS, Facebook, face-to-face, and the aggregation of the physical (face-to-face) and all the communication methodologies networks. As it is seen in this figure, each color represents a separate group. Figure 6.13 shows the quality of clustering (in terms of modularity) as a function of the number of clusters when applying the Grivan and Newman approach. We can see that SMS network is divided into a large number of groups than Facebook and face-to-face that gives a very high modularity results. However SMS is consist of a disconnected groups. In Figure 6.13 we can see at 15 to 20 clusters SMS has an effect in increasing the modularity result of the total network where Facebook and face-to-face has a steady value of modularity at this number of clusters.



(a) SMS network (b) Facebook network

(c) Face-to-face network (d) Online & physical network

Figure 6.12: Clustered Sub-networks at The Highest Modularity Value.

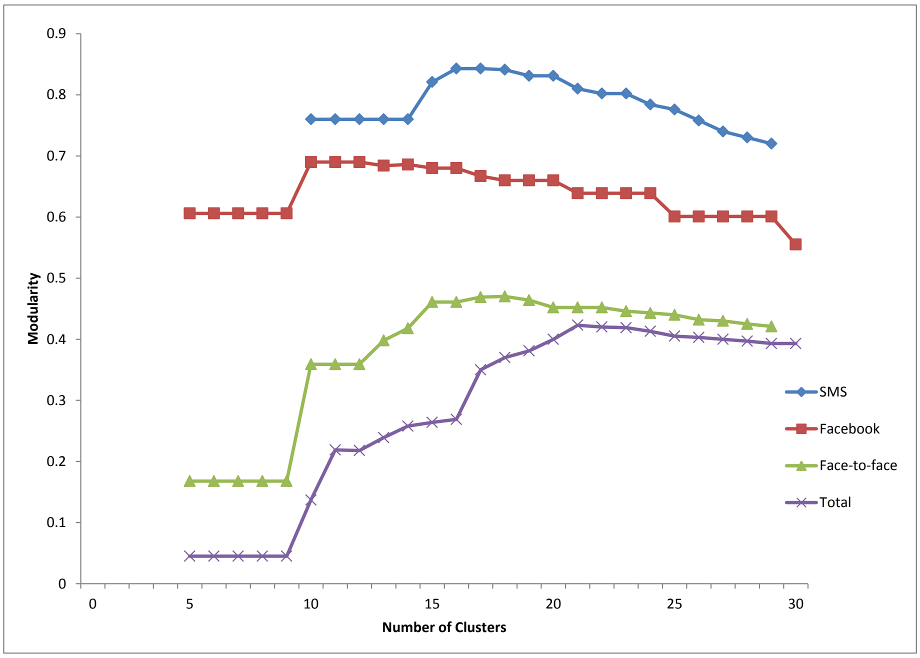


Figure 6.13: Modularity Values through Iteration of Different Methodologies.

6.5 Extracting Key-players of The Network

The analysis began by defining the essential information communication technologies in the social network followed by the significance of weaker links. The next analysis was related to clustering the social network to observe the community structure followed by the modularity in order to extract the most accurate number of community in the data set. In this section, the key-players algorithm was applied in the network to obtain the optimal number of nodes that were included in dominating set.

Key-players represent a minimal subset of nodes from which all others can be reached within a particular maximum path length. Introduced in [13], we employed key-player analysis to determine the trade-off between a minimum number of informed participants and extent of possible dissemination across the population within a given path length. This allows us to explore the susceptibility of a network to possible information spreading effects. We achieved this by searching for generalised dominating sets that maximise the proportion of the participants that are reachable from at least one key-player within a given path length. The heuristic technique that we adopted the key-player algorithm in Chapter 2 for this research. We adapt the algorithm to make it suitable to be applied on the directed network by adapt the reachability measure where the applied results is shown in Figure 6.14. We use the out-degree as a measure of reachability instead of D^R as we are looking to measure the fastest path of information dissemination in the network.

Figure 6.14 shows that Email, Chatting on the internet, SMS and Phone-call communication methodologies suffer from dis-connectivity, as clarified in the key-player results for different path length. on of the cause of this result is the unpopularity of these communication methodologies. In contrast, face-to-face communication shows a good connectivity through the different path lengths. Generally, Facebook shows better connectivity than SMS although SMS is more popular but it consists of a disconnected groups.

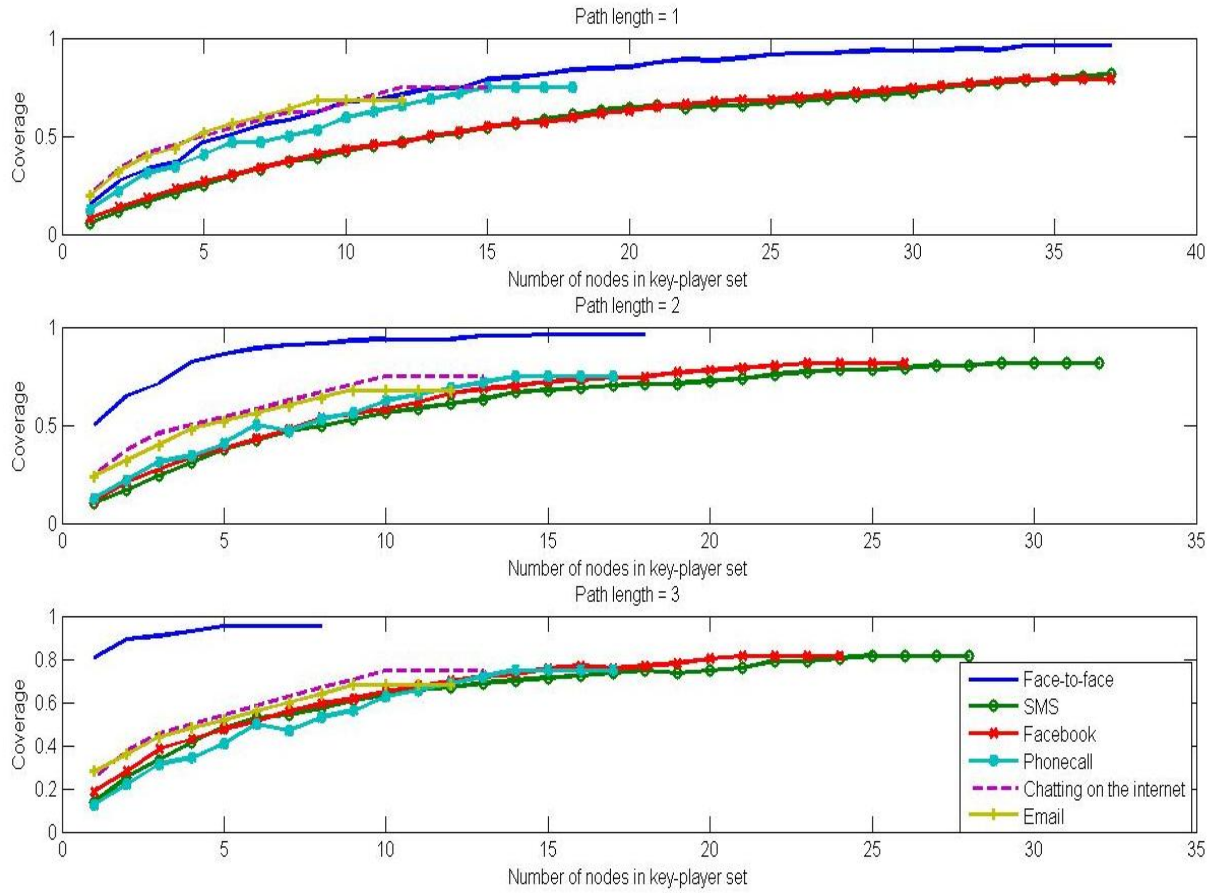


Figure 6.14: The Result of Key Players for Each Communication Technology at Different Path Length.

From Figure 6.15 it was noticed that, by combining the other communication methodologies with face-to-face, we have more connectivity than having face-to-face alone. This proves the theory that on-line communication methodologies support and increase the connectivity of the existing physical relationship. As we can see 100% of reachability is achieved at 23 nodes in the key players set. This is equal to the number of clusters which achieve the optimal modularity value for the key-players of combination between physical and other communication methodologies.

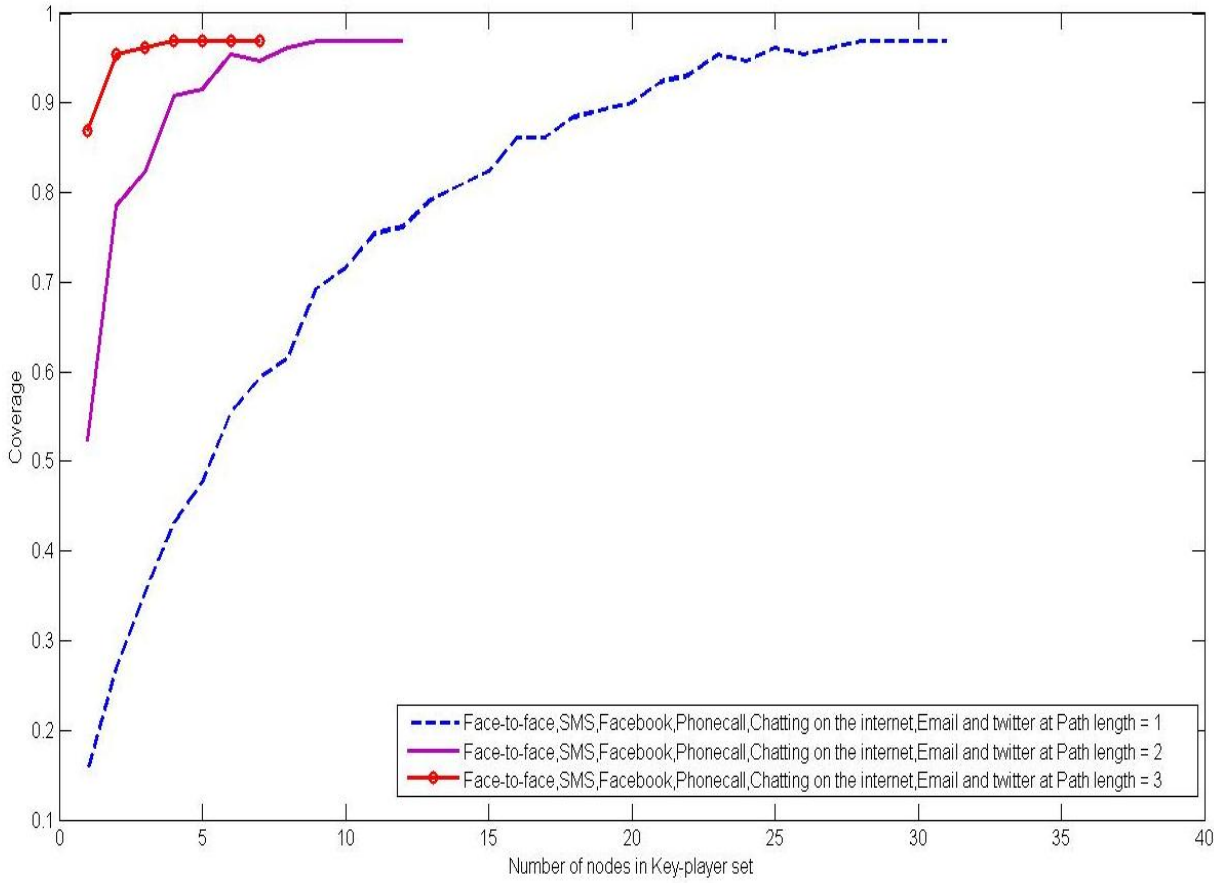


Figure 6.15: Keyplayers of Combination between Physical and Online Communication Technologies at Different Path Length .

6.6 Summary

This chapter focused on the analysis of the detected self-reported (physical) social networks. These networks were built based on the participants' perception by completing an online survey. Social network analyses were applied to the social networks to extract the characteristics. From these analyses, several interesting key findings were found.

- **Faceook and SMS are more popular communication methodology than other communication methodologies**

In terms of communication methodology and network connectivity, it was found that Facebook and SMS were the most popular methodologies in addition to the dominant face-to-face interaction. In Figure 6.1 it was found that these technologies improve the connectivity of the network by combining their links to the physical interaction network.

- **Weak Links adapt the network connectivity**

  From the relationship perspective, weak links play an important role in increasing network connectivity. The strength of the weak links come from having more unidirectional links. According to Figure 6.5 course-mates are unidirectional but add significantly to the overall connectivity of the network as the shortest path adapted from 5.046 to 3.39.

- **Weak links prefer to interact physically while strong links interact through communication methodologies**

  In Figure 6.6, SMS and phone-call are generally used in sustaining stronger links as compared to SMS and phone-call for friendships and course-mates, however face-to-face sustained weaker links as compared to face-to-face with stronger links

- **Positive correlation between relationship strength and frequency of interaction**

  There was a positive correlation between relationship strength and frequency of interaction; as the relationship strength decreases, the frequency of interaction declines, as seen in Figure 6.10.

- **Key-players and communication methodologies**

  Electronic communication mechanisms have a much greater effect on community clustering, especially Facebook. For key players, there is a strong effect in local (path length 1) spreading for Facebook interactions. For longer path lengths, Face to face interaction is still dominant, as seen in Figure 6.14.

- **Detected key networks**

  Different types of networks were detected in the survey study. These networks were directed networks that represent different aspects. Communication networks represent different communication technologies that were used from participants perspective to communicate together (namely face-to-face, Facebook, SMS, Twitter, phonecall). Relationship networks also represent the relationship strength between different participants from their perspective (course-mate, friendship, strong friendship). Frequency networks represent the frequency of interaction between different participants (daily, a few times a week, once a month, once a semester). A subset of these networks was extracted to represent the experiment participants only and to be an undirected graph. We focused on four key networks that were face-to-face networks which reflected the physical interaction between participants. In addition, relationship strength network found a correlation between the relationship strength and the experimental social networks.

All in all, it is found that corresponding to the relationship strength, participants prefer to communicate more with the new communication technologies than physically interact particularly the strong relationship strength as seen in Figure 6.6. This finding is concluded from the survey results where participants become less dependant on the face-to-face interaction and they are using portable communication technologies instead. This fact did not conflict with the fact that the physical interaction is the dominant method of communication as seen in Table 6.1.

Chapter 7

Comparison of Opportunistic and Social Networks

7.1 Introduction

In this chapter, we examine both the humans' social networks and the opportunistic networks that come from their mobility and interactions. We aim to see the similarities and differences there are between the opportunistic networks that devices see and the social networks that exist between humans. In Chapter Three, a mobility tracking system framework was designed to determine the opportunistic network that participants create from indoor mobility. An experimental study was conducted to collect the required data set about participants' opportunistic networks from their mobility patterns. This data set was analysed and three different assessments of opportunistic networks (co-location detection and the device, duration from mobility tracking in wireless cells) were extracted in Chapter Five. In Chapter Five, we made a number of observations on participant mobility. In particular:

- **Participants routine**

  We found that, for the first two weeks of the experimental study, the participants did not follow any pattern (such as being at the same location on the same day of week at the same time) in their movements throughout the school building. This finding was surprising as the first two weeks were normal study weeks that follow the school timetable.

- **Correlation between participants and the buildings**

  In tracking the participants' movements, the data was dominated by two important access points that cover the important labs and class rooms used by participants in the school timetable. It is not surprising that the participant movement patterns are consistent with the occurrence of events for students (coursework submission).

- **Extracted opportunistic network connectivity**

Generally, weak links always play an important role in the connectivity of social networks. It is interesting that the experimental network was still highly connected after removing weak links. As a result, these types of networks require a small number of key-players to cover the whole network.

- **Relationship between participants together**

  The co-located opportunistic network is a network that has been built from the recorded physical interaction between the participants. We found that the co-located and duration social networks had a similar relationship with nearly the same strength. This means that the co-location opportunistic network is able to represent the full picture of participants' interaction. It is surprising that both networks were similar although the duration network is a deduction from the participant's movements (as detected by the server).

- **Extracted key networks**

  In this experimental study, there are three useful views of opportunistic networks. H_{b1}, H_{b2} and H_{b3} are undirected graphs that explore the extracted social networks. H_{b1} is the mobility trajectory network that represents the participants' trajectory through the school building. H_{b2} is the duration network that represents the joint duration period of time between participants being at same place. H_{b3} is the co-located social network that represents the physical interactions between participants that have been recorded by a Bluetooth mote in system files (experimental data).

In Chapter Four, a questionnaire was conducted to collect the participants' social network from their viewpoint and to understand their behaviour in terms of a pattern of communication. A comprehensive analysis of the collected data is demonstrated in Chapter Six where the following findings were concluded:

- **Faceook and SMS are a more popular communication methodology than other communication methodologies**

  In terms of communication methodology and network connectivity, it was found that Facebook and SMS were the most popular methodologies in addition to the dominant face-to-face interaction. It was found that these technologies adapt the connectivity of the network by combining their links to the physical interaction network.

- **Weak Links adapt the network connectivity**

  From the relationship perspective, weak links play an important role in increasing network connectivity. The importance of the weak links come from having more unidirectional links. Course-mates are unidirectional but add significantly to the overall connectivity of the network as the shortest path adapted from 5.046 to 3.39.

- **Weak links prefer to interact physically while strong link interact through communication methodologies**

  Participants with strong relationship strength prefer to use communication such as SMS and phone call for communication. However participants with weak relatioship strength prefer to interact physically rather than using communication methodologies.

- **Positive correlation between relationship strength and frequency of interaction**

  There was a positive correlation between relationship strength and frequency of interaction; as the relationship strength decreases, the frequency of interaction declines.

- **Key-players and communication methodologies**

  Electronic communication mechanisms have a much greater effect on community clustering, especially Facebook. For key players, there is a strong effect in local (path length 1) spreading for Facebook interactions. For longer path lengths, face-to-face interaction is still dominant.

- **Detected key networks**

  Different types of networks were detected in the survey study. These networks were directed networks that represent different aspects. Communication networks represent different communication technologies that were used from participants perspective to communicate together (namely face-to-face, Facebook, SMS, Twitter, phonecall). Relationship networks also represent the relationship strength between different participants from their perspective (course-mate, friendship, strong friendship). Frequency networks represent the frequency of interaction between different participants (daily, a few times a week, once a month, once a semester). Subsets of these networks were extracted to represent the experiment participants only and to be an undirected graph. We focused on four key networks that were face-to-face networks which reflected the physical interaction between participants.

To conclude, it is found that participants prefer to communicate more using new communication technologies than physically interact particularly for strong relationships. This finding is concluded from the survey results where participants become less dependant on the face-to-face interaction and they are using portable communication technologies instead.

In Chapters Five and Six, the opportunistic networks and extracted social networks were studied separately and their characteristics were assessed. In this chapter, we aim to find the similarities and differences between the opportunistic and social networks. This is a challenging task because the networks exist for different purposes, with opportunistic networks being temporal in nature while social networks are a static representation based on perception of the individuals. Therefore, we will look at the difference between these networks using different techniques:

- In Section 7.2 we compare different abstracted structural representations of opportunistic networks with social networks. We adopt an evaluation technique to assess similarity of neighbourhoods.
- In Section 7.3 we assess the temporal nature of the opportunistic networks and social networks. This allows us to see how rapidly the human and machine based structures compare in terms of information dissemination.

These assessments are novel because they allow us to compare very different types of networks based on a common purpose (for example, information dissemination). An opportunistic network is a network that depends on *technologies* without a *brain*, where social network is the network that depends on the *brain* without *technologies* for the same group of people. It is interesting to find out the similarities and differences between these networks

7.2 Characteristics of Opportunistic and Social Networks

In this section we compare different abstracted structural representations of opportunistic networks with social networks. We adopt an evaluation technique to assess the similarity of neighbourhoods. There are a number of networks that were detailed in Chapters Five and Six for this evaluation. Three different networks were extracted from the experimental data set in Chapter Five. H_{b1}, H_{b2} and H_{b3} are undirected graphs that describe the opportunistic network in different ways. These are the mobility trajectory network, the duration network, and the physical interaction (co-located) network that was extracted from the Bluetooth system files (experimental data). G_{c1}, G_{R1}, G_{R2}, and G_{R3} are directed graphs that represent different aspects of the social networks, namely the face-to-face communication network, strong friendship relationship network, friendship relationship network and course-mate relationship network respectively.

Subsets of these networks were extracted that represent the experimental participants only. These networks were identified by adding an $*$ to their notation to differentiate that these are undirected subsets network (that represent the participants network only):

- G^*_{c1} is a undirected graph that shows the face-to-face communication network for experimental participants. It has 22 vertices with 26 edges that have only 5 connected groups.
- G^*_{R1} is a undirected graph that shows the strong friendship network for experimental participants. It has 5 vertices with 7 undirected edges that contain 2 connected groups.

- G^*_{R2} is a undirected graph that shows the friendship network for experimental participants. It has 16 vertices with 21 undirected edges that contain 4 connected groups.

- G^*_{R3} is a undirected graph that shows the course-mate network for experimental participants. It has 24 vertices with 35 undirected edges that include 2 connected groups.

Figure 7.1 shows the degree distribution for different kinds of subset networks that were extracted from survey study. It is noted that face-to-face is a scale-free network as its degree distribution follows the power low distribution. On the other hand, the relationship networks roughly follow a Gaussian distribution. This result for the subset of the network that represent experimental participants only. However this result conflict with the result when we study the whole network together in Figure 6.2

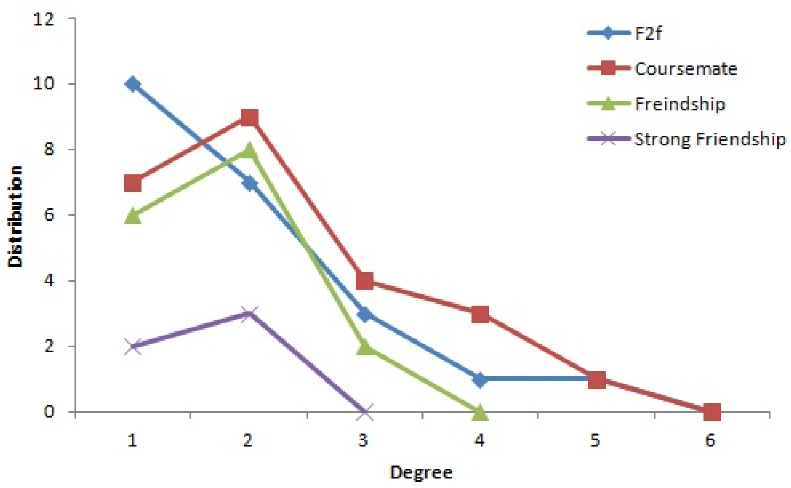


Figure 7.1: Degree Distribution for Different subset of The Networks.

7.2.1 Common Edges between Opportunistic and Social Networks

From the two case studies (experimental study and survey study) we have different types of networks that are very difficult to compare. However, we assessed the similarities and differences between different abstracted networks by using and amending an information retrieval evaluation concept. Here, precision and recall were used for the first time in comparing different networks together. These metrics are generally used to compute the accuracy of the retrieved documents. In particular:

Recall is the fraction of the relevant documents that are retrieved (the number of the correct results divided by the number of the results that have been returned). *Precision* is the fraction of the retrieved documents which is relevant (the number of correct results divided by the number of all retrieved results). There is a trade- off between precision and recall, as seen in Figure 7.2. *F-measure* (i.e. F_1) is the weighted average of precision and recall, where F_1 takes values between 0 and 1 with the best score being 1.

$$Precision = \frac{|\text{relevant documents} \cap \text{retrieved documents}|}{|\text{relevant documents}|} \tag{7.1}$$

$$Recall = \frac{|\text{relevant documents} \cap \text{retrieved documents}|}{|\text{retrieved documents}|} \tag{7.2}$$

$$F_1 = 2.\frac{Precision.Recall}{Precison + Recall} \tag{7.3}$$

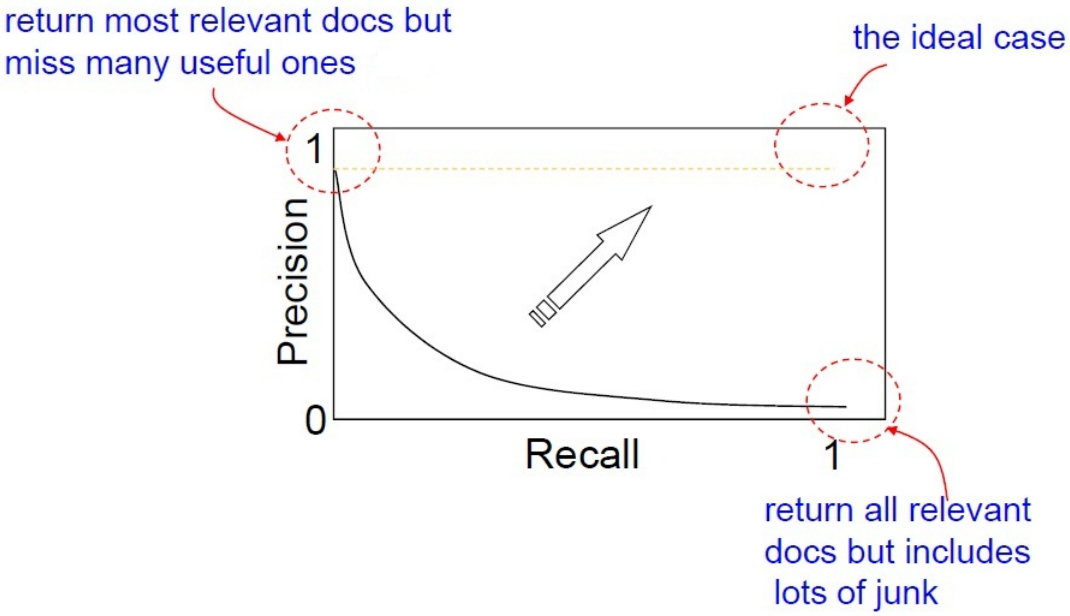


Figure 7.2: Recall and Precision.

Pantel and Lin [78] used precision and recall to evaluate their clustering algorithm that automatically discovers the word sense. On the other hand, precision and recall were used to evaluate ontology matching algorithms in [34], as precision and recall can be used to evaluate corrections and completeness of the results. Our application of precision and recall concept is to compare opportunistic and social networks.

We were interested in performing a comparison between different network structures. This comparison is in order to discuss the relation between static social networks and opportunistic networks. For the first time, precision and recall have been used to explore these differences.

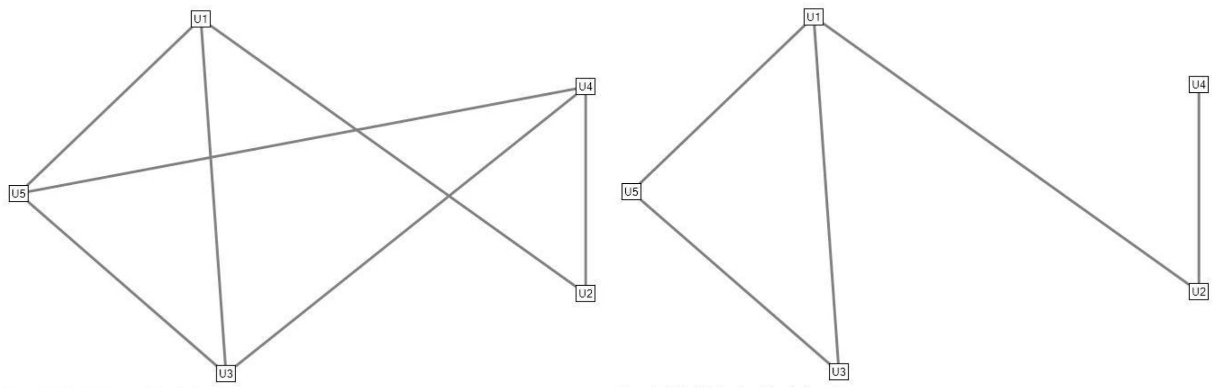
In this way *Precision* can be defined as the number of common edges retrieved between two graphs divided by the total number of edges retrieved in the source network. Furthermore, *Recall* can be defined as the number of common edges retrieved between graphs divided by the total detected edges from the secondary network. For example, the precision and recall between two networks A and B will be as follows: Let A **network**: is the network that is extracted from the experimental study. B^* **network**: is a subset network from B network. This subset represents the experimental participants' self-reported social network. $|E(A)|$: is the number of edges that A network includes. $|E(B^*)|$: is the number of edges that B^* network includes.

$$Pr(A, B^*) = \frac{\mid E(A) \cap E(B^*) \mid}{\mid E(A)) \mid} \tag{7.4}$$

$$R(A, B^*) = \frac{\mid E(A) \cap E(B^*) \mid}{\mid E(B^*) \mid} \tag{7.5}$$

$$F_1(A, B^*) = 2 * \frac{Pr(A, B^*).R(A, B^*)}{Pr(A, B^*) + R(A, B^*)} \tag{7.6}$$

Illustrative Example



(a) Network A (b) Network B^*

Figure 7.3: Example Networks.

In Figure 7.3, There are two networks where network A represents the network that has been extracted from the experimental study in Figure 7.3(a) and network B^* represents the network that has been detected from the survey study. In order to find the correlation between these two networks, precision and recall have been applied as follows:

$E(A) = \{e_{12}, e_{24}, e_{34}, e_{45}, e_{15}, e_{35}, e_{13}\}$

$E(B^*) = \{e_{12}, e_{24}, e_{15}, e_{13}, e_{35}\}$ So

$$Pr(A, B^*) = \frac{\mid E(A) \cap E(B^*) \mid}{\mid E(A)) \mid} = \frac{5}{7} = 0.714$$

$$R(A, B^*) = \frac{\mid E(A) \cap E(B^*) \mid}{\mid E(B^*) \mid} = \frac{5}{5} = 1$$

$$F_1(A, B^*) = 2 * \frac{Pr(A, B^*).R(A, B^*)}{Pr(A, B^*) + R(A, B^*)} = 2 * \frac{0.714}{1.714} = 0.833$$

In this example, we can see that the network A is able to recall all the edges in network B^*. However, not all the retrieved edges are relevant as they are not included in network B^*. In order to have accuracy of retrieval, F_1 measure has been calculated where F_1 has values between 0 and 1. As long as the result is near to the value 1, this means that the accuracy of the retrieval is good quality. In the example, F_1 has the value 0.8 which means the retrieval is good quality because only two edges have been retrieved that did not represent any edge of the relevant network.

7.2.2 Correlation between Opportunistic and Social Networks

For each of the social networks, precision and recall are applied in order to extract the correlation between these and the opportunistic networks. The following figures show the correlation of each social network separately with the three opportunistic networks. Figure 7.4 describes the correlation between the face-to-face social network G^*_{c1} as B^* and the opportunistic networks as A in Equation 7.4, 7.5 and 7.6. In Figure 7.5 explores the correlation between different opportunistic networks as A and the course-mate social network $G^*_{R_3}$ as B^*. The correlation between friendship social network $G^*_{R_2}$ as B^* and the opportunistic networks as A are seen in Figure 7.6. Figure 7.7 shows the correlation between strong friendship social network $G^*_{R_1}$ as B^* and opportunistic networks as A in Equation 7.4, 7.5 and 7.6. From applying these networks under precision and recall, the following was found:

- **Of the approximations to opportunistic networks the co-located opportunistic network is the best in similarity to the social network**

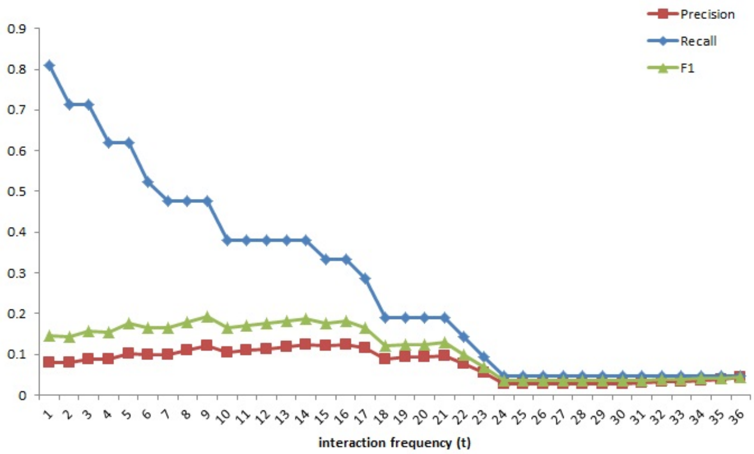
  As seen in Figure 7.4(a), it was found that the co-located opportunistic network retrieves most of the face-to-face social network where it recalls around 80% of the social network. This means that the co-located opportunistic network contains the face-to-face social network more than other experimental opportunistic networks as the co-located social network represents the actual interaction between the participants. The co-located opportunistic network is more reliable in representing the physical interaction to a greater degree than other approximations of the opportunistic network (trajectory and duration).

- **The co-located opportunistic network is successful in recalling more of the strong-friendships than other types of relationships**

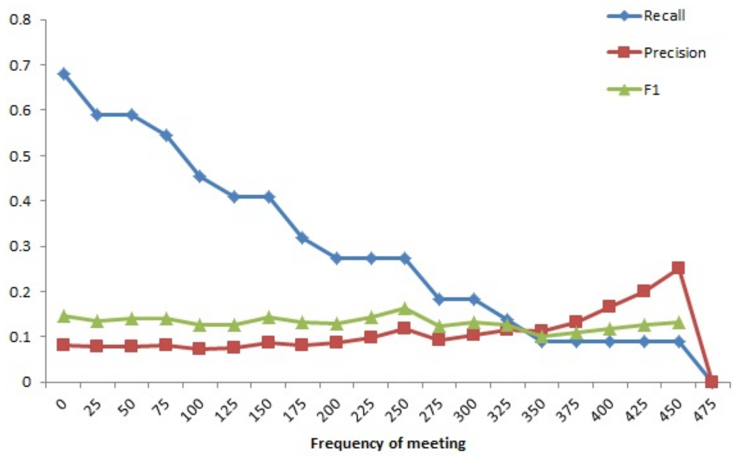
In Figure 7.6 it is shown that the co-located social network best captures the relationship of strong friendship. By increasing the threshold co-located opportunistic network, we are still able to recall more strong friendship relationships than other relationship strengths, as seen in Figures 7.5(a), 7.6(a) and 7.7(a). At the frequency 200, we can see that strong-friendship has been recalled steadily by increasing the frequency while other social networks are decreased in recalling at the level of frequencies. This indicates that strong friendships prefer to communicate physically outside the school building which the co-location opportunistic network captures.

- **All opportunistic networks have approximately the same accuracy of retrieving the face-to-face and relationship social networks.**

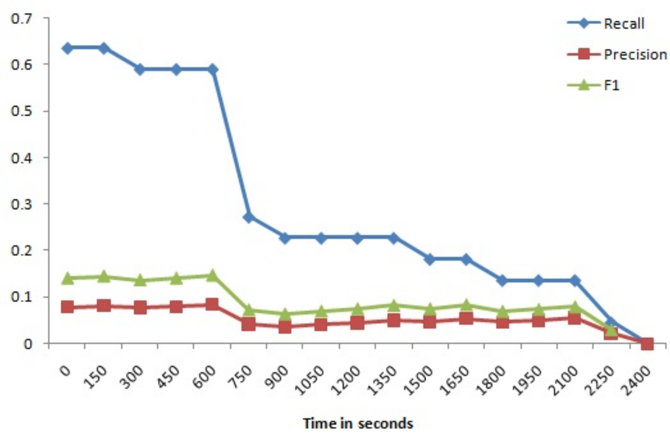
  From the F_1 metric, we found that all the F_1 values for different opportunistic networks ranged from 0.1 to 0.2 where all the self-reported social networks have the same level of retrieving accuracy which is 10-20%. This means that all the networks retrieve large numbers of edges to form opportunistic networks where some of these edges are common with the static social networks' edges. This indicates the collected data includes a lot of edges are included from the participants' mobility data.



(a) Co-located

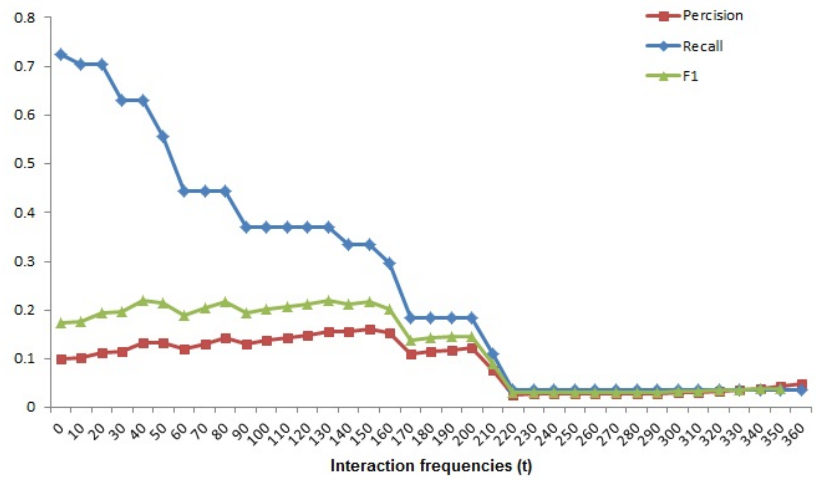


(b) Trajectory

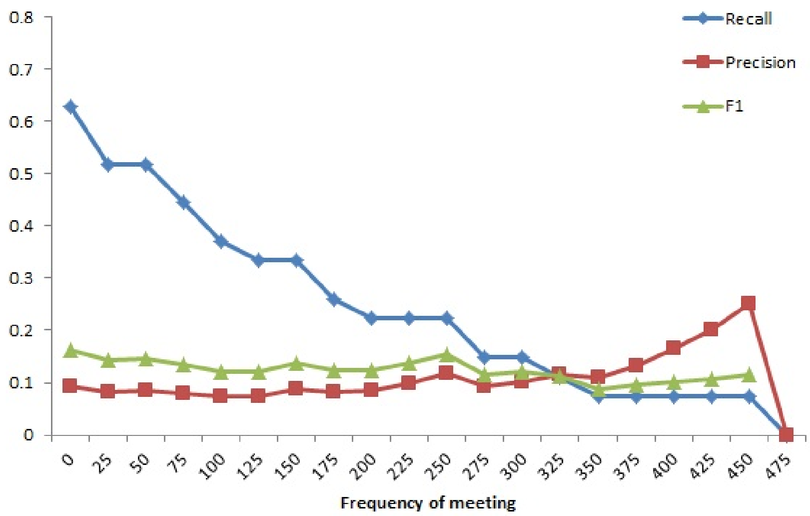


(c) Duration

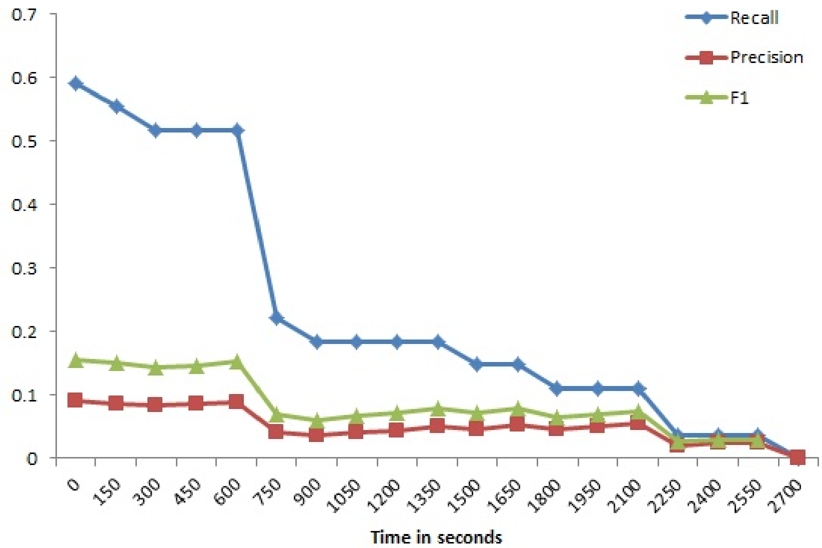
Figure 7.4: Precision/recall for face-to-face social network with Opportunistic Networks .



(a) Co-located

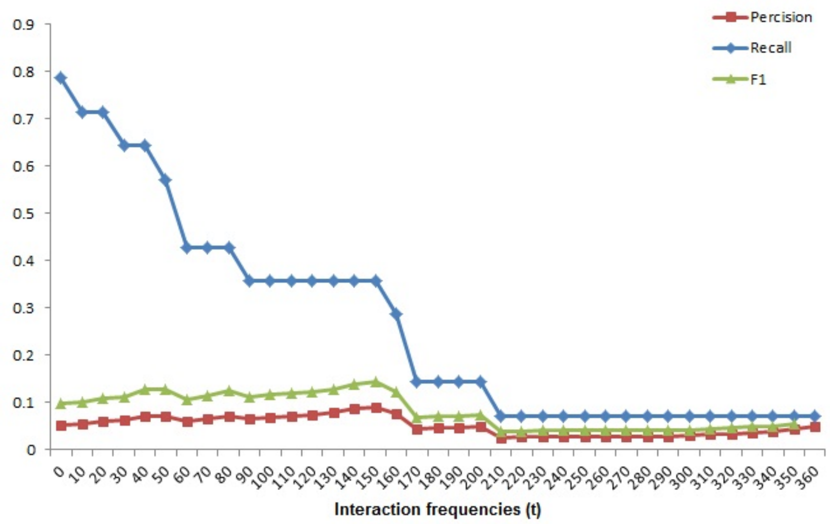


(b) Trajectory

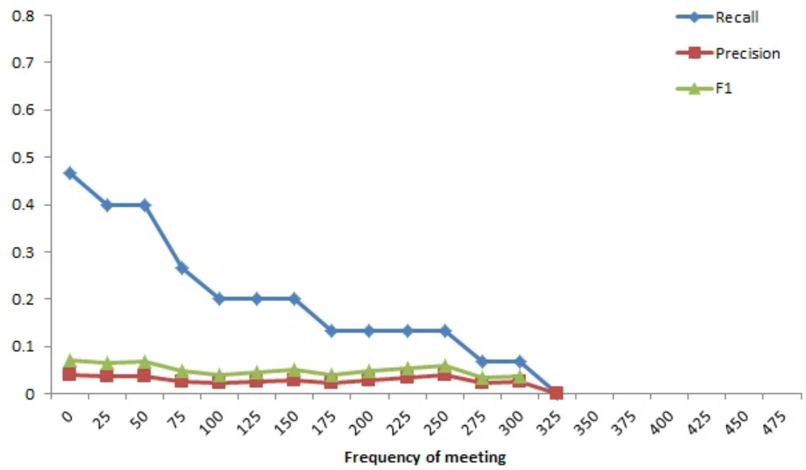


(c) Duration

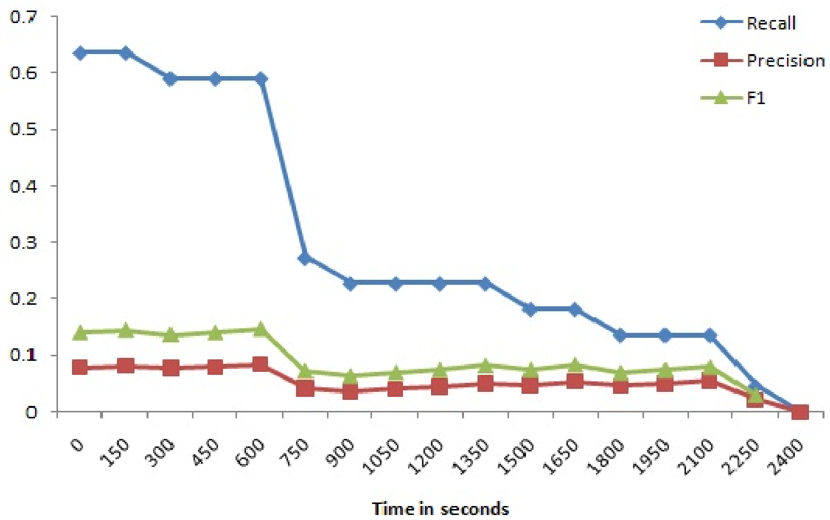
Figure 7.5: Precision/recall for Course-mate with Opportunistic Networks.



(a) Co-located

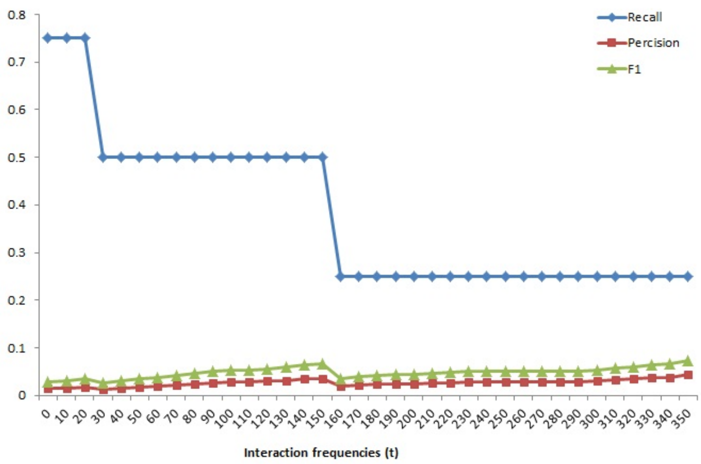


(b) Trajectory

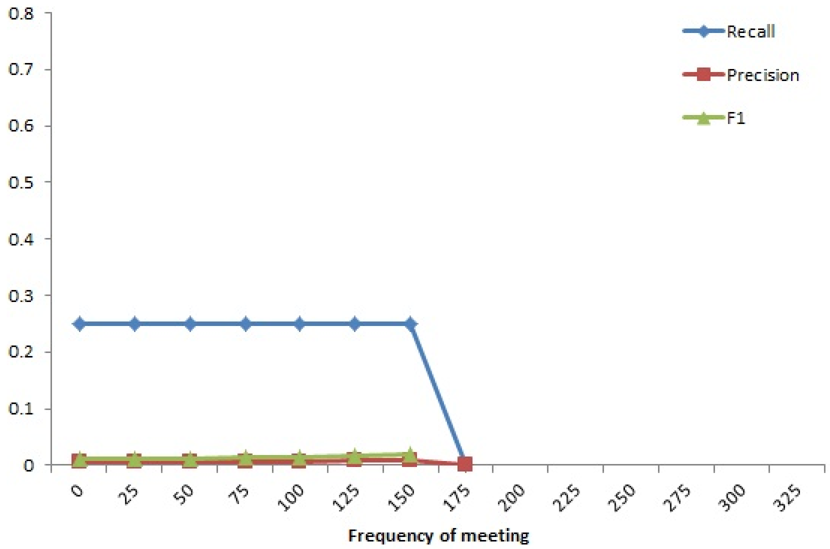


(c) Duration

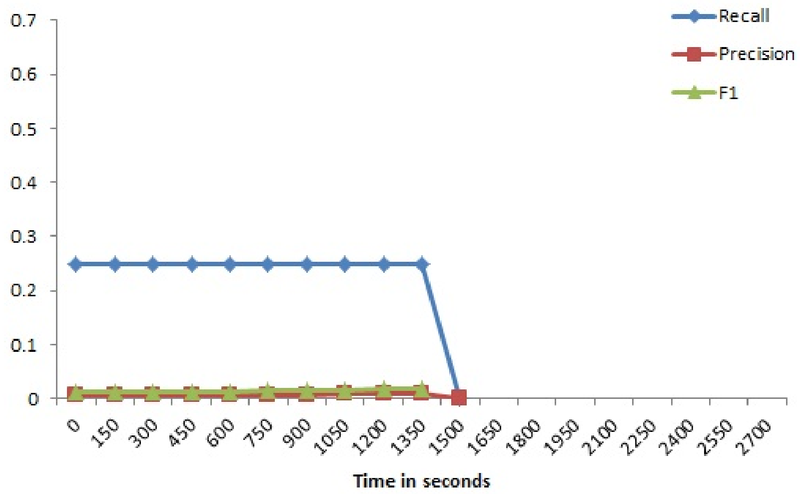
Figure 7.6: Precision/recall for Friendship with Opportunistic Networks.



(a) Co-located



(b) Trajectory

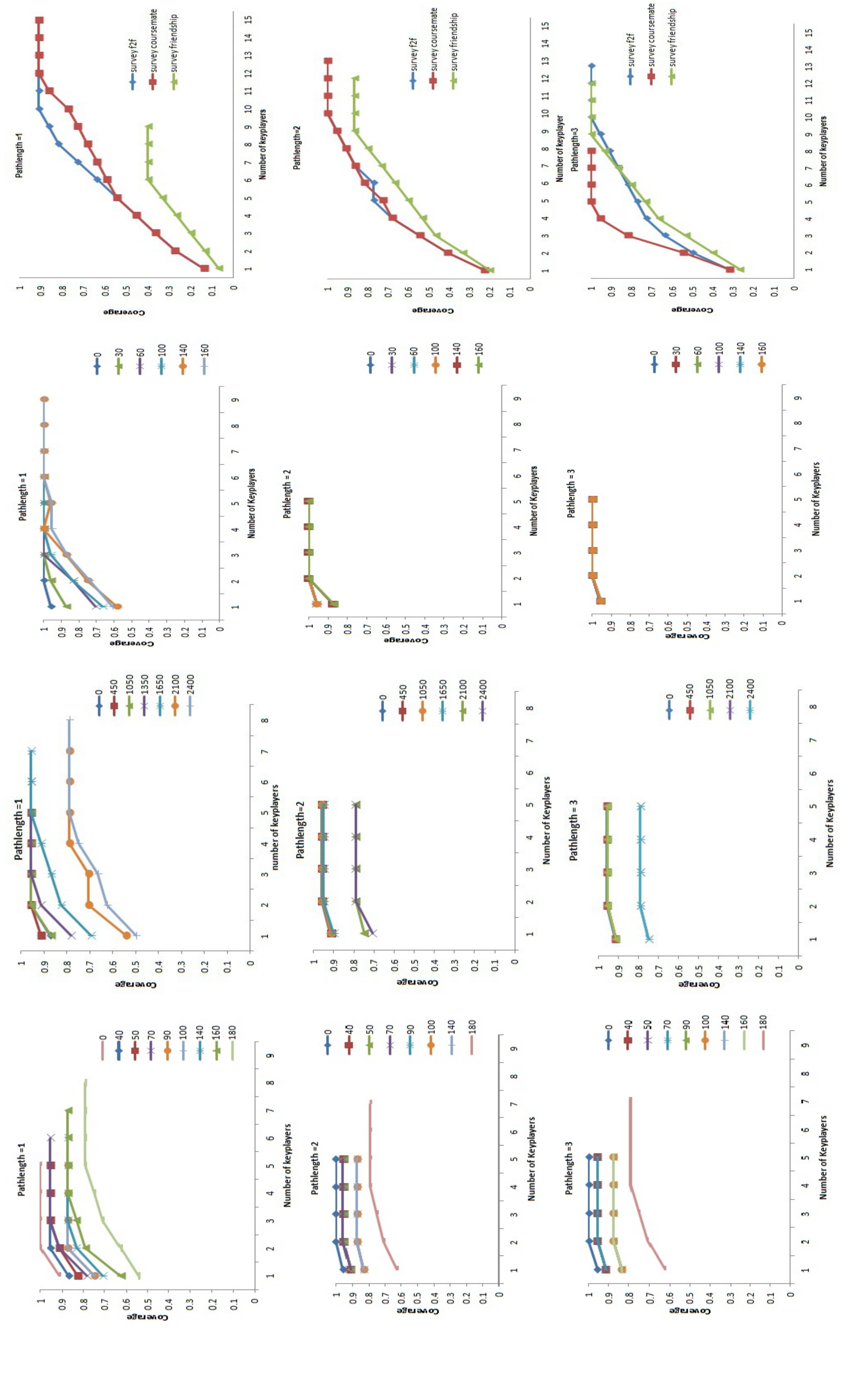


(c) Duration

Figure 7.7: Precision/recall for Strong Friendship with Opportunistic Networks.

7.3 Dissemination Properties of Opportunistic and Social Networks

As detailed in Chapters Five and Six, in comparing the connectivity of the opportunistic networks and self-reported social networks, we found that all the opportunistic networks are more highly connected than self-reported social networks, as seen in Figure 7.8. From this result, small numbers of key-players are required to cover the opportunistic networks. This means that information potentially disseminates more easily in the extracted opportunistic networks than in social network. To explore this further, we looked at the total number of hops required from any node to provide network flooding. This provides a way of comparing the different networks and we can approximately model time. Note that the previous two techniques of finding the similarities (precision and recall) were applied to the relationship and physical interaction social networks. In this section, we focus on the frequency of communication instead of the relationship strength between participants because this roughly models time and can be compared in some approximate way to opportunistic networks. The social network is the network that depends on a brain without technologies where opportunistic network is the network that depends on technologies without brain.



(a) Key-players of Trajectory Network (b) Key-players of Duration Network (c) Ke-yplayers of Co-located network (d) Keyplayers of Self-reported network

Figure 7.8: Key-players of Different Social Networks at Different Threshold.

From the analysis, we found that the first six days (first three slots) of the opportunistic network data collection had very active patterns of movements as compared to the last three slots where the timetable changed for the students. As a result, we focused on the first six days of the experimental data collection for the opportunistic networks. For each day of the experimental study, we had two different opportunistic networks. The first network represents the mobility of the participants inside the school building, known as *mobility opportunistic network*. This mobility network represents the duration of the opportunistic network as we need to assure that the participants spend time together because this analysis is based on transfer information that requires a time to be transferred. The second network represents the physical interaction between the participants inside and outside the school building, known as *co-located opportunistic network*.

In order to consider information dissemination for each node in each network, we imagine sending out a packet of information and calculate how many hops are required for this packet to reach everyone in the network: the network flooding scenario. Figure 7.9 , 7.10 and 7.11 show the output of the number of hops. Figure 7.11, shows the frequency distribution of delivering a packet to all networks' nodes for each day of the experiment (co-located opportunistic network). Figure 7.9 explores the results of the information flooding analysis on the social network, the structure of which is seen in Appendix A.2. For each day, a network structure graph was detected to assess the information flooding analysis, as seen in Appendix B.2. In Figure 7.10, frequency distribution of the information flooding analysis is applied to the mobility opportunistic network, where each day has a different network structure, as seen in Appendix B.3.

From Figure 7.9, A.10, A.11 and A.12 we can see that although the daily basis has a lot of one hop frequency but the daily basis network do not cover the whole nodes in the network. On the other hand, for "daily and few-times-a-week network" and "daily, few-times-a-week and monthly", these networks cover the whole nodes but the information need to travel for a long time through the network to be able to deliver the information for each node in the network.

From Figure 7.10 and 7.6(a), the one and two hops are dominant in the mobility opportunistic network where two and three hops are dominant in the co-located opportunistic network. From the different figures at Appendix B.2 and B.3, we find that the connectivity of the co-located network everyday are more connected than the mobility network. The findings of this analysis are as follows:

- **The co-located opportunistic network is the fastest in terms of disseminating information.**

  It was found that the co-located opportunistic network are likely to be the fastest network for sending information. This confirms the previous findings related to the connectivity

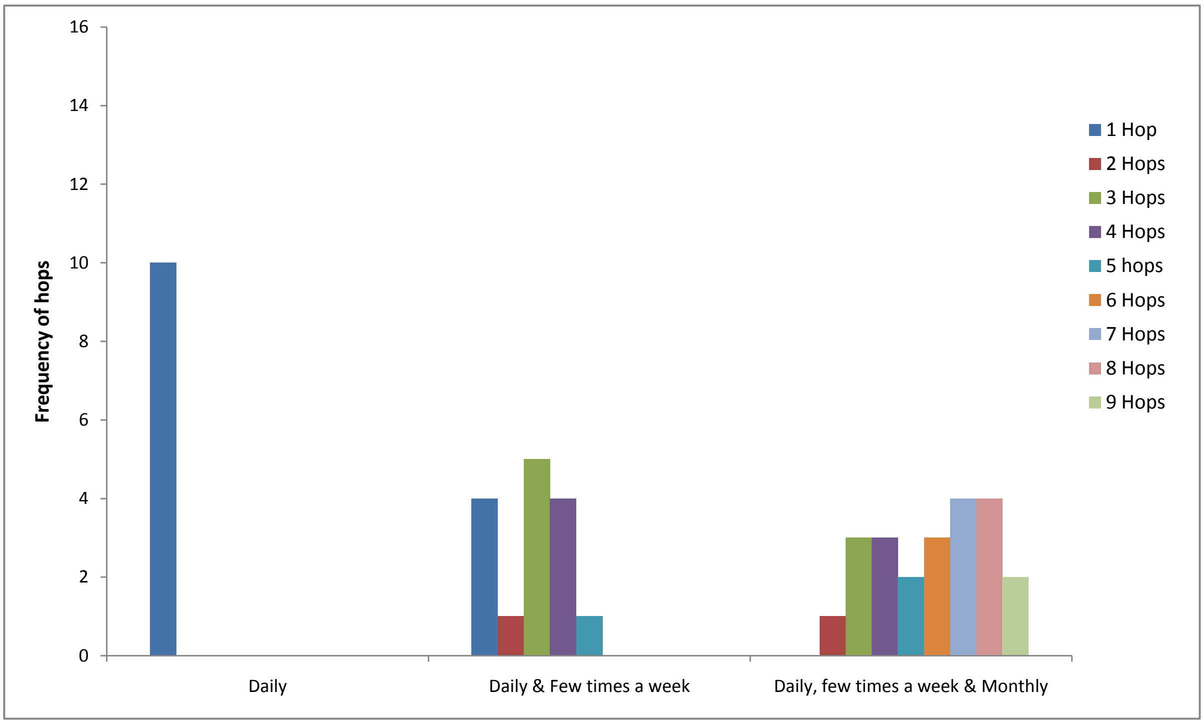


Figure 7.9: Hops Frequency Distribution for delivering a Packet to all Nodes in Self-reported Social Network.

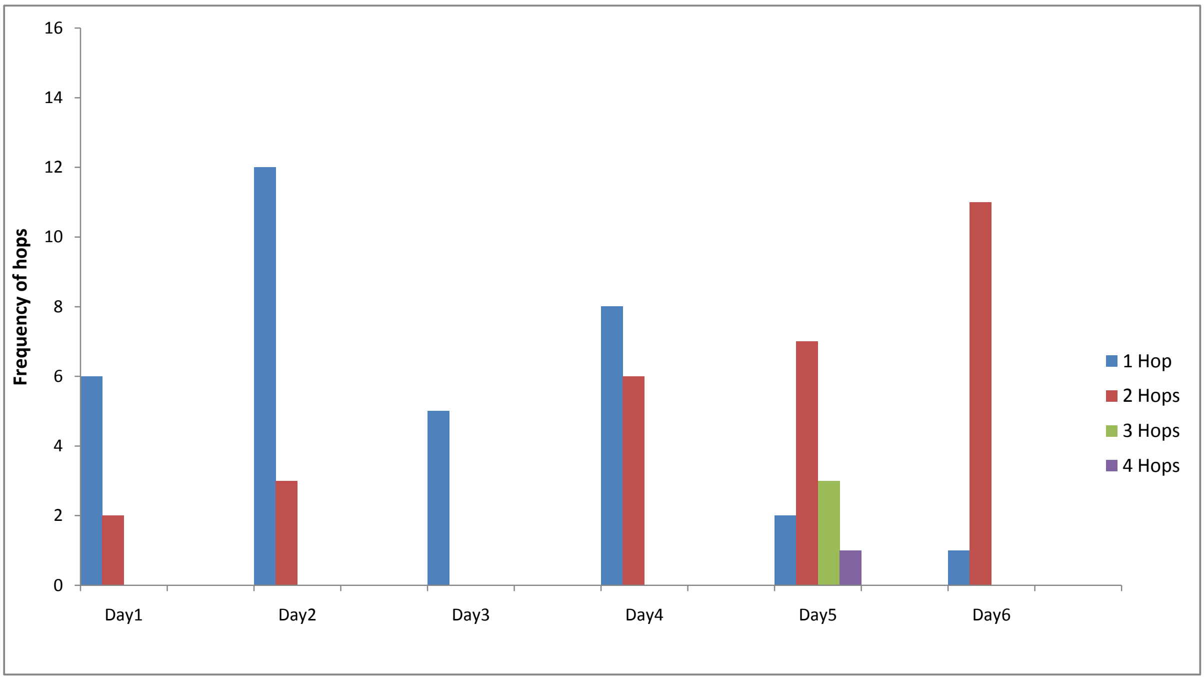


Figure 7.10: Hops Frequency Distribution for delivering a Packet to all Nodes in Mobility Network.

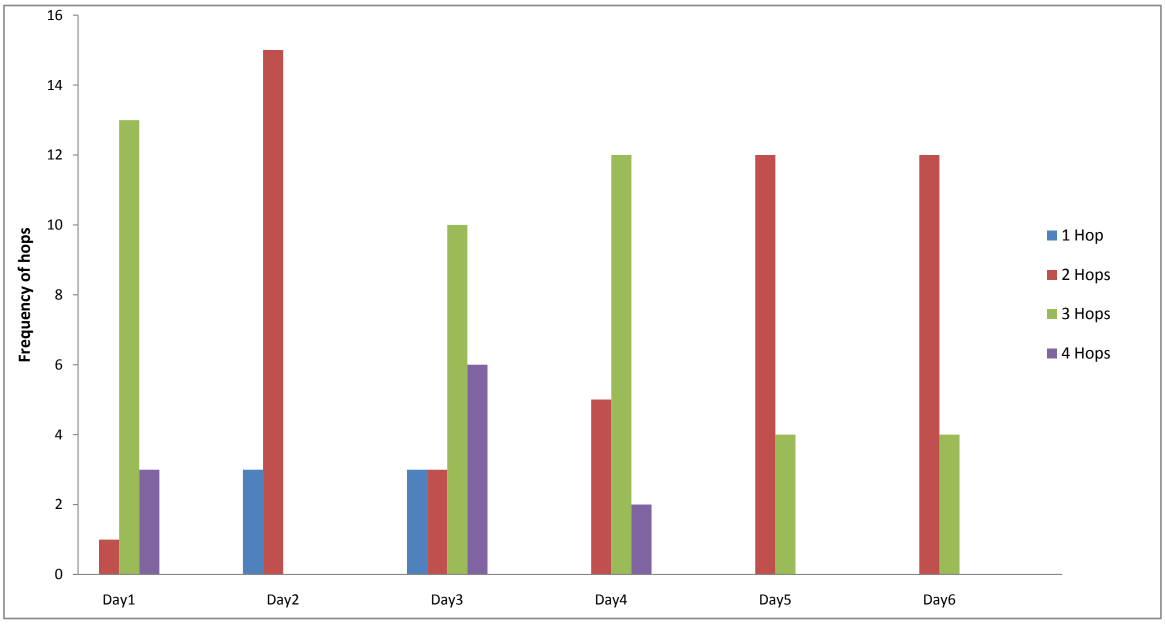


Figure 7.11: Hops Frequency Distribution for Delivering a Packet to All Nodes in Co-located Network.

of this network. This is because this network includes the participants' interaction in and out the school building, as seen in Figure 7.11.

- **Opportunistic networks are not affected by the strength of the relationship between the participants.**

  Not only are most of the relationship strengths recalled by the co-located opportunistic network, but also lots of indescribable weak links are included as well. This means that the wireless device records everything occurring in range where these devices are affected and supported by the weak/unknown relationship links from the participants' movements which increases its connectivity.

- **Weak links play a significant role in disseminating information.**

  In Figure 7.11 and 7.10, this analysis concludes that indescribable or unknown relationships (weak links) are able to effectively disseminate information indoors. This substantiates the finding that the relationship strength between participants does not affect information dissemination.

- **Mobility opportunistic network is not effective in disseminating information.**

It was found that tracking the indoor mobility pattern of the participants indoors is not sufficient as people normally interact together wherever they are (inside or outside the building). We found that carrying a wireless device is better method for information dissemination where devices can see each other easily than servers can do.

7.4 Summary

This chapter investigated the correlation between a human self reported social network and the opportunistic network that comes from their mobility and interactions. It explored the similarities and differences that exist between the opportunistic networks that devices see and the social networks that exist between humans. This was a challenging task because the networks exist for different purposes, with opportunistic networks being temporal in nature while social networks are a static representation based on the perception of individuals. Therefore, we looked at the differences between these networks using different techniques.

Using precision and recall, we found that 60 - 80 % of the static social networks were embedded into the opportunistic network in addition to other unknown relationship links between the participants extracted from their movements, as seen in Figures 7.4, 7.5, 7.6 and 7.7. On the other hand, a very small proportion 10-20% of the opportunistic network is embedded in the self-reported social network. This renders the opportunistic network to be highly connected in structure to disseminate information. From this result, a small number of key-players are required to cover the opportunistic networks, as seen in Figure 7.8. This means that information potentially disseminates more easily in the extracted opportunistic networks than in self-reported social networks.

A flooding analysis technique was used in order to detect the correlation between networks in terms of time. We focused on the frequency of communication instead of the relationship strength between participants because this roughly models time and can be compared in some approximate way to opportunistic networks. We found that device-to-device analysis provided a full picture in terms of information dissemination rather than server detection, as seen in Figure 7.11 and 7.10. This is because device to device records all communication regardless of the location of participants. Although the opportunistic network is approximate, it floods the network with information quickly. This is due to the remarkable role of the unknown relationship weak links in disseminating information. Also, as the co-located opportunistic network are more highly connected than the self-reported social network, information floods speedily through the opportunistic networks, as seen in Figure 7.11 and 7.9. Devices can be added to humans to speed up information dissemination.

Chapter 8

Conclusion and Future Work

Overview

To conclude this research, this chapter provides a summary of the thesis with the possible future direction of this research. Initially, the chapter explores the thesis contributions and challenges. It ends by focusing on future work that is related to this research.

8.1 Thesis Findings

Returning to the thesis hypothesis, this thesis investigated, in a local context, the relationship between a group's physical mobility and social structure. The aim was to uncover the extent to which opportunistic networking can utilize the relationships and frequencies of communication that exist between humans who carry wireless devices. In addition, the thesis aimed to reveal the differences between the opportunistic network that devices see and social networks between humans. In order to achieve this, two main stages were implemented. Firstly, we developed a bespoke distributed system that would be able to automatically record the mobility movement and physical interaction between users using Bluetooth technology (Bluetooth mote and dongle Bluetooth). Secondly, we conducted an electronic survey to detect the groups' social network from their perspective. This study was conducted with a large group of Computer Science students.

In Chapters Five and Six, it was concluded that opportunistic network is different in terms of structure and connectivity of the participants' social network. Chapter Six concluded that most strong relationship strengths prefer to utilize the advanced technologies for communication rather than using physical interaction. Additionally, in Chapter Five, there was only one opportunistic network. There were two ways to approximate it: one device-to-device and one device-to-server in building. While one network presents the mobility inside the building (device-to building), the other network presents the physical interaction anywhere (device-to-device).

From the conclusion drawn in Chapter Seven was that the social network that opportunistic devices (device-to-device and device-to-building) see is very different to human social networks. The human social network is embedded in the opportunistic networks and it sees many more weak links. This means that routing and dissemination algorithms for opportunistic networks need to be able to deal with cooperation, trust, privacy and security issues that help to harness the resources of devices that are seen on weak links.

8.2 Thesis Contribution

Some of the main contributions made by this thesis as follows

- **Developing Mobility Tracking System** - (Chapter Three) We presented, in detail, the design and development of an indoor mobility tracking system to meet all the user and system requirements that were generated for research purposes. It was a challenge to create an opportunistic network and, based on our knowledge, there are no opportunistic network platforms available other than the prototype developed in the haggle project. Therefore, we took a different approach and looked at developing a system to detect opportunities for peer-to-peer transmissions. A detailed framework of the system was proposed. We detailed the requirements, design and configuration of the system for deployment and capture of mobility data in Chapter Three.

- **Conducting Electronic Survey for Building Social Network** - (Chapter Four) We used the electronic survey as a research methodology to uncover the social network of a group of participants. It was required to understand the participants' social network from their perceptions. We mainly focused on the design and the implementation of the survey in order to collect the ego-centric social information about the participants.

- **Extraction of different Opportunistic network from Mobility Data** - (Chapter Five) From the mobility data collected in Chapter Three, different algorithms were developed to approximate the extraction of different opportunistic networks (devices seeing device and devices seen in range building access points). In the developing work, we also learned about participants' use of the building space.

- **Modification to Key-players Algorithm** - (Chapter Six) This algorithm was modified and used in the analysis of a self-reported data set. The algorithm was originally developed for an undirected social network. We modified the algorithm to be suitable for directed graph. We also learned about the participants' networking habits in this chapter.

- **System Evaluation** - (Chapter Seven) It was challenging to compare different types of networks together as we had social networks and opportunistic networks. Information retrieval evaluation concept was used to evaluate the system by understanding the correlation between the self-reported social network and the co-located and mobility opportunistic networks. Information dissemination technique was used to find the similarities between temporal opportunistic and frequency of interaction survey networks. The drawback regarding the opportunistic network was its approximation due to the limitation of the tiny Bluetooth device as it did not feature a clock to record the time of interaction. However, we found that utilizing this wireless device was the most effective way to detect the opportunistic network between the participants wherever they were.

8.3 Achievements

In conducting this research, many challenges have been overcome. Bluetooth technology was adapted as it proved its applicability in terms of localizing people indoors rather than GPS. A challenging aspect was that smart phones limit the discoverability option for Bluetooth services.

Following recent related research, we found that Bluetooth was suitable sensor for this particular research as it is able to sense the physical interaction between users besides the ability to be utilized for tracking users indoors. Bluetooth can be used to detect the presence of users inside the building with the accuracy of room level. Besides the previous features, we utilized Bluetooth as it is embedded in most personal devices, such as mobile phones, laptops and tablets.

Chapters Three, Four, Five and Six detailed the research study of this thesis and the results were evaluated in Chapter Seven. It was a challenge to determine the requirements of the mobility tracking system where this system was developed to be automatic. We were the only users of the system. Many difficulties were faced while developing the system, such as Bluetooth security that which has been added to new mobile phones for more privacy, Bluetooth programming on desktop machines instead of mobile phones, programming the Bluetooth mote when its manual was not useful, and finally, recording the physical interaction between users via FTP protocol.

Each of these challenges was managed and overcome and the system was piloted for the experimental study. Each component of the system was tested separately. The system with all its components was tested and it achieved its functionality perfectly. The developed system focused on two types of data: the users' indoor movements and their physical interaction. It included two different applications that collected the required data set. The collected data was stored in a server that included a central database.

A number of other challenges were faced during the survey study. This study mainly focused on the design and implementation of the survey in order to collect the ego-centric social information about the participants. A survey was chosen as a method of information collection as it is considered the cheapest and fastest method for collection. This survey was conducted on first year undergraduate students, therefore it was important to form questions that would be understandable for this particular group. It was challenging to convince participants to participate in this study and provide their personal social information.

Once the results of the survey were analysed, it was found that social network analysis plays an important role in understanding the characteristics of any type of data that represents interaction between people, such as phone-calls, email, Facebook and physical interaction. It was found that it is challenging to apply social network analysis in indoor mobility tracking to extract the users' social network that has not been found in other related research. In addition, other previous research did not apply social network analysis on the recorded physical interaction between users. Therefore, the dynamics and mobility patterns of users were studied along with the recorded physical interaction to find a correlation between users' social relationships and their opportunistic network. All of these challenges were addressed throughout this research.

One of novelties of this study was the comparison between different social networks and opportunistic networks for the same group of people. The challenge was to differentiate types of networks together as there are social networks and opportunistic networks. Information retrieval evaluation concept was used to evaluate the system by understanding the correlation between the self-reported social network and the co-located and mobility opportunistic networks. Precision and recall were applied to the relationship and physical interaction social networks.

Information flooding technique was used to find the similarities between temporal opportunistic and frequency of interaction survey networks. This was applied to the frequency of communication instead of the relationship strength between participants because this roughly models time and can be compared in an approximate way to opportunistic networks.

In terms of precision and recall, we found that 60 - 80 % of the static social networks are embedded into the opportunistic network in addition to other indescribable links between the participants extracted from their movements. This means that the opportunistic network is highly connected in structure. Based on this result, a small number of key-players are required to cover the opportunistic networks. This means that information potentially disseminates more easily in the extracted opportunistic networks than in social networks.

As a result, information dissemination technique was used in order to detect the correlation between networks in terms of time. We focused on the frequency of communication instead of the relationship strength between participants because this roughly models time and can be compared in an approximate way to opportunistic networks. We found that device-to- device provides a full picture of information dissemination rather than server detection. This is because device-to-device records all communication regardless of the location of participants. Although the opportunistic network is approximate, it quickly floods the network with information. This is due to the remarkable role of the indescribable weak links in disseminating information. Furthermore, as the co-located opportunistic network is highly connected than the social network, information speedily floods through the opportunistic network.

The research in this thesis provides a framework and analysis for the development of Mobility and Physical interaction tracking systems. For current direction of this research, it could be further advanced by utilize the finding of this research for the benefit of information dissemination inside organization. This could be established by using wireless devices such as Bluetooth mote. What's more, these interesting findings can be used to optimize the routing protocol and information provision of the opportunistic network. Where information provision in opportunistic network is challenges due to the lack of network topology. Also the finding can be utilized to achieve efficient and effective method for broadcasting in opportunistic network to minimise resources duplication and utilization of data transmission.

Bibliography

[1] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *Privacy Enhancing Technologies*, pages 36–58. Springer, 2006.

[2] Divyakant Agrawal, Amr El Abbadi, and Robert C Steinke. Epidemic algorithms in replicated databases. In *Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 161–172. ACM, 1997.

[3] Florian Alt, Jorg Muller, and Albrecht Schmidt. Advertising on public display networks. *Computer*, 45:50–56, 2012.

[4] G. Anastasi, R. Bandelloni, M. Conti, F. Delmastro, E. Gregori, and G. Mainetto. Experimenting an indoor bluetooth-based positioning service. In *Distributed Computing Systems Workshops, 2003. Proceedings. 23rd International Conference on*, pages 480–483. IEEE, 2003.

[5] M. Anne, J.L. Crowley, V. Devin, and G. Privat. Localisation intra-bâtiment multi-technologies: Rfid, wifi et vision. In *Proceedings of the 2nd French-speaking conference on Mobility and ubiquity computing*, pages 29–35. ACM, 2005.

[6] S. Aparicio, J. Perez, AM Bernardos, and JR Casar. A fusion method based on bluetooth and wlan technologies for indoor location. In *Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008. IEEE International Conference on*, pages 487–491. IEEE, 2008.

[7] P. Bahl and V. N. Padmanabhan. RADAR: an in-building RF-based user location and tracking system. In *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 2, pages 775–784, 2000.

[8] C. Ballester, A. Calvó-Armengol, and Y. Zenou. Who's who in networks. wanted: the key player. *Econometrica*, 74(5):1403–1417, 2006.

[9] Mortaza S. Bargh and Robert de Groote. Indoor localization based on response rate of bluetooth inquiries. In *Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments*, pages 49–54. ACM New York, NY, USA, 2008.

[10] Allan Beaufour, Martin Leopold, and Philippe Bonnet. Smart-tag based data dissemination. In *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*, pages 68–77. ACM, 2002.

[11] M.T. Blanche, K. Durrheim, and D. Painter. *Research in practice: Applied methods for the social sciences*. Juta Academic, 2008.

[12] Chiara Boldrini, Marco Conti, Jacopo Jacopini, and Andrea Passarella. Hibop: a history based routing protocol for opportunistic networks. In *World of Wireless, Mobile and Multimedia Networks, 2007. WoWMoM 2007. IEEE International Symposium on a*, pages 1–12. IEEE, 2007.

[13] S.P. Borgatti. Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory*, 12(1):21–34, 2006.

[14] R. Bruno and F. Delmastro. Design and analysis of a bluetooth-based indoor localization system. In *Personal Wireless Communications*, pages 711–725. Springer, 2003.

[15] J. J. Caffery and G. L. Stuber. Overview of radiolocation in CDMA cellular systems. *Communications Magazine, IEEE*, 36(4):38–45, 1998.

[16] S. S Chawathe. Beacon placement for indoor localization using bluetooth. In *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, pages 980–985, 2008.

[17] K.C. Cheung, S.S. Intille, and K. Larson. An inexpensive bluetooth-based indoor positioning hack. *Proc. UbiComp06 Extended Abstracts*, 2006.

[18] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1082–1090, New York, NY, USA, 2011. ACM.

[19] A. Clauset and N. Eagle. Persistence and periodicity in a dynamic proximity network. In *DIMACS Workshop on Computational Methods for Dynamic Interaction Networks*, pages 1–5, 2007.

[20] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, Ubicomp '10, pages 119–128, New York, NY, USA, 2010. ACM.

[21] R. Cross, A. Parker, L. Prusak, and S.P. Borgatti. Knowing what we know: Supporting knowledge creation and sharing in social networks. *Organizational Dynamics*, 30(2):100–120, 2001.

[22] A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the web. 2004.

[23] M.S. Dahl and C.Ø.R. Pedersen. Knowledge flows through informal contacts in industrial clusters: myth or reality? *Research Policy*, 33(10):1673–1686, 2004.

[24] E.W. Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[25] D.A. Dillman. *Mail and telephone surveys*. Wiley New York, 1978.

[26] D.A. Dillman. The design and administration of mail surveys. *Annual review of sociology*, pages 225–249, 1991.

[27] T.M.T. Do and D. Gatica-Perez. Contextual grouping: discovering real-life interaction types from longitudinal bluetooth data. In *Mobile Data Management (MDM), 2011 12th IEEE International Conference on*, volume 1, pages 256–265. IEEE, 2011.

[28] Zhengbin Dong, Guojie Song, Kunqing Xie, Yixian Sun, and Jingyao Wang. Adequacy of data for mining individual friendship pattern from cellular phone call logs. In Yixin Chen, Hepu Deng, Degan Zhang, and Yingyuan Xiao, editors, *FSKD (5)*, pages 573–577. IEEE Computer Society, 2009.

[29] N. Eagle, A. Pentland, and D. Lazer. Mobile phone data for inferring social network structure. *Social computing, behavioral modeling, and prediction*, pages 79–88, 2008.

[30] N. Eagle, A.S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[31] Nathan Eagle and Alex Sandy Pentland. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, April 2009.

[32] Nathan Eagle, Alex (Sandy) Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36):15274–15278, September 2009. PMID: 19706491 PMCID: PMC2741241.

[33] S. Escalera, X. Baró, J. Vitria, P. Radeva, and B. Raducanu. Social network extraction and analysis based on multimodal dyadic interaction. *Sensors*, 12(2):1702–1719, 2012.

[34] J. Euzenat. Semantic precision and recall for ontology alignment evaluation. In *Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 348–353, 2007.

[35] T. Falkowski. Community Analysis in Dynamic Social Networks. 2009.

[36] K. Fischbach, P.A. Gloor, C. Lassenius, D.O. Olguin, A.S. Pentland, J. Putzke, and D. Schoder. Analyzing the Flow of Knowledge with Sociometric Badges. *Procedia-Social and Behavioral Sciences*, 2(4):6389–6397, 2010.

[37] E. Furey, K. Curran, and P. Mc Kevitt. HABITS: a history a ware based wi-fi in door tracking s ystem. 2008.

[38] S. Gaito, E. Pagani, and G.P. Rossi. Strangers help friends to communicate in opportunistic networks. *Computer Networks*, 55(2):374–385, 2011.

[39] G. Gallo and S. Pallotino. Shortest path methods in transportation models. *Publication of: Elsevier Science Publishers BV*, 1984.

[40] M. Gaved and P. Mulholland. Grassroots initiated networked communities: A study of hybrid physical/virtual communities. 2005.

[41] J. Gips and A. Pentland. Mapping human networks. In *Pervasive Computing and Communications, 2006. PerCom 2006. Fourth Annual IEEE International Conference on*, page 10. IEEE, 2006.

[42] Natalie Glance, Dave Snowdon, and Jean-Luc Meunier. Pollen: using people as a communication medium. *Computer Networks*, 35(4):429–442, 2001.

[43] B. Guo, D. Zhang, Z. Yu, and F. Calabrese. Extracting social and community intelligence from digital footprints. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–2, 2012.

[44] D. Hahnel, W. Burgard, D. Fox, K. Fishkin, and M. Philipose. Mapping and localization with RFID technology. In *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, volume 1, pages 1015–1020, 2004.

[45] A. Harter, A. Hopper, P. Steggles, A. Ward, and P. Webster. The anatomy of a context-aware application. *Wireless Networks*, 8(2):187–197, 2002.

[46] S. Hay and R. Harle. Bluetooth tracking without discoverability. *Location and context awareness*, pages 120–137, 2009.

[47] C. Haythornthwaite. Online personal networks size, composition and media use among distance learners. *New Media & Society*, 2(2):195–226, 2000.

[48] C. Haythornthwaite. Social networks and internet connectivity effects. *Information, Community & Society*, 8(2):125–147, 2005.

[49] R. Helms and K. Buijsrogge. Application of knowledge network analysis to identify knowledge sharing bottlenecks at an engineering firm. In *Proceedings of the 14th European conference on information systems*, 2006.

[50] R. Helms, R. Ignacio, S. Brinkkemper, and A. Zonneveld. Limitations of Network Analysis for Studying Efficiency and Effectiveness of Knowledge Sharing. *Electronic Journal of Knowledge Management*, 8(1):53–68, 2010.

[51] J. Hightower, R. Want, and G. Borriello. SpotON: an indoor 3D location sensing technology based on RF signal strength. *UW CSE 00-02-02, University of Washington, Department of Computer Science and Engineering, Seattle, WA*, 1, 2000.

[52] Wei jen Hsu, Debojyoti Dutta, and Ahmed Helmy. Csi: A paradigm for behavior-oriented delivery services in mobile human networks. *CoRR*, abs/0807.1153, 2008.

[53] Wei jen Hsu, Debojyoti Dutta, and Ahmed Helmy. Csi: A paradigm for behavior-oriented profile-cast services in mobile networks. *Ad Hoc Networks*, 10(8):1586–1602, 2012.

[54] S.C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[55] Philo Juang, Hidekazu Oki, Yong Wang, Margaret Martonosi, Li Shiuan Peh, and Daniel Rubenstein. Energy-efficient computing for wildlife tracking: Design tradeoffs and early experiences with zebranet. In *ACM Sigplan Notices*, volume 37, pages 96–107. ACM, 2002.

[56] Gueorgi Kossinets, Jon Kleinberg, and Duncan Watts. The structure of information pathways in a social communication network. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 435–443, New York, NY, USA, 2008. ACM.

[57] V. Kulyukin, C. Gharpure, J. Nicholson, and S. Pavithran. RFID in robot-assisted indoor navigation for the visually impaired. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 2, page 1979âĂŞ1984, 2004.

[58] Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, GIS '08, pages 34:1–34:10, New York, NY, USA, 2008. ACM.

[59] H.L. Lim. Social network analysis: measuring symmetry of information flow in virtual learning groups. In *Proceedings of the 2009 conference on Information Science, Technology and Applications*, pages 40–44. ACM, 2009.

[60] J. Lim, D. Ross, R.S. Lin, and M.H. Yang. Incremental learning for visual tracking. *Advances in neural information processing systems*, 17:793–800, 2004.

[61] J. Lumsden and W. Morgan. Online-questionnaire design: establishing guidelines and evaluating existing support. 2005.

[62] S. Mardenfeld, D. Boston, S.J. Pan, Q. Jones, A. Iamntichi, and C. Borcea. Gdc: Group discovery using co-location traces. In *Social computing (SocialCom), 2010 IEEE second international conference on*, pages 641–648. IEEE, 2010.

[63] A. Martinez, Y. Dimitriadis, B. Rubia, E. Gómez, and P. De La Fuente. Combining qualitative evaluation and social network analysis for the study of classroom social interactions. *Computers & Education*, 41(4):353–368, 2003.

[64] U. Matzat. Reducing problems of sociability in online communities: Integrating online communication with offline interaction. *American Behavioral Scientist*, 53(8):1170, 2010.

[65] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[66] M. Morris, A.E. Kurth, D.T. Hamilton, J. Moody, and S. Wakefield. Concurrent partnerships and hiv prevalence disparities by race: linking science and public health practice. *American Journal of Public Health*, 99(6):1023, 2009.

[67] T. Mueller-Prothmann and I. Finke. SELaKT-social network analysis as a method for expert localisation and sustainable knowledge transfer. *Journal of Universal Computer Science*, 10(6):691–701, 2004.

[68] S. Murthy and J.J. Garcia-Luna-Aceves. An efficient routing protocol for wireless networks. *Mobile Networks and Applications*, 1(2):183–197, 1996.

[69] M. Y. Nam, M. Z. Al-Sabbagh, and C. G. Lee. Real-time indoor human/object tracking for inexpensive technology-based assisted living. In *Proceedings of IEEE real-time systems symposium (RTSS), Rio de Janeiro*, pages 5–8, 2006.

[70] F. Naya, H. Noma, R. Ohmura, and K. Kogure. Bluetooth-based indoor proximity sensing for nursing context awareness. In *Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium on*, pages 212–213. IEEE, 2005.

[71] M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002.

[72] M.E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.

[73] M.E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577, 2006.

[74] Sze-Yao Ni, Yu-Chee Tseng, Yuh-Shyan Chen, and Jang-Ping Sheu. The broadcast storm problem in a mobile ad hoc network. In *Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*, pages 151–162. ACM, 1999.

[75] The University of Texas at Austin. survey tables questiontypes, 2007.

[76] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A.L. Barabasi. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, May 2007.

[77] D. Pandya, R. Jain, and E. Lupu. Indoor location estimation using multiple wireless technologies. In *Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. 14th IEEE Proceedings on*, volume 3, pages 2208–2212. IEEE, 2003.

[78] P. Pantel and D. Lin. Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619. ACM, 2002.

[79] L. Pei, R. Chen, J. Liu, H. Kuusniemi, T. Tenhunen, and Y. Chen. Using inquiry-based bluetooth RSSI probability distributions for indoor positioning. *Journal of Global Positioning Systems*, 9(2):122–130, 2010.

[80] Alex Pentland, Richard Fletcher, and Amir Hasson. Daknet: rethinking connectivity in developing nations. *Computer*, 37(1):78–83, 2004.

[81] S. Phillips, M. Katchabaw, and H. Lutfiyya. WLocator: an indoor positioning system. In *Wireless and Mobile Computing, Networking and Communications, 2007. WiMOB 2007. Third IEEE International Conference on*, pages 33,, 2007.

[82] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan. The cricket location-support system. In *Proceedings of the 6th annual international conference on Mobile computing and networking*, pages 32–43, 2000.

[83] S.H. Robert Harle. Bluetooth Tracking without Discoverability. In *Location and Context Awareness: 4th International Symposium, LoCA 2009 Tokyo, Japan, May 7-8, 2009 Proceedings*, page 120. Springer, 2009.

[84] D. Ross, J. Lim, and M.H. Yang. Adaptive probabilistic visual tracking with incremental subspace update. *Computer Vision-ECCV 2004*, pages 470–482, 2004.

[85] R. Rowe, G. Creamer, S. Hershkop, and S.J. Stolfo. Automated social hierarchy detection through email network analysis. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 109–117. ACM, 2007.

[86] Y. Saab and M. VanPutte. Shortest path planning on topographical maps. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 29(1):139–150, 1999.

[87] Tara Small and Zygmunt J Haas. The shared wireless infostation model: a new ad hoc networking paradigm(or where there is a whale, there is a way). In *International Symposium on Mobile Ad Hoc Networking & Computing: Proceedings of the 4 th ACM international symposium on Mobile ad hoc networking & computing*, volume 1, pages 233–244, 2003.

[88] Jing Su, James Scott, Pan Hui, Jon Crowcroft, Eyal De Lara, Christophe Diot, Ashvin Goel, Meng Lim, and Eben Upton. Haggle: Seamless networking for mobile applications. *UbiComp 2007: Ubiquitous Computing*, pages 391–408, 2007.

[89] K. Subrahmanyam, S.M. Reich, N. Waechter, and G. Espinoza. Online and offline social networks: Use of social networking sites by emerging adults. *Journal of Applied Developmental Psychology*, 29(6):420–433, 2008.

[90] L. Tahvildari. Local positioning techniques with emphasis on bluetooth. *E&CE750: Topic*, 2001.

[91] J. Tang, T. Wang, and J. Wang. Information Flow Detection and Tracking on Web2. 0 BLOGS Based on Social Networks. In *The 9th International Conference for Young Computer Scientists*, pages 1664–1670. IEEE, 2008.

[92] Jie Tang, Tiancheng Lou, and Jon Kleinberg. Inferring social ties across heterogenous networks. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 743–752, New York, NY, USA, 2012. ACM.

[93] K. Thapa and S. Case. An indoor positioning service for bluetooth ad hoc networks. In *Midwest Instruction and Computing Symposium, MICS*, 2003.

[94] M.C. Thomas-Hunt, T.Y. Ogden, and M.A. Neale. Who's really sharing? Effects of social and expert status on knowledge exchange within groups. *Management science*, 49(4):464–477, 2003.

[95] Amin Vahdat, David Becker, et al. Epidemic routing for partially connected ad hoc networks. Technical report, Technical Report CS-200006, Duke University, 2000.

[96] P. Vorst, J. Sommer, C. Hoene, P. Schneider, C. Weiss, T. Schairer, W. Rosenstiel, A. Zell, and G. Carle. Indoor positioning via three different rf technologies. *ITG-Fachbericht-RFID SysTech 2008*, 2008.

[97] R. Want, A. Hopper, V. FalcÃčo, and J. Gibbons. The active badge location system. *ACM Transactions on Information Systems (TOIS)*, 10(1):91–102, 1992.

[98] K. Wendlandt, M. Berhig, and P. Robertson. Indoor localization with probability density functions based on bluetooth. In *Personal, Indoor and Mobile Radio Communications, 2005. PIMRC 2005. IEEE 16th International Symposium on*, volume 3, pages 2040–2044, 2005.

[99] F. Wu, B.A. Huberman, L.A. Adamic, and J.R. Tyler. Information flow in social groups. *Physica A: Statistical and Theoretical Physics*, 337(1-2):327–335, 2004.

[100] B. Xie. Using the internet for offline relationship formation. *Social Science Computer Review*, 25(3):396, 2007.

[101] Zhiwen Yu, Xingshe Zhou, Daqing Zhang, Gregor Schiele, and Christian Becker. Understanding social relationship evolution by using real-world sensing data. *World Wide Web*, pages 1–14.

[102] D. Zhang, B. Guo, B. Li, and Z. Yu. Extracting social and community intelligence from digital footprints: an emerging research area. *Ubiquitous Intelligence and Computing*, pages 4–18, 2010.

[103] S. Zhou and J. K. Pollard. Position measurement using bluetooth. *Consumer Electronics, IEEE Transactions on*, 52(2):555–558, 2006.

Appendix A

Case Studies Resources

A.1 Electronic survey

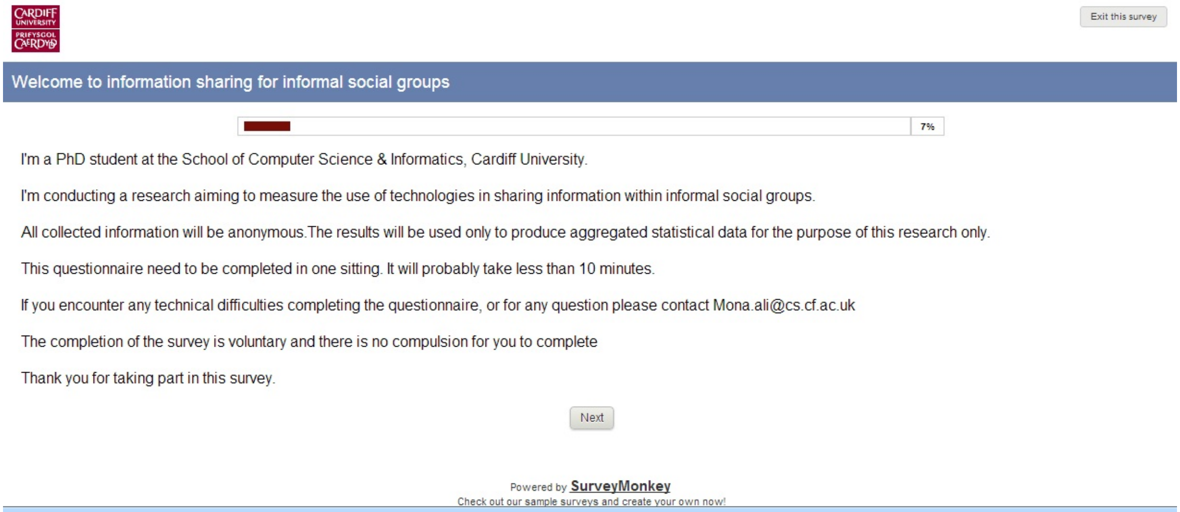
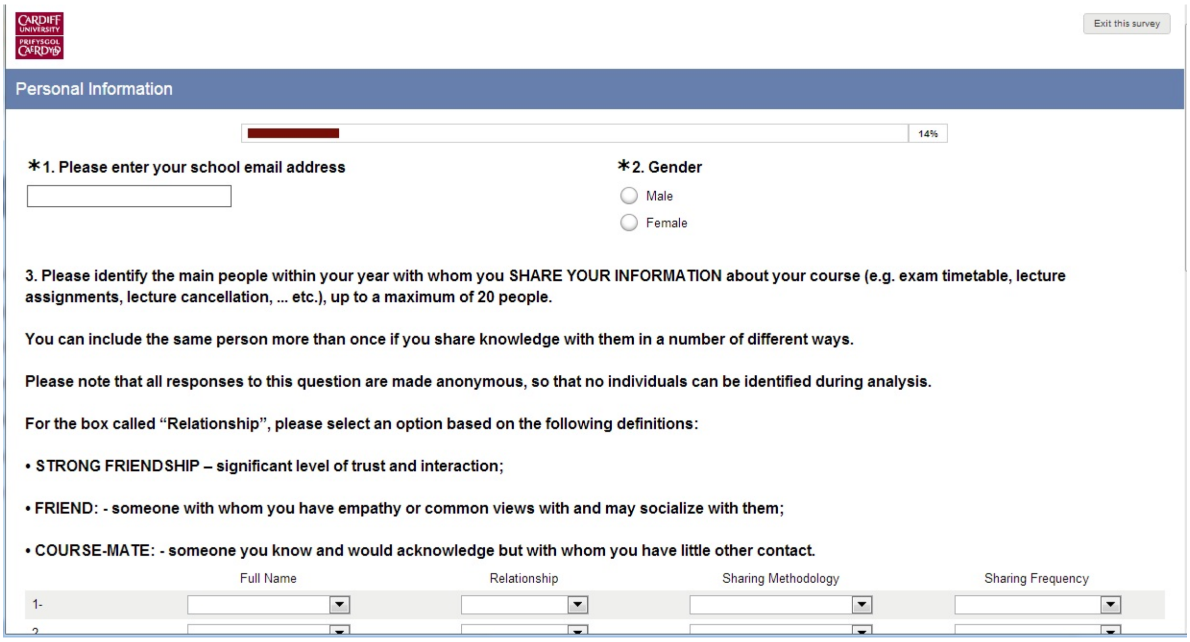


Figure A.1: Survey Welcome Page.



Exit this survey

Personal Information

14%

*1. Please enter your school email address

*2. Gender

Male

Female

3. Please identify the main people within your year with whom you SHARE YOUR INFORMATION about your course (e.g. exam timetable, lecture assignments, lecture cancellation, ... etc.), up to a maximum of 20 people.

You can include the same person more than once if you share knowledge with them in a number of different ways.

Please note that all responses to this question are made anonymous, so that no individuals can be identified during analysis.

For the box called "Relationship", please select an option based on the following definitions:

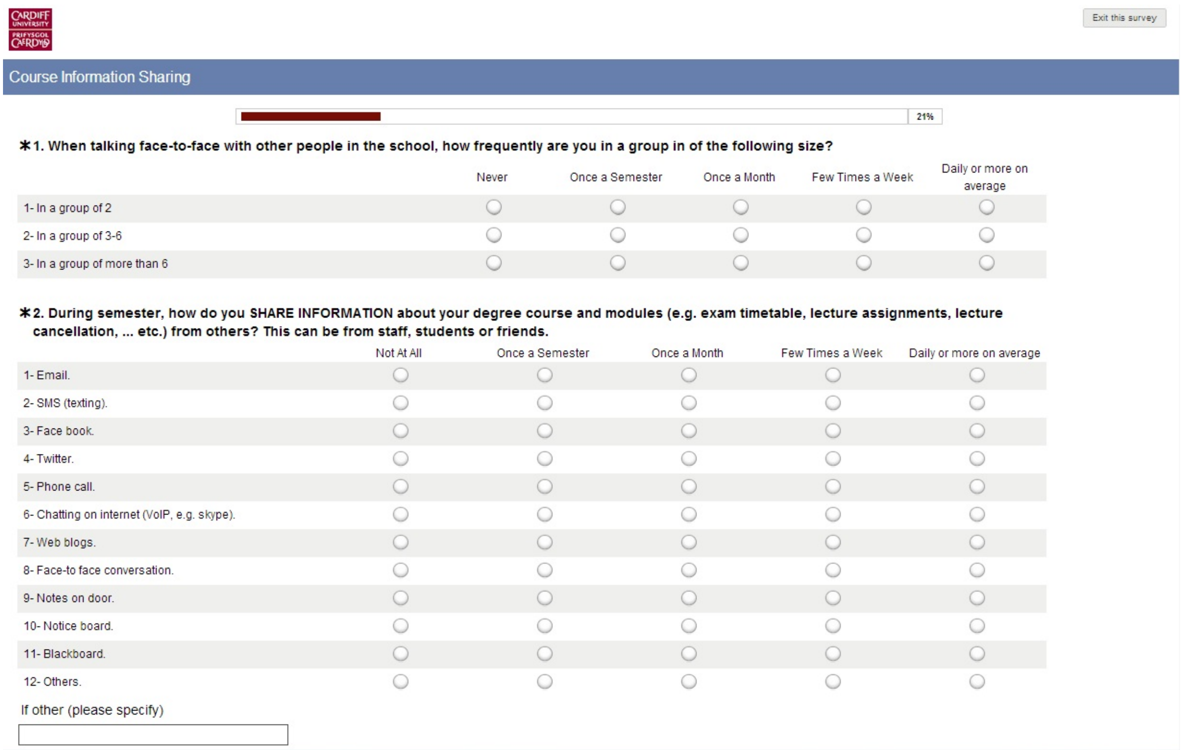
• STRONG FRIENDSHIP – significant level of trust and interaction;

• FRIEND: - someone with whom you have empathy or common views with and may socialize with them;

• COURSE-MATE: - someone you know and would acknowledge but with whom you have little other contact.

	Full Name	Relationship	Sharing Methodology	Sharing Frequency
1-				
2				

Figure A.2: Personal Information Page.



Exit this survey

Course Information Sharing

21%

*1. When talking face-to-face with other people in the school, how frequently are you in a group in of the following size?

	Never	Once a Semester	Once a Month	Few Times a Week	Daily or more on average
1- In a group of 2					
2- In a group of 3-6					
3- In a group of more than 6					

*2. During semester, how do you SHARE INFORMATION about your degree course and modules (e.g. exam timetable, lecture assignments, lecture cancellation, ... etc.) from others? This can be from staff, students or friends.

	Not At All	Once a Semester	Once a Month	Few Times a Week	Daily or more on average
1- Email.					
2- SMS (texting).					
3- Face book.					
4- Twitter.					
5- Phone call.					
6- Chatting on internet (VoIP, e.g. skype).					
7- Web blogs.					
8- Face-to face conversation.					
9- Notes on door.					
10- Notice board.					
11- Blackboard.					
12- Others.					

If other (please specify)

Figure A.3: Course Information Page.

Figure A.4: Bluetooth Registration Page.

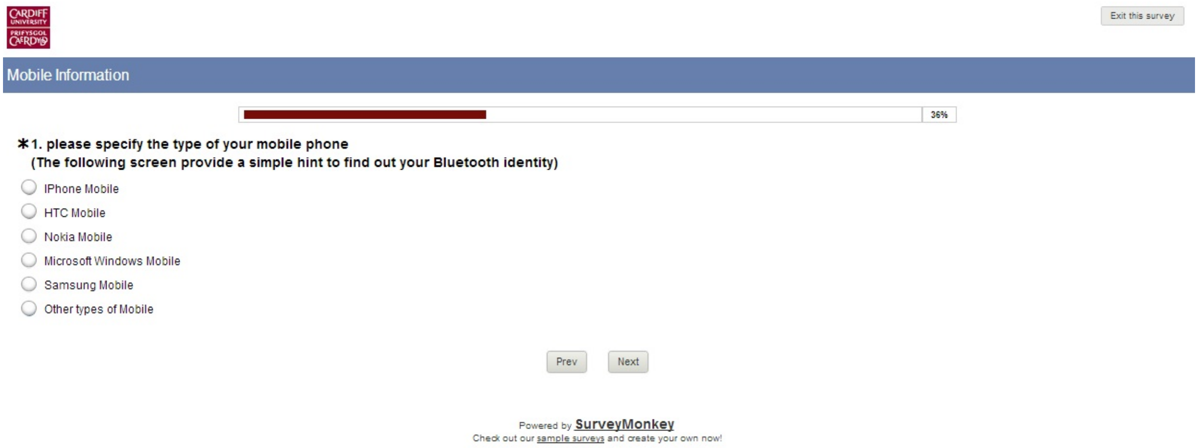


Figure A.5: Mobile Information Page.

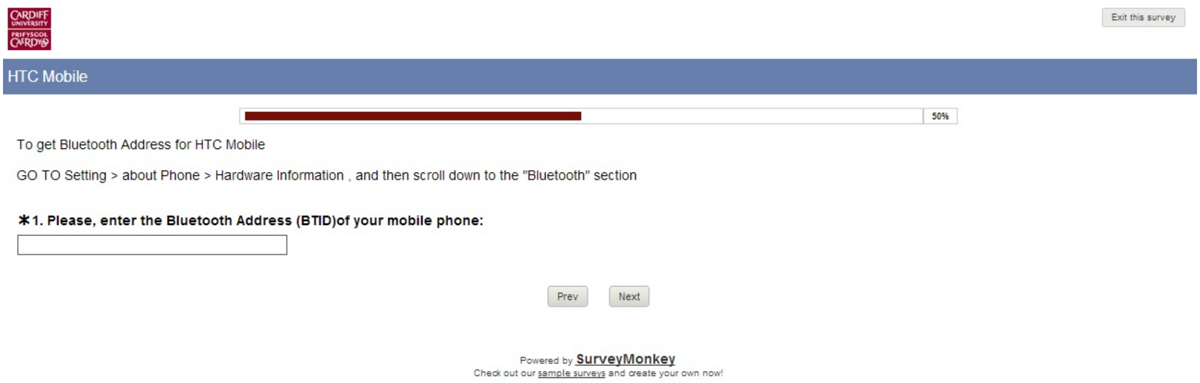


Figure A.6: HTC Page.

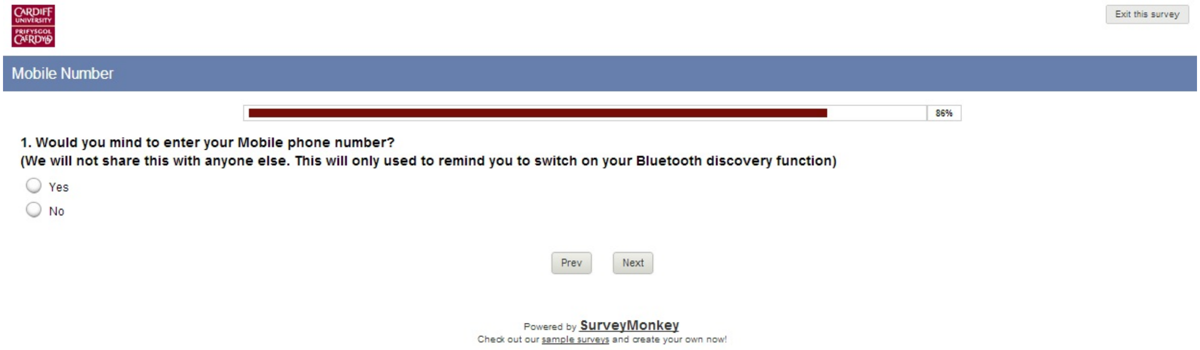


Figure A.7: Mobile Number Page.

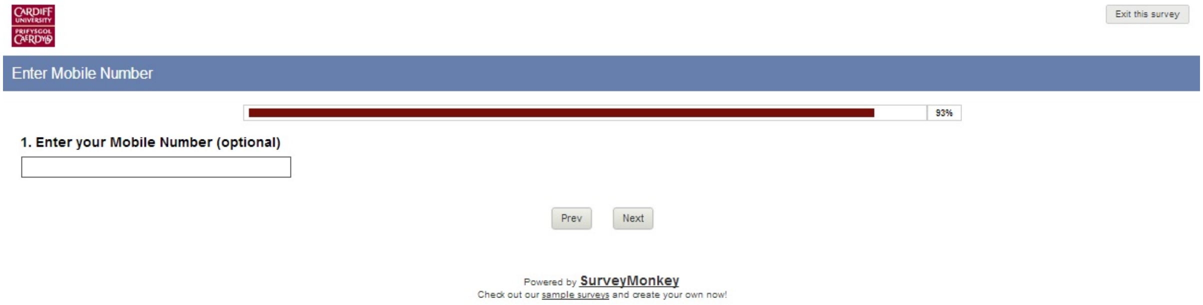


Figure A.8: Enter Mobile Number Page.

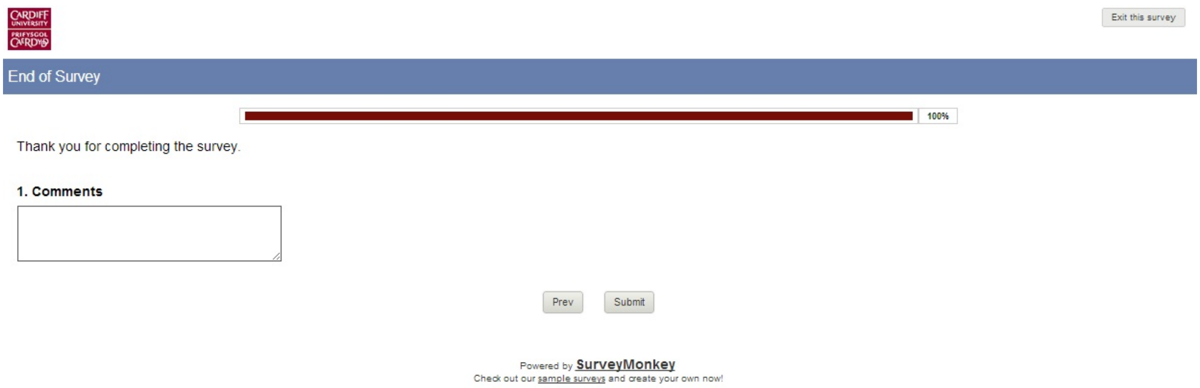


Figure A.9: End of Survey Page.

A.2 Survey Frequency Network Graph

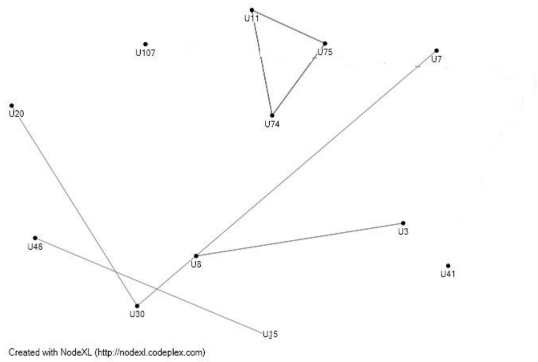


Figure A.10: Daily Interaction.

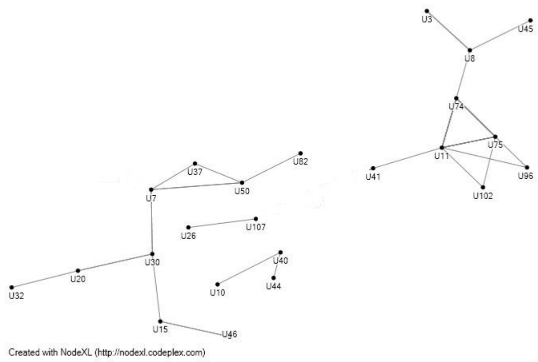


Figure A.11: Daily and Few-times-a-week Interaction.

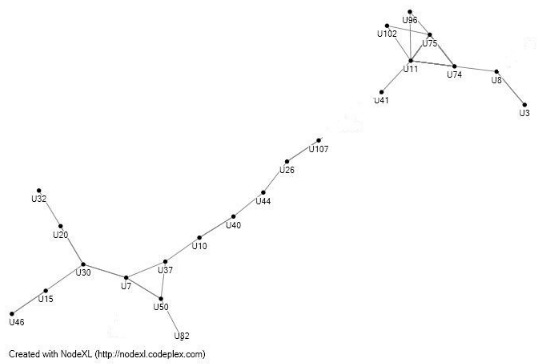


Figure A.12: Daily, Few-times-a-week and Monthly Interaction.

A.3 Terms and conditions for Experimental study

Bluetooth competition

Terms and conditions:

Please read carefully before you sign:

1- Carry the case where ever you are at the school building.

2- Each volunteer has his/her unique ID that will be appeared on the case cover. Please remember your ID number.

3- Do not swap the cases between participants.

4- Do not leave the case in the school.

5- Do not leave the case at home while you are at school (e.g. tie it to your bag!).

6- The case is your responsibility so please try to keep it safe. Do not leave the case with anyone else.

7- Do not forget to pick up your case and hand in on the slots mentioned below

8- The prize will be drawn after finishing the experiment by the end of this semester.

9- You will receive one entry in the draw for each slot of the experiment you complete.

10- Please do not open the case.

If you are accepted the terms and conditions please sign below:

Name (please write you name in capital letter):

Date:

Device Number:

Telephone number:

Pick up time		Hand in time	
Monday 23rd April from 8:30 to 10:00		Tuesday 24th April from 4:00 to 5:00	
Thursday 26th April from 8:30 to 10:00		Friday 27th April from 4:00 to 5:00	
Monday 30th April from 8:30 to 10:00		Tuesday 1st May from 4:00 to 5:00	
Thursday 3rd May from 8:30 to 10:00		Friday 4th May from 4:00 to 5:00	
Tuesday 8th May from 8:30 to 10:00		Tuesday 8th from 4:00 to 5:00	
Thursday 10th May from 8:30 to 10:00		Friday 11th May from 4:00 to 5:00	

Figure A.13: Terms and Conditions.

Appendix B

Experimental Social Networks

B.1 Adjacency Matrices for Experimental Social Networks

	U10	U102	U107	U11	U115	U15	U20	U26	U3	U30	U32	U37	U40	U41	U44	U46	U50	U7	U74	U8	U82	U9	U96	Total
U10	0	23	0	70	1	51	0	433	52	0	52	70	70	70	0	0	42	36	0	0	30	396	37	1433
U102	23	0	198	97	527	378	175	6	309	0	314	343	610	248	319	173	332	312	239	238	306	97	138	5382
U107	0	198	0	0	0	187	119	23	205	0	107	256	332	128	185	177	193	178	162	247	219	0	0	2916
U11	70	97	0	0	0	85	0	0	125	0	163	156	145	159	0	0	127	57	0	0	66	116	110	1476
U115	1	527	0	0	0	0	0	1	0	0	0	0	0	0	603	0	0	0	0	0	0	1	0	1133
U15	51	378	187	85	0	0	163	44	207	0	419	491	281	188	163	164	410	274	208	242	464	59	104	4582
U20	0	175	119	0	0	163	0	0	154	0	300	124	157	61	110	98	120	81	297	143	117	0	0	2219
U26	433	6	23	0	1	44	0	0	0	0	13	11	1	0	1	0	0	0	1	0	16	360	0	910
U3	52	309	205	125	0	207	154	0	0	0	308	375	467	274	190	146	305	261	219	307	308	152	80	4444
U30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	7
U32	52	314	107	163	0	419	300	13	308	0	0	512	311	306	92	96	364	227	232	125	332	180	163	4616
U37	70	343	256	156	0	491	124	11	375	0	512	0	427	340	143	183	460	338	136	242	550	176	146	5479
U40	70	610	332	145	0	281	157	1	467	0	311	427	0	351	194	149	338	290	213	265	342	156	123	5222
U41	70	248	128	159	0	188	61	0	274	0	306	340	351	0	87	70	276	235	112	113	273	170	114	3575
U44	0	319	185	0	603	163	110	1	190	0	92	143	194	87	0	151	161	170	175	230	137	0	0	3111
U46	0	173	177	0	0	164	98	0	146	0	96	183	149	70	151	0	126	139	120	205	149	0	0	2146
U50	42	332	193	127	0	410	120	0	305	0	364	460	338	276	161	126	0	258	168	189	452	100	123	4544
U7	36	312	178	57	0	274	81	0	261	7	227	338	290	235	170	139	258	0	147	211	312	67	86	3686
U74	0	239	162	0	0	208	297	1	219	0	232	136	213	112	175	120	168	147	0	254	170	0	0	2853
U8	0	238	247	0	0	242	143	0	307	0	125	242	265	113	230	205	189	211	254	0	206	0	0	3217
U82	30	306	219	66	0	464	117	16	308	0	332	550	342	273	137	149	452	312	170	206	0	111	85	4645
U9	396	97	0	116	1	59	0	360	152	0	180	176	156	170	0	0	100	67	0	0	111	0	91	2232
U96	37	138	0	110	0	104	0	0	80	0	163	146	123	114	0	0	123	86	0	0	85	91	0	1400
																								71228

adjMatrix.pdf Cell co-location

Figure B.1: Trajectory Social Matrix.

	U10	U102	U107	U11	U15	U115	U20	U26	U3	U30	U32	U37	U40	U41	U44	U46	U50	U7	U74	U75	U8	U82	U9	U96	Total
U10	0	1284	0	620	442	0	255	6665	457	0	458	620	2773	620	0	0	361	308	0	0	152	253	3492	308	19068
U102	1284	0	1716	1669	3648	4489	2285	1546	2894	0	3832	2965	6639	2136	13732	1498	2872	2710	2069	0	2154	2640	979	1269	65026
U107	0	1716	0	193	1715	0	1380	426	1760	0	916	2246	3992	1074	1627	1515	1654	1528	1406	0	2156	1914	0	193	27411
U11	620	1669	193	0	829	0	208	0	1450	0	1413	1374	1291	1403	65	0	1111	472	0	0	0	541	1022	1684	15345
U115	0	4489	0	0	0	0	0	0	186	0	0	0	0	0	5153	0	0	0	0	0	72	0	64	3	9967
U15	442	3648	1715	829	0	0	1649	382	1976	0	3984	4266	2443	1573	1827	1415	3544	2384	1787	0	2075	4040	474	1026	41479
U20	255	2285	1380	208	1649	0	0	223	1301	0	3716	1023	1869	504	2699	800	992	637	2576	0	1178	989	0	208	24492
U26	6665	1546	426	0	382	0	223	0	0	0	107	84	2826	0	0	0	0	0	0	0	0	128	3169	0	15556
U3	457	2894	1760	1450	1976	186	1301	0	0	0	2883	3296	4144	2394	1888	1259	2666	2249	1895	0	2644	2686	1349	857	40234
U30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	0	0	50
U32	458	3832	916	1413	3984	0	3716	107	2883	0	0	4499	2815	2666	2603	801	3112	1922	1955	0	1032	2878	1590	1437	44619
U37	620	2965	2246	1374	4266	0	1023	84	3296	0	4499	0	3777	2940	1232	1589	3965	2941	1136	0	2129	4782	1554	1284	47702
U40	2773	6639	3992	1291	2443	0	1869	2826	4144	0	2815	3777	0	3027	1770	1254	2954	2491	1834	0	2292	2991	1388	1078	53648
U41	620	2136	1074	1403	1573	0	504	0	2394	0	2666	2940	3027	0	743	560	2358	2018	971	0	962	2364	1511	963	30787
U44	0	13732	1627	65	1827	5153	2699	0	1888	0	2603	1232	1770	743	0	1329	1399	1485	1510	0	2057	1181	0	0	42300
U46	0	1498	1515	0	1415	0	800	0	1259	0	801	1589	1254	560	1329	0	1050	1187	989	0	1790	1297	0	0	18333
U50	361	2872	1654	1111	3544	0	992	0	2666	0	3112	3965	2954	2358	1399	1050	0	2192	1421	0	1578	3911	877	1067	39084
U7	308	2710	1528	472	2384	0	637	0	2249	50	1922	2941	2491	2018	1485	1187	2192	0	1260	0	1819	2726	583	741	31703
U74	0	2069	1406	0	1787	0	2576	0	1895	0	1955	1136	1834	971	1510	989	1421	1260	0	0	2170	1472	0	0	24451
U75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
U8	152	2154	2156	0	2075	72	1178	0	2644	0	1032	2129	2292	962	2057	1790	1578	1819	2170	0	0	1792	0	0	28052
U82	253	2640	1914	541	4040	0	989	128	2686	0	2878	4782	2991	2364	1181	1297	3911	2726	1472	0	1792	0	952	720	40257
U9	3492	979	0	1022	474	64	0	3169	1349	0	1590	1554	1388	1511	0	0	877	583	0	0	0	952	0	788	19792
U96	308	1269	193	1684	1026	3	208	0	857	0	1437	1284	1078	963	0	0	1067	741	0	0	0	720	788	0	13626
																									692982

Duration Social Network

adj Matrix.pdf

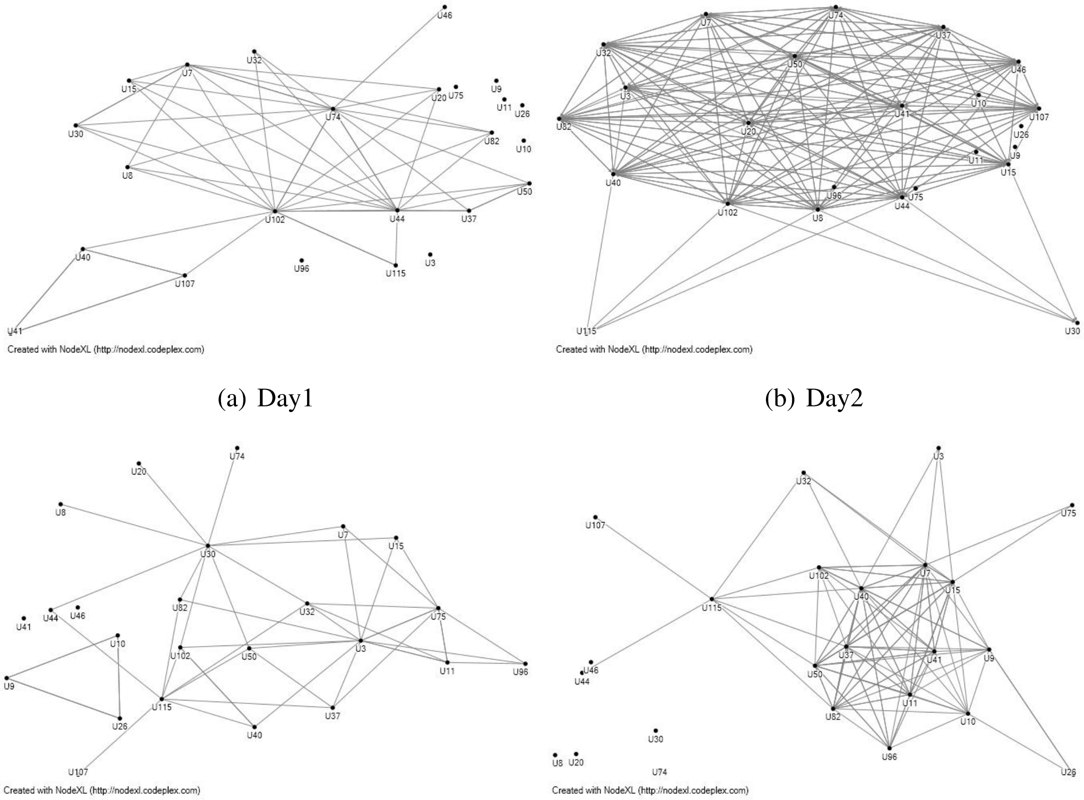
	U10	U102	U107	U11	U15	U115	U20	U26	U3	U30	U32	U37	U40	U41	U44	U46	U50	U7	U74	U75	U8	U82	U9	U96	Total
U10	0	33	0	0	0	0	1	243	0	0	0	0	53	0	0	0	0	0	0	0	3	0	93	0	426
U102	40	0	1359	133	139	1308	904	83	131	40	450	421	834	219	5355	39	885	299	644	0	339	305	4	33	13964
U107	0	38	0	78	0	0	38	0	27	0	25	181	515	3	31	3	1	3	15	0	4	169	0	36	1167
U11	25	0	4	0	87	0	2	0	142	0	215	35	12	0	0	0	21	23	8	546	0	17	17	287	1441
U115	0	1245	28	178	2	0	0	0	133		341	20	72	0	1378	0	132	0	0	0	0	4	0	29	3562
U15	61	74	12	91		0	353	10	54	203	343	192	57	136	15	156	159	351	320	10	1128	347	94	54	4220
U20	12	53	23	31	84	0	0	28	50	0	169	30	121	21	24	26	41	13	253	0	24	30	0	2	1035
U26	1498	43	82	0	313	0	4	0	0	0	0	0	381	0	0	0	0	0	0	0	0	108	486	0	2915
U3	0	355	24	1078	13	6	27	0	0		763	4	66	0	4	0	3	8	5	3	4	7	0	831	3201
U30	0	12	0	0	36	0	40	0	0	0	2	0	0	0	4	0	4	54	14	0	46	0	0	0	212
U32	6	56	31	149	66	0	156	0	188	0	0	57	63	7	29	3	10	4	32	0	6	39	6	77	985
U37	166	61	155	51	150	0	38	3	70	0	71	0	157	80	86	18	228	164	32	0	35	423	152	59	2199
U40	22	272	245	31	48	9	65	0	44		61	89	0	121	52	17	89	69	30	0	30	35	20	24	1373
U41	6	26	48	7	11		13		30		4	60	66	0	38	7	49	59	24		16	12	4	4	484
U44	0	2753	164	148	48	198	86	0	48	32	310	289	51	177	0	16	323	164	115	0	113	169	0	134	5338
U46	0	12	8	0	9	0	8	0	10	0	8	16	10	5	14	0	13	18	10	0	19	9	0	0	169
U50	83	43	51	22	117	0	22	0	31	0	0	172	65	66	47	21	0	88	37	0	12	14	97	19	1007
U7	20	87	73	31	106	0	61	0	67	49	35	132	110	101	85	34	73	0	49	4	78	44	21	26	1286
U74	0	56	59	22	65	0	301	0	45	14	78	60	42	50	55	21	61	63	0	0	44	27	0	28	1091
U75	0	0	59	563	16	0	0	0	8	0	120	4	0	0	0	0	0	36	0	0	0	0	0	208	1014
U8	9	18	4	0	125	1	42	0	4	17	45	7	1	2	13	19	10	28	45	0	0	18	0	0	408
U82	20	8	74	8	69	0	23	0	22	0	99	700	43	36	8	16	20	20	8	0	24	0	10	4	1212
U9	144	0	0	0	0	0	0	121	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	265
U96	1	36	0	134	0	12	0	0	70	0	60	1	2	0	32	0	1	0	0	0	0	0	0	0	349
																									49323

device co-located

adj Matrix.pdf

Figure B.3: Co-located Social matrix.

B.2 Day-by-Day Network Structure for Co-located Opportunistic Network

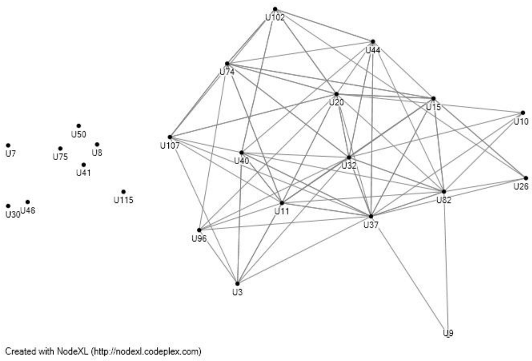


(a) Day1 (b) Day2

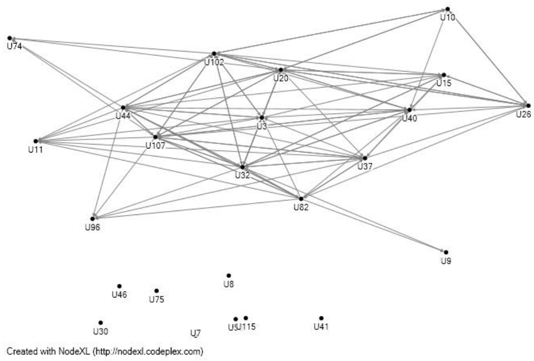
(c) Day3 (d) Day4

Figure B.4: Day-by-Day Co-located Opportunistic Network Week 1.

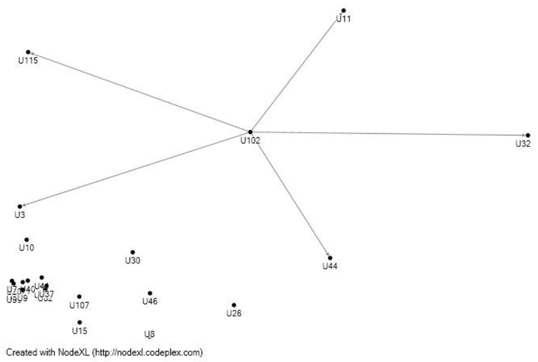
B.3 Day-by-Day Network Structure for Mobility Opportunistic Network



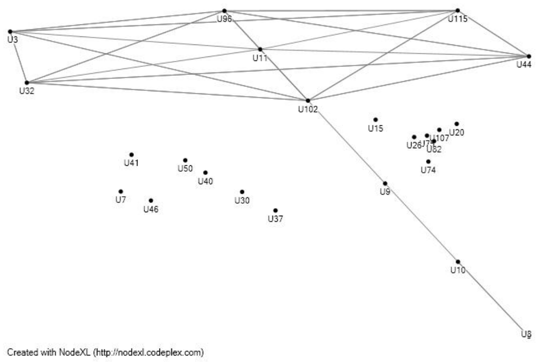
(a) Day5



(b) Day6

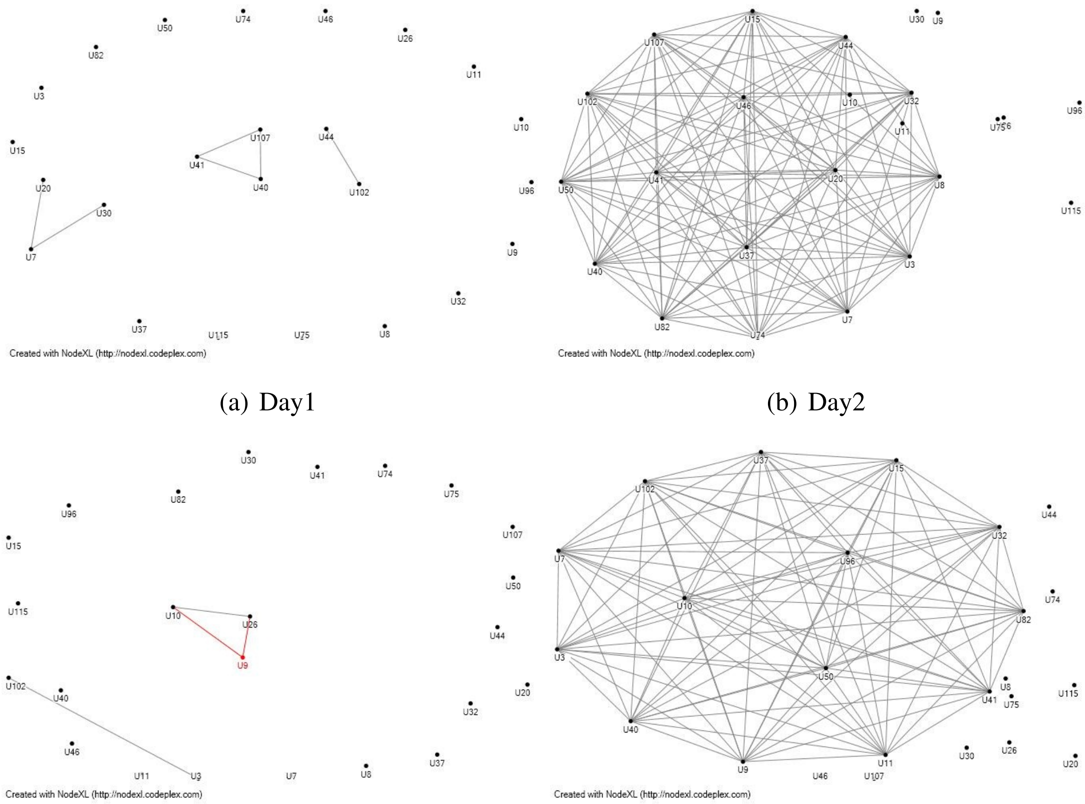


(c) Day7



(d) Day8

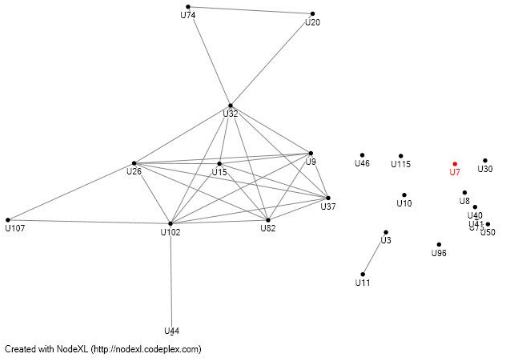
Figure B.5: Day-by-Day Co-located Opportunistic Network Week 2.



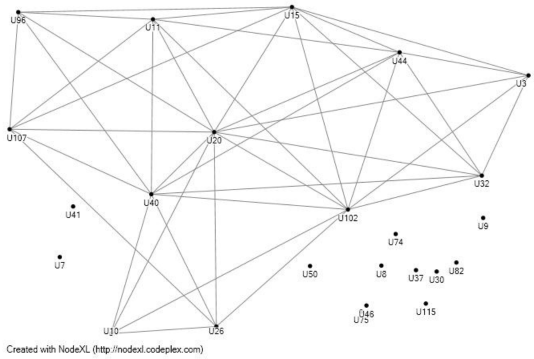
(a) Day1 (b) Day2

(c) Day3 (d) Day4

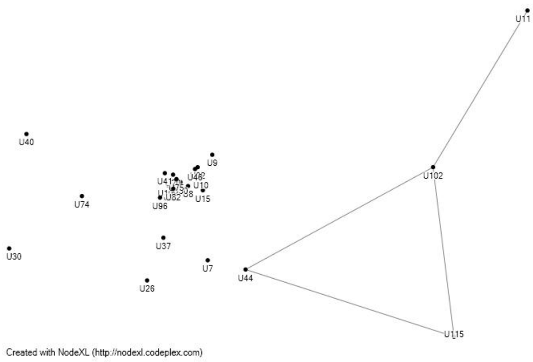
Figure B.6: Day-by-Day Mobility Opportunistic Network Week 1.



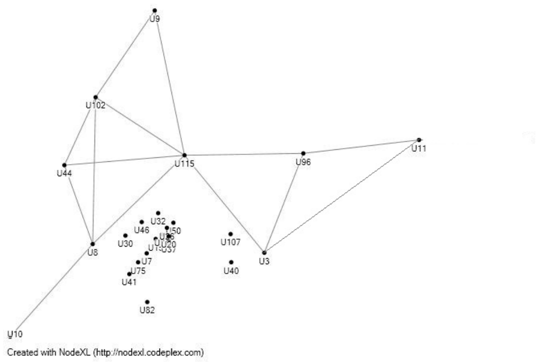
(a) Day5



(b) Day6



(c) Day7



(d) Day8

Figure B.7: Day-by-Day Mobility Opportunistic Network Week 2.

Appendix C

Journal Paper in Preparation

The role of online communication in offline social networks

Mona A. S. Ali, Roger M. Whitaker, Matthew J. W. Morgan

Abstract

Social networks are frequently categorized as online (supported electronically without necessitating participants meeting) or offline (defined though physically defined relationships). Often these social networks are treated as mutually exclusive but in many situations they coexist across the same community. In these circumstances online social networks operate as a communication mechanism to enhance information flow and support different types of relationship. In this paper we undertake an observational study of a social network of 98 undergraduate students to determine the role of on-line communication in offline social networks. We exploit social network analysis to examine the structure of the underlying communication. We examine the role of electronic social networks (web-based, email, telephone and SMS) to establish how different methods of communication and communication frequency support different types of relationship. Interesting trends emerge on how different technologies are used. The results reaffirm the importance of on-line social networks in facilitating structurally important weak links and reinforcing strong links.

1 Introduction

Social networks have become increasingly important for understanding and explaining complex behaviour within groups and communities. The importance of social and technological networks is now widely acknowledged [Kle08]. The Internet and widespread availability of electronic communication have added much greater opportunity for social relationships [LS05] to be established and maintained. As such the electronic social network has emerged as a popular and effective mechanism for communication. Popular approaches to such communication now include Facebook, Twitter, SMS as well as email and mobile telephone. It is often the case that online and offline social networks are studied independently. However in many cases an online social communication coexists with an offline social network across the same community. In these circumstances it is possible to study the effect that online communication has on the offline social network.

Electronic communication makes it much easier to connect people who may not have a strong offline relationship. This makes online communication technology a powerful phenomenon with possibly significant changes to information flow in a social network. We undertake an observational study of a social network of 98 undergraduate students over which online and offline social networks coexist. We seek to determine the role of online communication in the offline social network, in particular the role that online communication plays in augmenting face-to-face interactions. Through a participant questionnaire we are able to determine the communication structures and the patterns of communication supporting the underlying relationships. These are examined using extensive social network analysis.

Figure C.1: Journal Paper.

2 Social network analysis and information flow

Social networks provide a structure for information flow. Consequently much of the literature in this field is relevant and we highlight some of the most interesting ways in which social network analysis has been adopted and exploited.

Considering offline social networks based on participant (face to face) interaction, information flow has been of particular interest for business and organisational analysis. In many cases social networks are used and extended as an analysis tool that allows communication to be discovered and exploited. Of these contributions, Cross et al. [CPPB01] propose an approach that is based on creating a sociogram for an information flow network in an organisation. Also related to organisational needs, Mueller-Prothmann & Finke [MPF04] use social network analysis to develop a method for expert localisation and knowledge transfer. They adapt social network analysis to fit organisational practice where it provides a tool to identify knowledgeable communities and to analyse the structure of information flow within and between organizations. Their analysis uses a range of graph-based metrics to assess the social network structure. Dahel & Pedersen [DP04] use a questionnaire to examine the role of informal contacts in a specific cluster or sub-network. The authors analyse the knowledge flow and determine whether the employees actually acquire valuable knowledge through informal information networks.

Helms & Buijsrogge [HB06] seek to extend social network analysis and develop a technique called *knowledge network analysis*. In doing so they add various concepts that are aimed at making social network analysis more suitable for knowledge networks. These concepts include knowledge management roles, expertise levels, knowledge velocity and knowledge viscosity. Helms et al. [HIBZ10] focus on evaluating the limitations of network analysis for knowledge sharing. The evaluation is carried out through a case study at an international product software developer event. The authors compare and contrast qualitative and quantitative studies and use this to extract limitations of network analysis in the context of their work. Fischbach et al. [FGL$^+$10] also study information flow between workers within an organisation but extend their analysis to a range of different tools and technologies. The study uses dynamic social network analysis and includes face-to-face interaction, email and instant message communication. These works show how social networks can be exposed and exploited. It is noted that identifying scenarios within which both online and offline social networks coexist is a challenge.

Beyond the use of social network analysis in an organisational context, social network analysis has been employed to study information flow through diverse technologies. In [MDR$^+$03], Martinez et al study collaborative support in computer aided learning. A mixed evaluation method is used that integrates quantitative statistics, qualitative data analysis and social network analysis. They combine different type of data source to support their approach where the data sources include computer log events, face-to-face interaction, questionnaires and focus groups. Email communication is the focus of Kossinets et al. [KKW08] who study the temporal dynamics of communication using emails between the staff within a University, defining a network *back-bone* that maps the quickest information flows in the network.

Tang et al. [TWW08] propose a new approach for information detection and tracking on blogs using both social features and text. They focus on discovering hidden communities within a social network and develop a weighted graph representation showing closeness of relations. They combine the these methods to extract the information flow and investigate both temporal and spatial dimensions. Lim [Lim09] looks at patterns of information flow in the chat exchanges of two virtual learning groups using social network analysis. Quantitative statistical network analysis is applied to textual data (tutorial transcripts) containing the exchanges of information within online collaborative learning context. Information flow is also tackled from a physical perspective in [WHAT04]. Statistical calculation is used to analysis the information flow of

Figure C.2: .

instant message between employees in an organization, taking into account the observation that a people interested in an item are likely to belong to the same social groups.

Currently relatively few studies examine both online and offline social communication structures over a single community. The *hybrid* online-offline social network has been defined as a social network in which social links are maintained using both online and offline methods of communication [GM05]. Some of the few studies tackling the hybrid network include [AG06] where the authors consider the embedding of social networks in different technologies. Further work from the psychology viewpoint [SRWE08] establishes how networks of "friends" of young adults relate to their online social networks. These studies examine the social network structure using social network analysis approaches including an ego-centric viewpoint. Haythornthwaite et al [Hay05] studies the impact of social media on existing relationships and the influence on strength of relationship between parties.

A focus of a number of studies concerning hybrid social networks concerns the issues of trust and identity. This affects the sharing of information and in some cases it is claimed that a mixture of virtual and physical social networks may overcome these difficulties. For purposes of knowledge sharing, [Mat10] [GM05] investigate communities where physical interactions are extended in to the virtual world. Subrahmanyam et al [SRWE08] attempt to answer the question as to whether having physical interactions combined with virtual interactions reduces problems concerning trust and online knowledge sharing. In a reverse study Xie [Xie07] found that conversely, as well as enabling the creation of online social relationships, the Internet can affect offline relationship formation.

2.1 Our contribution

In this paper we explore the role of communication technologies in a social network of students in higher education. This is a well-defined and self-contained social network with hybrid online and offline properties. We explore the role of the online communication on the social network and we are seeking to address a number of questions. These include: the extent to which connectivity is increased from online social networks; the extent of linkage between different strengths of relationship and different types of technology usage; the frequency of interaction and the relationship with technology; the susceptibility of the network to spread of information. The results reveal the dependency on and preference for electronic modes of interaction when alternatives are available. The results give a measure of the potential "communication gain" that participants have over communication a solely offline network. Our analysis also reveals the clustering characteristics that are due to online social networks and the role of key players within the hybrid network for global information spreading. From these questions we are able to significantly increase our understanding of the interplay between online communication and offline social networks and the augmentation effect of online communication technologies.

2.2 The current study

Our study has been conducted using a cohort of mainly 18-19 year old undergraduate students in the first year (freshman) of bachelor degree programmes in Computer Science and Information Systems. This is a well-defined community that exists as an offline social network. A feature of this age group is their disposition toward online interaction which makes it easy to observe the impact of online communication. An online questionnaire was distributed to 137 subjects with a response rate of 76% (104 students). Of these 6 responses were incomplete and were discounted from analysis, leaving a sample of 98 subjects, with a 15%-85% female-male gender balance. The questionnaire was designed to establish the characteristics of the social network from each

Figure C.3: .

participant's personal view point (i.e., an ego-centric perspective). This required participants to express their perceived communication activity with other subjects. This approach allows us to ask questions about a variety of online communication technologies but a disadvantage is the potential for mis-perception of a subject's own activity (in certain circumstances this has been measurable by considering the different perceptions of two parties engaged in a relationship). Anonymity has been preserved by recoding personal identifiers in the data prior to analysis.

For each participant the questionnaire was designed to identify the relationships maintained with others and we investigated how these relationships were maintained. Specifically considered was the intensity of the relationship (relationship strength in three categories), the frequency of interaction and and the mode of interaction (offline, online and by which communication mode). Relationship strength was categorised as a "strong friendship" (someone with whom you have significant level of trust and interaction),"friendship" (someone with whom you have empathy or common views with and may socialize with them) or "course- mate" (someone you know and would acknowledge but with whom you have little other contact). Frequency of interaction was categorised as the most frequent option from daily, at least a few times per week, at least every month, or at least once per semester. Communication was categorised as either face-to-face, mobile phone text messaging, telephone, email, micro-blogging, chatting on the Internet (VoIP), or Facebook, which was a-priori known to be the dominant social networking service used by this group. Thus each possible relationship could be maintained in $3 \times 4 \times 7 = 78$ different combinations of relationship intensity, frequency of interaction and mode of interaction. The maximum reach by any participant was 18 incidences.

Each incidence of mode of interaction, relationship strength and frequency of interaction represents an edge (a, b) in a directed graph from the participant a who has completed the questionnaire to b with whom she communicates. Similarly other graphs can be created by considering a subset of the experimental variables (mode of interaction, relationship strength and frequency of interaction). We use this approach to create a range of graphs that allow us to explore the relationship between the experimental variables. A range of social network analysis techniques are used to assess the underlying network structures from which we are able to draw a range of conclusions. We seek to determine:

- *communication technologies and network structure.* In particular we are able to determine the dependency on and role of different communication structures (Section 3.1);
- *relationship strength and network structure.* In particular we determine the role of weak links and the modes of communication relative to relationship strength (Section 3.2);
- *frequency of interaction and network structure.* In particular we able to determine the critical structures that are facilitating most of the communication traffic and the frequency with which different technologies are used. We also assess the frequency of interaction across different strengths of relationship (Section 3.3);
- *community detection and clustering.* In particular we assess the extent to which clustering occurs within the network and how this occurs (Section 3.4).

Across a diverse range of graphs we apply a wide range of social network analysis techniques. These techniques are applied selectively to address the above issues, and includes shortest path analysis, clustering, centrality, assortativity and key player analysis.

Figure C.4: .

3 Results

Firstly we consider the social network structure. In Section 3.1 we look at the social network's dependency on different types of communication (electronic or physical). We decompose the network based on whether relationships are supported by particular modes of communication, ordered by popularity. In doing so we are able to determine the marginal effects of additional online communication for particular types of technology. In Section 3.2 we consider the effect of relationship strength on network structure. In Section 3.2.1 characteristics of networks formed from different intensity of relationship is assessed. As a result the structural effect of weak links in the network is observable. How different intensity relationships are maintained by different modes of communication is explored in Section 3.2.2. In Section 3.3 we explore temporal issues concerning participant interaction. Different structures support different frequencies of interaction. In Section 3.3.1 we consider the way in which different technologies are used with different frequency. We also consider the relationship strength and frequency of interaction in section 3.3.2. In Section 3.4 we determine the how social network is affected in terms of clustering by the combination of the physical and social networks. Finally in Section 3.5 we determine the characteristics of the social network from the perspective of key player analysis with particular interest in the effect on key players in the virtual network.

3.1 Communication technologies and network structure

We consider the different modes of communication through a directed network $G_C = (V, E)$ where V is the set of participants and an edge $(i, j) \in E$ exists if and only if node i communicates with node j via at least one mode of communication, where $C \subseteq \{c_1, \ldots, c_5\}$, the set of possible modes of communication. While keeping the population V fixed, we vary the set C and examine the effect on shortest path lengths between all pairs of nodes. We label the communication modes in C in order of popularity, the greatest first, where popularity is measured by the number of participants who indicate they use a particular mode of communication. This is displayed in Table 1.

Table 1: Popularity of Different Modes of Communication

Identifier	Communication Mode	% Participants
c_1	Face to Face	98.2%
c_2	Facebook	42.6%
c_3	SMS	28.7%
c_4	email	25.2%
c_5	Phone call	14.7%

We examine the effect of additional types of communication, starting with the most popular and progressively observing the effect of additional (less popular) modes of communication. This results in a progressively more dense network structure. We examine the effect of additional connectivity by considering the profile of all shortest path lengths from within the network. Table 2 shows the mean path progressively decreasing.

From Figure 1 we can see that the mode shortest path length of three dominates mainly due to the additional inclusion of both Facebook and email. However from Table 1 we can see that while Facebook is well adopted (42.6%), email is a much less popular choice for communication (in terms of number of adopters) than all other forms of communication with the exception of the phone call. From Table 2 the addition of email communication to the network only

Figure C.5: .

marginally increases participant inclusion, with the percentage of participants having no path between them decreasing from 19.4% to 17.0%. However the small number of participants who do use email (25.2%) make key connections across the network because they provide an important effect - Table 2 shows the largest decrease in the mean shortest path length (from 3.37 to 3.22) as compared to when other technologies were added. This is due to a small number of email users who have a relatively high out-degree. In this regard email has a powerful structural effect.

A further observation concerns the degree distribution for each of the communication methodologies (Figure 3). The offline network's degree distribution is approximately Gaussian as compared to each online network's degree distribution which are closer to a power law distribution and indicates that online networks are approximately scale free. Finally we also note differences in the mixing characteristics for networks resulting from different technologies. In particular the face-to-face and SMS communication resulted in weak assortative mixing (Table 3) where as the remaining electronic modes of communication exhibit stronger dis-assortative mixing, which Facebook having the strongest characteristics. This is possibly explained by the broadcast functionality that exists in technologies such as Facebook and email that connects nodes having widely varying degrees. Table 4 shows that the aggregation of different networks smoothes the dis-assortative effects.

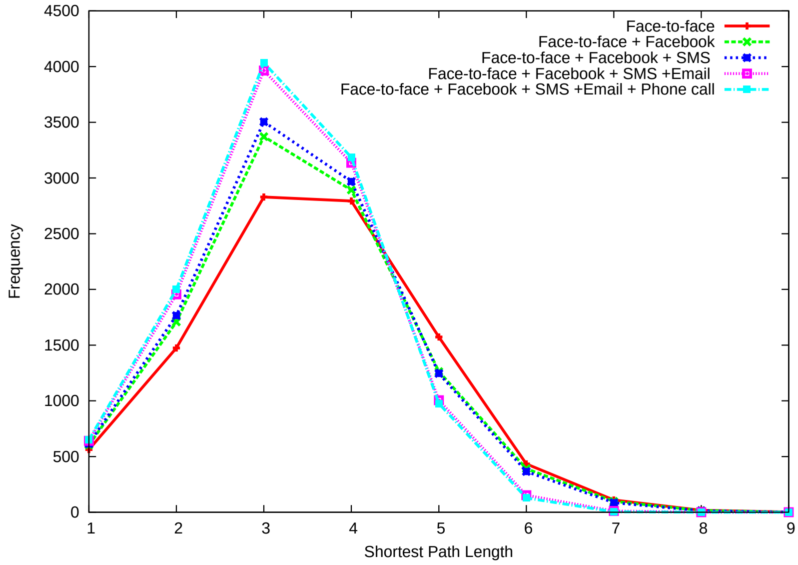


Figure 1: The distribution of shortest path lengths for different modes of communication.

Figure C.6: .

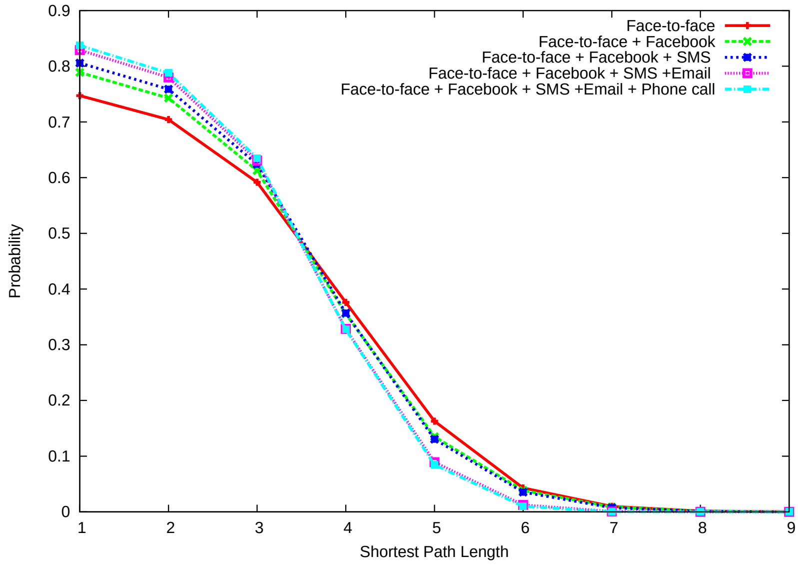


Figure 2: The Probability of shortest path lengths for different modes of communication.

Table 2: Statistics on Average Path Lengths in Graph G_C

Communication set C	Mean Shortest Path Length	Standard Deviation on Path Length	% with no path
$\{c_1\}$	3.53	1.287	25.3%
$\{c_1, c_2\}$	3.4	1.239	21.1%
$\{c_1, c_2, c_3\}$	3.37	1.20	19.4%
$\{c_1, c_2, c_3, c_4\}$	3.22	1.08	17.0%
$\{c_1, c_2, c_3, c_4, c_5\}$	3.2	1.06	16.0%

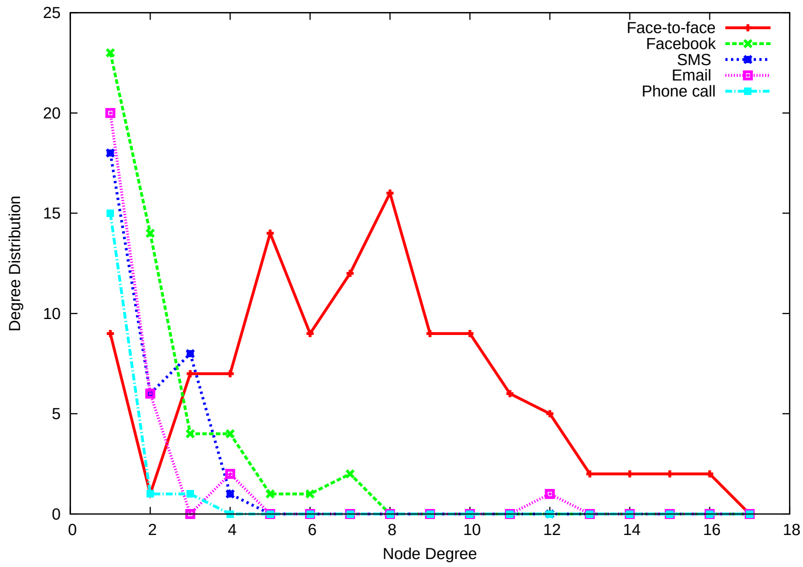


Figure 3: Degree Distribution for each communication methodologies

Figure C.7: .

7

Table 3: Assortative coefficient of Different Modes of Communication

Identifier	Communication Mode	Assortative coefficient
c_1	Face to Face	0.149199469
c_2	Facebook	-0.412720601
c_3	SMS	0.169995561
c_4	email	-0.419609665
c_5	Phone call	-0.296296296

Table 4: Assortative coefficient of combining different technologies G_C

Communication	Assortative coefficient
$\{c_1\}$	0.149199469
$\{c_1, c_2\}$	0.102050395
$\{c_1, c_2, c_3\}$	0.114130724
$\{c_1, c_2, c_3, c_4\}$	0.081515898
$\{c_1, c_2, c_3, c_4, c_5\}$	0.080460052

3.2 Relationship strength and network structure

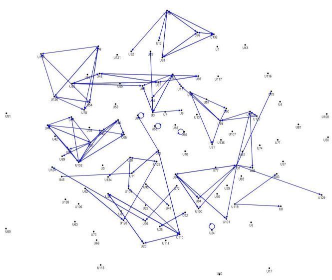
We examine different levels of relationship strength and assess their effect on the shortest path lengths. To do this we consider a directed network $G_R = (V, E)$, where set V is the set of participants and there exists an edge $(i, j) \in E$ if and only if node i perceives it has a relationship with node j with strength from set $R \subseteq \{R_1, R_2, R_3\}$. Here R_1 is the strongest relationship strength, denoted as "strong friendship" (someone with whom you have significant level of trust and interaction), R_2 is the medium level of relationship strength, denoted as "friendship" (someone with whom you have empathy or common views with and may socialize with them) and R_3 is the weakest relationship strength, denoted as "course- mate" (someone you know and would acknowledge but with whom you have little other contact) - see Table 5. We observe the effect of relationship strength by combining different relationship types together, starting with the strongest strength (which is least popular) and progressively observing the effect of adding weaker links. The effect of this additional connectivity is analysed using shortest paths.

3.2.1 Role of weak links in network connectivity

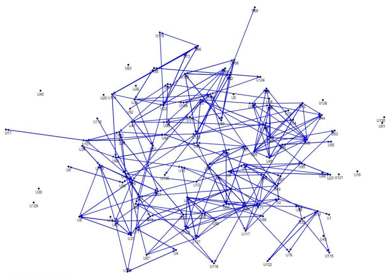
Figure 4 shows the different density and structure of relationship types. In isolation the strong friendships provide little overall connectivity (Figure 4a) with improvements for friendship (Figure 4b) and course-mate relationships (Figure 4c). In Figure 5 we also display the mean number of relationships held by each participant.

Table 5 shows statistics on the different types of relationship strength. For strong friendships, 61.7% of participants have at least one strong-friendship while 86.9% of participants have at least one friendship and 92.1% of participants have at least one course-mate. Interestingly there is also often a mismatch between reciprocation of friendship - for example if node i has a strong friendship with node j, the inverse relationship is not always a strong friendship. 43.5% of bidirectional relationship are mis-matched in this way. To count the different types of links we consider a graph of undirected links. An undirected link $\{i, j\}$ is defined if and only if either the links (i, j) or (j, i) exist. Consequently an undirected link indicates the existence of some relationships (in either direction or both ways) between two nodes. The number of undirected links for strong friendship and friendship nearly equals half of the number of directed links but

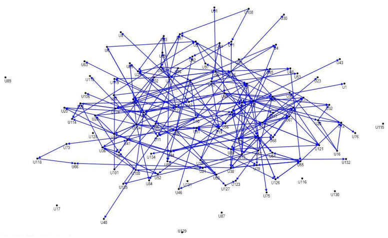
Figure C.8: .



(a) Strong-friendship



(b) Friendship



(c) Course-mate

Figure 4: Different relationships social network

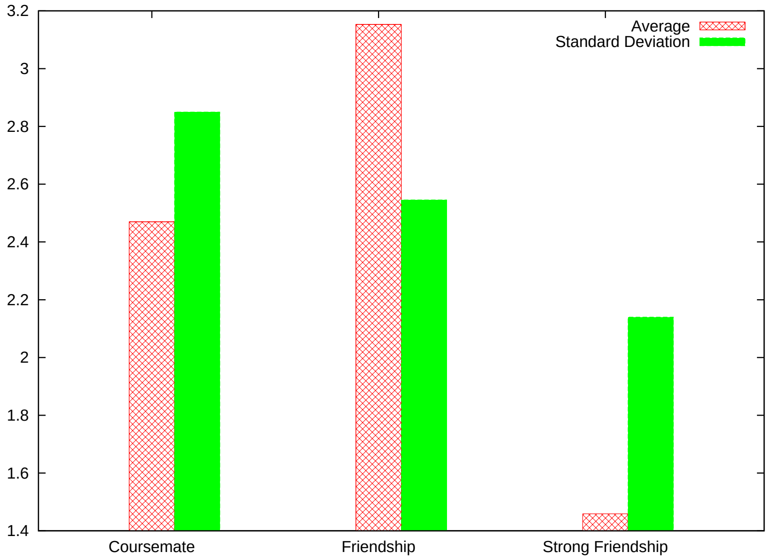


Figure 5: Different Strength of Relationship per person

Figure C.9: .

9

this is not the case for course-mate. The course-mates, as we can see in Table 5, the number of undirected links equals 205 however the number of directed links equals 240. This means that most course-mate links are uni-directional. This result reaffirms the significance of the course-mate relationship in adding more connectivity to the network rather than the other relationships (strong friendship and friendship).

Table 5: Popularity of Relationship Strength

Relationship Strength	% Included	# of Directed Links	# of Undirected Links
Strong Friendship (R_1)	61.7%	111	51
Friendship (R_2)	86.9%	273	166
Course-mate (R_3)	92.1%	240	205

The role of the weak relationships (i.e., course-mate) is notable. In Table 6 we look at the effect of combining different types of relationships, starting with the strongest relationships and progressively adding the weaker ones. The weakest relationships (i.e., course-mate) play a significant role in reducing the mean path length and improving connectivity. This is also shown in Figure 6 where the distribution of shortest path lengths between all pairs on nodes (within connected components) is given. The addition of the course-mate relationships to friendships and strong friendships leads to a reduction in mean shortest path length from 5.55 to 3.2. The path length reported for strong friendship (1.633) is low due to the graph having a small connected component.

Table 6: Statistics on Average Path Lengths within connected components

Communication set R	Mean Shortest Path Length	Standard Deviation on Path Length	% with no path
$\{R_1\}$	1.633	0.85	98.1%
$\{R_1, R_2\}$	5.55	2.79	70%
$\{R_1, R_2, R_3\}$	3.2	1.06	16%

3.2.2 Modes of communication and relationship strength

In Figure 7 we display the type of communication methodology that is used in sustaining each class of relationship. There is relatively little difference between the types of communication used with face-to-face interactions dominating. The only other notable observation is that email is used more in sustaining weaker links as compared to more use of SMS for friendships and strong friendships. This is consistent with email being a potentially more formal method of communication.

3.3 Frequency of interaction in the social network

Different interaction frequencies between the participants result in different communication topologies. These can be characterized from a directed network using four types of frequencies to select edges, namely: daily, few times a week, once a month and once a semester. Figure 10 displays the resulting network structures and Figure 11 displays the corresponding degree distribution for each interaction frequency. For daily interaction, the participants interact via a connection of clusters - see Figure 10a. For few-times-a-week interactions, the communication

Figure C.10: .

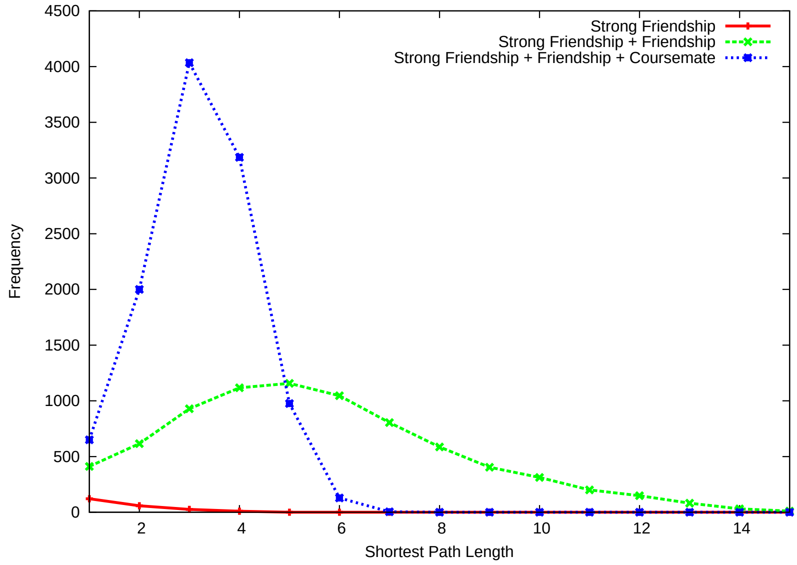
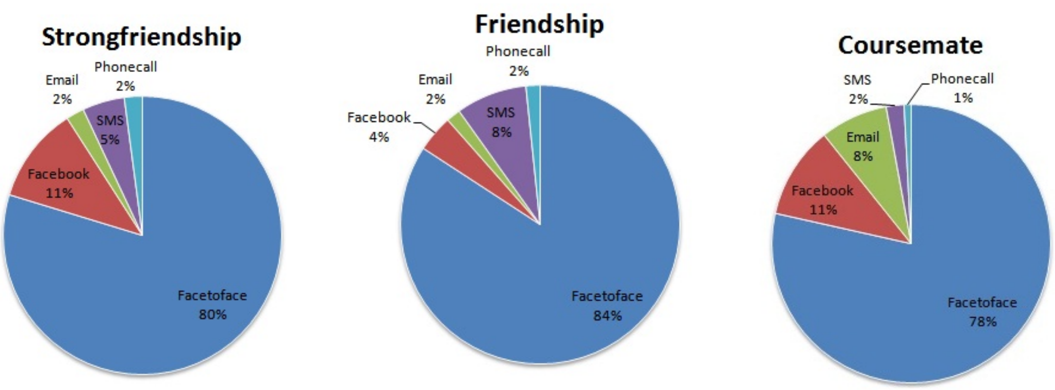


Figure 6: Shortest Path Length of Different Levels of Relationship within connected components

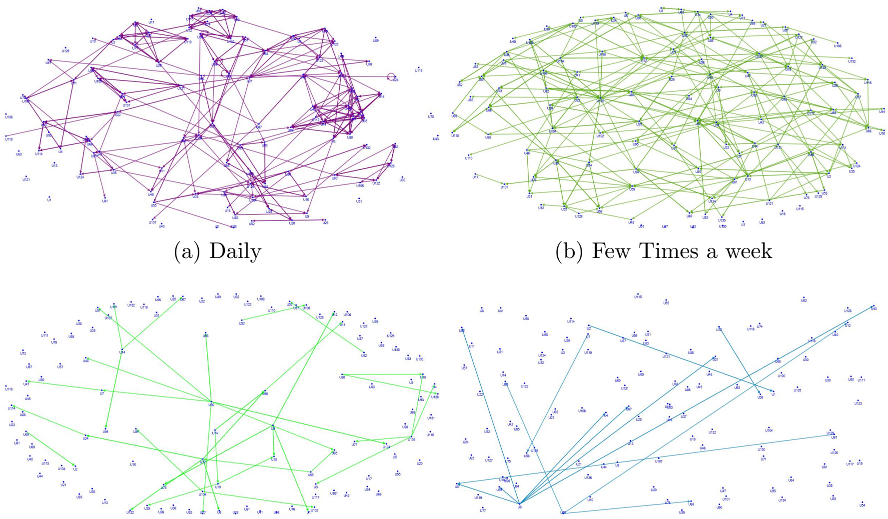


(a) (b) (c)

Figure 7: Relationship and Methodology usage

Figure C.11: .

between participants is more regularly distributed -see figure 10b. For the other two frequencies that represent weak-links in this context, the networks are disconnected and sparse.



(a) Daily (b) Few Times a week

(c) Once a month (d) Once a semester

Figure 8: Figures of Different communication Frequencies

Figure C.12: .

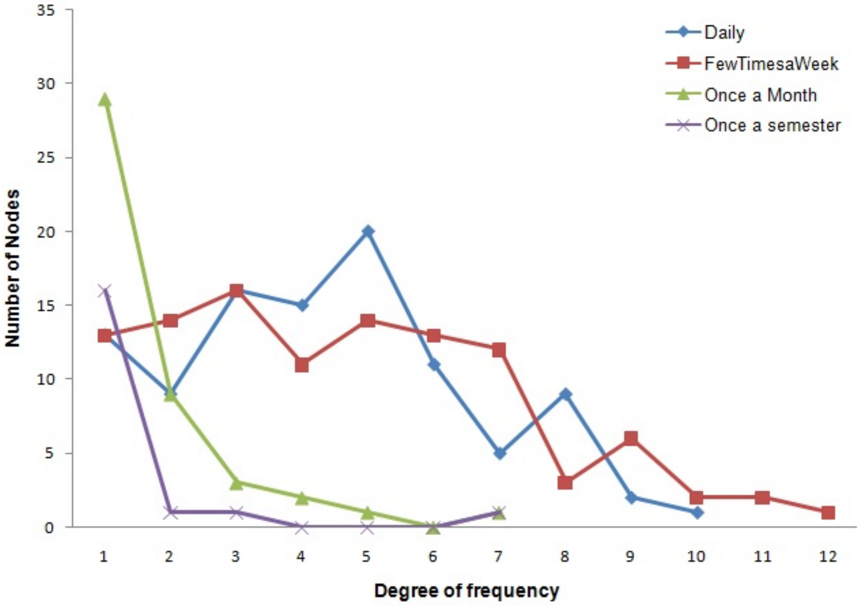


Figure 9: Degree of frequencies

3.3.1 Frequency of interaction and mode of communication

In Figure 12 we address the types of communication and the frequency with which they are used. A number of findings are notable. Firstly highly frequent communication (e.g., daily) are sustained primarily by face to face interactions. As the frequency of communication for a relationship is reduced, face to face interactions reduce and email interactions are increasingly used, especially in the case highly infrequent interactions (i.e., once a semester) which are primarily sustained by email interactions. At the same time, it seems that when excluding highly infrequent interactions, Facebook becomes a substitute for face to face communication.

3.3.2 Frequency of interaction and relationship strength

In Figure 13 we can see a correlation between friendship strength and frequency of interaction. Strong relationships are sustained by frequent interactions with the strong friendships and friendships seeing daily interactions. In contrast the weaker relationship (i.e., course-mate) are dominated by less frequent interaction albeit still quite frequent (i.e., few times a week).

3.4 Community sub-structures

An interesting feature of the hybrid online and offline social network is the extent to which different technologies support dense sub-structures and the role played by individuals in the network. To explore this we have analysed the clustering characteristics induced by different technologies. We apply the Grivan-Newman clustering method [New06b]. Using betweenness centrality Grivan and Newman focus on constructing a measure to indicate the edges which are least central to the cluster and they remove them. This divisive technique is repeatedly applied as described in [New06b]. To assess the strength of different clustering levels we use modularity as an external measure [New06b, New06a]. Modularity compares the number of

Figure C.13: .

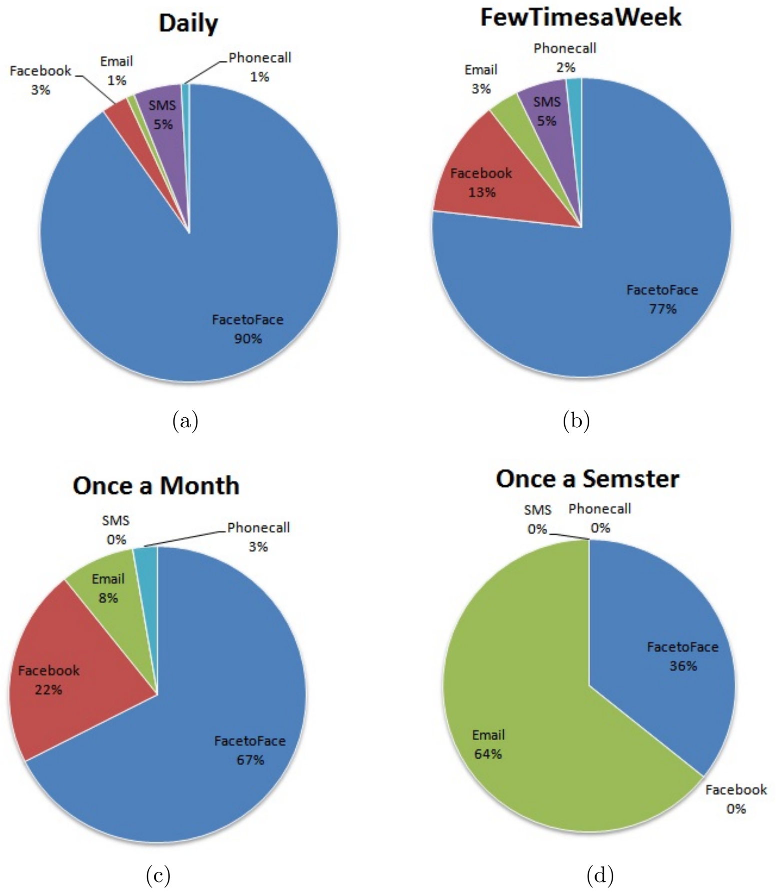


Figure 10: Frequency of usage for different communication technologies

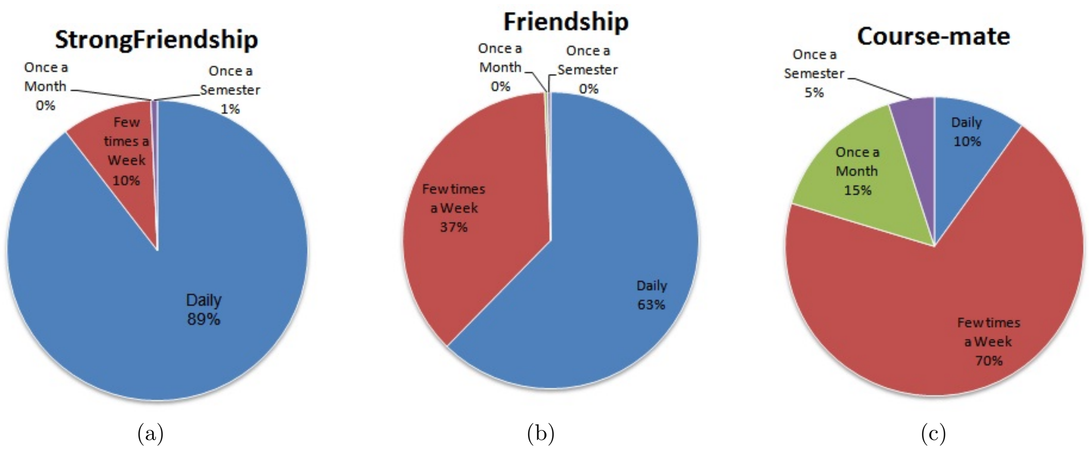
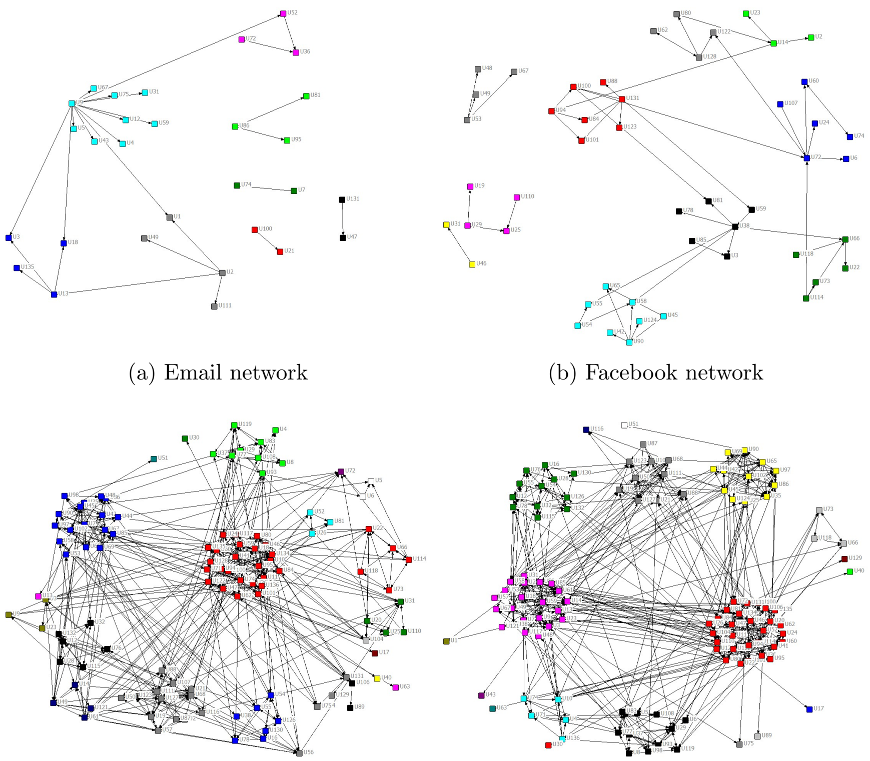


Figure 11: The frequency of interaction and relationship strength

Figure C.14: .

edges inside a cluster with the expected number of edges that one would find in the cluster if the network were a random network with the same number of nodes and where each node keeps its degree but edges are randomly connected. It does not provide a guide to how many clusters a network should ideally be split into but is a useful measure on the quality of a division of a network into clusters, with higher modularity measures indicating increased density within the clustering. Figure 15 shows the quality of clustering (in terms of modularity) as a function of number of clusters on applying the Grivan and Newman approach. Note that stronger clustering can be seen in the Facebook network, followed by the email network and the aggregated online and offline network. The email network most strongly partitions into a small number of clusters (X clusters), followed by the facebook network (X clusters), the aggregated aggregated online/offline network (X clusters) and finally the face-to-face network (X clusters). Interestingly the strongest clustering effect, as measured by modularity, is seen through Facebook closely followed by email. The aggregated online/offline network and face to face network exhibit similar levels of clustering strength that are significantly lower. It is likely that the strong clustering occurs for email and Facebook because of the "opt-in" nature of these technologies as compared to casual face-to-face interactions. Figures 14a, 14b, 14c and 14d show the clustered sub-networks that occur at maximum modularity.



(a) Email network

(b) Facebook network

(c) Face-to-face network

(d) Online & offline network

Figure 12: Clustered sub-networks at the highest modularity value

Figure C.15: .

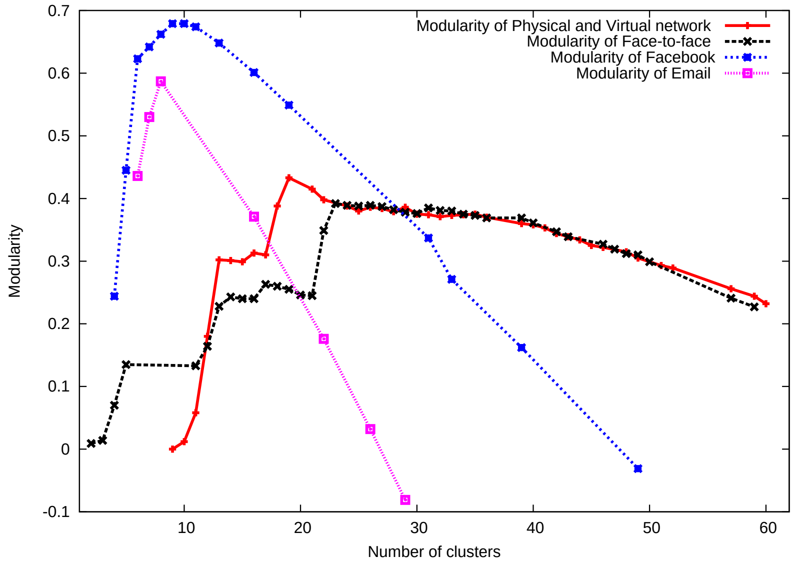


Figure 13: Modularity values through iteration of different methodologies

3.5 Key players

Key players represent a minimal subset of nodes from which all others can be reached within a particular maximum path length. Introduced in [Bor06], we employ key player analysis to determine the trade-off between a minimum number of informed participants and extent of possible dissemination across the population within a given path length. This allows us to explore the susceptibility of a network to possible information spreading effects. We achieve this by searching for generalised dominating sets that maximise the proportion of the participants that are reachable from at least one key player within a given path length. The heuristic technique that we adopt for this search is presented in Figure XXXXX and is run for a large number of iterations. This achieves local optima that form an upper bound on the global solution. Figure 16 shows the results for the aggregated

shows the different number of key players and the corresponding percentage of the population that is covered. Note that a only short maximum path length is needed to rapidly achieve maximum coverage. Moreover,

node that reached by the key players. As we can see 100% of reachaility achieved at 21 nodes in the key players set. That is equal to the number of clusters which achieve the optimal modularity value for the data set.

4 Discussion and Conclusions

Notes on key findings: need to write this from the perspective of the effect of electronic communication on the physical offline network.

structure: The low overall shortest path length. the important role of email in network structure. Evidence in different ways (including assortativity, difference in degree distributions). Different networks are providing different functions. Augmentation of the physical network by electronic means – Table 2.

relationship strength: course-mates are unidirectional but add significantly to the overall

Figure C.16: .

Algorithm 1 Keyplayers greedy optimization

Require: $Graph$ of the social network (adjacency matrix)

```
Select k nodes at random to populate set S
Set F = fit using appropriate key player metric
for each node u in S and each node v not in S do
    DELTAF = improvement in fit if u and v were swapped
    Select pair with largest DELTAF;
    if DELTAF <= then
        terminate
    else
        swap pair with greatest improvement in fit
        Set F = F + DELTAF
    end if
end for
```

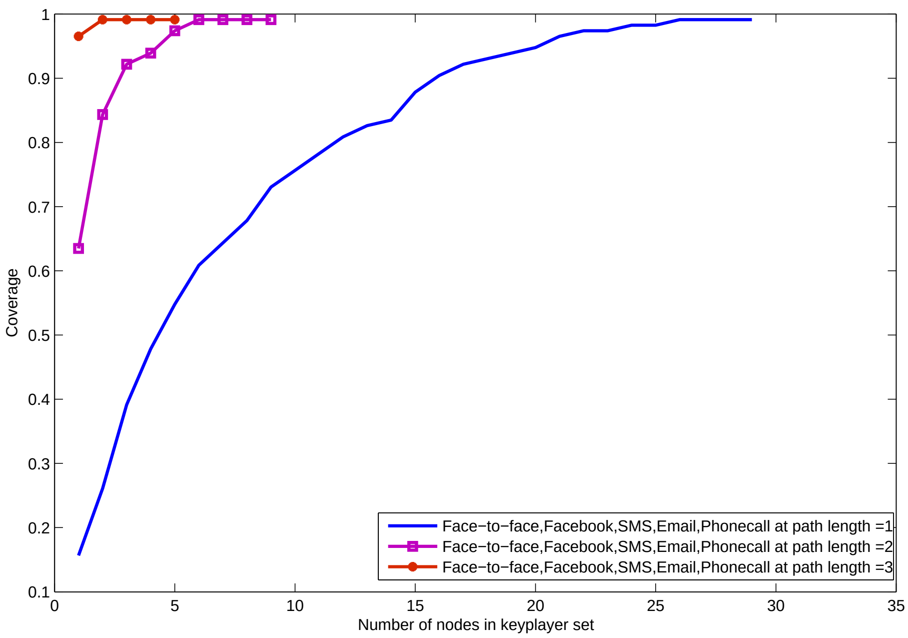


Figure 14: Key players and coverage for offline and online communication technologies

Figure C.17: .

17