Estimating Social Networks using Communications Metadata Gathered from Mobile Devices

Jamie A. Banford

A thesis submitted for the degree of Doctor of Philosophy
to the
Centre for Intelligent and Dynamic Communications,
Department of Electronic and Electrical Engineering,
University of Strathclyde.

Submitted for examination, September 2011.

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed:
Date:

'…because he was one of the people, tied to them by a thousand invisible strings…'

—William Golding,
The Inheritors

Abstract

Mobile communication devices are now truly ubiquitous; they are present everywhere in the modern world. They are also the first human artefacts capable of automatically detecting the subtle ways in which people reveal the nature of the relationships between them. This information is contained within the communications metadata available on these devices. By analysing these communications metadata certain tie signs become discernible and it becomes possible to estimate the current state of the social relationships of the user of the device. However, although this information is available on mobile communication devices few established techniques for gathering, and interpreting it have been defined.

This thesis presents empirical investigations into detecting and categorising social relationships using mobile devices. It introduces mechanisms to detect the social ties between the users of mobile devices, based on the interactions between them, and explores techniques to accurately categorise these ties. The ability to detect social ties allows the construction of a social graph with out any prior knowledge of the social relationships between the users of mobile devices.

The results of the investigations reported in this thesis show that, although large amounts of data are lost while gathering social social data using mobile devices, estimated ties are confirmed to be correct in the majority of cases.

Acknowledgements

This work would not have been possible without the help, support, and assistance of many people.

I would like to thank the following people especially: Dr James Irvine for giving me the opportunity to undertake this work, and for his advice and guidance throughout the last three and a half years; Dr Alisdair McDiarmid for sharing his expertise as a researcher and software engineer, without which this work would have taken considerably longer to complete; and my colleague Stephen Bell who through many illuminating discussions contributed to this work indirectly, and as the main developer of the Nodobo software also contributed directly to the results this thesis depends on.

I would also like to thank the Mobile Virtual Centre of Excellence for funding this research.

Lastly, I would like to thank my family and friends whose support and encouragement was a constant source of strength and inspiration to me during my time reading this degree.

Thank you all!

Contents

List of Publications

[1] Jamie Banford, Alisdair McDiarmid, and James Irvine. Multigraph Representation of Relationships for Enterprise Knowledge Networks. In *Proceedings of Wireless World Research Forum Meeting 22*, 2009.

[2] Jamie Banford, Alisdair McDiarmid, and James Irvine. Relationship Mapping for Proactive Growth of Knowledge Networks. In *Proceedings of IEEE 70th Vehicular Technology Conference*, 2009.

[3] Jamie Banford, Alisdair McDiarmid, and James Irvine. Estimating the Strength of Ties in Communication Networks with a Small Number of Users. In *Proceedings of The Sixth International Conference on Wireless and Mobile Communications*, 2010.

[4] Jamie Banford, Alisdair McDiarmid, and James Irvine. FOAF: Improving Detected Social Network Accuracy. In *Adjunct Proceedings of the 2010 ACM Conference on Ubiquitous Computing*, 2010.

[5] Jamie Banford, Alisdair McDiarmid, and James Irvine. FOAF Introductions: Automatically Growing and Improving the Accuracy of Detected Social Graphs. In *Proceedings of The Fifth International Conference on Pervasive Computing and Applications*, 2010.

[6] Alisdair McDiarmid, Stephen Bell, Jamie Banford, and James Irvine. Nodobo: Mobile Phone as a Software Sensor for Social Network Research. In *Proceedings of Wireless World Research Forum Meeting 25*, 2010.

[7] Jamie Banford and James Irvine. Estimating Social Graphs in an Education Environment. In *Proceedings of Wireless World Research Forum Meeting 27*, 2011.

[8] Jamie Banford and James Irvine. Estimating Social Graphs in an Education Environment. *IEEE Vehicular Technology Magazine*, 7(1), 2012.

1. Introduction

Contents

1.1. Motivation

This work is motivated by the a need for a tool to improve research into social relationships.

Social scientists are still conducting analysis of interpersonal ties on data gathered by observation in the first half of the 20th century [1, 2]. The widespread adoption of the mobile phone in the last decade and the ubiquitous nature of many mobile devices in the present day creates a potential opportunity to gather vast amounts of human social data [3, 2] but the difficulty of gathering this data from mobile devices, and ensuring the veracity of gathered data still provide challenging research questions.

A huge increase in the use of Social Network Services has occurred in recent years. Services such as Facebook and LinkedIn now have more registered users than many of the countries in the world have citizens: Facebook has over 500 million registered users [4], and LinkedIn has over 100 million members [5].

Surprisingly, networking—the creating and maintaining of social relationships—is not the primary use of these services [6]. Although people use social networking services to communicate with a small subset of their contacts, the primary use of these services appears to be the dissemination of information among peer groups. The details of activities, the other people involved, and associated media (photos, videos, and hyperlinks), are all shared through users' *activity streams* [7].

Social networking services are becoming popular on smartphones and tablets. Existing Social Network Services have ported their services to mobile environments, and new social network services have been launched specifically for mobile devices. However, these services are limited. They rely on the user to complete the task of updating their information; have only one dichotomous method of denoting complex, changeable social relationships; and fail to take advantage of the wealth of information available from communications networks which could add significant value to these services.

Much of the success of social networking sites could be attributed to some users' desire to be seen to be widely socially connected. The designers of these systems initially assumed that people would mark their real-life contacts, expecting that contacts marked on some social web service would have a corresponding real-world social relationship [8]. However, social network sites consider ties between users as bidirectional: if you are my friend, then I am also your friend. Therefore, a small group of power users can create a large number of ties in the network, and even increase the network size through inviting new participants, simply by pursuing this goal of an artificially inflated 'friends' list. This phenomenon also applies at the

smaller level: users may derive some satisfaction from the experience of listing their friends and demonstrating that they are popular [8].

The act of declaring 'friends' on an Social Network Service is also inaccurate: cognitive networks—social networks as perceived by the individual–are not always accurate, with one study finding that people could only recall their social network with 50% accuracy [9]. Therefore, a user's social network as listed on an Social Network Service is not necessarily representative of their actual social network. A tie to another person on Facebook or LinkedIn does not always match with a tie in real life. This inaccuracy is not obviously damaging to the SNS themselves, who benefit from increased usage of their site through advertising revenue from page-views, but it devalues the network for its users. If the social networking sites do not accurately represent their participants' broad social network, then they cannot be reliably exploited to benefit the users in their everyday working or personal life.

Social Network Services which are based on a network of socially-aware mobile devices, which will enable people not only to communicate, but help manage their manage their social ties allowing them to discover and share information with those in their social network more easily, have the potential to influence human interaction in the workplace and beyond.

1.2. Thesis Argument and Problem Statements

The thesis argument put forward by this document is given below. It is followed by three problem statements which elucidate the thesis argument.

Thesis Argument: *Metadata associated with the telecommunications protocols available on mobile devices contains information about communicants' social relationships. The existence and certain aspects of the nature of those relationships can be revealed by analysis of this metadata.*

Gathering Communications Metadata. Mobile devices are potentially useful social data gathering instruments. However, the veracity of any data gathered must be established in order that it can be used for analysis.

Detecting Social Relationships. Communications metadata, such as the time or duration of a voice call, contains latent information on the existence of social relationships between communicants, but, at present, there are no clearly established techniques for accessing this information.

Comparing Detected Relationships. All social relationships are not equal. Quantitative methods to distinguish between different detected relationships analogous to the qualitative 'stronger' and 'weaker' social relationships proposed in related social science literature are required.

1.3. Contributions

In addressing each of these problem statements, this thesis makes four contributions:

A Dataset. A new dataset of social interaction data using mobile phones is presented. Communications metadata from a group of twenty-seven high school students was gathered over a period of five months. Data on phone calls and text messages, Bluetooth device discovery, WiFi access point, and cell tower ID was gathered. The dataset has been made freely available.

Dataset Accuracy. The veracity of two freely available datasets is assessed by analysing the corroborating records in the datasets. It is found that at best only half of communications metadata gathered on mobile devices is corroborated by corresponding data gathered on other devices.

Reliable Detection of Social Ties. A ruleset which can reliably infer the presence of a social relationship based only on data gathered using mobile devices, and with no prior knowledge of the relationships between communicants, is presented.

Classification of Social Ties. Methods to classify detected social relationships based on patterns in communication metadata are investigated.

1.4. Structure of Thesis Document

This thesis contains six chapters. The remainder of this, the first, chapter provides some background information about the project on which the author worked as a research student which lead to this thesis, and discusses the suitability of mobile devices as sensors for social relationships.

Chapter 2 gives an overview of detecting co-proximity. It discusses the importance of co-proximity, detecting co-proximity with bespoke devices, and with mobile devices. It introduces a key piece of related work—the Reality Mining study [10]—and outlines key results from that work on using Bluetooth enabled mobile devices to detect co-proximate individuals.

Chapter 3 examines the interdependence between mediated communications and social relationships. The concept of communication network graphs is introduced and the necessary system architecture to create them is discussed. Communication network graphs derived from real data are analysed, and the need for corroboration in datasets is discussed.

Chapter 4 presents a novel estimated social graph derived from co-proximity data and communications metadata, and presents comparisons of the estimated social graph with the social graph reported by the participants and the social graph observed by the researchers.

Chapter 5 examines tie strength: investigating the phenomenon of triadic closure and how it relates to neighbourhood overlap in small mobile network datasets.

Chapter 6 draws conclusions and discusses future work.

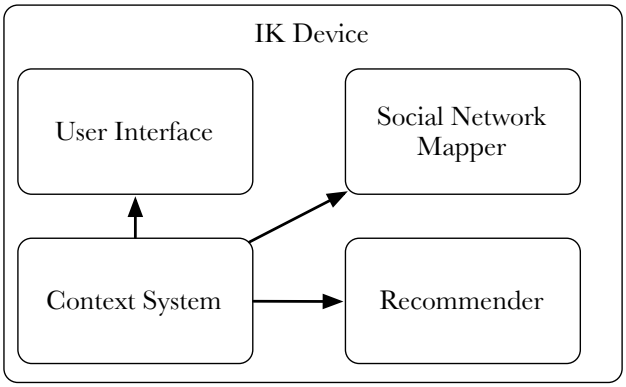
1.5. PhD Background

The work reported in this thesis document was undertaken as part of the Mobile Virtual Centre of Excellence Core Four Project: Instant Knowledge [11, 12]. The Instant Knowledge project aimed to monitor the device state and sensor data from the handheld devices carried by employees in modern workplaces, to foster informal workgroups, and gather information about employees' knowledge and skills.

The Instant Knowledge system consists of several components on each device [11] as shown in Figure 1.1. A context system gathers pertinent context data from the device and send it to the other system components as appropriate. A social network mapper automatically gathers and shares the information required to build an interpersonal network, mapping the links between communicants. A recommender system provides context-triggered recommendations by employing machine learning techniques which pro-actively provide the user with useful contact suggestions from the interpersonal network. Distributed recommendations can also be requested from peers, so that the system still works in a local ad-hoc environment where wider Internet access is temporarily unavailable. A third component ensures that the user feels in control of the information being communicated between devices, unobtrusively prompting the user for human input to guide the machine learning components, and ensuring privacy and security.

The Instant Knowledge project was undertaken by the University of Strathclyde in conjunction with two partner institutions, Royal Holloway, University of London, and The University of Southampton. The project was divided into three work packages, and each partner institution had responsibility for a specific work package.

Figure 1.1.
The Components of the IK System.



The University of Strathclyde was responsible for work involving mobile devices; Royal Holloway, University of London for security and privacy research, and The University of Southampton for machine learning and recommender systems.

The research undertaken at the University of Strathclyde was further subdivided into three work tasks: network implementation, communications metadata, and user control and usability. All work on the Instant Knowledge project undertaken at the University of Strathclyde was carried out in the Mobile Communications Group, Centre for Intelligent and Dynamic Communications, Department of Electrical and Electronic Engineering under the supervision of Dr James Irvine.

The work undertaken by the author was on the network implementation work task. It focused on building and growing the interpersonal network of a user of the Instant Knowledge system. The work in this task developed protocols which enabled the exchange of contact information among users of the system, enabling establishment and expansion of the number of connections in the network. The research also defined the meaning of having such a link between two people, and determined how to describe their relationship. Once methods of establishing a network were defined, the task examined techniques to pro-actively apply the protocols to build the network without user effort.

The communication metadata work task was undertaken by Dr Alisdair McDiarmid. It developed techniques to extract and describe useful context information from ongoing mobile communications, in order to retrieve implied social links between communicants. This communications metadata must be presented in a manner which allowed it to be understood by several other system components. Therefore, an ontology specification was a key outcome of this area of work. The challenges in this work task were in determining how much information can be found

within communications sessions, defining that which may be useful for building and extending the interpersonal network, and creating a syntax for describing this information which will scale to future communications systems.

Lastly, the user control and usability work task was undertaken by Stephen Bell. The interpersonal network in the Instant Knowledge system was designed to be largely automatic in its operation, requiring no direct user input. However, it was vital that the user felt in control of the information being communicated between devices. Automation is only useful so long as it is actually helpful; all intelligent assistant systems require human input to guide them in the correct direction. This task examined the human interface issues raised by this new system. Exposing the network information to the user in a usable manner is vital to the success of the system, and the research determined how the user should be able to control the interpersonal network.

While each of these work tasks pursued a different thread of research, there was some overlap between work tasks. Moreover, a key output of the Instant Knowledge project was a demonstration system, and as the project progressed the disparate threads of research became intertwined into a holistic demonstration of all of the research undertaken as part of the Instant Knowledge project. The thesis work presented here is the work of the author within the collaborative working environment described above, and can be considered the author's own work in all cases except for the study described in Section 3.5. In this case work undertaken on a side project involving the author and Stephen Bell is presented. Relevant, background information about the study, the research methodology employed, and other actors involved and the role the played in the study is given in Section 3.5.1.

1.6. Social Relationships and Mobile Devices

Mobile devices can be used to detect human relationships. However, before considering the technical details of how this is possible, and what can be achieved by doing so, it is necessary to examine two more fundamental issues: clarification of what is meant by a social relationship, and an explanation of why mobile devices are relevant when attempting to detect them.

In related work the term *social tie* is often preferred to social relationship. The word relationship can be misleading. The more common usage, which describes romantic involvement or kinship, is not the sense in which it is used here. Instead the word tie is used to describe the presence of any social link or bond, and avoid the more emotive connotations of relationships.

To define a social tie we begin with the work of Erving Goffman. In his book *Relations in Public: Microstudies of the Public Order* [13] Goffman describes a set of rules which outline how the two individuals—or *ends* in Goffman's terminology—who participate in the tie should behave when they are face-to-face. These rules define a tie as a set of mutual standards of behaviour based on roles. The ends of a tie engage in specific activities in established situations, each situation has a set of necessary interactions between the individuals, and repeated dealings between two individuals in one type of situation result in a role relationship.

The correct behaviour between the ends of a tie is based first and foremost on identification. Goffman separates identity into two categories: *social* identity and *personal* identity [13].

The social identity of a individual is perceived during simple social interactions. It is derived from the broad social categories to which they can belong, for example age-grade, race, sex, or class. Organisations and groups can also be considered as social categories in this sense. Knowledge of an individual's personal identity however, requires that you know them personally. It is what Goffman calls the 'unique organic continuity' attributed to each individual, and is established through distinguishing features such as name and appearance, and elaborated by knowledge of their biography and social attributes. The distinction between social and personal identity results in two main categories of relationship: anonymous relations, and anchored relations [13].

Anonymous relations are the patterned interactions between two individuals who identify each other solely on the basis of instantly perceived social identity. The ends engage in fleeting and distant interactions which are so commonplace they may even pass unnoticed. (Goffman describes 'an individual courteously passing a stranger in the street' as an example.) Anonymous relations might not undergo much development, and may simply be passing interactions between individuals who never meet again [13].

Anchored relations involve each individual identifying each other personally and knowing the other does likewise. There are three fundamental aspects to this process: the individuals involved openly acknowledge the relationship to one another; the process of personal identification is irrevocable (individuals cannot become unacquainted, and will make a faux-pas if they forget that they are acquainted); and the personal acquaintance begun between them is mutual and understood to be so. This creates a well defined starting point for anchored social relationships. From this starting point each anchored tie develops over time and gains a history shared by both ends of the tie [13].

The status of a relationship at any time is displayed as a *tie-sign*; a signal which allows the status of the relationship between two people to be determined by an observer. Tie-signs are part of what Goffman called the *interaction order*; the area of all face-to-face interactions between people. They include all evidence about relationships involving objects, acts, expressions and only exclude the literal aspects of explicit documentary statements. These kinds of signals provide information about the current state of a relationship but do not allow the entire history of the tie to be determined. They are usually explicit through body placement, posture, gesture, and vocal expression of the individuals currently present in a situation [13].

When used in this work the term *social tie* refers to the anchored relations defined by Goffman. They are unique social bonds that begin in an irrevocable and mutual introduction between individuals, and whose current status is displayed via numerous tie-signs.

The interaction order described by Goffman has been the basis for other work investigating social interactions between mobile phone users, and although Goffman was not particularly influential in his lifetime, in the decades since his death the number of his adherents has increased and his importance is noted in sociological circles as well as in other disciplines [14].

In *New Tech, New Ties* [15] Rich Ling expands on Goffman's theories on face-to-face interaction when he discusses the changes to social cohesion caused by mobile communication. He notes that although

> Goffman is a fruitful source of insight into the use of mobile telephones in co-present situations. The question remains, however, how his very physically co-present analysis can be applied to mediated situations [15].

Goffman's work considers face-to-face interaction almost exclusively, although, as Ling points out, this is because the level of mobile interaction common today was unheard of in Goffman's time. Ling states that it is therefore unnecessary to limit interaction to the physically co-present but only limit it to the *perceptually* co-present. He argues that Goffman's ideas remain relevant to mobile communication because mobile communication allows individuals to maintain interactions when not physically co-present [15]. He cites empirical studies of the behaviour of mobile phone users in France [16] and Japan [17] which find that co-present interactions can blend seamlessly into mediated interactions.

The interaction order described by Goffman is also cited by Mark A. Smith. In his essay *From Hyperlinks to Hyperties* [18] he describes how the analogue and ephemeral nature of tie-signs makes detecting and analysing them extremely difficult. He proposes that the widespread use of computer-mediated communication channels

has created a class of ties which are significantly more easy to detect and analyse, and describes how

> the social world is becoming 'machine readable'. Social networking sites […], Web discussion boards, e-mail lists, private instant messaging, and such emerging channels as graphical worlds are all examples of the expansion of the interaction order into machine-readable media. But they also illustrate the limits of these tools for impacting the primary interaction order of face-to-face encounters. Some edge toward the interaction order, as when people use mobile phones or laptops to instant message or e-mail one another while in the same meeting or room. But much of the activity of the face-to-face interaction order is not inscribed in a systematic and widespread manner.
>
> Ties in computational media take on new attributes that are distinct from ties in the physical world. Computational ties are machine readable; can be collected from a wide range of ongoing events and systems; and can be aggregated, searched, and analysed in ways that reveal patterns and connections not previously visible [18].

The postulation that tie-signs are machine readable is central to the work contained in this thesis. The explosion in the availability and popularity of computer-mediated communication channels in recent years has led to unprecedented levels of digital interactions which are available for analysis. Modern mobile devices are relevant to any attempts to discover social relationships because they are potential repositories of *all* of the examples of machine-readable ties discussed by Smith, with all of the communication channels he describes potentially available on modern mobile devices. Furthermore the short range communication channels available almost universally on these devices allow insights into the face-to-face interactions of users as well as mediated interactions. Mobile communications is a discipline perfectly placed to build tools to explore social relationships: they allow access to many new digital tie-signs from many different communication channels; and their mobility and ubiquity combined with their ability to detect other local devices create the possibility of analysis of face-to-face interactions too.

2. Detecting Co-proximity

Contents

2.1. The Importance of Co-proximity

The interaction order described by Erving Goffman is based on face-to-face interactions and although, as discussed in the previous chapter, Goffman's ideas can be extended to apply to mediated interactions, knowledge of physical co-proximity is also relevant to the study of social ties. Mobile devices can create new possibilities for social interaction, but many social interactions still take place between people who are face-to-face, particularly when they are exchanging complex information [19].

In fact face-to-face interactions are still the dominant mode of interaction between people. In *Social Interactions Across Media: Internet, Telephone, and Face-to-Face* [20] Baym et al. report results from two studies into the social interactions for college students conducted at two large midwestern universities. They compare interactions on the phone, via the internet, and face-to-face and find that of all interactions reported by the students 64% were face-to-face. Interactions were rarely exclusively face-to-face however: 64% of participants conduct interactions face-to-face, on the phone, and online. No participants interact exclusively using the telephone, and only one participant reported exclusively face-to-face or internet interactions respectively.

In this work, the term *co-proximity* is defined as 'corporeal co-presence' [21]. That is, where individuals are physically present in the same place at the same time, and sense that they are close enough to be perceived in whatever they are doing, and to be perceived in this sensing of being perceived [22, 23].

The proximity of individuals can be itself a kind of tie-sign, revealing information about the nature of the tie between them. In *The Hidden Dimension* [24] Edward T. Hall proposes four 'distances in man' which signal some information about the relationships between co-proximate individuals and the activities that they are undertaking. They are intimate distance, personal distance, social distance, and public distance. Each of these zones is then further subdivided into a close and a far phase.

At intimate distance the presence of the other person is unmistakable. In the close phase individuals are touching, and in the far phase they are between 15cm and 45cm apart. Personal distance corresponds to the term of the same name in Anthropology [24], and designates the distance consistently separating members of a non-contact species. Known colloquially as 'personal space' it creates a small protective bubble that individuals maintain between themselves and others. In humans, the close phase is within touching distance between 45cm and 75cm apart, and the far phase is 'at arm's length' from 75cm to 120cm. Social distance is far enough away that intimate visual detail in the face is not perceived. Nobody touches or expects to touch another unless there is some special effort, but conversation is

possible with normal voice level. It ranges from 120cm to 2m in the close phase to 2m to 4m in the far phase. Public distance is well outside the circle of involvement. Faces can be perceived but not in detail, and voice level must be risen to converse. The close phase is from 4m to 7m and the far phase greater than 7m [24]. The diferences between each of the eight phases is shown in Figure 2.1.

2.2. Detecting Co-proximity with Electronic Devices

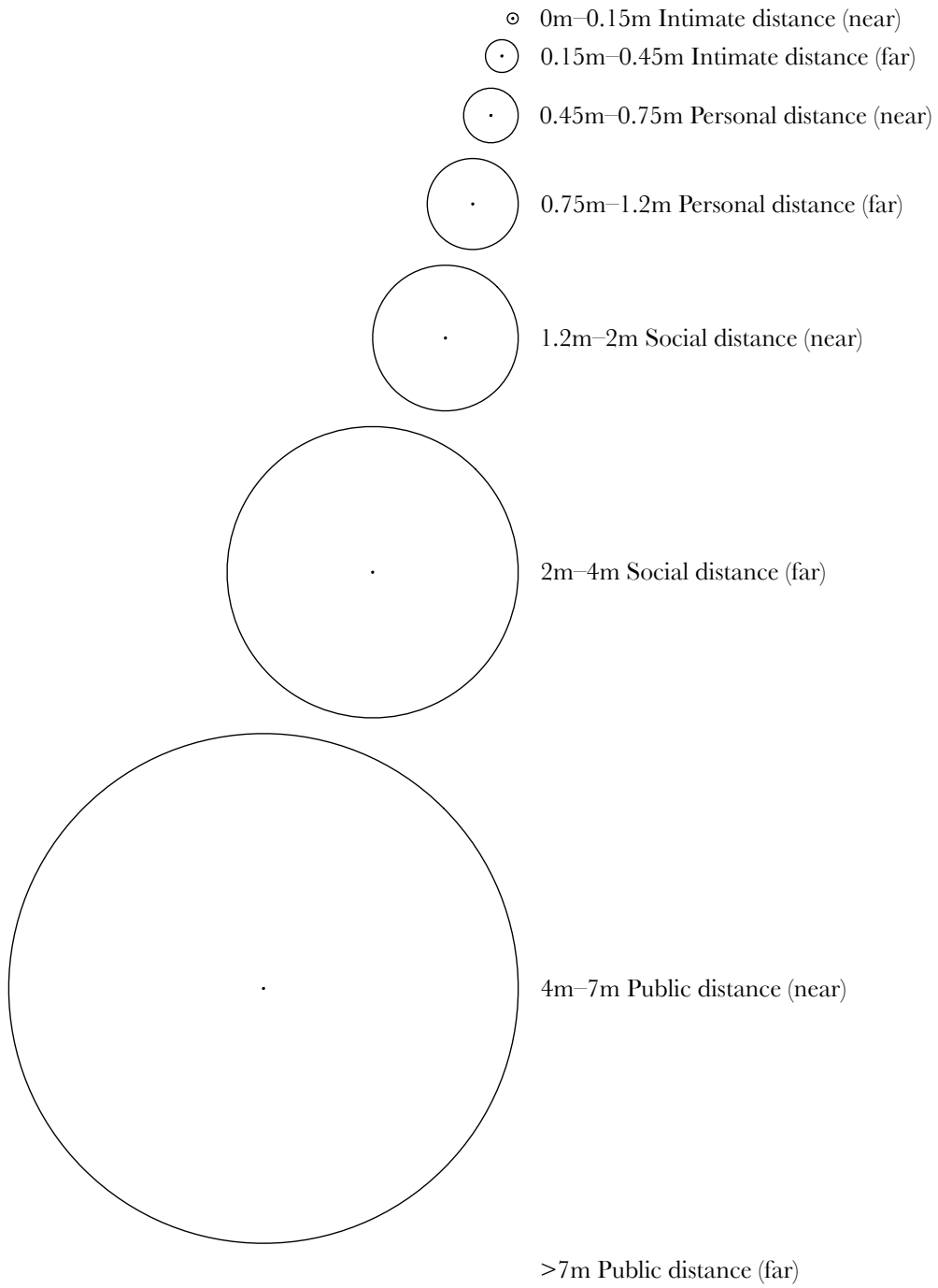
Research has been undertaken in the field of ubiquitous computing into building devices which detect the distances between individuals [25, 26, 27, 28]. These devices are typically worn or carried, are powered by microcontrollers, and communicate with other devices using short-range communications protocols.

In *Supporting Group Collaboration with Interpersonal Awareness Devices* [25] Holmquist et al. propose the development of 'interpersonal awareness devices' (IPADs). These are self-sufficient devices identified with a particular user which are carried or worn at all times. They utilise the relationship between the devices (such as co-proximity) to convey awareness of other IPAD users. However, these devices do not mediate any communication between users; they simply communicate some information about other nearby devices.

A prototype IPAD–the Hummingbird–is discussed. It creates awareness of other co-present devices by producing an audible 'hum' when two or more Hummingbirds belonging to users in the same group are within 100 meters of one another. At this distance the granularity of co-proximity detected is solidly in the the far-phase of the public distance described by Hall, and as such the Hummingbird only begins to illuminate the ties between users. However, these devices do start to augment the interaction order slightly. Users become aware of the presence of others through otherwise impenetrable physical obstacles, and in locations where they would not expect to meet [25].

In *The Familiar Stranger: Anxiety, Comfort, and Play in Public Spaces* [26], Paulos and Goodman propose developing mobile devices which explore social phenomena between co-present individuals. The focus of the work is on 'familiar strangers', a term introduced by Stanley Milgram [29] to describe individuals who are repeatedly observed but never interacted with. Milgram uses the example of commuters, describing standing next to the same people at a train station every day for years but never conversing with them. Familiar strangers are a specific type of anonymous relation which form a border zone between an individual's acquaintances and complete strangers, and, because they are repeatedly observed, establish connections

Figure 2.1.
Hall's 'Distances in Man'.



to specific locations [26].

Paulos and Goodman discuss the the *Jabberwocky*: a device for detecting familiar strangers [26]. They are personal, wearable, devices which use Bluetooth to detect other nearby devices. As two people approach each Jabberwocky detects and records the presence of the other. Jabberwockys can also be placed on objects and at certain locations allowing connections between familiar strangers and locations that they inhabit to be made. The range of connectivity is stated to be up to 30 meters giving better resolution to detected co-proximity than the Hummingbird, but still the far-phase of public distance. The ability to vary the power of the radio on a Jabberwocky is mentioned in passing, alluding to the possibility of more sensitive detection of co-proximity [26].

Devices in the form of 'active badges' [30] have also been used to attempt to infer the presence of social ties from detected proximity. While the original active badge concept focused on location-aware applications using badges equipped with infrared sensors [30], badges have also been used to detect relationships between people [27, 28].

In the GroupWear project [27], Borovoy et al. discuss using badges to display the similarity between users when they are face to face. User co-proximity is detected by infrared sensors on badges, and similarity is measured by comparing the number of matching responses to multiple-choice questions. In order to ensure GroupWear tags correctly identified situations where users were face-to-face, the authors designed the tags so that they only detected another tag when the users were facing each other, and at a normal conversation distance [27].

Although the range of the GroupWear badges is not explicitly stated, the phrase 'normal conversational distance' implies the close phase of social distance and therefore a range of up to 2 meters. In related earlier work Borovoy et al. describe the implementation of a similar device, the *Thinking Tag* [28], in detail. In these devices the transmit power of the infrared transmitter was limited to 'about five feet' (1.5 meters), exactly in the close phase of social distance [28].

Devices which detect co-proximity with infrared sensors and gather data on social ties have also been proposed. In their paper *Sensing and Modelling Human Networks Using the Sociometer* [31] Choudhury and Pentland describe wearable devices which detect conversations. Sociometers are personal devices which uniquely identify individuals. They use infrared to detect co-proximity, and a microphone to detect speech. Like Jabberwocky devices Sociometers also store information about the interactions they detect locally in memory on board the device [31].

Four low-power infrared transmitters are used in the Sociometer. They create a

cone shaped detection area in front of the Sociometer with a range of 'approximately six feet' (1.8 meters) [32]. Like the GroupWear tag the Sociometer detected other devices in the close phase of social distance. Focusing the detection area to the front of the device ensures that only individuals directly in front are detected, not those in a circle around the device (as with devices with radio transmitters like the Hummingbird). Infrared sensors provide information about face-to-face co-proximity but the microphone is required to detect conversations. The microphone is placed slightly below the user's mouth on the chest, this close placement allows speech by the user to be separated from other speech and ambient noise by thresholding. The data gathered by the Sociometers can then be analysed to detect 'episodes'; continuous periods of time where users are both co-present and speaking and hence presumed to be conversing [32].

Sociometers were deployed in an experiment in the MIT Media Lab where 23 people wore the devices for six hours a day for eleven days [31]. After the deployment, pairs of conversations were identified and the links between individuals mapped. Links were calculated from the number of episodes which were greater than 5% of the total interaction time for each participant, and were used to map the network of interactions which were detected. Choudhury and Pentland presented their estimations of conversations to the participants to gain a measure of ground truth, and correctly identified 87.5% of conversations of length one minute or greater [31].

The Sociometer is the first device which truly attempts to detect the presence of social ties, and appears to be capable of detecting co-proximity and conversation interactions with a high level of accuracy. However, implementation issues such as how the data was retrieved from each Sociometer and collated for further analysis, and design details such as the precise definition of what constituted a link are not discussed, making more in depth analysis impossible. (For example, did a conversation have to be detected by both Sociometers to be considered to contribute to a link or is detection by one sufficient?)

Although extensive details of the implementation and design of the Sociometer studies are not available, Choudhury et al. describe the technical specification of a related device: the *Mobile Sensing Platform* [33]. The Mobile Sensing Platform gathers more general context data on individuals, not specifically information on social interactions, but the type of technical problems encountered are similar. When discussing the lessons learned from deployments of the Mobile Sensing Platform Choudhury et al. highlight issues relating to transferring data gathered locally on devices to external devices for preprocessing, logging, and analysis. In particular, they

they highlight the problem that

> connectivity wasn't reliable enough to continuously stream sensor data […]. The packet losses and intermittent connection drops required us to switch to a wired solution where the sensor board was physically attached to a PDA via a USB cable.
>
> We used this [combination of equipment] to collect [our data]. This solution worked as a temporary research prototype, but it clearly wasn't feasible longer term because it required participants to carry a bulky, wired device combination simply to collect data [33].

Although wearable sensors can detect co-proximity, the technical problems associated with processing data on the device and transmitting data from one device to another limit the usefulness of wearable sensors to all but the simplest of research applications. The addition of a mobile device allowed proof of concept to be established but was ultimately unworkable.

In addition to the technical and logistical problems with bespoke wearable sensors, cost also becomes a limiting factor when attempting more sophisticated applications. The MIThril platform [34] was developed in the MIT Media Lab at the end of the 1990s. It was a prototyping system for wearable pervasive computing applications which linked a range of sensors to a CPU by a common bus. Initially, like the Sociometer and Mobile Sensing Platform, custom hardware was used, but at a cost of more than $3000 per unit in 2001 it was apparent that it would be difficult to build and maintain more than a handful of devices [34]. However, by 2003

> the availability of inexpensive, Linux-capable PDAs with significant signal processing and communications capabilities [allowed the development of devices] with many of the capabilities of the original MIThril system, but at a fraction of the original MIThril 2000 system complexity and cost [34].

More complex applications require additional processing power and more robust communication channels. Rather than extending the functionality of the bespoke devices already developed researchers turned instead to mobile communication devices, which by the early 2000s had significant processing power, memory, and various communication channels available making them ideal tools for researchers to build on the early work on detecting co-proximity described here.

The potential of mobile devices as platforms to detect co-proximate interactions was also noted in earlier work. Paulos and Goodman [26] mention the suitability of mobile phones to replace custom iMote devices:

> Jabberwockies require a low power localised radio, limited processing, and small storage. Today's Bluetooth enabled mobile phones satisfy these constraints and make an ideal platform to develop a personally carried Jabberwocky application. These Bluetooth enabled mobile phones support the same interactions and metaphors as personally carried iMotes.

Although it is possible to investigate co-proximate interactions by building bespoke hardware, the complexity and cost of such an approach is unfeasible. Applications which require the storage and processing of data, and the availability of robust communications channels require the use of PDAs or mobile phones. These devices have the processing power, memory, and communications capability required for more complex investigations of co-proximate interactions. Often the various sensors available on these devices are sufficient to replace the external sensors used in earlier platforms. Mobile devices, albeit with some modifications to the operating system and with additional software, can be used as complete sensor nodes when attempting to detect co-proximate interactions.

2.3. Detecting Co-proximity with Mobile Phones

In their paper *VibeFones: Socially Aware Mobile Phones* [35], Madan and Pentland propose that mobile phones are

> really the researcher's wearable computer in disguise. People carry their mobile phones for most of their day, which makes them ubiquitous wearable sensors that can collect continuous, long-term, behavioural and social data […]. [However, mobile phones] are *social* in a very limited sense of the word—while they connect users and support sharing of information, they understand very little about the user or the nature of the interaction itself [35].

It is from such a premise—that mobile phones can be considered to be ubiquitous conveyers of social information, but have no understanding of the nature of the interactions they facilitate—the authors postulate that mobile devices can be made to be socially aware, that is, capable of somehow perceiving the social ties between their users [35]. Like the Sociometer Madan and Pentland propose to combine speech and proximity information in order to detect social interactions. Rather than only detecting the presence of conversations they also seek to classify them according to elements such as rhythm, stress, and intonation which can give insights into the emotional state of the speaker. Unlike the Sociometer however, proximity is detected by Bluetooth radios integrated into the phone. This means that although

no additional sensors are required, face-to-face interactions cannot be directly detected [35].

The concept of a socially aware mobile device is a compelling idea, and Madan and Pentland outline how it could be made possible with Linux-based mobile phones and various software enhancements, but they only allude to the implementation of their concept and no results are presented. A more detailed implementation of mobile phones as social sensors is given in another project undertaken at the MIT Media Lab (in which Pentland was also involved), Reality Mining.

2.3.1. The Reality Mining Study

The Reality Mining Project was carried out in the MIT Media Lab in 2004. One of the main aims of the project was to show that mobile phones alone could be used to gather data on social ties. No additional sensors need to be integrated with the mobile device like those in MIThril or the Mobile Sensing Platform. Reality Mining also aimed to show that this kind of data gathering exercise was scalable, and hence practical [2].

One hundred Nokia 6600 mobile phones with custom operating systems were deployed over the course of an academic year. Seventy-five participants were either students or faculty at the MIT Media Lab, and the remaining twenty-five were students at the adjacent Sloan Business School. The dataset includes call logs, Bluetooth proximity data, and cell tower IDs of all participants, a total of approximately 450,000 hours of co-proximity, communication, and device usage behaviour [10]. A dataset containing data for all of the participants over a nine month period was made available to the research community.

The choice of handset was important to the study. Although the phones acted as sensor nodes without any additional hardware, custom software was required to gather the study data. The software used was ContextPhone which can only be run on Symbian Series 60 mobile phones [2].

ContextPhone was developed by Raento et al. and is described in detail in their paper *ContextPhone: A Prototyping Application Platform for Context-Aware Mobile Applications* [36]. It is a set of open-source C++ libraries and other software components which provide the context data available on mobile devices as a resource for developers. The platform is divided into four modules: sensors, communications, customisable applications, and system services. Sensors gather context data, including location data, user interaction data, and communication data. The communications module implements the standard communications protocols expected on the phone, local area with infrared and Bluetooth and wide area with GSM and GPRS. SMS

and MMS are also supported, as is XMPP. Customisable applications allows developers to build custom applications to a much greater degree than was possible using Symbian at the time. The systems services module provides the ability to start and run background services on series 60 phones (a feature which was not available previously).

ContextPhone enabled Series 60 devices allowed Eagle and Pentland to gather significant amounts of data during the Reality Mining Study. Reality Mining is noteworthy for two reasons: it was the first study to use mobile phones as social sensors where more than a handful of devices were deployed, and the availability of the data allowed many other researchers access to data that would have otherwise been unavailable. However, the dataset is not perfect.

In his PhD Thesis, Eagle identifies three ways in which errors were introduced into the dataset: data corruption, device detection failures, and what he refers to as 'human error'. Initial versions of the software repeatedly wrote over the same cells in flash memory cards in the phones, some of these cards failed due to their limited number of read-write cycles. Data for ten participants was lost due to corruption caused by this error in the first two months of the study.

The ability of the phones to accurately detect other co-proximate devices was also an issue. The range of Bluetooth doesn't guarantee that detected devices are co-proximate, and the scanning period of five minutes means that short interactions may be missed. A small percentage of interactions—between 1 and 3 percent—are not detected due to the Bluetooth server crashing.

Two classes of 'human error' are identified by Eagle: the phone being off and the phone being separated from the user[1]. The phone is either off because it was switched off by the user deliberately or because of exhausted batteries. In both cases it is possible that potential interactions were missed. When the phone is off the user may still interact with other people, but the phone has no way of detecting these interactions, and when the phone is separate from the user other co-present devices may be detected, but the phone is no longer an accurate proxy for the user.

2.3.2. You Are Your Cell Phone (Most of the Time)

The question of the applicability of a mobile device as an accurate proxy for a user is an interesting one. If one assumes that it is, and by extension that knowledge of the location or other context of a mobile device is synonymous with the context of the user of that device, then the possibilities are enormous. In his essay *You Are Your Cell*

[1] A user turning their phone off or not having their phone with them at all times is not necessarily an *error* on the part of the user. However, these actions by the human users of devices will introduce errors into the dataset, and this is presumably why Eagle chose the name human error.

Phone [37] Roy Want argues that a

> characteristic that sets mobile computers apart from desktop computers is that they're intimately associated with their user's daily life and experiences. […] Now that cell phones have become mobile and ubiquitous in their own right, we can take the proxy concept to a new level. […] A person's cell phone experiences almost all of the physical parameters that the person experiences—it feels the same forces, travels at the same velocity, is about the temperature, is exposed to the same sounds and pollution levels, and near the same people and equipment. By recording the state of sensors attached to a mobile phone, you're effectively recording its owner's experiences across a rich set of dimensions [37].

However, if a mobile device is not an accurate proxy for a user, then user context may be significantly different from device context. As Patel et al. [38] point out in *Farther Than You May Think: An Empirical Investigation of the proximity of Users to Their Mobile Phones*,

> this approach assumes that the mobile phone is an accurate proxy for the location of its owner. Intuitively and anecdotally, we know that people do in fact carry their mobile phones with them *much* of the time, but these same phones are not physically on their bodies nor within arm's reach at *all* times [38].

Patel et al. investigated users physical relationship to the mobile phone in an empirical study in 2006. Sixteen participants were given small plastic beacons to wear around their necks for three weeks. The beacons were to be worn almost all of the time—most subjects wore them even while sleeping—and are therefore assumed to be a reliable proxy for the individual. The distance between the phone and the beacon could then be measured and used to provide an estimate of the distance between the phone and the individual [38]. The beacon, worn around the neck as a pendant,

> emits a Bluetooth signal detected by a custom-built application on the user's Bluetooth-enabled phone. The application on the phone pings the tag every 60 seconds and approximates the distance based on the strength of the signal received. This method allows for the determination of three levels of proximity: within arm's reach (strong signal); in the same room (signal is weak or varied); or unavailable (signal could not be detected) [38].

The beacon uses a class 2 Bluetooth module which is reduced to -22 db to extend battery life and limit the maximum at which the phone can detect the beacon to around 5 or 6 meters. A signal loss of 5 db is assumed to account for absorption by the human body [38].

Rather than use a Received Signal Strength Indicator, which is inconsistently implemented across mobile phones if at all, Patel et al. implemented their own simpler signal strength indicator for proximity detection.

> In this solution, the round trip time of the Service Discovery Protocol packets are used to estimate the distance between the tag, the link quality should degrade. The lower link quality then increases the bit error rate and thus the number of packet retransmissions. The retransmits in turn increase the service discovery time.
>
> […] A phone within arm's reach typically shows a service discovery time of about 2000–4000 ms, room level distance of about 4000–7000, and no returned service discovery information is interpreted as the phone being out of range or further than room level.
>
> In practice, physical room level distance can result in fluctuating values between 4000 ms and no discovery. This fluctuation is likely due to a BER so high that the bluetooth module times out and does not report a successful discovery. One serious issue with this phenomenon is the difficulty in determining whether the phone transitioning from 'room level' and truly out of range or whether the phone is consistently at room level with erroneous fluctuation described. [38]

If high rates of fluctuation were observed—for example, where the range transitions with every reading—for more than five minutes Patel et al. classified the reading as room level.

The proximity levels detected varied. Phones are within arm's reach between 17% and 85% of the time, 58% on average. However, a significant increase in the average time the phone was within arm's reach is seen during times that they were away from home: over 70% of the time on average. Conversely, users were less likely to have their phone at arm's length while at home; only 50% of the time on average.

The assumption that mobile devices are an accurate proxy for users is valid for more than half of the time. It is more robust when the user is known to be somewhere other than the home, although some uncertainty as to the validity of the assumption remains. This uncertainty must be considered at all times during any attempt to infer user context from mobile device context, and care must be taken not to allow the assumption that a mobile device is an accurate proxy for a user to become specious and undermine the validity of the work.

2.3.3. Suitability of Bluetooth for Co-presence Detection

Discovery services are a crucial part of the Bluetooth framework. Using the service discovery protocol, device information, services and the characteristics of the services

can be queried before a connection between two or more Bluetooth devices is established [39]. By repeatedly running the Service Discovery Protocol on mobile devices it is possible to detect all the co-proximate devices at a given time.

The maximum range of a class 2 Bluetooth device is 10m. In practice it will be less than this, albeit, with high variance in bit error rates beyond distances of six meters reported [40]. If we assume that the effective range of Bluetooth is six meters, co-proximate users discovered using Bluetooth will be in the near phase of public distance at least, if not within social distance. Figure 2.2 shows the overlap between social and public distances, and the effective range of a class 2 Bluetooth radio.

One concern with using Bluetooth to continuously scan for co-proximate devices is the effect that this may have on the battery life of the devices used. However, related work by Kukkonen et al. [41] shows that repeated, regular use of Bluetooth does not significantly affect battery performance.

In *BeTelGeuse: A Platform for Gathering and Processing Situational Data* they report

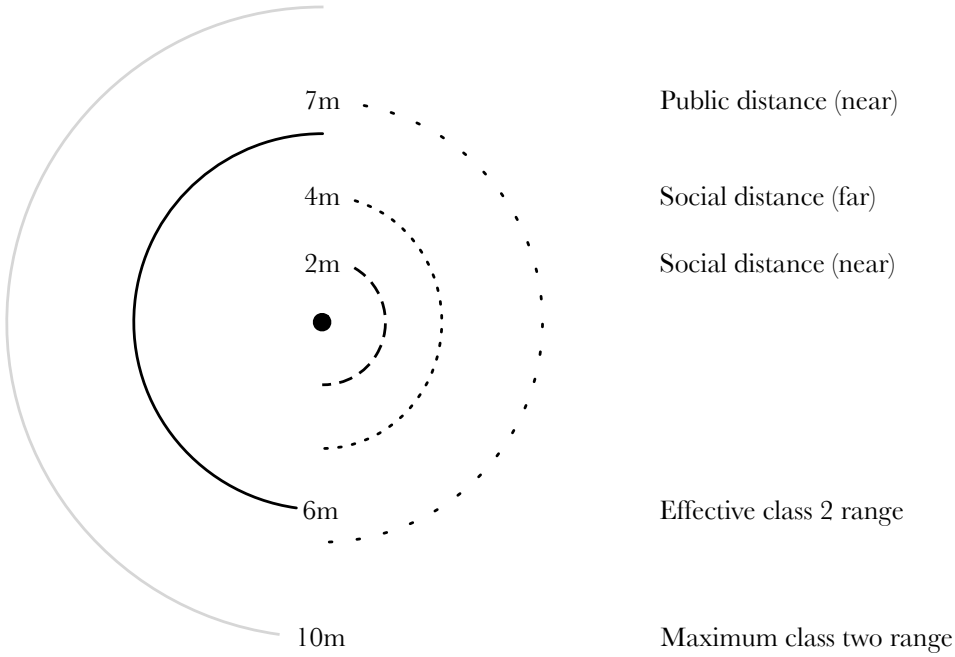
> a set of experiments in which [they] used BeTelGeuse under different configurations and measured time it took to drain the battery of five fully-charged, brand-new Nokia E61i devices (with standard 1500mAh batteries). […] As [a baseline they] used a version in which only the [data gathering software ran]. This version lasted between 35 and 36 hours. Adding a GPS device that was read once per minute decreased that battery life to 34 hours, running Bluetooth scans on top of this had only minor impact. These values span well over a day, which makes these setups well suited for long term data collection. Changing the GPS from periodic reading mode to continuous streaming had a more significant effect on the battery lifetime, with the mean lifetime decreasing to 25.7 hours. Again, the Bluetooth scanning had only minor impact on the performance (mean 25 hours). Thus, BeTelGeuse's battery usage is well suited to Bluetooth.

Bluetooth is well suited to detecting co-present devices. Discovered devices will be in the near phase of public distance, and may well be in social distance. Although it would be preferable to be certain that co-present devices are within social distance, the availability of Bluetooth on many mobile devices makes it an attractive choice. Moreover, the power requirements of running the Service Discovery Protocol continuously are not so strenuous as to adversely affect the battery life of the device.

2.3.4. Relationship Inference from Co-proximity Data

In his PhD thesis [2] Eagle discusses how

Figure 2.2.
Overlap Between Effective Bluetooth Range and Social and Public Distance.



> knowledge of the shared context of two users can provide insight into the nature of their association. For example, being near someone at 3pm by the coffee machines confers different meaning than being near her at 11pm at a local bar. However, even simple proximity patterns provide an indication of the structure of the underlying friendship network […]. [Eagle] trained a [machine learning process] to detect patterns in proximity between users and correlate them with the type of relationship. The labels for this model came from a survey taken by all of the experimental subjects at the end of two months of data collection […] [2].

This technique picked up the

> common sense phenomenon that office acquaintances are frequently seen in the workplace, but rarely outside the workplace. Conversely, friends are often seen outside the workplace, even if they are co-workers [2].

Further analysis of the correlation between co-proximity and social ties is given by Eagle et al. in *Inferring Social Network Structure using Mobile Phone Data* [42].

> Proximity is generally much higher for friends, but time and location are important predictors as well, where the ratio of proximity off [*sic*] hours outside work is much higher for friends than nonfriends. [Eagle et al.] therefore

> divided proximity into variables corresponding to on campus/off campus, daytime/nights (separated at 8 a.m. and 8p.m.), weekend proximity, and phone communication. A factor analysis revealed that two factors capture most of the variance in these variables. The first factor, which loads most heavily on proximity at work during the daytime is labeled 'in-role', as it represents the traditional behaviour between colleagues. The second factor, which loads most heavily on off-campus proximity in the evening and on weekends, is labeled 'extra-role' and is representative of behaviours outside the work environment. […] [It] is possible with a single parameter to accurately predict 96% of […] self-reported friendships based only on objective measurements of behaviour because the strong cultural norms associated with social constructs such as friendship produce differentiated and recognisable patterns of behaviour.

2.4. Summary

Face-to-face interactions are still the dominant mode of interaction between people. The proximity of individuals can be itself a kind of tie-sign, revealing information about the nature of the tie between them, for example individuals interacting at social distance.

Devices with a badge form factor and infrared sensors are capable of detecting social distance between individuals but they do not attempt to discover any further information about the ties between users. Instead they augment the existing interaction order by highlighting the presence of other users, and in doing so create the possibility of new types of social interaction.

Although it is possible to investigate co-proximate interactions by building bespoke hardware, the complexity and cost of such an approach makes it difficult, particularly for large studies. Applications which require the storage and processing of data, and the availability of robust communications channels require the use of PDAs or mobile phones.

Socially aware mobile devices could be made possible with smart phones and various software enhancements. These devices have the processing power, memory, and communications capability required for more complex investigations of co-proximate interactions. Often the various sensors available on these devices are sufficient to replace the external sensors used in earlier platforms. Mobile devices, albeit with some modifications to the operating system and with additional software, can be used as complete sensor nodes when attempting to detect co-proximate interactions.

The Reality Mining study was an important milestone in the field: the use of

mobile phones to gather large amounts of social interaction data proved that the approach was possible, and the accurate identification of many social ties proved that mobile devices can be suitable proxies for users, and that Bluetooth can be successfully used to detect co-proximate individuals.

3. Communication Network Data

Contents

3.1. Social Ties and Communications

Mobile devices are first and foremost communication devices, and human communication is intrinsically linked to social ties. In *Pragmatics of Human Communication* Watzlawick et al. assert that it is impossible not to communicate. The assertion is derived from the fact that behaviour has no opposite—that there is no such thing as non-behaviour—and that all behaviour during human interaction has some intrinsic message value, meaning it can be considered to be a communication. It follows that if one cannot not behave then one cannot not communicate [43].

> Activity or inactivity, words or silence all have message value: they influence others and these others, in turn, cannot *not* respond to these communications and are thus themselves communicating [43].

This fundamental coupling between behaviour and communication has interesting implications for our previous definition of a relationship: if a communication imposes behaviour as well as conveying information then any communication implies some commitment or tie and thereby defines the role relationship [43]. Building on earlier work by Bateson[1], the content and relationship aspects of communication are defined.

> The report aspect of a message conveys information and is therefore synonymous in human communication with the *content* of the message. It may be about anything communicable regardless of wether the particular information is true or false, valid, invalid, or undecidable. The command aspect, on the other hand refers to what sort of message it is to be taken as, and, therefore, ultimately to the *relationship* between the communicants [43].

If every communication has a content and a relationship aspect, relationships can classify the communications between the ends [43]. It follows that analysis of communications will yield information about the relationships between communicants. Detecting the presence of social ties can therefore be achieved by detecting communications and attempting to estimate the relationships associated with them.

Both email and mobile phone data have been used in attempts to infer the social networks of communicants. Kossinets and Watts [44] discuss the use of email communications of students, faculty and staff from a large university to investigate the underlying network of social ties in *Empirical Analysis of an Evolving Social Network*. They recorded the timestamp, sender, and list of recipients (but not the content)

[1] Here Watzlawick et al. cite Bateson, Gregory, and Jackson, Don D. 'Some Varieties of Pathogenic Organisation', in David McRioch (ed.). *Disorders of Communication*, vol. 42 (Research Publications, Association for Research in Nervous and Mental Disease, 1964), pp. 270–283. This article was not available to the author.

of 14,584,423 email messages over 355 days. All messages with more than four recipients were removed in order to remove mailing lists and other mass mailings and to ensure that interpersonal communications were accurately reflected. Eckmann et al. [45] also use email data to analyse the interactions between people at a university. In *Entropy of Dialogues Creates Coherent Structures in Email Traffic* they show that spikes in email traffic can be used to identify groups of collaborating individuals, and, like Kossinets and Watts, they use the reciprocal exchange of email messages to denote a potential social tie.

In *Structure and Tie Strengths in Mobile Communication Networks* [46] Onnela et al. use call data from a mobile network provider over an 18 week period to investigate the social networks of 4.6 million mobile subscribers. Users are considered to have a tie if they have at least one reciprocal exchange of phone calls during the 18 week period. The need for reciprocal calls eliminates a large number of one-way calls, most of which correspond to single events, suggesting that they typically reach individuals that the caller does not know personally [46]. Palla et al. [47] also use mobile call data to infer the presence of social ties. In *Quantifying Social Group Evolution* the authors examine call patterns between over 4 million subscribers over the course of a year.

Mobile phone data is also used in smaller scale studies. Laursen [48] discusses the relationship inferences which can be drawn from analysis of the content of calls and SMS messages in *Please Reply! The Replying Norm in Adolescent SMS Communication*. The dataset used contains 511 SMS messages and 287 calls between six 14 year old Danish school children (three boys and three girls) over the period of one week, gathered with the help of the network provider.

Although Laursen's study aimed to correlate patterns in communications with social ties which are already known to exist, not to infer the presence of new ties, the work is still relevant. She finds that if an SMS message is sent to someone with an important social tie then typically a reply is received within three minutes [48]. This suggests that reply time is another tie sign visible in communication metadata, and moreover, if messages are consistently replied to within three minutes then there is an important social tie between communicants. Reply time is therefore a useful indication of the presence of social ties.

The estimation of social ties from analysis of communications is an established method of estimating the presence of social ties, although studies vary in both the number of communicants involved, duration of the study time period, and the communication channels analysed. Despite this variation they have common features too: all are done retrospectively on data that is not normally available, and in order to gather this data some special permission is required to access it. Studies using

email data received permission to access the data stored in university email servers, and studies using mobile data use data from mobile network providers. The fact that the datasets used are rarely made public, potentially creates severe constants on researchers: those researchers wishing to study mobile communications data usually require the coöperation of a mobile network provider to access the data.

3.2. Communication Network Graphs

Social Network Analysis is concerned with detecting and interpreting patterns of social ties between people, organisations, or nations involved in a social relation. It is applied in various social sciences: to study kinship, friendship, and gift giving in anthropology; to investigate affective relationships in social psychology; to analyse power relations in political science; and to examine trade and organisational ties among firms in economics [49]. In this work, concepts from Social Network Analysis are used to provide a common language to describe both communication networks (and later the social networks estimated from them), and to provide tools for the analysis of these networks.

3.2.1. Relations and Network Analysis

There are three principal data types in the social sciences: attribute data, relational data, and ideational data [50]. Attribute data relate to the attitudes, opinions, and behaviour of agents. Relational data are the contacts, ties and connections, and group attachments and meetings which relate one agent to another (and so cannot be reduced to the properties of the individual agents themselves). Ideational data describes meanings, motives, definitions and typifications themselves.

Since the main goal of Social Network Analysis is detecting and interpreting patterns of social ties between people involved in a social relation it is primarily concerned with relational data. *Relations* are 'substantive connections', between the members of the network and a *network analysis* consists of a body of quantitate measures of network structure [51].

Wasserman and Faust [51] define a number of relations which may be measured in a network analysis: individual evaluations (friendship, liking, respect, etc.); transactions or the transfer of material resources (lending, borrowing, buying, and selling); transfer of non-material resources (communications and sending or receiving of information); interactions; movement (both physical movement, from place to place and social movement, between occupations or statuses); formal roles; kinship (marriage, descent) [51].

In this thesis, the relations of interest are social ties and as such the primary areas of interest in any network analysis are communications and interactions. One advantage of using mobile devices to capture relational data is access to communications metadata, for example the time, duration and frequency of phone calls between users can be recorded. Another—as discussed in Chapter 2—is the ability of mobile devices to detect co-presence.

3.2.2. Graph Theory: Notation and Terminology

Graph theory has been widely used in social network analysis as a means of formally representing social relations and quantifying important social structural properties [52]. Different terminology and notation is used in different related work. For clarity, the graph theoretic notation and terminology used in this work is discussed here.

Definition of a Graph A *graph* $\mathscr{G}$ consists of two sets of information: a finite nonempty set $\mathscr{N}$ of N *nodes*, $\mathscr{N} = \{n_1, n_2, \ldots, n_N\}$, and a proscribed set $\mathscr{L}$ of L *lines*, $\mathscr{L} = \{l_1, l_2, \ldots, l_L\}$. Each line is an unordered pair of distinct nodes, $l_x = (n_i, n_j)$ and is said to *join* n_i and n_j. If nodes n_i and n_j are joined by a line l_x they are *adjacent nodes*, and the line l_x and node n_i are *incident* with each other, as are n_j and l_x. If two distinct lines l_x and l_y are incident with common nodes, then they are *adjacent lines*. A graph with $\mathscr{N}$ nodes and $\mathscr{L}$ lines is called a $(\mathscr{N}, \mathscr{L})$ graph, denoted $\mathscr{G}(\mathscr{N}, \mathscr{L})$. The $(1, 0)$ graph is *trivial*; all other graphs are nontrivial. A graph that contains N nodes and no lines, $\mathscr{G}(N, 0)$ is *empty*. A graph is *labeled* when points are distinguished from one another by names [52, 53].

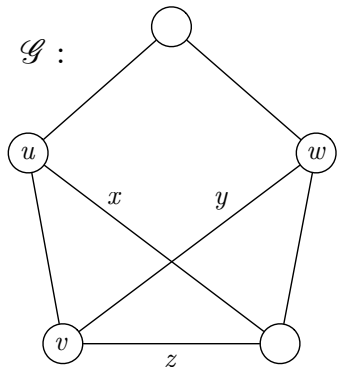
A *component* is the set of nodes to which a node belongs that can be reached from it by paths running along the lines of the graph [54].

It is customary to represent a graph by means of a diagram and to refer to it as the graph. Thus, in the graph $\mathscr{G}$ shown in Figure 3.1, the nodes u and v are adjacent but u and w are not: lines x and y are adjacent but x and z are not. Although the lines x and z intersect on the diagram, their intersection is not a point of the graph [53].

A graph is a model for a social network of undirected dichotomous ties. In a social graph, the nodes represent actors and lines represent ties between actors [52].

There are several variations of graphs. Note that the definition of a graph permits no *loop* that is, no line joining a node to itself. In a *multigraph*, no loops are allowed but more than one line can join two nodes; these are called *multiple lines*. If both loops and multiple lines are permitted, we have a *pseudograph* [53].

Figure 3.1.
A graph to Illustrate Adjacency. (**Adapted from [53].**)



If more than one relation is measured on the same set of actors, then the graph representing the network must allow each pair of nodes to be connected in more than one way [55]. For example, if a network analysis considered actors who called one another as well as actors who were co-proximate with one another, then two relations are measured on one set of actors.

A multigraph $\mathscr{G}$ consists of a set of nodes, $\mathscr{N} = \{n_1, n_2, ..., n_g\}$ and two or more sets of lines, $\mathscr{L}^+ = \{\mathscr{L}_1, \mathscr{L}_2, ...\mathscr{L}_R\}$. Let R be the number of sets of lines in the multigraph, and subscript the lines to denote to which set it belongs. If each relation is nondirectional, each line in each of the R sets is an unordered pair of distinct nodes, $l_{kr} = (n_i, n_j)$. A pair of nodes may be included in more than one set of lines. Since there are R sets of lines, each unordered pair of nodes may have from 0 up to R lines between them [55].

Many relations are directional—the ties are orientated from one person to an other. Mediated communications between people are examples of directional relations, for example 'made a call *to*' or 'received an SMS *from*'. In these cases information is sent from one person to another; one person is the source and the other is the destination of the information. In the network analysis of electronic communications the ties are directional and therefore the graph representing such ties must be directed [56].

A *directed graph* or *digraph* $\mathscr{D}$ consists of a finite nonempty set $\mathscr{N}$ of nodes together with a prescribed collection $\mathscr{L}$ of ordered pairs of distinct nodes. The elements of $\mathscr{L}$ are *directed lines* or *arcs*. An arc $l_x = \langle n_i, n_j \rangle$ is directed from n_i to n_j. The arc from node n_i to node n_j is not the same as the arc from node n_j to n_i ($l_x = \langle n_i, n_j \rangle \neq l_y = \langle n_j, n_i \rangle$), there are two distinct possible arcs for each pair of nodes. By definition,

a digraph has no loops or multiple arcs [53, 56].

Often social network data consist of valued relations in which the strength or intensity of each tie is recorded. Examples of valued relations include the frequency of interaction among pairs of people, or the rating of friendship between people in a group. Relations of this kind cannot be represented on a graph or digraph, since the lines or arcs in a graph or digraph are dichotomous [57].

A *valued graph* or a *valued directed graph* is a graph (or digraph) in which each line (or arc) carries a value, allowing valued relations to be represented. A valued graph consists of three sets of information: a set of nodes, $\mathscr{N} = \{n_1, n_2, ..., n_N\}$, a set of lines (or arcs), $\mathscr{L} = \{l_1, l_2, ..., l_L\}$, and a set of values, $\mathscr{V} = \{v_1, v_2, ..., v_L\}$, attached to the lines or arcs. Associated with each line or arc is a value from the set of real numbers. A valued graph is denoted by $\mathscr{G}_V(\mathscr{N}, \mathscr{L}, \mathscr{V})$, or simply $\mathscr{G}_V$ [57].

3.2.3. Dyads and Star Graphs

A *dyad* is the simplest nontrivial graph consisting of a pair of nodes and the possible line between the nodes. In a graph, an unordered pair of nodes can be in only one of two states: either the nodes are adjacent or they are not. Thus dyadic states for an undirected relation are dichotomous; either the actors have a tie present, or they do not [58]. The graph $\mathscr{G}_D(2, 1)$ is shown in Figure 3.2.

Figure 3.2.
A Dyad.

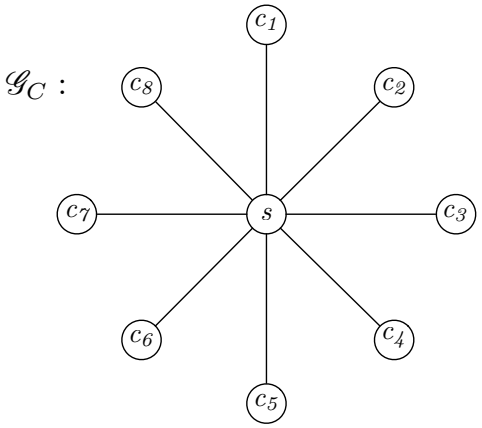
A dyad can show any dichotomous relation between two actors: if a mobile subscriber *s* has a contact *c* stored in the address book of their phone then the dyad would look like that shown in Figure 3.3.

Figure 3.3.
A Dyad Showing a Mobile Subscriber and a Contact.

If the address book of *s*'s phone has eight contacts, each dyad can be added to the graph to create a graph with a star topology. The graph $\mathscr{G}_C(\mathscr{N}_C, \mathscr{L}_C)$ has a set of

nodes $\mathscr{N}_C = \{s, c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8\}$ representing the subscriber and each contact, and a set of lines $\mathscr{L}_C = \{l_1, l_2, l_3, l_4, l_5, l_6, l_7, l_8\}$ where $l_n = (s, c_n)$. The graph $\mathscr{G}_C$ is shown in Figure 3.4.

Figure 3.4.
A Star Graph Showing a Mobile Subscriber and Eight Contacts.



Digraph dyads consisting of two nodes and the possible arcs between them are also possible. Since there may or may not be an arc in either direction for a pair of nodes n_i and n_j, there are four possible states for each dyad: null, two asymmetric states, and reciprocal.

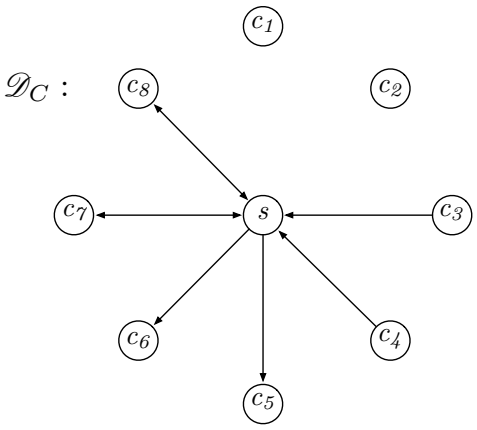
The *null* dyad has no arcs, in either direction between the nodes. The dyad for nodes n_i and n_j is null if neither the arc $\langle n_i, n_j \rangle$ nor $\langle n_j, n_i \rangle$ is contained in the set of arcs $\mathscr{L}$. An *asymmetric* dyad has an arc between the nodes in one direction or the other, but not both. The dyad for nodes n_i and n_j is asymmetric if either the arc $\langle n_i, n_j \rangle$ or the arc $\langle n_j, n_i \rangle$, but not both, is contained in the set of arcs $\mathscr{L}$. Thus, there are two possible asymmetric dyads. A *reciprocal* dyad has two arcs between the nodes, one going in one direction the other going in the opposite direction. The dyad for nodes n_i and n_j is reciprocal if both arcs $\langle n_i, n_j \rangle$ and $\langle n_j, n_i \rangle$ are contained in the set of arcs $\mathscr{L}$ [59].

If the digraph represents voice calls made between mobile subscribers, a null dyad represents neither subscriber calling the other, an asymmetric dyad represents one subscriber calling the other, and a reciprocal dyad represents both calling each other.

Instead of analysing the address book of subscriber s, if the call log is used as our data source a digraph $\mathscr{D}_C(\mathscr{N}_C, \mathscr{L}_C)$ may be created. The set of nodes $\mathscr{N}_C$ is identical to that in $\mathscr{G}_C$ but the set of arcs $\mathscr{L}_C$ now contains directed lines where $l_{nin} = \langle s, c_n \rangle$ or $l_{nout} = \langle c_n, s \rangle$. $\mathscr{D}_C$ is shown in Figure 3.5. There are no calls between s and c_1 or

c_2, relations which are represented with null dyads. Calls are received from c_3 and c_4 and made to c_5 and c_6, relations which are represented with asymmetric dyads. Calls are both received from and made to c_7 and c_8, relations which are represented with reciprocal dyads.

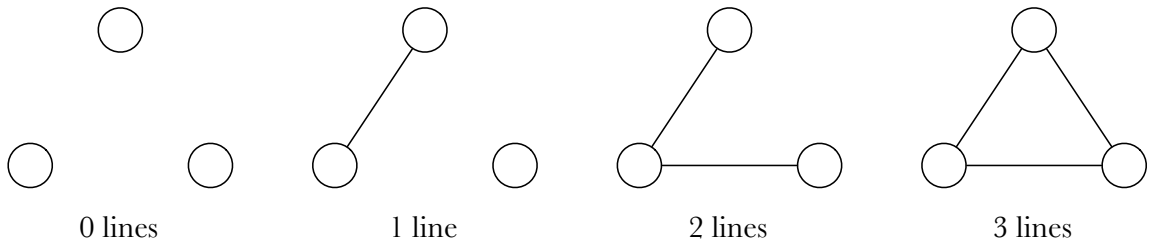
Figure 3.5.
A Directed Graph Showing the Calls Between a Subscriber and Contacts.



3.2.4. Triads and Ego Graphs

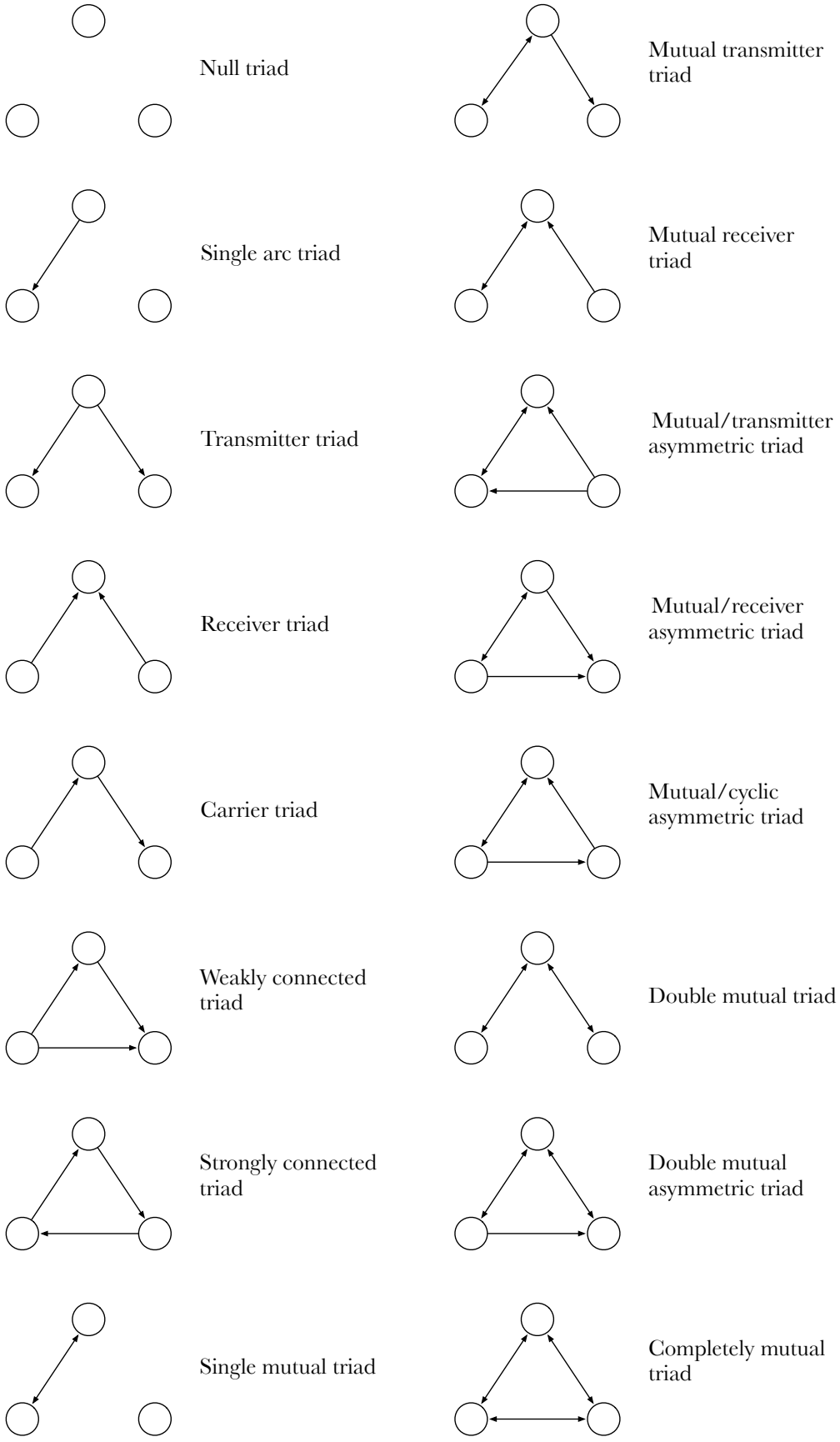
A graph consisting of three nodes and the possible lines between them is called a *triad*. A triad may be in one of four possible states, depending on whether none, one, two, or three lines are present among the nodes of the triad. These four possible triadic states are shown in Figure 3.6 [58].

Figure 3.6.
The Four Possible Triad States. (**Adapted from [58].**)



Triads may also exist in digraphs. Consider any three nodes of a digraph n_i, n_j, and n_k, where $i \neq j \neq k$. The set of three nodes without the lines which may exist

Figure 3.7.
The Sixteen Digraph Triad States.

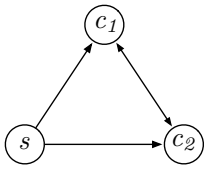


between them is called a *triple*, when the lines which join the nodes in the triple are considered there is a triad. T_{ijk} is the triad involving n_i, n_j, and n_k. In a triad the ordering of the nodes matters so we will always let $i < j < k$ [60].

For a set of N actors, there are $\binom{N}{3}$ triads. Let $\mathscr{T}$ denote the set of all triads: $\mathscr{T} = \{T_{123}, T_{124}, \ldots, T_{(N-2),(N-1),N}\}$. This set is of size $\binom{N}{3} = (1/6)N(N-1)(N-2)$ [60].

Consider now how many ties can be present in a digraph triad. There are three nodes in the triad, and each node can join to two others, giving six possible arcs. There are $2^6 = 64$ states for a triad if the digraph is labeled. (If the digraph is not labeled then some of the states will be isomorphic—structurally indistinguishable from one and other—in the same way that unlabelled asymmetric dyad states are structurally indistinguishable [60].) The sixteen (unlabelled) digraph triad states are shown in Figure 3.7.

Figure 3.8.
A Triad Showing Calls Between a Subscriber and Two Contacts.



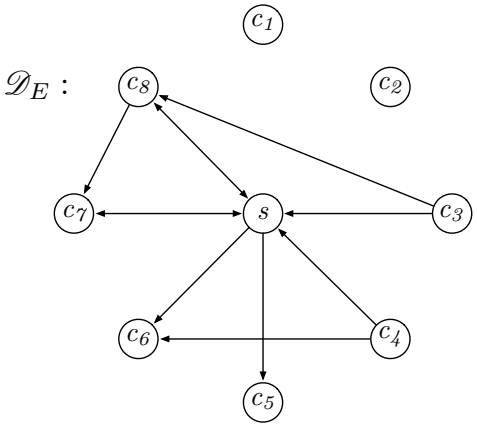
If a mobile subscriber *s* has made calls to two contacts c_1 and c_2 the relations between *s* and c_1 and c_2 create a transmitter triad—a special case of the star graph with only three nodes. If, however c_1 then calls c_2 the triad becomes a weakly connected triad like that shown in Figure 3.8. By analysing the relations between contacts in addition to the relations between the contacts and the subscriber shown in the star graphs of Figures 3.4 and 3.5 we create another social graph topology: the *ego graph*.

An ego graph consists of a focal node, termed *ego*, a set of nodes who have some relation to ego, termed *alters*, and the relations between the alters [61]. Ego graphs are used by anthropologists to study the social environment surrounding individuals and families, and are also used to study 'social support' relationships which aid the health or well-being of an individual [61].

Figure 3.9 shows an ego graph $\mathscr{D}_E$ of the calls between a subscriber *s* and eight contacts c_1 to c_8. All of the arcs present in $\mathscr{D}_C$ (shown in Figure 3.5) are present in $\mathscr{D}_E$ as well as the additional arcs $\langle c_3, c_8 \rangle$, $\langle c_4, c_6 \rangle$, $\langle c_8, c_7 \rangle$ between the alters.

When using data from mobile devices to create ego graphs more than one data source is needed: it is longer possible to use the address book or call logs from a single

Figure 3.9.
An Ego Graph Showing the Calls Between a Subscriber and Contacts.



mobile device as these sources of data will not contain any information about the alters' contacts or phone calls. To construct a complete social graph it is therefore necessary to gather data from more than one mobile device.

3.3. Mapping Communication Networks

The data required to map social graphs is communication network data. This section discusses what can be achieved by analysing communications, and how this data can be gathered on real mobile devices and combined to create a communications network graph. A method for determining the veracity of the gathered communications network data is also discussed.

Traffic Analysis

In *Secrets and Lies: Digital Security in a Networked World* Bruce Schneier defines traffic analysis as

> the study of communications patterns. Not the content of the messages themselves but, characteristics about them. Who communicates with whom? When? How long are the messages? How quickly are the replies sent, and how long are they? What kinds of communication happen after a certain message is received? These are all traffic analysis questions and their answers can reveal a lot of information.

In the intelligence community the practice of studying the characteristics of messages (but not their content) to infer patterns in communications is often

used [62]. A similar approach is used in this work. No analysis of the content of messages is considered—neither in voice or text-based communications—instead only characteristics of the messages, in the form of communications metadata, is used.

Communications Metadata

Metadata is structured data which describes the characteristics of a resource. In the case of a call made from a mobile device it might include the number the call is from, the number the call is made to, the direction of the call, the time the call is made, and the duration of the call. Text-based communications such as SMS messages or email will have similar metadata although there will be no duration as these communications are not real-time, instead a length measured in bytes or the number of characters could be measured. Analysis of communications metadata of this kind allows traffic analysis to be performed on communications devices.

Pervasive Computing and Context

Pervasive computing aims to enhance computer use by making many computers available throughout the physical environment which are effectively invisible to the user [63], and envisions a world of fully interconnected wireless devices, with cheap wireless networks everywhere.

Today, some aspects of communication infrastructure are ubiquitous: there is appropriate bandwidth for large-scale wireless networking in large parts of the developed world [64], the appropriate communications protocols to handle mobility have seen large-scale deployment [65], and the number of people adopting wirelessly enabled mobile devices is increasing quickly [65].

Pervasive computing systems take advantage of environmental information, or context, to enhance the interaction with the user. Context is usually defined as information that is part of the operating environment of a system that can be sensed by the system. This can include the location, identity, activity, and state of people, groups and objects as well as information relating to places or the computing environment [66].

Communications metadata can be considered to be context data

Mobile devices which gather data on users' social interactions require context information pertaining to the nature and state of users' social relationships. Communications metadata provides 'social' context because of the inherent link between communications and social ties.

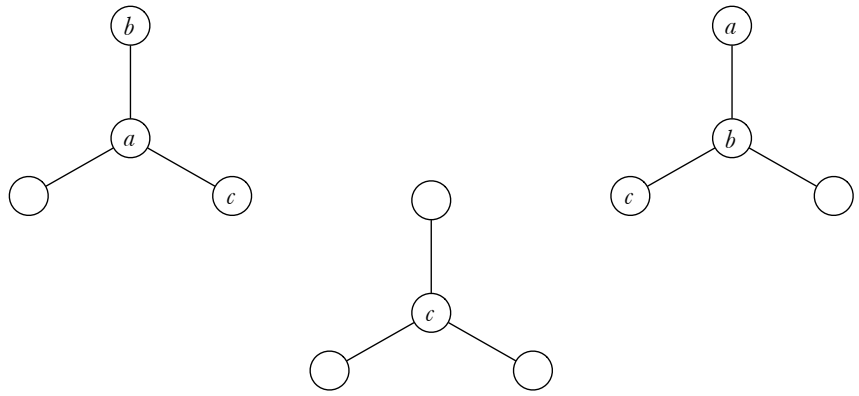
The following communications channels are available on most mobile devices: voice and video calls, and SMS and MMS messages via cellular networks; and

communications via standard internet protocols such as email, instant messaging, VoIP calls, file transfers, and resource sharing via WiFi and 3G mobile networks. Each communication attempt, whether or not it was successful or reciprocated, and information about the timing of communications such as the frequency, the time elapsed since the last communication, or the duration of any conversations could all be gathered and used as 'social' context data.

3.3.1. System Architecture

Communication Metadata is used to map the communications between mobile users. If the relation to be map is *has called* then by analysing the call logs on a mobile device it is possible to create star graphs like those discussed in Section 3.2. Three star graphs for three subscribers *a*, *b*, and *c* are shown in Figure 3.10. Each subscriber has called three other people. Both *a* and *b* have called each other, and they have also called another person—subscriber *c*.

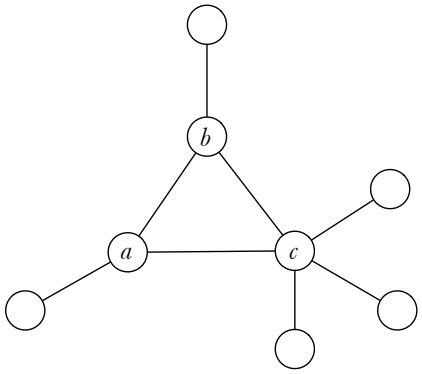
Figure 3.10.
Star Graphs for Three Mobile Subscribers.



It is clear that there is some redundancy in representing the relations between *a*, *b*, and *c* using only star graphs. The line (a, b) is shown twice as are the lines (a, c) and (b, c). Each of the star graphs can be combined to make a complete network graph for the three subscribers. The complete communications network graph is shown in Figure 3.11.

There are two possible architectures for a system which is able map to mobile communications networks in this way. Either a centralised architecture is used, where each star graph is sent to a central point (a dedicated server for example) to be combined into a complete network graph in a similar way to the Home Location Register of a cellular network: a given device knows its own location but the HLR

Figure 3.11.
Complete Communications Network Graph for Three Subscribers.



knows the location of all devices on that network. Or a distributed architecture is used, where each star graph is shared between devices in some way and devices create a local version of the complete network graph in a similar way to the sharing of information in mobile ad-hoc networks [67].

In this work only centralised networks are considered as the focus is on the problem of detecting social ties between the users of mobile devices. The problem of efficiently sharing data across a distributed network is a second order issue and not considered here.

3.3.2. A Note on Gathering Social Context

Some of the nodes shown in Figures 3.10 and 3.11 are not labeled. The identity of these nodes is not known because they have no context gathering software running on their devices. Any relation exists between two actors and hence requires two actors in order to exist. The two ends of a social tie may be the sender and recipient of an SMS message, for example.

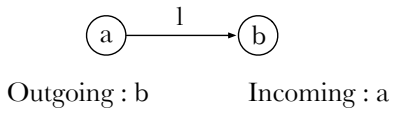
When one attempts to gather communications metadata from mobile devices two distinct classes of data emerge: data pertaining to interactions between users whose mobile devices are gathering context data, and data pertaining to interactions between users of mobile devices and others. The others may be users of mobile devices with whom social ties exist, but they may also be automated services, PTSN phones in homes or offices, or even other devices belonging to the same person. Without context aware software running on a device, and hence some knowledge of that device, it is difficult to make inferences about any interactions with that device. For this reason, only data pertaining to interactions between users whose mobile

devices are gathering context data is of interest in this work.

3.3.3. Data Corroboration

When performing traffic analysis on a centralised dataset of context gathered by devices each mediated interaction made between two mobile devices will be recorded twice: once on the sending device and once on the receiving device. This means that for every call or SMS shown on the communication network graph there are two corroborating records. The valued digraph in Figure 3.12 represents a single call between two subscribers *a* and *b*. As *a* made the call to *b* it is recorded as an outgoing call to *b* on device *a*, and similarly as an incoming call from *a* on device *b*.

Figure 3.12.
Two Corroborating Call Records Between *a* and *b*.



By analysing the corroborating pairs of calls it is possible to determine the reliability of the communications data gathered. If a mediated interaction recorded on the sending device is corroborated on the receiving device then we can be sure that the data is correct. However, if a mediated interaction recorded on the sending device is not corroborated on the receiving device (or vice versa) then the accuracy of the data gathered is brought into question. Clearly some data is missing, but it is impossible to know how much. The amount of corroborating data is used to give an indication of the completeness of the dataset: the higher the corroboration, the more complete the dataset.

3.4. The Reality Mining Communication Graph

The Reality Mining study [10], discussed in Chapter 2 during the review of systems capable of detecting human co-proximity, also gathered some communication metadata from the mobile phones used in the study. This section gives more detail of the communications metadata gathered, discusses the communications network graph derived from this communications information, and investigates the veracity of the data gathered.

3.4.1. The Dataset

The Reality Mining dataset includes data pertaining to study participants' calls, SMS messages, and cell tower data, as well as Bluetooth device discovery. Details of what data was gathered are given below, and further details such as the database schema used can be found in Michael Lambert's Masters thesis [68].

The cell phones used in the study recorded 'interesting' events to log files. The timestamps used when recording these events had the following format: `YYYYMMDDTHHMMSS`. (The character T denotes time and acts as a delimiter between date and time information in self-explanatory formats.) When the phones were on, three ASCII-formatted logs were created using the current timestamp as part of the filename `starter-TIMESTAMP.txt`, `call-TIMESTAMP.txt`, and `log-TIMESTAMP.txt`.

`starter-TIMESTAMP.txt` contains status and debug information about the logging application, but information about the participants. `call-TIMESTAMP.txt` contains a log of phone calls, SMS messages, and data transfers including incoming/outgoing status, the number dialled, the duration of the call, and other information. `log-TIMESTAMP.txt` contains a log of other device context information: current cell tower, nearby Bluetooth devices, current foreground application, idle/active status, and charging status. This data was collated and stored in a MySQL database.

The database schema refers to the data stored in `call-TIMESTAMP.txt` files as *callspans*. The following call and SMS message data is available:

Voice Calls

- Start time,
- End time,
- ID of the *person* whose phone is recording the call,
- ID of the *interface* at the other end of the call (i.e the phonenumber),
- Direction with respect to the device logging the call, and
- Call duration.

Short Messages

- Start time,
- End time,
- ID of the *person* whose phone is recording the message,
- ID of the *interface* at the other end of the message (i.e the phonenumber),
- Direction with respect to the device logging the call, and
- Status of message: sent or delivered.

The data contains 897,921 call and SMS records, 39,206 (4.4%) of which are between the ninety-seven study participants. There are 599,097 calls, 108,693 SMS messages, and 190,131 data transfers. Data transfers are ignored in the remainder of this analysis as none were detected within the group of study participants.

34,742 calls (5.8%) and 4,464 messages (4.1%) are between the study participants. (All values for calls and SMSs within the study group do not include self-calls—calls a participant made to themselves—some of the total number will include self-calls.)

Many of the call and SMS records in the dataset are duplicated: manual inspection of the data found that many of the call and short message records have the same person (recorder of the interaction), phone number (opposite end), start time, end time, and duration. In order to continue the analysis, duplicate calls were removed from the dataset by the author of this thesis.

After removing duplicate records 4,240 unique voice call records and 657 short message records between study participants remain. Therefore 87.8% of voice call records and 98.11% of short messages between Reality Mining participants in the original dataset are duplicates.

3.4.2. Reality Mining Communication Graphs

The unique call and SMS records in the Reality Mining dataset can be used to create graphs of the communications between the participants in the Reality Mining study. Digraphs showing the calls between study participants and SMS messages between study participants are shown in Figures 3.13 and 3.14 respectively. Each arc on the digraph represents one or more call or SMS message between participants, with the arc directed towards the receiver of the call or the message. Study participants who sent or received voice calls or messages are represented by the nodes on the graph, study participants who did not send or receive any calls or messages are not represented. The nodes are numbered according to the `person_id` numbering scheme in the dataset.

The call network graph shown in Figure 3.13 indicates calls were made between seventy-seven of the ninety-seven study participants. Similar to the analysis of the proximity interactions between study participants performed by Eagle [2], two main components are clearly visible although four unconnected dyad pairs can also be seen. The degree of most nodes is in the order of 1–3 with some nodes, such as 29, 83, and 86, having a much higher degree.

The SMS network graph shown in Figure 3.14 is considerably more sparse than the call graph. (The call graph is shown greyed out in the background for reference.) The smaller of the two components breaks into two further components, one of

which is a transmitter triad and the larger of the two components into three smaller components, two of which are dyad pairs. Most arcs are similar if not the same as those in the call network graph, although some only appear on one. The arcs $\langle 63, 97 \rangle$ and $\langle 63, 87 \rangle$ appear in both graphs, $\langle 29, 86 \rangle$ and $\langle 86, 29 \rangle$ are on the call graph but only $\langle 86, 29 \rangle$ is on the message graph. The edge $\langle 75, 73 \rangle$ appears on the call network graph but not the message network graph, and the edge $\langle 84, 75 \rangle$ appears on the message network graph but not the call network graph.

3.4.3. Corroboration in the Reality Mining Dataset

The unique call and SMS records in the Reality Mining dataset allow the communications between study participants to be mapped. However, although each of these records is unique they have not been corroborated as described in Section 3.3.3.

When two members of the Reality Mining study exchange a mediated interaction (sending or receiving a message or phone call), there should be two records of this event: a log of an outgoing communication from the sender, and a log of an incoming communications from the receiver. The Reality Mining dataset does not exhibit this expected corroboration to any significant degree.

Every (unique) call and SMS message record in the Reality Mining dataset was tested for a corroborating record. To corroborate one record, another record must be found where the sender and receiver matched and vice versa, the start time is ± 30 seconds (to allow for clock de-synchronisation), and the duration of a call is within ± 5 seconds.

Of the 4,240 unique call records in the data set 426 (10.05%) are corroborated meaning that 213 calls are reliably recorded. Similarly, 36 (5.48%) of the 657 unique SMS message records are corroborated meaning that only 18 records are reliably recorded.

To account for larger values of clock de-synchronisation, the corroboration test was run for a window size of up to 10 minutes. The results are shown in Figure 3.15.

Allowing for clock de-synchronisation of ten minutes increases the number of corroborated calls to 15.5% and the number of corroborated messages to 11.57%. These results are significant: the communications metadata in the Reality Mining data set cannot be considered reliable, and as such, it is not suitable for use as the basis for an investigation into inferring social ties from communication patterns. A better dataset is required.

Figure 3.13.
Reality Mining Call Network Graph.

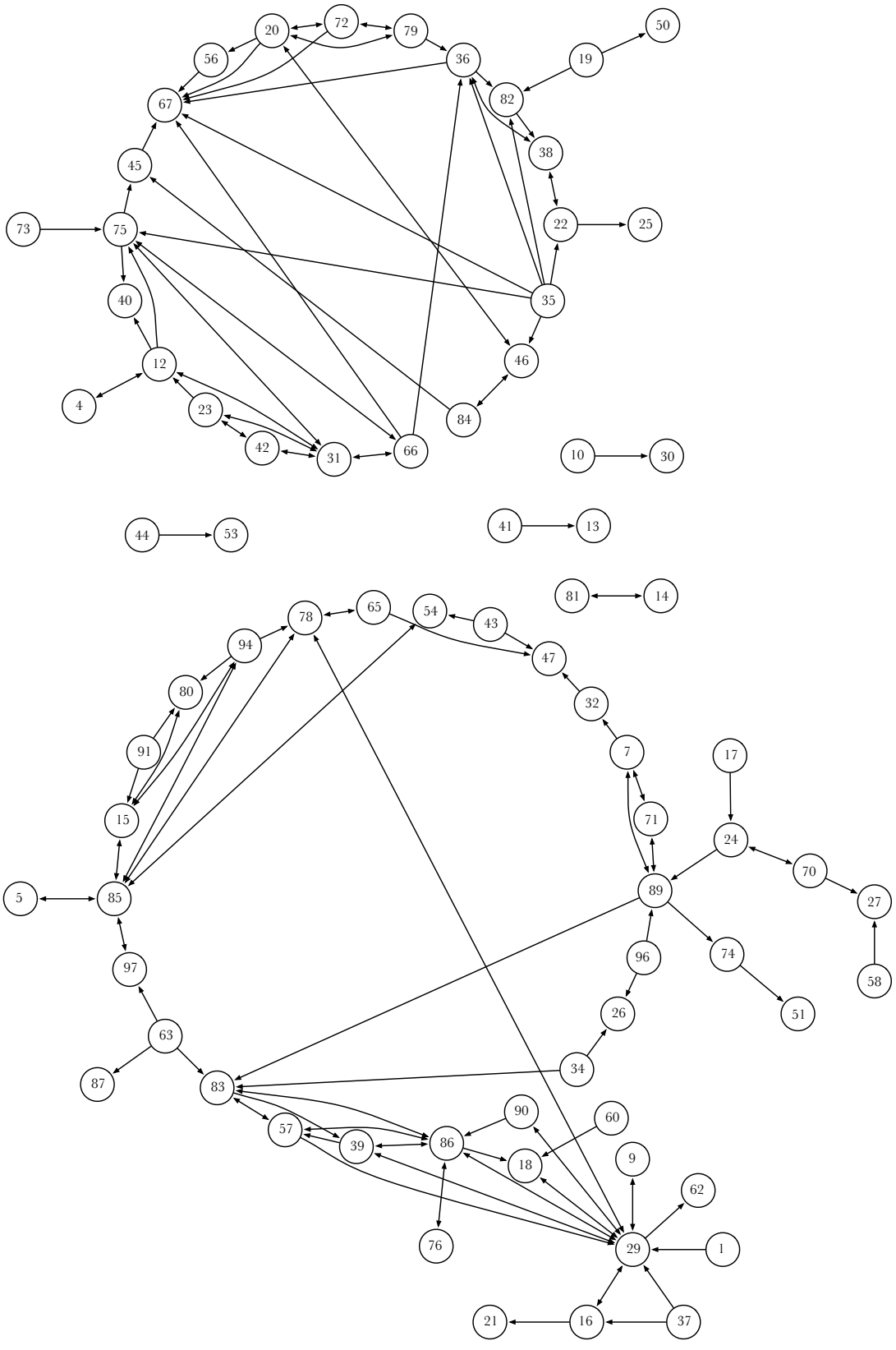


Figure 3.14.
Reality Mining SMS Network Graph.

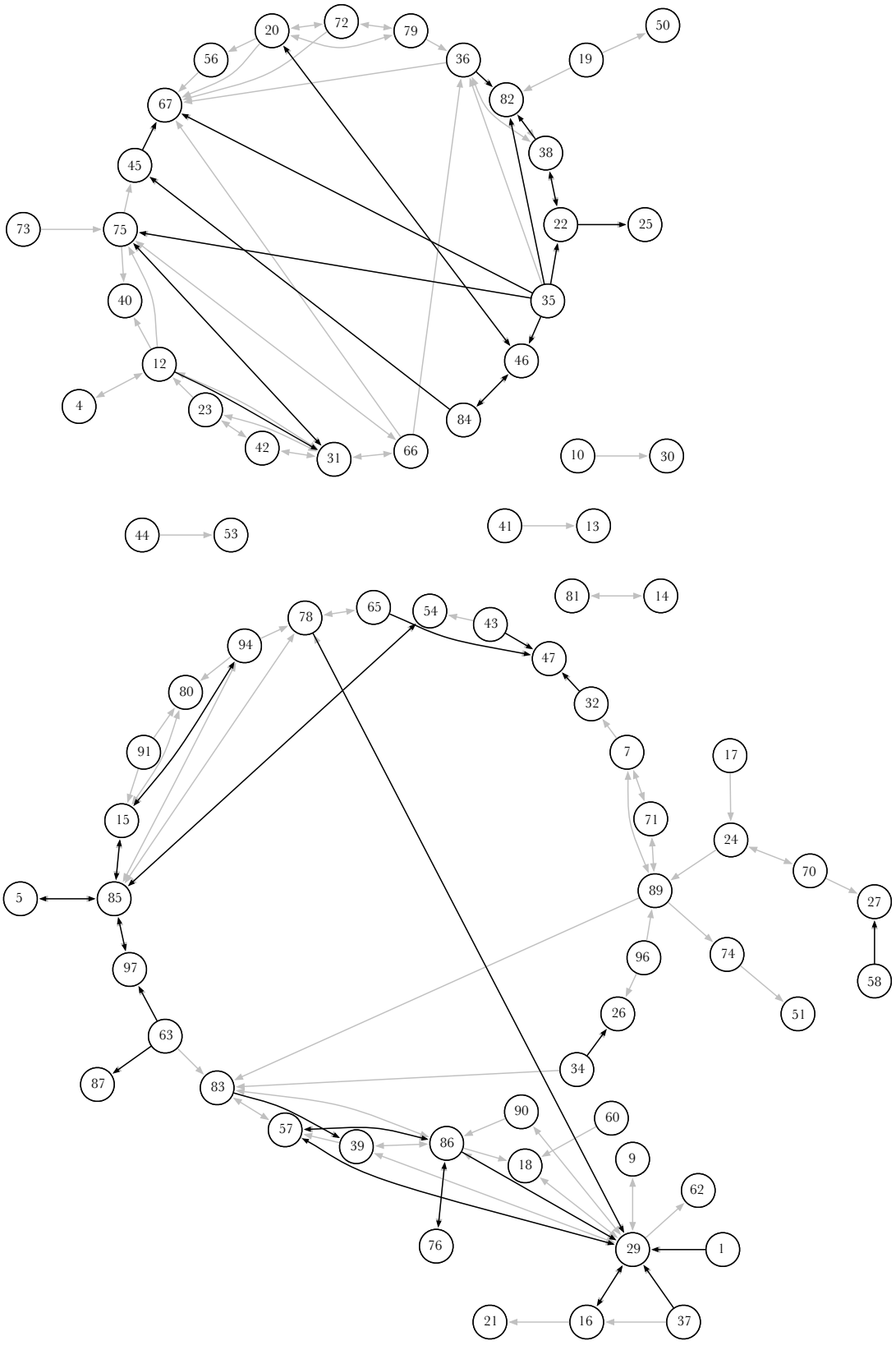
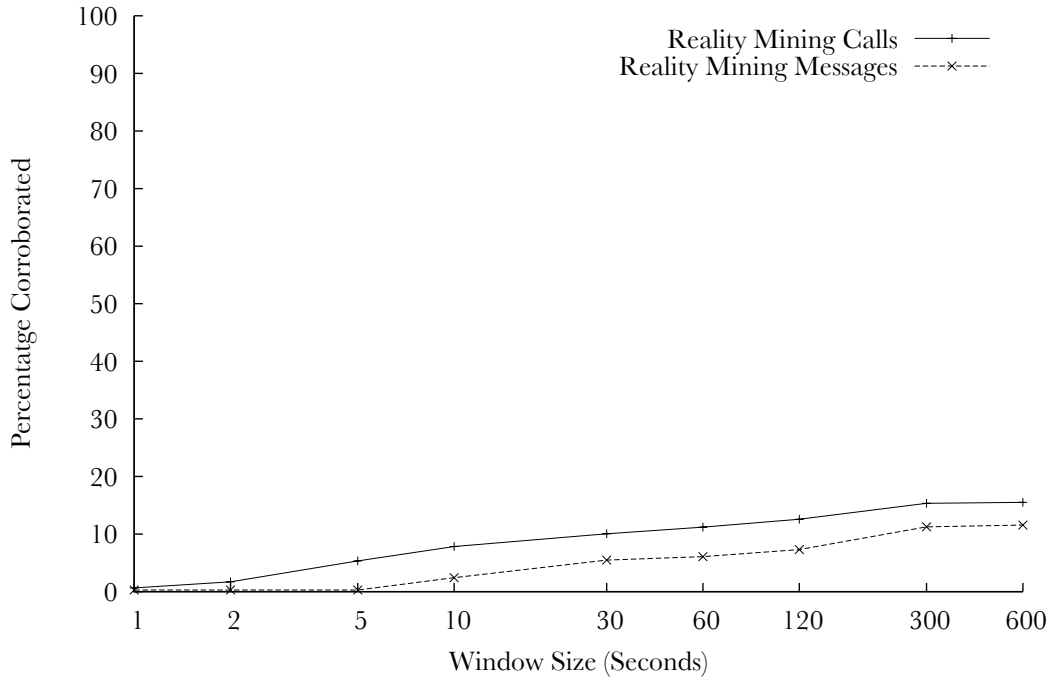


Figure 3.15.
Percentage of Corroborated Calls and Messages in the Reality Mining Dataset.



3.5. The Nodobo Dataset

The communications metadata in the Reality Mining dataset is so sparse it is not suitable for use as the basis for an investigation into inferring social ties from communication patterns. As no other suitable datasets were available, a project to gather another dataset was undertaken. The Nodobo project gathered communications metadata from a group of twenty-seven students at Springburn Academy in Glasgow. The resulting dataset is freely available online[2].

3.5.1. Background

The Nodobo dataset was gathered during a University of Strathclyde funded project entitled 'Use of Social Networks to Support Education' undertaken with colleagues in the Applied Educational Research Centre, University of Strathclyde. The project explored how social networks and mobile devices are being used by secondary school

[2] http://uk.crawdad.org/meta.php?name=strath/nodobo

students to enhance their learning experience.

A group of three researchers from the Applied Educational Research Centre, led by Alistair Wilson, worked with the students at the school and gathered qualitative study data. At the same time a group of three researchers, including the author, led by Dr Alisdair McDiarmid developed a system to gather social interaction data using smartphones. This allowed traffic analysis to be performed on the voice calls and SMS messages sent and received by the study participants, and from this traffic analysis a social graph was derived.

Study participants were all in the fifth year of high school (the second last in the Scottish system) and were part of a group of discerning students identified by the school as having a high chance of gaining university admission. Students were required to attend school from Monday to Friday between 08:30 and 15:45 and encouraged to attend optional study groups after school on some days. Not all students attended all of the same classes, but there was some overlap with some study participants attending some of the same classes.

Ethical Issues

Due to the sensitive nature of the personal information gathered during the study, ethical consent was sought from the University Ethics Committee. Ethical approval for the study was given based on the following conditions:

- All participants were required to provide informed consent to take part in the research.
- To guarantee informed consent there was an opportunity for the parents/carers and pupils interested in becoming involved to attend an information session with the research team before the research commenced. In addition the research team distributed detailed information sheets to each pupil and parent/carer.
- Only those pupils that attended the information sessions were invited to take part in the research project.
- Consent forms were given to students, parents (for information), and teachers. As all of the students are over the age of 12, parental consent was not needed, however their agreement was sought in addition to consent from the student.
- A researcher had ongoing contact with all participants ensuring they have an opportunity to withdraw from the project at any time.
- During the course of the project pupils were represented by allocated ID numbers and there was no display of their names or numbers to other pupils.

- Where participants contact other mobile devices, outside of the research group, one way data encryption meant that all external numbers were encrypted and not visible or knowable to the research team.
- During the course of the project all digital data were held on a secure University server. Other forms of data including completed consent forms were stored in locked university accommodation.
- Data was maintained in an anonymised state beyond the duration of the project with access restricted to the members of the research team, all of whom had Enhanced Disclosure. Data gathered was only used for the stated purposes of the research.

System Development

The software used to gather data in the Nodobo study was developed by Stephen Bell, Dr Alisdair McDiarmid, and the author. It is based on existing software developed by Stephen Bell and Alisdair McDiarmid which allows precise capture and replay of smartphone user interactions sessions to enable new usability testing experiments [69, 70]. This software gathered context data pertaining to users' interactions with smartphones—such as touching the touch-screen or pressing buttons—and was used to examine how smartphone applications or mobile websites are used in the real world.

Working from requirements elicitation performed by the author, Stephen Bell was able to augment this existing software to gather communications metadata as well as user interactions context data. The resulting software was then tested by the author and Stephen Bell before being deployed.

3.5.2. Data Collection

Study Participants were given a handset with some modifications made to the operating system. The Nodobo Study gathered data using Google Nexus One handsets running Android 2.3.

The Nexus One was selected for two reasons: it is open enough to allow the appropriate modifications to the operating system for capturing the social interactions metadata to be made, and the phone is powerful enough to replace all of the functionality of the users' existing phone meaning that they will use it in place of their old device. (Although at the time of writing the Nexus One is becoming a little outdated and a new, more powerful device will be required in the future.)

Nodobo Social is a set of software extensions to the Android operating system, which monitor the applications running on a mobile device to record a variety

of communications and usage context data. The records are stored in a SQLite3 database on the device's SD card, which is synchronised periodically over the air with a web services data store, ensuring that data is continuously received from devices throughout a deployment, broken or out of contact devices can be detected quickly, and the SD card on the device acts as a back up of the data gathered in a deployment rather than the only data store [71].

The software on the phone captures data using a variety of software sensors, logging phone calls and text messages, Bluetooth device discovery, WiFi access point, and cell tower ID. The direction of calls and text messages is recorded, along with the associated phone number, and the duration of the call or length of the message. Bluetooth proximity is recorded every minute, and includes all phones in the study as well as any other devices which respond to service discovery. Basic positioning is achieved through WiFi hotspot and cell tower ID records.

From September 2010 to the end of January 2011 the Nodobo study recorded 13,035 call records, 83,542 SMS records, and 5,292,103 proximity records. There were 1,309 calls within the study, 25,982 messages within the study, and no duplicate records.

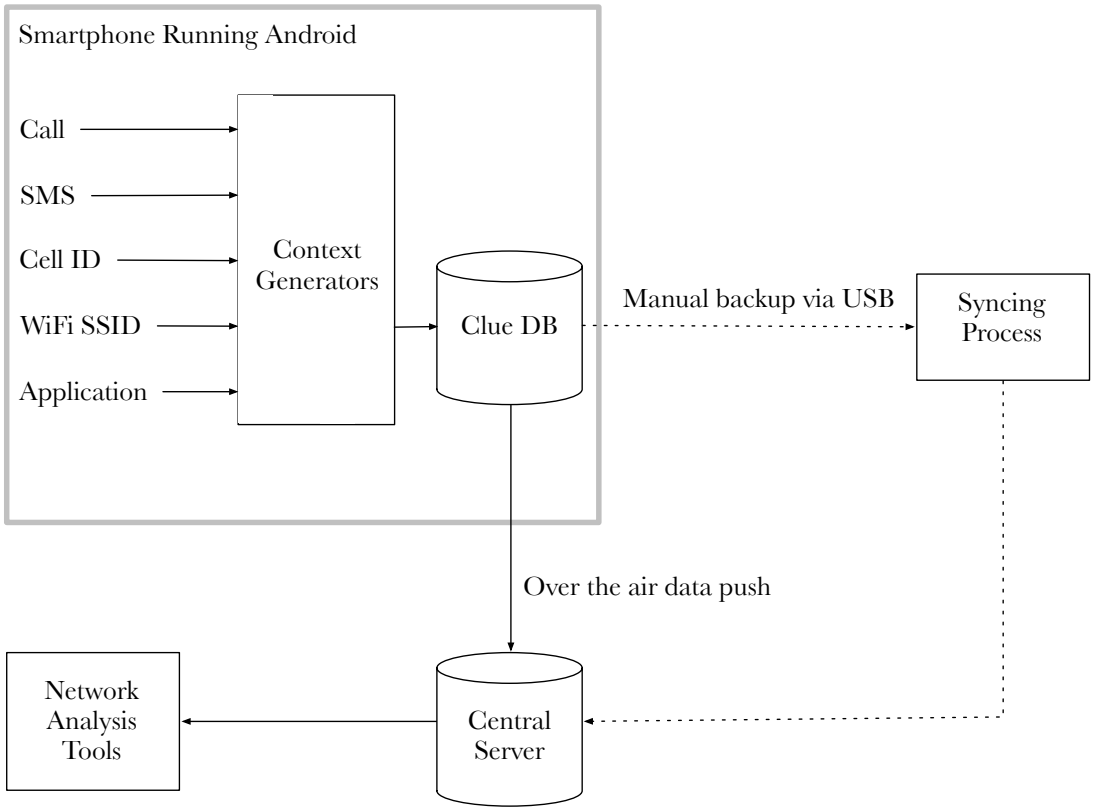
An overview of the Nodobo Social architecture is shown in Figure 3.16. Context generators create new context records or *clues* which are stored on the devices in a clue database. The contents of the clue database are pushed over the air to a central server, and are then available for analysis. If data is for some reason not being sent to the central server the option of manually collecting the contents of the clue database via USB is available.

Android stores past calls and SMS messages in SQLite3 databases called `contacts.db` and `mmssms.db` respectively. Objects called 'content providers' and 'content observers' provide a mechanism for interacting with these databases; specifically, an application can register to receive notifications when the contents of these databases change. By registering a content observer for each of the databases, it is possible to record when calls and SMS messages are sent and received.

The simplest mechanism for synchronising the call log, or the SMS log, with the interactions database is to insert the most recent call, or SMS, in the respective database into the interactions database directly after a notification is received. However, this approach assumes a single notification for each interaction which is not the case. There are three notifications sent for calls, and two notifications sent for SMS messages, meaning that each call or SMS message would be inserted into the interactions database multiple times [71].

Instead, when a call or SMS notification is received, the handler selects the

Figure 3.16.
Nodobo Social Architecture.



original ID of the latest call in the interactions database. If this original ID is less than the ID of the newest entry, the newest entry can be inserted into the interactions database only when the notification is called the first time [71].

Bluetooth is used to detect nearby co-proximate devices. Nodobo uses service discovery to discover devices for 12 seconds each minute. As the Bluetooth modem is unavailable when the device is asleep, the application must wake the device in order to discover devices. Every minute an alarm is registered which wakes the device and registers a wake lock so the processor will not go back to sleep once the discovery process is started [71].

3.5.3. Initial Examinations of the Nodobo data

Analysis of the timing of interactions provides an initial common sense check of the data gathered during the Nodobo study. The timestamps of the interaction data gathered during the study were filtered by hour and by day of the week, and

the cumulative percentages were then plotted. They show broadly what would be expected—most interactions occur during school hours on weekdays, with some interactions on weekday evenings and weekends—but there are also some interesting features in the data.

Figure 3.17.
Percentage of Detected Nodobo Interactions by Day.

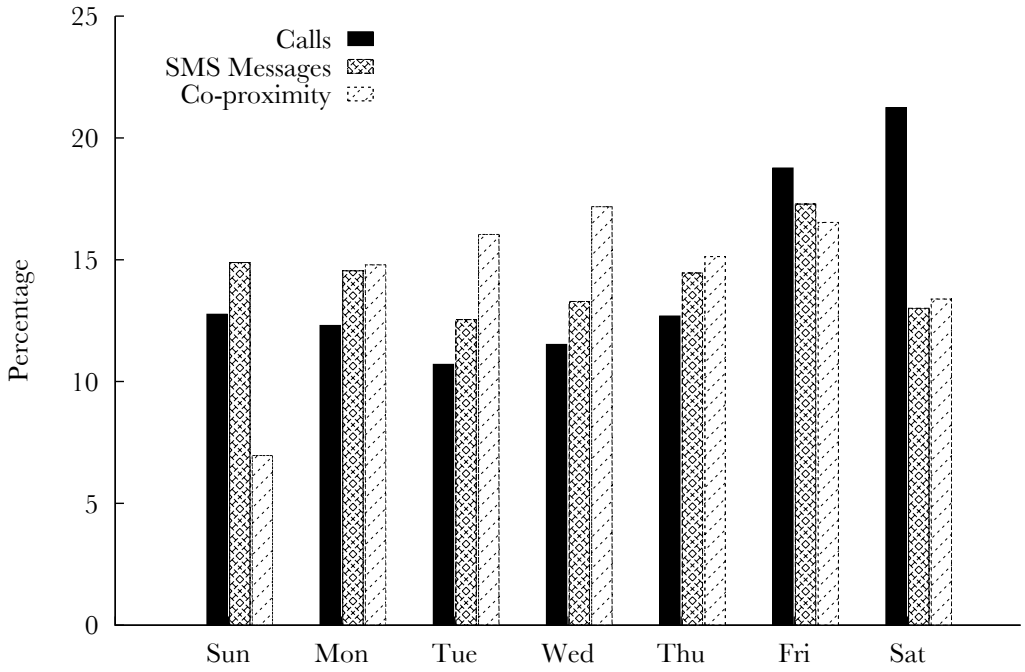
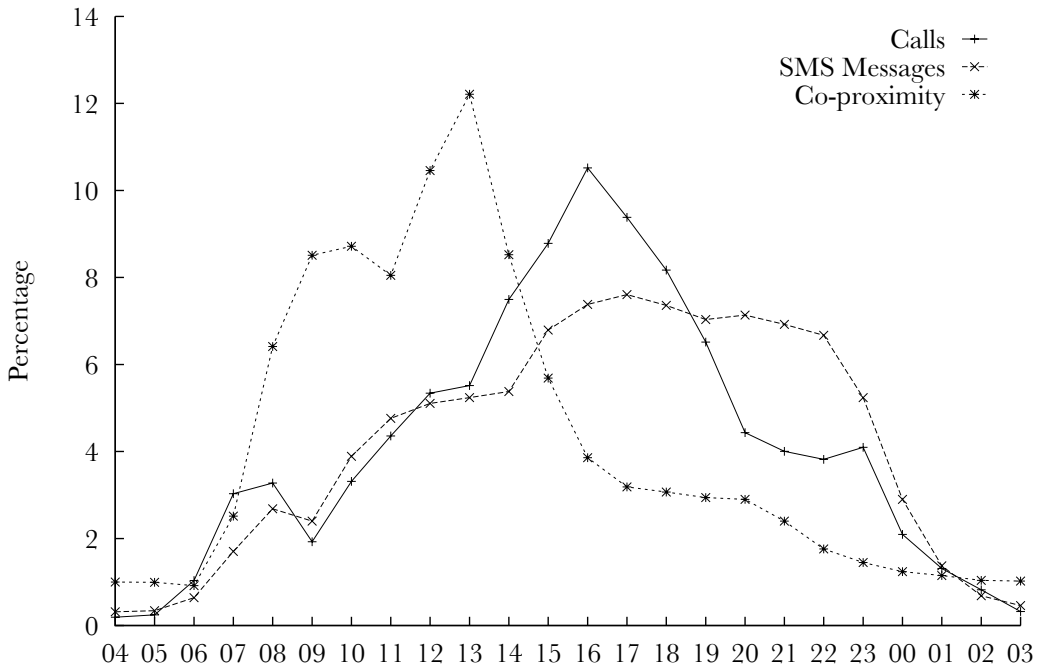


Figure 3.17 shows shows the change in the number of calls, SMS messages, and co-proximate interactions, for study participants, over the course of each week in the study. More calls are made on Friday and Saturday compared to the rest of the week, SMS messages are fairly evenly distributed throughout the week, and co-proximity is detected almost always between Monday and Friday when the participants should have been attending school.

Figure 3.18 shows the change in the number of calls, SMS messages, and co-proximate interactions, for study participants, over the course of each day of the study. Calls peak in the morning before school, and in the evening after school, although there are some calls being made throughout the day hours and the evening. SMS messages increase throughout the day and peak in the evening. Interestingly, SMSs are sent late into the night and early hours of the next morning. Co-proximity

is detected almost exclusively between 0700 hours and 1600 hours although there is a long tale suggesting a small amount of co-proximity in the evening. Significant peaks in the amount of co-proximate interactions are seen at 0900 hours and 1300 hours during the first class of the day and the lunch break respectively.

Figure 3.18.
Percentage of Detected Nodobo Interactions by Hour (Adapted from [71]).



Study participants appear to attend school as expected on weekdays: they arrive each day around 9:00am, spend their lunch break together, and leave school in the late afternoon. There is some co-proximity on Saturdays, perhaps for sporting or social events, and very little co-proximity on Sundays. SMS seems to be the common mode of communication, and is used throughout the week including during school hours and late into the night. Voice calls on the other hand are made more often on Friday and Saturday than other days, perhaps when organising weekend activities. Voice calls are also made less frequently during school hours than SMS messages but do show spikes before and after the school day, this is presumably because they harder to conceal than SMS messages when in class!

3.5.4. Nodobo Communication Graphs

The call and SMS records in the Nodobo dataset can be used to create graphs of the communications between the study participants like those of the Reality Mining Study in Section 3.4.2. Again, each arc on the digraph represents one or more voice call or SMS message between participants, with the arc directed towards the receiver of the call or the message. Study participants who sent or received voice calls or messages are represented by the nodes on the graph, study participants who did not send or receive any calls or messages are not represented. The nodes are numbered according to the `user_id` numbering scheme in the dataset.

The Nodobo SMS message network graph $\mathscr{D}_{message}$ consists of a set of twenty-five nodes $\mathscr{N}_{message}$ and a set of eighty-four arcs $\mathscr{L}_{message}$ where

$$\begin{aligned}\mathscr{N}_{message} = \{&1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22,\\ &23, 24, 25, 26, 27\}\end{aligned}$$

and

$$\begin{aligned}\mathscr{L}_{message} = \{&\langle 1,6\rangle, \langle 1,8\rangle, \langle 1,11\rangle, \langle 1,15\rangle, \langle 1,27\rangle, \langle 3,11\rangle, \langle 3,27\rangle, \langle 4,7\rangle,\\ &\langle 4,13\rangle, \langle 4,17\rangle, \langle 4,23\rangle, \langle 5,24\rangle, \langle 6,1\rangle, \langle 6,11\rangle, \langle 6,15\rangle, \langle 6,27\rangle,\\ &\langle 7,4\rangle, \langle 7,13\rangle, \langle 7,23\rangle, \langle 7,27\rangle, \langle 8,10\rangle, \langle 8,13\rangle, \langle 8,23\rangle, \langle 8,27\rangle,\\ &\langle 9,22\rangle, \langle 10,8\rangle, \langle 11,1\rangle, \langle 11,3\rangle, \langle 11,4\rangle, \langle 11,6\rangle, \langle 11,8\rangle,\\ &\langle 11,15\rangle, \langle 11,18\rangle, \langle 11,23\rangle, \langle 11,26\rangle, \langle 11,27\rangle, \langle 13,4\rangle, \langle 13,7\rangle,\\ &\langle 13,8\rangle, \langle 13,12\rangle, \langle 13,23\rangle, \langle 13,26\rangle, \langle 13,27\rangle, \langle 14,19\rangle, \langle 14,21\rangle,\\ &\langle 14,25\rangle, \langle 15,1\rangle, \langle 15,6\rangle, \langle 15,8\rangle, \langle 15,11\rangle, \langle 15,27\rangle, \langle 16,19\rangle,\\ &\langle 17,4\rangle, \langle 17,6\rangle, \langle 17,11\rangle, \langle 17,12\rangle, \langle 17,18\rangle, \langle 18,11\rangle, \langle 18,17\rangle,\\ &\langle 19,14\rangle, \langle 19,16\rangle, \langle 19,21\rangle, \langle 19,25\rangle, \langle 21,14\rangle, \langle 21,19\rangle, \langle 21,25\rangle,\\ &\langle 22,9\rangle, \langle 23,8\rangle, \langle 23,13\rangle, \langle 25,14\rangle, \langle 25,21\rangle, \langle 26,11\rangle, \langle 26,13\rangle,\\ &\langle 26,23\rangle, \langle 26,27\rangle, \langle 27,1\rangle, \langle 27,3\rangle, \langle 27,6\rangle, \langle 27,7\rangle, \langle 27,8\rangle,\\ &\langle 27,11\rangle, \langle 27,13\rangle, \langle 27,23\rangle, \langle 27,26\rangle\}.\end{aligned}$$

The SMS network graph in Figure 3.19 shows messages were sent by twenty-five of the twenty-seven study participants. The graph has two distinct components, one with sixteen nodes and one with five. The remaining four nodes form the same disconnected dyad pairs as in the call network graph. The five node component contains the same nodes as the weakly connected sub-component seen in the call network graph.

The Nodobo call network graph $\mathscr{D}_{calls}$ consists of a set of twenty-three nodes $\mathscr{N}_{call}$ and a set of fifty-five arcs $\mathscr{L}_{calls}$ where

$$\mathscr{N}_{call} = \{1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 19, 21, 22, 23, 24, 25, 26, 27\}$$

and

$$\begin{aligned}\mathscr{L}_{calls} = \{&\langle 1, 15\rangle, \langle 3, 11\rangle, \langle 4, 7\rangle, \langle 4, 8\rangle, \langle 4, 23\rangle, \langle 5, 24\rangle, \langle 6, 1\rangle, \langle 6, 11\rangle, \\ &\langle 6, 15\rangle, \langle 7, 4\rangle, \langle 7, 13\rangle, \langle 7, 27\rangle, \langle 8, 10\rangle, \langle 8, 13\rangle, \langle 8, 23\rangle, \langle 8, 27\rangle, \\ &\langle 9, 22\rangle, \langle 11, 3\rangle, \langle 11, 15\rangle, \langle 11, 23\rangle, \langle 11, 26\rangle, \langle 11, 27\rangle, \langle 13, 4\rangle, \\ &\langle 13, 8\rangle, \langle 13, 23\rangle, \langle 13, 26\rangle, \langle 13, 27\rangle, \langle 14, 19\rangle, \langle 14, 25\rangle, \langle 15, 1\rangle, \\ &\langle 15, 6\rangle, \langle 16, 19\rangle, \langle 17, 4\rangle, \langle 17, 11\rangle, \langle 19, 14\rangle, \langle 19, 16\rangle, \langle 19, 21\rangle, \\ &\langle 19, 25\rangle, \langle 21, 19\rangle, \langle 21, 25\rangle, \langle 22, 9\rangle, \langle 23, 4\rangle, \langle 23, 7\rangle, \langle 23, 8\rangle, \\ &\langle 23, 13\rangle, \langle 25, 21\rangle, \langle 26, 11\rangle, \langle 26, 13\rangle, \langle 26, 25\rangle, \langle 26, 27\rangle, \langle 27, 7\rangle, \\ &\langle 27, 8\rangle, \langle 27, 11\rangle, \langle 27, 13\rangle, \langle 27, 26\rangle\}.\end{aligned}$$

The call network graph in Figure 3.20 shows the calls made between twenty-three of the twenty-seven study participants. The graph is dominated by a large component containing nineteen of the twenty-three nodes with the remaining four forming two disconnected dyad pairs. The main component of the graph also contains a weakly connected sub-component (nodes 14, 16, 19, 21, and 25) connected to the main component by the arc $\langle 26, 25\rangle$ alone.

The superset of all nodes in both the call and message network graphs contains twenty-five nodes

$$\mathscr{N}_{call} \cup \mathscr{N}_{message} = \{1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27\},$$

and twenty-three nodes are common to both graphs

$$\mathscr{N}_{call} \cap \mathscr{N}_{message} = \{1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 19, 21, 22, 23, 24, 25, 26, 27\}.$$

There are no nodes which appear on the call network graph but not the message

Figure 3.19.
Nodobo Message Network Graph.

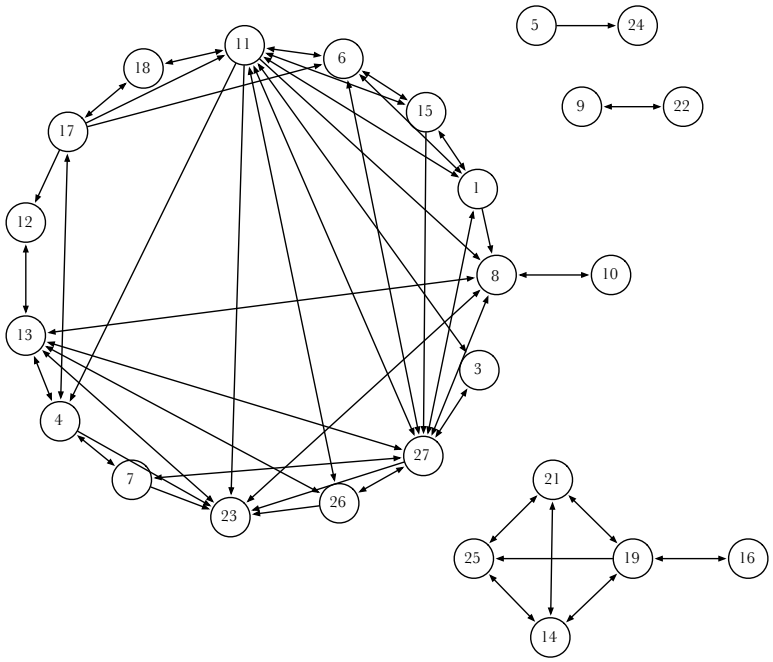
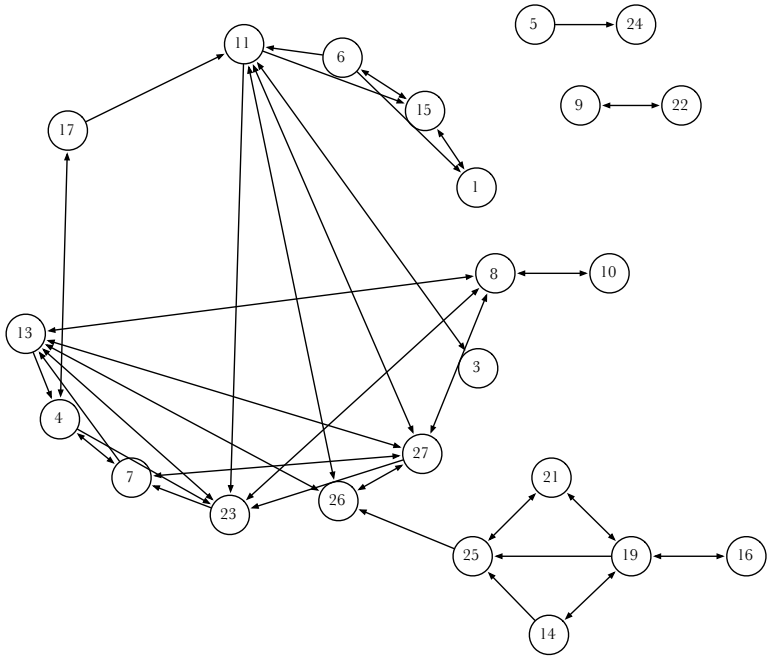


Figure 3.20.
Nodobo Call Network Graph.



network graph

$$\mathscr{N}_{call} \setminus \mathscr{N}_{message} = \emptyset,$$

and there are two nodes which appear on the message network graph but not the call network graph

$$\mathscr{N}_{message} \setminus \mathscr{N}_{call} = \{12, 18\}.$$

The superset of all arcs in both the call and message network graphs contains eighty-eight arcs

$$\begin{aligned}\mathscr{L}_{call} \cup \mathscr{L}_{message} = \{&\langle 1,6\rangle, \langle 1,8\rangle, \langle 1,11\rangle, \langle 1,15\rangle, \langle 1,27\rangle, \langle 3,11\rangle, \langle 3,27\rangle,\\ &\langle 4,7\rangle, \langle 4,8\rangle, \langle 4,13\rangle, \langle 4,17\rangle, \langle 4,23\rangle, \langle 5,24\rangle, \langle 6,1\rangle,\\ &\langle 6,11\rangle, \langle 6,15\rangle, \langle 6,27\rangle, \langle 7,4\rangle, \langle 7,13\rangle, \langle 7,23\rangle, \langle 7,27\rangle,\\ &\langle 8,10\rangle, \langle 8,13\rangle, \langle 8,23\rangle, \langle 8,27\rangle, \langle 9,22\rangle, \langle 10,8\rangle, \langle 11,1\rangle,\\ &\langle 11,3\rangle, \langle 11,4\rangle, \langle 11,6\rangle, \langle 11,8\rangle, \langle 11,15\rangle, \langle 11,18\rangle,\\ &\langle 11,23\rangle, \langle 11,26\rangle, \langle 11,27\rangle, \langle 13,4\rangle, \langle 13,7\rangle, \langle 13,8\rangle,\\ &\langle 13,12\rangle, \langle 13,23\rangle, \langle 13,26\rangle, \langle 13,27\rangle, \langle 14,19\rangle, \langle 14,21\rangle,\\ &\langle 14,25\rangle, \langle 15,1\rangle, \langle 15,6\rangle, \langle 15,8\rangle, \langle 15,11\rangle, \langle 15,27\rangle,\\ &\langle 16,19\rangle, \langle 17,4\rangle, \langle 17,6\rangle, \langle 17,11\rangle, \langle 17,12\rangle, \langle 17,18\rangle,\\ &\langle 18,11\rangle, \langle 18,17\rangle, \langle 19,14\rangle, \langle 19,16\rangle, \langle 19,21\rangle, \langle 19,25\rangle,\\ &\langle 21,14\rangle, \langle 21,19\rangle, \langle 21,25\rangle, \langle 22,9\rangle, \langle 23,4\rangle, \langle 23,7\rangle,\\ &\langle 23,8\rangle, \langle 23,13\rangle, \langle 25,14\rangle, \langle 25,21\rangle, \langle 26,11\rangle, \langle 26,13\rangle,\\ &\langle 26,23\rangle, \langle 26,25\rangle, \langle 26,27\rangle, \langle 27,1\rangle, \langle 27,3\rangle, \langle 27,6\rangle,\\ &\langle 27,7\rangle, \langle 27,8\rangle, \langle 27,11\rangle, \langle 27,13\rangle, \langle 27,23\rangle, \langle 27,26\rangle\}\end{aligned}$$

of which fifty-one arcs are common to both graphs

$$\begin{aligned}\mathscr{L}_{call} \cap \mathscr{L}_{message} = \{&\langle 1,15\rangle, \langle 3,11\rangle, \langle 4,7\rangle, \langle 4,23\rangle, \langle 5,24\rangle, \langle 6,1\rangle, \langle 6,11\rangle,\\ &\langle 6,15\rangle, \langle 7,4\rangle, \langle 7,13\rangle, \langle 7,27\rangle, \langle 8,10\rangle, \langle 8,13\rangle, \langle 8,23\rangle,\\ &\langle 8,27\rangle, \langle 9,22\rangle, \langle 11,3\rangle, \langle 11,15\rangle, \langle 11,23\rangle, \langle 11,26\rangle,\\ &\langle 11,27\rangle, \langle 13,4\rangle, \langle 13,8\rangle, \langle 13,23\rangle, \langle 13,26\rangle, \langle 13,27\rangle,\\ &\langle 14,19\rangle, \langle 14,25\rangle, \langle 15,1\rangle, \langle 15,6\rangle, \langle 16,19\rangle, \langle 17,4\rangle,\\ &\langle 17,11\rangle, \langle 19,14\rangle, \langle 19,16\rangle, \langle 19,21\rangle, \langle 19,25\rangle, \langle 21,19\rangle,\end{aligned}$$

$$\langle 21, 25\rangle, \langle 22, 9\rangle, \langle 23, 8\rangle, \langle 23, 13\rangle, \langle 25, 21\rangle, \langle 26, 11\rangle, \langle 26, 13\rangle, \langle 26, 27\rangle, \langle 27, 7\rangle, \langle 27, 8\rangle, \langle 27, 11\rangle, \langle 27, 13\rangle, \langle 27, 26\rangle\}.$$

Four arcs appear on the call network graph but not on the message network graph

$$\mathscr{L}_{call} \setminus \mathscr{L}_{message} = \{\langle 4, 8\rangle, \langle 23, 4\rangle, \langle 23, 7\rangle, \langle 26, 25\rangle\},$$

and thirty-three arcs appear on the message network graph but not on the call network graph

$$\begin{aligned}\mathscr{L}_{message} \setminus \mathscr{L}_{call} = \{&\langle 1, 6\rangle, \langle 1, 8\rangle, \langle 1, 11\rangle, \langle 1, 27\rangle, \langle 3, 27\rangle, \langle 4, 13\rangle, \langle 4, 17\rangle, \\ &\langle 6, 27\rangle, \langle 7, 23\rangle, \langle 10, 8\rangle, \langle 11, 1\rangle, \langle 11, 4\rangle, \langle 11, 6\rangle, \langle 11, 8\rangle, \\ &\langle 11, 18\rangle, \langle 13, 7\rangle, \langle 13, 12\rangle, \langle 14, 21\rangle, \langle 15, 8\rangle, \langle 15, 11\rangle, \\ &\langle 15, 27\rangle, \langle 17, 6\rangle, \langle 17, 12\rangle, \langle 17, 18\rangle, \langle 18, 11\rangle, \langle 18, 17\rangle, \\ &\langle 21, 14\rangle, \langle 25, 14\rangle, \langle 26, 23\rangle, \langle 27, 1\rangle, \langle 27, 3\rangle, \langle 27, 6\rangle, \\ &\langle 27, 23\rangle\}.\end{aligned}$$

3.5.5. Corroboration in the Nodobo Dataset

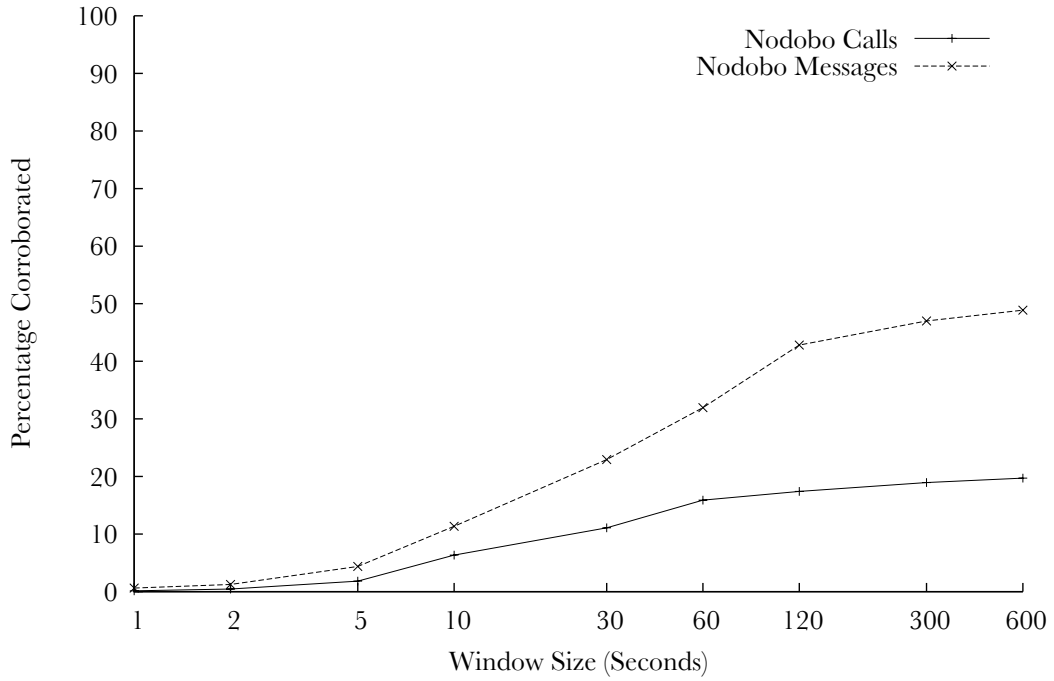
The mediated communications records in the Nodobo dataset were tested for corroborating records in a similar manner to those in the Reality Mining dataset. To corroborate one record another record must be found where the sender and receiver matched and vice versa, the start time is ± 30 seconds (to allow for clock de-synchronisation), and the duration of a call is within ± 5 seconds.

The corroboration observed was better than in the Reality Mining data, but still far from perfect, particularly that of voice calls. 145 (11.08%) of calls corroborated and 5958 (22.93%) of messages corroborated.

To account for larger values of clock de-synchronisation, the corroboration test was run for a window size of up to 10 minutes. The results are shown in Figure 3.21. Allowing for a maximum clock de-synchronisation of ten minutes, approximately 20% of voice calls and 50% of SMS messages in the Nodobo dataset were corroborated by another record.

The level of corroboration observed suggests that either the context gathering software used in the study was not functioning correctly, and, therefore all participants have similar amounts of corroborating interactions, or the context gathering software was functioning correctly but the mechanisms for transferring the data

Figure 3.21.
Percentage of Corroborated Calls and Messages in the Nodobo Dataset.



from the device to the server were not, meaning records were missing for some study participants.

To further investigate the low levels of corroboration, the corroboration for each Nodobo participant who recorded mediated interactions was calculated. The results are shown in Table 3.1.

Levels of corroboration ranged from none to complete corroboration for both calls and SMS messages. Corroboration of messages was generally better than calls. Only one participant (5) has no corroborated calls or messages, one (16) has full corroboration for calls, and one (22) full corroboration for text messages, although the count of interactions in these cases was small. Participant 18 has almost full corroboration of messages with a total of 154 messages, and participant 26 has 90% corroboration of 3,423.

The fact that some participants have high corroboration and others low, suggests that although the corroboration within the dataset as a whole is not high, the context gathering software was working correctly. The lack of corroboration is instead caused by some devices being unable to send data back to the server, perhaps because of lack

Table 3.1.
Corroboration of Calls and Messages for Individual Nodobo Participants.

Participant ID	Corroborating Calls	Corroborating Mesasges
1	6 (75.00%)	178 (88.12%)
3	1 (50.00%)	57 (90.48%)
4	14 (48.28%)	294 (46.74%)
5	0 (0.00%)	0 (0.00%)
6	1 (12.50%)	121 (63.02%)
7	18 (4.48%)	425 (7.66%)
8	7 (5.38%)	28 (5.18%)
9	1 (50.00%)	4 (57.14%)
10	-	12 (27.27%)
11	5 (35.71%)	530 (63.25%)
13	37 (14.62%)	3,830 (52.75%)
14	0 (0.00%)	104 (30.77%)
15	7 (87.50%)	63 (40.13%)
16	1 (100.00%)	1 (50.00%)
17	0 (0.00%)	31 (9.81%)
18	-	154 (99.35%)
19	55 (41.67%)	1,346 (70.55%)
21	55 (32.16%)	1,364 (45.50%)
22	1 (20.00%)	8 (100.00%)
23	6 (30.00%)	1 (16.67%)
25	1 (50.00%)	31 (75.61%)
26	28 (66.67%)	3,423 (90.13%)
27	14 (21.88%)	695 (75.05%)

of WiFi connectivity, or a data tariff, or because some users put their SIM cards into other phones during the study period meaning that interactions were recorded at one end but not the other.

3.6. Summary

Mobile devices are first and foremost communication devices, and human communication is intrinsically linked to social ties. It follows that analysis of communications will yield information about the relationships between communicants. Detecting the presence of social ties can therefore be achieved by detecting communications and attempting to estimate the relationships associated with them.

Previous studies have used both email and mobile phone data in attempts to infer the social networks of communicants. Although all have been done retrospectively on data that is not normally available, and in order to gather this data some special permission is required to access it. Mining the data available on mobile phones allows

access to communications metadata without the need for special access to a particular silo of data.

Concepts from social network analysis are used to provide a common language to describe detected communication networks. In particular, the concept of a relation is used to identify important connections between the members of the network. The relations applied to communications metadata are used to perform traffic analysis studying the characteristics of messages to infer patterns in communications without analysing the content of the messages.

Graph theory can be used to represent the results of network analyses performed on communications metadata. The resulting communication network graphs show the relations between the actors as a set of lines connecting a set of nodes. Voice call and SMS message communication graphs created with the Reality Mining dataset show clustering similar to that observed in analysis of co-proximate interactions in related work, but highlighted the sparsity of the data available, particularly SMS data.

The reliability of datasets can be established by testing records of mediated interactions for corroborating pairs of interactions between communicants. If few corroborating pairs of calls or SMS messages are found then data must be missing from the dataset and therefore cannot be considered accurate. The Reality Mining dataset displays low levels of corroboration for both calls and SMS messages. The lack of corroboration combined with the sparsity of mediated data in the dataset meant that a different dataset was required. Since no other freely available datasets were suitable for studies attempting to infer social ties from communications metadata, the author and his colleagues undertook a project to create a new dataset.

The Nodobo project gathered communications metadata from a group of 27 students at Springburn Academy in Glasgow. From September 2010 to the end of January 2011 the study recorded 13,035 call records, 83,542 SMS records, and 5,292,103 proximity records as well as cell tower IDs and WiFi SSIDs.

The data gathered during the Nodobo study gives an insight in to the behaviour of the study participants. Analysis of the timing of the interactions detected by day of the week and by hour shows that participants were co-proximate most often on weekdays during school hours, that SMS messaging is commonly used throughout the day and the evening on both weekdays and weekends, and that voice calls are more often made before and after school and on Fridays and Saturdays. Voice call and SMS message communication graphs created with the Nodobo dataset show strong similarity although they are not identical suggesting that most of the participants who send SMS messages also call.

Allowing for a maximum clock de-synchronisation of ten minutes, approximately

50% of SMS messages and 20% of voice calls corroborated in the Nodobo dataset. Although these results show that there is clearly some data missing from the Nodobo dataset, the amount of corroboration is significantly higher than in the Reality Mining data. While this improvement is welcome the lack of corroboration is worrying: communication network graphs derived from incomplete metadata may be not be sufficient to accurately infer social ties due to incomplete data.

4. Social Graphs

Contents

4.1. Detected and Reported Social Network Data

A social graph—or *sociogram* in Social Network Analysis literature—is a network of social ties between actors [72]. The actors are represented by the nodes on the graph and, like the communication graphs in Chapter 3, are usually people. (Although network analyses have been performed on relations between companies and nation states [73].)

Until recently social network data was gathered primarily using questionnaires or interviews [74] and more recently data from Social Network Services has also been used [75]. These sources rely on data which is *reported* by individuals rather than collected independently by researchers. This can be problematic: one study carried out by Bernard et al. [9] found that in a blank page report of their social network individuals were only 50% accurate.

Social network data can also be gathered by observation and from archival records [74]. Although these methods are useful for gathering affiliation data—who attended a specific event, for example—observations limit the data to the researchers' impressions of a given situation and appropriate archival records may not exist for the desired data.

The popularity of computer-mediated communication channels, and in particular the mobile phone in the last decade, and the ubiquitous nature of many mobile devices in the present day creates a potential opportunity to gather vast amounts of human social data [2, 18]. This *detected* data allows the inference of social network data which is implicit in the interactions between users, and does not require users to explicitly report who their 'friends' are.

In this chapter a novel estimated social graph of social ties derived from both mobile communications metadata and co-proximity data is presented.

4.2. Estimating Social Ties

In order to accurately estimate the existence of a social tie between two participants in the Nodobo study a set of relations which are an accurate proxy for a social tie must be defined. It is not enough to assume the presence of a social tie based on a single mediated or co-proximate interaction.

Like related work which investigates social ties in both mobile phone and email communication networks [46, 47, 44, 45] this thesis considers study participants to have a tie if they have at least one reciprocal exchange of phone calls during the study period. SMS messages are also considered in this analysis, and the definition of the a relation is extended to include them: study participants are considered to have a tie

if they have had at least one reciprocal exchange of mediated communications. That is, participants A and B will be considered to have a tie if they have exchanged voice calls, SMS messages, or a combination of both.

Proximity interactions are more complicated. There is no need for reciprocity because, by definition, if participant A is co-proximate to participant B the reverse must also be true. However, a single interaction may last less than a minute and for that reason mean little. Moreover, isolated longer interactions may also not be significant—standing close to someone on public transport or in a queue for example, or attending the same class once or twice a week but not having any other social tie.

In order to avoid false positives from either of these potential errors, proximity interactions between participants are only considered pertinent if they satisfy two criteria: proximity for a 'meaningful' period of time, and proximity for a 'regular' period of time.

As an initial estimate, a period of at least thirty minutes total proximity in a given day is used to represent meaningful co-proximity. This is an arbitrary value based on the assumption that this period of time is long enough that insignificant, incidental interactions will be avoided but short enough that significant interactions which happen in passing will be detected.

Similarly, participants are only considered to have a tie if they have been meaningfully co-proximate on four out of every seven days when both participants phones were active. This value is again arbitrary, but it is hoped that many of the longer incidental interactions between participants will be filtered out while regular interactions between participants will be detected.

Social ties are considered to exist between any participants who satisfy either of these relations: either they have exchanged mediated interactions or they have repeatedly been co-proximate for a meaningful period of time, or both. The social ties detected are therefore dichotomous in accordance with the definition of a social tie given in Chapter 1.

4.3. The Nodobo Social Graph

The estimation of social ties between the participants in the Nodobo study is achieved by applying the relations defined above to the data gathered in the Nodobo study. The resulting graph is referred to as the *estimated* social graph.

During the Nodobo study, two additional sets of social network data were obtained: one which was reported by the study participants, and one which was observed by researchers. These datasets were used to create alternative representations

of the Nodobo social graph—the *reported* and *observed* social graphs respectively—and are compared with the estimated social graph later in this chapter to test the accuracy of the relations which are used to estimate ties.

4.3.1. The Estimated Nodobo Social Graph

The estimated Nodobo social graph $\mathscr{G}_{estimated}$ has a set of twenty-seven nodes: one for each of the study participants

$$\begin{aligned}\mathscr{N}_{estimated} &= \mathscr{N}_{nodobo} \\ &= \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, \\ &\quad 21, 22, 23, 24, 25, 26, 27\},\end{aligned}$$

and a set of fifty-four lines joining twenty-three participants

$$\begin{aligned}\mathscr{L}_{estimated} = \{&(1,4), (1,15), (5,24), (6,1), (6,15), (6,23), (7,4), (7,13), \\ &(8,10), (8,15), (11,1), (11,3), (11,4), (11,6), (11,8), (11,17), \\ &(11,18), (11,23), (11,26), (12,13), (13,4), (13,8), (13,26), \\ &(14,21), (15,4), (16,23), (17,4), (17,6), (17,18), (19,14), \\ &(19,21), (19,25), (23,1), (23,3), (23,4), (23,7), (23,8), \\ &(23,13), (23,15), (23,21), (23,26), (25,14), (25,21), (26,4), \\ &(27,1), (27,3), (27,6), (27,7), (27,8), (27,11), (27,13), \\ &(27,15), (27,23), (27,26)\}.\end{aligned}$$

The estimated social graph is shown in Figure 4.1. The graph is dominated by one component containing twenty-one of the twenty-seven participants, two participants form an unconnected dyad pair, and four are unconnected to any of the other participants.

4.3.2. The Reported Social Graph

The reported social graph was derived from interviews with the study participants after the completion of the study.

Participants were given a graphical representation of our estimation of their star network based upon the estimated social ties discussed in Section 4.2. They were then asked to reject any ties with participants who they did not consider to be a friend or school friend and add any ties to participants who they did consider to be a friend or school friend. Estimated ties rejected by participants are considered false positives,

Figure 4.1.
The Estimated Nodobo Social Graph.

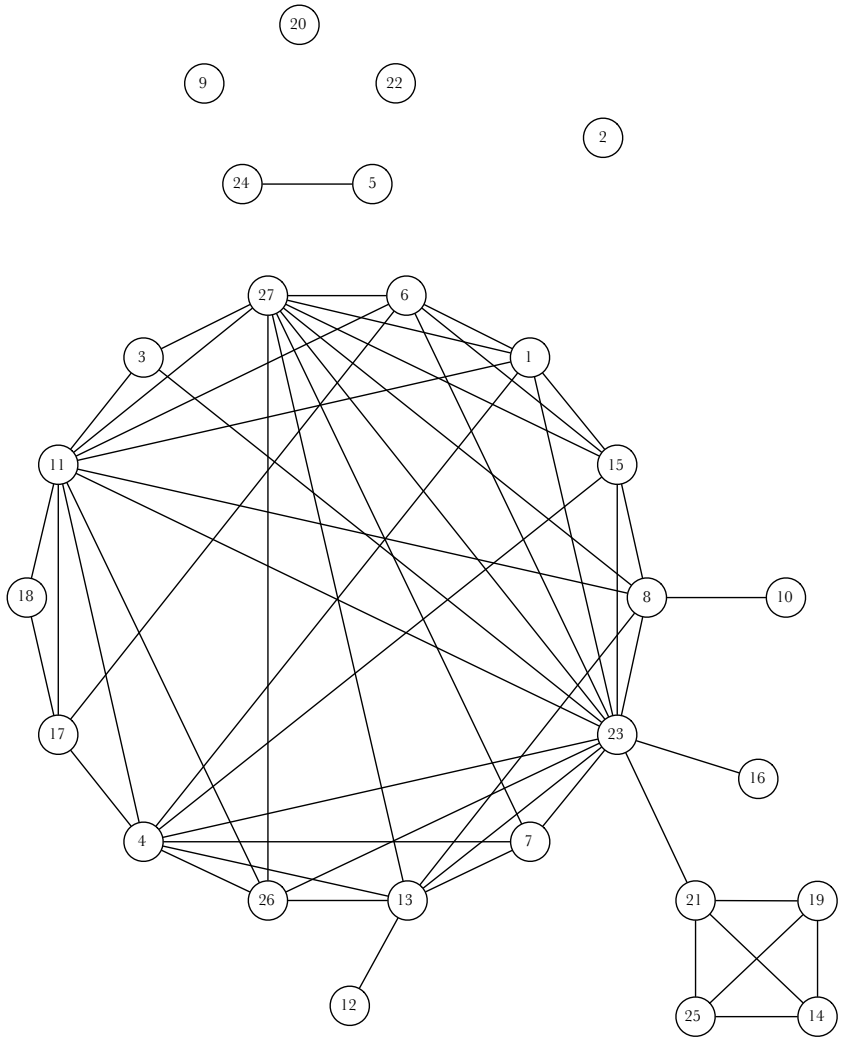


Figure 4.2.
The Reported Nodobo Social Graph.

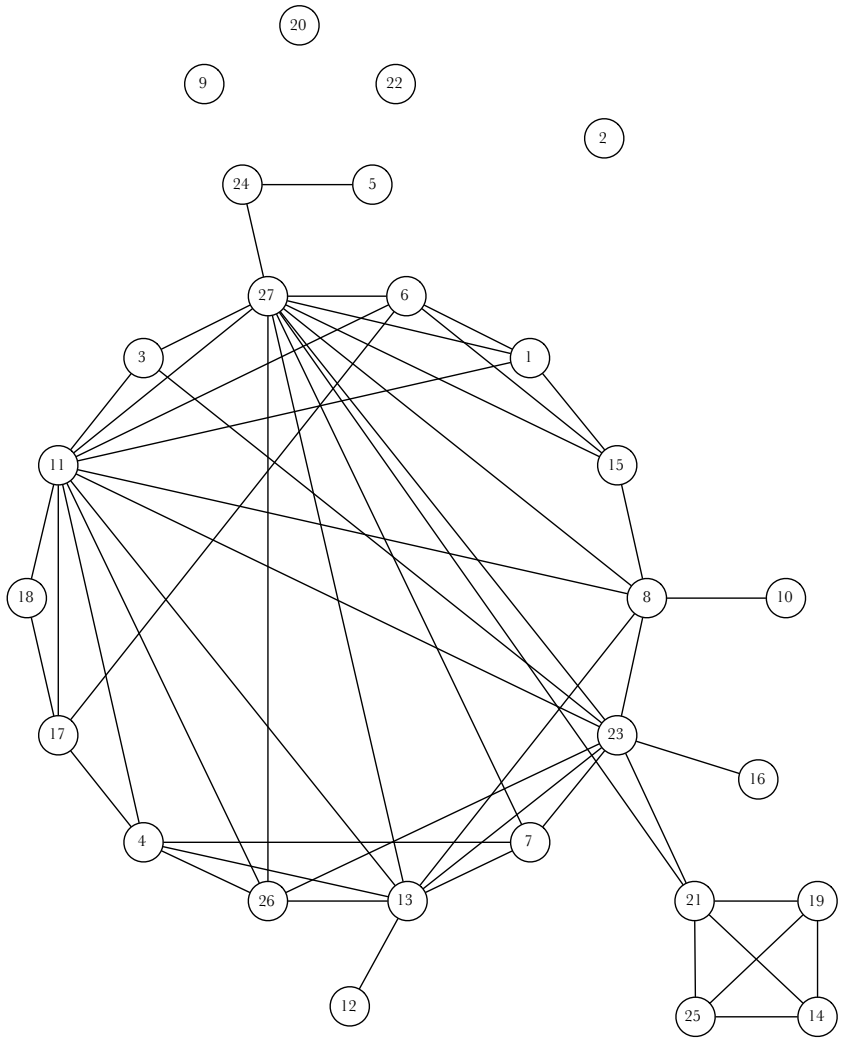


Figure 4.3.
The Observed Nodobo Social Graph.

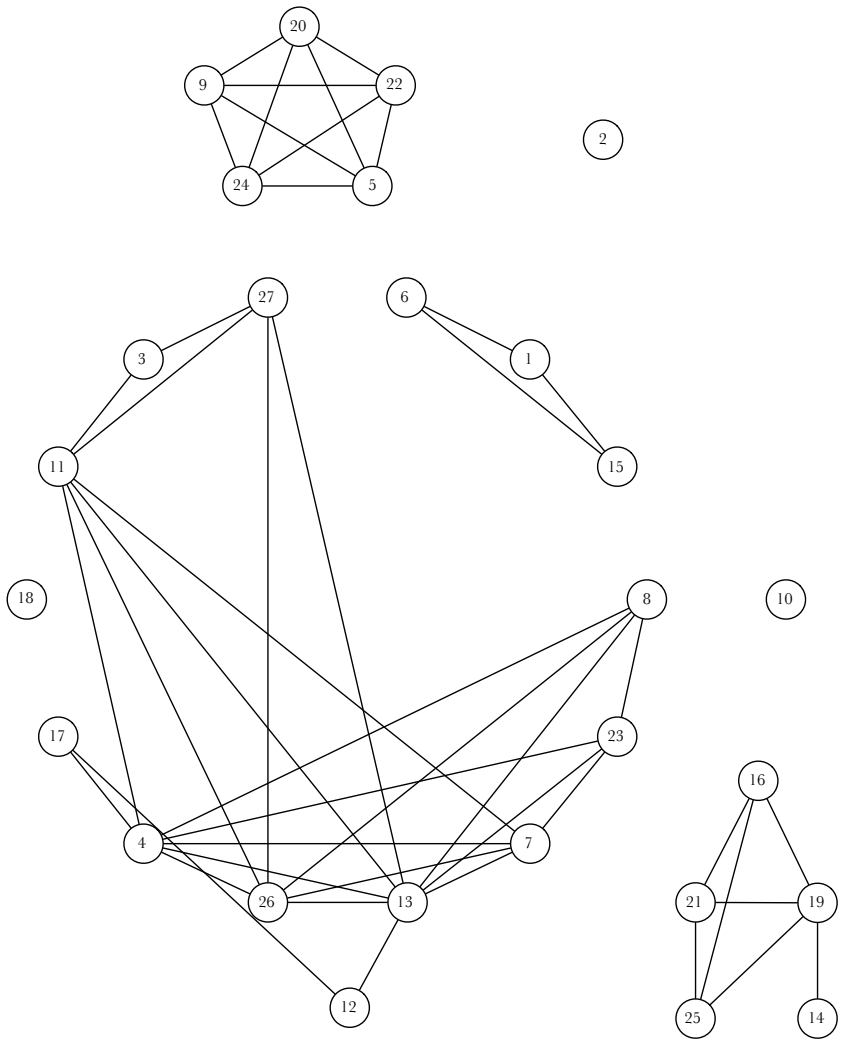


Table 4.1.
Reported Inaccuracies in the Estimated Social Graph.

Participant ID	End IDs of False Positive Ties	End IDs of False Negative Ties
1	-	**4**, **23**
3	-	**23**
4	-	**1**, **15**
5	-	-
6	-	17, **23**
7	-	-
11	**13**	-
13	**11**	-
15	-	**4**, 11, **23**
17	12	-
19	16	-
21	16, **27**	-
23	-	**1**, **3**, **6**, **15**, 16, 21
24	**27**	-
26	-	-
27	**21**, **24**	-

(Inaccuracies confirmed by both ends of the tie shown in bold.)

and estimated ties which were missed are considered false negatives.

Due to participants having left school, and otherwise poor attendance, only fifteen of the twenty-seven participants were interviewed. The results of the interviews are shown in Table 4.1. Three of the participants interviewed reported no false positive ties and no false negative ties. Seven participants reported false negative ties, and six reported false positive ties. No participants had both false positive and false negative ties. The number of false negative ties for a single participant is at most two, and the number of false positive ties for a single participant is between one and six.

The reported Nodobo social graph $\mathscr{G}_{reported}$ also has a set of twenty-seven nodes one for each of the participants

$$\begin{aligned}\mathscr{N}_{reported} &= \mathscr{N}_{nodobo} \\ &= \mathscr{N}_{estimated} \\ &= \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, \\ &\quad 21, 22, 23, 24, 25, 26, 27\}.\end{aligned}$$

As the estimated ties are dichotomous the incomplete interview data complicated the analysis slightly: only discrepancies in the estimated graph agreed upon by both ends of the tie can be considered to be ground truth. The asymmetric interview data

may suggest directed ties which cannot be directly compared with the undirected edges of the estimated graph. Therefore a set of fifty-one lines joining twenty-three participants exist in the reported social graph, which is shown in Figure 4.2.

$$\begin{aligned}\mathscr{L}_{reported} =&\{(1,15),(5,24),(6,1),(6,15),(7,4),(7,13),(8,10),(8,15),\\ &(11,1),(11,3),(11,4),(11,6),(11,8),(11,13),(11,17),\\ &(11,18),(11,23),(11,26),(12,13),(13,4),(13,8),(13,26),\\ &(14,21),(16,23),(17,4),(17,6),(17,18),(19,14),(19,21),\\ &(19,25),(23,3),(23,7),(23,8),(23,13),(23,21),(23,26),\\ &(25,14),(25,21),(26,4),(27,1),(27,3),(27,6),(27,7),\\ &(27,8),(27,11),(27,13),(27,15),(27,21),(27,23),(27,24),\\ &(27,26)\}.\end{aligned}$$

4.3.3. The Observed Social Graph

The observed social graph was derived from observations made in the first few weeks of the study by researchers working with the study participants. They observed the interactions between the participants while spending time with them in school and informally interviewed them both alone and in groups.

The observed Nodobo social graph $\mathscr{G}_{observed}$ again has a set of twenty-seven nodes one for each of the participants

$$\begin{aligned}\mathscr{N}_{observed} =&\mathscr{N}_{nodobo}\\ =&\mathscr{N}_{estimated} = \mathscr{N}_{reported}\\ =&\{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,\\ &21,22,23,24,25,26,27\}.\end{aligned}$$

and a set of forty-five lines joining twenty-four of the participants

$$\begin{aligned}\mathscr{L}_{observed} =&\{(1,6),(1,15),(3,11),(3,27),(4,7),(4,8),(4,11),(4,13),\\ &(4,17),(4,23),(4,26),(5,9),(5,20),(5,22),(5,24),(6,15),\\ &(7,11),(7,13),(7,23),(7,26),(8,13),(8,23),(8,26),(9,20),\\ &(9,22),(9,24),(11,13),(11,26),(11,27),(12,13),(12,17),\\ &(13,23),(13,26),(13,27),(14,19),(16,19),(16,21),(16,25),\\ &(19,21),(19,25),(20,22),(20,24),(21,25),(22,24),(26,27)\}.\end{aligned}$$

The observed social graph is shown in Figure 4.3.

4.3.4. Additional Data in the Observed Graph

In addition to observing where ties existed between study participants some further observations were made about the interactions between them. Four participants are identified as either not attending school regularly or having left school altogether

$$\mathscr{N}_{absent} = \{2, 10, 18, 20\}.$$

Two of this group are unconnected in both the estimated and reported graphs and two are on the periphery of these graphs. This information is interesting but, if required, the archived attendance records from the school would be a better source of data.

Two ties described as 'friends but mostly via text' are also identified

$$\mathscr{L}_{SMS\ ties} = \{(12, 13), (12, 17)\},$$

and one participant is described as 'texting a lot of people'

$$\mathscr{N}_{SMSer} = \{13\}.$$

The Nodobo message network graph (Figure 3.19) shows SMS messages were detected from participant 12 to 13 and from 13 to 12. There is also a tie between them on both the reported and estimated graphs. However, the message network graph only shows messages sent from participant 17 to participant 12 and not from 12 to 17, and no tie between these participants appears on either the reported graph or the estimated graph.

Participant 13 has an out degree of 6 on the Nodobo message network graph, higher than most other participants, but participant 27 has an out degree of 9 and participant 11 has an out degree of 10. Participant 13 does not send messages to an especially high number of the study group but is higher than average.

This additional observed data does show that some observations made of the participants' interaction behaviour are broadly accurate for some participants, but analysis of the gathered communications metadata is a more reliable source of interaction data across the study group as a whole.

The observed data also begins to classify ties between participants. Some participants are identified as being 'not as close as others' and two romantic relationships are also identified. This kind of classification of social ties cannot be

done using the dichotomous social graphs above which are concerned simply with establishing whether or not a social tie exists between participants. It is a logical next step to take however, and will be considered in Chapter 5.

4.4. Comparing the Estimated and Reported Graphs

The set of participants connected to at least one other participant $\mathscr{N}_{joined}$ is the same in both the estimated and reported social graphs

$$\begin{aligned}\mathscr{N}_{joined} =& \mathscr{N}_{est\ joined} = \mathscr{N}_{rep\ joined} \\ =& \{1, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 23, 24, \\ & 25, 26, 27\}.\end{aligned}$$

The superset of all lines in both the estimated and reported social graphs contains fifty-seven lines,

$$\begin{aligned}\mathscr{L}_{estimated} \cup \mathscr{L}_{reported} =& \{(1, 4), (1, 6), (1, 11), (1, 15), (1, 23), (1, 27), \\ & (3, 11), (3, 23), (3, 27), (4, 7), (4, 11), (4, 13), \\ & (4, 15), (4, 17), (4, 23), (4, 26), (5, 24), (6, 11), \\ & (6, 15), (6, 17), (6, 23), (6, 27), (7, 13), (7, 23), \\ & (7, 27), (8, 10), (8, 11), (8, 13), (8, 15), (8, 23), \\ & (11, 13), (11, 17), (11, 18), (11, 23), (11, 26), \\ & (8, 27), (11, 27), (12, 13), (13, 23), (13, 26), \\ & (13, 27), (14, 19), (14, 21), (14, 25), (15, 23), \\ & (15, 27), (16, 23), (17, 18), (19, 21), (19, 25), \\ & (21, 23), (21, 25), (21, 27), (23, 26), (23, 27), \\ & (24, 27), (26, 27)\},\end{aligned}$$

forty-eight lines are common to both graphs

$$\begin{aligned}\mathscr{L}_{estimated} \cap \mathscr{L}_{reported} =& \{(1, 6), (1, 11), (1, 15), (1, 27), (3, 11), (3, 23), \\ & (3, 27), (4, 7), (4, 11), (4, 13), (4, 17), (4, 26), \\ & (5, 24), (6, 11), (6, 15), (6, 17), (6, 27), (7, 13), \\ & (7, 23), (7, 27), (8, 10), (8, 11), (8, 13), (8, 15), \\ & (8, 23), (8, 27), (11, 17), (11, 18), (11, 23),\end{aligned}$$

$$(11,26),(11,27),(12,13),(13,23),(13,26),$$
$$(13,27),(14,19),(14,21),(14,25),(15,27),$$
$$(16,23),(17,18),(19,21),(19,25),(21,23),$$
$$(21,25),(23,26),(23,27),(26,27)\}.$$

The set of six lines which appear on the estimated social graph but not reported social graph is the set of false positive ties $\mathscr{L}_{fpr}$

$$\begin{aligned}\mathscr{L}_{fpr} =& \mathscr{L}_e \setminus \mathscr{L}_r \\ =& \{(1,4),(1,23),(4,15),(4,23),(6,23),(15,23)\},\end{aligned}$$

and the set of three lines which appear on the reported social graph but not the estimated social graph is the set of false negative ties $\mathscr{L}_{fnr}$

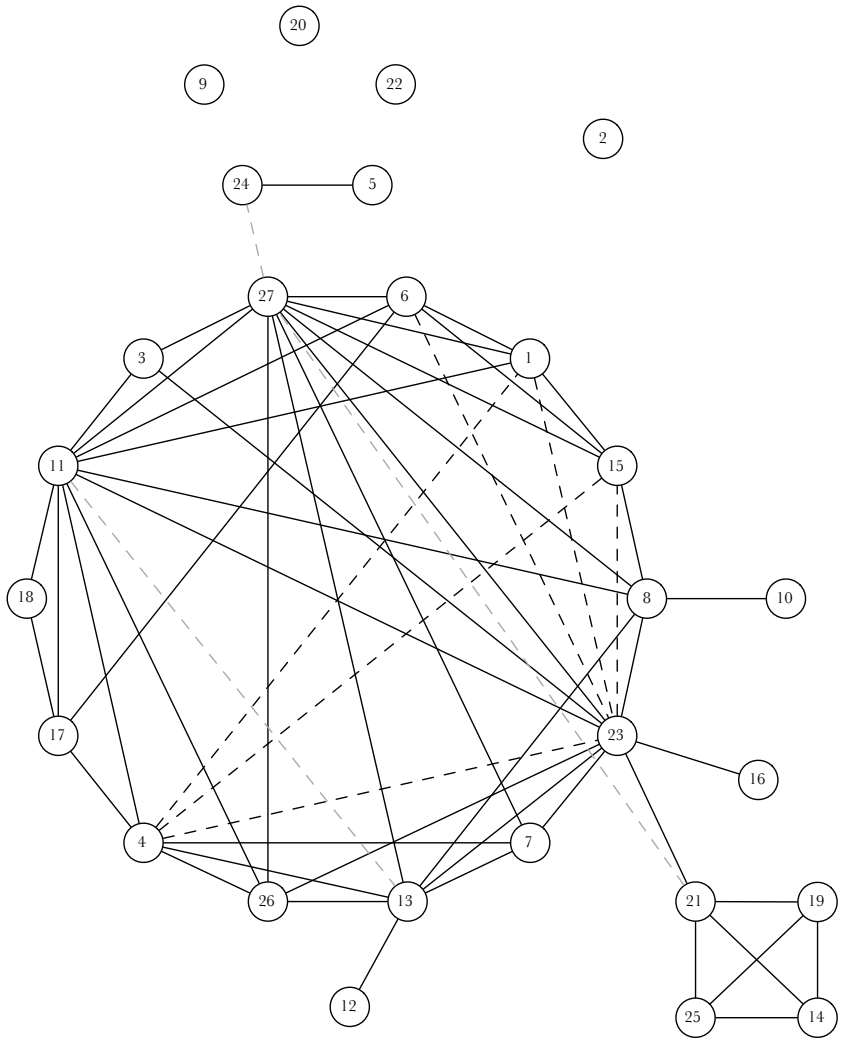
$$\begin{aligned}\mathscr{L}_{fnr} =& \mathscr{L}_r \setminus \mathscr{L}_e \\ =& \{(11,13),(2,27),(24,27)\}.\end{aligned}$$

The differences between the estimated and reported social graphs are shown in Figure 4.4. Ties common to both are shown as solid black edges, false positive ties are shown as dashed black edges, and false negative ties are shown as dashed grey edges.

Both graphs are broadly similar with one large component containing most of the nodes and four nodes completely disconnected from the graph. The reported graph joins nodes 5 and 24 to the main component with the line $(24,27)$, joins the small component of nodes $\{14,19,21,25\}$ to the main component by the lines $(21,27)$ and $(21,23)$, and adds an additional line in the large component between nodes 11 and 13. All of the false positive ties are interconnections within the largest component.

The similarity of both graphs is encouraging: 89% of estimated ties are confirmed as being correctly identified. However, the accuracy of estimated ties was varied, with the estimation of participant 23's ties particularly inaccurate. Moreover, the lack of interview data for twelve of the twenty-seven participants makes a complete comparison impossible because not all reported ties have been confirmed by both ends of the tie. By examining data used to derive confirmed ties, false positive ties, and false negative ties, it is possible to determine the cause of some of the inaccuracies in the estimated social graph.

Figure 4.4.
Differences Between the Estimated and Reported Graphs.



4.4.1. Confirmed Ties

Forty-eight ties were confirmed. They show a large range in the number of mediated interactions between participants: some have no calls, many have a few, and some have hundreds, with 378 calls recorded between participants 4 and 7. SMS messages are more common: only a few ties have none, most have dozens, and some have thousands. Participants 13 and 26 exchanged 10,439 SMS messages in the 144 days of the study: an average of 72.5 messages a day.

All estimated ties show some proximity interactions between the ends. This is expected as all of the participants attended the same school. Some have only hundreds of recorded interactions: there are 185 instances of co-presence between participants 19 and 25, approximately three hours in total (device discovery scans run once a minute), meaning that the tie is most likely based on the calls and messages exchanged by the ends, the co-proximity is incidental. On the other hand, some confirmed ties have many thousands of interactions: there are 36,704 instances of co-presence (approx. 25 days) between participants 13 and 26, for example. This suggests that there is a variation in the proportion of mediated and co-present interactions: while many ties involve both mediated and co-present interaction, some have many more co-present interactions.

Sixteen of the forty-eight confirmed ties are estimated based on co-proximity. Nine of these are estimated based on both mediated interactions and co-proximate interactions,

$$\{(1,15),(4,7),(7,27),(8,23),(11,23),(13,23),(13,26),(23,27),(26,27)\},$$

and the remaining seven are estimated on co-proximity alone

$$\{(3,32),(4,26),(7,23),(15,27),(16,23),(21,23),(23,26)\}$$

Although ties derived only from co-proximity are less common, they appear throughout the graph. In addition to increasing the number of false negative ties to thirteen—making 27% of all estimated ties false positives—removing ties derived only from co-proximity significantly alters the topology of the social graph. Participant 16 becomes unconnected from the graph, as does the small cluster of the main component $\{14, 19, 21, 25\}$. The connectedness of the large cluster of the main component also decreases: the subgroup $\{3, 11, 23, 27\}$ is no longer fully connected, and ties such as $(4, 26)$ and $(15, 27)$ are also lost. Removing ties derived from co-proximate interactions shows the necessity of co-presence data: mediated interactions are not enough to accurately estimate the social graph as social ties exist been people

who do not communicate via voice call or SMS.

For more detailed information, a count of the voice calls, SMS messages, and proximity interactions between the ends of the forty-eight estimated ties confirmed by the reported social graph are given in Appendix A.

4.4.2. False Positive Ties

The existence of false positive ties suggests that the relations used to estimate ties between the Nodobo participants are too lenient. Either the ends have exchanged calls or SMSs but do not consider that there is a tie between them, or there are no calls or SMSs between the ends and the false positive tie is based on flaws in the rules for estimating ties based on co-presence.

A count of the total calls, SMS messages, and co-proximate interactions between each end of the six false positive ties is shown in Table 4.2. There are no calls or SMS messages between the ends of five of the ties—they are comprised only of proximity interactions—but the tie (4, 23) includes both calls and SMS messages.

Table 4.2.
Details of False Positive Ties.

Tie	Call Count	SMS Count	Proximity Count
(1,4)	0	0	2,066
(1,23)	0	0	1,967
(4,15)	0	0	924
(4,23)	5	31	598
(6,23)	0	0	2,652
(15,23)	0	0	3,162

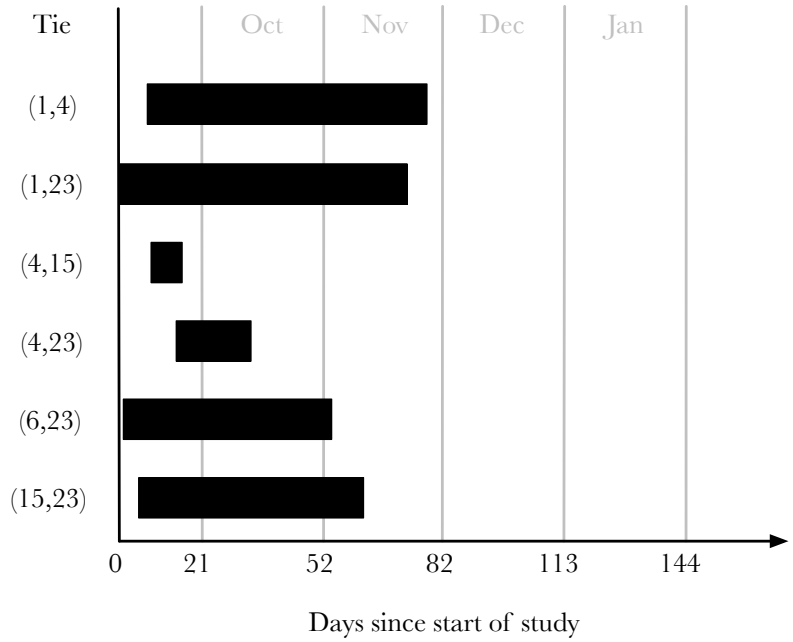
Participants 4 and 23 did not identify a tie between themselves despite exchanging both calls and SMS messages. This is the only instance of the co-proximity relation incorrectly identifying a tie. Further investigation showed that all calls between these participants were made on one day shortly after the beginning of the study—the ninth of September—and that SMS messages were sent on four days between the 17th of September and 16th of October, with all mediated interactions stopping long before the end of the study. This false positive suggests that either: the assumption that reciprocated mediated interaction always signals a tie is incorrect; some participants who do not identify a tie between themselves do occasionally use mediated channels to communicate, or that the dynamic nature of the social graph is beginning to show flaws in this static graph-theoretic analysis; participants 4 and 23 may have had a social tie at some point in the past, but some weeks later when the tie between them

was estimated, no longer consider that they do.

The participants at each end of the remaining five false positive ties do not exchange any mediated communications and therefore false positive ties must be created by a problem with the co-proximity relation. Although these users do not identify each other as friends they do spend some time in proximity to one and other resulting in the estimated tie. This effect may be caused by the fact that although some of the participants will be within social distance in classes they are compelled to be there, and given the choice they may not choose to socialise with each other.

The relatively low amount of co-proximity between participants 4 and 15 is a cause for concern: if a tie can be estimated based only on fifteen hours of co-proximity then it is likely that either one or both of these participants' phones were not active for a significant part of the study period. The time period that co-presence was detected for each end of the false positive ties was calculated. The results are shown in Figure 4.5.

Figure 4.5.
Time Periods When Co-presence was Detected for False Positive Ties.

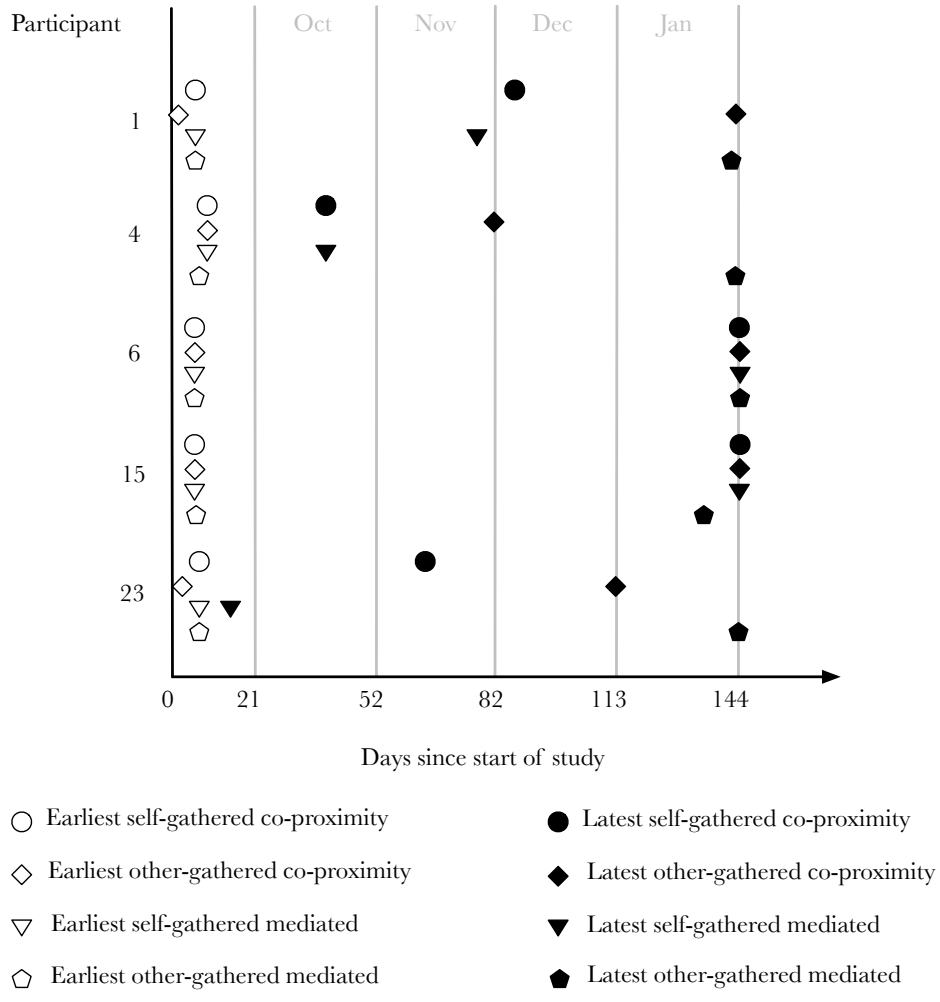


The periods of time where the ends of false positive ties are co-present are all significantly shorter than the duration of the study—participants 4 and 15 are co-present for only a handful of days at the beginning of the study. This suggests that either these false positive ties are also examples of a friendship which was no longer considered to exist when the study participants were interviewed, or that a

malfunction in the devices used by these participants prevented accurate data being gathered causing false positive ties.

The timestamps of all the interactions for each of the study participants who reported false positive ties show whether or not data was received throughout the duration of the study. Five participants reported false positive ties: 1, 4, 6, 15, and 23. Interactions data for a specific participant can be added to the dataset either by that participant or the participant at the other end of the interaction, denoted self-gathered and other-gathered in this analysis. The earliest and latest timestamps of both self-gathered and other-gathered mediated or co-proximate interactions for participants who reported false positive ties are shown in Figure 4.6.

Figure 4.6.
The Earliest and Latest Timestamps of Interactions of Participants who Reported False Positive Ties.



Three different patterns are visible, each of which suggest a different source of

errors. The first, as seen in data for participants 6 and 15, shows participants for whom both self-gathered and other-gathered mediated and co-proximate interactions are present in the dataset. This suggests that the ties were incorrectly identified based on problems with co-proximity relation and not technical problems gathering data or participants using a different handset or SIM card. Self-gathered and other-gathered co-proximity data for the duration of the study show that participants were carrying their devices throughout, and self-gathered and other-gathered mediated interactions show that they were using them to communicate.

The second pattern, exhibited by participant 1, shows other-gathered interactions which span the duration of the study, but self-gathered interactions for part of the study only. This suggests that the participant was using their device for the duration of the study—a fact verified by other participants who gathered interactions with that device—but that a technical problem prevented data being sent over the air to the server. This may have been due to a malfunction context gathering software on the device or simply the lack of access to either a cellular or WiFi interface. Whatever the cause, some data from this participant is missing which may have contributed to the estimation of false positive ties.

The third pattern, seen in participants 4 and 23, shows participants whose other-gathered mediated interactions span the duration of the study, but self-gathered mediated interactions and all co-proximate interactions are only gathered for part of the study. Mediated interactions were gathered by other devices for the duration of the study, but only self-gathered (by the participants' study devices) for part of it, suggesting that participants used the same SIM card in another phone for part of the study. Their study devices did report some interactions, suggesting that the study device was used for some time, and the fact that co-proximity was detected for a large part of the study suggests that the study device may have been carried in addition to another phone. Participants using SIM cards in other handsets may have lead to many, if not all, of the false positive ties as participants 4 and 23 are one of the ends of all of the false positive ties (and both of the ends in one tie).

Using the same SIM card in another handset also accounts for some of the low corroboration discussed in Section 3.3.3. When the SIM is removed from the study device the phone number used to identify the participant (part of the tuple participant id, phone number, and Bluetooth MAC address) will stay the same, but only one end of any mediated communications sent or received at that time will be recored and sent to the database.

4.4.3. False Negative Ties

False negative ties presumably exist because there were no recorded interactions between these participants and therefore no tie could be estimated. Either there was no reciprocal exchange of calls or SMS messages or an insufficient amount of co-presence was detected. It is also possible that calls or SMSs were sent from one end of the tie but not the other resulting in an unbalanced, 'one-sided' tie.

Table 4.3 shows total counts for calls, SMS messages, and proximity between each end of the three confirmed false negative ties.

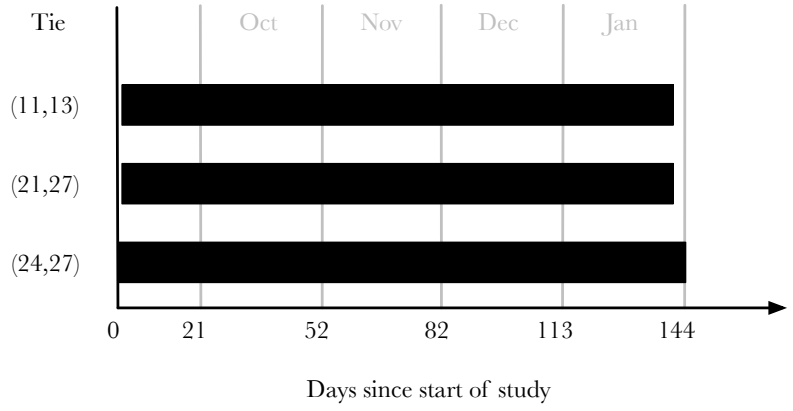
Table 4.3.
Details of False Negative Ties.

Tie	Calls Count	SMS Count	Proximity Count
(11,13)	0	0	10,176
(21,27)	0	0	13,799
(24,27)	0	0	2,713

These results show that, like the majority of false positives, ties inferred solely from proximity interactions cause all of the false negative ties. False negative ties are not caused by one-sided mediated communication, meaning that there are no participants who do identify a tie between themselves but do not have a reciprocal exchange of mediated interactions.

The time period that co-presence was detected for each end of the false negative ties was calculated in the same way as that for false positives. The results are shown in Figure 4.7.

Figure 4.7.
Time Periods When Co-presence was Detected for False Negative Ties.



Unlike false positive ties, co-proximity is detected for false negative ties throughout the duration of the study. This suggests that the participants who reported false negative ties are using the study devices and that there are no problems retrieving the data gathered on those devices. The fact that they contain no mediated interactions—like most false positives—suggests that the co-proximity relation is too straightforward and must be improved to detect more ties. It is also possible that participants who identified false negative ties communicate using channels on which interactions could not be detected by the devices used in the study, such as instant messaging or social network services. The periods of time where the ends of false positive ties are co-present are all significantly shorter than the duration of the study, with participants 4 and 15 being co-present for only a handful of days. Similarly, the date of the last proximity interaction record for one end of the tie consistently comes significantly before the end of the study, with the ties $(1, 4)$ and $(1, 23)$ both having no detected co-proximity between each other after the end of November and none at all after the beginning of December, nearly two months before the end of the study.

Modified smartphones are useful tools for gathering data on social interactions. Providing participants with handsets with some software modifications allows both meditated and co-proximate interactions to be detected. Gathered data is then sent over the air to a central server were it is collated and then queried for social ties. However, testing the corroboration of records in the resulting dataset shows that significant data loss is experienced from some devices.

Despite the loss of some data it is still possible to accurately estimate many social ties between the participants. Using only a few simple rules we were able to correctly identify ties between participants 89% of the time. Ties were estimated based on two relations: one for mediated interactions, and the other for co-proximate interactions.

The mediated relation requires a reciprocal exchange of calls or SMS messages and are a strong indicator of the presence of social ties. Thirty-two of forty-eight confirmed ties were estimated in this way, with one false positive, in the analysis presented here. The need for reciprocity also indicates that confirmed ties are not based on one-sided communication: no false negative ties were identified in which participants identified a tie between them but only communicated in one direction.

The co-proximity relation requires regular proximity for a meaningful period of time. For the analysis shown here this was arbitrarily defined as proximity on four out of seven days for a total of thirty minutes. Seven confirmed ties were estimated based on this relation with a further nine satisfying both relations. This shows the importance of co-proximity when identifying social interactions. However, the co-proximity relation is the source of most incorrectly identified ties. Five of the six false

positive ties satisfied the co-proximity relation, and all three false negative ties show co-proximity is detected throughout for the study participants involved.

These errors may be due to missing data, and examples have been shown of both devices which failed to push data to the server and participants who removed their SIM cards from study devices during the study, but in future work a more precise co-proximity relation should be employed.

4.5. Comparing the Estimated and Observed Graphs

The estimated social graph and the observed social graph have considerable differences. The set of nodes which are joined by at least one tie in observed social graph $\mathscr{N}_{obs\ joined}$ is not equal to the set of participants connected by at least one tie in the estimated and reported graphs

$$\begin{aligned}\mathscr{N}_{obs\ joined} =& \{1, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22,\\ & 23, 24, 25, 26, 27\}\\ \neq & \mathscr{N}_{joined}.\end{aligned}$$

Two participants are joined by ties in the estimated and reported graphs which were not joined in the observed social graph

$$\mathscr{N}_{joined} \setminus \mathscr{N}_{obs\ joined} = \{10, 18\},$$

and three participants are joined in the observed social graph but not in the estimated or reported social graphs

$$\mathscr{N}_{obs\ joined} \setminus \mathscr{N}_{joined} = \{9, 20, 22\}.$$

The superset of all lines in both the estimated and observed social graphs contains seventy-two lines,

$$\begin{aligned}\mathscr{L}_{estimated} \cup \mathscr{L}_{observed} =& \{(1, 4), (1, 6), (1, 11), (1, 15), (1, 23), (1, 27),\\ & (3, 11), (3, 23), (3, 27), (4, 7), (4, 8), (4, 11),\\ & (4, 13), (4, 15), (4, 17), (4, 23), (4, 26), (5, 9),\\ & (5, 20), (5, 22), (5, 24), (6, 11), (6, 15), (6, 17),\\ & (6, 23), (6, 27), (7, 11), (7, 13), (7, 23), (7, 26),\\ & (7, 27), (8, 10), (8, 11), (8, 13), (8, 15), (8, 23),\end{aligned}$$

$$(8,26),(8,27),(9,20),(9,22),(9,24),(11,13),$$
$$(11,17),(11,18),(11,23),(11,26),(11,27),$$
$$(12,13),(12,17),(13,23),(13,26),(13,27),$$
$$(14,19),(14,21),(14,25),(15,23),(15,27),$$
$$(16,19),(16,21),(16,23),(16,25),(17,18),$$
$$(19,21),(19,25),(20,22),(20,24),(21,23),$$
$$(21,25),(22,24),(23,26),(23,27),(26,27)\},$$

with twenty-seven lines common to both graphs

$$\begin{aligned}\mathscr{L}_{estimated} \cap \mathscr{L}_{observed} =\{&(1,6),(1,15),(3,11),(3,27),(4,7),(4,11),(4,13),\\ &(4,17),(4,23),(4,26),(5,24),(6,15),(7,13),\\ &(7,23),(8,13),(8,23),(11,26),(11,27),(12,13),\\ &(13,23),(13,26),(13,27),(14,19),(19,21),\\ &(19,25),(21,25),(26,27)\}.\end{aligned}$$

The set of twenty-seven ties which appear on the estimated graph which are not on the observed graph is the set of false positive ties $\mathscr{L}_{fpo}$

$$\begin{aligned}\mathscr{L}_{fpo} =&\mathscr{L}_{estimated} \setminus \mathscr{L}_{observed}\\ =&\{(1,4),(1,11),(1,23),(1,27),(3,23),(4,15),(6,11),(6,17),\\ &(6,23),(6,27),(7,27),(8,10),(8,11),(8,15),(8,27),(11,17),\\ &(11,18),(11,23),(14,21),(14,25),(15,23),(15,27),(16,23),\\ &(17,18),(21,23),(23,26),(23,27)\},\end{aligned}$$

and the set of eighteen lines which appear on the observed social graph but not the estimated graph social graph is the set of false negative ties $\mathscr{L}_{fno}$

$$\begin{aligned}\mathscr{L}_{fno} =&\mathscr{L}_{observed} \setminus \mathscr{L}_{estimated}\\ =&\{(4,8),(5,9),(5,20),(5,22),(7,11),(7,26),(8,26),(9,20),\\ &(9,22),(9,24),(11,13),(12,17),(16,19),(16,21),(16,25),\\ &(20,22),(20,24),(22,24)\}.\end{aligned}$$

Although twenty-seven observed ties are confirmed by the reported social graph there are equally many false positive ties on the observed graph suggesting that casual observation is a highly ineffective method of estimating the presence of social ties

between individuals. Furthermore, as noted in Section 4.1, observation data are limited to the impressions of the observers and because they cannot be observing all participants at all times the data gathered is consequently incomplete. This effect presumably creates the eighteen false negative ties observed and further reinforces the ineffectiveness of observational data gathering.

4.6. Summary

A social graph is a network of social ties between actors. Using estimated ties to create social graphs allows the inference of social network data which is implicit in the interactions between users, and does not require users to explicitly report who their 'friends' are.

In this chapter a novel estimated social graph of social ties derived from both mobile communications metadata and co-proximity data is presented.

Three graphs derived by different techniques are analysed: the estimated graph derived from social interactions detected using mobile devices, the reported social graph derived from interviews with the study participants after the completion of the study, and the observed graph derived from observations made during the study by researchers working with the study participants

Two relations were defined which are used to create the estimated social graph: *Study participants are considered to have a tie if they have had at least on reciprocal exchange of mediated communications*, and *Study participants are considered to have a tie if they have been co-proximate for a period of at least thirty minutes on four out of every seven days when both participants phones were active.* Social ties are considered to exist between any participants who satisfy either or both of these relations.

The reported social graph was created by giving each participant a graphical representation of their star network based upon the estimated social graph. Participants were asked to reject any ties with participants who they did not consider to be a friend or school friend and add any tie to participants who they did consider to be a friend or school friend. The reported social graph is considered to be ground truth, and as such, an accurate representation of the social network between participants at the time that they were interviewed.

The observed social graph was created collating impressions of the social ties between participants made in the first few weeks of the study by researchers. They observed the interactions between the participants while spending time with them in school and informally interviewed them both alone and in groups.

The estimated and reported graphs are broadly similar but not identical. Forty-

eight of fifty-four estimated ties are confirmed by comparison with the reported social graph meaning that 89% percent of estimated ties were estimated correctly. There were six false positive estimated ties and three false negatives.

The high percentage of correctly estimated ties is encouraging, and this initial attempt at estimating the social graph accurate. However, the accuracy of estimated ties was varied for some participants, with the estimation of participant 23's ties particularly inaccurate. Furthermore, the lack of interview data for twelve of the twenty-seven participants makes a complete comparison impossible.

Thirty-two of forty-eight confirmed ties were estimated using the mediated relation, with one false positive. The need for reciprocity also indicates that confirmed ties are not based on one-sided communication: no false negative ties were identified in which participants identified a tie between them but only communicated in one direction. Seven confirmed ties were estimated based on this relation with a further nine satisfying both relations. This shows the importance of co-proximity when identifying social interactions. However, the co-proximity relation is the source of most incorrectly identified ties. Five of the six false positive ties satisfied the co-proximity relation, and all three false negative ties show co-proximity is detected throughout for the study participants involved.

The existence of false positive ties suggests that the relations used to estimate ties between the participants are not stringent enough. By examining data used to derive accurately estimated ties and the false positive and false negative ties it is possible to determine the cause of the inaccuracies in the estimated social graph. Three potential causes of errors were identified: incomplete data, participants not using the study devices (or using the study device and another device), and changes in the social graph between the last time some data was received and the time the reported graph was compiled.

There is much less similarity between the estimated graph and the observed graph. Although there are some ties in common between the observed and estimated graphs the large number of both false positive and false negative ties makes any meaningful comparison impossible. Observation data are limited to the impressions of the observers and because they cannot be observing all participants at all times the data gathered is consequently incomplete. The strong similarity between the estimated and reported graphs also suggests that the observed graphs is not as accurate over the complete set of participants as either of the other two social graphs.

However, the observed graph does begin to classify the ties between participants. Some participants are identified as being 'not as close as others' and two romantic relationships are also identified. More detailed information such as this is not

available on graphs of dichotomous ties, thus a method of valuing each tie is required.

5. Tie Strength

Contents

5.1. The Concept of Tie Strength

The concept of tie strength, and how to measure it, has been studied by sociologists for some time. In his influential paper *The Strength of Weak Ties* Mark Granovetter defines tie strength in terms of the amount of time, emotional intensity, intimacy, and reciprocal services which characterise the tie [76]. Ties are broadly categorised into three types: strong, weak, and absent. Strong ties are longstanding, close social bonds between people. Weak ties are acquaintances who connect with one another, but overall are not close friends. Absent ties are those without substantial significance, but imply some occasional, fleeting interaction.

Strong ties are the basis for relationships involving trust [77], and are essential in helping organisations to cope with crises [78]. Weak ties can be used to find the *local bridges* which are the sources of new information to each clique in the network [76], and aid the transfer of non-complex information through the network [79].

However, the ties discussed in Granovetter's work are intangible. The definition of tie strength as a combination of the amount of time, emotional intensity, intimacy, and reciprocal services which characterise the tie [76] is difficult—if not impossible—to quantify and measure.

Marsden and Campbell [80] attempt to use *indicators* of strength based on Granovetter's original definition, and additionally *predictors* of tie strength. The notion of *closeness* is used as a measure of the intensity of a relationship: a survey choice of the tie end as an acquaintance, good friend, or very close friend; duration and frequency of contact index the amount of time spent in a tie; and measures of the breadth of topics discussed by friends and the extent of mutual confiding are used to represent intimacy. Predictors used are kinship, neighbour and co-worker status, and overlapping organisational memberships [80].

Petróczi et al. [81] gather details of tie strength in online networks using questionnaires. Information on various tie strength components such as frequency, intimacy/closeness, voluntary investment in the tie, advice given/received, desire for companionship, multiple social context (breadth of topics), long period of time (duration), reciprocity, provide support/emotional intensity, mutual confiding (trust), sociability/conviviality is gathered [81].

Similarly, Gilbert and Karahalios used questionnaires to gather information of the ties of Facebook users. Importantly, they also automatically captured data about study participants and their Facebook friends including: predictive intensity variables such as wall words exchanged and inbox messages exchanged; intimacy variables such as participants' number of friends, friends' number of friends, days since last communication, inbox intimacy words, appearances together in photos,

and friends' relationship statuses; a duration variable–days since first communication; reciprocal services variables—links exchanged by wall post and applications in common; structural variables including the number of mutual friends and groups in common; emotional support variables such as wall and inbox positive emotion words and wall and inbox negative emotion words; and social distance variables including age difference, number of occupations difference, and educational difference [75].

The intimate association of many mobile devices with individuals makes it possible to use them, and associated technologies, to sense the presence of people and the creation and modification of their social ties [18]. This chapter investigates the possibility of using metadata from mobile communications networks to estimate communicants' tie strength.

5.1.1. Tie Strengths Observed in the Nodobo Study

In Section 4.3.4 some additional data about the social ties between the participants in the Nodobo study is discussed. The observed graph also contains some information about the strength of the ties between some participants: seven ties are described as 'not as close as others' and two romantic relationships are identified. If it is assumed that the romantic relationships are examples of strong social ties then a sample set of weaker ties and sample set of stronger ties can be defined:

$$\mathscr{L}_{weaker} = \{(5, 20), (9, 20), (16, 21), (16, 19), (16, 25), (20, 22), (20, 24)\},$$

and

$$\mathscr{L}_{stronger} = \{(4, 7), (13, 26)\}.$$

Some potential proxies for tie strength—total number of SMS messages, aggregated SMS message length, total number of calls, aggregated call duration, and total number of proximity interactions—were compared for both sample sets. If the weaker ties have significantly lower values of tie strength for each measure than the stronger ties, then it may be possible to use these measures to determine the strength of ties of estimated social graphs.

The results are shown in Tables 5.1 and 5.2. Weaker ties have significantly lower values of all tie strength proxies than stronger ties. In fact, all but one of the weaker ties have no calls or SMS exchanged between the ends of the tie, with $(16, 19)$ having only three of each, while the average values for SMS count and call count of stronger ties are 8,220.5 and 291 respectively. Similarly, the average number of co-proximity

Table 5.1.
Weaker Ties from the Observed Social Graph.

Tie	SMS Count	SMS length	Call Count	Call Duration	Proximity Count
(16,21)	0	0	0	0	3,493
(16,19)	3	170	3	96	7,239
(16,25)	0	0	0	0	696
(5,20)	0	0	0	0	690
(9,20)	0	0	0	0	139
(20,22)	0	0	0	0	520
(20,24)	0	0	0	0	258
Average	0.43	24.29	0.43	13.71	1,862.14

Table 5.2.
Stronger Ties from the Observed Social Graph.

Tie	SMS Count	SMS length	Call Count	Call Duration	Proximity Count
(4,7)	6,002	131,520	378	110,026	6,973
(13,26)	10,439	450,657	204	93,144	36,704
Average	8,220.5	291,088.5	291.0	101,585.0	21,838.5

interactions is over ten times larger for stronger ties than for weaker, although the strong tie (4, 7) has fewer proximity interactions than (16, 19): a further indication that co-proximity interactions cannot simply be thresholded.

These results suggest, anecdotally at least, that tie strength can be inferred from communication metadata: a stark difference is seen between the stronger and weaker ties for all of the potential proxies for tie strength tested. However, only the dyadic ties themselves are considered here; any relationship with the social graph as a whole is not considered, and therefore the results cannot be generalised beyond the dyadic ties considered.

5.2. Triadic Closure and Forbidden Triads

> If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future [82].

This quote from a paper written by Anatol Rapoport more than half a century ago succinctly describes an intuitive process which is known in sociology as *triadic closure*. Triadic closure is a key concept in *The Strength of Weak Ties* as it enables the

strength of dyadic ties to be related to the larger social graph [76]. Granovetter asserts that

> the stronger the tie between A and B, the larger the proportion of individuals […] to whom they will both be tied, that is, connected by a weak or strong tie. This overlap in their friendship circles is predicted to be least when their tie is absent, most when it is strong, and intermediate when it is weak.

Instead of attempting to define tie strength precisely, and considering instead the 'friend-acquaintance dichotomy' proposed in Easly and Kleinberg's discussion of the strength of weak ties hypothesis, where strong ties represent closer friendships with a greater frequency of interaction and weak ties acquaintances with few interactions [83], then it may be possible to estimate the strength of ties by measuring the number of common neighbours of each tie.

The number of closed triads which include a given line in the network graph is equal to the number of common neighbours between the nodes at each end of the line. This is illustrated in Figure 5.1: each common neighbour adds an other closed triad to the graph.

5.2.1. Triadic Closure in the Reality Mining Dataset

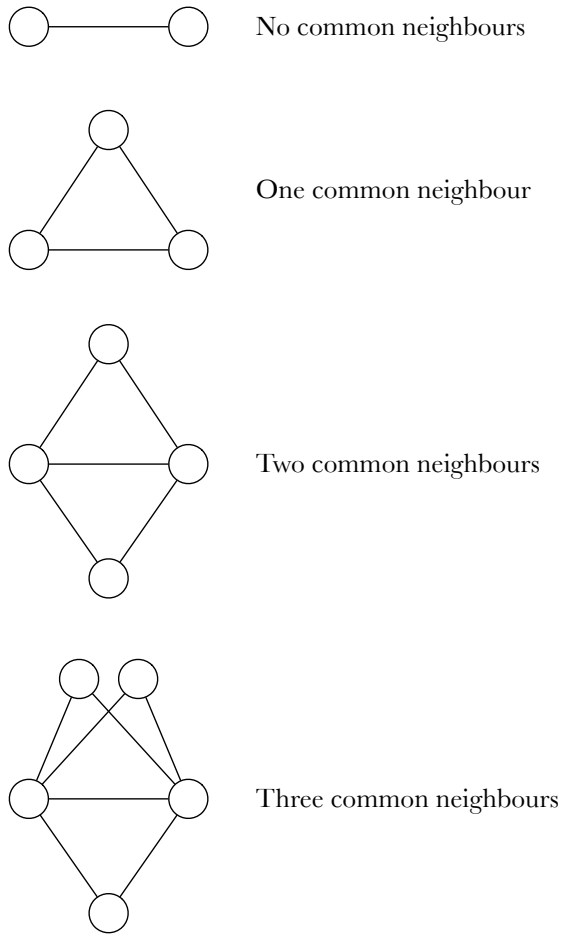
There are 112 ties based on directed mediated communications in the Reality Mining data. These relationships are considered to be undirected, there are therefore 56 ties with mediated interactions between study participants. These interactions are between 54 people, and there are 8 triads with mediated interactions on all sides.

Table 5.3 presents details of the significant triads present in the Reality Mining dataset. For each triad some detail of each of the three ties is given: the number of participants at each end of the tie, and the count of calls, short messages, and detected proximities.

Table 5.3.
Triads Found in the Reality Mining Dataset.

Triad			(a, b)			(a, c)			(b, c)		
a	b	c	Call	SMS	Prox	Call	SMS	Prox	Call	SMS	Prox
15	80	94	6	0	100	62	0	585	2	0	53
15	85	94	8	2	68	62	0	585	21	0	72
20	72	79	42	0	163	6	0	45	12	0	61
7	71	89	4	0	22	5	0	61	1	0	105
35	46	73	4	1	65	2	0	73	1	0	111
29	57	86	10	3	429	66	0	560	17	0	353

Figure 5.1.
Triads and Number of Common Neighbours.



Although it is clear that triads do form in the Reality Mining data, and that some Reality Mining participants have common ties to others, there are not many triads. Furthermore, only two triads share a common tie meaning that only one tie has more than one common neighbour, and the maximum value of number of common neighbours is two.

This suggests that the few ties that exist, are not especially strong, and exist between small groups of participants. The common tie between the overlapping triads is the tie between participants 15 and 94. Interestingly, this tie has the second highest count of calls and the highest count of proximity interactions of all the ties which form triads, providing further inconclusive evidence that counts of interactions act as a proxy for tie strength.

5.2.2. Triadic Closure in the Nodobo Dataset

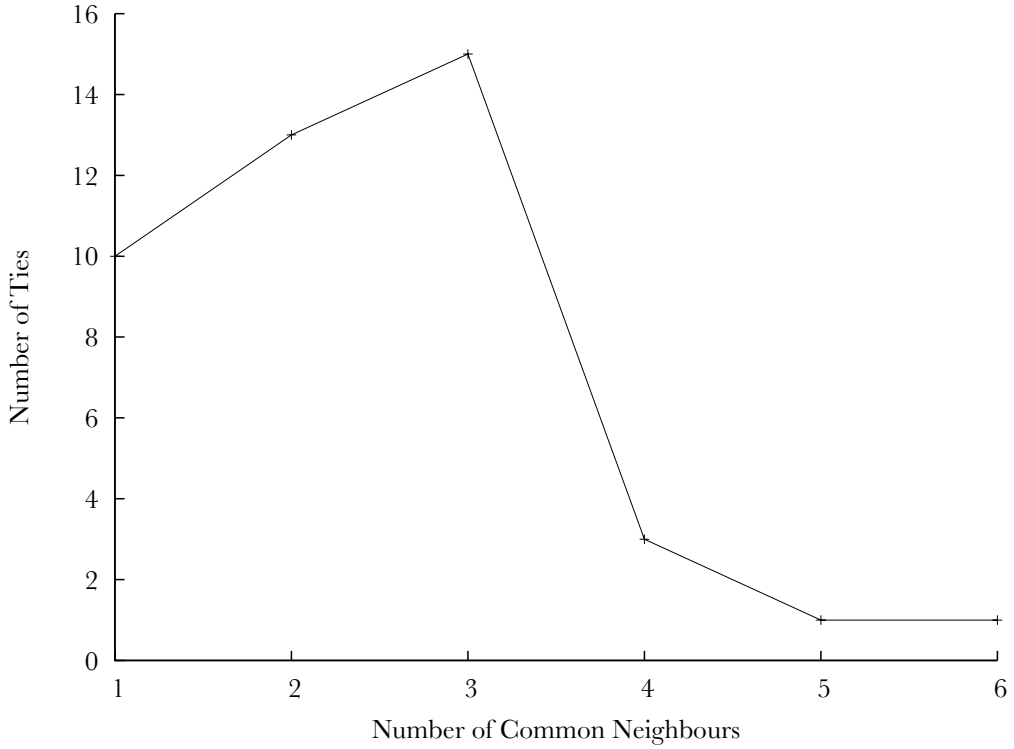
Table 5.4.
Triads Found in the Nodobo Dataset.

Triad			(a, b)			(a, c)			(b, c)		
a	b	c	Call	SMS	Prox	Call	SMS	Prox	Call	SMS	Prox
1	6	11	5	241	10730	0	40	7422	1	20	8680
1	6	15	5	241	10730	15	199	32168	3	17	20371
1	6	27	5	241	10730	0	114	9915	0	54	13344
1	8	27	0	1	6686	0	114	9915	10	38	7343
1	11	15	0	40	7422	15	199	32168	1	6	12628
1	11	27	0	40	7422	0	114	9915	10	856	15999
1	15	27	15	199	32168	0	114	9915	0	1	18906
3	11	27	6	38	10998	0	86	9333	10	856	15999
4	7	13	378	6002	6973	3	7	2253	1	13	13695
4	7	23	378	6002	6973	5	31	598	1	1	3773
4	8	13	2	0	945	3	7	2253	30	36	9108
4	8	23	2	0	945	5	31	598	120	472	7609
4	11	17	0	5	866	3	198	1193	1	67	6311
4	13	23	3	7	2253	5	31	598	71	328	5142
6	11	15	1	20	8680	3	17	20371	1	6	12628
6	11	27	1	20	8680	0	54	13344	10	856	15999
6	15	27	3	17	20371	0	54	13344	0	1	18906
7	13	23	1	13	13695	1	1	3773	71	328	5142
7	13	27	1	13	13695	97	239	18113	4	134	18959
7	23	27	1	1	3773	97	239	18113	0	104	4450
8	11	27	0	2	7587	10	38	7343	10	856	15999
8	13	23	30	36	9108	120	472	7609	71	328	5142
8	13	27	30	36	9108	10	38	7343	4	134	18959
8	23	27	120	472	7609	10	38	7343	0	104	4450
11	15	27	1	6	12628	10	856	15999	0	1	18906
11	17	18	1	67	6311	0	243	2639	0	88	2710
11	26	27	4	20	13540	10	856	15999	5	339	21997
13	23	26	71	328	5142	204	10439	36704	0	1	3148
13	23	27	71	328	5142	4	134	18959	0	104	4450
13	26	27	204	10439	36704	4	134	18959	5	339	21997
14	19	21	5	237	5750	0	10	3559	285	4616	23712
14	19	25	5	237	5750	10	291	434	2	16	185
14	21	25	0	10	3559	10	291	434	12	114	345
19	21	25	285	4616	23712	2	16	185	12	114	345

A similar study performed using the Nodobo dataset (the only difference being that ties require a reciprocal exchange of mediated interactions). An examination of the mediated relations in the dataset found thirty-four triads in the Nodobo data. Counts of calls, SMS messages, and co-proximity interactions are shown in Table 5.4.

The number of common neighbours in the Nodobo dataset is shown in Figure 5.2. Like Reality Mining, some triads do not overlap any others (showing participants with number of common neighbours of one) and in some cases number of common neighbours of five or six can be seen.

Figure 5.2.
The Number of Common Neighbours in the Nodobo Dataset.



The increase in the number of common neighbours seen in the Nodobo data shows that the Nodobo social graph is much more clustered than the Reality Mining graph. This is possibly because the participants in the Nodobo study, who were all in the same year-group at school, are more similar than than the Reality Mining participants, who were students and faculty from two separate schools in MIT [10].

5.3. Using Communications Metadata to Estimate Tie Strength

By using a large corpus of mobile phone user data ($N = 4.6 \times 10^6$) Onnella et al. [46], established that call duration can be used as a proxy for tie strength in large, complex social networks. A coupling between the neighbourhood overlap of two mobile phone users' social graphs and the cumulative distribution of the aggregated duration of the calls between is observed.

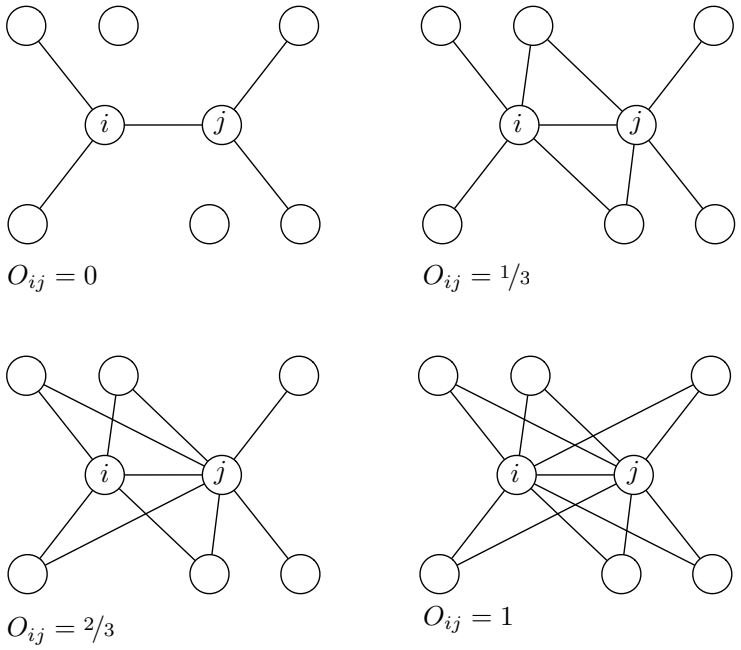
The strength of weak ties hypothesis establishes a link between neighbourhood overlap and tie strength: the greater the neighbourhood overlap between two people, the stronger the tie between them. Overlap is represented quantitatively as the proportion of common neighbouring nodes [46], and is calculated as shown in equation (5.1),

$$\begin{aligned} O_{ij} &= \frac{\text{Number of nodes who are neighbours of both } i \text{ and } j}{\text{Number of nodes who are neighbours of at least } i \text{ or } j} \\ &= \frac{n_{ij}}{((k_i - 1) + (k_j - 1) - n_{ij})} \end{aligned} \tag{5.1}$$

where n_{ij} is the number of common neighbours of two nodes i and j, and k_i and k_j are the degrees of nodes i and j respectively.

This idea is illustrated in Figure 5.3. Nodes i and j are joined by a line, and four different states of overlap are shown. $O_{ij} = 0$ when nodes i and j have no common neighbours, $O_{ij} = 1/3$ when two of a possible six neighbours are common to i and j, $O_{ij} = 2/3$ when four of a possible six neighbours are common to i and j, and $O_{ij} = 1$ when all six possible neighbours are common to i and j.

Figure 5.3.
An Illustration of Varying Overlap Between Two Nodes. (Adapted from [46].)



By using a large corpus of mobile phone user data, Onnella et al. [46] established that certain communications metadata can be used as a proxy for tie strength in large,

complex social networks. A coupling between the neighbourhood overlap of two mobile phone users' social graphs and the cumulative distribution of the aggregated duration of the calls between them is observed. This shows that tie strength can be estimated from the communications metadata of large numbers of users if data pertaining to all users is collated.

Measuring the neighbourhood overlap of two users is possible with existing sets of test data, but it may not be possible or desirable to do this in an application in real-time:

- the need to send and receive all contacts every time we want to calculate overlap is not efficient,
- the need to share contacts to calculate overlap endangers users' privacy,
- users may not be available; they need to be online, or have a proxy policy enforcer to share up-to-date contact lists.

Therefore suitable proxies are required to estimate overlap using only data available locally on mobile devices.

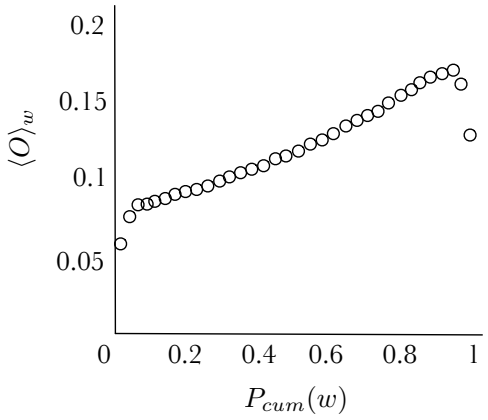
Mobile devices provide users with many communications channels. Voice calls, SMS and MMS messages are possible with every mobile phone. Smartphones add email, instant messaging, VoIP and video calls, and other internet communications. Local-area communications and proximity detection are possible with Bluetooth and WiFi on many smartphones and PDAs.

Metadata from each communication attempt, whether or not it was successful or reciprocated, and information about the timing of communications such as the frequency, the time elapsed since the last communication, or the duration of any conversations can all be gathered from mobile devices and possibly could be used to infer the strength of ties between communicants.

5.3.1. Aggregated Call Duration as a Proxy for Tie Strength

Onnela et al. [46] ask how the neighbourhood overlap of a tie (line on the social graph) depends on the strength of the tie. The strength of weak ties hypothesis predicts that overlap should increase as tie strength increases [83]. This relationship is seen in the data. Figure 5.4 shows the neighbourhood overlap of edges as a function of their percentile in the sorted order of all edges by tie strength [83]. Moving along the x-axis from left to right shows ties of greater and greater strength, and because of the linear relationship visible on the plot, overlap also increases with increasing tie strength.

Figure 5.4.
Overlap as a Function of Cumulative Tie Strength. (From [46])



5.3.2. Tie Strength in the Reality Mining Dataset

When the methodology employed by Onnela et al. to estimate tie strength is applied to the Reality Mining data, the results are mixed. The strength of the ties between participants who exchanged calls with another participant are considered, although due to the small number of corroborated calls in the dataset, reciprocity is not required .

In Figure 5.5, two proxies using call metadata are considered: aggregated call duration and total number of calls made. For ties between participants, overlap generally increases as a function of cumulative aggregated call duration, although the increase is not as smooth as that seen in Figure 5.4, and the range of values of overlap is an order of magnitude smaller. This significant difference in scale may be due to the small number (N = 97) of participants in Reality Mining compared to the large number in the corpus used by Onnela et al. ($N = 4.6 \times 10^6$), and the lack of call data within the dataset.

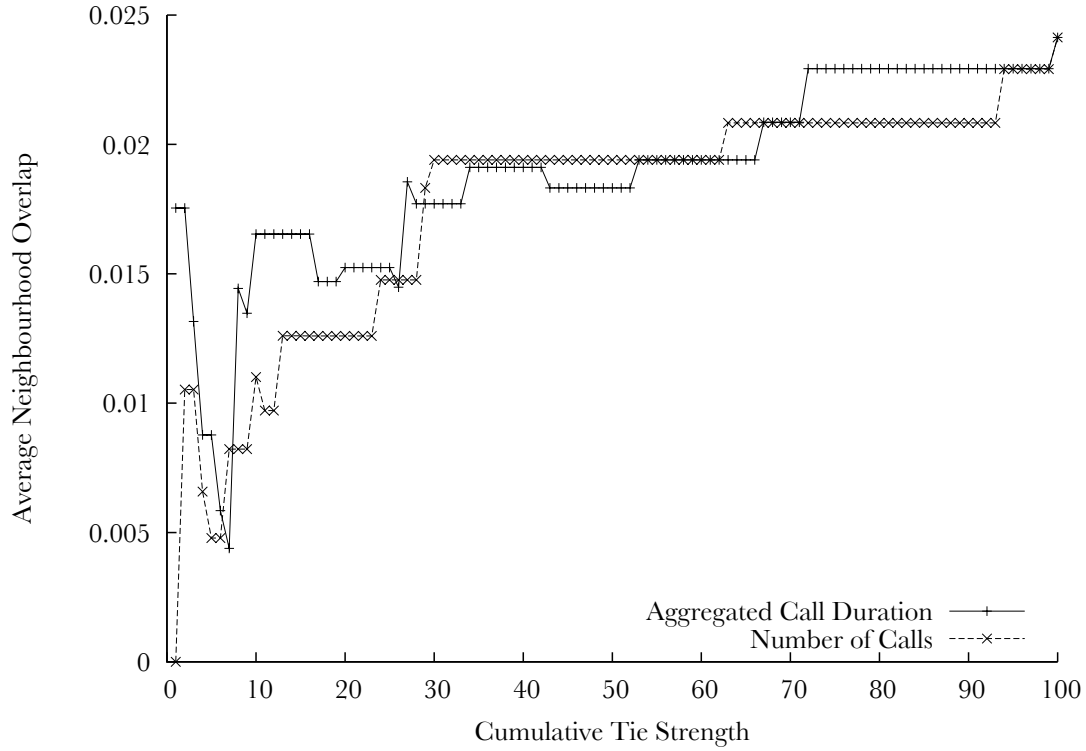
This analysis suggests that aggregated call duration and call count may be used as proxies for tie strength in small mobile communication networks. However, neither is completely conclusive, and high fluctuation is visible especially for weaker ties.

5.3.3. Tie Strength in the Nodobo Dataset

Similar investigations testing the suitability of certain communications metadata as proxies for tie strength were performed using the Nodobo data. Because of the high number of SMS messages in the dataset, SMS data as well as call data was tested.

Figure 5.6 shows the results of SMS metadata tie strength proxies plotted against

Figure 5.5.
Overlap as a Function of Cumulative Call Metadata in Reality Mining.

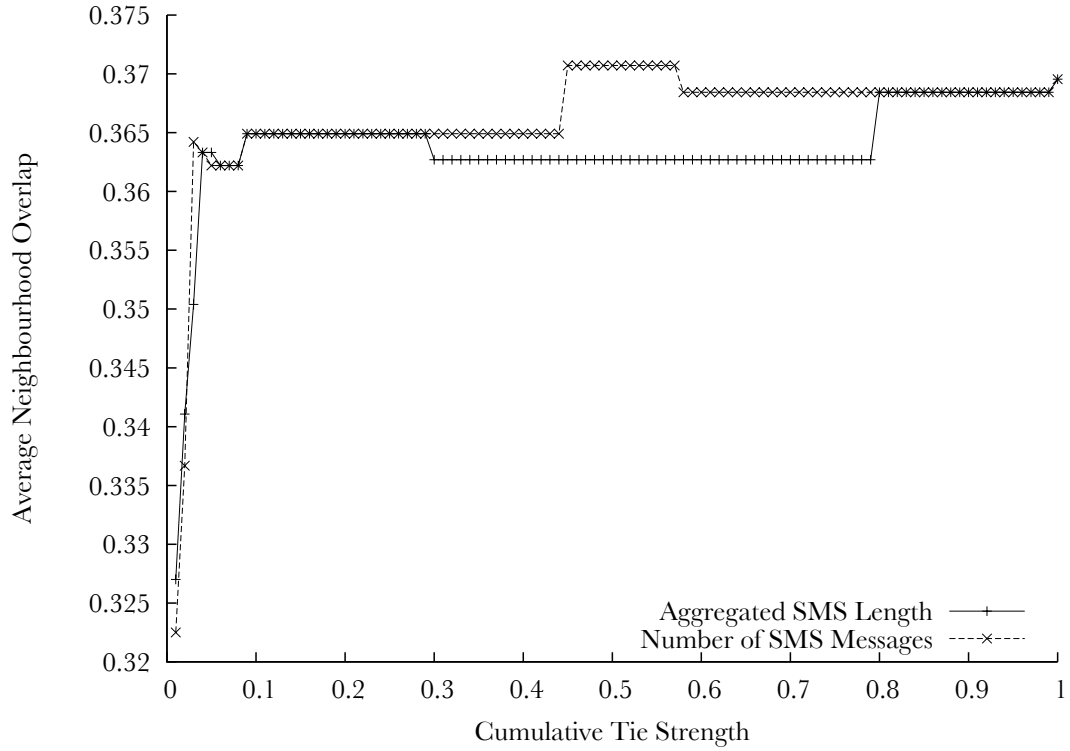


overlap. Both the aggregated SMS length in characters and the total number of SMS messages are shown. No linear relationship between tie strength and overlap is observed. Both curves rise sharply before flattening meaning that the overlap in the tenth percentile of tie strength is approximately the same as the overlap in the ninetieth percentile.

Figure 5.7 shows the results of voice call proxies for tie strength plotted against overlap. The results are similar to those seen for SMS proxies: no linear relationship is observed. Total number of calls initially rises sharply before levelling, and overlap against aggregated call duration is almost completely flat: average neighbourhood overlap is constant for all values of tie strength.

These results show that neighbourhood overlap, calculated using both call and SMS metadata, does not necessarily give any indication of tie strength. The calculated values of overlap for the Nodobo data are generally constant in most cases, with only smallest aggregated totals of mediated communication metadata being slightly smaller. The participants in the Nodobo study are all members of the same year group in the same school, and as such may have similar relationships with one another. Certainly, the heterogeneity of individuals (and hence, social ties)

Figure 5.6.
Overlap as a Function of Cumulative SMS Metadata in Nodobo.



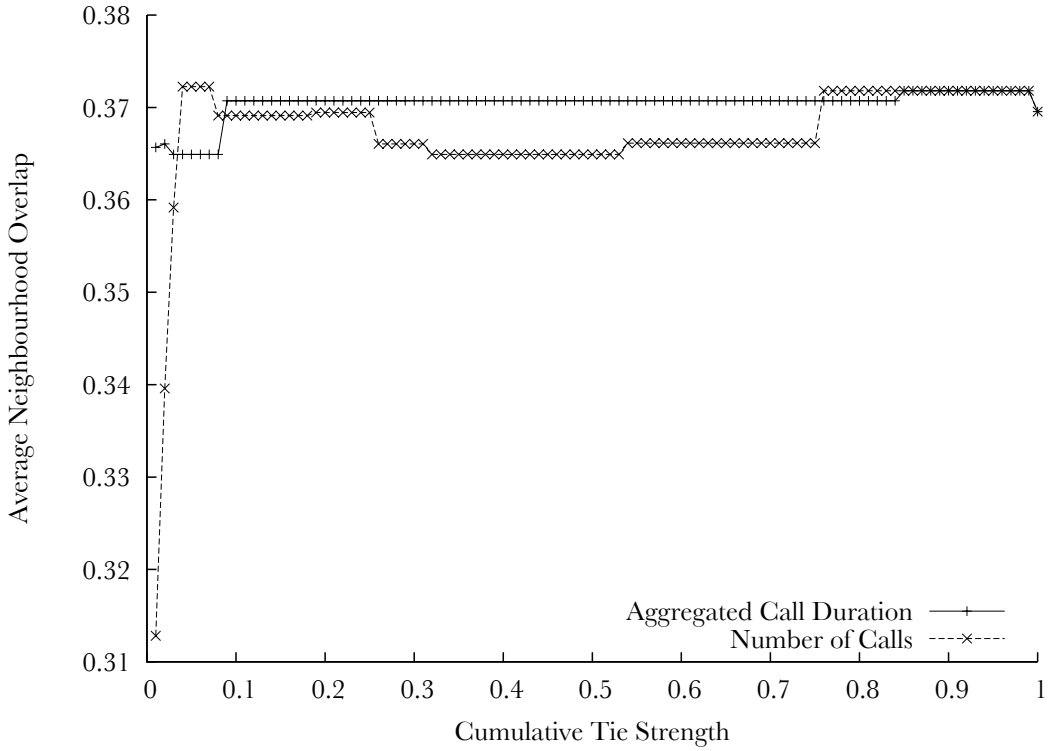
recorded in the data used by Onnella et al. is significantly different to Nodobo, whose participants are similar.

The data used by Onnela et al. is from a European mobile network provider and contains the call metadata of millions of people meaning that a huge variety of different people and therefore different social ties are represented. The same is also true, albeit to a lesser extent, with the Reality Mining data. Reality Mining participants were selected from a larger group than Nodobo and although some similar 'classmate' relationships undoubtedly exist in the Reality Mining data, the wider variety of participants probably creates a wider variety of social ties explaining why the linear relationship between tie strength and neighbourhood overlap still holds loosely.

5.4. Summary

Although the existence of social ties can be inferred from communications metadata, this inference give no indication as to the current state of the ties. All social ties are not equal and change over time. The concept of tie strength seeks to differentiate

Figure 5.7.
Overlap as a Function of Cumulative Call Metadata in Nodobo.



social ties based on the strength of the bond between individuals. Ties are broadly classified into three types: strong, weak, and absent by Mark Granovetter. Strong ties are longstanding, close social bonds between people. Weak ties are acquaintances who connect with one another, but overall are not close friends. Absent ties are those without substantial significance, but imply some occasional, fleeting interaction.

Triadic closure is the process through which social networks develop neighbourhood overlap. Examining both the Reality Mining and the Nodobo datasets found instances of triadic closure occurring in a graph of reciprocated mediated interactions. Of the thirty-six triads present in the Nodobo data twenty-one are closed by proximity interactions before mediated, with the proximity interactions happening on average a month before the mediated interactions.

Few instances of closed triads are seen in the Reality Mining data. This lack of common neighbours shows low clustering for the Reality Mining network. The Nobobo data contains many examples of closed triads meaning that many nodes on the graph have common neighbours. The Nodobo graph is therefore more highly clustered than the Reality Mining graph. The difference in the demographic of the study participants may account for the difference clustering: the Nodobo participants,

who as part of the same year-group in the same school, can be considered to be more homogeneous than the Reality Mining participants, who were both students and faculty from different schools within the same university.

Neighbourhood overlap is considered an indicator of tie strength: the greater the number of common acquaintances, the stronger the tie. In large mobile networks there is a linear correlation between the average neighbourhood overlap and the aggregated duration of calls. This allows aggregated call duration to be used as a proxy for tie strength.

This approach to estimating tie strength was tested on two significantly smaller datasets: the Reality Mining dataset which has ninety-seven participants and the Nodobo dataset which has twenty-seven participants. We found that a loose linear correlation between aggregated call duration and neighbourhood overlap is seen in the Reality Mining data but not in the Nodobo data. The number of calls were also tested and gave similar results. SMS message metadata from the Nodobo dataset was also tested and again no increase in overlap as aggregated SMS length or SMS count increased.

These results suggest that the techniques used to estimate tie strength in large networks cannot be applied to smaller networks because the increase in tie strength seen does not correspond to a clear linear increase in neighbourhood overlap. However, testing tie strength proxies for weaker and stronger ties observed by researchers working with the Nodobo study participants did show larger values for all tie strength proxies for stronger ties.

It does seem that the intuitive idea of stronger social ties between communicants being visible in increased mediated interaction does have some basis. However, on the evidence of the datasets tested here tie strength cannot be said to be directly coupled with neighbourhood overlap.

6. Conclusions and Future Work

Contents

6.1. Summary and Conclusions

The results presented in this thesis show that mobile devices can be used to detect the social ties between communicants. The explosion in the availability and popularity of computer-mediated communication channels in recent years has led to unprecedented levels of digital interactions which are available for analysis. Modern mobile devices are relevant to any attempts to discover social ties because they are potential repositories of many machine-readable ties. Furthermore, the short range communication channels almost universally available on these devices allow insights into the face-to-face interactions of users as well as mediated interactions. Mobile communications is a domain perfectly placed to build tools to explore social relationships: they allow access to many new digital tie-signs from many different communication channels; and their mobility and ubiquity combined with their ability to detect other local devices create the possibility of analysis of face-to-face interactions too.

Face-to-face interactions are still the dominant mode of interaction between people. The proximity of individuals can be itself a kind of tie-sign, revealing information about the nature of the tie between them, for example individuals interacting at social distance. These *co-proximate* interactions are important when attempting to detect social ties. Mobile devices, albeit with some modifications to the operating system and with additional software, can be used as complete sensor nodes when attempting to detect co-proximate interactions.

Mobile devices are first and foremost communication devices, and human communication is intrinsically linked to social ties. It follows that analysis of communications will yield information about the relationships between communicants. Detecting the presence of social ties can therefore be achieved by detecting *mediated* interactions and attempting to estimate the social ties associated with them.

Previous studies have used both email and mobile phone data in attempts to infer the social networks of communicants. These studies have been carried out retrospectively on data that is not normally available and some special permission was required to access it. Mining the data available on mobile phones allows access to communications metadata without the need for special access to a particular silo of data.

This approach has both pros and cons. On the one hand, using rootable, SIM-unlocked handsets allows modified handsets in which study participants can use their existing SIM cards to be deployed. It provides a (relatively) straightforward way to detect both mediated and co-proximate interactions between small ($N < 100$) groups of existing subscribers over various mobile networks. However, testing the

corroboration of records in the resulting dataset shows that significant data loss is experienced from some devices. Furthermore, deploying customised devices is expensive in both money and time, and does not easily scale to large numbers of participants.

Despite the loss of some data it is still possible to accurately estimate many social ties between the participants. Using only a few simple rules we were able to correctly identify ties between participants 89% of the time. Ties were estimated based on two relations: one for mediated interactions, and the other for co-proximate interactions.

The mediated relation requires a reciprocal exchange of calls or SMS messages and are a strong indicator of the presence of social ties. Thirty-two of forty-eight confirmed ties were estimated in this way, with one false positive, in the analysis presented here. The need for reciprocity also indicates that confirmed ties are not based on one-sided communication: no false negative ties were identified in which participants identified a tie between them but only communicated in one direction.

The co-proximity relation requires regular proximity for a meaningful period of time. For the analysis shown here this was arbitrarily defined as proximity on four out of seven days for a total of thirty minutes. Seven confirmed ties were estimated based on this relation with a further nine satisfying both relations. This shows the importance of co-proximity when identifying social interactions. However, the co-proximity relation is the source of most incorrectly identified ties. Five of the six false positive ties satisfied the co-proximity relation, and all three false negative ties show co-proximity is detected throughout for the study participants involved.

These errors may be due to missing data, and examples have been shown of both devices which failed to push data to the server and participants who removed their SIM cards from study devices during the study, but in future work a more precise co-proximity relation should be employed.

Although the existence of social ties can be inferred from communications metadata, this inference give no indication as to the current state of the ties. All social ties are not equal and change over time. The concept of tie strength seeks to differentiate social ties based on the strength of the bond between individuals. However, the techniques used to estimate tie strength in large networks cannot be applied to smaller networks because the increase in tie strength seen does not correspond to a clear linear increase in neighbourhood overlap.

Testing tie strength proxies for weaker and stronger ties observed by researchers working with the Nodobo study participants did show larger values for all tie strength proxies for stronger ties, suggesting that the intuitive idea of stronger social ties between communicants being visible in increased mediated interaction does have

some basis. Although, using the datasets tested here tie strength cannot be said to be directly coupled with neighbourhood overlap.

6.2. Contributions

The contributions of this thesis are

1. A dataset;
2. Dataset veracity;
3. Reliable detection of social ties; and
4. Classification of social ties;

each of which is summarised in the following subsections.

6.2.1. A Dataset

A new dataset of social interaction data using mobile phones has been compiled. The Nodobo project gathered communications metadata from a group of 27 students at Springburn Academy in Glasgow. From September 2010 to the end of January 2011 the study recorded 13,035 call records, 83,542 SMS records, and 5,292,103 proximity records as well as cell tower IDs and WiFi SSIDs. The dataset has been made freely available.

6.2.2. Dataset Veracity

The reliability of datasets can be established by testing records of mediated interactions for corroborating pairs of interactions between communicants. If few corroborating pairs of calls or SMS messages are found then data must be missing from the dataset and therefore the data gathered cannot be reliable.

Allowing for a maximum clock de-synchronisation of ten minutes, approximately 20% of voice calls and 50% of SMS messages corroborated in the Nodobo dataset. These results show that there is clearly some data missing from the Nodobo dataset, and suggest that using mobile devices to gather social data is a method susceptible to errors.

6.2.3. Reliable Detection of Social Ties

Two relations were defined to create an estimated social graph: *Study participants are considered to have a tie if they have had at least on reciprocal exchange of mediated communications*,

and *Study participants are considered to have a tie if they have been co-proximate for a period of at least thirty minutes on four out of every seven days when both participants phones were active*. Social ties were considered to exist between any participants who satisfy either or both of these relations.

The estimated graph derived from these rules was compared to a self-reported social graph and was found to be broadly accurate. Forty-eight social ties were accurately identified with six false positive ties and three false negative ties.

By applying simple rules grounded in common sense thinking to communications metadata gathered by mobile devices, it is possible to accurately estimate the social graph of a small group despite the loss of large amounts of data.

6.2.4. Classification of Social Ties

Although we have seen that the existence of social ties can be inferred from communications metadata this inference gives no indication as to the current state of the tie. Neighbourhood overlap is considered an indicator of tie strength and in large mobile networks there is a linear correlation between the average neighbourhood overlap and the aggregated duration of calls. This allows aggregated call duration to be used as a proxy for tie strength.

This approach to estimating tie strength was tested on the Reality Mining and Nodobo datasets—both of which have fewer than one hundred participants—and found that a loose linear correlation between aggregated call duration and neighbourhood overlap is seen in the Reality Mining data but not in the Nodobo data. The number of calls were also tested and gave similar results. SMS message metadata from the Nodobo dataset was also tested and again no increase in overlap as aggregated SMS length or SMS count increased.

These results suggest that the techniques used to estimate tie strength in large networks cannot be applied to smaller networks, however testing tie strength proxies for weaker and stronger ties observed by researchers working with the Nodobo study participants did show larger values for all tie strength proxies for stronger ties suggesting that all mediated communications are in fact examples of strong ties.

6.3. Future Work

Although it is possible to infer the existence of social ties using communications metadata gathered from mobile devices, this work is only a beginning. Further to the method proposed here many additional steps should be taken in the future.

Scale The study conducted as part of this work was on a small scale. Although the results were valuable, if the work was to be repeated on a larger scale assistance from OEMs, or network operators would be required. OEMs could provide APIs on mobile device operating systems which researchers could utilise, thus making the process of gathering social data an exercise in application development. Network providers could provide datasets for research purposes. Although such datasets would only contain mediated interactions between subscribers on a single network they would be significantly larger than any dataset gathered by individual deployments of custom devices.

Co-proximity Any data provided by network providers would not contain co-proximity interactions. It may be possible to infer the co-proximity of subscribers from cell tower IDs, WiFi SSIDs, or location systems such as GPS and one aspect of future work should investigate this, but the use of Bluetooth to detect co-proximity is also not perfect. Future work into new methods of detecting co-proximate mobile devices using radio, or perhaps other mediums such as ultrasound, is required.

Tie Strength The ability to classify detected ties is important. Further work into the differences between the strong ties in the Nodobo data should be carried out. However, as the main purpose of the Nodobo study was to determine the existence of social ties from communications metadata, additional studies may be required to investigate tie strength thoroughly. Further investigation of alternative tie strength metrics should be a priority. Weak ties should also be investigated more thoroughly but this may not be possible without improvements in the ability of mobile devices to detect co-proximate devices discussed above.

Dynamic Network Analysis The network analyses discussed in this work are static. That is they are conducted on a fixed dataset and produce a single social graph. In real-world situations social networks are dynamic: ties are formed and experience changes from one day to the next. The work methods for detecting ties and estimating tie strengths in this work should be extended to include the dynamic nature of social networks allowing the changes in social ties which take place over time to be studied.

Security and Privacy The need to ensure that personal information is not leaked to malicious outsiders by socially aware mobile devices is paramount. It may be assumed that users (and their devices) are trustworthy, but there is still a threat from external attackers attempting to harvest social network data. Rogue devices may attempt to illicitly gather social network data from surrounding users, or a malicious

insider may attempt to spoof social ties in order to connect with someone who is not an acquaintance for example. These and other security issues must be addressed to ensure that in any deployment of a socially aware system users are properly authenticated and that sensitive information is stored and shared with the appropriate level of security.

Prototyping and Deployment Lastly, future work should consider the prototyping and deployment of socially aware mobile devices. By attempting to solve real-world problems in an organisation by social network analysis of mobile communications data much can be learned about the ideas proposed in this thesis, and many of the previous suggestions of areas for further work will be addressed simultaneously. The host institution may be an enterprise, care facility, school, or university but each deployment will allow more data to be gathered and studied. Moreover aspects of the scale of the deployment, co-proximity detection, tie strength, temporal changes in the social graph, and security and privacy will be addressed in each deployment making the lessons learned in these deployments extremely valuable to researchers and academics in this field as well as to the stakeholders in the host institutions themselves.

Bibliography

[1] Linton C Freeman. *Dynamic Social Network Modelling and Analysis: Workshop Summary and Papers*, chapter Finding Social Groups: A Meta-Analysis of the Southern Women Data, pages 39–50. The National Academies Press, 2003.

[2] Nathan Eagle. *Machine Perception and Learning of Complex Social Systems*. PhD thesis, Massachusetts School of Technology, May 2005.

[3] Johan Himberg, Kalle Korpiaho, Heikki Mannila, and Hannu Toivonen Johanna Tikanmäki. Time Series Segmentation for Context Recognition in Mobile Devices. In *Proceedings of the IEEE International Conference on Data Mining*, pages 203–210, 2001.

[4] Mark Zuckerberg. 500 Million Stories [online]. July 2010. Available from: `http://www.facebook.com/blog.php?post=409753352130` [cited 2011-09-12].

[5] Jeff Weiner. 100 Milion Members and Counting [online]. March 2010. Available from: `http://blog.linkedin.com/2011/03/22/linkedin-100-million/` [cited 2001-09-12].

[6] News Article. Primates on Facebook. *The Economist*, February 2009.

[7] M. Atkins, D. Recordon, Six Apart, C. Messina, DiSo Project, M. Keller, MySpace, A. Steinberg, and Facebook. Atom Activity Extensions. Internet-Draft atomactivity-00, Internet Engineering Task Force, 2009. Work in progress. Available from: `http://martin.atkins.me.uk/specs/activitystreams/atomactivity`.

[8] Dana Boyd. Friends, Friendsters, and MySpace Top 8: Writing Community Into Being on Social Network Sites. *First Monday*, 11(12), December 2006.

[9] H. Russel Bernard, Peter Kilworth, David Kronenfeld, and Lee Sailer. The Problem of Informant Accuracy. *Annual Review of Anthropology*, 13:495–517, 1984.

[10] Nathan Eagle and Alex Pentland. Reality Mining: Sensing Complex Social Systems. *Personal Ubiquitous Computing*, 10(4):255–268, 2006.

[11] James Irvine, Alisdair McDiarmid, Craig Saunders, Allan Tomlinson, and Nigel Jefferies. Instant Knowledge: Secure Autonomous Business Collaboration. In *Wireless World Research Forum Meeting 20*. Wireless World Research Forum, 2008.

[12] [online]Available from: `http://www.mobilevce.com/instant-knowledge` [cited 20th February 2012].

[13] Erving Goffman. *Relations in Public: Microstudies of the Public Order*. Penguin, 1972.

[14] Simon Johnson Williams. Appraising Goffman. *The British Journal of Scoiology*, 37(3):348–369, 1986.

[15] Rich Ling. *New Tech, New Ties*. MIT Press, 2008.

[16] Christian Licoppe. Connected Presence: The Emergence of a New Repertoire for Managing Social Relationships in a Changing Communications Technoscape. *Environment and Planning D: Society and Space*, 22(1):135–156, 2004.

[17] Mizuko Ito. *Mobile Communications*, chapter Mobile Phones, Japanese Youth, and the Replacement of Social Contact. Springer, 2005.

[18] Marc A. Smith. *The Hyperlinked Society: Questioning Connections in the Digital Age*, chapter From Hyperlinks to Hyperties, pages 165–176. digitalculturebooks, 2009.

[19] Thomas J. Allen. Architecture and Communication Among Product Development Engineers. Working Papers 165–97, Massachusetts Institute of Technology (MIT), Sloan School of Management, 1997.

[20] Nancy K. Baym, Yan Bing Zhang, and Mei-Chen Lin. Social Interactions Across Media: Internet, Telephone, and Face-to-Face. *New Media and Society*, 6(3):299–318, 2004.

[21] Shanyang Zhao. Toward a Taxonomy of Co-Presence. *Presence*, 12(5):445–455, 2003.

[22] Erving Goffman. *Behaviour in Puplic Places*. The Free Press, 1966.

[23] Jamie Lawrence, Terry R. Payne, and David De Roure. Co-Presence Communities: Using Pervasive Computing to Support Weak Social Networks.

In *WETICE '06: Proceedings of the 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*, pages 149–156, 2006.

[24] Edward T. Hall. *The Hidden Dimension*. Anchor Books, 1966.

[25] Lars Eric Holmquist, Jennica Falk, and Joakim Wigstrom. Supporting Group Collaboration with Interpersonal Awareness Devices. *Journal of Personal Technologies*, 3(1-2):105–124, 1999.

[26] Eric Paulos and Elizabeth Goodman. The Familiar Stranger: Anxiety, Comfort, and Play in Public Places. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 223–230. ACM, 2004.

[27] Richard Borovoy, Fred Martin, Sunil Vermuri, and Brian Silverman. GroupWear: Nametags that Tell about Relationships. In *CHI '98 Summary*, pages 329–330. ACM Press, 1998.

[28] Richard Borovoy, Michelle McDonald, Fred Martin, and Mitchel Resnick. Things That Blink: Computationaly Augmented Name Tags. *IBM Systems Journal*, 35(3 & 4):488–495, 1996.

[29] Stanley Milgram. *The Individual in a Social World*, chapter The Familiar Stranger: An Aspect of Urban Anomimity, pages 68–71. McGraw Hill, 1992.

[30] R Want, A Hopper, V Falcao, and J Gibbons. The Active Badge Location System. *ACM Transactions on Information Systems*, 10(1):91–102, 1992.

[31] Tanzeem Choudhury and Alex Pentland. Characterizing Social Networks using the Sociometer. In *Proceedings of NAACOS 2004*, 2004.

[32] Tanzeem Choudhury and Alex Pentland. Sensing and Modeling Human Networks using the Sociometer. In *Proceedings of 7th IEEE International Symposium on Wearable Computers (ISWC'05)*, pages 216–222. IEEE, 2005.

[33] Tanzeem Choudhury, Sunny Consolvo, Beverly Harrison, Jeffery Hightower, Anthony LaMarca, Louis LeGrand, Ali Rahimi, Adam Rea, Gaetano Bordello, Bruce Hemingway, Predrag Klasnja, Karl Koscher, James A. Landay, Jonathan Lester, Danny Wyatt, and Dirk Haehnel. The Mobile Sensing Platform: An Embedded Activity Recognition System. *IEEE Pervasive Computing*, 7(2):32–41, 2008.

[34] Rich DeVaul, Michael Sung, Jonathan Gips, and Alex "Sandy" Pentland. MIThril 2003: Applications and Architecture. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers*, ISWC '03, pages 4–11, 2003.

[35] Anmol Madan and Alexander Pentland. VibeFones: Socially Aware Mobile Phones. In *10th IEEE International Symposium on Wearable Computers*, pages 109 –112, 2006.

[36] Mika Raento, Antti Oulasvirta, Renaud Petit, and Hannu Toivonen. ContextPhone: a Prototyping Platform for Context-aware Mobile Applications. *IEEE Pervasive Computing*, 4(2):51–59, 2005.

[37] Roy Want. You Are Your Cell Phone. *IEEE Pervasive Computing*, 7(2):2–4, 2008.

[38] Shwetak N. Patel, Julie A. Kientz, Gillian R. Hayes, Sooraj Bhat, and Gregory D. Abowd. Farther Than You May Think: An Impirical Investigation of the Proximity of Users to Their Mobile Phones. In P Dourish and A Friday, editors, *In Proceedings of UbiComp 2006*, pages 123–140, 2006.

[39] Riku Mettala. Bluetooth Protocol Architecture. White Paper, August 1999.

[40] Anil Madhavapeddy and Alastair Tse. A Study of Bluetooth Propagation Using Accurate Indoor Location Mapping. In *Seventh International Conference on Ubiquitous Computing*, volume 3660 of *Lecture Notes on Computer Science*, pages 105–122. Springer, August 2005.

[41] Joonas Kukkonen, Eemil Lagerspetz, Petteri Nurmi, and Mikael Andersson. BeTelGeuse: A Platform for Gathering and Processing Situational Data. *Pervasive Computing*, 09:49–56, 2009.

[42] Nathan Eagle, Alex Pentland, and David Lazer. Inferring Social Network Structure using Mobile Phone Data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[43] Paul Watzlawick, Janet Beavin Bavelas, and Don D Jackson. *Pragmatics of Human Communication*. W W Norton and Co. Inc, 1962.

[44] Gueorgi Kossinets and Duncan J. Watts. Empirical Analysis of an Evolving Social Network. *Science*, 311:88–90, January 2006.

[45] Jean-Pierre Eckmann, Elisha Moses, and Danilo Sergi. Entropy of Dialogues Creates Coherent Structures in Email Traffic. *Proceedings of the National Academy of Sciences*, 101(40):14333–14337, 2004.

[46] Jukka-Pekka Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and Albert-László Barabási. Structure and Tie Strengths in Mobile Communication Networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, May 2007.

[47] Gergely Palla, Albert-László Barabási, and Tamás Vicsek. Quantifying Social Group Evolution. *Nature*, 446:664–667, April 2007.

[48] Ditte Laursen. *The Inside Text: Social, Cultural and Design Perspectives on SMS*, volume 4 of *The Kluwer International Series on Computer Supported Cooperative Work*, chapter Please Reply! The Replying Norm in Adolescent SMS Communication, pages 53–73. Springer Netherlands, 2006.

[49] Wouter de Nooy, Andrej Mrvar, and Vladimir Batagelj. *Exploratory Social Network Analysis with Pajek*, chapter Looking for Social Structure, page 5. Cambridge University Press, 2005.

[50] John Scott. *Social Network Analysis: A Handbook*, chapter Networks and Relations, pages 2–3. SAGE Publications Ltd, 1991.

[51] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*, chapter Social Network Data: Collection and Applications, pages 37–38. Cambridge University Press, 1994.

[52] Dawn Iacobucci. *Social Network Analysis: Methods and Applications*, chapter Graphs and Matrices, pages 93–94. Cambridge University Press, 1994.

[53] Frank Harary. *Graph Theory*, chapter Graphs, pages 9–10. Number 2. Addison-Wesley, 1969.

[54] Mark E. J. Newman. The Structure and Function of Complex Networks. *Society of Industrial and Applied Mathematics*, 45(2):167–256, 2003.

[55] Dawn Iacobucci. *Social Network Analysis: Methods and Applications*, chapter Graphs and Matrices, pages 145–146. Cambridge University Press, 1994.

[56] Dawn Iacobucci. *Social Network Analysis: Methods and Applications*, chapter Graphs and Matrices, pages 140–141. Cambridge University Press, 1994.

[57] Dawn Iacobucci. *Social Network Analysis: Methods and Applications*, chapter Graphs and Matrices, pages 121–122. Cambridge University Press, 1994.

[58] Dawn Iacobucci. *Social Network Analysis: Methods and Applications*, chapter Graphs and Matrices, pages 99–100. Cambridge University Press, 1994.

[59] Dawn Iacobucci. *Social Network Analysis: Methods and Applications*, chapter Graphs and Matrices, page 124. Cambridge University Press, 1994.

[60] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*, chapter Triads, pages 559–564. Cambridge University Press, 1994.

[61] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*, chapter Social Network Data: Collection and Applications, page 42. Cambridge University Press, 1994.

[62] Whitfield Diffie and Susan Landau. *Whit Privacy on the Line: The Politics of Wiretapping and Encryption.*, chapter Cryptography, page 44. MIT Press, 2007.

[63] Mark Weiser. Ubiqitous Computing. In *IEEE Computer Hot Topics*, 1993.

[64] Ofcom. The International Communications Market 2008. Research Document, November 2008.

[65] Ofcom. Communications Market Report. Research Document, August 2008.

[66] Daniel Salber, Anind K. Dey, and Gregory D. Abowd. The Context Toolkit: Aiding the Development of Context-Enabled Appliactions. In *CHI '99: Conference of Human Factors in Computing Systems*, pages 434–441. ACM, May 1999.

[67] Avinash Srinivasan, Joshua Teitelbaum, Huigang Liang, Jie Wu, and Mihaela Cardei. *In Algorithms and Protocols for Wireless Ad-Hoc and Sensor Networks*, chapter Reputation and Trust-based Systems for Ad Hoc and Sensor Networks. Wiley & Sons, 2008.

[68] Micheal Joseph Lambert. Visualising and Analysing Human-Centered Data Streams. Master's thesis, Massachusetts Institute of Technology, May 2005.

[69] Stephen Bell, Alisdair McDiarmid, James Irvine, and Nigel Jeffries. Usability Evaluation with Minimal Observation Impact. In *Wireless World Research Forum Meeting 23*. Wireless World Research Forum, 2009.

[70] Stephen Bell. *Towards an End-to-End Solution for Performing Usability Evaluation for Self-recording Devices*. Submitted for examination, University of Strathclyde, 2011.

[71] Stephen Bell, Alisdair McDiarmid, and James Irvine. Nodobo: Mobile Phones as a Software Sensor for Social Network Research. In *Proceedings 73rd IEEE Vehicular Technology Conference*, 2011.

[72] John Scott. *Social Network Analysis: A Handbook*, chapter Networks and Relations, page 10. SAGE Publications Ltd, 1991.

[73] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*, chapter Social Network Data: Collection and Applications, page 17. Cambridge University Press, 1994.

[74] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*, chapter Social Network Data: Collection and Applications, pages 45–51. Cambridge University Press, 1994.

[75] Eric Gilbert and Karrie Karahalios. Predicting Tie Strength with Social Media. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 211–220, 2009.

[76] Mark S. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360 – 1380, 1973.

[77] David Krackhardt. *Networks and Organisations: Structure Form and Actions*, chapter The Strength of Strong Ties: The Importance of *Philos* in Organisations, pages 216–238. Harvard Business School Press, 1992.

[78] David Krackhardt and N Stern, Robert. Informal Networks and Organizational Crises: An Experimental Simulation. *Social Psychology Quarterly*, 51(2):123–140, 1988.

[79] Morten T. Hansen. The Search-Transfer: The Role of Weak Ties in Sharing Knowledge across Organization Subunits. *Administrative Science Quarterly*, 44(1):82–111, 1999.

[80] Peter V. Marsden and Karen E. Campbell. Measuring tie strength. *Social Forces*, 63(2):482–501, 1984.

[81] Andrea Petroczi, Tamas Nepusz, and Fulop Bazso. Measuring Tie-Strength in Virtual Social Networks. *Connections*, 27(2):39–52, 2007. Available from: `http://eprints.kingston.ac.uk/2396/`.

[82] Anatol Rapoport. Spread of Information Through a Population with Socio-Structural Bias I: Assumption of Transitivity. *Bulletin of Mathematical Bio-physics*, 15(4):523–533, 1953.

[83] David Easly and Jon Kleinberg. *Networks, Crowds, and Markets:Reasoning about a Highly Connected World*, chapter Strong and Weak Ties, pages 52–53. Cambridge University Press, 2010.

A. Details of Confirmed Ties

Some details of the interactions between the ends of the forty-eight estimated ties confirmed by the reported social graph are given in Table A.1. For each tie a count of the voice calls, SMS messages, and proximity interactions is given.

Table A.1.
Details of the 48 Estimated Ties Confirmed by the Reported Graph.

Tie	Call Count	SMS Count	Proximity Count
(1,6)	5	241	10,730
(1,11)	0	40	7,422
(1,15)	15	199	32,168
(1,27)	0	114	9,915
(3,11)	6	38	10,998
(3,23)	0	0	2,258
(3,27)	0	86	9,333
(4,7)	378	6,002	6,973
(4,11)	0	5	866
(4,13)	3	7	2,253
(4,17)	3	198	1,193
(4,26)	0	0	2,048
(5,24)	1	3	620
(6,11)	1	20	8,680
(6,15)	3	17	20,371
(6,17)	0	5	10,050
(6,27)	0	54	13,344
(7,13)	1	13	13,695
(7,23)	1	1	3,773
(7,27)	97	239	18,113
(8,10)	2	79	841
(8,11)	0	2	7,587
(8,13)	30	36	9,108
(8,15)	0	2	15,286
(8,23)	120	472	7,609

(Continued overleaf)

Tie	Call Count	SMS Count	Proximity Count
(8,27)	10	38	7,343
(11,17)	1	67	6,311
(11,18)	0	243	2,639
(11,23)	1	103	3,864
(11,26)	4	20	13,540
(11,27)	10	856	15,999
(12,13)	0	7	1,958
(13,23)	71	328	5,142
(13,26)	204	10,439	36,704
(13,27)	4	134	18,959
(14,19)	5	237	5,750
(14,21)	0	10	3,559
(14,25)	10	291	434
(15,27)	0	1	18,906
(16,23)	0	0	1,228
(17,18)	0	88	2,710
(19,21)	285	4,616	23,712
(19,25)	2	16	185
(21,23)	0	0	2,130
(21,25)	12	114	345
(23,26)	0	1	3,148
(23,27)	0	104	4,450
(26,27)	5	339	21,997