A Proposed Framework for Crowd-Sourced Social Network Data Collected over Bluetooth

By

Julian Benavides

A thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Department of Electrical & Computer Engineering
University of Manitoba
Winnipeg, Canada

Abstract

Currently, mobile computing is mandating or influencing the direction of new developments in information technology. The high level of adoption that mobile devices have among individuals allows for multiple opportunities for new developments applicable to academic communities, governments and businesses. Data of various types can be collected in a crowd-sourced manner. As such, this thesis examines the collection and application of data collected through a purpose-designed app relying on Bluetooth and geo-location technologies on mobile devices. Through three distinct development iterations and using Bluetooth connectivity, information about connectivity to other mobile devices can be obtained, and in this way the number, type, and device names of "connecting" devices are gathered and stored. Another interesting aspect associated with this type of data collection is that the mobile device may be either moving or stationary during the data collection process. Information can be collected and mined to help map real-life events such as traffic patterns or crowd movement within mass gatherings, as well as ethereal social interactions, and these data can in turn be used as input to various models and simulators. When geo-location technologies are incorporated, a higher level of detail can be obtained on the location of devices. This technology allows for mapping movement and contacts made between people, allowing for the gathering of more detailed social patterns of individuals. As part of this study, the technology developed using Bluetooth connectivity and geo-location is then taken to an additional iteration to develop a mobile system that is able to find and establish direct connections with other individuals and initiate real-life interactions. The work demonstrates that mobile technologies can provide a broad framework of action for the generation and collection of valuable data that can be used for behavioural studies, simulations and other type of research that involves real-life social interactions.

Acknowledgements

I would like to thank my advisor, Professor Bob McLeod. He has been a real mentor, a great advisor and a continuous guide. I would also like to thank my co-advisor Dr. Marcia Friesen for her help during my courses and through my thesis writing process. I must also thank my colleague and friend B.C.P Demianyk for all his help, and advice during the research phase of my masters and for his valuable collaboration during the project development.

Special thanks to my examining committee for their time and efforts reviewing this thesis.

Dedication

I would like to dedicate this effort to

- *My lovely wife Viviana, for her patience and continuous encouragement to finish my graduate studies and for her never-ending love.*
- *Martin who is my gift from god, who always gives me a reason to be a better dad.*
- *My parents who taught me the importance of knowledge and who always supported my undergraduate and graduate studies.*

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1. Overview

This section provides a general explanation of basic issues involved in this research project. The main objective of the project is to design and implement a technology capable of gathering massive amounts of crowd sourced data through the development of a Smartphone centric platform for data collection and a data server for posterior data modeling and analysis. The modeled data could be later used to predict and simulate human behavior during events such as vehicular traffic jams, epidemics and other activities that involve human interaction. This overview begins with a descriptive section defining the main motivation for the research investigation in more detail. The second section presents the main goals for the project as defined when it was first envisioned. The third section of this overview introduces and discusses other research related to this project, followed by a brief summary of the introduction.

Chapter 2 presents a general description of the concepts used during the research and development phases of this study, including how they were used and their rationale.

The following section provides, a detailed description of the technology used for data gathering, processing and data storing along with some architectural notions. Chapter 3 describes in detail the first iteration done for the mobile application developed for this project, denoted the Bluetooth Agent Tracker (BAT), along with the context in which it was deployed and its potential uses. In Chapter 4, the second iteration of the mobile application is presented (Face2Face) describing the changes made, what was expected for this development and several samples of data gathered along with the analysis applied to it. Chapter 5 describes the final iteration of the mobile application prototype (Friend Finder), the objectives of this design and some analysis of non-trivial usage of the data that could be gathered using the application. Chapters 6 and 7 provide some detail how the data modeling and data analysis were done and proposes some future ideas and general conclusions for the research.

1.2. Motivation

The increased levels of adoption and market penetration that Smartphone devices have experienced over the past seven years make the leverage of Smartphone usage throughmobile computing applications an interesting and relevant topic of research. What makes it even more relevant is the fact that Smartphone and handheld device proliferation is expected to further increase [**1**]. Nonetheless, it is also clear that the specific issues related to data in general and Smartphone data specifically is a complex one [**2**]. The increasing availability of handheld devices coupled with the amount of sensors embedded within those devices and the increasing number of applications built for them create high levels of complexity in performing focused research on this topic. Scientific enquiry on this issue has commonly been carried out within corporations or large academic institutions with high level of access to different data [**3**] and resources [**4**]; for that reason, developing academic research that yields relevant conclusions on this topic certainly presents some challenges. For example, as expressed in [**5**], the proliferation *itself* of Smartphones and handheld devices has created challenges for researchers. These authors find that the unprecedented rate at which Smartphone demand and usage has grown has led to a proliferation of the types of devices and platforms available in the market. They concede that for that reason, with a multitude of devices in use by the public and dozens of different software platforms, realizing what an ideal platform is, or researching Smartphone solutions that are valuable, are complex issues that cannot easily be resolved. These types of challenges fuel considerable motivation in performing research that is relevant, useful, and whose conclusions will have impacts on significant as well as popular issues.

When exploring the current state of mobile networking data collection technologies one arrives at the realization that research should have a well-defined structure, which involves gathering information on people's regular behavior in everyday life, and then using this information to model and predict future behavior based on the patterns encountered through data collection. This naturally would lead the research toward the analysis of both individual and crowd behaviour. Finding patterns of behaviour through analysis and pattern-recognition of large amounts of data is a valid form of knowledge generation because enormous amounts of detail and evidence can be found from relationships and patterns in present in the data [**5**]. As such, the ordinary, everyday usage of wireless technologies for common purposes such as voice calling and SMS, can lead to significant amounts of knowledge generation through the collection of information regarding the context within which Smartphones are used or the environment in which they are located, from which that data then becomes available for analyses and other applications. These data lead to information that could be interpreted as a model of human interactions including, but not limited to proximity, short and long term contact or interactions, and geographic commuting behaviors [**6**].

Perhaps the most important consideration when performing this research is that it makes a direct contribution – through knowledge generation - on improving societal conditions. The technology deployed, the software developed, the data gathered and the analysis made through studies on Smartphone usage can help improve people's living conditions by modeling solutions related to relevant aspects such as disease epidemics or pandemics. As another example, using Smartphone data to model of how automobile traffic flows during peak hours can also help improve road design or traffic strategies. As a result, a wide range of daily activities that involve human interaction and present potential large-scale challenges fuel this research and are the main motivation to start a project with the characteristics described within this thesis.

1.3. Project Goals

The main goal of this thesis project is to design and develop a set of mobile applications that allow crowd sourced data gathering and processing for follow-on modeling and analysis. Within this goal, specific objectives to enrich and refine the project were:

1. To design and develop a Smartphone application that will use the wireless capabilities of the device to gather a defined set of information about an individual and their surroundings.
2. To design and deploy a server side infrastructure to handle large amounts of data and its processing. This includes the development of interfaces for remote/mobile data aggregation, end-to-end data transformation and all the functionalities related to server handling processes.
3. To design a software enhancement methodology to allow other developers to reuse and improve the platform and applications created during the research.
4. To ensure a scalable design of the server and mobile software developed during the research.

5. Ensure security of data transfer processes between the mobile devices and the servers and the anonymity of all the users/devices used during the data gathering phase.

All the goals proposed above are intimately related to the desire of generating a useful set of tools, methodologies and infrastructures to analyze, model and transform data with the purpose of understanding the way individuals behave (act) and interact. The potential to contribute to topics such as pandemics, traffic data collection, and human interaction provide a tangible purpose to every goal of this research.

1.4. Related Work

Crowdsourcing is a data collection mechanism that has been in use for several years. This term was first coined by Jeff Howe in 2006 in Wired Magazine. An informal definition of crowdsourcing indicates that it describes a new Web-based business model that harnesses the creative solutions of a distributed network of individuals through what amounts to an open call for proposals or contributions of some form [**7**]. Howe indicates that as a term, crowdsourcing is adaptable, and is used for several different activities that take on various forms. This concept involves, in part, the notion of regular citizens participating in knowledge generation through regular everyday activities. Crowdsourcing is in some ways similar to open-source software development, where the value of the end result lies in a mostly anonymous crowd working toward a common end result. Knowledge generation through internet sites such as Wikipedia is another form of applying the crowd's contributions toward a valuable goal. Even for-profit organizations such as Threadless, iStockphoto, and InnoCentive currently use this model. Crowdsourcing is a popular model nowadays as it is considered a democratic, team-oriented process for distributed problem solving, which younger generations tend to prefer [**8**]; however, one of the most relevant characteristics of data generated through crowdsourcing is verifying the authenticity and validity of data generated through this model because of the potential scale involved. To generate crowd sourced data, a researcher collects data on peoples' regular behaviours where such data is provided by the people through their day-to-day activities rather than through a controlled, artificial data collection process, and the researcher uses such data to form patterns through the discovery of relationships among variables. This method of research became very popular over the last decade as the ubiquity of the internet and the availability of mobile technologies allowed data gathering of people's activities without interrupting their regular lives [**7**]. Crowdsourcing has faced a certain level of skepticism among the scientific community; however, as Sabou, Bontcheva, and Scharl state [**9**], this is more due to a lack of awareness and a certain bias in lessons-learned reports on projects that have used this method, which overshadow the major positive outcomes and opportunities that crowdsourcing brings. The popularity of crowdsourcing initiatives in science rises from the fact that not only does it outsource natural human work to gather data in a non-expensive manner as research agents do not have to perform additional tasks to get the job done, but also because it gives to any project the approach to a real life scenario that simulations or artificially generated data can rarely achieve [**10**]. There is a factor that essentially differentiates a single-source framework and a crowd sourced one, and it is the uncertainty of human erratic behavior. Crowdsourcing exploits valuable knowledge-generating factors such as the unpredictable behaviour of humans [**11**] through filtering. Collecting data using crowdsourcing strategies is currently a popular topic in literature especially as this methodology enhances data collection for natural and social scientific purposes. For that reason, crowdsourcing of data has generated considerable levels of attention among the academic community.

Studies that use of crowd sourced data to attempt to resolve societal issues and improve human conditions are becoming increasingly encountered in the literature as well. For example, Zook et al. [**12**] describes the ways in which crowdsourcing web-based mapping were used for Haiti Relief efforts during the earthquake of 2010. Using crowdsourcing, various agencies were able to map evacuees and focus relief efforts in a more efficient way. In addition, another piece of research on a similar topic, which yielded important information using crowdsourcing as a data collection method, was the work by Bengtsson et al., [**13**] who described how crowdsourcing enhanced relief efforts through mapping affected population movements after the Haiti earthquake. To this extent, as there are about 3.5 million mobile phone users in Haiti, the authors used data from the largest mobile telephony provider in Haiti to track population movement, and were able to relieve earthquake effects, and prevent or mitigate cholera outbreaks.

One outstanding example of a study that yielded results on network analysis was done by Choffnes, Bustamante, and Ge [**14**]. They presented crowdsourcing-type data collection to detect network use satisfaction from users of network service providers on mobile services. The authors presented what they called a crowd sourced data network collection as an alternative to traditional data collection methods from network service providers to detect network errors. The authors concede that crowdsourcing is a better methodology for data analysis due to its flexibility, scalability, quickness of data collection, and respect of agent privacy. They exhaustively researched the topic of network event monitoring using crowdsourcing as its data collection framework. Particularly, Choffnes, Bustamante, and Ge proposed the use of alternative/additional methods to enhance the way data is gathered and processed for the project's benefit.

Crowdsourcing as a valid data collection method has also been scrutinized. In the Kittur, Shi and Suh study [**15**], the quality of data collected through crowdsourcing was compared with traditional data collection methods such as surveys and laboratory studies. The authors concede that in addition to the data collected by crowdsourcing methods the use of additional sources of data to complement the gaps that crowd sourced data may have can enhance the accuracy of the project results. The work described a situation in which there are gaps of information at certain times due to factors such as inactivity, inexperience, or lack of interest, and how they filled those gaps by using additional sources of data and additional data gathering methods.

In addition, the work by Ekici, and Bozdag [**16**] provided insight into the concept of using wireless devices for non-transmission related tasks. Those authors studied and explained the idea behind using every day elements with wireless capabilities to gather data through the use of sensors, which maximize network lifetime, and would work even when connectivity is not available all the time. The authors provide an overview of different wireless sensor networks to collect data. This article had a significant influence on the present project to investigate the implementation of a similar infrastructure using a wireless protocol such as Bluetooth for the purposes of data collection. In addition, it is important to mention the work by Wiggins and Crowston [**17**], who studied citizen science as a form of research collaboration that involves members of the general public in scientific research projects. Their work suggests the correlation between human beings and electronic devices, which can be used for useful scientific data to build knowledge, which inspired the idea of individuals and their personal devices as digital representations of them, made possible by finding patterns of behaviour through monitoring of those devices.

This project also found inspiration on the work by Sultan [**18**]. Through investigating novel ways to gather crowd sourced data for traffic safety investigation, Sultan's work presents a similar approach to the one implemented in this project. Sultan proposed a new vehicular safety platform by integrating emergency context-aware management with mobile computing by manipulating wide spatial data using crowdsourcing. The platform used internet connectivity in Smartphones to use mobile telephone users as sources of vehicular safety context. Their work is related to the present project as the present study focuses on finding novel methods for data harvesting through crowdsourcing using wireless radio protocols to identify individuals in their surroundings for posterior analysis. Nonetheless, the most relevant difference between the work proposed in this thesis and the one made by the referenced author is the desire to create an end-to-end framework to gather, store, analyze, model and use real human interaction data for human related simulations. Also, this work provides a novel method of modeling real social behaviors such as proximity, contact duration and geographical location for input to Agent Based Model and their simulations.

Collection data about human behaviour using sensor networks is another topic that was researched for the current thesis project. Olguín and Pentland [**19**] propose a data collection method that uses sensors to detect social interactions in an automatic way. The device that the authors propose would be wearable, and as data is gathered and transmitted automatically without affecting a person’s regular life, large amounts of data can be stored quickly and accurately. In this way, activities such as physical proximity to other people, conversational dynamics, and physical activity levels can be described. They describe the advantage of using such a framework, undertake a few experiments with it, and conclude that using sensor networks where sensors can be carried by individuals can represent several advantages to scientists, as they enhance the quality of data captured, the uses of it, and minimize the effort on obtaining such high quality data.

1.5. Chapter Summary

This chapter provided an introductory view of the study. The motivation to perform this study was based on observations on different trends in wireless usage. Smartphone adoption is growing faster than adoption for any other technological device and leveraging this proliferation is an objective of the research. In addition, there was a realization that such popular usage of wireless technologies can lead to significant amounts of knowledge generation by collecting information regarding the context within which Smartphones are used. This can be used toward understanding of certain societal conditions as diverse as disease epidemics to pandemics and traffic monitoring, through the recognition of human proximity.

The work draws on several other studies found in literature. Related studies include projects on data gathering using crowdsourcing strategies, collecting data using sensor networks, measuring proximity using wireless technologies, and Smartphone application usage. Related studies provided a foundation for this research in both a theoretical and a practical way. In addition, they represented benchmarks upon which this project was developed.

CHAPTER 2

Background Concepts

Providing some information regarding topics related to the study is essential to the understanding of it. For that reason, this section introduces a notion of different situations, and concepts used to structure the research that gave way to this thesis. Concepts related to the human-technology interaction scenarios are illustrated as the setting in which the project was intended to be deployed. In addition, the technology behind each of the interactions, the proposed system model, and the desired data output are also presented. Conceptual explanations based on current literature on the topic are also described in this section.

2.1. Project Universe

Figure 1 illustrates one of the human-technology interaction scenarios in which the project was intended to be deployed and the proposed system model. In this figure, different receivers and transmitter setups are shown at different locations; those interconnected devices describe a unique possibility for interaction and data harvesting, which was one of the main objectives for this work. Individuals performing regular everyday activities generate an indirect interaction to one another without even knowing, and based on the fact that a vast majority of the population have/use a Smartphone, it is acceptable to assume their location could be registered in real time adding rich geographical location information to the data gathered as most of the devices described as Smartphone include GPS capabilities or other means of localization. Data collected could include shopping trends, public transit usage, and work commuting or any other regular activity even as the user travels. The Smartphone device could be used as an “agent” to track and report the information to a central entity for analysis and posterior data modeling.

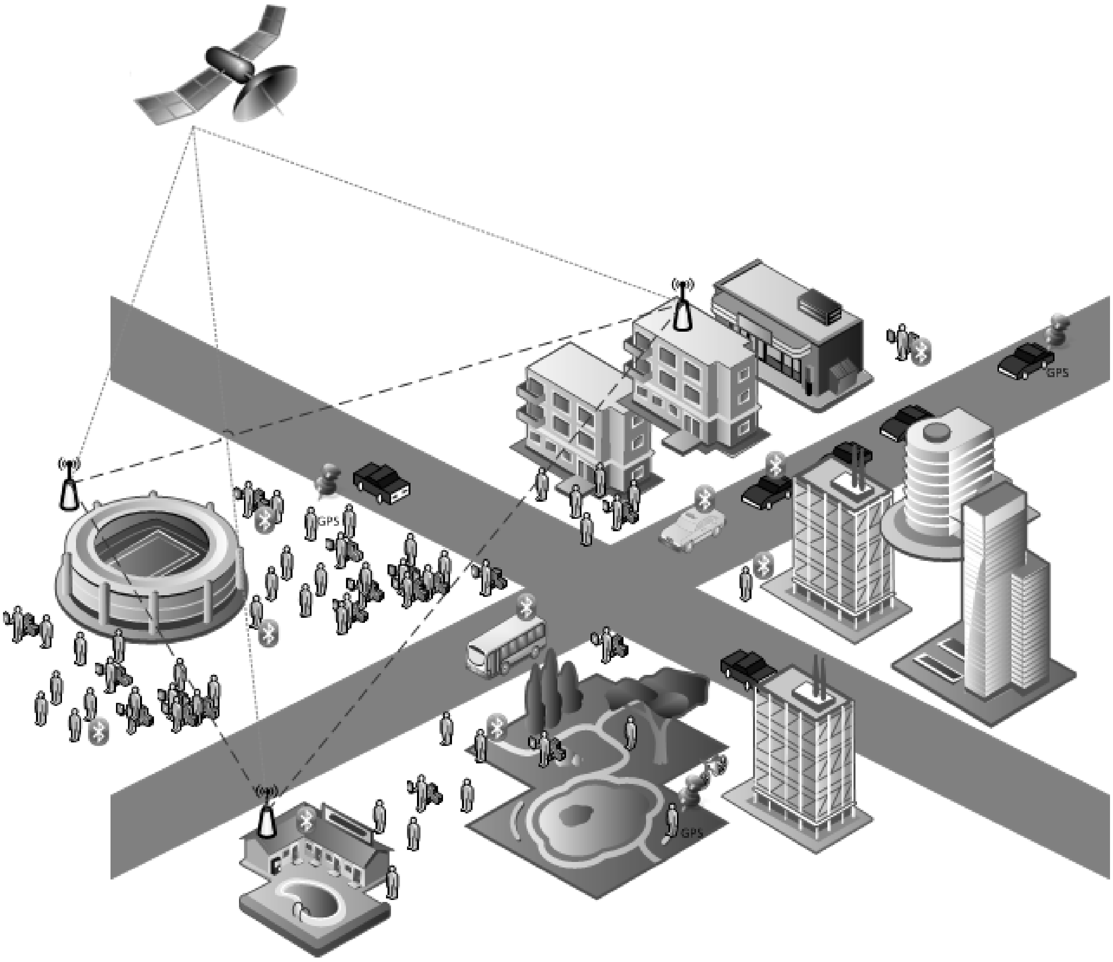


Figure 1 - General Scenario

Trying to understand ordinary human interactions such as peer-to-peer contact and proximity duration at public places is an essential issue to consider as a basis for the present study. Using data transmitted over sensor networks is the most valid and viable option to attempt to understand and model this. As Oliver [**20**] states: “as the user base of Smartphones continues to grow, the study of the relationship that exists between humans and their personal handheld devices can provide valuable insights and benefits to the research area.” Networking technologies, the ubiquity of the internet, and the vast popularity of handheld devices that are connected to the internet make it possible to address this problem using a crowdsourcing approach. As such, Smartphones being connected devices able to represent a human being in the digital spectrum, can be used as a point of reference for each individual that carries one; this represents a significant opportunity to gather information about individuals in their natural environment.

Research based on Smartphone computing has demonstrated that Smartphone devices have become the most used digital gadget by humans not only based on the population density measurement but also on the amount of time a person uses the Smartphone device during the day [**21**]. The fact that people constantly remain in close proximity with their Smartphone device even when they are in places such as a bathroom, a bedroom, a bus, or an office, or during times when they are moving from one location to another one, allows for researchers to presume a solid correlation between an individual’s location and that of their Smartphone device. Therefore, performing research on a human’s location based on the location of their Smartphone has a high level of validity. Chen and Kotz [**22**] analyze the concept of contextual information, which represents information about the environment around the situation when the Smartphone is being used. This includes a user location, time of day, nearby people and devices that can be gathered during the research process. The authors then concede that having mobile applications that are context-aware, that is, applications that are able to act to changes in contextual environment; can certainly enhance any project that involves collecting data from those applications. Contextual information is what will be gathered and used during the present project coupled with user activity within that context.

There are, however, further considerations that must be taken into account when trying to collect usable data for research purposes. The infrastructure needed to handle massive amounts of data collected is something that should be considered beforehand. Gathering data from multiple sources is not a simple task to develop and manage so platform deployment scope is something that should be considered. Transmission protocols were the base of this research because without them it would be impossible to determine or transmit anything for further analysis; understanding the protocol used during the research became imperative in order to maximize the way they were used.

2.2. Backend Infrastructure

Figure 2 illustrates the cumbersome process of data gathering, transmission, aggregation, processing and storage that was conceived when the project guidelines were defined; a proposed backend infrastructure was selected for deployment as it combined three key factors that were mandatory: It was extendible, the development and implementation was feasible in a short period of time and it was low-cost (it did not require licences or special conditions). The project objectives included maximizing the potential use of different communication protocols within a defined device to gather valuable data with minimal or no interaction from the user side; this presented a challenge in the design phase that will be described in more detail in the next chapter.

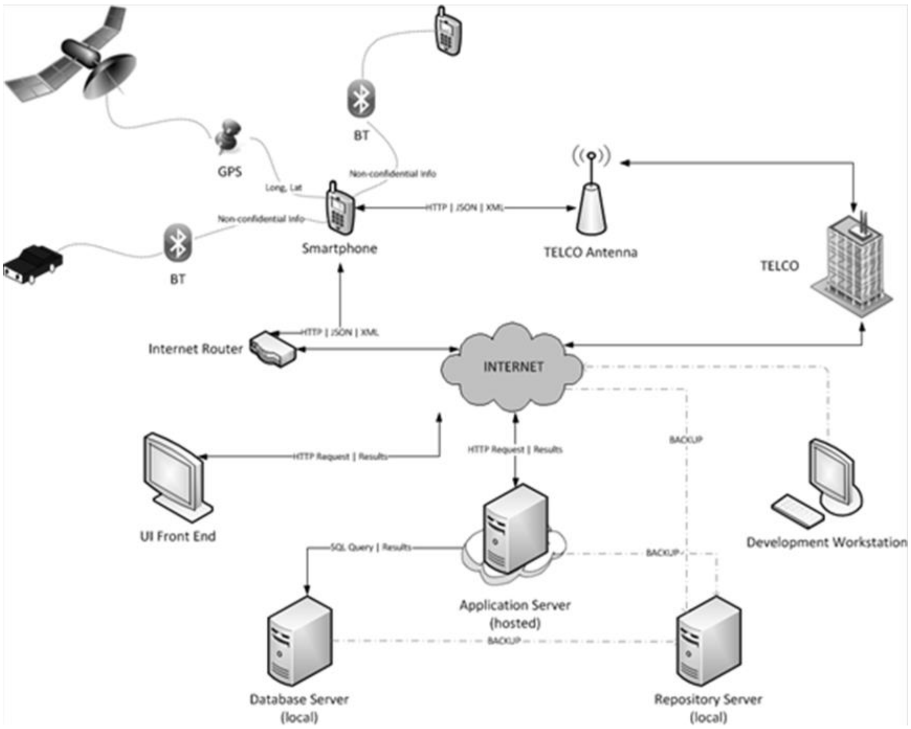


Figure 2 - Infrastructure and Communications

A crowd sourced project immediately implies having considered a strong, scalable infrastructure – both at a hardware and software level - to hold and process the amounts of information that could be generated at any given time. This section defines and explains the type of technology, communications, processes and all the components used to make the posterior data collection and analysis a smooth task.

As Figure 2 clearly describes, there are several individuals interacting in the data building process and along the interaction cycle, well known communication protocols are utilized; those protocols are categorized and explained in the following section.

2.3. Transmission Protocols

While trying to find novel ways to use a specific set of technologies, a researcher must be in constant interaction with them; he or she needs to be able to understand them, be aware of their limitations and take advantage of the small functionalities and capabilities other researches may have overlooked or have taken for granted. Many times those technologies are complemented with others, making a project flow in the right direction. This project includes the use of several technology structures, with communication protocols being a very important aspect of this project's success.

Figure 3 shows the general idea behind data collection using mobile devices, in a simplified manner. Using GPS and assisted GPS for geo-location information and cellular (CDMA,GSM,GPRS) and Wi-Fi for data transmission allows this project to exploit Bluetooth and other short range protocols to harvest data using Meta information to infer human interactions such as proximity and estimated contact duration, augmenting more familiar location based data.

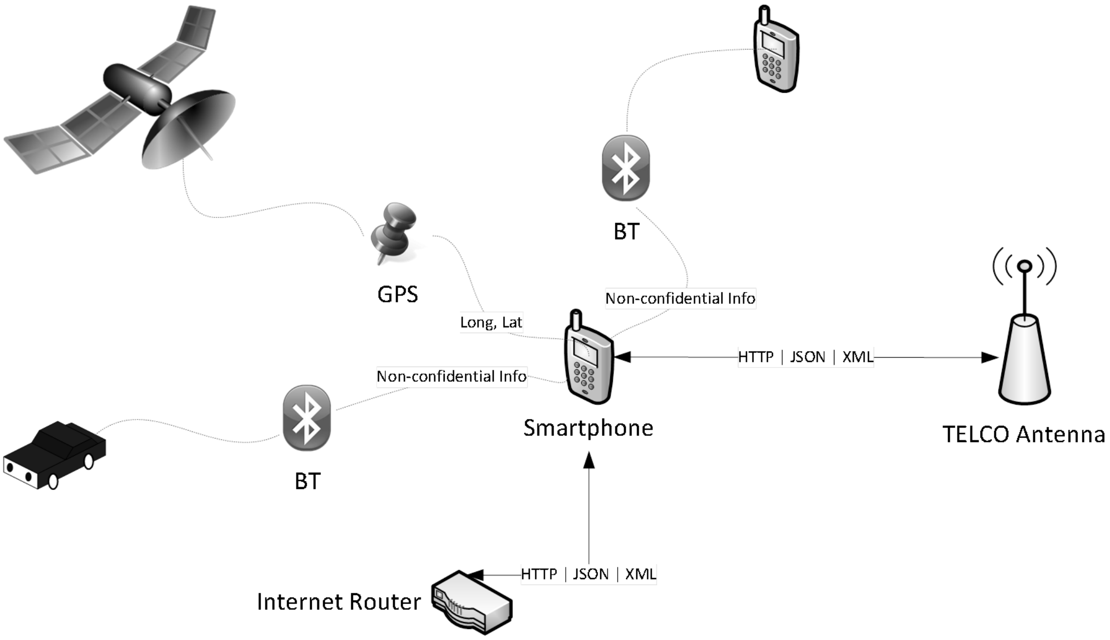


Figure 3 Transmission Protocols Interaction Diagram

Table 1 describes how some of the transmission protocols selected to implement this project were used by defining three categories of use: transmission, location and interaction.

The Transmission category holds the protocols that are responsible for data exchange from the devices gathering information to the central processing entity (server); it was decided to use both Wi-Fi and cellular for transmission because the vast majority of mobile devices defined as "Smartphones" include both technologies and knowing that most of the time their use overlaps (cellular is used if and only if Wi-Fi connectivity for data transmission is unavailable). This gave a certain level of continuity and system efficiency to the data gathering process.

Wi-Fi is a technology that allows an electronic device to exchange data wirelessly (using radio waves) over a computer network.; it is also defined as any wireless local area network products that are based on (IEEE) 802.11 standards [**23**]. Cellular technology such as Code division multiple access (CDMA) is a channel access method used by various radio communication technologies for data and voice transmission over cellular [**24**].

For the location category, GPS and A-GPS were the technologies selected for mobile device positioning discovery as these two technologies are widely used by all the major Smartphone manufacturers. They provide the project with a good accuracy level for the purposes that were required.

The Global Positioning System (GPS) is a space-based satellite navigation system that provides location and time information in all weather conditions, anywhere on or near the Earth where there is an unobstructed line of sight to four or more GPS satellites. Assisted GPS, generally abbreviated as A-GPS, is a system that can under certain conditions improve the start-up performance of a GPS satellite-based positioning system. It is used extensively with GPS-capable cellular phones [**25**].

Table 1 Example Transmission Protocols Categorization

Transmission	Location	Interaction
Wi-Fi *IEEE 802.11* CDMA *W-CDMA (UMTS)*	GPS *NMEA 0183* A-GPS *RRLP*	Bluetooth *IEEE 802.15* NFC *NFCIP-1*

Bluetooth is a wireless technology standard for exchanging data over short distances (using radio transmissions in the ISM band from 2400–2480 MHz) from fixed and mobile devices [**26**]. Near field communication (NFC) is a set of standards for Smartphones and similar devices to establish radio communication with each other by physical contact of the devices or bringing them into very close proximity, usually no more than a few centimeters. Present and anticipated applications include contactless transactions, data exchange, and simplified setup of more complex communications such as Wi-Fi [**27**].

Human interaction is the backbone of this research, so naturally this category holds the key technologies used in this research; Bluetooth was selected for this purpose because although it is somewhat limited in range, that limitation also pushed this research towards using technology for different purposes than the ones for which they were initially created. Although NFC (Near Field Communication also listed in Table 1) was not actually used during the development phase of this research project, it was identified as a potential tool to reach the possibility of gathering very close, intimate interactions between devices which would lead future extensions of this project to a finer granular analysis of individual interaction; based on the potential benefit of its usage.

2.4. Back and Front-End Technology

Managing great amounts of information is a complex matter; it becomes even more complex when the scope includes managing the whole data lifecycle (gathering, processing, storing and analysing). In this vein, the back-end infrastructure was conceived as a central entity that manages the vast majority of processes related to the data. This central entity is formed by a two-layer server arrangement that splits the processing duties into application and database sub entities.

Figure 5 illustrates how the processes were distributed across the servers involved in the project, aiming to optimize the resource usage and the application performance.

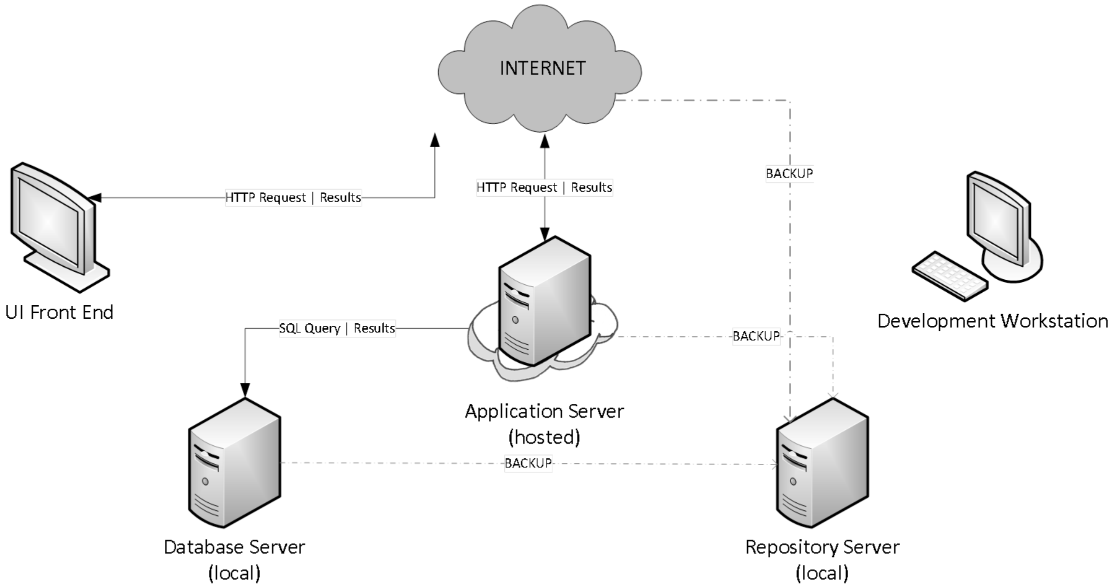


Figure 4 Back-end Infrastructure

Figure 4 illustrates the way processing duties were divided at a server level; a first layer application server managed all the requests related to the application side, leaving enough resources free for the second layer database server to manage heavy duty processing for data queries both in the gathering/storing phase and the analysis phase. A third server is planned as a repository to maintain the potential enhancements that could be implemented in the application, in an organized way.

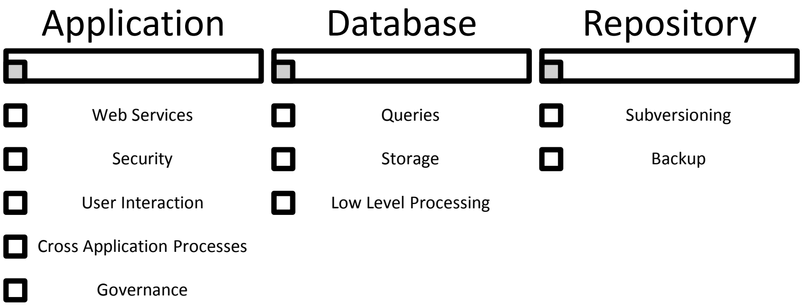


Figure 5 Server Schema

Application Server

Web Services: This refers to programs, scripts or macros developed to allow interaction between external entities (mobile devices, users, servers) with the processes and the servers that are part of this project. Some examples of this are the web service in charge of listening for external requests from the mobile devices used sending tracking information at a defined period of time using the HTTP_GET method.

Security: This refers to the processes defined for authentication at the user or device level to grant or prevent access to the different services provided.

User Interaction: This refers to the programs that in one way or another have direct interactions with a user to provide or request information at any given time.

Cross Application Processes: This refers to the scripts that interact with third party applications involving different processes to produce a particular result. Some examples for this could include CRON jobs running daily to process batched-unprocessed information.

Governance: This refers to the tools developed to manage the end-to-end processes involving the server applications, scripts and tools, the database processes and the mobile devices configuration; although those are separate-independent entities, some configurations and processes should be governed end-to-end to prevent communication inconsistencies between the parties.

Database Server

Queries: This refers to all the scripts, routines and procedures established at database level to manage the information stored or to-be stored.

Storage: This refers to all the procedures defined to segment data stored in the database to allow a more manageable dataset for analysis queries.

Low Level Processing: This refers to all the processes related to filtering, sorting and data calculation done at the database level.

Repository Server

Sub versioning: This refers to the tool used to maintain an organized application enhancement project, allowing a control version of the tools, script, configurations and everything related to the project at a code level.

Backup: This refers to all the procedures defined to keep the source of the application and tools secured and outside from the application and the database server.

Processing Servers Technical Specifications

Hardware:

- Shared Processor Genuine Intel(R) Xeon(R) CPU E5620 @ 2.40GHz – Cache 12288 KB
- Shared Memory: 12348980k/13369344k available (2009k kernel code, 114160k reserved, 909k data, 356k init, 11552236k highmem)
- Attached scsi generic sg0 type 0 sd 0:0:0:0: [sda] 286494720 512-byte logical blocks: (146 GB/136 GiB)10/100/1000
- Shared Ethernet port

Software:

- Backend administration system: CPanel
- Backend DB Administrator: PHPMyAdmin
- MySQL Server: Localhost via UNIX socket
- Server version: 5.1.56
- Protocol version: 10
- MySQL charset: UTF-8 Unicode (utf8)

- Scripting Language: PHP 4.3

Software

- Content Management System: Joomla
- Framework 2.5.4
- Jumi connector for mixed platform interfaces.

2.5. Platform Deployment

Current mobile technologies offer a great variety of Smartphone devices on which to develop; this immediately presents the need for a straight-forward, extendible software architecture for the application developed to be able to work on multiple platforms without a major re-development process.

This thesis describes in detail three different versions of the application developed, showing its evolution and the benefits achieved with every enhancement made to it during the development phase; for every version, one or more platforms were used.

For the Bluetooth Agent Tracker the development was only done on the Blackberry Smartphone; it was a proof of concept development so it was not designed to be ported to any other device (In addition, the devices were made available from RIM). The second version, face to face (F2F) was redesigned and deployed on a Blackberry Smartphone and also on an Android Smartphone; the Android API gave much more control over the radios to work with geographical location and data transmission over WI-FI or 3G. Both APIs used Java as a primary programming language so code portability was feasible. The last version of the application was only prototyped and not developed but it is designed to be ported to Blackberry, Android and iOS, representing the principal Smartphone providers in the market at time of writing. The platform selection was made because of the crowd sourced nature of the project; having the application on these three platforms ensures (somehow) that people will be more likely to use these devices and, ergo more data could be harvested in support of one of the goals for this thesis.

2.6. Chapter Summary

This chapter provided a general description of the concepts and technology used for the development of this project. Different scenarios that are part of the human interaction used to map the basis for the data collection model that will be presented, including several ordinary activities performed by individuals, who by using their Smartphones will be providing information to the network about activities that can include physical proximity to other people, conversational dynamics, and physical activity levels. Understanding the context within which data is collected using a Smartphone device is also significant as Smartphones are devices that highly relate to a person's daily activities, and are effective devices for tracking data about a person. The infrastructure required to achieve such purposes includes an assortment of wireless technologies which include Bluetooth technologies, the back-end server infrastructure and a web-client user/device front-end infrastructure.

CHAPTER 3

The Bluetooth Agent Tracker

3.1. Overview

After presenting a general introduction to this thesis in Chapter One, including the motivation for this research, the project aims and objectives, an introduction to the main topics, and the outline of some sources of related work, Chapter Two presented an introduction to the development performed to create the Bluetooth Agent Tracker (BAT), the technology behind it, and some potential uses.

The first section of this current chapter provides an introduction to Bluetooth technology as a whole. Some details of the network, its topology, benefits, and limitations are described. Then, the chapter introduces a description of the Bluetooth Agent Tracker (BAT) - the first-iteration mobile application that was developed and upon which much of this research was based. Technical details of the development of this application are provided, in addition to the development platform, a description of the data collected, and originally intended uses of the application. The chapter concludes with summary of potential uses of the application that go beyond the originally intended application of the system.

3.2. The Bluetooth Protocol

This research relied on a mobile application developed for data collection purposes; the first iteration mobile app was named the *Bluetooth Agent Tracker (BAT)*. This is a standalone mobile application which tracks information about Bluetooth devices that are within the range of connectivity of the device running the application. As the wireless technology previously described is a key component of the aforementioned research, a proper definition is outlined below.

Bluetooth is a wireless connectivity standard that was originally conceived in 1994 by engineers at Ericsson labs [**28**]. The idea was to create a method of wireless communication that would provide short range connectivity that required minimal infrastructure and resources, and efficient battery use. Due to the great potential realized in the technology, a special interest group (SIG) for the development of this technology was created. This was a group formed by several telecommunications organizations that agreed to work collaboratively on developing an open specification for wireless connectivity. Those companies were Ericsson, Nokia, Toshiba, IBM, and Intel. Since its inception, the SIG has been working on the development and promotion of the technology to make it universally available.

The Bluetooth connectivity standard uses radio frequency (RF) for signal transmission, and allows for wireless communication between devices that are located within a short distance of one another (up to 50 metres) over a personal area network established via Bluetooth. It is not used for long range radio broadcasts [**29**]. The developers of this technology utilized a license-free band in the spectrum that would be useful for transmission of Bluetooth radio signals. The license-free band is the 2.4GHz industrial, scientific and medical (ISM) band, which is available in most countries. Due to the free band that Bluetooth uses, any device that has the technology enabled can connect without having to pay any licensing costs. Bluetooth does not require direct line-of-sight, which means that the radio signals emitted by devices do not need to be in an unobstructed straight line for the devices to connect [**30**]. This provides an important advantage for Bluetooth connectivity because this means that devices can connect through walls, bodies, clothes, and many non-metal objects without the limitation presented in other wireless technologies that may require direct line-of-sight. When Bluetooth was initially developed, this fact generated great attention over other leading short-range connectivity technologies at the time such as Infrared Links (IrDA), which typically require direct line of sight. A third outstanding advantage that Bluetooth presents is the fact that many devices can connect together at the same time, and be part of the same network.

This standard supports ad-hoc device connection and automatic connection discovery. Due to the fact that ad-hoc networks can be formed with Bluetooth-enabled devices, no pre-existing configuration is necessary, therefore eliminating the need to have a pre-set infrastructure to create a network of devices [**31**]. This characteristic makes Bluetooth an ideal technology for studies using crowdsourcing, and maximizes the potential uses of data collected. Moreover, Bluetooth presents a very low set-up cost, and it is relatively energy-efficient, providing an advantage for battery lifetime on handheld devices.

3.2.1. Network Specifications

As stated before, Bluetooth uses RF signals for data transmission. Two or more devices can be connected to form a network, and the group of these interconnected devices that use the Bluetooth protocol for communications is called a *Piconet* [**29**]. Devices in a *Piconet* form ad-hoc networks as they discover service availability, so a *Piconet* is a dynamic array of network agents. In general, the first device that connects to a Piconet is a master. This device controls traffic over the formed ad-hoc network. All other devices that subsequently connect are slaves. Slaves on a *Piconet* can be active or inactive depending on the status of their connection. To establish a connection, agents share an access code that will allow one device to maintain synchronization with the same radio frequency as another device. All Bluetooth-enabled devices are constantly listening for page messages. Units constantly page other units, and when the paging and receiving unit select the same frequency hop sequence, then a connection can be established between the two, and the device is said to have joined the *Piconet*. Devices can be part of more than one *Piconet* at the same time, so a master in one *Piconet*, can be a slave in another *Piconet* at the same time. These devices that belong to more than one *Piconet* may act as bridges between *Piconets*, and this connection of more than one *Piconet* is called a *Scatternet*. Scatternets are there to more efficiently utilize the radio channel [**29**]. The figure below shows the typical layout of a *Scatternet*.

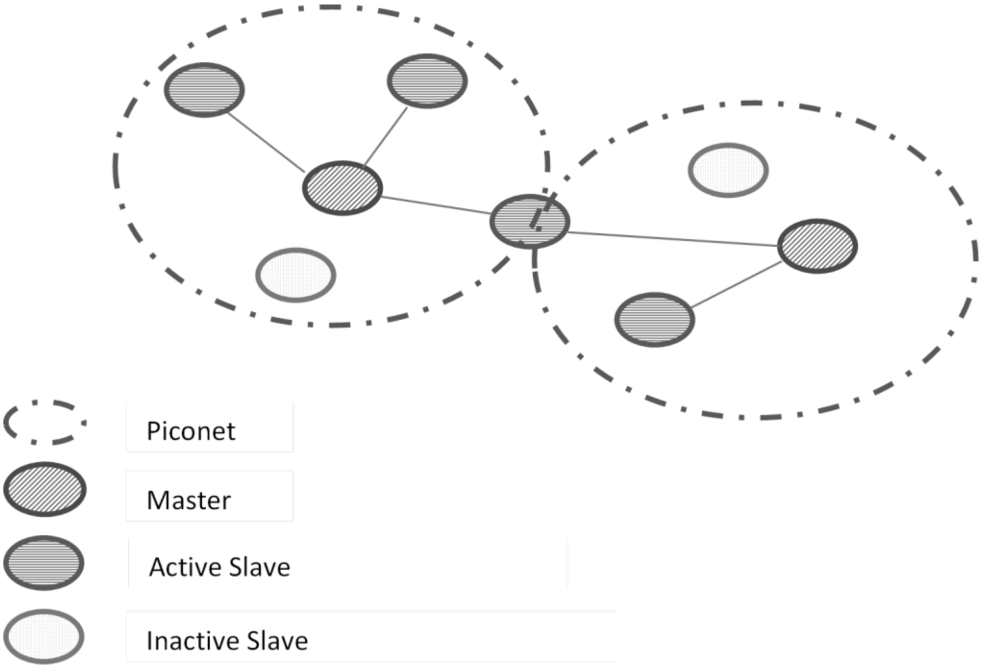


Figure 6 - Bluetooth Scatternet

Due to the fact that the band is limited – approximately 80 MHz at 2.4 GHz, as well as shared, it can become quite saturated, especially in networks where many devices can be searching for service and connecting dynamically as they find it. In order to avoid problems with interference on an overly-populated radio band, which is available to any device, Bluetooth uses frequency hopping. Under the Bluetooth standard, *Piconets* constantly "hop" from one channel frequency to another [**32**]. This type of spectrum spreading is used to supress interference [**29**]. Spectrum spreading divides the frequency band into various channels, and during a connection, radio transceivers jump from one channel to another one pseudo-randomly (albeit with a predetermined and known pattern).

3.2.2. Limitations of Bluetooth Technology

The nature of Bluetooth provides certain unique characteristics to connectivity using this standard. For example, the fact that the technology is meant to be for short-range connectivity allows Bluetooth to be implemented for a very specific set of uses. This very nature of Bluetooth technology was also the starting point of many of the challenges that were identified for the purposes of this research.

The level of familiarity of devices that join a network could be hard to define and determine using Bluetooth connectivity. Due to the nature of this protocol, where the connecting signal can sometimes go through walls or windows, this allows devices to form *Scatternets* with other devices even if they are in a different building, when adjacent. The challenge with this is that when establishing network familiarity, it is not possible to distinguish or differentiate devices that are directly connected to the master device in the same room from those which are in different rooms or even different buildings, but due to proximity can still connect [**33**]. This will affect some of the potential applications of sensor networks using Bluetooth connectivity. In addition, a related limitation of Bluetooth connectivity and its use in sensor networks is the fact that Bluetooth does not provide the chance to determine the exact distance between two devices that are connecting. As Bluetooth is a short-range connectivity standard, when devices are connected to one another under this protocol, it is possible to determine that the two devices are within close proximity to one another, however, it is not possible to establish the exact distance nor the directionality between the two with any degree of precision.

Bluetooth might raise some concerns over data privacy and security. As the band used is a free band, any device can access the same frequency. Even though this was the original intention of Bluetooth as an open wireless connectivity standard, and by design all devices that generate discovery signals can join the ad-hoc network, the fact that any device can connect and transmit information in the network to other devices that are connected raises some privacy concerns. The privacy concerns place a lot of responsibility on the researchers in particular as the ones who plan on making use of Bluetooth networks for public research purposes. The type of information collected from different sources and the crowdsourcing approach to data collection generate privacy-related concerns for researchers who must at all times act ethically in the collection of scientific data. In any case, it is worth mentioning that the Bluetooth technology does provide link layer security between connecting devices. It provides authentication security, data between two devices is encrypted, and network agents cannot exchange information before establishing a connection.

3.3. The Bluetooth Agent Tracker

A Bluetooth Agent Tracker (BAT) application for mobile devices was conceived and developed after the characteristics of the Bluetooth technology were fully considered and the BAT became the backbone of this research. This application was originally envisioned as a method to identify devices that were within the Bluetooth range of the device that was running the application; the device running the BAT app was denoted an Agent. The information gathered from the devices discovered (found) by the Agent was then recorded in a log for later analysis. Every device with Bluetooth capabilities is able of perform a process called device discovery, this means that all devices can find information about other Bluetooth devices within range. The information that was collected by the Agent running the BAT application upon device discovery of other devices included the other device's Bluetooth radio MAC address, its "friendly name" (i.e. Blackberry 9800 or John Doe's Android), date and time of the interaction, and basic information of the Agent running the BAT application and performing the tracking (name, Bluetooth MAC address, and battery level).

The first objective was using the Bluetooth Agent Tracker application to find out how many times a Bluetooth device made an appearance in the Agent's log, within a specified timeframe. For example, a device running the BAT would store a record of any Bluetooth device within its range every 20-30 seconds for a defined amount of time; this log would later be imported into a Server for analysis. At that point, the application was merely meant to track nearby devices that established a preliminary connection, and no use beyond this was identified for the application.

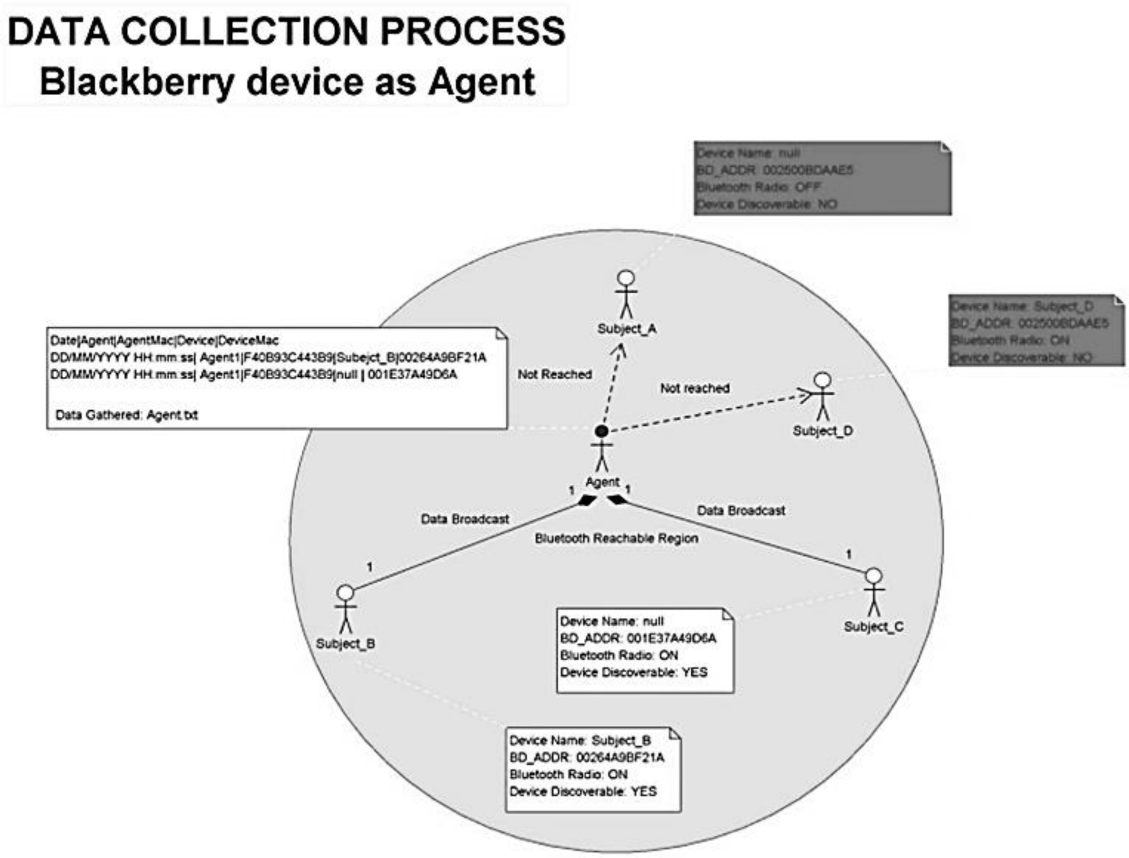


Figure 7 - Use Case Scenario for data collection from probe agents

Figure 7 describes a simplified model of how the BAT application worked; the devices surrounding the *Agent* device could be configured to broadcast or not broadcast its information, and during the discovery process triggered on the Agent device they would show and be logged or not logged, depending on their broadcast information configuration. This process continues for the defined time for the probe, and the results would be later processed on a web server database for further analysis.

Figure 8 depicts the sequence of events and the entities involved in the process of gathering data from the Agent device and the process of adding the data gathered into the server for processing. On the diagram, the Agent Tracker starts the *Discovery Agent* process and stores the information of the devices tracked locally in a temporary text file; this file holds the information gathered; the *Record Data* entity appends every new record generated during the discovery process to the text file until the probe time-lapse finishes. Once the process has finished, the user in charge of the Agent device manually gets the text file out of the Smartphone (via USB) and imports it into the database using a web tool that was developed for that purpose. The sequence of events ends when the data are validated from the server side and marked as valid.

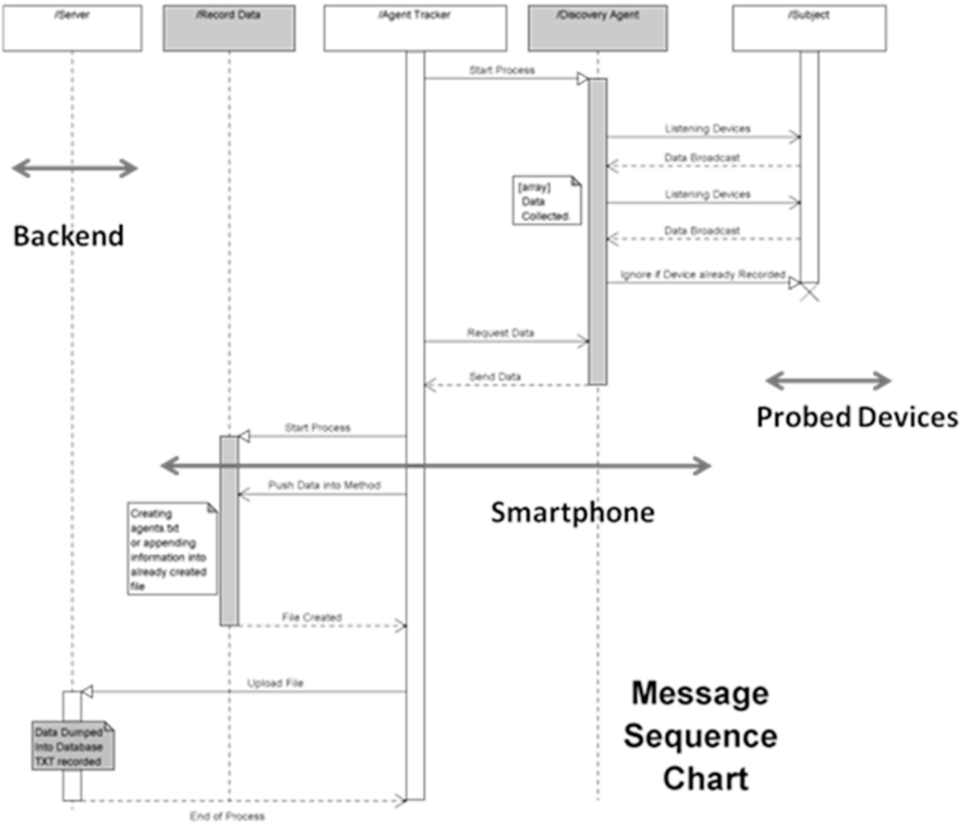


Figure 8 - BAT Sequence Chart

Figure 9 describes how the Bluetooth Agent Tracker interacts with devices in its vicinity during the data gathering process; the circle shapes represent the Bluetooth radio range of each device interacting with the Agent device and the solid lines represent a connection between the Agent device and the developer's workstation, and between the workstation to the data processing server. It is important to note that although the diagram describes a static point in time in which the devices are within discoverable range, the reality of the model involved dynamic devices moving around the Agent device at any given point in time;

this dynamic behavior drove the next stage of development of this application that included geographic location capabilities to help the researchers understand the implications of having Agents in close proximity with several other devices in different places. This is discussed in more detail in Chapter 4.

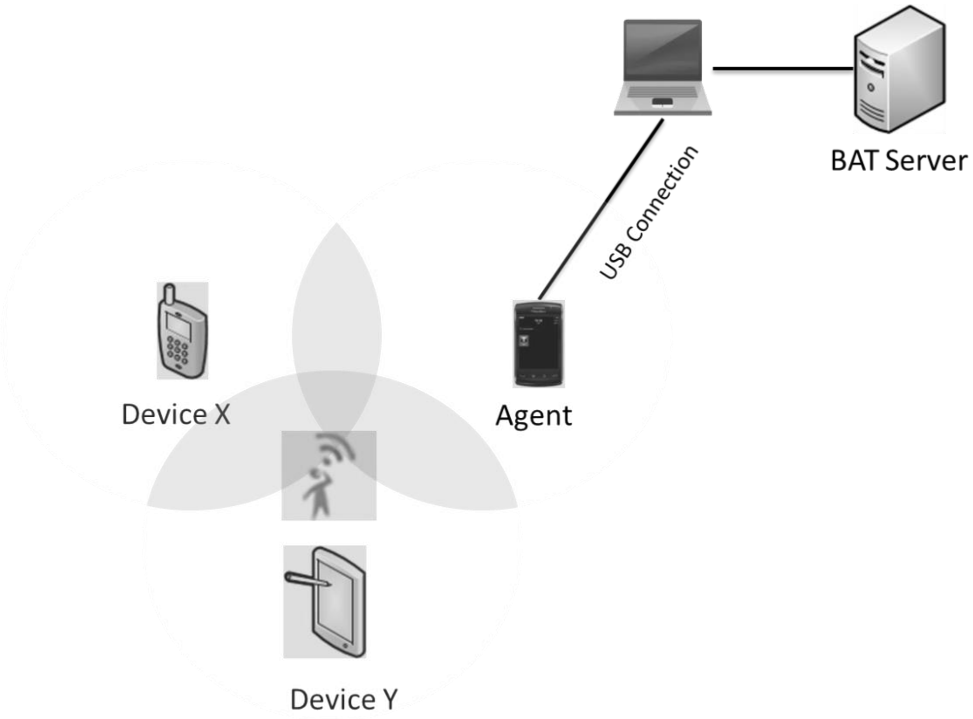


Figure 9 - BAT Interaction Scheme

3.3.1. BAT: The Application

The Bluetooth Agent Tracker application ran as a background service in the Blackberry Smartphone but was started on the front end of the device as shown in Figure 10; It used the Smartphone LED to notify the user when the discovery process was taking place, showing a dark blue blinking LED that represented the process starting and a fixed red LED if the application crashed. The interface created for this first prototype displayed basic information to allow the user to know where to find the collected data and to be able to stop the service at any time. Figure 11 describes this interface.

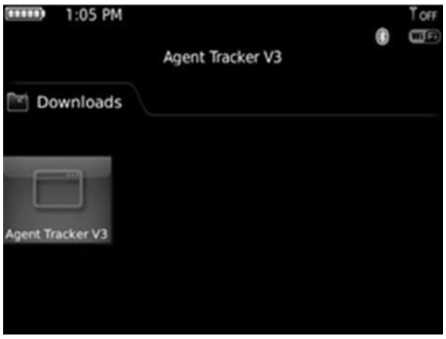
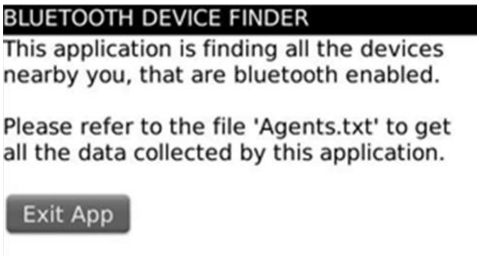


Figure 10 - BAT Starting Point

Once the test was finished, the BAT user was able to stop the service from the user interface and gather the information from the text file created during the process. Initially, the application was designed to display the information gathered on the front end interface (directly from the application screen) but after some tests it was determined that showing great amounts of information on the user interface would not represent any benefit to the process and would rather consume resources unnecessarily. The information was stored in a text file for easier data handling.


BLUETOOTH DEVICE FINDER

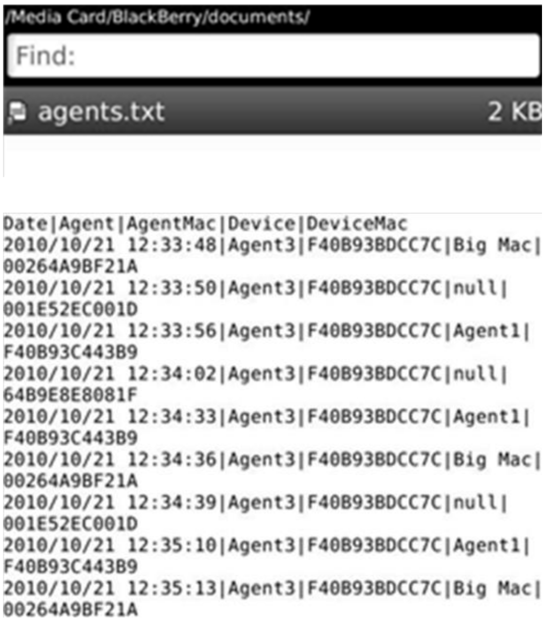
This application is finding all the devices nearby you, that are bluetooth enabled.

Please refer to the file 'Agents.txt' to get all the data collected by this application.

Figure 11 - BAT Interface

On the data processing side, the user was able to navigate using the Smartphone operating system to get into the *agents.txt* file to be able to verify the information gathered during the testing period and to later import that information into the database. The first approach for these data gathering model raised the concern of having some sort of security to handle the information because of its sensitive nature (information from other individuals transmitted over the air); the information was later serialized to keep the information as abstract as possible in case of intrusion when the data transmission from the Smartphone to the server was done via Wi-Fi (described in detail in Chapter 4).

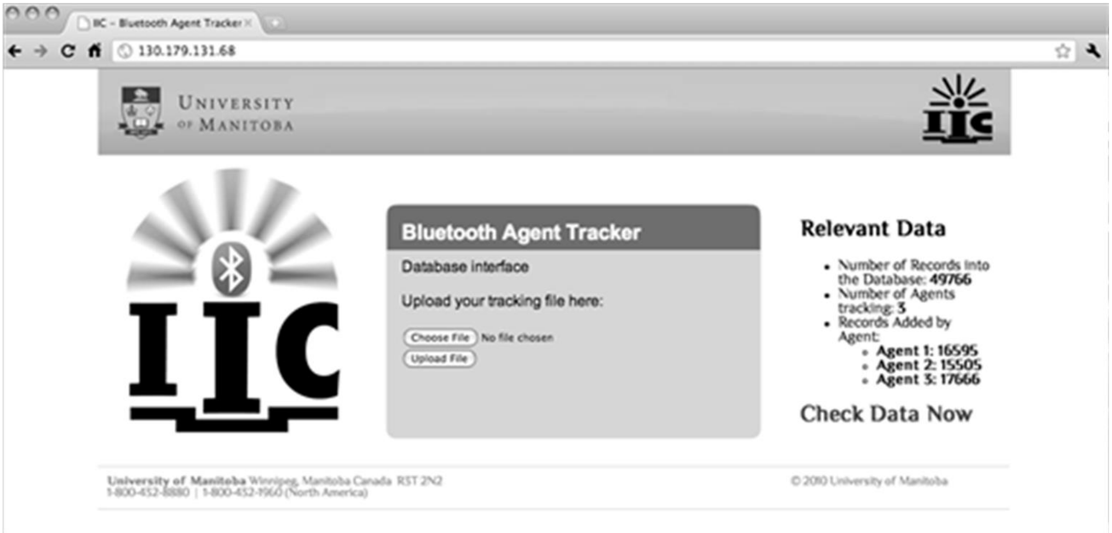
Figure 12 describes the path in which the text file with the tracking records was residing and a sample of the data gathered by the *Agent* devices.



```
Date|Agent|AgentMac|Device|DeviceMac
2010/10/21 12:33:48|Agent3|F40B93BDCC7C|Big Mac|
00264A9BF21A
2010/10/21 12:33:50|Agent3|F40B93BDCC7C|null|
001E52EC001D
2010/10/21 12:33:56|Agent3|F40B93BDCC7C|Agent1|
F40B93C443B9
2010/10/21 12:34:02|Agent3|F40B93BDCC7C|null|
64B9E8E8081F
2010/10/21 12:34:33|Agent3|F40B93BDCC7C|Agent1|
F40B93C443B9
2010/10/21 12:34:36|Agent3|F40B93BDCC7C|Big Mac|
00264A9BF21A
2010/10/21 12:34:39|Agent3|F40B93BDCC7C|null|
001E52EC001D
2010/10/21 12:35:10|Agent3|F40B93BDCC7C|Agent1|
F40B93C443B9
2010/10/21 12:35:13|Agent3|F40B93BDCC7C|Big Mac|
00264A9BF21A
```

Figure 12 - BAT Tracking Records

The last step of the data gathering process involve the data storage process into a database for posterior analysis; for this purpose, a simple web user interface (Figure 13) was developed for retrieve, validate and store the information recorded in the text file into the Bluetooth Agent Tracker database.



Bluetooth Agent Tracker

Database interface

Upload your tracking file here:

No file chosen

Relevant Data

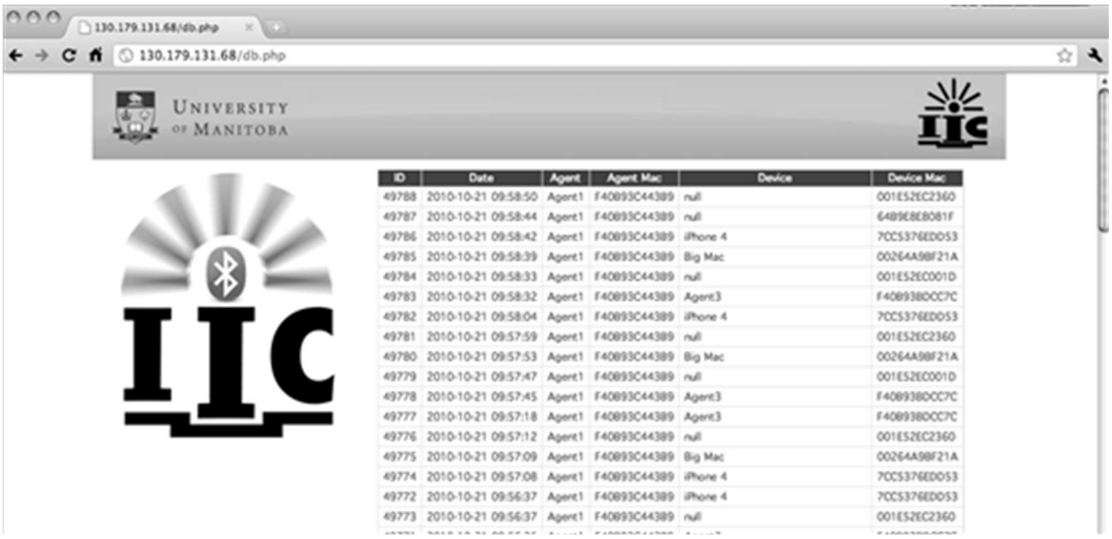
- Number of Records into the Database: **49766**
- Number of Agents tracking: **3**
- Records Added by Agent:
  - **Agent 1: 16595**
  - **Agent 2: 15505**
  - **Agent 3: 17666**

University of Manitoba Winnipeg, Manitoba Canada R3T 2N2
1-800-432-8880 | 1-800-432-1960 (North America)

© 2010 University of Manitoba

Figure 13 - BAT Web Interface

The web interface also provided information about the Agents that were part of the test and the number of records gathered by them. An additional screen shown on Figure 14 was developed to create a user-friendly display of the devices recorded during the test, which allowed the researchers to start inferring new potential uses for the Bluetooth Agent Tracker.



ID	Date	Agent	Agent Mac	Device	Device Mac
49788	2010-10-21 09:58:50	Agent1	F40893C44389	null	001E52EC2360
49787	2010-10-21 09:58:44	Agent1	F40893C44389	null	6489E8E8081F
49786	2010-10-21 09:58:42	Agent1	F40893C44389	iPhone 4	7CC5376ED053
49785	2010-10-21 09:58:39	Agent1	F40893C44389	Big Mac	00264A98F21A
49784	2010-10-21 09:58:33	Agent1	F40893C44389	null	001E52EC001D
49783	2010-10-21 09:58:32	Agent1	F40893C44389	Agent3	F40B938DCC7C
49782	2010-10-21 09:58:04	Agent1	F40893C44389	iPhone 4	7CC5376ED053
49781	2010-10-21 09:57:59	Agent1	F40893C44389	null	001E52EC2360
49780	2010-10-21 09:57:53	Agent1	F40893C44389	Big Mac	00264A98F21A
49779	2010-10-21 09:57:47	Agent1	F40893C44389	null	001E52EC001D
49778	2010-10-21 09:57:45	Agent1	F40893C44389	Agent3	F40B938DCC7C
49777	2010-10-21 09:57:18	Agent1	F40893C44389	Agent3	F40B938DCC7C
49776	2010-10-21 09:57:12	Agent1	F40893C44389	null	001E52EC2360
49775	2010-10-21 09:57:09	Agent1	F40893C44389	Big Mac	00264A98F21A
49774	2010-10-21 09:57:08	Agent1	F40893C44389	iPhone 4	7CC5376ED053
49772	2010-10-21 09:56:37	Agent1	F40893C44389	iPhone 4	7CC5376ED053
49773	2010-10-21 09:56:37	Agent1	F40893C44389	null	001E52EC2360

Figure 14 - BAT Data Gathered Sample

The BAT application process was very simple yet very useful as it let the author gather a great amount of information in a short period of time. As stated in [**34**], in the case of the BlackBerry Storm devices used in BAT application prototyping, with just over three months of data collection and with just five probe devices, approximately 500,000 contact (connection) records were collected.

Figure 15 describes the data flow and transmission protocols used between devices during the data transit process. On the data gathering segment, only devices that have a discoverable Bluetooth radio will be discovered with the Agent device which will record the information broadcasted by the devices over the ISM (industrial, scientific and medical) band from 2400–2480 MHz.

The data transfer segment involves the interaction between the *Agent* device, the development workstation and the user workstation that was used to retrieve the data harvested by the *Agent* device in order to get it for processing and further analysis via USB cable; this process is later re-engineered in order to get the data automatically pushed to the server over a Wi-Fi or CDMA network, which supported data integrity by not relying on a user to remember to manually upload data. The initial reason for using the USB to upload data was that the devices provided by RIM did not have service contracts for cellular. In addition as this was a proof of concept the temporary file data storage option was deemed to be sufficient at that point in time.

Finally, in the data processing segment, the data are validated, stored and later queried during the analysis phase. The application that interfaces the user with the database entity uses an ODBC driver to interact with the relational database.

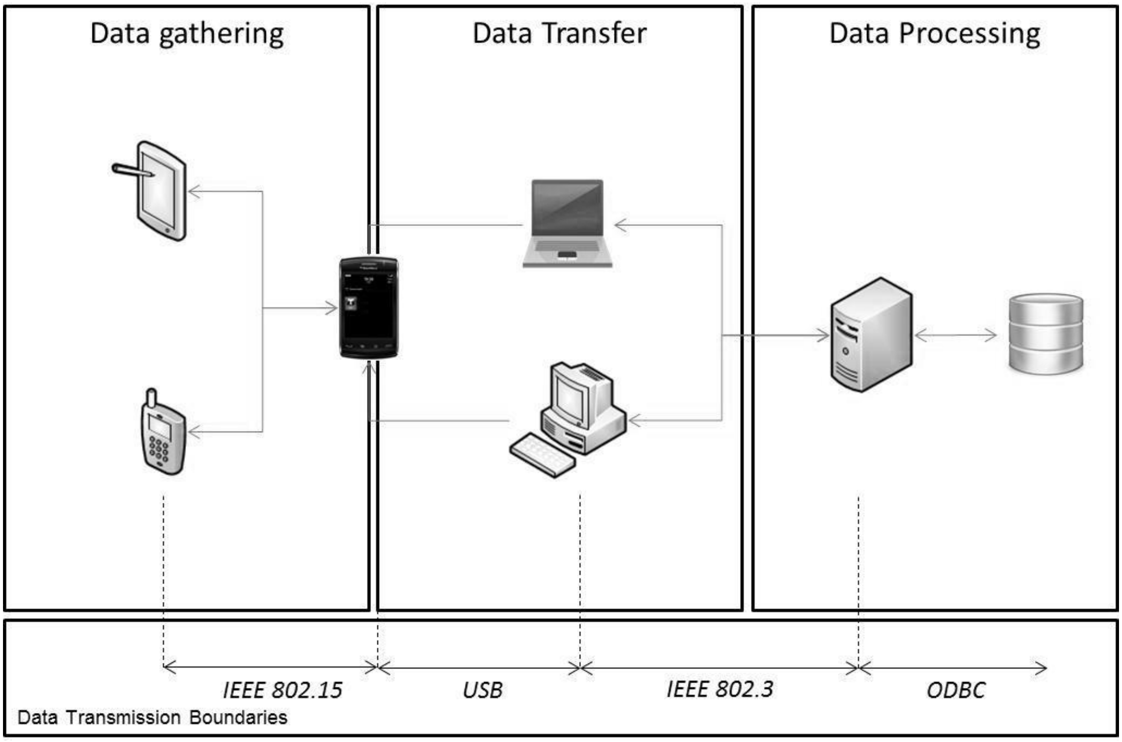


Figure 15 - Data Transmission Scheme

3.3.2. Development Framework

The Bluetooth Agent Tracker application was initially developed for one platform (Blackberry) and was later re-developed for the Android OS. The application was originally created for Blackberry due to the fact that Blackberry's API and devices were more accessible and the Bluetooth functionality was flexible enough to reach the objectives defined at the project's starting point. However, the application was later developed under an Android platform because as the Android API grew in popularity, the platform allowed a more extensive use of the radio APIs and it provided more flexibility than the Blackberry APIs.

There were several considerations made before the BAT mobile application development started. Battery usage was one of the most important concerns as it would determine the amount of time users would be willing to run the application, which would be directly related to the amount of data that could be collected [**35**]. Another consideration was CPU and RAM usage because excessive use of the device resources would immediately turn the application into a NO for large-scale adoption usage (people don't use applications that drain their cellphone's resources). In addition, the selected platforms that were to be used to develop the Bluetooth Agent Tracker application had to include versatile mobile capabilities, the ability to operate through wireless channels, and the ability to always be connected using Bluetooth networks as research required the ability to trigger local Bluetooth events to generate the whole discovery process required to achieve the objective of this first development prototype.

The first device running a Bluetooth Agent Tracker instance was a Blackberry Storm 2 which was called *Agent 1*. The Blackberry API allowed a straight-forward use of the Bluetooth radio connectivity capabilities of the Smartphone. This API is based on the Java Micro edition and makes use of the Java runtime environment included on the Smartphone operating system. The Java Micro edition has a rich library of classes that made the development process relatively straight-forward. As far as Bluetooth requirements, the Blackberry platform allowed Bluetooth device pairing and fully supported the Java Community Process JSR82 Bluetooth specification; thus, applications built under this platform could scan for devices in close proximity, initiate discovery of services, and create Bluetooth services [**36**]. Blackberry also provided persistence storage that was very beneficial for research uses [**5**]. The platform provides standard methods to access files and to manipulate the file system. With regard to memory management, the Blackberry OS (at that time) did not automatically kill processes when the device ran out of memory and the Garbage collector for the Java version used during the development phase was not working efficiently, which at times would lead to an out-of-memory error. As such, during the development phase we made sure the applications managed memory in an appropriate way.

After the application was developed on a Blackberry platform, the Android API was pursued to develop the Bluetooth Agent Tracker under this platform. Nexus HTC devices were originally used for running the application. The Android platform is based on Linux, and is currently one of the most widespread platforms due to its versatility and its openness. Android is based on the Java API. Applications installed under the Android platform run their own Linux processes and resources, never sharing resources with other applications [**37**]. This platform allows for battery status query, and provides capabilities to manipulate the battery profile. The initial release of this API did not have Bluetooth support, but this capability was later added. Today the Android platform allows for Bluetooth activities such as scanning for other Bluetooth devices, querying the local Bluetooth adapter for paired devices, connecting to other devices through service discovery, transferring data to and from other devices, and managing multiple connections. The OS shuts down processes in cases of the system running out of memory, which is a detriment of this platform when it needs to be used for research [**5**]. As the platform is open to the community, the API has made significant progress in becoming incredibly comprehensive.

3.3.3. A Brief Analysis of the Data Gathered

Portions of this section have previously appeared in [**38**]. Generally, in the case of data services with voluntary data input or contribution, the participants must perceive value – either to themselves and/or to the collective organization – in offering the information. In a study of a similar project at MIT, the benefit perceived by the participants was that they were provided with a cell phone and cellular service by which their contacts were tracked [**39**]. In this work and facing different resource constraints, we developed an application that could be used by a relatively small number of participants as probes allowing them to gather proximity data from a variety of consumer electronic devices that were Bluetooth-discoverable, including Smartphones, laptops, tablets, and earpieces.

Initially there wasn't a good estimate of the effectiveness of proposed BAT in terms of collecting proximity data. There are a number of issues related to radio control and Bluetooth pairing mentioned earlier that were in question. However, in many cases even if the device is not in discoverable mode it can still be found as long as its Bluetooth radio is on. In terms of identifying the device there are look-up engines and repositories that are readily available to help infer the device type [**40**].

Initial attempts to visualize the data were based upon the following ideas, some of which were realized to various degrees. A difficulty with the BAT data alone in addition to the relatively small number of participants was that there was no geo-spatial component associated with the data. Some of these visualization difficulties were minimized as a degree of spatial data was included in subsequent development. Using the collected and stored data, a contact graph was generated from a sample of the data (Figure 16). Figure 16 illustrates a weighted graph of Agent devices alone. The graphs are drawn with edge weights representing aggregate connectivity duration between the probe devices. The aggregates are taken over three distinct time periods. The graphs represent associations on a typical work day (e.g. Wednesday), a weekend day (e.g. Saturday), as well as an aggregate over a 7 day period.

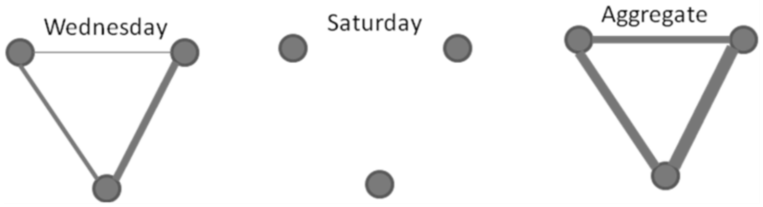


Figure 16 - A Contact Graph Collected from Several Agents

Figure 17 presents the data from the perspective of an individual Agent. These figures are similar to those used in presenting social contact graphs or graphs of knowledge. In Fig. 17, the *Agents* are illustrated as in Figure 17 (Wednesday) but with their contacts drawn local to the *Agent tracker*. The weight is illustrated as the distance from the *Agent tracker*. In the scenario for each individual node, the closer it was to the adjacent node, the greater the edge weight or duration of contact.

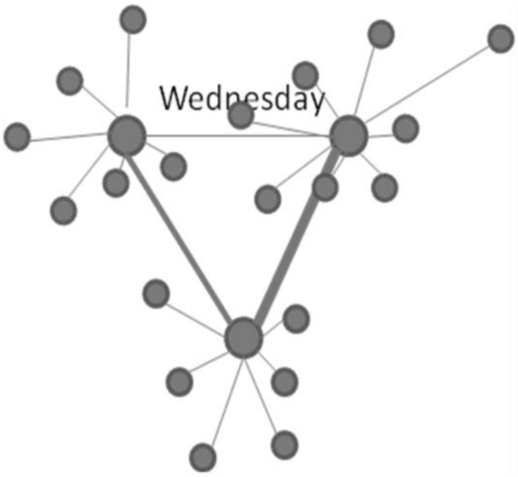


Figure 17 - A Contact Graph Collected from Several Agents and their Connections

As expected, there are devices that are detected by more than one Agent at or near the same time. Figure 17 treats all Agent devices as independent. In the case where a device is seen by two Agents it was elected to represent that node as being collinear proportionally located between the two interrogating phones. This can be generalized to a device interrogated by more than two Agents easily.

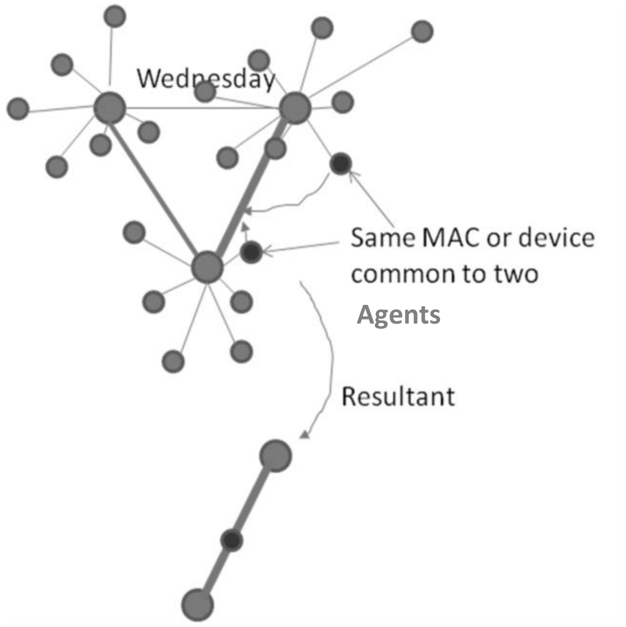


Figure 18 - Representation with Common Devices Tracked

The virtue of the dependence of *tracked* devices is that it in effect increases the connection between two *Agent* devices as discussed in the next section.

Figure 19 illustrates the data collected over a 24 hour period with all connections illustrated.

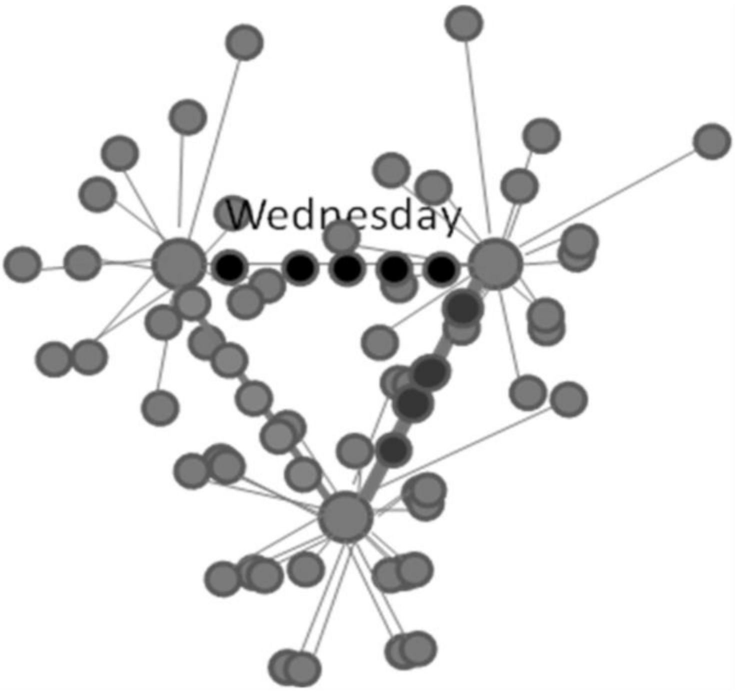


Figure 19 - Representation with All Connections Illustrated

Visualization of massive amounts of data is difficult from a number of perspectives. The difficulty associated with making a graph as planar as possible is an NPC problem. However, graph drawing lends itself to a number of approximation techniques that provide aesthetically pleasing and semantically 'correct' layouts.

There is essentially one means of representing the data graphically, parameterized by a contact window. The contact window presents and aggregate of contacts and duration over a specified time interval. For certain applications, the contact window may be a 24 hour period and would be relatively static over routine schedules. If the contact window were 10 minutes, the contact graphs would be considerably more dynamic, albeit capturing similar aggregate data over a shorter time period. This is seen by the data collected on a weekend as opposed to a work day. Also, in this particular scenario Agents 1 and 2 work in relatively close proximity while Agent 3 has a more irregular work pattern.

In addition to visualization, data of this type can be used within models for disease spread that are based on contact patterns. It is important to note that the BAT application represents nodes on the graph where clearly not all nodes are created equal. Smartphones running the BAT application have considerably more data associated with them than those that are queried (discovered). In an attempt to collect significantly more data, the backend data repository allows anyone running the application to upload their data, and in return provided with a graph of their connectivity social proximity contact patterns. For this to be truly useful, a significant percentage of a population would be required to participate. One final aspect associated with the BAT application is that contact with *non-mobile* objects can also be collected and analyzed. For example, devices that are somewhat permanent without mobility such as desktop workstations with a Bluetooth transceiver are often discoverable without being able to infer that a person is associated with the object. These allow for landmarks to be mined and identified that are useful in localization of the Agent devices. In a real deployment these features could be learned of mined from the data itself.

3.4. Potential use of the Bluetooth Agent Tracker

As the Bluetooth Agent Tracker application was developed, the initial intended use of it was to trace devices that were discoverable by the Agent device running the application. This information was traced for more than six weeks. The results of these data registration, after the analysis, demonstrated a potential to use this information in a way that went beyond the initial idea of purely tracking devices with Bluetooth connectivity for proximity. The information collected by the BAT application could be very useful for other purposes. The information that was originally collected displayed continuous and direct contact among devices; in addition, it was possible to infer with a good level of confidence that there was close proximity to some devices due to the fact that the connectivity logs from the Agent device would show the same devices at the same times on different days during several weeks. As such, the previously identified limitation of Bluetooth technology became a factor that would provide wider opportunities for use of the Bluetooth Agent Tracker. The very reason that motivated the BAT application which was the fact that Bluetooth enabled devices can be used as sensors in a network, allowing the researcher to find more uses of the application in social situations. Mobile devices can be considered wearable sensors that will be identifiers of people's daily activities. Thus, the initial analysis allowed the researcher to consider the collected data as a tool for development of models and simulations that required an interaction among individuals who hold those devices – more specifically, that the data can serve as inputs into models that may include epidemic disease spread, traffic flow, disaster relief, and other similar models that rely on accurate and precise resolutions of individuals' movements and interactions.

Some of the models envisioned by this preliminary analysis are briefly described on the following sections.

3.4.1. Establishing Social Patterns

Using Bluetooth devices to establish social patterns of individuals that carry Bluetooth enabled devices is one of the most evident uses of tracking applications – where the Bluetooth device (usually, a Smartphone) effectively serves as a proxy for its owner, and the device's movement patterns can essentially be defined as the owner's movement patterns. As the Bluetooth Agent Tracker would, on further developments, be able to locate itself with the use of the geo-location capabilities on the mobile device that is running the application, it would be possible to establish with a good level of accuracy, the location of devices that join the network and get paired with the devices running the application. As a matter of fact, it has been demonstrated that Bluetooth devices can be used to establish social patterns in daily activity [**4**]. As such, with collection of social patterns, certain types of social relationships could be inferred, socially significant locations such as shopping malls, parks, arenas, and clubs could be discovered, and organizational rhythms could be modelled.

3.4.2. Epidemic Simulation

Epidemics have the potential to cause significant damage to society due to the rapid and unpredictable nature of their spreading and the high health cost of associated with infection spread. When models are created to simulate epidemics, the infection spread behaviour can be established to a certain level, and mitigating strategies could be conceived and tested within the model. As the Bluetooth Agent Tracker could be potentially used to register population displacement (once the geo-location capabilities were implemented) which influences the spread of diseases [**41**], the spread of an epidemic based on patterns of mobility of individuals could be modelled. Agent-based and network-based simulation procedures could be used for epidemiological modelling, where network interactions are modelled based on network theory [**42**]. For example, different agents could be traced, and by assigning probabilities of transmission, and infection to each agent that is discovered by the Bluetooth Agent Tracker, models for contact based disease spread can be created. As a futuristic application in epidemiology, the efficacy of a limited vaccine could be improved by targeting persons who are most highly connected (super spreaders) and identified via Bluetooth proximity records.

3.4.3. Traffic Models

Bluetooth tracking could be used to gather different types of information related to vehicle and pedestrian traffic in cities. Bluetooth signals have been used to track travel times between two points through the use of static sensors and device discovery [**43**]. By assuming that most discovered devices will be related to people's Smartphones and thereby serve as a proxy for the individual, the Bluetooth Agent Tracker could be used to collect information about how congested a particular area of a city is, through counting the number of devices that make a connection to it in a specified timeframe. In addition, the direction that vehicles or pedestrians are traveling relatively to the device running the BAT could also be gathered with the use of this application. Being able to track devices through traffic could also introduce business uses of the application such as fleet tracking, and establishing the most direct routes for distribution of goods. In addition, people's patterns of movement within a big store or shopping centre could be traced as well.

3.4.4. Natural Disaster Relief

After a natural disaster occurs in any particular geographic area, there are vast movements of people that can occur, as people move away from insecure areas and move towards safety. Agent tracking through Bluetooth could be used to identify and track the directions that people would take after a natural disaster, and also to identify people in close proximity. People movements can cause important increases in morbidity and mortality if relief assistance cannot be directed to the locations with the highest concentrations of people [**13**]. As such, the application could be implemented on devices in disaster zones in order to track movements of an affected population. It is also possible to use Bluetooth tracking as a means to locate a victim who may be trapped as a result of a disaster [**44**].

3.5. Chapter Summary

This project is backboned on an application using Bluetooth technology. Bluetooth is a close range, battery-efficient connectivity standard that does not require line of sight to generate a connection. Enabled devices are recorded on the Agent device logs. Through service discovery, Bluetooth-enable devices can and are able to communicate with one another. Bluetooth presents certain limitations that are a result of the nature of the technology. However, the nature and characteristics of this connectivity standard allow for unique solutions of networking problems and patterns.

The first iteration of this research was based on a software application that was designed and developed for mobile devices, called Bluetooth Agent Tracker (BAT), and was built initially on the Blackberry platform. The application is meant to be installed on a device, allowing it to track (discover) nearby devices that have Bluetooth connectivity. Initially it was unclear as to the type and quality of the data that could be collected. The original problem that the application tried to resolve was to identify devices that were in close proximity to the device that runs the BAT application, however, once the value of the data gathered was realized, the researcher realized that through data collected, contactability and proximity of devices could be deduced, and that many other uses could be derived from the application including discovery of social patterns, input for contact based epidemic simulations, traffic mapping, and natural disaster relief support.

CHAPTER 4

Face2Face – Second Version of the Bluetooth Agent Tracker

4.1. Overview

Chapter 3 presented the Bluetooth Agent Tracker (BAT), including the Bluetooth protocol, its characteristics, and limitations. The BAT application itself was introduced, and the backend of the application as an essential part of the infrastructure required for it to yield successful results was described. Enhancements were identified which could enhance the uses to the application beyond what BAT was originally designed to achieve.

The current chapter describe Face2Face, a second-iteration or upgraded version of the BAT application and it represents an evolution of the application as it was originally conceived by including geo-location data collection. This chapter firstly provides a discussion on geo-location services that modern Smartphones provide today. This includes descriptions of the protocol that is being used, and the way in which it operates, which usually refers to GPS and assisted GPS services. In addition, the history and some of the limitations of this technology will be examined. Later in the chapter, the Face2Face application is described along with its relationship with Bluetooth/GPS technologies; the applications of Face2Face to traffic and epidemic simulation are also discussed, where spatial location and proximity are clearly established.

4.2. Geo-location Technologies

Geo-location is the process by which the physical location of a device can be determined; it is used by advanced geographic area systems to determine meaningful locations, by assessing the position of an object according to its coordinates. A meaningful location can be the distance between the device and a landmark, the distance to another device, or a location on a map. Specifically, geo-location technology refers to the use of GPS technologies and other associated technologies that will allow establishing of accurate location of people or devices [**45**]. There are many considerations to establish accurate geo-location, such as the type of terrain in the location of the device or person, or whether the person holding the device has direct line of sight to the sources of information, such as satellites, or the density of population of the area where the device is located. For example, a flat terrain will have different considerations than hilly terrain. As such, many different technologies are available for these types of location services, each of which is best suited to specific situations and uses.

4.2.1. Evolution of Geo-location Technologies

The history of geo-location can be traced back to ancient times, when humans began moving to explore new latitudes, and realized the need to understand where they were located on the planet. The first technique that humans used for geo-location was through the triangulation of stars. In the 1920s, radio signals were first used for calculating distances between ships, and the Russians introduced the idea of using satellites. Russian scientists realized that they could pin down the location of a satellite by taking note of radio frequency (RF) signals and adjusting for the Doppler effect, where the frequency is higher when an object approaches and reduced when the object is moving away [**46**]. In 1973, the United States Air Force, using this concept, launched 24 satellites for global tracking, and called them NAVSTARS. They began using digital signals, and placed the satellites at higher orbits to achieve greater coverage. They used ranging measurements of the satellites for greater precision as well [**47**]. This is how the first GPS as it is known today was conceived. The technology has evolved, and today GPS has been enhanced with other technologies to achieve true geo-location.

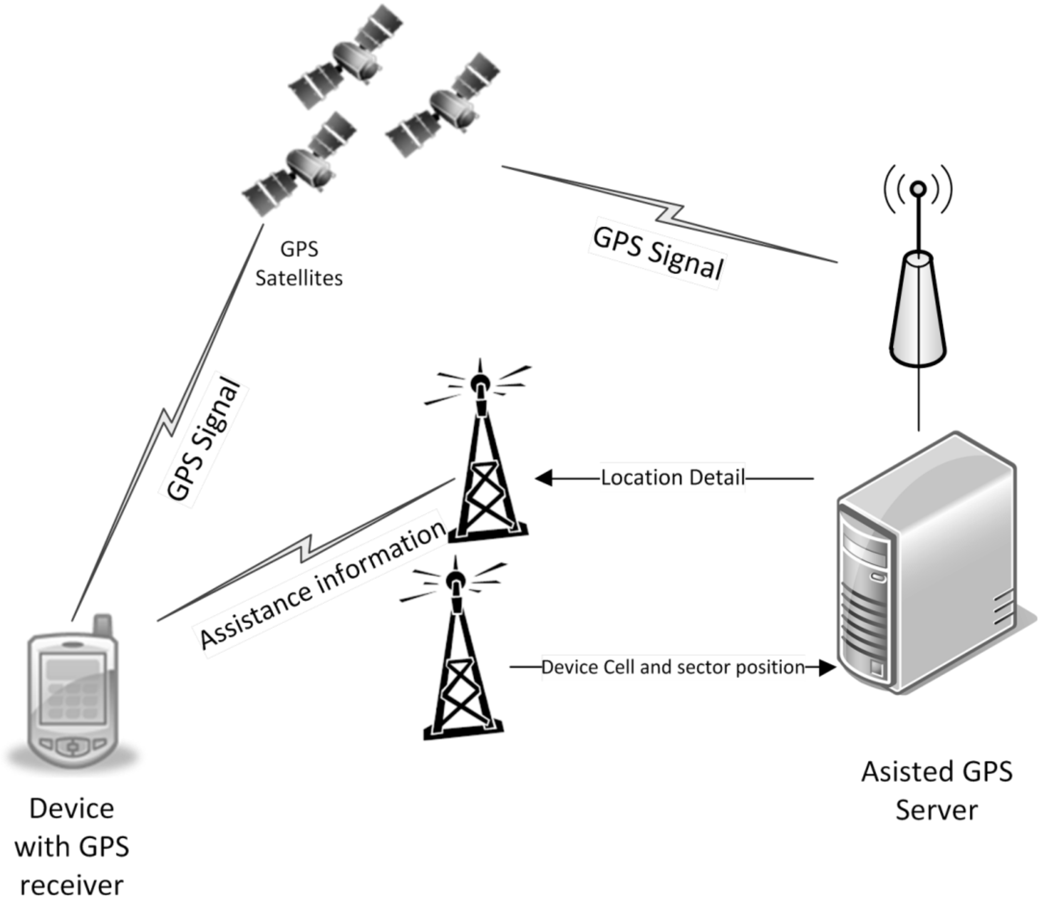


Figure 20 - Assisted GPS Architecture

Assisted GPS (A-GPS) is the technology that allows GPS location with a higher level of precision and response time. It is a convergence of technologies that works in conjunction to

GPS to enhance location accuracy by reducing the search space through the use of the wireless network to which the device is connected by, using the device's own GPS receivers, and an estimation of the device's location down to the cell or sector in which it is operating [**45**] [**48**]. It has been demonstrated that there is significant inaccuracy on GPS technologies when operating in urban environments [**49**]; for that reason, other methods are required through the use of assisted GPS in order to achieve higher levels of accuracy. The latest developments in geo-location introduce many different technologies for mobile location-based services.

As stated previously, the use of location-based services provides many advantages for different aspects of human and social living. There are many options to achieve location identification of a device, and many of them have pros and cons associated with them. Using radio frequency signals from the closest cell tower to the phone is a simple and economical technique to locate devices; however, this technique is inexact, and impossible to pin down the exact location of a device. GPS allows the estimation of position but may require a significant delay as the receiver needs to obtain almanac data before estimating a position fix. Assisted GPS, on the other hand, provides the ability to more precisely locate a device, however, it is more expensive (it usually requires a data plan), it may requires special chips to be included with the device (e.g. SIM card), and the device must be in within the area of coverage of the company providing the cellular network service or have roaming enabled (usually costly).

In spite of the great advantages that it provides, A-GPS is not a perfect technology. It is known not to work well indoors and in high-density urban areas [**50**], and for that reason, other complementary technologies could be implemented in order to complement the location/contact services of a certain device when it is out of range; technologies such as Wi-Fi, Bluetooth, ultrasound, and infrared could be also used to enhance location services on these devices. Face2Face, an application that represents the evolution of the previously deployed Bluetooth Agent Tracker (BAT), used a balanced mix of these technologies to gather a very specific set of information using the full potential of A-GPS, the unique connectivity features of Bluetooth and the simplicity and availability of Wi-Fi. Other technologies incorporated into geo-location services include TOA (Time of arrival), where a signal is sent from the device to different Cell towers, and each one of those send that data to the location server, which calculates the device's location according to the length of time the signal took to arrive to a tower [**45**]. Other techniques used to enhance geo-location precision involve AOA (angle of arrival), which calculates the angle at which the signal arrived to the receiver, and RSS (Received Signal Strength). However, it is generally recognized that TOA is the most precise of the complementary measurements for areas with low or no direct line of sight for GPS location services [**51**]. In addition, Wi-Fi positioning is one of the latest techniques used to enhance accuracy of location information. Wi-Fi positioning uses terrestrial-based access points (AP) and the information they broadcast to surrounding areas to determine their location, and the overlapping signals used by thousands of AP's today, especially in urban areas, create a natural grid that allows for location referencing [**50**]. As such, the evolution of geo-location services has been characterized by the convergence of multiple technologies that complement each other to improve the accuracy of location information.

Today, geo-location has evolved in a way that it is able to more accurately locate entities both indoors and outdoors, and even in underground locations such as mining sites through the use of wireless sensor networks in combination with LAN networks [**52**]. In addition, the accuracy and reliability of geo-location using various techniques and technology continues to grow. Today, high-accuracy geo-location methodologies capable of centimeter-level relative accuracy have been developed. For example, in a design based on a quadruple integration of the Global Positioning System (GPS), the inertial measurement unit (IMU) system, the terrestrial radio-frequency (RF) system pseudolite (PL), and terrestrial laser scanning (TLS) have been developed, which enhances high-accuracy geo-location for environments where GPS is limited [**53**]. These types of developments demonstrate the trend toward convergence of technologies to develop accurate, robust, and ubiquitous geo-location systems for multiple uses. As geo-location evolves, more technologies will converge to establish advanced geo-location.

4.2.2. Geo-location on Smartphones

The essence of geo-location and the mobile nature of Smartphones make the two technologies a strong match to complement one another. There are many applications today that were developed specifically for Smartphones, which heavily rely on geo-location technologies so that those can be valuable, and even useful. Today, the real significance of a high percentage of applications on mobile phones relies on the geo-location services they offer [**54**]. Even applications that do not implement geo-location services as their main feature use some kind of geo-location to enhance the service they offer. This is the case with social media applications such as Facebook, Twitter, and Instagram, or weather applications, all of which use geo-location services. In the specific case of this project, geo-location was not crucial for the development of BAT; however, it was of the essence for the development of its second iteration, Face2Face. This uncovers a particular importance on generating architectures that facilitate geo-location services on Smartphones and other personal devices.

Commonly, architectures used on Smartphones to generate location information involve a mobile device sending and receiving signals from a cell tower, and getting information from a satellite (GPS). This mobile device sends information to a database that holds location data, and processes the information, and sends back location coordinates to be displayed on the mobile device screen [**55**]. Nevertheless, as stated earlier, GPS connection is not necessary for localization. Some Smartphones applications and local infrastructures today are able to generate location data even without GPS.

Applications such as Google Maps on mobile devices use different techniques and protocols to achieve data connection including wireless networks and cell tower information. Smartphones then can send HTTP requests to different sources requesting location data. Thus, TOA or Wi-Fi positioning information can be received from this request. The data used to identify location can include cell tower id, location area code, mobile network code, or mobile country code [**56**]. These data can then be translated to coordinate data by posting an HTTP requests to geo-location databases such as Google's database, with cell data information. These databases can then return actual coordinate data. As stated earlier, while using cell data or signals for geo-location, does not prove to be as accurate as GPS, it can be used to enhance GPS location or estimate location in the absence of GPS. Once coordinate information is obtained, the device can easily locate and display the coordinates on a map for the user to identify the location.

One of the most common challenges encountered when developing geo-location technologies on Smartphones is energy consumption. Battery usage on the devices is extensive with this technology, which has been known as a common deterrent to the technology. Smartphones that use location services using GPS experience an enormous battery drain due to the fact that the device's application is constantly receiving and processing GPS signals for location updates [**49**]. In addition, as mobile platforms are increasingly being equipped with multiple radio interfaces to enhance geo-location operating in several different locations while constantly switching between them, connections ranging from Bluetooth for personal-area links, Wi-Fi for local-area connectivity, and cellular connections also place significant strain on battery usage [**57**]. With geo-location technologies using multiple ways for wireless data transfer to enhance detailed location information, battery consumption becomes a concern that cannot be ignored.

Problems such as using geo-location sensors when the device is static, the absence of power-efficient sensors, or having multiple applications that use geo-location services, impact the battery usage on Smartphones [**58**]. For these reasons, location-based application developers for Smartphones must utilize techniques that will make the most efficient use of energy on the Smartphone. In fact, when developing the BAT and the Face2Face applications, this author found energy utilization to be one of the most important challenges to overcome in creating a useful system. Several application designs and principles were put in practice for improving battery efficiency. For example, using techniques to determine when to turn on the GPS radio, and when to use other technologies for geo-location when the GPS was not the most accurate option (such as indoors, and in urban areas) is one of the identified techniques to conserve battery [**49**]. Using these and other techniques such as substitution, suppression, and piggybacking, has been demonstrated to improve battery life on Smartphones by up to 75% [**58**]. As such, with battery consumption being identified as one of the most important issues when developing location-based applications, it was important to determine several design techniques that could be used when building a system in order to achieve a more energy-efficient result.

In addition to battery usage, there were other issues when creating geo-location based services for mobile phones. The need for standardization and data integration is one of them. Geographic information needs to be common and standardized for any mobile device to be able to supply location-based services anywhere on the planet [**55**]. In addition, location-based services require a common database and centralized server that processes location information sent by mobile devices. This creates network congestion and poses questions as to what the capacity of GIS servers should be in order to handle an uncertain amount of mobile devices sending requests for location data processing [**55**].

There are also non-technical issues related to privacy and protection of private information, as people could potentially use collected information in a malicious and illegal way. Even though most of the time the data shared do not necessarily identify the user of the device, sometimes services could include names or other personally identifiable information, generating a clear risk for the device holder. There are issues regarding the use of geo-location data from a privacy perspective, raising questions such as who has access to the information once it is processed, how this information is used, and the timeline for retention and destruction [**59**]. These issues raise concerns for developers regarding design decisions that would minimize risks regarding data misuse. For those reasons, regulators (governing bodies) and legislation have been created in order to protect personally identifiable information that is stored on Smartphones, where telecommunications providers are specifically required to make efforts to protect such information [**60**]. These were a few of the design considerations, which make the process of creating a useful location-based service for Smartphones challenging.

4.3. From BAT to Face2Face

The Bluetooth Agent Tracker described in previous chapter evolved into an application that additionally implements geo-location technologies in order to accurately provide location information of a Smartphone device and captures this as part of the data collected. The second iteration of the application is called Face2Face. The enhanced system still uses the Bluetooth discovery process to scan devices in the vicinity, and additionally, it establishes the device's location information using the device's geo-location capabilities. The application component that handles the data transfer procedures, uses wireless communication radios (3/4G, Wi-Fi) to push the information to the server automatically; the server side also grew in capabilities, it can map the movement of different entities within a geographical area, and can establish aspects such as commuting paths, traffic trends and proximity with anticipated useful level of accuracy. During the application's Bluetooth discovery process, the system has the capacity to log information such as MAC address, and the device “friendly” name of any device that is within the area of range and has the Bluetooth discovery mode set to ‘on’ as well as a timestamp. Bluetooth protocol has been used as part of surveillance systems in the past to track specific devices throughout determined areas. As each device has its own unique 48-bit identifier, it is possible to then search for and locate other Bluetooth devices and their movement throughout a geographical area. The Face2Face application takes advantage of the fact that almost all the Smartphones today implement Bluetooth as a close range technology protocol, which allows a massive exposure of information if the Bluetooth radio is on and set on discoverable mode. Once these data are collected, they can be used to generate behavioral models, traffic simulations and in general, to determine the collective behavior of individuals in relation to one another in a defined space and time. It is clear that the intention of this researcher is to capture the behavior of masses rather than extracting potential private information from individuals.

Face2Face uses mobile technology to harvest contact or proximity data within a given workplace, social gatherings or organization in an automated fashion. The objective of the application is to generate enough data to be able to model person-to-person (Face2Face) interactions and subsequently simulate such interactions with a significant level of accuracy (who meets with whom, for how long, and where) in order to be able to simulate disease spread patterns, traffic behavior and/or other interactions between individuals. The application developed could be used even by a relatively small number of participants as probes, or a representative sample of the larger population, allowing them to “vampire” proximity data from a variety of consumer electronic devices inclusive of other Smartphones. These ideas are well entrenched in the wired world and used by system administrators as a means of monitoring their networks. The Face2Face probes are capable of monitoring a user’s social network as well as their sub-social network (proximity contacts that a person has not explicitly made but that the probe device has detected).

The Face2Face Smartphone application was prototyped on five probe devices that maintain explicit location data when available, by having the device GPS-enabled, augmented with connection attempts to close-proximity devices that are discoverable via Bluetooth (or equipped with a Bluetooth transceiver that is on and/or discoverable). These connection attempts to close-proximity, Bluetooth-enabled devices log information including: date and time; MAC address (BD_ADDR); any user information or device meta-information a person may have provided (inadvertently or not); and, geographic location if the probe device is GPS-enabled. MAC address and meta-information are logged for both the probe device and the close-proximity (discovered) device. Meta-information refers to factory-programmed or user-programmed device IDs, e.g. “BlackBerry 5660” or “Jay’s iPhone”, respectively. The data collected is then logged to a database where it can be mined for contact durations, distributions and associations.

Initially there was not enough data to estimate the effectiveness of the proposed technique in terms of collecting proximity data. There are a number of issues related to radio control and Bluetooth pairing in question. These devices used for the experiment support Bluetooth v2.0 with consumer applications primarily being hands free operation and wireless stereo headsets. As most people now are prohibited from using a cell phone while driving, the Bluetooth headset is most often discoverable. Thus, even if the handset is not discoverable, the probe applications are able to discover a user’s accessories. There is really no means of circumventing or preventing someone from being able to attempt a Bluetooth connection once the device is discoverable. In all cases, at some level of a standard protocol some information is necessarily sent in plain sight. In many cases even if the device is not discoverable, it can still be found as long as its Bluetooth radio is on. In terms of identifying the device there are look-up engines and repositories that are readily available to help infer the device type. In the case of the BlackBerry devices used with this project, a device can only be detected if the Bluetooth option is set to discoverable. Figure 21 describes the Face2Face automated process.

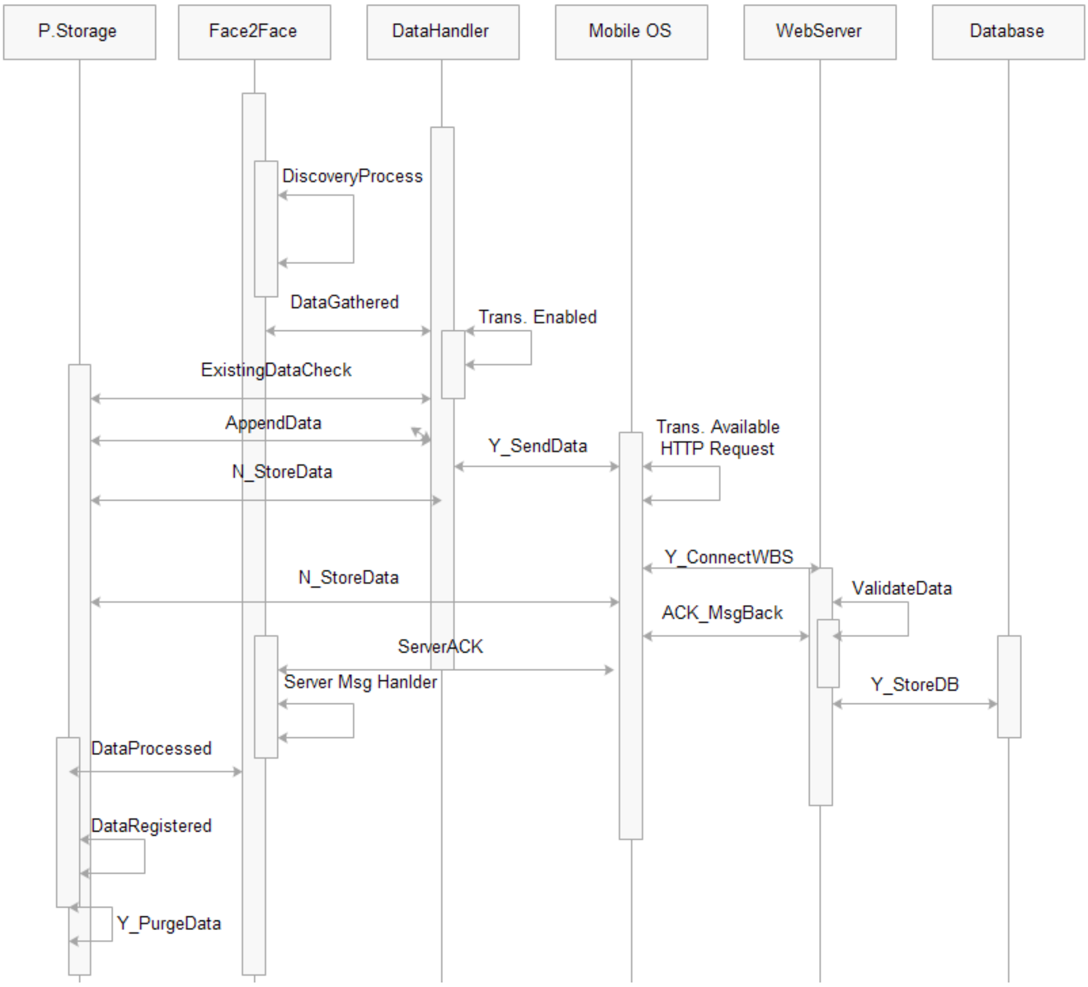
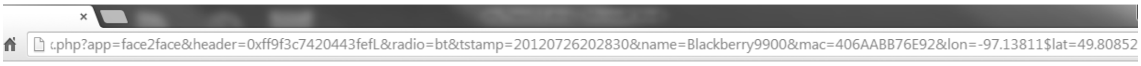


Figure 21 - F2F Sequence Diagram

4.3.1. F2F Web Service and Data Handling Process

For the second iteration of the Bluetooth Agent Tracker, some changes were included in the way the data were handled from the Smartphone and the server perspective. It was in the best interest to have a clean, non-intrusive data transmission process using wireless connectivity and a common web data transmission standard that will prevent performance degradation on the user's end so it was decided to push the information using http headers to reach the server and an on-demand script on the server that would be capable of interpreting the request from the Smartphone device through the HTTP GET method. This method was chosen because it was desirable to have a straight forward way to generate requests from several Smartphone technologies without having to figure out a way to interpret the internal data handling differences on each Smartphone operating system. Pushing a text string (URI address) had the most balanced efficiency-complexity ratio for achieving this.

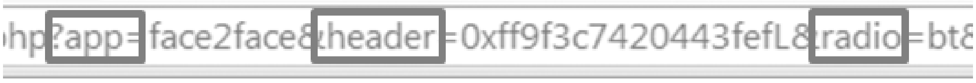


```
‹.php?app=face2face&header=0xff9f3c7420443fefL&radio=bt&tstamp=20120726202830&name=Blackberry9900&mac=406AABB76E92&lon=-97.13811$lat=49.80852
```

Figure 22 - F2F URI String Sample

Figure 22 displays the string that was built by the different Smartphones using Face2Face to push data gathered through the Web Service enabled on the IIC server. As the string is quite long the URI will be separated to analyze its composition.

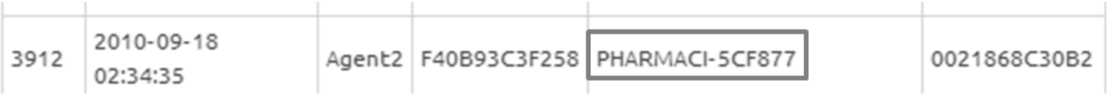
The figure below describes the first three attributes of the string: the application, the header, and the radio. The first attribute that notifies the server that the application that will be invoked is Face2Face (attribute value face2face) because the web server was in fact providing services to several applications that used similar approach for pushing information; having this attribute allowed the script to define the correct database connection parameters for further data insertion.



```
hp?app=face2face&header=0xff9f3c7420443fefL&radio=bt&
```

Figure 23 - F2F Attributes Sample I

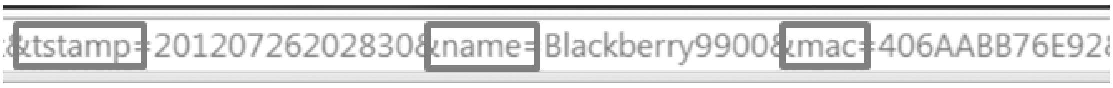
The second attribute, the string Header (attribute value 0xff9f3c7420443fefL) is a unique hex value pre-shared by the server to the Smartphone when it starts a session; this value is verified by the server before data are pushed into the database to ensure data consistency; it was decided to use this measure because using the GET method for handling information is not the most secure channel to do so, and people could potentially use this to inject or retrieve data from the database; having a valid header allows data transactions to be processed in the database. The third attribute is the Wireless Radio (attribute value bt or Bluetooth). This attribute was defined because the scope defined for next releases of the application would also include data gathering capabilities for other radios such as Bluetooth Low Energy, Wi-Fi and/or NFC. In this particular release, Wi-Fi information was available but it was decided not to work with it because the research was aiming to exploit the Bluetooth radio as much as possible and including additional information about other radio protocols could lead the research to lose focus and subsequent relevance.



3912	2010-09-18 02:34:35	Agent2	F40B93C3F258	PHARMACI-5CF877	0021868C30B2

Figure 24 - F2F Meta Information gathered

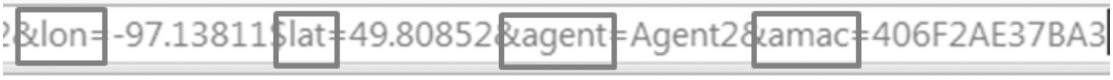
The next three attributes were: *timestamp, name* and *mac.* The timestamp was included to get the specific moment the devices were discovered for subsequent analysis; the format was stripped from any non-numeric symbol to make it easier to implement in any platform and it was re-structured on the server side (i.e. 20120726202830 → YYYYMMDDHHMMSS → Web Service → YYYY-MM-DD HH:MM:SS → 2012-07-26 20:20:30 → Database). The next attribute, name also referred as “Friendly Name” was included to gather some meta-information about the devices gathered. Much of the time, users tend to use that name to describe their personalities, roles, or other forms of personal information that could be further analyzed. A good example of this is shown in Figure 24. The third attribute was the mac (referred as MAC address). This was and still is the most important attribute we can gather from the spectrum because it gives a unique value to the entity being discovered. There cannot be two Bluetooth devices with the same MAC and the information subscribed within that mac can give us a good understanding of what type of device is being discovered.



```
&tstamp=20120726202830&name=Blackberry9900&mac=406AABB76E92&
```

Figure 25 - F2F Attributes Sample II

The final sets of attributes were: *lon*, *lat*, *agent*, and *amac*. The first attribute *lon*, also referred to as Longitude was the first piece of information for us to be able to enable geo-location capabilities into the analysis; *lat* also referred to as Latitude gave us the second piece to establish the approximate location of an individual at a certain time. The APIs for Blackberry and Android gave two extra attributes to be exploited (Altitude and Speed) but they were disregarded in this iteration to avoid unnecessary data processing on the Smartphone side. The last two attributes, *agent* and *amac*, were also pushed to determine which agent was retrieving the information provided for future analysis; the mac was also pushed to verify that it was a registered agent and to avoid possible erroneous data injections.



```
2&lon=-97.13811&lat=49.80852&agent=Agent2&amac=406F2AE37BA3
```

Figure 26 - F2F Attributes Sample III

One final aspect associated with the automated means of contact data collection is that contact with non-mobile objects can also be collected and analyzed. For example, devices that are somewhat stationary such as desktop workstations are often discoverable without being able to infer that a person is associated with the object. These however allow for landmarks to be identified that are useful in localization of the probe devices.

4.4. Uses and advantages of the Face2Face application

Incorporating actual location based services considerably enhances aspects of the probe application and adds value to the applications identified previously that involve social contact networks, including epidemic modeling and traffic modeling. Bluetooth location services have been used for several purposes in the past, including geo-location services. For example, one of the first times Bluetooth was used for entity positioning and tracking was at the largest zoo in Denmark: the Aalborg Zoo. A Bluetooth system was installed to make tags available to parents who wished to easily locate their children within the park. This made parents able to attach tags to their children to track the child's movement within the zoo. Through Bluetooth receivers that were installed throughout the zoo that would trace each of the tag's movements [**61**]. Bluetooth surveillance has also been used for business purposes to trace customer trends and preferences. However, the combination of geo-location information and information that is shared through Bluetooth discovery, several new objectives can be achieved, such as spatial location, proximity among individuals throughout time. These early positioning and tracking systems were individual-centric as in the examples above; subsequent generations as well as Face2Face are attempts at systems that are tending to be more appropriately considered cohort-centric. At this time the usefulness of cohort-centric technologies are seen "through a glass darkly".

The Smartphones that were used to test the Face2Face application supported GPS as well as assisted GPS services. Within this context, wherever and whenever possible, location based data is appended to the proximity records. Users who run the Face2Face application immediately are able to obtain information regarding contacts that are in proximity as they move around a geographical area such as a campus or a city. The graphical record of the contacts in proximity can then be illustrated as an XML mash-up overlaid on Google Maps.

With the introduction of GPS and assisted GPS technologies to the Bluetooth location services developed on the application, there were several limitations that were overcome. For example, the introduction of A-GPS allowed a fairly accurate device positioning (less than 1.5m radial error). In addition, once a device was located, the proximity duration was measured by the amount of time the device was reachable by the Agent; the greater the time it was reachable, the longer the device was placed in relation with the Agent.

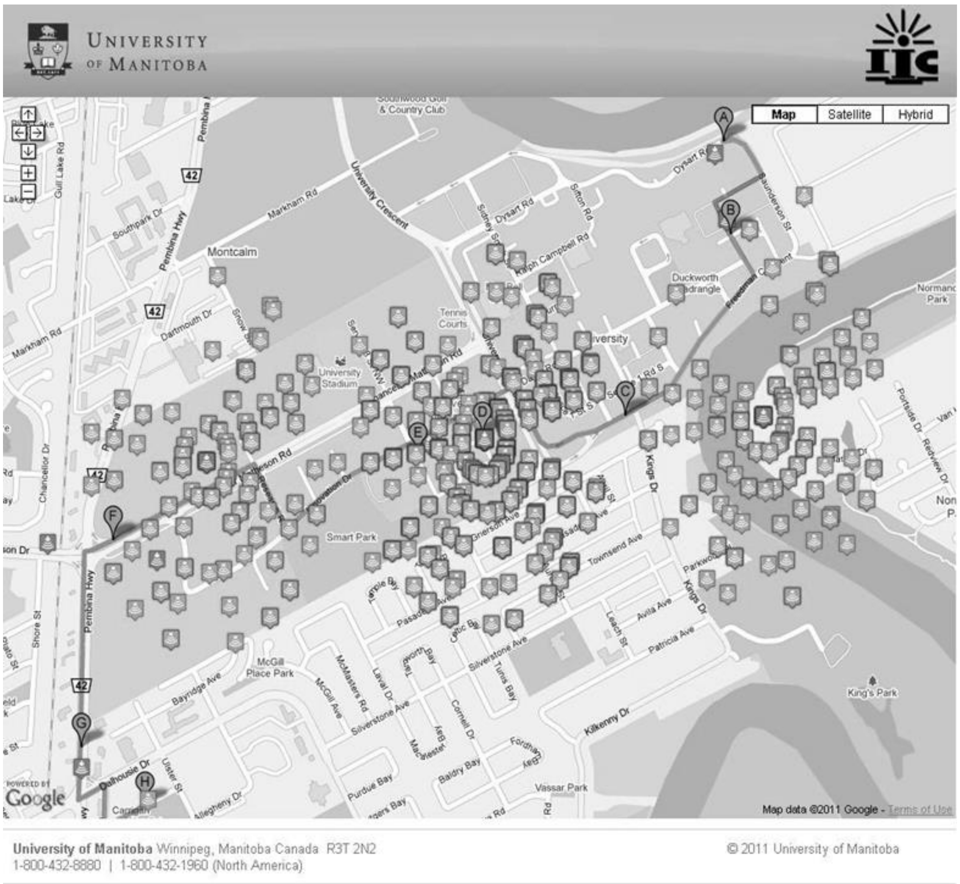


Figure 27 - Agent's Path

In addition, from an efficiency perspective, using GPS to integrate it with the BAT does not incur major additional costs due to the fact that the GPS technology has long been developed, and implemented on almost every Smartphone device available in the market. Implementing an application that makes use of location technologies does not impact its usability and adoptability because it is embedded within the current infrastructure, it can, however impact other aspects such as energy consumption.

Face2Face is able to collect data and map the movement of devices used as probes. For the specific development of this research, data was collected using the HTC Hero probe device associated with one of the participants. Figure 27 illustrates a mash up of proximity contact data incorporating spatial as well as temporal data collection. Figure 27 also illustrates some of the inaccuracies of this particular instance which used the HTC Hero as the data collection agent (the geo-location data as well as the devices in proximity).

The Face2Face application provides the ability to store data collected on the user's mobile device. In addition, data can be backhauled over the Internet to a database and subsequently analyzed or displayed. For the purposes of the research, data that is stored on users' devices would be voluntarily provided. This application enhances the usability of technologies consolidated to allow for flexible geo-location services. Location over several different types of terrain, as well as indoor and outdoor are made available by the convergence of technologies. The use of application such as Face2face can be a cornerstone of multiple services and applications including traffic, epidemic, or migration simulations, by applying the concept of crowdsourcing discussed in previous chapters of this thesis. In addition, business intelligence can be enhanced through more accurate and immediate tracking of consumer movements and trends. For example, shopping malls could make use of Face2Face applications to trace potential customer movements throughout the mall. The system could generate a precise map of customer's path and even record information of areas where customers spend more time as well as those in close contact with. These capabilities certainly enhance and make an evolution of the concept of a digital footprint, whereby the digital footprint of a person will go beyond the trail left by visiting websites, shopping patterns on mobile commerce, or interaction on social media applications. Sensors can be used to register the activities of an individual in addition to collecting personal information such as heart rate, body temperature, and even physical social interactions [**62**]. This concept will now additionally include physical proximity to other individuals, patterns of movement within specific geographical areas, speed of movement, and friends who are constantly in physical proximity to them.

4.5. Chapter Summary

The use of Bluetooth and geo-location technologies clearly enables spatial location, proximity among individuals across time, and movement and speed of the movement of individuals. Face2face was the second iteration of the Bluetooth Agent Tracker, and it was not the final version of it. Including A-GPS services combined with Bluetooth discovery capabilities provides an incredible opportunity to implement the system in critical congregations of people such as on mobs, concerts, festivals, or sporting events. This can easily be achieved as an increasingly larger percentage of people tend to carry a Bluetooth and/or a GPS-enabled device. The final evolution of this application is FriendFinder, which allows the application to be used as more of a “social” Agent. The FriendFinder application allows for the traceability of interactions among individuals who are in proximity, and it will be described in detail on the next chapter.

The application that provided a foundation to this project, the Bluetooth Agent Tracker, evolved into an application that uses geo-location technologies to broaden the usability and to enhance the value of the application. Geo-location technologies involve the use of Global Positioning System-enabled devices to converge with other technologies in order to establish accurate, precise, and meaningful location information. Assisted GPS technologies (A-GPS) use the cellular infrastructure to aid in establishing location and reducing the time and processing requirements to do so. Due to the high level of wireless infrastructure available today including Wi-Fi, , and infrared technology, multiple techniques can be used to enhance geo-location information beyond GPS. Geo-location information today has become a foundation of many services, which include navigation, emergency services, commercial services, recreation, tracking and networking [**50**].

In an effort to provide an example to highlight the capabilities of the technologies described above, Peterson presents an interesting approach that discusses a similar methodology and system for obtaining traffic information from mobile Bluetooth detectors [**63**]. The mobile Bluetooth detector transmits the collected data to a remote facility where the data are processed to generate traffic information. This shows a clear trend in the research community towards using these types of technologies to gather aggregate information from a large number of individuals.

CHAPTER 5

FriendFinder Big Data Analysis

5.1. Overview

Previous chapters of this thesis have presented the evolution of an application that uses different types of technologies evolving towards the third and final iteration: the FriendFinder application. The system followed an iterative approach, starting from an idea to implement a technology that allowed the collection of large amounts of crowd-sourced data through the use of mobile computing technologies, and specifically, Smartphones as end devices. The initial iteration of the system presented in this thesis was the Bluetooth Agent Tracker (BAT), which involved the use of Bluetooth technology to track nearby devices that are within range of the Bluetooth network of the scanning device. The second iteration was the Face2Face (F2F) application, which included the implementation of geo-location technologies to enhance the application’s ability to provide accurate, more-detailed information about individual commuting patterns combined with proximity. In that mode, fairly accurate approximate location and basic contact identification could also be gathered with the application when the Bluetooth discovery process was engaged. This chapter describes the third and final iteration of the system prototype, called FriendFinder; it describes in detail how the technology concepts applied to previous iterations of the system can now be put into practice and extended with an emphasis on a personal social network application.

This chapter provides a detailed description of the FriendFinder application, including a description of its functionality, how the technology is applied to the application, how the application would work in real-life scenarios, and what type of information would be gathered. Then, a brief explanation of the "Big Data" concept is provided along with an initial analysis of the relationship between its concept and nature, and the data model used while working with the different releases of the Bluetooth Agent Tracker.

Following that, data gathered via the various versions of the application is analyzed, including some computational algorithmic simulation to fill the data gaps uncovered during the data assimilation process. Finally, a conclusion of the work will be presented, including a summary of the project, and considerations for future work that can be based on this study.

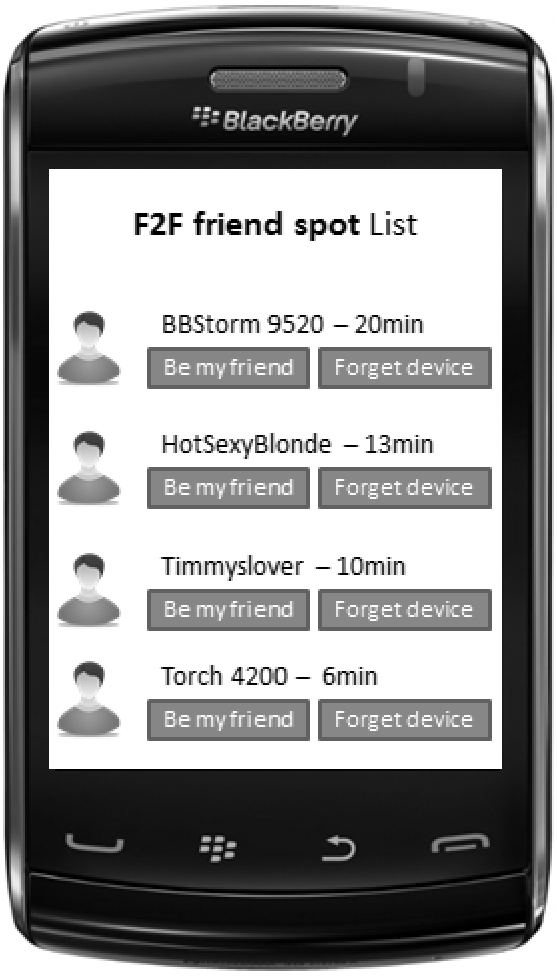
5.2. FriendFinder

As presented in the previous chapter where the Face2Face (F2F) application was introduced, the next iteration of this system is an ability to identify and make real connections with people who are in close proximity. This release is code-named FriendFinder. Developing a system in different iterations provided several benefits. First, it allowed one to resolve smaller development and deployment problems such as hardware/OS compatibility and address issues early in the project; in addition, it allowed the entire development process to be “much more flexible and able to address new requirements throughout the project” [**64**]. As such, this project started from an abstract concept such as crowdsourcing and evolved through three iterations of evolution, , and this allowed the researcher to uncover the real value of the research. Based on the same principles and concepts used for the BAT and F2F systems, which include the use of crowd sourced data, Bluetooth identification of nearby devices, and high-accuracy geospatial localization (F2F), the FriendFinder application evolves as another application of the same concept, which provides the user with an interface to engage in real peer-to-peer social relationship interactions based upon geospatial location and proximity of other users.

To reiterate the evolution of the application to the current iteration, it's important to recall that the objective of the Bluetooth Agent Tracker was to find out how many times within a specified timeframe a Bluetooth device made an appearance and was registered by the application. Data collected through this identification could be used to establish social patterns such as number of people in close proximity. The second iteration of the system then was the F2F application, which introduced geo-location capabilities to the already established Bluetooth infrastructure in order to obtain more accurate location information about devices and movement patterns. In the F2F application, agents could identify a person/device in close proximity only when they checked the database contacts list (Figure 14) which was an after-the-fact process, not in real time. FriendFinder evolves into an application that provides a more social aspect and is a system conceived to trace and simulate real-life social activities among individuals. FriendFinder makes use of the previously built system features such as Bluetooth identification and geo-location, to allow a person to walk into any crowded place, and identify other individuals in the same place who (i) have their Bluetooth devices in discoverable mode and (ii) are running the same application, and with whom the person could potentially make real digital contact.

The FriendFinder application is explained through the exposition of a real-life social situation or use case. Suppose that a random person arrives into a crowded space where other people are scattered, such as at a sporting event. This person is using a Bluetooth-enabled mobile device running the FriendFinder application, which is be able to identify and communicate with devices that have the same technology capabilities. FriendFinder can search for, and identify nearby Bluetooth devices, establish a connection with those who also have FriendFinder installed, and determine its own geospatial location to provide an accurate scenario for further analysis. Due to the fact that a Bluetooth pairing connection (required to secure peer-to-peer communication between devices) is available and engaged, the proximity between devices cannot exceed the Bluetooth protocol range specs which places the digital interaction as a "close" or in-range interaction that will be considered close proximity during the data analysis process. In addition, consider then that in the place where the person arrives, there is another person who also has FriendFinder installed.

Once the FriendFinder application establishes the existence of other devices running the FriendFinder app in the same location, it notifies the users of the proximity of other devices. For the purposes of actually finding relevant connections through this application, the FriendFinder application would only notify users when a connection is established for a long enough period of time; therefore, discovered devices that are in range for short periods of time and then disappear would be disregarded. As Murphy et al. conclude, Bluetooth technologies are suitable for these types of device-to-device connections [**65**]. When a user has been in the room for a period of time, a list of the devices that have been continually discovered by their device can be provided along with certain information – these could be considered potential FriendFinder friends.



F2F friend spot List

BBStorm 9520 – 20min

HotSexyBlonde – 13min

Timmyslover – 10min

Torch 4200 – 6min

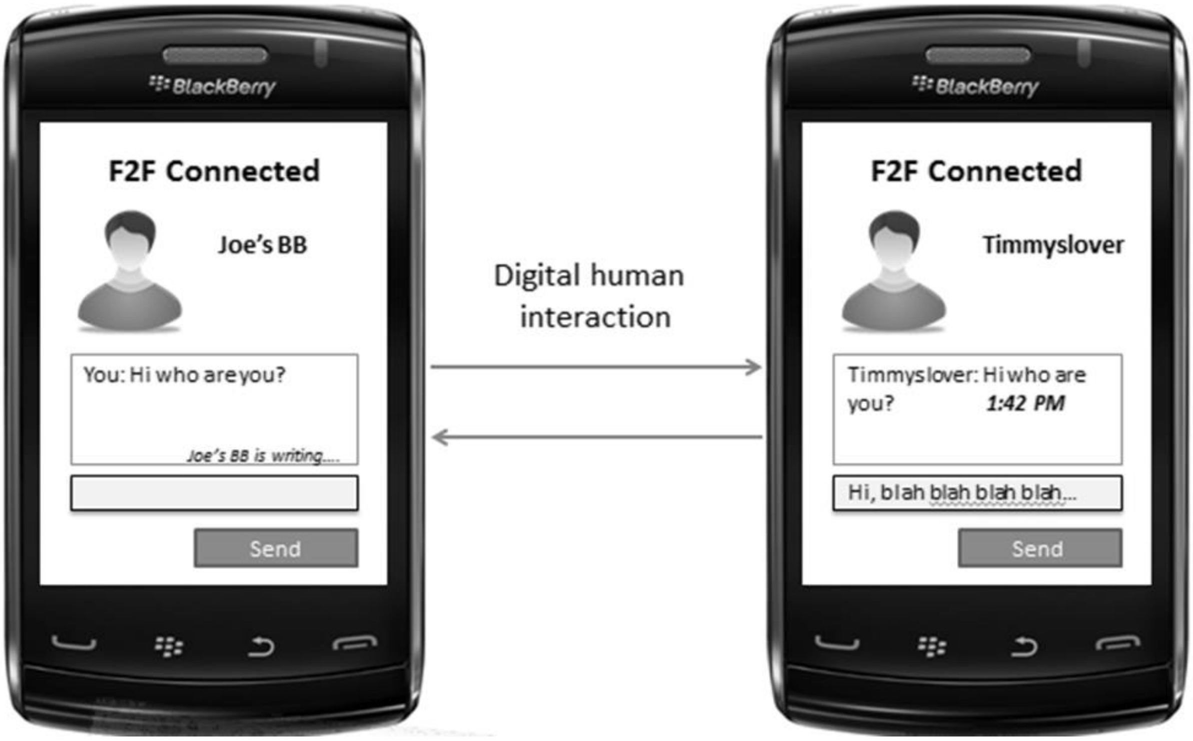
Figure 28 - Friend Spot List Screen

Privacy is a very important concern in these types of applications, and broad adoption of a mobile application today highly depends on the extent to which privacy concerns are adequately addressed [**66**]. For the purposes of maintaining privacy within the application, no personal information about the owners of either of the devices would be presented. The only information that would be presented would be the device *friendly* name, and the length of the connection, as shown on Figure 28.

When the list of users within range has been presented to the user, the user can select a contact with whom he or she intends to establish a social connection. Subsequently, the device would try to pair the device of the user who made the selection with the device of the user whose profile was selected. Then, a request would be sent to the selected person to make a connection, in a process similar to requesting a Bluetooth pairing of the devices. An important security feature of the FriendFinder app was that the connection and subsequent communication would not involve any service provider or network interaction.

If the person whose profile was selected accepts the connection, the devices are paired, and then direct interaction can begin. Both users can now communicate in a text-based manner. Once this connection is established, one user can request further information from the other, and the information will be shared only if the user accepts.

Two users can remain connected while exchanging messages. Once the communication has been maintained for a period of time, and both users present an interest in one another, they can share information that can progressively get more personal, such as sharing a picture of each other, or other information about them such as age or gender, or simply agree on a meeting place. Naturally, if the communication continues to progress, the two parties, knowing that they are in close proximity to one another, would want to follow up with a direct, personal connection. In this way, Friend Finder has generated, enabled, and initiated a real-life relationship.



You: Hi who are you?

Joe's BB is writing....

Timmyslover: Hi who are you? 1:42 PM

Hi, blah blah blah blah...

Figure 29 - FriendFinder Connection Established

FriendFinder represents an evolution over other mobile text messaging applications such as BBM or Whatsapp in various aspects. For instance, other applications depend on a server connection that manages the communication between devices. FriendFinder applies the principle of direct connection between the two devices. As such, the information that is shared only resides on each other's devices without storing the information in any other third party servers. In addition, the transition from a purely digital to a physical interaction is entirely controlled by both parties, and such transition can occur immediately after the digital interaction has begun.

FriendFinder represents a significant evolution to the concept of social networking in particular from the perspective of data ownership. Modern social network theory is different from traditional definitions of social interaction theory where societal groups were built up of human individuals. By contrast, social networking starts from the relationships between individuals and models social interactions as constituted of networks made up of sets of the relations or ties between the nodes [**67**]. Traditionally, in social network theory, ties between nodes, which can be represented by individuals, companies, families, or in general any group of people, were formed by some type of natural ties such as a close knit tie or an actual social contact. It was always direct contact by an individual that initiated the interaction, either by directly approaching a second person, or by an actual physical introduction by other individuals. Technology has enhanced the ability to build social networks through electronic social media tools such a Facebook and twitter, or mobile applications such as Whatsapp. However, the newly proposed technology implemented through applications such as FriendFinder enhance the generation of new relationships with individuals that are in close proximity, and potentially suggest new connections for the user. The application has the capability of sorting through other individuals that are in close proximity based on a simple decision-making script to display the devices that have been localized on the sample place, or nearby for a certain amount of time. A modified, enhanced version of FriendFinder could potentially establish the probability of achieving a successful relationship even before contact has been made (to be described later). In addition, this new technology allows users to find new connections and start social interactions in places that do not necessarily provide an internet connection such as airports, concerts, and areas not covered by a wireless networks. because the application relies on peer-to-peer connections made autonomously between two devices.

CHAPTER 6

Analysis and Application Trials

6.1. Trial run with the application

The following section provides descriptive statistics of the information that was gathered during the trials. Descriptive statistics summarize the information in various ways but do not provide statistical analysis of correlational or causative relationships or inter-dependencies. Since the purpose was to collect a significant volume of data via crowdsourcing, descriptive statistics highlight non-trivial features of the data collected in a quantitative way. In this way, the methodology used in this project could be compared with similar work in an effort to make the results and conclusions of the work more robust and supported by statistical evidence.

6.1.1. Data Collection

During the development of this framework, a data gathering process was conducted, applying the concepts covered in this thesis which included crowdsourcing, Bluetooth identification, and geo-location. Using the Bluetooth Agent Tracker (BAT), which was introduced in Chapter 2, the researcher collected data for several months by installing BAT on five mobile devices (*agents)*, and collecting broadcast data about devices within range.

The initial data collection project involved a simple form of collecting and analyzing data to generate contact graphs and distributions of personal social networks. Data collected was gathered using the BAT application (Chapter 2). The data are representative of the ensemble of many consumer electronic products and can be used to infer considerable information in terms of proximity, contact duration, as well as geographical location information. In order to augment data, meta-information as well as some landmarks on the terrain was used for localization. Data generated from networks of personal contact are then demonstrated to be of utility in estimating the potential of the spread of an infection such as respiratory diseases where a vector of transmission is the proximity to infected agents.

The data collected in this trial was stored in a database table which included six attributes identifying the data. The data collected included the following structure:

Table 2- BAT Data Fields

Field Name	Data Type	Description
ID	Number	Unique ID of the record within the database
Date	DateTime	The date and time when the connection was made
Agent	Varchar	The number of the agent (1 to 5) that was used to make the connection
Agent Mac	Varchar	MAC address of the agent that made the connection
Device	Varchar	Name of the Bluetooth device that made the connection
Device Mac	Varchar	Mac Address of the device with which the agent established connection

The data are presented in terms of distributions and visualization tools for evolving contact graphs. Some weeks the experiment was run continuously, and some other weeks the experiment was run with interruptions. Due to the nature of the experiment, the quantity of data collected was not uniform on certain days. For example, there could be differences in the number of people encountered on different days. On Sundays, public transit buses are a lot less crowded than they are on other days of the week, where at least one of the agents was a regular public transit user. Another reason for data not to be uniform is that on certain days some devices were left home, and data was not collected during those days. Even for this very simple trial, compliance was an issue. This is one of the characteristics of crowd sourced data, and as such, when making an attempt to generate a time series graph that shows data collected, where the function of the number of contacts collected by devices for each day is diagramed, the graph will contain outliers as days when no or very little data is collected. Perhaps the greatest reason that data was sporadic was the entirely voluntary nature of the BAT application. As it was initially developed as a BlackBerry prototype for reasons discussed previously, the volunteers essentially had to carry a Smartphone with no additional features or data plan. As such it was difficult to enforce any compliance to regularly collecting and reporting. This is a primary difficulty with any system that collects data with little benefit seen by the data provider and/or through a device that has no other purpose. In any case, it was still desirable to demonstrate that the data still held value even if irregularly collected by a very small group. An additional difficulty with the BAT application was that it was initially fairly cumbersome to upload data collected. As with any type of crowd sourced application, it is highly desirable to be as noninvasive as possible, similar to versions based on the Android (F2F).

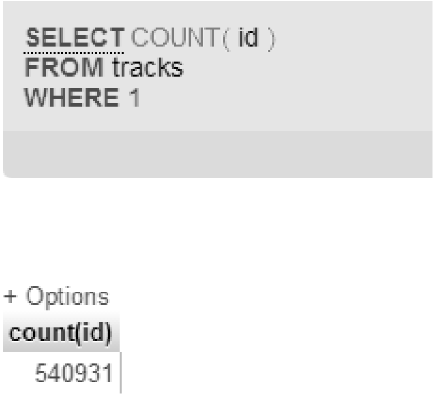
In an attempt to use the data gathered to predict values for datasets not recorded (not in the past nor in the future), a Monte Carlo simulation process was developed to calculate non nil data values for timeframes in which the trials performed had very few or no records at all. This was done in an effort to generate relevant data that could enable meaningful deductions from the analysis. For example, the n^{th} series could be calculated, and such series could yield a computed number of contacts after a large number of tests (>10000) predicting future or past contact patterns with a good level of accuracy. Using the Monte Carlo method, the value obtained could be used in order to fill the gaps on the original graph, and it could then be stated with a certain degree of confidence that the resulting numbers represented the number of devices contacted on a particular day by each device (agent) that this experiment employed if it had been active. The basic premise is that the gaps in the data were artificial and thus could justifiably be filled in. Consequently, as a result of habitual activities, gaps in the record can be synthesized. The suitability of this process is context-specific and may not be justified in all situations. In addition, the trend with personal communication devices like Smartphones is that they are quite often always carried or in proximity of the owner, making the requirement for filling in gaps even less necessary in a full-scale implementation of the application.

Filling in gaps within data collected is somewhat precarious; leaving known erroneous gaps in data can lead to misleading interpretations and predictions and introducing synthesized and artificial data can also lead to similar problems. [**68**] discusses means of filling in gaps in the data associated with practical time series forecasting. Similar ideas are presented here although admittedly the data set used here is at the edge of where artificial data may be reliably synthesized.

6.1.2. Trial data analysis

Based on the fact that data gathered during the first rounds, which as explained in previous chapters was collected using five mobile devices (agents) using F2F during several weeks, large samples of data, scattered through the weeks, were not being recorded or reported due to system inactivity. This led the researcher to consider a Monte Carlo series approach to minimize nil-value records on the dataset in an effort to avoid data gaps caused by system inactivity from altering further statistical analysis. The gaps shown in the data gathered during the application trials could represent several scenarios including having the Smartphone off or leaving the device at home

The first approach was to verify the exact amount of records registered on the BAT/F2F database, and the data limits by date. For this, two simple queries were performed (Figure 30 and Figure 31):

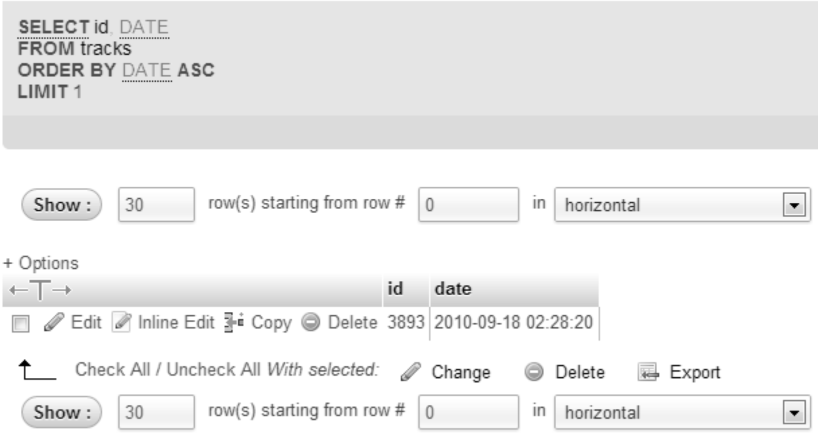


```
SELECT COUNT( id )
FROM tracks
WHERE 1
```

count(id)
540931

Figure 30 - Total amount of records F2F

The total number of records (contacts) registered in the BAT/F2F database was 540,931 (Figure 30). This in itself was considered quite surprising as there were only five trackers (probes) with considerably variation in recording and reporting compliance.



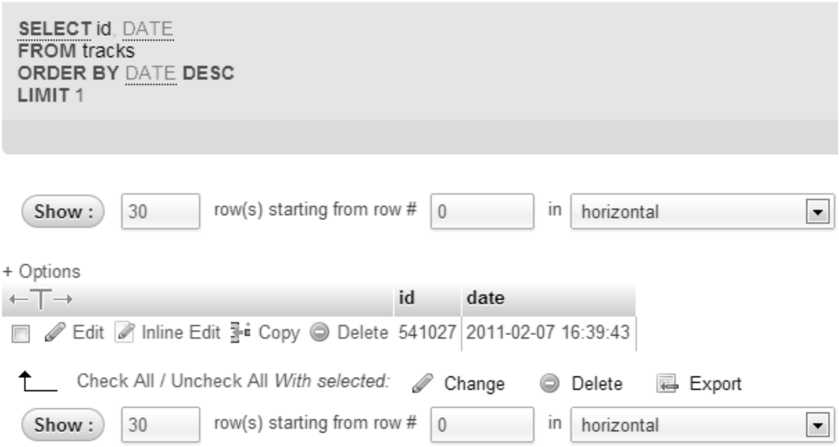
```
SELECT id, DATE
FROM tracks
ORDER BY DATE ASC
LIMIT 1
```

Show : 30 row(s) starting from row # 0 in horizontal

	id	date
□ Edit Inline Edit Copy Delete	3893	2010-09-18 02:28:20

Check All / Uncheck All *With selected:* Change Delete Export

Show : 30 row(s) starting from row # 0 in horizontal



```
SELECT id, DATE
FROM tracks
ORDER BY DATE DESC
LIMIT 1
```

id	date
541027	2011-02-07 16:39:43

Show : 30 row(s) starting from row # 0 in horizontal
+ Options
←T→
☐ Edit Inline Edit Copy Delete
Check All / Uncheck All With selected: Change Delete Export
Show : 30 row(s) starting from row # 0 in horizontal

Figure 31- F2F oldest and newest record

The limit dates were September 18th and February 7th (almost 5 months), so in preparation for characterization of the data collected, the data was grouped by day/week/month to be able to identify contact patterns at first sight. This step was undertaken to see if the data contained expected or stylized patterns based on days of the week, and constituted an initial descriptive rather than statistical analysis of the dataset. An SQL query along a PHP script was written to encapsulate and group the data.

The most general view of the data was the overall count by day. The researcher wanted to assess the contact trend per day overall to see which days were "stronger" in terms of contactability (Table 3; Figure 32).

Table 3 - Contacts per day (overall)

Day of the week	count(id)
Sunday	30,768
Monday	84,712
Tuesday	110,179
Wednesday	105,032
Thursday	98,779
Friday	79,566
Saturday	31,895
TOTAL	540,931

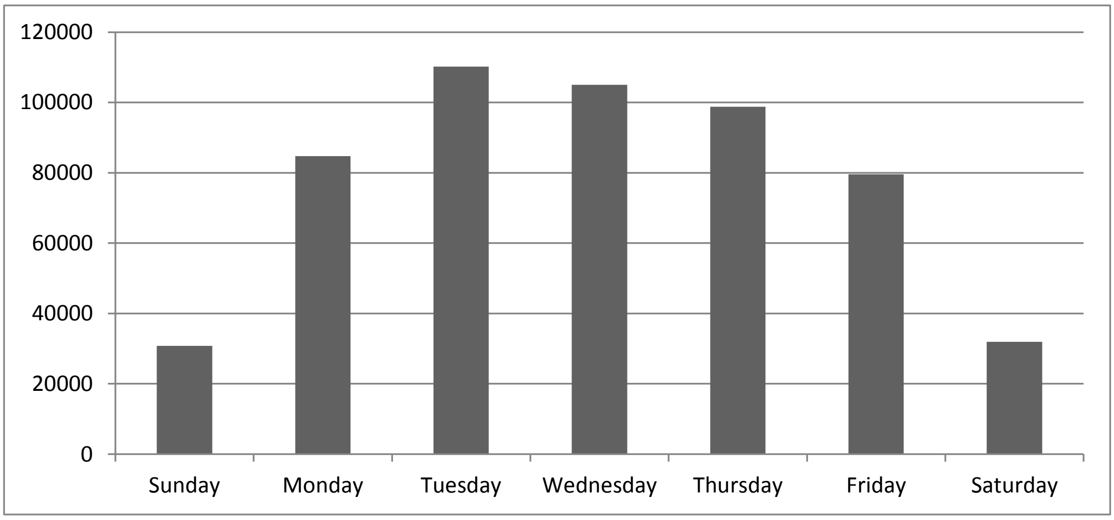
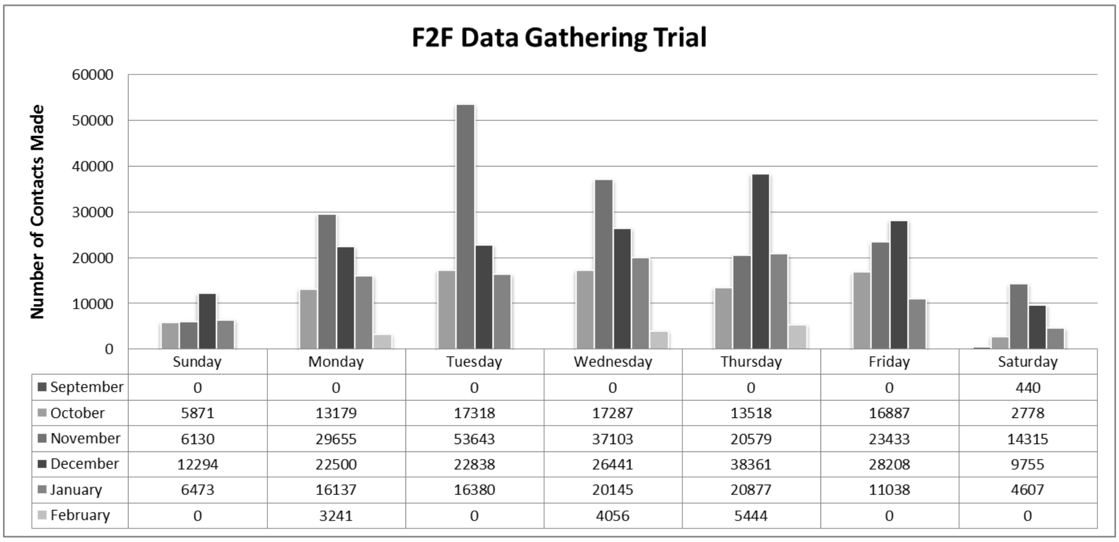


Figure 32 - Total contacts per day (overall)

The results illustrate highest contact ranges mid-week and lower ranges over the weekend as expected A finer-grained analysis was required in order to get a better qualitative characterization of the data. Had this type of weekly pattern not been found, the data collected would have been considerably more questionable, based on the fact that individuals (in this case, agent) are not always together nor they are always in social interaction scenarios.

Data were queried by month, then by week, and then by day over the entire period of collection (Figure 33-45),

By Day/Month



	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
■ September	0	0	0	0	0	0	440
■ October	5871	13179	17318	17287	13518	16887	2778
■ November	6130	29655	53643	37103	20579	23433	14315
■ December	12294	22500	22838	26441	38361	28208	9755
■ January	6473	16137	16380	20145	20877	11038	4607
■ February	0	3241	0	4056	5444	0	0

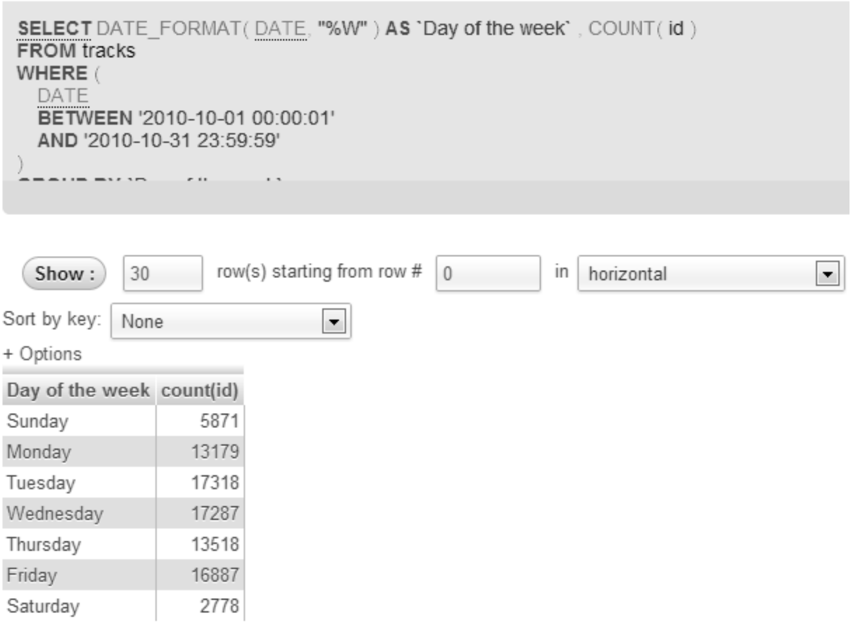
Figure 33 - F2F Grouped by Day/Month

September

The first month of the trial only showed 440 contacts made on Saturdays and no other records on any other day of the week. These data was collected by one tracking device during testing by the author. This is an outlier month value on the overall measure and the end values would be affected accordingly.

October

October presents a good data sample as the contact data is scattered through the days of the week and the average contacts per day are relatively large. Based on the amount of data gathered per day in this month, the values of this month will be added to other significant values from other months in order to generate enough synthetic data using the Monte Carlo approach to back fill values for the months with data gaps.



```
SELECT DATE_FORMAT( DATE, "%W" ) AS `Day of the week` , COUNT( id )
FROM tracks
WHERE (
   DATE
   BETWEEN '2010-10-01 00:00:01'
   AND '2010-10-31 23:59:59'
)
```

Show : 30 row(s) starting from row # 0 in horizontal

Sort by key: None

Day of the week	count(id)
Sunday	5871
Monday	13179
Tuesday	17318
Wednesday	17287
Thursday	13518
Friday	16887
Saturday	2778

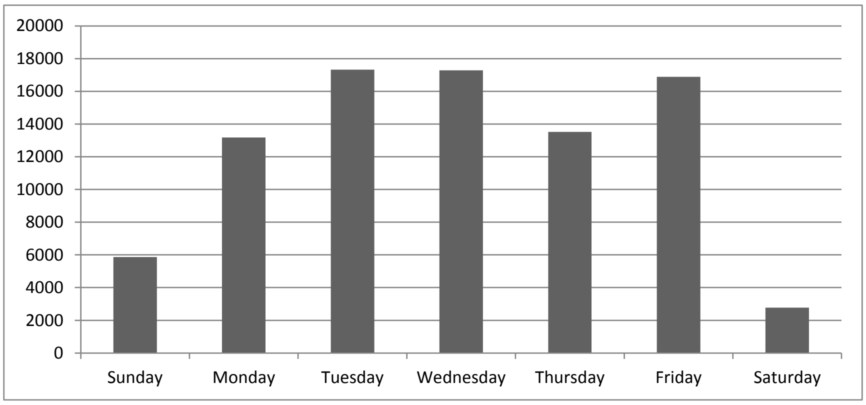
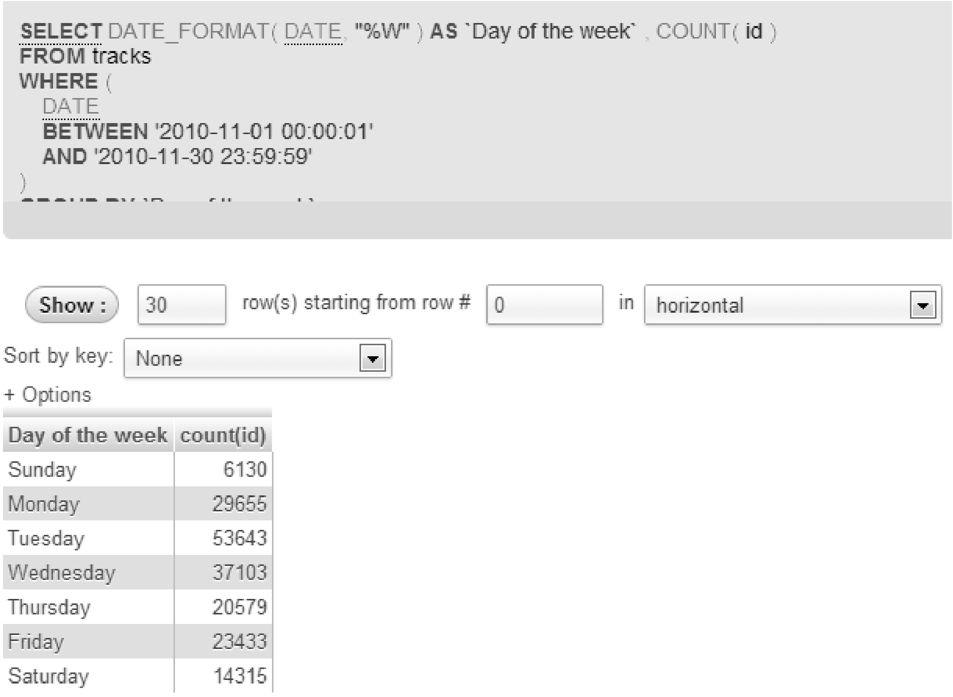


Figure 34 - F2F October Query/Graph

November

This month shows a disparity of contacts during the week/weekends, nevertheless the amount of data is enough to help create the series along with October values. A potential anomaly is possibly Thursday.



```
SELECT DATE_FORMAT( DATE, "%W" ) AS `Day of the week` , COUNT( id )
FROM tracks
WHERE (
   DATE
   BETWEEN '2010-11-01 00:00:01'
   AND '2010-11-30 23:59:59'
)
```

Day of the week	count(id)
Sunday	6130
Monday	29655
Tuesday	53643
Wednesday	37103
Thursday	20579
Friday	23433
Saturday	14315

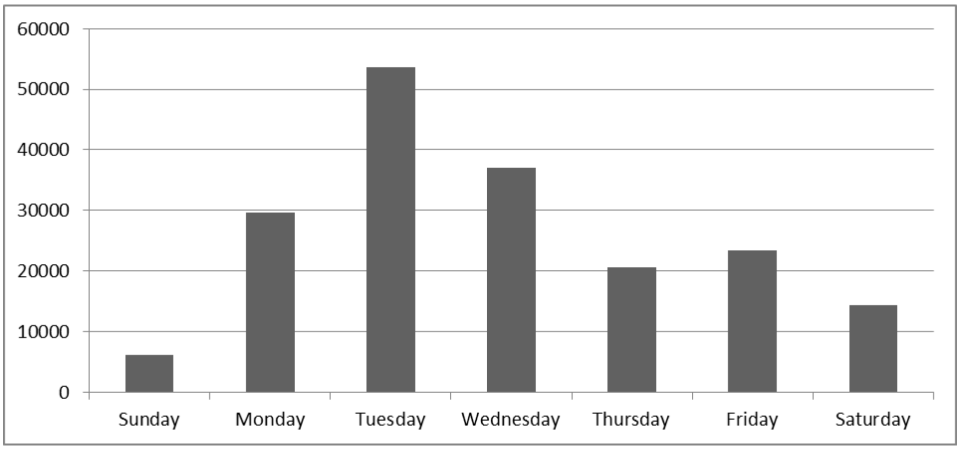
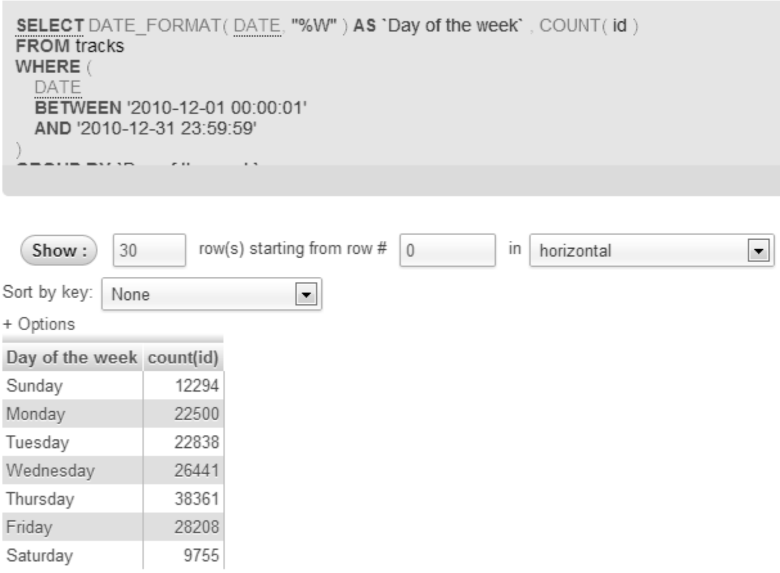


Figure 35 - F2F November Query/Graph

December

December likely presents a valuable set of data for the overall data gathering trial, both for the number of records generated and for their distribution across the collection period, with almost 160,400 contacts in the month. Again it appears to be consistent with stylized expectations, except Thursday perhaps presenting a qualitative difference compared to previous months. This set of data could be even more detailed (contacts by hour or time of the day) to evaluate contact patterns according to the time of the day.



```
SELECT DATE_FORMAT( DATE, "%W" ) AS `Day of the week` , COUNT( id )
FROM tracks
WHERE (
  DATE
  BETWEEN '2010-12-01 00:00:01'
  AND '2010-12-31 23:59:59'
)
```

Show : 30 row(s) starting from row # 0 in horizontal

Sort by key: None

Day of the week	count(id)
Sunday	12294
Monday	22500
Tuesday	22838
Wednesday	26441
Thursday	38361
Friday	28208
Saturday	9755

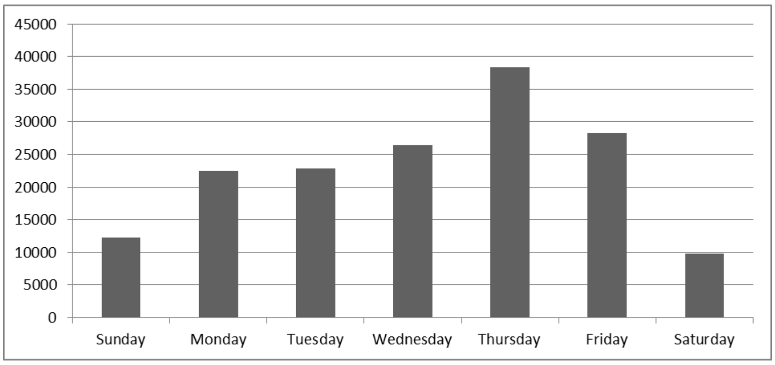
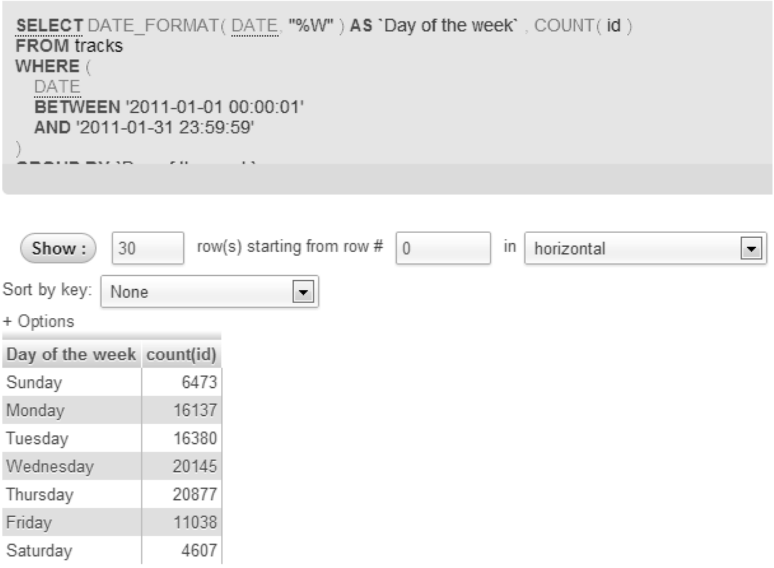


Figure 36 - F2F December Query/Graph

January

The January data also included one of the five probes attending two NHL Jets games which were both held on Thursday which may account for some of the reasons why Thursday appeared to have slightly more contacts than other months.



```
SELECT DATE_FORMAT( DATE, "%W" ) AS `Day of the week` , COUNT( id )
FROM tracks
WHERE (
  DATE
  BETWEEN '2011-01-01 00:00:01'
  AND '2011-01-31 23:59:59'
)
```

Show : 30 row(s) starting from row # 0 in horizontal

Sort by key: None

+ Options

Day of the week	count(id)
Sunday	6473
Monday	16137
Tuesday	16380
Wednesday	20145
Thursday	20877
Friday	11038
Saturday	4607

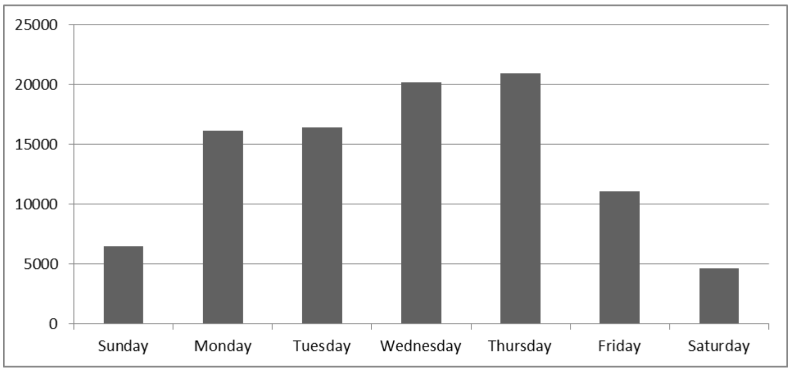
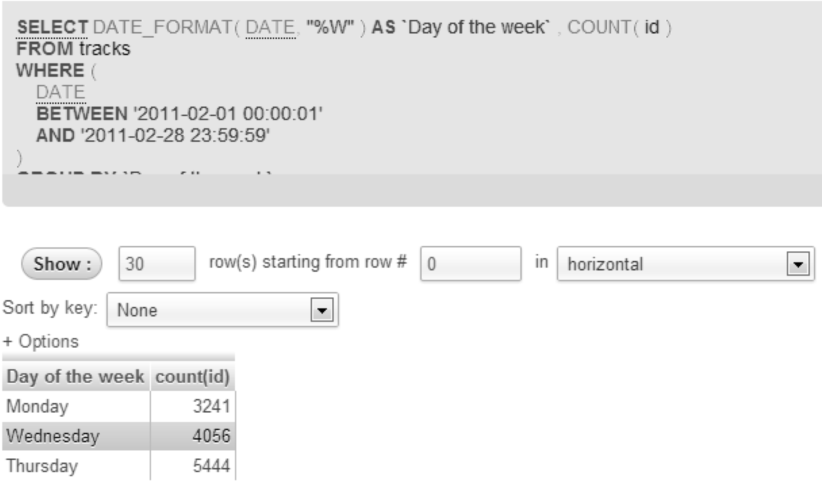


Figure 37 - F2F January Query/Graph

February

The February dataset provides a useful target for the Monte Carlo simulation because it has a fairly large amount of contacts registered through the month, distributed through the weekdays in a similar way to other months, which will most likely minimize error introduction while trying to predict significant contact values. February also represents the month when least compliance from volunteers was maintained which contributed to the anomalous nature of February's data



```
SELECT DATE_FORMAT( DATE, "%W" ) AS `Day of the week` , COUNT( id )
FROM tracks
WHERE (
   DATE
   BETWEEN '2011-02-01 00:00:01'
   AND '2011-02-28 23:59:59'
)
```

Day of the week	count(id)
Monday	3241
Wednesday	4056
Thursday	5444

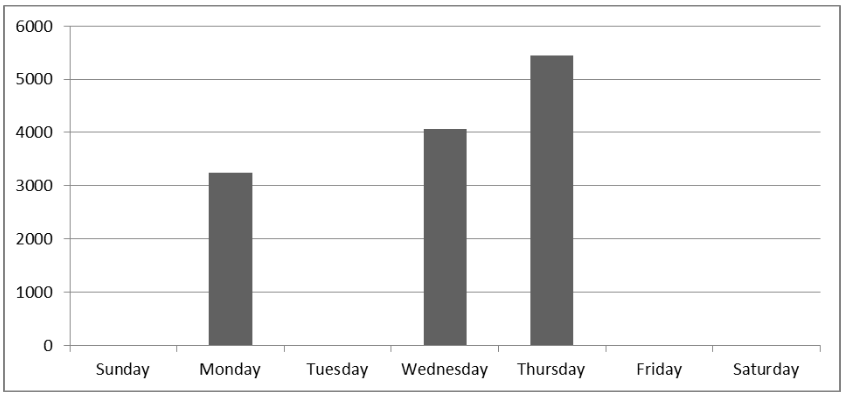


Figure 38 F2F February Query/Graph

Table 4 - F2F Data provides the numeric information of the contacts made by F2F agents during the applications trials during the four to five month period.

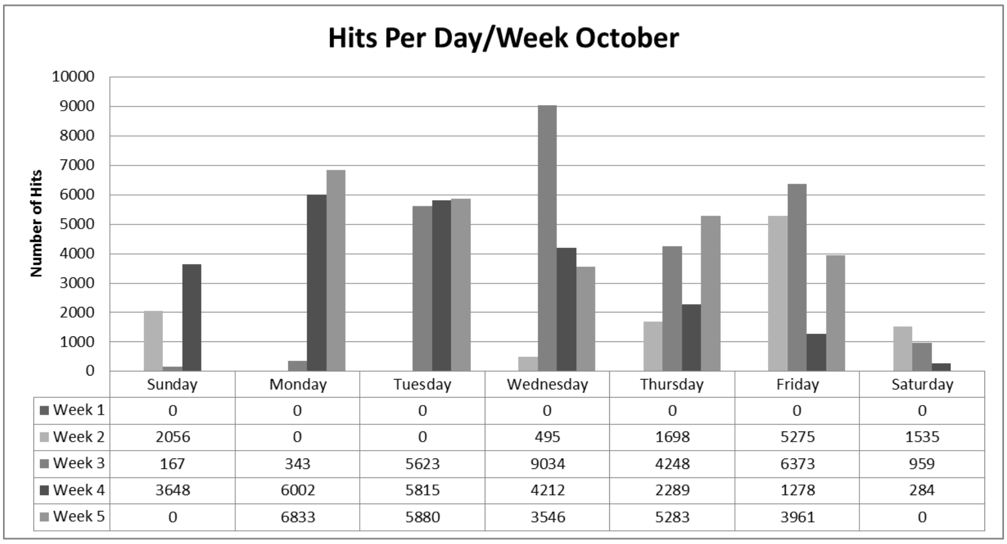
Table 4 - F2F Data

	September	October	November	December	January	February

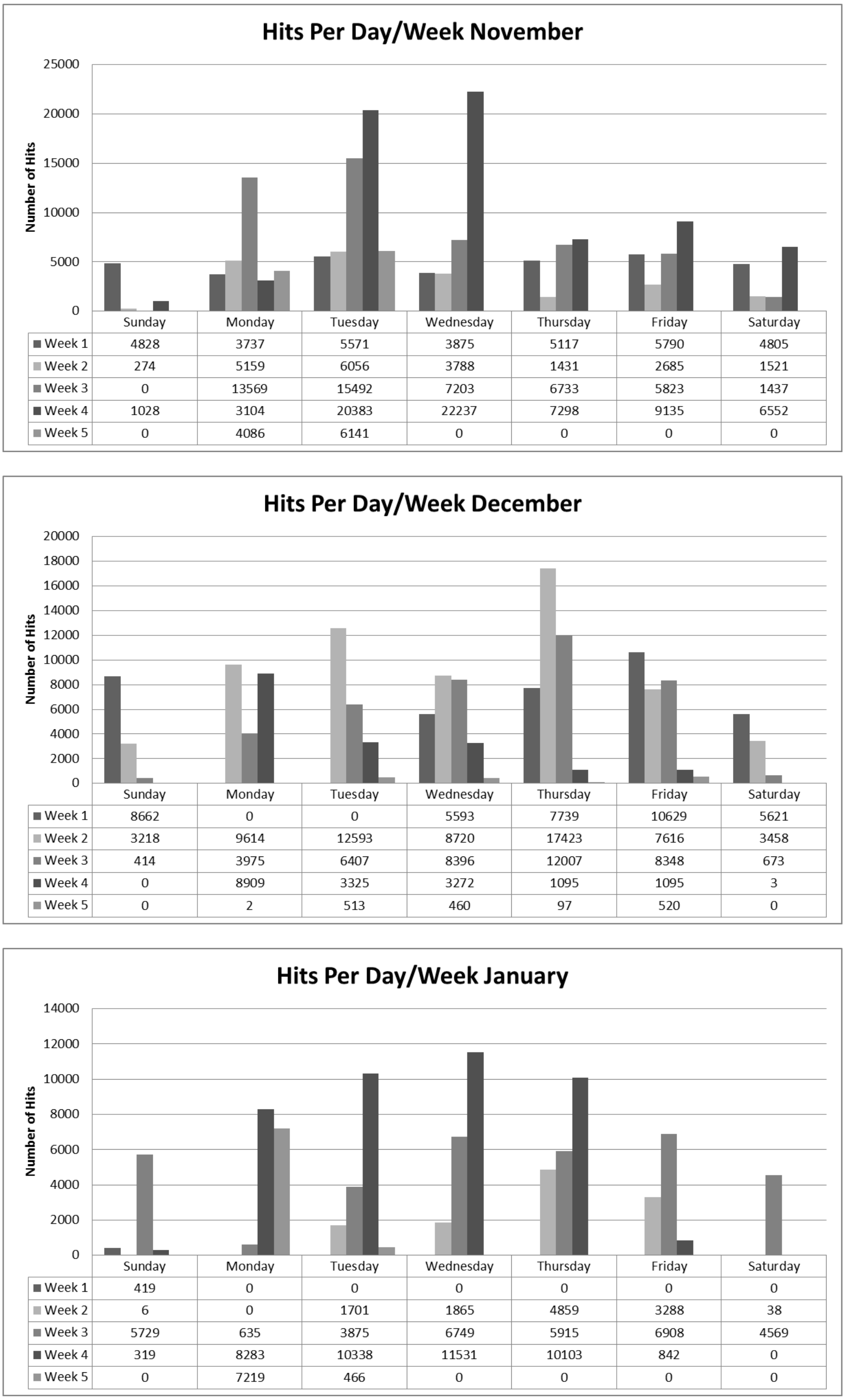
Sunday	0	5871	6130	12294	6473	0
Monday	0	13179	29655	22500	16137	3241
Tuesday	0	17318	53643	22838	16380	0
Wednesday	0	17287	37103	26441	20145	4056
Thursday	0	13518	20579	38361	20877	5444
Friday	0	16887	23433	28208	11038	0
Saturday	440	2778	14315	9755	4607	0
Totals	**440**	**86838**	**184858**	**160397**	**95657**	**12741**

Similarly, the data were analyzed by week to discern the most significant months data-wise; according to the table above the months to be taken into account should be October to January. The information is summarized in an attempt of not repeating steps or information that has already been presented.

By Day/Week



	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
■ Week 1	0	0	0	0	0	0	0
■ Week 2	2056	0	0	495	1698	5275	1535
■ Week 3	167	343	5623	9034	4248	6373	959
■ Week 4	3648	6002	5815	4212	2289	1278	284
■ Week 5	0	6833	5880	3546	5283	3961	0



	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
■ Week 1	4828	3737	5571	3875	5117	5790	4805
■ Week 2	274	5159	6056	3788	1431	2685	1521
■ Week 3	0	13569	15492	7203	6733	5823	1437
■ Week 4	1028	3104	20383	22237	7298	9135	6552
■ Week 5	0	4086	6141	0	0	0	0

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
■ Week 1	8662	0	0	5593	7739	10629	5621
■ Week 2	3218	9614	12593	8720	17423	7616	3458
■ Week 3	414	3975	6407	8396	12007	8348	673
■ Week 4	0	8909	3325	3272	1095	1095	3
■ Week 5	0	2	513	460	97	520	0

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
■ Week 1	419	0	0	0	0	0	0
■ Week 2	6	0	1701	1865	4859	3288	38
■ Week 3	5729	635	3875	6749	5915	6908	4569
■ Week 4	319	8283	10338	11531	10103	842	0
■ Week 5	0	7219	466	0	0	0	0

Figure 39 - Summary Hits per Day/Week Oct to Jan

While the detail of the data gathered during the application trial would allow an analysis by hour/day, this was not undertaken here. Generating approximated values using the Monte Carlo method is a rather straight-forward approach to generate artificial and statistically significant values; however the risk of introducing error into the analysis is inherent. In this case, it was known that the lack of data in certain time frames was evidently caused by non-compliance with carrying the device rather than actual and legitimate agent inactivity, and these data absence could probably introduce even greater error to the overall analysis. This first-hand knowledge of the trial agents' compliance justified the application of the Monte Carlo simulation in this case.

The gaps were filled independently on each run and then included on the next run to help generate new values. This iterative method was used to fill all the gaps found on the probe.

The following steps were implemented:

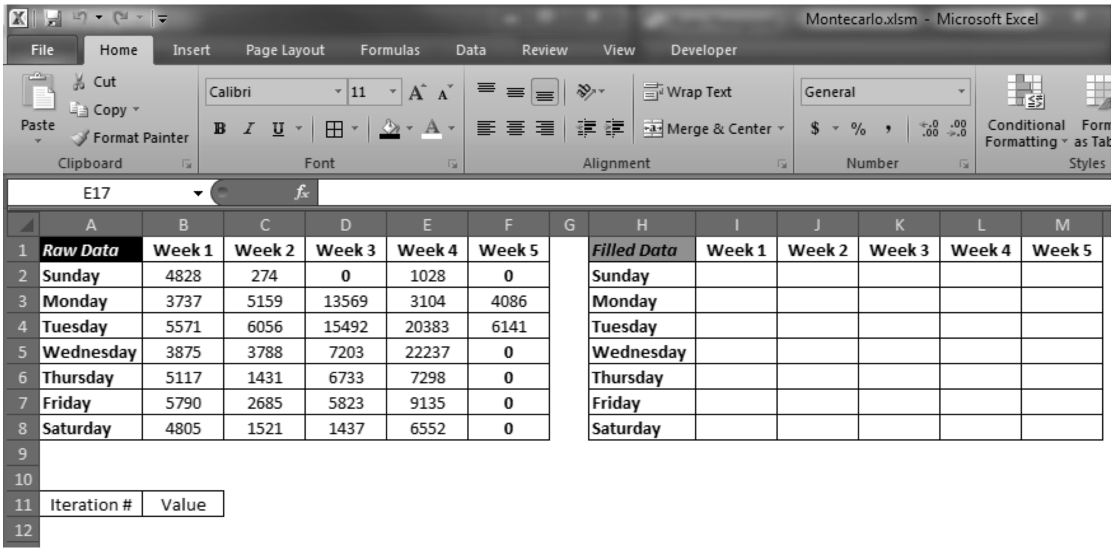
1. Determine a set of records of the day to work with that show an average value, and use it as a seed to start the Monte Carlo simulation
2. Calculate upper-lower limits based on the seed values
3. Determine the required number of iterations required to achieve an acceptable level of error in the calculation (2%)

   $$\varepsilon = \frac{3\sigma}{\sqrt{N}} \rightarrow N = \frac{9\sigma^2}{\varepsilon^2}$$

   Where ε represents the error, σ is the standard deviation of the random variable, and *N* is the number of iterations.
4. Assuming that our set of data is normally distributed, use a Random function (i.e. rand ( ) in excel) to generate random numbers between the upper and lower limits already calculated.
5. Execute the script to generate the desired amount of iterations.
6. Once calculated (averaging the values generated on step 5) verify that the calculated values doesn't exceed the maximum error percentage defined.

7. Iterate

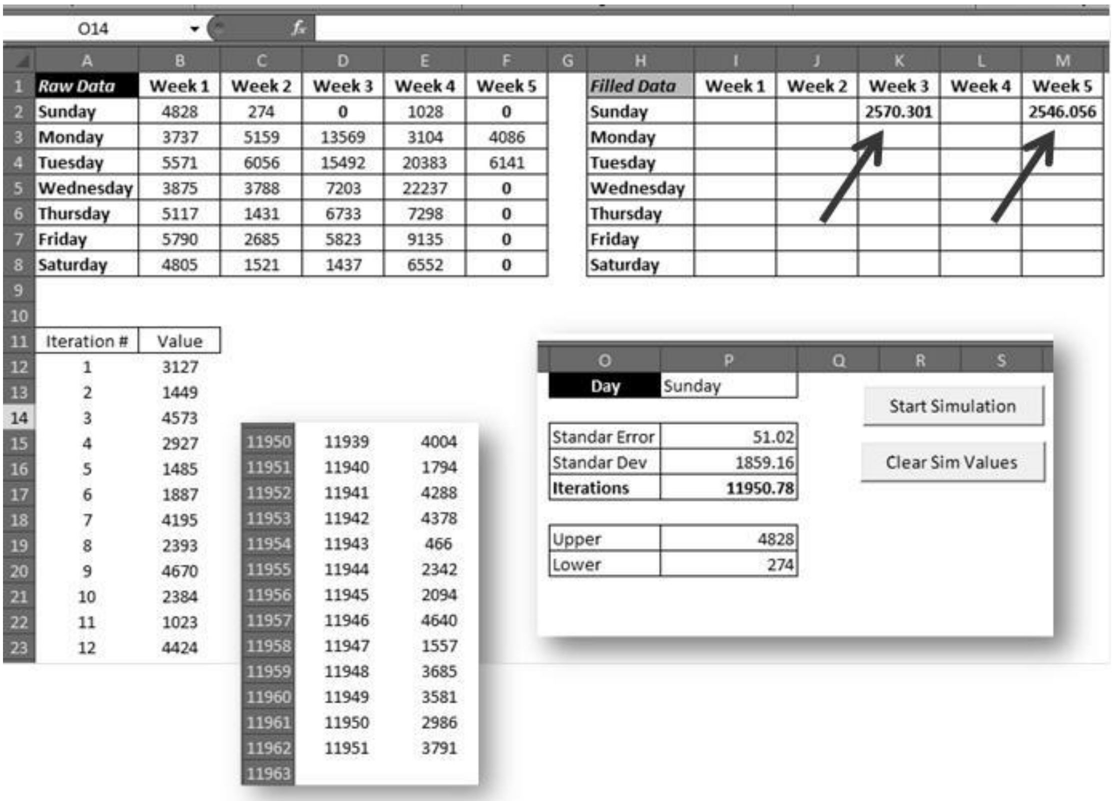
The simulation was done using VBA embedded on an excel sheet because the simulation didn't require massive processing power to fulfill the objective. The simulation was run for each month of the probe selected and the results along with a description of the process are presented below:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	*Raw Data*	Week 1	Week 2	Week 3	Week 4	Week 5		*Filled Data*	Week 1	Week 2	Week 3	Week 4	Week 5
2	Sunday	4828	274	0	1028	0		Sunday					
3	Monday	3737	5159	13569	3104	4086		Monday					
4	Tuesday	5571	6056	15492	20383	6141		Tuesday					
5	Wednesday	3875	3788	7203	22237	0		Wednesday					
6	Thursday	5117	1431	6733	7298	0		Thursday					
7	Friday	5790	2685	5823	9135	0		Friday					
8	Saturday	4805	1521	1437	6552	0		Saturday					
9													
10													
11	Iteration #	Value											
12													

Figure 40 – Monte Carlo Simulation

The vba script used to generate the simulation focused on the empty values for any given day, and based on the calculated values for upper and lower limits, standard error and upper bound standard deviation were identified by calculating the standard deviation between the upper, lower and average values of the seeded data. After selecting any given day, the script would then calculate the number of iterations (generating random numbers between the upper and lower limits) and after that it would calculate the average number on the "filled data" table. Basically an attempt was made to make Mondays resemble typical Mondays, Tuesdays resemble typical Tuesdays etc.



Raw Data	Week 1	Week 2	Week 3	Week 4	Week 5
Sunday	4828	274	0	1028	0
Monday	3737	5159	13569	3104	4086
Tuesday	5571	6056	15492	20383	6141
Wednesday	3875	3788	7203	22237	0
Thursday	5117	1431	6733	7298	0
Friday	5790	2685	5823	9135	0
Saturday	4805	1521	1437	6552	0

Filled Data	Week 1	Week 2	Week 3	Week 4	Week 5
Sunday			2570.301		2546.056
Monday					
Tuesday					
Wednesday					
Thursday					
Friday					
Saturday					

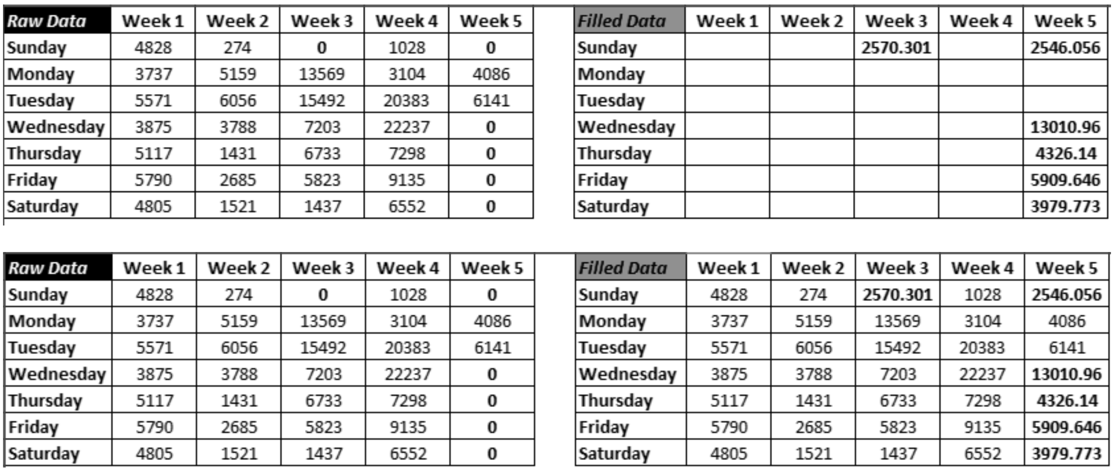
Iteration #	Value
1	3127
2	1449
3	4573
4	2927
5	1485
6	1887
7	4195
8	2393
9	4670
10	2384
11	1023
12	4424

11950	11939	4004
11951	11940	1794
11952	11941	4288
11953	11942	4378
11954	11943	466
11955	11944	2342
11956	11945	2094
11957	11946	4640
11958	11947	1557
11959	11948	3685
11960	11949	3581
11961	11950	2986
11962	11951	3791
11963		

Day	Sunday
Standar Error	51.02
Standar Dev	1859.16
Iterations	11950.78
Upper	4828
Lower	274

Figure 41 - Simulation Run

Once finished, the simulation successfully filled the zero valued fields for all the days / months in the probe, providing another view of the data after being transformed.



Raw Data	Week 1	Week 2	Week 3	Week 4	Week 5
Sunday	4828	274	0	1028	0
Monday	3737	5159	13569	3104	4086
Tuesday	5571	6056	15492	20383	6141
Wednesday	3875	3788	7203	22237	0
Thursday	5117	1431	6733	7298	0
Friday	5790	2685	5823	9135	0
Saturday	4805	1521	1437	6552	0

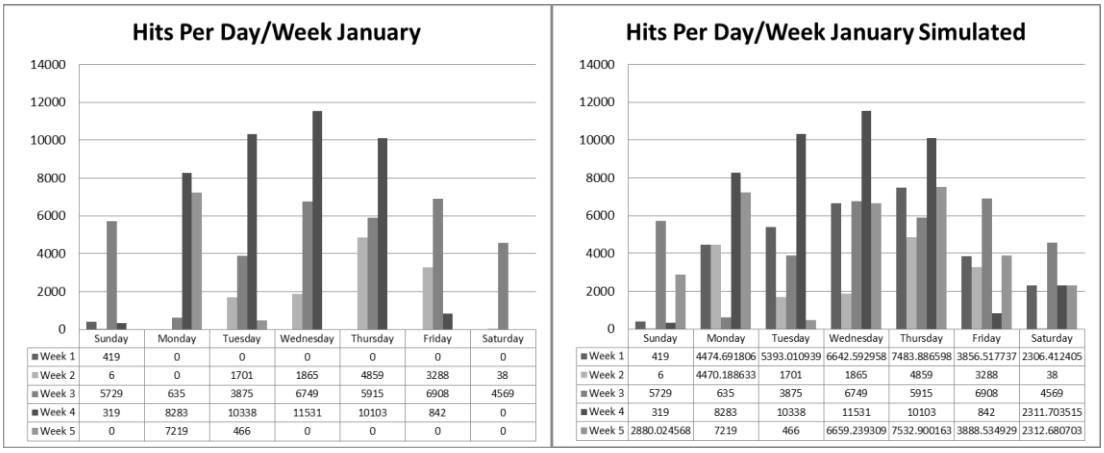
Filled Data	Week 1	Week 2	Week 3	Week 4	Week 5
Sunday			2570.301		2546.056
Monday					
Tuesday					
Wednesday					13010.96
Thursday					4326.14
Friday					5909.646
Saturday					3979.773

Raw Data	Week 1	Week 2	Week 3	Week 4	Week 5
Sunday	4828	274	0	1028	0
Monday	3737	5159	13569	3104	4086
Tuesday	5571	6056	15492	20383	6141
Wednesday	3875	3788	7203	22237	0
Thursday	5117	1431	6733	7298	0
Friday	5790	2685	5823	9135	0
Saturday	4805	1521	1437	6552	0

Filled Data	Week 1	Week 2	Week 3	Week 4	Week 5
Sunday	4828	274	2570.301	1028	2546.056
Monday	3737	5159	13569	3104	4086
Tuesday	5571	6056	15492	20383	6141
Wednesday	3875	3788	7203	22237	13010.96
Thursday	5117	1431	6733	7298	4326.14
Friday	5790	2685	5823	9135	5909.646
Saturday	4805	1521	1437	6552	3979.773

Figure 42 - F2F Data transformed

The following samples describe the transformation made to the data for several months; however the simulation was run for the entire dataset, filling the zero value days in an effort to try to homogenize the dataset.



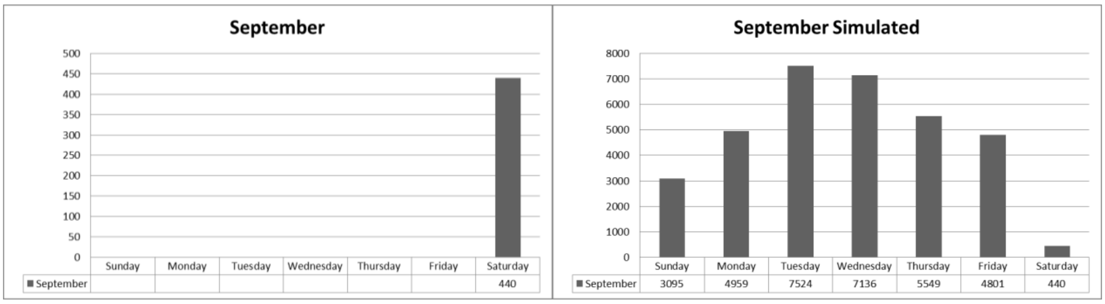
	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
■ Week 1	419	0	0	0	0	0	0
■ Week 2	6	0	1701	1865	4859	3288	38
■ Week 3	5729	635	3875	6749	5915	6908	4569
■ Week 4	319	8283	10338	11531	10103	842	0
■ Week 5	0	7219	466	0	0	0	0

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
■ Week 1	419	4474.691806	5393.010939	6642.592958	7483.886598	3856.517737	2306.412405
■ Week 2	6	4470.188633	1701	1865	4859	3288	38
■ Week 3	5729	635	3875	6749	5915	6908	4569
■ Week 4	319	8283	10338	11531	10103	842	2311.703515
■ Week 5	2880.024568	7219	466	6659.239309	7532.900163	3888.534929	2312.680703

Figure 43 - Data Comparison

It is apparent that outliers in the dataset (values not equal to zero) are still present after the simulations, which reinforce the idea of having time lapses with almost nil contactability, which is behavior that accommodates to a real-life scenario. If the data analysis required the elimination of these outliers, different parameters could be potentially used during the simulation preparation to create an even smoother contactability trend, however, this will naturally include the error factor that comes with it, and which has been previously discussed.

Going back to the original scenario (September), the researcher used all the data gathered, plus the data generated during the simulation in an effort to use the information already gathered to predict the contactability trend in a month with a considerable lack of data. The same simulation approach was used, averaging the contacts per day / month of each other month and generating the simulation for September gaps. The results are discussed below.



	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
■ September							440

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
■ September	3095	4959	7524	7136	5549	4801	440

Figure 44 - Data Transformation

September (the beginning of this project's trial), is a good example of the benefit provided by a simulated, statistically significant dataset predicted using the methods mentioned previously, to minimize data gaps present on a determined timeframe using data contained on the original data gathered during the entire project. A potential approach to determine the accuracy of this simulation and determine the level of error that simulated data adds to real, crowd sourced data analysis- would be to compare it real data gathered during the same time period.

Table 5 - Table of Queries F2F

Queries Performed on the F2F Database	
Query	**Description**
SELECT COUNT(id) FROM tracks WHERE 1	Total number of records on the database
SELECT id, DATE FROM tracks ordered BY DATE ASC LIMIT 1	Get earliest record based on date
SELECT id, DATE FROM tracks ordered BY DATE DESC LIMIT 1	Get latest record based on date
SELECT DATE_FORMAT(DATE, "%W") AS `Day of the week` , COUNT(id) FROM tracks GROUP BY `Day of the week` ORDER BY DATE_FORMAT(DATE, "%w") LIMIT 0 , 30	Get total number of contacts by day overall
SELECT DATE_FORMAT(DATE, "%W") AS `Day of the week` , COUNT(id) FROM tracks WHERE (DATE BETWEEN 'YYYY-MM-DD HH:MM:SS' AND 'YYYY-MM-DD HH:MM:SS') GROUP BY `Day of the week` ORDER BY DATE_FORMAT(DATE, "%w") LIMIT 0 , 30	Get records grouped by day between date 1 and date 2

The main purpose of the above trials was to determine the extent to which one may be able to reasonably sample from a large but limited data set to fill in gaps in the data. The particular data collected here was likely too sparse to reliably fill in missing gaps but one could envision a more complete and larger scale system whereby the gaps as a percentage of the overall sampling period is low enough that the data could be estimated (see, for example [**68**].

6.1.3. Cross Validation

The collected dataset was also considered relative to the known qualities of social networks, as a means to further assess its credibility as a representation of actual social contact patterns.

Overall, the data collected appeared to be representative of the type of contacts one may expect. Although the data collection was sporadic, the fact that the data were collected over a considerable timeframe made the overall trends consistent with expectations when assessed from the perspective of Pareto type of characteristics [**69**] and [**38**], adding confidence that the data collected by the BAT/F2F application displayed characteristics of real social contact patterns.

In general, social contact data are conjectured to fall under the types of data that can be described by empirical laws and/or distributions. The dataset can be queried to generate the probe device and a rank ordering of contact durations. For illustration purposes, Agent 3 was selected for demonstration as it had recorded the largest number of contacts overall (147,000). The most straight-forward distribution is that associated with Zipf's law, which refers to the size of an event relative to its rank order as expressed below:

$$D(r) \sim r^{-z}$$

Here D(r) refers to the cumulative duration of contact with rth entity. Figure 46 illustrates this relationship in a graphical manner.

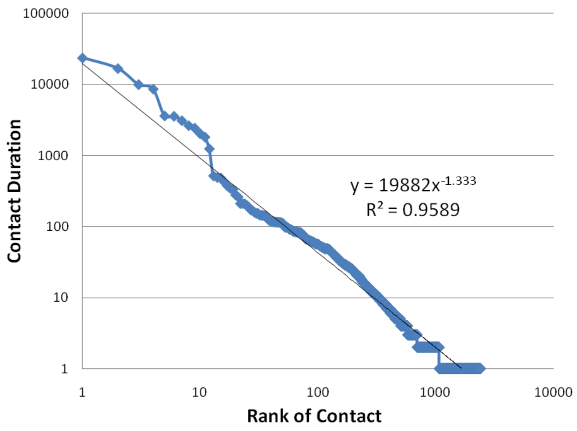


Figure 45 - Proximity contacts duration plotted in rank order (Zipf).

A closely related distribution follows Pareto law. Pareto's law is given in terms of the cumulative distribution function (CDF), i.e. in this case, the number of contacts (N_c) with duration larger than or equal to the duration is an inverse power of the duration as expressed below:

$$P[N_c > D] \sim D^{-p}$$

In general, p should be inversely related to z. This is not precisely the case here, as there is a considerable number of unit durations that tend to skew the rank ordering in a somewhat artificial manner. Figure 47 illustrates the Pareto relationship in a graphical manner.

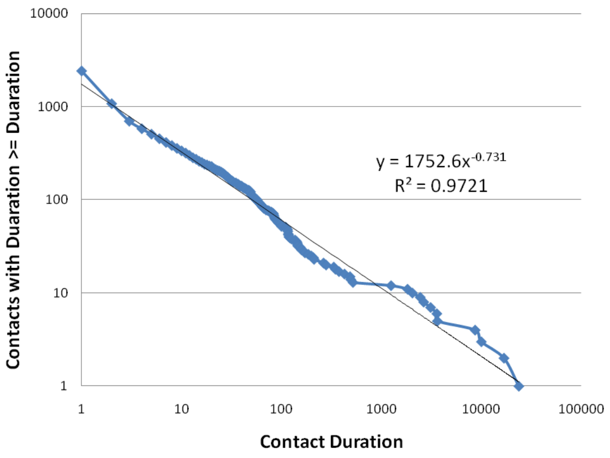


Figure 46 - Contact cumulative distribution function (Pareto).

The associated power law distribution is derived from the probability distribution function (PDF) associated with the CDF given by Pareto's Law. This is essentially a distribution of the number of contact durations of precisely duration D. As such, the power law exponent k = 1 + p.

A power law exponent less than two implies that there is no first moment or mean associated with the distribution. However, as the data obtained from the probe devices is finite, a mean can be calculated.

An interesting (albeit not surprising) parameter that can be extracted from the Pareto principle is the 80-20 rule. From the data collected, the 80-20 rule indicates the number of contacts with which the probe is in contact for 80% of the total contact durations. For Agent 3 this was calculated to be 14 contacts of 2417 or approximately 0.58% of the total contacts. Table 6 illustrates a number of parameters and estimates associated with the four most significant probe devices.

Table 6 - Exponents of the Probe Devices

Agent (All data)	Zipf Exponent	Pareto Exponent	Power Law Exponent (calculated)	PDF Duration “mode” “mean”	80/20 rule
Agent0	2.01	0.41	1.41	168.3	7/399

(student)	R^2=0.95	R^2=0.95			
Agent1 **(faculty)**	1.58 R^2=0.97	0.74 R^2=0.97	1.74	163.4	19/1826
Agent2 **(faculty)**	1.39 R^2=0.97	0.63 R^2=0.98	1.63	145.5	11/1234
Agent3 **(student)**	1.33 R^2=0.96	0.73 R^2=0.98	1.73	141.8	14/2417

The findings summarized in the above figures and table demonstrates a good fit with the 80-20 rule, further adding credibility to the robustness and relevance of the data and the data collection method. The identification of these parameters lends insights into contact and interaction patterns used in models and simulations associated with the spread of contact based infectious disease. As expected, the distributions associated with personal proximity contact display a heavy tail, and exponents can be extracted and used in larger scale modeling. Other observations that can be mined from the data can also be inferences to mobility. A probe with a large number of unit durations is arguably more mobile than one with relatively fewer unit durations.

As a further exploration to establish the relevance and robustness of the data, the current efforts in crowd-sourced data and posterior data analysis were related to similar work using different methodologies, in order to position the current work relative to agent-based modelling (ABM) and simulation.

In a different study[1], an agent based model was developed to study an infectious disease outbreak in a representative small town [**69**]. The town selected was Morden, Manitoba. Figure 47 illustrates the aggregate rank ordering as extracted from the BAT/F2F data.

[1] The agent based modeling and simulation was done by Marek Laskowski.

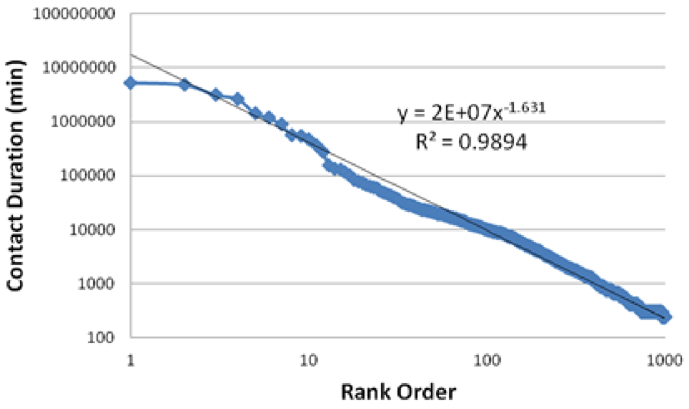


Figure 47 - Rank ordering of all agents (aggregated).

In the agent-based modeling study, the rank order exponent (Zipf's law) of agent contacts is approximately 1.63. This yields an estimated power law exponent of approximately 1.61. The implication is that an agent's contact pattern follows a power law distribution (heavy tail) without finite moments. This result aligns with the current work as well as intuitive perceptions of real face-to-face contact patterns. From these ABM simulations and the aggregated rank orderings, an 80/20 rule can be estimated. In this case, 80% of the contact durations are spent with approximately 4% of a person's contacts (25/670). This again is qualitatively consistent with data extracted from the current work.

Figure 48 illustrates the rank ordering of contacts parameterized by demographics, extracted from the ABM simulation in [**69**]. Intuitively, these profiles appear reasonable. School-age children spend considerable time with three groups, namely household members, school classmates, and friends. The knee in the curve of school-age children is between 20 and 32. For samples of age groups, the exponents associated with Zipf's law are presented in Table 7. It is also intuitive that a 2-year-old and a 70-year-old have similar contact patterns due to their relatively low mobility. The distribution of the adults reflects patterns of conformance consistent with actual survey reports [**70**] and synthesized social contacts [**71**]. In that work, the consequence of the rank ordering implies that the coefficient associated with the corresponding Pareto distribution is between 0 and 1. The lack of a finite mean in the corresponding contact PDF approximation implies that a few long duration contacts are a significant vector of infection spread. In these cases, the (heavy) tail of the distribution may be a significant factor in contact-based disease spread modeling.

The general contact patterns extrapolated from the current work were similar to those seen within the ABM.

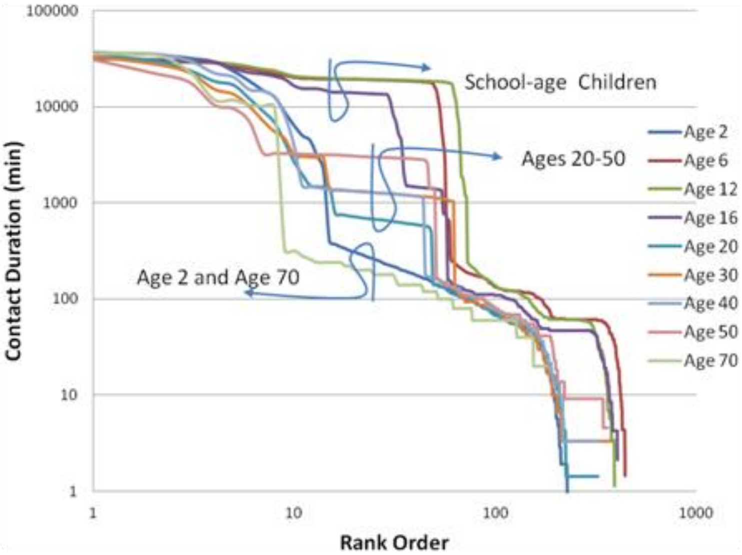


Figure 48 - Rank ordering of agents of various demographics.

Table 7- Zipf Exponents for Various Demographics

Age	Zipf exponent	R^2
2	1.86	0.76
6	1.51	0.80
12	1.85	0.78
16	1.66	0.80
20	2.0	0.94
30	1.87	0.96
40	1.95	0.95
50	1.95	0.97
70	1.50	0.85

6.1.4. Big Data

The importance of data analytics is increasingly highlighted by scholars. It is said that for businesses, an ability to record, store, and analyze large datasets will be an important factor for competitive advantage, supporting productivity, and innovation [**72**]. It is now evident that data can change society as it has a deep importance in areas that go well beyond business, and it can make a big impact in areas such as research, politics, and sports [**73**]. Big Data refers to enormous data sets, which are gathered from multiple locations such as mainframe systems, ERP systems, marketing systems, or in a crowd-sourced way. Data today is being collected from a heterogeneous group of sources. In addition to the traditional means of collecting information, "today millions of networked sensors are being embedded in the physical world in devices such as mobile phones and automobiles, sensing, creating, and communicating data" [**72**]. Organizations of different nature collect terabytes of information about suppliers, customers, operational efficiency from many sources [**72**]. Important characteristics of Big Data are that the entire dataset is considered (N=all) and that the data are often unstructured and requires high levels of analytics in order to gain value from it. Appropriate Big Data management allows for organizations to obtain greater results in having a better understanding of a business and all the factors that affect it, or in general it allows for a better understanding of the events in nature, and provides explanation. In addition, with a better understanding of present events, predicting future outcomes to the same events becomes a feasible task. For example, a better understanding of customers can be generated, better business trends can be generated, or better metrics for business performance can be obtained.

Another characteristic of Big Data is that data are often available and usable for purposes completely unrelated to those for which they were originally collected.

Furthermore, Big Data focusses on trends, patterns, connections, and correlations within and between datasets, without large concern for causative or explanatory factors.

The Bluetooth Agent Tracker system and all the subsequent iterations can represent an important contributor to the collection of large amounts of data that can be used for multiple uses. Within the context of Big Data, the BAT, F2F and FriendFinder applications are highly relevant.

This work represents an important addition to the means for companies, governments, and in general all types of organizations to generate data. It is a new source of data that can be quantified and analyzed, accurate via its crowd sourced origins which map directly to the individual. While the data generated through the applications is not personal, it can generate valuable data about users without crossing any privacy line. For example, the mood, preferences, or any personally identifying information cannot be generated about the user, but meta-data can be made available through BAT or F2F; this is data such as location, time and date of contact, the number of devices that are in near proximity, or traveling time between locations. This information could be correlated to other datasets and used to deduce more information about a user. For example, if a user goes to an ice cream shop every day, it could be inferred that the user is a potential ice cream consumer, and the number of people that are in the ice cream shop at the same time with them. When this concept is applied for a period of time, and applied over several people who provide this information, the amount of information generated would be a lot greater, and the deductions that could be made on the data generated could be considerable. As such, this project enabled the generation of large amounts of data, and for big data analytical concepts to be applied over the data generated.

6.1.5. Enhanced online partner matching services

Throughout history and as societies evolve; humans have progressively developed channels to try to generate romantic/intimate encounters. This includes matchmaking, placing personal ads in newspapers, and more recently online dating. Online dating today is one of the most popular services on the internet, and one of its most significant revenue generators. Online Dating Services generated more revenue than any other online content service in 2004, and it was estimated that the amount of spending for future years continued to grow [**74**]. Partner finding applications are progressively gaining popularity. This is in part due to the change in the pace of life in recent times. With the hectic pace at which professionals today live, the time left for social interactions is minimal. In addition, new generations are raised with a culture of online social relationships where social networking services are part of people's lives from very early stages in their childhood.

However, the current model of online matchmaking services presents many challenges, which include the presentation of untruthful information from participants in the service. One of the issues with the initial physical contact still presents risks for those involved in the relationship where the person they expect might not be the same as the one they encounter. Studies reveal that around 80% of online participants lie about something on their online dating profile [**75**]. This leads to a high level of skepticism, which limits the growth of the industry. The third iteration of the application developed in this project, the FriendFinder application, could potentially represent a feasible solution to address this issue.

The ability of FriendFinder to identify a person's contact that is in short range can be used by online dating services to enhance the system, and identify potential partners of a person when they are nearby and provide the user with an alert. The application would use the user's online dating profile, and use the same engine that online dating services use to match profiles and find potential partners for users. The enhancement would come with the ability of FriendFinder to locate contacts that are in near proximity. As such, a person who is using the online service enhanced by the FriendFinder application could be alerted when a person who is a potential partner according to the online service's matching engine, notified that this person is in near proximity, and prompt the user if a contact was to be made. Both parties could be notified, and a real and personal connection could be made at that moment. This would enhance the dating and matchmaking service as the person with whom the user is to be connected has already been identified as someone who is a good match, and deception can be minimized as the two people will be present then to corroborate the veracity of the online profile.

A practical example illustrates the above scenario. If person X has an online profile on a dating services site such as eHarmony or match.com and this service provides enhanced matchmaking services with the FriendFinder application, then person X can find profile matches in real time when he or she is in a crowded place and special event or even while commuting. For example, person X could be at the shopping mall, and as he/she is running the FriendFinder enhancement on his/her mobile device, FriendFinder is roaming and searching for contacts in the vicinity that could potentially match with person X. For every person with a mobile device on discoverable mode that would also have the application installed, FriendFinder would search to see if their dating profile matches with that of person X's. For every contact found, the online dating engine hub would compare profiles to see if any of the contacts found by FriendFinder is a potential dating match. As soon as a match with certain percentage of likelihood is found (this likelihood percentage would be configure by the user), person X would be notified about the potential match in the vicinity. The person X would be asked to engage on a first digital approach with the potential match, as well as the other person, and they could be prompted for their willingness to make a contact at this point. The two people's exact location at the moment together with the online profile could be then presented to the users, and their initial encounter can be made.

In general, there are some other potential uses and application of the concepts and technology used on FriendFinder, which include uses for business, for science, and for governments. For example, the concept of targeted marketing can move into a whole new dimension with the application of FriendFinder. Using a service similar to FriendFinder, businesses could find “friends” who are nearby and modify their marketing activities according to the needs of potential customers with certain types of profiles nearby. FriendFinder shifts the concept of online contacts, it closes the gap between internet “virtual” friends and real friends, and it also allows for an easier, more specific profiling of user of mobile devices.

CHAPTER 7

Conclusions

7.1. Conclusions

This thesis demonstrated an iterative process to develop an application that used Bluetooth connectivity on mobile devices to develop a new way to generate personal social contact data via device pairing. As mobile device usage has experienced an exponential growth in adoption, research that enhances the uses and benefits of mobile device usage is relevant. Based on the concept of crowdsourcing, the work developed a mobile application named Bluetooth Agent Tracker (BAT) which would use Bluetooth connectivity in order to collect information about other Bluetooth-enabled devices in its proximity. The data collecting device, also called the agent, can collect data including time of the connection, mac address of the device, length of the connection, and device *friendly* name. The collected dataset can reach a significant size and can be derived for multiple uses. Once scaled, "Big

Data" can be generated and analyzed in order to model, simulate, and predict different types of social situations and interactions.

Using the same basic conceptualization and application infrastructure, a second iteration of the BAT was the Face2Face application, where capabilities to accurately locate the connected devices in time and space were added by including geo-location technologies to the application. The principal aim of this iteration of the application was to derive the statistical properties of person-to-person contacts and to demonstrate the usefulness of locating and identifying peers in proximity using the application. The data collected could be logged to a database where it could be mined for contact durations, distributions and associations, all representatives of real-life social interactions.

The third and final iteration of the application involved the addition of system functionality to the application which would enhance social interactions of individuals. This final iteration built on the concept used for identifying nearby devices using Bluetooth connectivity enhanced with geo-location capabilities, and added new ways to find friends and to generate a type of social interaction that is unlike some of the most common forms of social networking. This final application prototype, which was called FriendFinder, bridges the gap between "virtual" and real-life relationships, and created a new way to meet people.

This iterative development moved from device-to-device connections to direct communication and creating new paradigms to generate social relationships. The main reason for this is the current popularity and high levels of adoption that mobile devices. In addition, as devices are so closely equated with people today (usually, one phone = one person and always the same person), data collected can easily be used to simulate real-life situations and to generate a better understanding about these situations. Ultimately, the potential to gather significant amounts of data was demonstrated as relatively few trials of the application on several test devices generated a considerable amount of data in a short period of time. The cross-validation process allowed the researcher to reinforce the credibility of the data collection technology and the validity of the data gathered by aligning it with existing work with similar objectives and using the same statistical analysis that demonstrated the significance of their results. Posterior uses of the data for pandemic simulations or other types of agent based models will be feasible and relatively easy to deploy.

The thesis also illustrates some of the challenges and limitations of collecting crowd sourced proximity data using Smartphones and Bluetooth connectivity. Among these, app distribution and compliance are central challenges facing anyone developing Smartphone apps. App distribution imposes challenges in potentially collecting large amounts of data while compliance introduces data vagaries which directly impact the quality of the data collected.

It is fair to say that at the outset it was uncertain if any useful proximity data could be collected through the type of applications developed here. This work has demonstrated that large amounts of data can be generated and that generating and analyzing data extracted through wireless electronic proximate contact is still quite embryonic but likely will have an interesting and quite likely very useful function and role to play in the future.

7.2. Considerations for Future Research

This thesis has demonstrated how simple connectivity among devices can be taken all the way to possibly enhancing social relationships among humans. The research generated through this thesis, however, does not stop here. Several routes for further research can be taken.

One further direction of research would be to analyze and establish more precisely the extent to which devices actually are able to represent people’s real-life patterns. Potential research questions in this regard would be what is the percentage of time that individuals carry their mobile device with them? -Cross-sectional as well as longitudinal studies investigating individuals' use of their mobile device might help in resolving this question. Such a study would complement the present study by allowing it to determine the extent to which data derived from mobility data of the device can be used to represent actual mobility and connections of a human being.

Privacy issues can also be a derived study arising from this research. The extent to which data collected can be stored and shared either in real time or at a later time can be studied. For example, the extent to which privacy and trust concerns of individuals who use this system would influence any of the interactions within the FriendFinder application could be studied. Such factors would also influence adoption of the application, which would determine the successful deployment and implementation of the system. While privacy concerns may have been included on other studies, the uniqueness of data gathered through the FriendFinder application would allow for a whole new study on privacy concerns on meeting strangers in public places. Such study could also drive some further developments on the FriendFinder application, and define which new features to incorporate, such as the type of personal information that could be shared on first contact.

Related to the above concept, the willingness of people to connect with others who they do not know, but only know that they are nearby, and connected to the same application can also be further studied. There might be other factors that might generate that initial interest for an interaction on both parties such as knowing that they have something in common. Identifying algorithms to establish personality matches of people who are in public places, and who establish a Bluetooth pairing might be an area of research that can establish bases for further iterations of the application.

As Bluetooth has been the technology used upon which direct connection between devices is established, it is also pertinent to study other forms of mobile connectivity that might provide other ways to create direct connection among devices.

Further research can also be based on data that is collected with BAT or F2F applications. Different types of actions and interactions can be modeled with data collected through using those applications. Models can simulate the characteristics of the behavior of social entities in individual settings or as part of a community. For example, agent translation, proximity to other agents, length of interactions, and patterns of localization over time can all be modeled and deduced from data collected through the applications built as part of this thesis.

Bibliography

[1] J White et al., "R&D Challenges and Solutions for Mobile Cyber-Physical Applications and Supporting Internet Services," *Internet Services and Applications*, vol. 1, no. 1, pp. 45-56, May 2010.

[2] B. C.P Demianyk, M Laskowski, and R. D McLeod, "A Distributed Mobile Application for Data Collection with Intelligent Agent Based Data Management Policy," in *SmartData,* Springer, New York, 2013, pp. 139-148.

[3] J. K Laurila et al., "The Mobile Data Challenge: Big Data for Mobile Computing Research," *Mobile Data Challenge by Nokia Workshop, in conjunction with Int. Conf. on Pervasive Computing*, 2012.

[4] N Eagle and A Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, 2006.

[5] E Oliver, "A Survey of Platforms for Mobile Networks Research," *Mobile Computing and Communications Review*, vol. 12, no. 4, 2008.

[6] K. S. Kung, S. Sobolevsky, and C. Ratti. (2013, Nov.) Cornell University Library. [Online]. http://arxiv.org/abs/1311.2911

[7] D. C Brabham, "Crowdsourcing as a Model for Problem Solving," *Convergence: The International Journal of Research into New Media Technologies*, vol. 14, no. 75, 2008.

[8] C. A Martin, "From high maintenance to high productivity," *INDUSTRIAL AND COMMERCIAL TRAINING*, vol. 37, no. 1, pp. 39-44, 2005.

[9] M Sabou, K Bontcheva, and A Schar, "Crowdsourcing Research Opportunities: Lessons from Natural Language Processing," in *12th International Conference on Knowledge Management and Knowledge Technologies (I-KNOW)*, 2012, pp. 17-74.

[10] M Berland and W Rand, "Participatory Simulation as a Tool for Agent-Based Simulation," in *Berland, Matthew, and William Rand. "Participatory simulation as a tool for agent-based simulation." Proceedings of the International Conference on Agents and Artificial Intelligence (ICAART-09)*, Porto, Portugal, 2009.

[11] L Von Ahn, "Human computation," in *Design Automation Conference DAC'09. 46th ACM/IEEE*, 2009.

[12] M Zook, M Graham, T Shelton, and S Gorman, "Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake," *World Medical & Health Policy*, vol. 2, no. 2, 2010.

[13] L Bengsston, X Lu, A Thorson, R Garfield, and J. V Schreeb, "Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti," *PLoS Med*, vol. 8, no. 8, 2011.

[14] D. R Choffnes, F. E Bustamante, and Zihui Ge, "Crowdsourcing Service-Level Network Event Monitoring," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 4, pp. 387-398, 2010.

[15] A Kittur, E. H Chi, and B Suh, "Crowdsourcing User Studies With Mechanical Turk," in *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, 2008.

[16] E Ekik, Y Gu, and D Bozdag, "Mobility-Based Communication in Wireless

Sensor Networks," *IEEE Communications Magazine*, July 2006.

[17] A Wiggins and K Crowston, "From Conservation to Crowdsourcing: A Typology of Citizen Science," in *System Sciences (HICSS) 44th Hawaii International Conference*, 2011, pp. 1-10.

[18] D. M. S. Sultan, "Vehicular Safety Platform Integrating MCC Based Emergency Context-Aware Management Architecture Enhanced With Crowdsourcing Filter," *International Journal of the computer, the Internet, and Management*, vol. 2, no. 1, 2012.

[19] D Olguin and A Pentland, "Social Sensors for Automatic Data Collection," in *Proceedings of the Fourteenth Americas Conference on Information Systems*, Toronto, 2008.

[20] E Oliver, "The Challenges in Large-Scale Smartphone User Studies," in *Proceedings of the 2nd ACM International Workshop on Hot Topics in Planet-scale Measurement*, 2010.

[21] P. Ross, "Top 11 technologies of the decade," *Spectrum*, vol. 48, no. 1, pp. 27-63, 2011.

[22] D Kotz and G Chen, "A survey of context-aware mobile computing research," *Technical Report TR2000-381, Dept. of Computer Science, Dartmouth College*, vol. 1, no. 21, 2000.

[23] A Aleviou, D Antonellis, and C Bouras, "Wi-Fi Technology," in *Encyclopedia of Internet Technologies and Applications*. Hershey, NY: InformatIon ScIence reference, 2008.

[24] M Buehrer, *Code Division Multiple Access (CDMA)*.: Morgan & Claypool, 2006.

[25] B Hofmann-Wellenhof, H Lichtenegger, and J Collins, *Global Positioning System. Theory and practice*. Wien (Austria): Springer, 1993.

[26] B. A Miller and C Bisdikian, *Bluetooth Revealed*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.

[27] H Al-Ofeishat and M Rababah, "Near Field Communication (NFC)," *IJCSNS International Journal of Computer Science and Network Security*, vol. 12, no. 2, pp. 93-99, 2012.

[28] K. Sairam, N. Gunasekaran, and S. Rama Reddy, "Bluetooth in Wireless Communication," *IEEE Communications Magazine*, June 2002.

[29] J Haartsen, "Bluetooth—The universal radio interface for ad hoc, wireless connectivity," *Ericsson Review*, no. 3, 1998.

[30] C Haslett, *Essentials of Radio Wave Propagation*. Cambridge, UK: Cambridge University Press, 2008.

[31] J Wu and I Stojmenovic, "Ad Hoc Networks," *IEEE Computer Society*, pp. 29-31, February 2004.

[32] P Popovsky, H Yomo, and R Prasad, "Strategies for adaptive frequency hopping in the unlicensed bands," *IEEE Wireless Communications*, December 2006.

[33] B Lavelle, D Byrne, C Gurrin, A. F Smeaton, and G. J.F Jones, "Bluetooth Familiarity: Methods of Calculation, Applications and Limitations," in *In MIRW - Workshop at MobileHCI07: 9th International Conference on HCI with Mobile Devices*, 2007.

[34] R. D McLeod, M. R Friesen, K Ferens, B. C.P Demianyk, and J Benavides, "3G Smartphone Technologies for Generating Personal Social Network Contact

Distributions and Graphs," Winnipeg, 2010.

[35] D Ferreira, A Dey, and V Kostakos, "Understanding Human-Smartphone Concerns: A Study of Battery Life," in *Proceedings of the 9th international conference on Pervasive (Pervasive '11)*, 2011, pp. 19-33.

[36] (2013) BlackBerry Developer. [Online]. http://developer.blackberry.com/

[37] (2012) Android Developer Guide. [Online]. http://developer.android.com

[38] J. H Benavides et al., "3G Smartphone Technologies for Generating Personal Social Network contact Distributions and Graphs," in *IEEE HelathCom 2011*, San Jose, CA, USA, June 2011.

[39] R. J. Finton et al. A Best Practice Collaboration Model between Georgia Power and State/District Public Health. [Online]. http://www.ama-assn.org/ama1/pub/upload/mm/415/atlanta_ga.pdf

[40] Vendor/Ethernet/Bluetooth MAC Address Lookup and Search. [Online]. http://www.coffer.com/mac_find/?string=f4-0b-93

[41] S Funk, M Salathe, and V Jansen, "Modelling the influence of human behaviour on the spread of infectious diseases: a review," *Journal of the Royal Society Interface*, vol. 7, no. 50, pp. 1247-1256, 2010.

[42] A.T. Skvortsov, R.B. Connell, P. Dawson, and R. Gailis, "Epidemic Modelling: Validation of Agent-based Simulation by Using Simple Mathematical Models," in *Proceedings of Land Warfare Conference*, 2007, pp. 221-227.

[43] C. A Quiroga and D Bullock, "Travel time studies with global positioning and geographic information systems: an integrated methodology," *Transportation Research Part C: Emerging Technologies*, vol. 6, no. 1-2, pp. 101-127, 1998.

[44] A Byagowi, S Malektaji, and R. D McLeod, "Bluetooth as a Victim Detection Sensor," *IEEE International Symposium on Safety, Security, and Rescue Robotics*, pp. 5-8, Nov. 2012.

[45] G. M Djuknic and R. E Richton, "Geolocation and Assisted-GPS," *IEEE Computer*, vol. 2, pp. 123-125, 2001.

[46] Graziadio School of Business and Management. (2011) Graziadio e-Learning. [Online]. https://wikis.pepperdine.edu/display/GSBME/History+of+Geolocation

[47] B Parkinson, "Introduction and Heritage of NAVSTAR, the Global Positioning system," in *Global Positioning System: Theory and Applications*. Danvers, MA: American Institute of Aeronautics and Astronautics Inc., 1996, pp. 3-28.

[48] F. V Digglen, *A-GPS: Assisted GPS, GNSS, and SBAS*. Norwood, MA: Artech House, 2009.

[49] J Paek, J Kim, and R Govindan, "Energy-Efficient Rate-Adaptive GPS-based Positioning forSmartphones," in *ACM MobiSys '10*, San Francisco, 2010.

[50] P. A Zandbergen, "Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning," *Transactions in GIS*, vol. 13, no. s1, pp. 5-26, 2009.

[51] K Pahlavan and X Li, "Indoor Geolocation Science and Technology," *IEEE Communications Magazine*, February 2002.

[52] M Ndoh and G. Y Delisle, "Geolocation in Underground Mines Using Wireless Sensor Networks," in *Antennas and Propagation Society International Symposium*, 2005.

[53] D Grejner-Brzezinska, C. K Toth, H Sun, X Wang, and C Rizos, "A Robust Solution to High-Accuracy Geolocation: Quadruple Integration of GPS, IMU,

Pseudolite, and Terrestrial Laser Scanning," *IEEE Transactions on Instrumentation & Measurement*, vol. 60, no. 11, pp. 3694 - 3708, 2011.

[54] M Enis, "Mobile Evolution," *Library Journal*, vol. 38, no. 2, pp. 34-36, 2013.

[55] R Beaubrun, B Moulin, and N Jabeur, "An Architecture for Delivering Location-Based Services," *International Journal of Computer Science and Network S 160 ecurity*, vol. 7, no. 7, pp. 160-166, July 2007.

[56] T Brinks. (2009, Feb) Learn how to find GPS Location on Any Smartphone and then make it relevant. [Online]. http://www.codeproject.com/Articles/31965/Learn-How-to-Find-GPS-Location-on-Any-SmartPhone-a#4

[57] T Pering, Y Agarwal, R Gupta, and R Want, "CoolSpots: Reducing the Power Consumption of Wireless Mobile Devices with Multiple Radio Interfaces," in *MobiSys*, Uppsala, Sweden, 2006.

[58] Z Zhuang, K Kim, and J. P Singh, "**Improving** Energy Efficiency of Location Sensing on Smartphones," in *MobiSys '10: Proceedings of the 8th Annual conference on mobile systems, applications, and services*, San Francisco, 2010.

[59] ISACA, "Geolocation: Risk, Issues, and Strategies," 2011.

[60] R Magallanes. (2013, June) Satellite Today. [Online]. http://www.satellitetoday.com/publications/via-satellite-magazine/2013/06/01/geolocation-the-next-privacy-frontier/

[61] R Goodwins. (2003, June) ZDNet. [Online]. http://www.zdnet.com/zoo-keeps-bluetooth-tags-on-wild-children-3002136631/

[62] A Kapadia, T Henderson, J. J Fielding, and D Kotz, "Virtual Walls: Protecting Digital Privacy in Pervasive Environments," *Lecture Notes in Computer Science*, vol. 4480, pp. 162-179, 2007.

[63] L. M. Peterson, "Obtaining vehicle traffic information using mobile Bluetooth detectors," Telecommunications 20120276847, Nov. 01, 2012.

[64] J Satzinger, R Jackson, and S Burd, *Systems Analysis and Design in a Changing World*. Boston, MA: Cengage Learning, 2012.

[65] P Murphy, E Welsh, and J. P Frantz, "Using Bluetooth for Short-Term Ad Hoc Connections Between Moving Vehicles: A Feasibility Study," Houston, TX, 2002. [Online]. http://scholarship.rice.edu/bitstream/handle/1911/20121/Mur2002May5UsingBluet.PDF?sequence=1

[66] N Sadeh et al., "Understanding and Capturing People's Privacy Policies in a Mobile Social Networking Application," *Human-Computer Interaction Institute*, vol. 80, 2007. [Online]. http://repository.cmu.edu/hcii/80

[67] K Williams and J. C Durrance, "Social Networks and Social Capital: Rethinking Theory in Community Informatics," *The journal of Community informatics*, vol. 4, no. 3, 2008. [Online]. http://ci-journal.net/index.php/ciej/article/view/465/430

[68] G Shmueli, *Practical Time Series Forecasting*. United States: Axelrod Schnall Publishers, 2012.

[69] J. H Benavides et al., "Smartphone Technologies for Social Network data Generation and Infectious Disease Modeling," *Journal of Medical and Biological Engineering*, vol. 32, no. 4, pp. 235-244, 2011.

[70] J Mossong et al., "Social contacts and mixing patterns relevant to the spread of infectious diseases," *PLoS Med.*, vol. 5, no. e74, 2008.

[71] S Del Valle, J. M Hyman, H. W Hethcote, and S. G Eubank, "Mixing patterns between age groups in social networks," *Social Networks*, no. 29, pp. 539-554, 2007.

[72] J Manyika et al., "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, 2011.

[73] S Lohr, "Big Data's Impact in the World," *The New York Times*, February 2012.

[74] J Manning, "Construction of Values in Online and Offline Dating Discourses: Comparing Presentational and Articulated Rhetorics of Relationship Seeking," *Journal of Computer-Mediated Communication*, 2013.

[75] J. T Hancock, C Toma, and N Ellison, "The Truth about Lying in Online Dating Profiles," in *CHI - Online Representation of Self*, San Jose, CA, 2007.