Smartphone Technologies for Social Network Data Generation and Infectious Disease Modeling

Julian Benavides Bryan C.P. Demianyk Shamir N. Mukhi Marek Laskowski
Marcia Friesen Robert D. McLeod*

Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Manitoba, R3T 2N2, Canada

Received 13 Jul 2011; Accepted 12 Jan 2012; doi: 10.5405/jmbe.2011.07.4

Abstract

This paper presents a means of collecting and analyzing data related to personal social contact networks. A custom application is developed for smartphones that support Bluetooth connectivity, as representative of the ensemble of many consumer electronic products, to infer users’ location and proximity to one another, the duration of such proximity (‘contact’), and GPS-based information. In many instances of testing the application in this work, this is augmented by device meta-identity. The smartphone application and data storage and retrieval are discussed in detail. Preliminary data were collected (device-device proximity, proximity duration, and location) in pilot testing on the Blackberry Storm and HTC Hero (Android) smartphones. Data are presented as distributions and visualization tools for evolving contact graphs, including Pareto distributions and power law exponents representing face-to-face contacts. Extracted parameters are useful for estimating the potential of infection spread (e.g., respiratory illness), where a key transmission vector is person-person contact. A variant of the standard SEIR individual-based model is developed, with individual contact patterns guided by contact distributions extracted from the smartphone proximity data. Finally, a detailed agent-based model (ABM) of a small community is developed and the spread of an infectious disease is simulated. The data from the ABM is then analyzed in terms of proximity distributions across various demographic profiles, illustrating the utility of the proposed data collection technologies in supporting advancing modeling and simulation efforts associated with infectious diseases.

Keywords: Contact graph, Social network, Infection spread modeling, Smartphone wireless sensor network, Agent-based model, Bluetooth

1. Introduction

Social network analysis is a field devoted to the study of the systems of human interaction, including patterns of interactions (who interacts with whom and for how long), networks that emerge among individuals, and patterns of interaction within and between networks [1]. Data collection has traditionally relied on self-reported data from small numbers of people (“where I was, whom I was with, and for how long”). Known limitations are that this type of data generally provide only snapshots in time and are influenced by known, systematic biases in self-reporting. The emergence of personal mobile communications has opened up new possibilities in collecting behavioral data from larger populations, over continuous periods of time, and with higher accuracy than self-reported data. The follow-on uses of the data—beyond theoretical insights into human behaviour and interaction patterns—include applications in fields as diverse as organizational management, city planning, and public health management.

This work developed a smartphone application called Face2Face (or F2F) for gathering interaction and location data logged by Bluetooth connections between personal mobile devices. The combined data forms a database of individuals’ contact data. Computational techniques were developed to generate and display meaningful social contact graphs from these data. The social contact graphs were input into the Susceptible Infected Recovered (SIR) model [2] (and its variants, such as individual-based Susceptible Exposed Infected Recovered (SEIR)) as well as agent-based models (ABMs) of disease spread to understand contact-based disease spread through a population and the role automated contact data collection may serve in improving modeling and simulation.

* Corresponding author: Robert McLeod
Tel: +1-204-474-7360; Fax: +1-204-261-4639
E-mail: mcleod@ee.umanitoba.ca
The novelty of the work is derived in part from the integration of social networks generated from smartphones with ABMs, specifically for the purpose of infection spread modeling. In a related study, a similar smartphone application collected contact data from smartphones running the application, denoting the contact data as intragroup contacts [1]; F2F extends this to also record proximity contact with devices not running the application. Previous studies have used social network data to simulate information flow during epidemic spread using an Susceptible Infected Susceptible (SIS) model [5], in contrast to the current work that models infection spread using the SIR and SIR variant disease spread model used in this work. Sophisticated ABMs have been developed for infection spread modeling at various scales [6]. However, the application of generated social network data to the simulation of infection spread within an ABM framework is still relatively new [7]. The small-town ABM developed in this study captures contact distributions that are used as a sanity check on the ABM itself, in that the contact distributions are compared with data extracted from F2F. Recently, new relevance and novelty of ABMs is derived, firstly, from the increasing opportunities to use real data (such as those captured by F2F) and secondly, from the ability incorporate a simulation fidelity that can deal with each individual agent explicitly, rather than a simulation fidelity that previously could only deal with compartmentalized simulations.

2. Materials and methods

2.1 Generation of contact data from smartphone sensors

The contact data collection application (F2F) was developed on Android and BlackBerry platforms. In pilot testing, four probe devices ran F2F. In this work, the term ‘probe device’ is borrowed from the concept of ‘probe vehicle’ in intelligent transportation systems, where representative vehicles in the general traffic stream are unique in that they act as mobile sensors, running applications designated for collecting data in real time; this is a means of proxy-based estimating of the characteristics of the entire set of devices. A probe device running F2F records the explicit time, date, and location data (GPS-enabled devices), augmented with connection attempts at intervals (in this case, set to 30 seconds, but can be adjusted by the user) to locate other discoverable devices within Bluetooth range. The connection data includes the probe device ID (physical MAC address of the device and device meta-identity) and the device ID of other devices within Bluetooth connectivity range (MAC address and device meta-identity, if available). The MAC address is set by the manufacturer, and the device meta-identity is an additional identifier that may be set by the user (for example, ‘Jane’s iPhone’). The collected data are then submitted to a web database service where they can be mined for contact durations and associations.

In pilot testing with the four probe devices, over 500,000 contact records were collected over a four-month duration. A larger-scale implementation with 100 probe devices has received approval by the university’s research ethics board, and is planned for implementation in the winter of 2012. The pilot data generated from this work form the basis for ongoing work in computational techniques for data visualization and contact graph generation and visualization, including location-based extensions such as overlaying contact networks on map utilities. The value of contact graphs or contact networks is as an input to disease modeling tools, such as a stochastic SIR model and its variants, and ABMs. This allows the modeling of the qualitative impacts of disease intervention strategies such as vaccination, quarantine, cohorting, and other contact-oriented interventions. This work also advances our understanding of the uncertainty and error associated with statistical data mining, and proposes analytical techniques for reducing error.

2.1.1 F2F application

The proposed application (F2F) runs on Bluetooth-enabled consumer devices such as a BlackBerry, Android, or iPhone. The application runs as an autonomous proxy for the user (usually also the owner) of the Smartphone or other Bluetooth-enabled device, capturing data from other discoverable Bluetooth-enabled devices.

F2F collects information from other devices within its immediate Bluetooth radio service area by polling for other devices at 30-second intervals. The Bluetooth radio service area is typically less than five meter radius, with obstructions impacting the range and signal strength characteristics, as illustrated in the use case of Fig. 1. In the figure, a number of devices are in the range of Agent 1 and various types of information are extracted. It is a statistical representation of proximity contacts, or a statistical representation of one’s environment. Some proportion of the population would not have discoverable devices or may not carry devices, whereas some people may carry more than one device.

2.1.2 Contact graph visualization

Contact graph visualization is useful for data portrayal. Visualizations used here allow one to track a probe device (agent) and estimate contacts in proximity while in transit. Figure 2 represents a probe device in transit on the University of Manitoba campus to a location just off campus. It illustrates a mash-up of proximity contact data, incorporating both spatial and temporal data.
In this case, the contact data are extracted from a database and overlaid on Google maps. Again, contact duration windows and durations may be configured as required. Figure 2 illustrates a mash-up of proximity contact data incorporating spatial as well as temporal data collection, and uses geo-coordinates extracted from GPS data on the smartphone, with most of the GPS data being cellular-assisted.

The route taken by the probe phone (Agent 0) is outlined as a solid line with markers A through G. The concentric rings of icons represent probe queries received from Bluetooth-enabled devices in the proximity of the probe on route during a given time window. The data were collected using an HTC Hero smartphone, proxying for one of the authors. This visualization is included for illustrative purposes in this paper, and is not used within the simulations presented in following sections.

For the study of 100 probe devices, a data set will be generated that will allow more sophisticated and thorough analysis of the social network, such as clustering and centrality. An initial conjecture to be explored in a larger study is that the data would reflect that one lives in a small world network.

2.1.3. Related smartphone technologies for proximity data collection

Within the overall framework, it is also possible to incorporate sources such as WiFi access points. An advantage of scanning for WiFi access points is that although of considerably greater range and thus decreasing the granularity of the data, they can be readily identified as landmarks. This feature has been added to F2F. Data scanned from WiFi devices is inherently different from those collected from Bluetooth, as WiFi connections are inherently person to access point, whereas Bluetooth connections are inherently person to person.

Additionally, one may integrate the probe data with data generated by the service providers. This type of data is also considerably different from those associated with Bluetooth and WiFi scanning. At present, service providers maintain records of cellular usage inclusive of cell tower sector text or call origin as well as duration of a call. These data are primarily used for accounting and billing purposes. Table 1 illustrates a range of wireless technologies that may be used to generate personal contact graphs and mobility patterns.

<table>
<thead>
<tr>
<th>Technology</th>
<th>Range</th>
<th>Type of Data</th>
<th>Type of Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFC (RFID)</td>
<td>~ 20 cm</td>
<td>P2P</td>
<td>Core Personal</td>
</tr>
<tr>
<td>Bluetooth</td>
<td>1-100 m (class-dependent)</td>
<td>P2P</td>
<td>Regular/Core</td>
</tr>
<tr>
<td>WiFi</td>
<td>5-100 m (typical)</td>
<td>Access point</td>
<td>Acquaintance</td>
</tr>
<tr>
<td>Cellular</td>
<td>Km+ (cell sector, typical)</td>
<td>Individual</td>
<td>Community</td>
</tr>
</tbody>
</table>

Near field communication (NFC) is being developed as short-range communication technology for a variety of consumer applications. A mode useful for personal contact generation includes a peer-to-peer (P2P) mode for information exchange between peers. The short range of NFC is suitable for very close personal contact and the corresponding duration of close personal contact. This has not been implemented in the data collection discussed here but is recognized as a potential augmenting technology, although untested for the application of generating proximity data usable for an ABM.

Bluetooth allows for similar P2P data capture, albeit with greater range and device diversity. WiFi is well-suited for a limited objective of locating access points. This is largely a consequence of the WiFi APIs and the fact that 802.11 capture mode is very platform/network adapter/driver/libpcap-dependent and may only capture a limited number of the packets transmitted through a wireless local area network. Nevertheless, WiFi access point information provides coarse-scale indoor localization. Cellular technology is primarily used in conjunction with the probe devices to provide cellular-assisted GPS positioning (as opposed to using cell tower information, which requires engaging the service provider).

A potential limitation of the Bluetooth agent tracker as implemented here is that it increases the power consumption of the device. This can be mitigated to some degree by improvements made to the protocol stacks, such as with Bluetooth low energy (BLE) as outlined within the Bluetooth v4.0 specification [8]. The smartphone wireless technology, range, duration, and time of day may also be used to supply additional information on the type of social contact. These types often include ‘acquaintance’, ‘regular contact or support’, and ‘core personal’, with durations of 1000, 10, and below 10 s, respectively, and can vary in size and interactivity [9].

Although simplified, the various smartphone wireless technologies can be used to automatically estimate the type of contact. For the purpose of using automated contact data in disease spread modeling, the contact type ‘community’ was added to provide information at the community level (smartphones associated with a particular cell tower or sector).

As with most technology-generated social contact methods, there is significant sample bias. Health Canada reports that there were 24 million cell phone users by the end of 2010, representing approximately 72% of Canada’s population. The F2F application was developed for BlackBerry and Android platforms, and will be ported to iOS platform for the iPhone/iPad/iPod. Since the F2F application was developed for BlackBerry and Android users, this creates a bias in the data that will be addressed in time, due to the possibility that the typical profile of BlackBerry and Android users may be statistically different than iPhone users. In addition,

![Figure 2. Contacts in transit.](image-url)
non-smartphone Bluetooth-enabled devices (e.g., laptops) can be used. Generally, there are demographic biases associated with early technology adopters. For smartphones and consumer electronics, these include age group, income, gender, and type and extent of usage, in that early adopters are generally also technophiles and will use technology beyond its core purposes (e.g., voice calls, texting, email). These biases tend to disappear as a technology is adopted. For consumer electronics, this period is often less than three years. The F2F application is meant to serve as an example of a data collection method that allows for the collection of proximate contacts with Bluetooth-enabled devices serving as proxies for their owners.

2.1.4. Proximity contact distributions

Contact data is conjectured to fall under types of data that can be described by empirical laws and/or distributions. In this subsection, several aspects associated with contact data are presented [10]. A web-based database allows for queries and data retrieval from all probe devices. In one instance, the data generates the identity of the probe device together with and a rank ordering of contact durations to this probe device. For illustration, Agent 3 was selected since it had recorded the largest number of total contacts (147,000).

The most straight-forward distribution is that associated with Zipf’s law, which relates the size of an event relative to its rank order as:

\[ D(r) \sim r^{-\alpha} \]  

where \( D(r) \) refers to the cumulative duration of contact with the \( r \)th entity, or the total amount of time spent with each unique contact over one or multiple meetings. Figure 3 illustrates the Zipf relationship.

A closely related distribution follows Pareto’s law. Pareto’s law is given in terms of the cumulative distribution function (CDF). In this case, the number of contacts \( N_c \) plotted on the y-axis with a duration larger than or equal to a given duration on the x-axis duration is an inverse power of the given duration on the x-axis:

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

In general, the exponent \( p \) should be inversely related to \( z \). This is not precisely the case here, as there are a considerable number of unit durations that tend to skew the rank ordering in a somewhat artificial manner. Figure 4 illustrates the Pareto relationship.

![Figure 4. Contact cumulative distribution function (Pareto).](image)

The associated power law distribution is derived from the probability distribution function (PDF) associated with the CDF given by Pareto’s law [10]. This is essentially a distribution of the number of contact durations of duration \( D \). As such, the power law exponent equals \( 1 + p \).

A power law exponent of 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 parameter for a given dataset that can be extracted from the Pareto principle is the 80-20 rule applied to the given dataset. 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 the 2417 total contacts (approximately 0.58%). Table 2 lists a number of parameters and estimates associated with the four probe devices.

<table>
<thead>
<tr>
<th>Agent</th>
<th>Zipf exponent</th>
<th>Pareto exponent</th>
<th>Power law exponent (calculated)</th>
<th>PDF duration mode and mean</th>
<th>80/20 rule (contacts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.01</td>
<td>0.41</td>
<td>1.41</td>
<td>68.3</td>
<td>7/399</td>
</tr>
<tr>
<td>(student)</td>
<td>R^2 = 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.58</td>
<td>0.74</td>
<td>1.74</td>
<td>63.4</td>
<td>19/1826</td>
</tr>
<tr>
<td>(faculty)</td>
<td>R^2 = 0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.39</td>
<td>0.63</td>
<td>1.63</td>
<td>45.5</td>
<td>11/1234</td>
</tr>
<tr>
<td>(faculty)</td>
<td>R = 0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.33</td>
<td>0.73</td>
<td>1.73</td>
<td>41.8</td>
<td>14/2417</td>
</tr>
<tr>
<td>(student)</td>
<td>R^2 = 0.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These parameters give insight into contact and interaction patterns, which can be further applied to models and simulations associated with the spread of contact-based infectious disease. As expected, the distributions associated with personal proximity contact exhibit a heavy tail. The exponents can be extracted and used in large-scale modeling. Mobility can also be inferred from the data. A probe with a large number of unit durations is more mobile than one with few unit durations.

![Figure 3. Proximity contact duration plotted in rank order (Zipf).](image)
2.2 Infectious disease spread models

Although interesting, personal contact data generation methods alone do not directly provide insight into the ramifications of social contacts as a predictive tool for disease outbreak beyond a cautionary indication of the anticipated connectivity between people. The value of contact network data is increased when they are input into modeling tools that can be used in decision support systems. In the present study, the contact graph was input into an individual- or agent-based methodology that models the spread of contact-based diseases such as influenza-like illness (ILI) within a population. The disease spread models were performed with the contact data and/or the PDFs of a person’s contact proximity.

2.2.1 SEIR compartmentalized/Individual model

A simple mathematical model was used to represent the health state of individuals. The model is a stochastic process where an individual’s health is represented by its state, namely Susceptible, Exposed, Infectious, or Recovered (SEIR), as a variant of the more common SIR models of disease spread. The SIR model has been frequently altered or modified, reflecting various additional states and/or policies such as vaccinations [11]. Using contact graphs, one can run simulations of infection spread within a population with a more realistic model of contraction based on the extents and durations of contacts of each individual or group of individuals. In the SEIR model, infected persons are not able to transmit the infection until a certain incubation period had elapsed; the Exposed state represents the time in an incubation period.

The conceptual framework of an individual-based SEIR model is shown in Fig. 5 with typical values explained in Table 3. Figure 5 illustrates the stochastic process representing the states of an individual. The most significant departure from a compartmental- or differential-equation-based analysis comes from the input extracted from the contact graph. The values in Table 3 and the stochastic process model of Fig. 5 can be refined and extended depending upon the characteristics of the infection. As epidemiologists continually improve the characterization of a disease and its infection and transmission vectors, the state diagram representing individuals’ states and durations can also be modified and improved accordingly.

![Figure 5. Stochastic process governing an individual’s health state.](image)

Although the probe data used here are limited, data collected on a larger scale can be input into individual-based SEIR models to provide insight into an infection surge. Modeling the impact of interventions such as quarantine, vaccination, or promoting absenteeism upon first symptoms can be easily implemented within this type of predictive framework. Data that could immediately be mined would include inference of an agent’s health state. That is, once a probe becomes inactive during an outbreak, it may be inferred that the individual has become ill and has had their mobility concomitantly reduced. In this case, an inactive probe infers inactivity in F2F; that is, a probe that is not recording proximity contacts via F2F is assumed to be stationary or sedentary, or to have highly reduced mobility. This is thus inferred that the individual is nominally sick at home. It does not imply overall network inactivity, as the ill individual may continue to use their smartphone for other functions while recovering from their illness at home.

The work is currently focused on scaling up the number of probe devices (from 4 to 100), as well as making the smartphone application more widely available, through BlackBerry App World and Android Marketplace, and extended to the Apple App Store once a compatible version is developed. The credibility and validity of individual-based SEIR disease spread models rely on the nature and degree of real-world data available, and thus modelers should make efforts to acquire as much real world data as technically and practically feasible [12]. Specifically, the stochastic SEIR model can be modified using data from the probe devices, such as the individual PDFs or Pareto distributions for their contacts and durations. As these are probe devices, they are individual devices (proxies) within a large population of users with similar devices. Thus, the probe device PDFs would in effect guide the sampling of the population before updating a person’s health state within the model. This would allow for Monte Carlo sampling and the simulation of a considerably large population without requiring complete contact graph connectivity information.

Here, the Infected state consists of two phases, namely work and home. In general, a person may be infected and become infectious at work prior to being ill and at home (immobile). Each person has essentially two contact lists: one associated with their day-to-day business activities with parameters governed by the 80/20 rule derived from the probe devices, and a home contact list which consists of household members. This extended model allows the contact data to be extracted at certain times of the day. For the simulation, the agents simply resided at residences with an average number four people. The same people would thus spend the night in close proximity, while during the day, each person’s cohort group would be different with sizes governed by the guidance provided by the 80/20 rule as inferred from the smartphone contact data.

| Table 3. SEIR model health state transition diagram parameters |
|-----------------|------------------|
| Parameter       | Value            |
| \( \Delta _{E} \) | Duration of incubation period |
| \( \Delta _{IW} \) | Duration of infectious period (at work) |
| \( \Delta _{IH} \) | Duration of infectious period (at home) |
| \( \Delta _{R} \) | Duration of recovered (immune) period |
| \( \alpha , p \) | Contact graph transmission probability from home contacts |
| \( \beta , p \) | Contact graph transmission probability from work contacts |

Parameter and Value
2.2.2 Agent-based SEIR model

The ABM resides within an ABM simulation framework geared towards high-fidelity modeling of human institutions of varying scales. The broad design goals of this framework, called Simstitution, are based on the collective experience of the authors gained while developing ABMs of human institutions. Originally, models of hospital emergency departments [13] and cities [14] were implemented in simulators strongly coupled to the specific modeling application [15]. Publishing results from a series of models built upon a common simulator framework, combined with verification of model components (or sub-models), is commonly used for building confidence in simulator frameworks for epidemiological modeling [16,17].

For rapid model construction, an ABM framework should facilitate the incorporation of real-time data feeds to improve data-driven simulation. A tool for visualization and graphically interacting with the model facilitates model development, validation, and debugging. Visualization is a key component in communicating results with experts and stakeholders [18]. A visualization tool can also be extended to allow model construction or editing model parameters imported from real data.

2.2.2.1 Simstitution design details

Simulated entities within Simstitution fall into one of two major categories; Agents (SimAgent), which are the autonomous entities that make decisions and interact with the environment, and instances of the SimRegion class, which represent spatially partitioned subdivisions of the environment. These relationships are summarized in Fig. 6.

![Class diagram for core Simstitution class hierarchy.](image)

SimRegion is a unit of spatial decomposition as well as a convenient unit of computation. In the latter role, it can be considered as a container for agents that need to have their next state computed. Figure 7 illustrates the details of this relationship. A particular instance of SimRegion can be the parent container of SimAgents or SimRegions but not both types at the same time. This restriction will in practice result in tree hierarchies of SimRegions, with SimAgents contained in the leaf SimRegions, and the top region at the root of the tree. The SimRegion spatial decomposition granularity becomes increasingly fine away from the root and towards the leaf regions of the tree.

Time advances in the simulation when the simulator advances the time of the top region (root of the tree) by some discrete time step. The top region will then advance the time of its children by the same time step in a recursive fashion such that the tree is traversed in a depth-first manner until all the SimAgents in the leaf regions have been simulated for that time step. The simulator repeats this process until a certain number of time steps have elapsed.

IndividualPolicy is a modular unit that affects the behavior of the subscribed SimAgent, which may also require the IndividualPolicy to store encapsulated SimAgent state data specific to itself. Examples are a schedule policy which causes the SimAgent to observe a particular day/night work/home schedule, or in the case of a hospital being modeled, a doctor policy which causes the SimAgent to treat patients within a hospital. Within a SimRegion, each possible derived IndividualPolicy class has a corresponding GroupPolicy for that SimRegion. The GroupPolicy acts as a factory for the corresponding IndividualPolicy and, if required, facilitates coordination between one or more derived IndividualPolicy classes (e.g., healthcare worker policy in a hospital that coordinates interaction between the nurse and doctor IndividualPolicies). It is assumed that the properties of the local environment constrain the behavior of agents (e.g., airport security lineup, swimming pool, hospital, bank). The associations between SimRegion, SimAgent, GroupPolicy, and IndividualPolicy are shown in Fig. 8.

![Relationships among modular agent policies.](image)

Communication or interaction between SimAgents is exclusively through messages passed between them. Messages received by a SimAgent are relayed to its IndividualPolicies which can lead to an internal change of state, or an action to be taken which could lead to additional messages being sent to other IndividualPolicies on the same subscribed SimAgent, or messages sent to other SimAgents. Message passing fits well with the agent paradigm, since the relationship between external events and internal agent state remains loosely coupled, in accordance with the principle of agent autonomy [19].

![Relationships between core class instances that form a tree.](image)
2.2.2.2 Details of small town model

An ABM of the town of Morden, Manitoba (population 6600), was modeled. The ABM incorporates the framework features mentioned in the previous section, and includes visualization capabilities for observing emergent model behavior during execution. The model is fairly basic so the SimRegion tree consists of only two layers, namely the root or top SimRegion (Morden) and the leaf SimRegions which represent the home, school, and work locations that agents occupy. The leaf SimRegions are arranged in a grid with empty spaces between structures to allow for SimAgent travel. Agents are assigned work, school, and home locations based on demographic data [20]. Figure 9 shows a screenshot of the Morden simulation at a particular time step. The entire town is shown on the left. A detailed view of six classrooms in the center of town, in which individual SimAgent details can be seen, is shown on the right. Details include the gender and age of the SimAgent, as well as disease status. Disease status is indicated by the color of the SimAgent icon. When run interactively, the icon changes color, with green indicating a susceptible state. Once the agent is infected, the icon becomes yellow, orange, or red depending on how long it has been in the infected state. Finally, recovered SimAgents are blue. The leaf SimRegions are depicted as colored squares, whose color represents the aggregated disease state of the SimAgents within that region. SimRegions with no SimAgents are white.

Four IndividualPolicy subclasses were used to generate the SimAgent behavior in the Morden model. The SchedulePolicy determines whether a particular agent wants to be at its assigned work, school, or home, depending on its demographic profile, and the current time which advances in increments of one hour. The SchedulePolicy sends messages containing the desired destination to the SimAgent’s MovementPolicy which handles the actual movement. The InfluenzaPolicy maintains the particular SimAgent’s disease state, and if in the Infected state, sends “infection” messages to other SimAgents in the same SimRegion, which is how disease spreads between SimAgents. Finally, the BluetoothTrackingPolicy emulates the Bluetooth smartphone contact application, and is the source of the synthetic contact data. The corresponding GroupPolicies are used to facilitate the aggregation of data in a spatially explicit manner to achieve the tiling effect shown in Fig. 9.

![Figure 9. Screenshot of a running simulation. Morden (left) and close-up of 6 classrooms (right).](image)

3. Results and discussion

3.1 Simulation results from SEIR compartmental model governed by contact data

Using data collected from the four probe devices to govern contact patterns between agents, Fig. 10 illustrates the results of infection spread through a population of 5000 persons. The simulation was exploratory and a set of basic parameters was used. A population of 5000 was of interest as it represents small communities, including remote and northern Canadian communities which were particularly hard hit in the 2009 influenza pandemic [21]. Each person had a contact list of approximately 10 close contacts reflecting the 80/20 rule derived from the Pareto distribution. The simulation, although coarse, included a circadian rhythm where each individual was also provided with a contact list of two persons during the night (every other 12-hour cycle) in addition to their daytime contacts. The probability of becoming infected was on average a 0.0025 probability of becoming infected if one of the close contacts was infected per hour of contact with that infected contact. This type of infection probability is an adjustable parameter; its current value is associated with this particular simulation but is also consistent with considerably larger models. As the model develops, the infection rate is derived from the probability distribution. In the simulation of Fig. 10, the probability of infection was calibrated to a reasonable secondary infection reproductive rate ($R_0$) of 1.9.

![Figure 10. SEIR individual model.](image)

The curves of Fig. 10 are typical of compartmental SEIR models. The only difference here is that these simulations are the result of individual stochastic models with contact lists governed by the observation of the 80/20 rule derived from the Pareto distribution of inferred contacts from an automated proximity contact pattern generator. In epidemiology, $R_0$ is the basic reproduction number of the infection and is the number of secondary infections that a single infected case will cause. In the case of an influenza strain (e.g., 1918), $R_0$ has been estimated to be between 2-3.

To explore conditions that may be representative of remote northern Canadian communities, the number of close-
proximity contacts during the at home cycle was varied from 2 to 5 (Fig. 11), reflecting census data for typical household sizes for both urban centers, smaller communities, and remote communities of Canadian Aboriginal peoples [20]. This represents a tendency towards overcrowding in homes. Qualitatively, the simulations indicate that a major contributing factor in the spread of an ILI is overcrowding, which exacerbates the infection spread as a consequence of increased exposure due to increased contact [22]. Another factor of potential public health concern is thus associated with an intervention that recommends for an infected person to stay home. In situations or environments with severe overcrowding in homes, this recommendation may have a detrimental impact. A similar result was found from simulations conducted by Episim [16]. In these scenarios, it may be worthwhile to set up temporary mobile facilities to house and treat persons infected as opposed to recommending that they stay home.

3.2 Simulation results from ABM-Generated contact data

This section discusses how the ABM can be improved and validated to some degree through the inclusion of as many data sources as is practically possible. The first and most obvious enhancement would be to use demographic data that is as accurate as possible. The ABM developed here is based on data obtained through the federal census by Statistics Canada. In addition, models of schools have been refined to provide for reasonable class sizes. The school data have been estimated, but would benefit from using real data for this parameter. With this model, a disease spread simulation was run to provide a baseline for modeling the spread of a respiratory infection or ILI. Figure 12 illustrates the spread of a disease among the community of Morden, Manitoba, in isolation.

In the first effort to improve the basic ABM, the agent contacts and durations were made to reflect the patterns in data extracted from the Bluetooth probe devices. The objective was to determine how well the model reflected real person-person networks. For the baseline simulations of the single-town ABM, typical contact patterns for all agents were used. The results of this analysis are summarized below.

Figure 11. Impact of overcrowding.

Figure 12. SIR disease spread simulation.

Figure 13 illustrates the rank ordering aggregated over all agents. The rank order exponent (Zipf’s law) 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 is expected from both the Bluetooth proximity pilot as well as intuitive perceptions of real face-to-face contact patterns. This instrumentation of the ABM which feeds back data while the simulation is running helps validate it as approximating real-world 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 consistent with data extracted from the Bluetooth data collection pilot. Although it appears that the curve of Fig. 13 can be divided into three distinct regions, each governed by a different exponent, this is apparent in many power law distributions; the best-fit exponent is usually selected.

Figure 13. Rank ordering of all agents (aggregated).

Figure 14 illustrates the rank ordering of contacts parameterized by demographics. 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 4. 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 [23] and synthesized social contacts [24].

Figure 14. Rank order of contacts parameterized by demographics.
The data from an ABM simulation can be validated using other types of published data. For example, in [23], contact patterns derived from a large population survey were analyzed. The results indicated that for their preliminary modeling of 5- to 19-year-olds are expected to suffer the highest incidence during the initial epidemic phase of an emerging infection transmitted through social contacts when the population is completely susceptible. These expectations are consistent with the contact patterns generated by the ABM developed here and consistent with the distributions obtained from the smartphone proximity data collection application.

4. Conclusion

This paper presented an autonomous data collection application for smartphones. The application records device identities, time stamps, and physical locations. The application polls for data at regular intervals (30 seconds) to proximate discoverable Bluetooth devices. Bluetooth was selected due to its use in many smartphones and other devices. Social contact graphs were generated and displayed using several computational techniques. Methods of visualizing this type of data require further development. The networks are inherently stochastic and non-planar, making the problem difficult. However, it should be noted that visualization is only one means of using the data. The data collected were also analyzed through estimates of distribution-governing exponents and parameters.

The utility of the contact data was illustrated within an individual-based SEIR model to provide insight into how disease spread may be influenced through personal contact. The individual-based predictive disease spread model is a stochastic process model with transitions influenced by the degree of contact people have with one another. The degree of contact (extent and duration for each person) is generated to be individual-specific and can be collected via the developed technologies and application.

A more sophisticated ABM was run to simulate a small town. In addition to the familiar SEIR model outputs, the contact distributions of various individuals were used. They were found to be in good agreement with the data collected from the Bluetooth proximity contact data.

The developed smartphone application and models also shed light on the uncertainty and error that can be expected in mining this type of data. Errors introduced in the data collection methods are largely a consequence of the inherent uncertainty of the radio signaling and the configuration of Bluetooth devices. In each case, the data is at best statistical and should be evaluated in that context.

Acknowledgements

The authors thank Tony Florio at Research in Motion for the donation of 5 BlackBerry Storm devices and Mathew Crowley of MTS Allstream for the donation of an HTC Hero device. Financial support from Manitoba Hydro is gratefully acknowledged.

References


