3G Smartphone Technologies for Generating Personal Social Network Contact Distributions and Graphs

J. Benavides, B. Demianyk, R.D. McLeod,
M.R. Friesen and K. Ferens
Electrical & Computer Engineering
University of Manitoba, Winnipeg, Manitoba, Canada
mcleod@ee.umanitoba.ca

S. N. Mukhi
Canadian Network for Public Health Intelligence
1015 Arlington Street,
Winnipeg, Manitoba, Canada

Abstract— **This paper presents a novel means of collecting and analyzing personal social network data, using 3G Smartphone that support Bluetooth connectivity. The application discovers and logs other close-proximity Bluetooth-discoverable devices, and this can be used to infer individuals' proximity, contact duration, as well as geographical location information. The contact data are presented in terms of distributions and visualization tools for evolving contact graphs. Finally, contact data are then demonstrated to be of utility in estimating the potential of the spread of a contact-based infectious disease, where a vector of transmission is proximity to infected agents.**

Keywords- contact graph; social network; modeling infection spread; wireless sensor network; radio frequency identification

I. INTRODUCTION

Social network analysis is the field devoted to the study of the systems of human interaction, including patterns of individual interactions (who interacts with whom and for how long), networks that emerge among individuals, and patterns of interaction within and between networks. The emergence of personal mobile communications has opened up new possibilities in collecting interaction data from larger populations, over continuous periods of time, and with higher accuracy than self-reported data.

The objective of this study is to automate the acquisition of individuals' contact (interaction) data using personal mobile devices (i.e. Smartphones). The follow-on objectives are to develop computational techniques that generate and visually display meaningful social contact graphs from the data, and to investigate simulated disease spread models (SIR and variants) on the data, in the interest of understanding disease spread through a population.

Personal contact is the primary means of transmission of influenza-like illness (ILI) and many other respiratory infections between people. There have been substantial efforts to model this type of infection spread at the scale of an entire population, using mathematical models in an attempt to understand the spread dynamics and to evaluate various infection control measures [1][2].

This work introduces 3G Smartphone application technologies to generate contact data within a given workplace or organization in an automated fashion. The objective of the application is to obtain statistical properties of person-person contacts and to subsequently demonstrate the data utility in the application of infection spread models. The infection spread models are stochastic phase-type models, and the primary means of analysis is that of an individual based SEIR (Susceptible, Exposed, Infectious, Recovered) model, as a derivative of an SIR model [3][4].

II. CONTACT DATA GENERATION/ANALYSIS

A. Contact Data Generation

There have been a number of research efforts oriented at estimating personal social contacts [5]. In this work, we developed an application that could be used by a relatively small number of participants as probes, allowing them to "vampire" proximity data from a variety of consumer electronic devices inclusive of other 3G Smartphones; this is a distinct difference from the work in [6]. These ideas are well entrenched in the wired world and used by system administrators as a means of monitoring their networks. The 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 3G Smartphone application was run on five probe devices that maintain explicit location data when available (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, (4) 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.

Figure 1 presents a "use case" more clearly illustrating the role of the probe agents among other Bluetooth devices. In this scenario, Agent 1 is the probe device interrogating a number of other mobile devices within Bluetooth transceiver range.

978-1-61284-696-5/11/$26.00 ©2011 IEEE

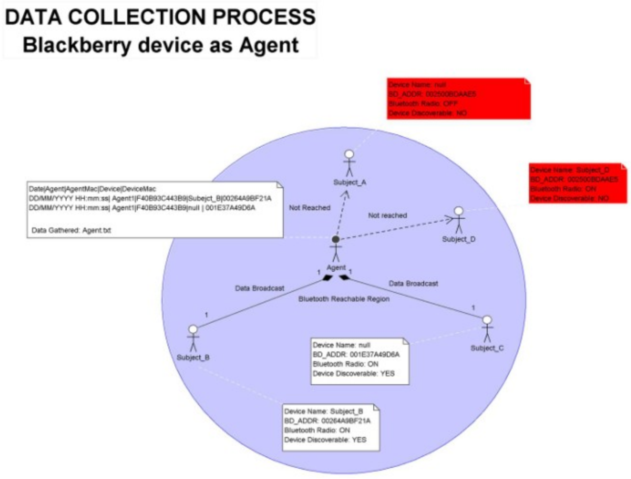

Figure 1: A "use case" scenario for data collection from probe agents.

The use case also illustrates issues with ranging. At the time of writing, the API for the Bluetooth radios is not as complete as that for WiFi for example, with the latter having more monitoring and control over parameters such as received signal strength indication (RSSI). As such, actual distances of subjects are difficult to quantify.

Initially we did not have a good estimate of 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 that were in question. As mentioned, the devices used were BlackBerry Storm 3G Smartphones and the HTC Hero (Android). These devices 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 [7].

In the case of the BlackBerry Storm devices used here, a phone can only be detected if the Bluetooth option is set to discoverable. With just over three months of data collection and with just five probe devices, approximately 500,000 contact / connection records were collected.

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 desk top 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.

B. *Location Based Extensions*

Incorporating actual location based services considerably enhances aspects of the probe application. The Smartphones we are using support GPS as well as assisted GPS services. Within this context, wherever and whenever possible, location based data is appended to the proximity records. An immediate benefit to a user then is to be able to obtain a graphical record of one's proximity contacts as they traverse a campus or city. This is illustrated as a XML mash-up overlaid on GoogleMaps. The data for Figure 3 was collected using the HTC Hero probe device that is represented by one of the authors. Figure 2 illustrates a mash up of proximity contact data incorporating spatial as well as temporal data collection.

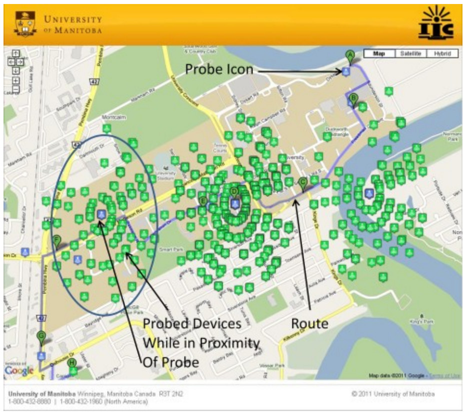

Figure 2: Proximity contacts collected from Smartphone probes incorporating location based GPS information.

C. *Contact distributions*

Contact data is conjectured to fall under types of data that can be described by empirical laws and/or distributions. In this sub-section, several aspects associated with contact data are presented. The web based database allows for queries and data retrieval for all probe devices. In one instance, the data generates 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 (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 r^{th} entity. Figure 3 illustrates this relationship in a graphical manner.

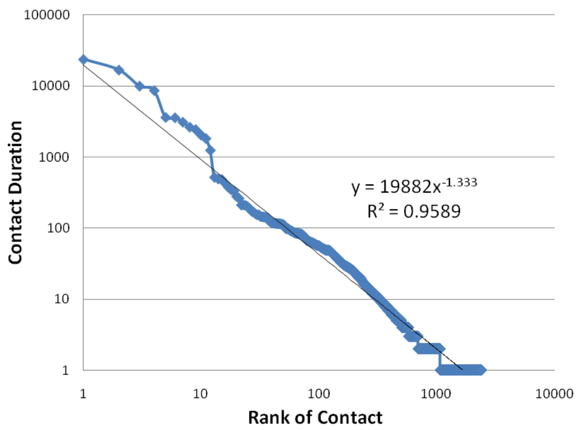

Figure 3: Proximity contacts duration plotted in rank order (Zipf).

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 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 in a graphical manner.

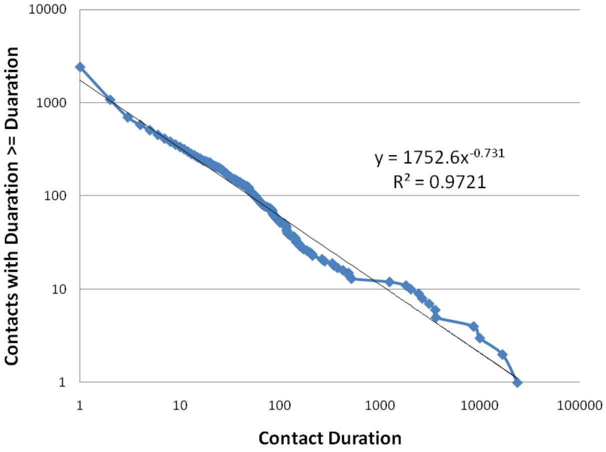

Figure 4: Contact cumulative distribution function (Pareto).

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 upon 2417 or approximately 0.58% of the total contacts. Table I illustrates a number of parameters and estimates associated with the four most significant probe devices.

TABLE I:
Exponents of the Probe Devices

Agent (All data)	**Zipf Exponent**	**Pareto Exponent**	**Power Law Exponent (calculated)**	**PDF Duration "mode" "mean"**	**80/20 rule**
Agent0 (student)	2.01 R^2=0.95	0.41 R^2=0.95	1.41	1 68.3	7/399
Agent1 (faculty)	1.58 R^2=0.97	0.74 R^2=0.97	1.74	1 63.4	19/1826
Agent2 (faculty)	1.39 R^2=0.97	0.63 R^2=0.98	1.63	1 45.5	11/1234
Agent3 (student)	1.33 R^2=0.96	0.73 R^2=0.98	1.73	1 41.8	14/2417

The knowledge 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 notions 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.

III. Contact Data Application

While interesting, personal contact data generation methods alone do not directly provide insight into the ramifications of social contacts as a predictive tool in the event of disease outbreak, beyond a cautionary indication of the anticipated connectivity between people. The value of contact network data is extended when they are used as input to additional modeling tools that can be used in decision support systems. In our case, we adapted the contact data as an input to an individual- or agent-based methodology that models the spread of influenza-like illness, or ILI within a population. In this case, a simple mathematical model represents the health state of individuals. The model is a stochastic process where the individual's health is represented by their state: Susceptible, Exposed, Infectious, and Recovered (SEIR), as a variant of more common SIR models of disease spread. In the SEIR model, infected persons are not able to transmit the infection until a certain incubation period had elapsed. As such the Exposed state represents the time in an incubation period. The basic SEIR model is shown in Figure 5 with the values explained in Table II. The values are illustrative, and can be tuned to specific types of infections.

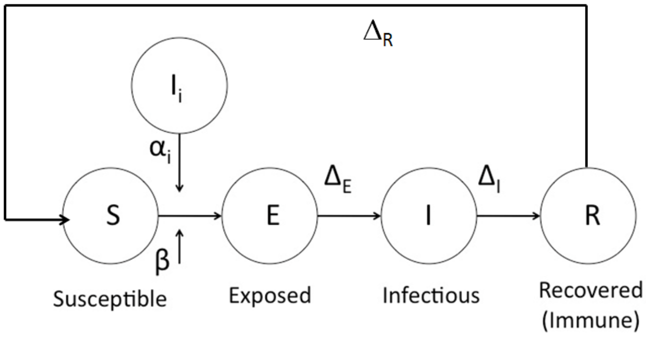

Figure 5: The stochastic process governing an individual's health state.

TABLE II
SEIR model health state transition diagram parameters

Parameter	Value
Δ_E	Duration of Incubation Period (e.g. 24 hours)
Δ_I	Duration of Infectious Period (e.g. 7 days)
Δ_R	Duration of Recovered (immune) Period (e.g. 200 days)
α_i	Contact graph edge weight to an another ContactNode
β	Contact graph transmission probability

The stochastic process model of Figure 5 can be refined and extended, depending upon the characteristics of the infection. As epidemiologists better characterize a disease and its infection and transmission vectors, the state diagram representing individuals' different states and durations can be modified accordingly.

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 data.

Although the probe data at present is limited, data collected with a larger set of devices and used as input to individual SEIR models could shed light on an infection surge as would be anticipated from an underlying contact-transmitted infectious disease. 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 actually would include inferencing of state. That is, once a probe becomes inactive during an outbreak it may be inferred that the individual has become ill and their mobility concomitantly reduced.

IV. Summary

This paper presents a novel means of collecting person-to-person contact data via 3G Smartphones running simple network application services. Data collected was analyzed visually, as well as through estimates of distribution governing exponents and parameters. The utility of the contact data was illustrated within an individual-based model to provide insight into how disease spread may be influenced through personal contact within a specific organization. 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 organization-specific and collected via the technologies outlined.

The models also shed considerable light on the uncertainty and error that can be expected in mining this type of data. Errors are 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.

Methods of visualizing this type of data are primitive and are an underdeveloped area of research in terms of effective visualization. The networks are inherently stochastic and non planar, making the problem even more difficult. However, it should be noted that visualization is only one means of using the data, and it is far from obvious how to show best to use this type of data when it comes to extracting utility. The SEIR model is only one means of further analyzing contact data.

Acknowledgment

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

References

[1] Grassly, N.C., and Fraser, C., "Mathematical models of infectious disease transmission". Nature Review Microbiology, 6, 477–487, 2008.

[2] Mossong J., Hens N., Jit M., Beutels P., Auranen K., Mikolajczyk R., Massari M., Salmas S., Tomba G.S., Wallinga J., Heijne J, Sadkowska-Todys M., Rosinska M., and Edmunds W.J., "Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases," PLoS Med 5(3): e74. doi:10.1371/journal.pmed.0050074, 2008.

[3] Weisstein, E.W. (n.d.). Kermack-McKendrick Model, in MathWorld--A Wolfram Web Resource. Retrieved 16 September 2009 from http://mathworld.wolfram.com/Kermack-McKendrickModel.html

[4] Wikipedia, Compartmental models in epidemiology, http://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology

[5] Srinivasan, V., Motani, M., and Ooi, W. T., "Analysis and implications of student contact patterns derived from campus schedules," Proceedings of the 12th Annual International Conference on Mobile Computing and Networking (Los Angeles, CA, USA, 23 - 29, 2006). MobiCom '06. ACM http://doi.acm.org/10.1145/1161089.1161100 pp. 23-29, September , 2006.

[6] Eagle N., Pentland A., & Lazer D., "Inferring Social Network Structure using Mobile Phone Data," Proceedings of the National Academy of Sciences, 106(36), pp. 15274-15278, 2009

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