3G TECHNOLOGIES FOR SOCIAL NETWORK DATA GENERATION

Julian Benavides, Bryan C.P. Demianyk, Shamir N. Mukhi[1], Ken Ferens,
Marcia R. Friesen and Robert D. McLeod
Electrical & Computer Engineering, University of Manitoba
[1]Canadian Network for Public Health Intelligence

INTRODUCTION

Social network analysis is the 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 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 disease spread as a public health concern. This work developed a 3G/4G Smartphone application to gather interaction and location data logged by Bluetooth connectivity between individuals' personal mobile devices (e.g.Smartphones). The combined data formed a database of contact data, from which computational techniques were developed to generate and display meaningful social contact graphs, and further, on which disease spread models (SIR and variants) were simulated in the interest of understanding disease spread through a population.

METHODS

The application was developed on Android and BlackBerry platforms. In pilot testing, several probes ran the application which maintains explicit time, date, and location data (device GPS-enabled), augmented with connection attempts to proximate devices that are discoverable via Bluetooth at regular intervals (e.g. 30 sec). The connection data includes the probe device ID (physical MAC address of the device, plus the device meta-identity) and the device ID of other devices within Bluetooth connectivity range (MAC address and meta-identity, if available). The collected data is then logged to a web database service where it can be mined for contact durations and associations. The application was pilot tested with four probe devices, in total collecting over 500,000 contact records over a four month duration. Implementation of a larger-scale implementation with 100 probe devices is currently underway. The pilot data form the basis for ongoing work in computational techniques for contact graph generation and visualization, including location-based extensions such as overlaying contact networks on map utilities. Extensions to other wireless contact data sources, such as WiFi and RFID scanning are also in development, as well as augmenting the data with location-based data from cellular service providers. The value of contact graphs (contact network graphs) is derived when used as input to disease modeling tools, such as a stochastic SIR model and variants. This allows for modeling of qualitative impacts of disease intervention strategies such as vaccination, quarantine, cohorting, and other contact-based interventions. The work also advances our understanding of the uncertainty and error associated with statistical data mining, and analytical techniques to reduce error.

THE APPLICATION

The application is denoted Face2Face (F2F), using Bluetooth enabled consumer devices (primarily Smartphones) to proxy for the user).

Figure 1: The F2F app

The F2F application polls for other devices within its proximate Bluetooth radio service area. This is typically less than 5 meters with obstructions and similar impacting range and signal strength characteristics as illustrated in the use case of Figure 2.

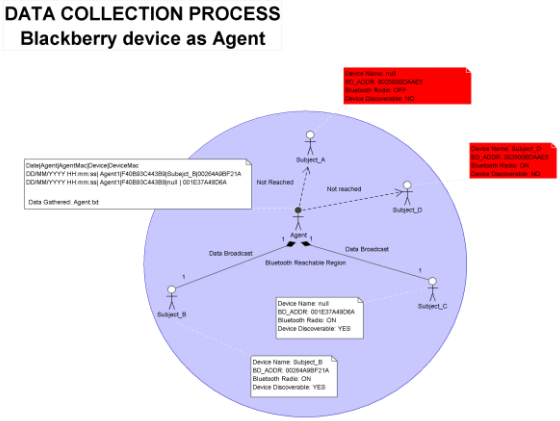

Figure 2: The F2F Application Modus Operandi

CONTACT GRAPH VISUALIZATION

Contact graph visualization is useful as a mean of data portrayal, and is a computational challenge in and of itself. Figure 3 illustrates the contacts acquired over a time window of two months with the short duration connections excluded for visualization purposes.

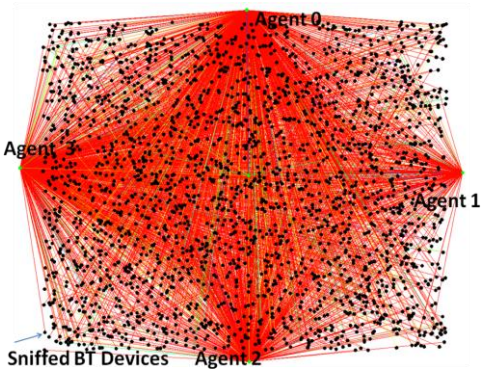

Figure 3: Contact Graph Visualization

Other forms of visualization allow one to track a probe device (user) and estimate contacts in proximity while en route. Figure 4 represents one probe device in transit on the University of Manitoba campus.

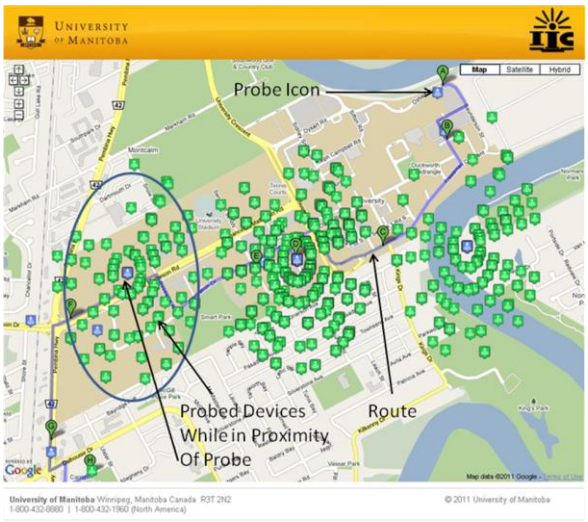

Figure 4: Contacts in Transit

The contacts are extracted from a database and overlaid on Googlemaps. Again, contact duration windows and durations may configured as required.

ANALYSIS

Personal contact graphs are believed to display characteristics associated with heavy tail distributions, and this discussion uses the familiar concepts of small-world networks and 80/20 rules. The data collected from the probe phones, using Bluetooth proxies for personal contacts, display similar characteristics. The clearest representation of this feature is extracted from plots and analysis of the

cumulative probability distribution of a given probe phone as shown in Figure 5.

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:

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

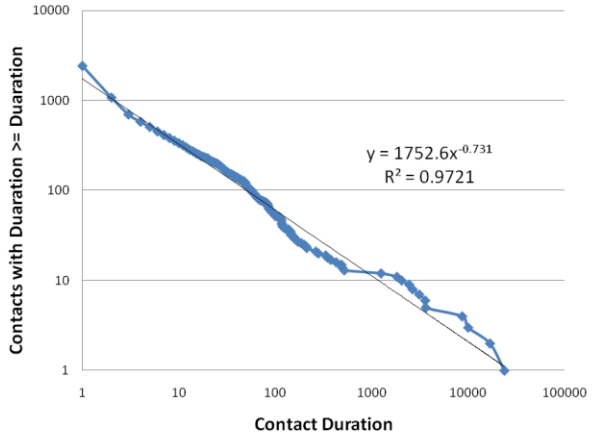

Figure 5: Contact cumulative distribution function (Pareto)

From this type of data, the exponent associated with the power law distribution can be calculated. In the case of Figure 3, the exponent is calculated to be 1.731 indicating a heavy tail. Not unexpectedly, the majority of a probe's contact duration is spent with a small number of devices probed. In the case of the data collected for the probe of Figure 3, 80% of the time was spent with just 14 devices, out of a total number of 2400 contacts made.

These patterns and characteristics of the contact data can be used in models of disease spread, particularly for contact- or proximity-based infections such as influenza or other respiratory illness. The heavy tail and exponents can be extracted and used in larger scale modeling.

Using notions of the 80/20 rule extracted from the data, an SEIR disease spread model (SIR variant) was built and run [2]. The infection was an influenza-like illness (ILI). In the case of Manitoba, isolated northern communities were particularly hard hit during the first wave of the 2009-2010 influenza outbreak. Although we are not attempting to replicate a particular community, the population we considered was a model on the order of 5000 people in relative geographic isolation. This provides a closed system for modeling purposes.

The model used as a base was a simple SEIR agent-based or discrete model. It is a phase type model where a person can be in any one of several health states. These states are typically denoted Susceptible, Exposed, Infected, Recovered. This is a minimal type phase space and is illustrated in Figure 6.

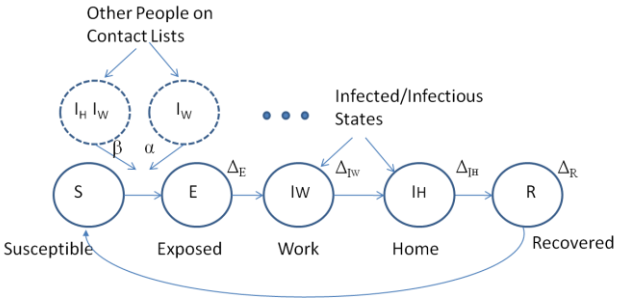

Figure 6: SEIR Compartmental Model.

In this work, the Infected state consists of two phases (work and home). In general, a person may be infected and infectious at work prior to a period where they may 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" and derived from the probe devices; and, their home contact list representative of family members/housemates. Figure 7 illustrates the SEIR model over a two month period, using data collected from the four probe devices and illustrating the spread of an infection through a population of 5000 persons. During this simulation, each person had a contact list of approximately 10 close contacts, reflecting the 80/20 rule found from the Pareto distribution associated with the F2F contacts. 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 second 12 hour cycle) in addition to their daytime contacts. The probability of becoming infected was p=0.0025. This was implemented as there was a 0.0025 probability or becoming infected if one of your close contacts was infected, per hour of contact. This infection probability is an

adjustable parameter associated with the simulation but consistent with considerably larger models.

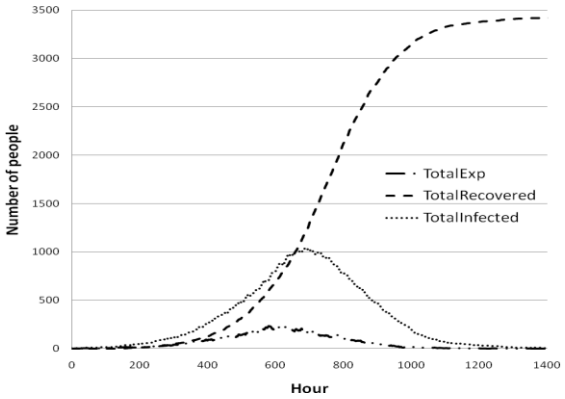

Figure 7: SEIR Compartmental Model.

The resulting curves are typical of compartmental SIR models. The only real 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" arising from the Pareto distribution of inferred contacts from an automated proximity contact pattern generator. In epidemiology, Ro is denoted the basic reproduction number of the infection and is the number of secondary infections a single infected case will cause. In the case of an nfluenza strain (e.g. 1918) Ro has been estimated to be between 2-3. In Figure 7, Ro is approximately 1.9.

The unique contribution of the work is the insights into technologies that gather adequate amounts of real data, non-intrusively, to provide meaningful input into disease spread models, where these models have typically relied on simulated data.

To further explore conditions that may be representative of remote northern communities, the number of close prximity contacts during the "home cycle" was varied from 2 to 5 as illustrated in Figure 8. This represents a tendency toward large households and/or overcrowding in homes. Qualitatively, the simulations indicate that a major contributing factor in the spread of an ILI would be overcrowding. The overcrowding exacerbates the infection spread as a consequence of increased exposure due to in increased contact[3]. Another potential public health concern is thus associated with an intervention that recommends for a infected person to stay home. In environments with severe overcrowding in homes, this may in fact be deleterious. In these scenarios it may be well worth setting up temporary mobile facilities to house and treat persons infected as opposed to recommending they stay home. Further investigations into the sensitivity of household size as a factor in infection spread is also warranted.

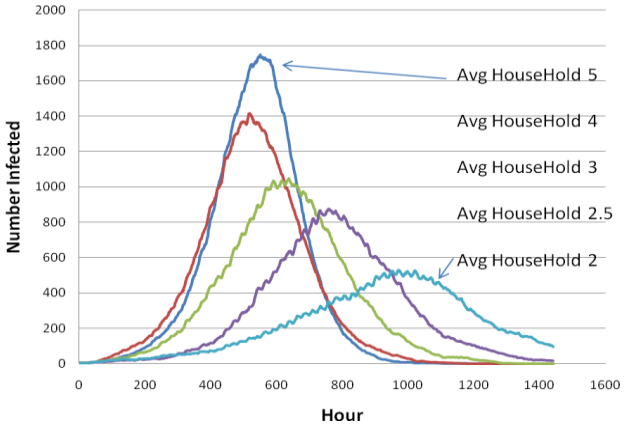

Figure 8: Impact of Overcrowding.

SUMMARY

This paper summarized a research program where a disease spread model was influenced by real data as an inference for personal contact. While limitations abound, the work highlighted the opportunities inherent in personal mobile devices as means to effectively and non-intrusively gather real data on social contact patterns.

ACKNOWLEDGEMENTS

The authors thank Manitoba Hydro, RIM, MTS Allstream and NSERC for support.

REFERENCES

[1] N. Eagle, A. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data", PNAS 2009 vol. 106 no. 36, 15274-15278

[2] http://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology#The_SIR_model

[3] A. J. McMichael, "Environmental and social influences on emerging infectious diseases: past, present and future Phil. Trans. R. Soc. Lond. B 2004 359, 1049-1058