Human Network Data Collection in the Wild: The Epidemiological Utility of Micro-contact and Location Data

Mohammad Hashemian[a], Kevin G. Stanley[a], Dylan L. Knowles[a],
Jonathan Calver[a], Nathaniel D. Osgood[a,b]
Dept. of Computer Science[a], School of Public Health[b]
University of Saskatchewan
Saskatoon, SK, Canada
{first.last@usask.ca}

ABSTRACT

Contagions – either pathogens spread through contact networks or societal memes spread through social networks – impact the occurrence and character of both epidemic and endemic diseases. While computational models explore disease parameters in the context of a given contact network, these models are always subject to the caveat that reality may not be consistent with the simplified assumptions regarding contact, contagion or network structure. More - and more accurate - data on the contact dynamics between people and places could alleviate some uncertainties, and make models more robust tools for policy-makers and researchers. Properly applied, consumer electronics can serve as a valuable source of this data. Using smartphones as sensor platforms rather than personal communications devices, it is possible to record high fidelity information on a participant's location, activity level, and contacts between both people and places. This paper describes the design, architecture and a preliminary deployment of a general smartphone-based epidemiological data collection system. The dataset, gathered over one month, contains over 45 million records related to the behavioral patterns of 39 participants. We provide an initial analysis of aggregate level statistics to demonstrate the power and scope of the technique for capturing relevant data. Demonstrating the potential for such data to inform decision-making, we further perform an agent-based simulation of a flu-like illness that uses the dataset to capture aspects of both person-person and environmental (place-person) transmission. We demonstrate that the data collection is possible, valuable, and scalable and that the data can be leveraged to inform detailed models capturing more complex physical interactions than were previously feasible.

Categories and Subject Descriptors

J.3 Life and Medical Sciences

General Terms

Measurement, Experimentation, Human Factors, Verification.

Keywords

Sensor-based data collection, Human contact pattern, Epidemiological modeling.

IHI'12, January 28–30, 2012, Miami, Florida, USA.
Copyright 2012 ACM 978-1-4503-0781-9/12/01...$10.00.

1. INTRODUCTION

The dynamics of contagion spread have been studied in systems ranging from the stock market to YouTube video popularity. Because the fundamental data, stock prices and news items, or video names and number of views are digitally archived in the public domain, these systems are readily analyzed. Contact dynamics in contagious diseases – and particularly respiratory infections – have not been as well studied because they primarily depend on fluctuating physical proximity networks, which are difficult to measure. Despite the availability of strong modeling approaches to evaluate health interventions, epidemiologists commonly lack sufficiently detailed empirical data to make strong predictions for the outcomes of interventions in systems whose evolution exhibits a strong dependence on contact dynamics, either between people or people and places.

Despite the similar nomenclature, viral videos spread across different networks than viral pathogens. Viral videos transit through a quasi-static social network, which is unlikely to change during the brief half-life of the video's popularity. Viral pathogens transmit through susceptible individuals being physically exposed to infectious carriers, usually other people, or environmental reservoirs of infection. These contacts are stochastic [6, 14] and can be classified depending on the nature of the infection [26]. A better understanding of the dynamics of contact networks – and of their associated reservoirs – could lead to a better understanding of how pathogens transfer in epidemic spread and remain viable in endemic scenarios.

A dynamic contact network is simply a mapping of the time a set of pairs of individuals are within certain proximity of each other. It is used to answer the question "was agent A in contact with agent B at time T?" Time is important: a longer exposure time increases the chance of pathogen transmission [26] and the stage of the infection often changes the transmission hazard. Traditionally, contact dynamics were estimated by manual contact tracing or diary which confers great value for infections with long latent periods (such as tuberculosis), or infections with clearly defined contact events (such as most sexually transmitted infections), but of insufficient resolution for more rapid and virulent diseases such as SARS or H1N1 flu.

In the past decade, researchers in communications and zoology have been using automated contact tracing systems to better understand the role of human mobility in communications systems [19, 30] or the interaction patterns of animals [28]. However, it was not until very recently that this was applied to the field of human health [11, 26]. Automated contact tracing of populations in epidemiological modeling and human health is in its infancy, and many areas remain unexplored, such as the role of contact with sensed locations in the spread of disease, the sampling time scale required, or the role of multisensory data to the cross-validation of conclusions from the sensed data.

In this paper we describe and validate a novel large-scale data collection system, which provides minute-level resolution measurements of participants' activity, location, person-person, and person-place contacts. We illustrate heterogeneities in contact patterns through aggregate data analysis and demonstrate the utility of model-dataset integration by combining contact dynamics data and agent-based simulations.

The dataset - called the Saskatchewan Human Ethology Dataset 1 (SHED1) - was collected over a period of 5 weeks with 39 participants and includes accelerometer, GPS, Bluetooth, WiFi and battery state information, culminating in over 45 million sensor records and hundreds of millions of individual measurements. We provide an initial analysis of this data in three results sections. Section 2 provides a literature review. The system architecture, data collection and simulation setup are described in Section 3. Section 4.1 gives an overview of the resulting contact dynamics in aggregate form. Section 4.2 provides analysis and visualizations on the interaction between contact dynamics of participants and the places they visit. Section 4.3 presents the results of an agent-based simulation that used the gathered contact dynamics as a temporal contact pattern and was studied using a Monte Carlo ensemble. Discussion, future work and conclusions are outlined in Section 5, 6, and 7, respectively.

2. RELATED WORK

Since the inception of mathematical epidemiology, human infection transmission models have provided a control to the representation of person-person contact processes. In recent decades, research has highlighted the importance of population heterogeneity and network structure in shaping outbreak emergence and progression, and endemic persistence of pathogens [13, 24]. Researchers noted that pathogens for whom population extinction would have been anticipated have managed to survive – and even flourish – in core regions of the network [13]. Research also identified the tremendous levels of heterogeneity seen in human contact patterns [27].

Insight into the importance of heterogeneity and network structure on pathogen transmission and survival has elevated the attractiveness of individual-level models, explicitly depicting static or dynamic contact networks, characterizing the impact of network structure on infection spread, and evaluating network-informed interventions. Despite recent contributions suggesting both the existence of great heterogeneity in duration [11] and the importance of such contact duration for infection transmission [26, 29], there remains a relative paucity of data on contact duration, largely due to the difficulty of collecting such information [24]. Sensor-based approaches have been identified as a significant opportunity for collecting this information [26, 24]

Many human pathogens impose a risk of transmission not only on a person-person basis, but also through the environment. Environmental reservoirs vary from air within an enclosed space, to surfaces, and aquatic environments. Compared to the heavy emphasis of mathematical epidemiology models on understanding the role of carriers, the role of place-based environmental reservoirs has been the focus of less modeling effort. In the social network community, place is recognized as informing understanding of the context and significance of contact patterns [35, 18, 17]. While place has featured prominently in partial-differential equation models for characterizing geographic spread of illnesses such as rabies [2, 3], West Nile Virus [2] and pandemic flu [33] as well as in geographically rooted agent-based models [5], and while environmental reservoirs are more commonly represented for models of some zoonoses [10, 22], relatively few human health models examine the impact of environmental reservoirs. Most such models concentrate on highly aggregated representation of a single aquatic environment [4, 25, 15], but others have portrayed highly aggregate characterizations of alternative environments, such as surfaces in health-care facilities [21]. As with person-person contact, the absence of detailed contact patterns (here, between people and places) has proven a major barrier for the construction of more detailed models.

While epidemiologists have found a great deal of utility in quasi-static contact networks, other disciplines have investigated human contact dynamics for communications purposes. Reality Mining [6] instrumented students and staff at MIT to study their inter-contact dynamics. They found that contact duration tended to follow a truncated power law distribution. Others have employed similar techniques to investigate the utility of Delay Tolerant [8] or Pocket Switched Networks (DTN, PSN) [14], a networking paradigm which routes low priority messages through contact between mobile agents rather than over fixed network infrastructure. Their characterization of human contact patterns was broadly consistent with Reality Mining, and added confirmatory analysis. Processes in PSN–like networks exhibit patterns similar to pathogen propagation in that they transit from person to person based on contact frequency and duration. However, PSNs route messages to minimize power consumption and maximize delivery ratio, whereas pathogens behave in a more stochastic manner. More recently, authors have examined the impact of location in DTN routing [12] leveraging datasets that contained location information. This research is partially enabled by datasets which incorporate location, such as the original Reality Mining [6], which employ cellular tower occupancy as a proxy for position, and [19] which records GPS positions for various subjects within single 24 hour periods. Neither of these methodologies is particularly suitable for the study of environmental reservoirs for contagious disease, however, because the resolution of cellular tower localization is overly coarse, and GPS is only reliable outdoors.

Quasi-static assumptions regarding contact dynamics are suitable for infections which move with very slow contagion dynamics, but even systems with easily identifiable distinct contacts such as most sexually transmitted diseases have been demonstrated to be strongly affected by network dynamics [23]. For infections with greater virulence and shorter duration, quasi-static analysis may not be adequate [26]. However, researchers have recently leveraged automated contact tracing as described in the DTN literature with simple agent based models [6, 12]. The work in [26] represents an important step forward in the integration of detailed micro-contact data and epidemiological modeling of contagious disease. However, it is only a first step and has a number of shortcomings, particularly due to the nature of the micro-contact data gathered and the choice of infectious disease model. The dataset collected contained over 800 participants – larger than most publish micro-contact data sets (e.g. [6, 11]) – but only over a single day, which is an exceptionally short duration. A generic infectious disease model was then applied to the dataset over and over, as the same contacts replayed day after day. This substantial simplification has validity for the population under study: high-school students and staff, a population also of particular epidemiological interest, but also a population that generalizes poorly. Additionally, this research did not consider the impact of place, likely because the entire study took place within a single institution with regimented location-occupancy for almost all participants. We address both the shortcoming of time and the neglect of place in the dataset presented here, but sacrifice number of participants as a tradeoff.

3. EXPERIMENTAL SETUP

3.1 System Design

SHED1's data acquisition backbone, iEpi, is a custom Android program written in the Java language to provide extensible sensor data acquisition components, stable encryption and opportunistic uploading. iEpi has been designed to allow ease of code reuse and extensibility as described below, but here we only provide a brief description of its major components.

3.1.1 Major System Components

iEpi can be thought of as a system composed of five parts: tasks, streams, data loggers, data senders, and servers. *Tasks* are pieces of work or duties to be carried out periodically, including sampling data and initiating dialogues with servers. *Streams* are feeds gathering data from on-device sensors or data derived from them. In this study, we used five streams: accelerometer, Bluetooth scans, WiFi scans, GPS, and battery status. More complex streams capable of producing data derived from several sources are also possible but not currently implemented. *Data loggers* encrypt and store data collected by tasks in on-device nonvolatile storage. *Data senders* take stored data and transmit it to servers. *Servers* decrypt, sort, combine, and store data for later use by researchers.

3.1.2 High Level System Operations

At device boot, iEpi automatically starts and creates tasks based on expected system behavior defined in a configuration file. Each task then performs their respective duties at the frequency and duration specified in the file. Tasks that collect data pass it to a data logger, which encrypts and stores the data in non-volatile storage. Tasks that control data transmission periodically invoke a data sender, which sends stored data opportunistically to the indicated server(s) via WiFi. In the interest of preserving battery life, between bouts of work, the device is permitted to enter a power-saving mode. This behavior does not interrupt normal device operation, only suspending the processor when work is not scheduled. Figure 1 depicts the iEpi architecture.

3.1.3 iEpi Configuration

As previously mentioned, iEpi is configurable; researchers can specify a variety of tasks to be performed, how often to perform them, and for what duration to perform them. Currently, iEpi operates on a simple duty cycle model. Researchers specify the length of a *duty cycle* (in this case, 5 minutes) over which all measurements will be repeated. Within each duty cycle, each sensor collects data for a duration specified in the configuration file, which is less than the overall duty cycle. This schema allows researchers to determine the amount and frequency of data to collect for each stream.

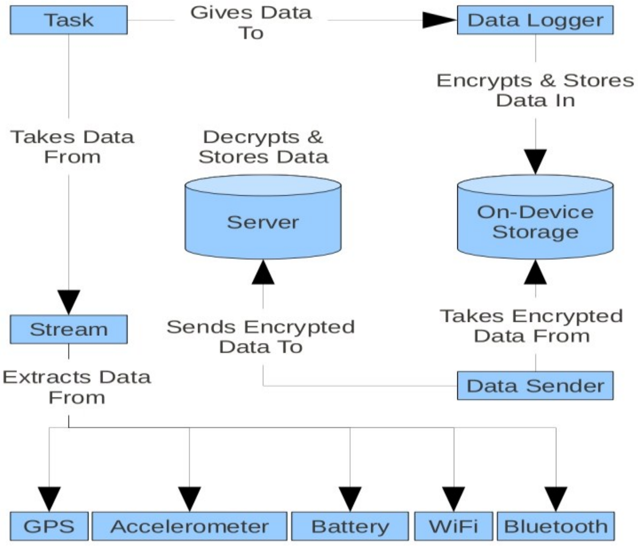

Figure 1. Simplified iEpi architecture and operations

3.1.4 Data Security

iEpi has the potential to record highly sensitive participant data. It is therefore important to ensure that participant data is virtually unreadable on-device and during transit to servers. Data is encrypted while on the phone, and only sent across secure wireless links. Data collection can be partially disabled by the user via a "snooze button," which can be accessed by opening the application. The button causes iEpi to redact all data but timestamp and participant ID helping to differentiate iEpi failure from a participant's desire for privacy.

3.2 Experimental Design

After receiving approval from our Research Ethics Board and employing the software described in the previous section, we ran a pilot deployment for 5 weeks during April and May of 2011 using Android DevPhone 2s running a custom version of the Android 2.1 operating system. Forty participants were recruited from the Computer Science Department, consisting of graduate students from several laboratories, technical staff and administrative staff. One participant withdrew within the first week, leaving 39 participants who completed the entire study. Results are presented here for these 39 participants. Phones were deployed incrementally over 3 days leading up to the experiment. Participants met one-on-one with at least one study organizer, and were walked through the experimental protocols and use of the phone, filled out consent forms, and had the opportunity to ask questions. Participants were requested to carry the phones with them at all times during the day, unless the phone was low on batteries, in which event they were requested to plug it into a computer near them. Participants were also requested to take the phone home with them at night, and to initiate charging just before going to bed. Participants were allowed to use the Android phone as their primary phone if they had a compatible SIM card. The phones were also pre-loaded with pay-as-you-go data plans with sufficient value for unlimited use over the entire study.

The phone was programmed to collect data in bursts every 5 minutes (defined duty cycle length) to manage data size and battery life. Every duty cycle, the phone logged 1 minute of accelerometer records, 1 minute of Bluetooth contacts, 3 seconds of WiFi contacts and 10 records of battery state. GPS records

Table 1. Data recorded by each sensor

Parameter	Variables recorded
ALL	Participant ID, time stamp
GPS	Latitude, longitude, velocity, accuracy
Acceleration	Acceleration in x, y, z
BlueTooth	MAC address, signal strength
WiFi	BSSID (MAC address), SSID (Network Name), signal strength, frequency, security protocol
Battery	Battery level, plugged status, battery status.

were collected for 2 minutes, but given that the GPS required significant time to acquire satellites to achieve position lock, the first approximately 90 seconds did not contain data. The information recorded by each sensor is summarized in Table 1. Values in the "ALL" row correspond to common variables.

Data was opportunistically uploaded by phones over the university's secure wireless network whenever the phone had accumulated at least 3000 records. Data on the server was accumulated in flat files and parsed at regular intervals and inserted into a MySQL database. Overall compliance was monitored by examining the number of records returned by participants. Participants with low compliance and those whom their return rate dropped significantly were notified through email.

At the conclusion of the study, participants returned their phones and filled out a questionnaire, which contained basic demographic information, information about perceived compliance and lab/office affiliations. No health data was collected in the survey as a condition of our ethics approval.

3.3 Simulation Setup

The collected dataset captures the high-resolution behavior patterns of participants during the experiment period. We used part of the dataset which represented participants contact patterns (using Bluetooth proximity) and their location information (based on WiFi-Router connectivity) to simulate the spread of a flu-like infection through proximity contacts and location-specific environmental reservoirs. Note that the transmission of infection simulated here does not reflect an actual pathogen, and the parameters are mainly for demonstration.

The simulated model classified each individual in the sample population into one of six states: *Susceptible*, *Latent*, *Asymptomatically Infectious*, *Symptomatic Infectious*, *Symptomatic Non-Infectious*, and *Recovered*. All of the agents in the model started in the *Susceptible* state. A susceptible individual could contract the infection either from exogenous or endogenous sources. Exogenous sources are defined as the population outside the study who were in contact with SHED1 participants and could transmit the infection to the monitored individuals. Assuming that 0.03% of population receive the infection per week during a pandemic [9], the exogenous infection probability per person per time unit (duty cycle) would be set to 0.0003. The endogenous source of infections is divided in two parts: contact with other study participants in an infectious state, or contact with a location-specific reservoir of infection.

Receiving the infection from either exogenous or endogenous sources transitions a susceptible agent to the latent state. Before starting the *Latent* period, the model computed the duration for each of the subsequent four stages of illness (i.e. Latent, Asymptomatically Infectious, Symptomatic Infectious, and Symptomatic Non-Infectious). In determining these durations, we sought to reproduce the observed variability in H1N1 progression by computing these durations using parameters from [32].

Each infected agent experienced the four illness states sequentially with the passage of time. A person in the *Asymptomatically Infectious* or *Symptomatic Infectious* state was considered infective, and could infect other susceptible adjacent individuals and their current location. The probability of infecting a susceptible individual in proximity per duty cycle was set to 0.00730, which is aligned with R0 reported in [32]. Unlike person-person infection transmission, an infective would spread pathogen to a location according to contagious events. An infective caused a contagious event on average once every three duty cycles. Each contagious event was modeled as increasing the level of pathogen in the current location of the infective by an amount sufficient to have a per-duty-cycle risk of transmission from the location to any susceptible visiting that location of 0.021. This increase was to be designed sufficiently large that the cumulative chance of a given susceptible becoming infected by an infectious event strictly after it has occurred was equal to the chance of a person present at the time of the being infected through the event itself. This per-duty-cycle probability could be saturated at 1 due to a high rate of contagious events. Infectivity of a location decreased exponentially and disappeared after 12 duty cycles (1 hour), assuming no additional contagious events occurred in this period.

We implemented the model in Network-Simulator 3, a discrete event simulator, using 39 agents and simulation period of 9792 duty cycles. Each agent represented one of the study participants, and participants' Bluetooth contacts and WiFi-based location information for each duty cycle was imported from SHED1 dataset to the related agent prior to the simulation. Three different scenarios were simulated: transmitting infection only via person-person contact, only via person-location contact, and both. Each scenario simulated using 100,000 Monte-Carlo realizations.

4. RESULTS

The kind of dataset described in this paper can be used to infer many aspects of human behavior which have an impact on health. In this manuscript we focus on the relationship between human-human contact and place. The role of contact dynamics – and, in particular, contact duration – is particularly important for the spread of communicable disease [26]. Location also plays an important role in the spread of contagious disease, both as a proxy for locales where person-person contact is likely, in norm-setting [20], and by hosting pathogen reservoirs, through mechanisms such as contaminated surfaces or suspended aerosols [34]. High-fidelity contact data of the sort collected in this study can be used to derive more accurate contact parameters for aggregate models, provide direct empirical insight into the behavior of the monitored population and as a source for aggregate models' mixing matrixes and dynamic contact patterns in agent-based models.

4.1 Aggregate Data Properties

With 45 million individual records and hundreds of millions of individual data points, the full implications of rich datasets like SHED1 can be difficult to assess. Aggregate measures can provide simple snapshots to represent the data as values, distributions or functional parameters. These values can represent broad trends in the data and can be employed as direct health

Table 2. Aggregate information on SHED1 dataset

Parameter	Value
Total GPS records	1,348,024
Total WiFi records	9,285,061
Unique WiFi routers	20,069
Unique WiFi locations	34,048
Total BT records	1,630,519
Unique BT MACs	9934
Contacts (duty cycle) between participants	74056
Contacts (duty cycle) with non-participant mobile BT devices	511038

measures or as the input to population level models. In this section we present measures of overall dataset scope, particularly with respect to the person and place contact data contained in the WiFi and Bluetooth tables. Table 2 contains simple aggregate data relating to the scope of the dataset.

In Table 2, a WiFi location is distinguished by a unique combination of routers visible to a participant with RSSI values of at least -80 dB. A participant contact is registered when one participant's phone discovers another with a MAC address in the list of participant devices. A non-participant contact is recorded when a participant comes into contact with a node which is discoverable, has a device class of cellular or smartphone and is not in the MAC address list of participant devices. It is possible for multiple contacts to happen in a single timeslot with a single node, if other nodes observe it simultaneously. The distribution of contact count is shown in Figure 3.

Participants were significantly more likely to have seen someone a very few number of times, and most of those isolated contacts were with non-participant devices. Because we selected participants to be from the same department, they had a reasonably high chance of contacting one another. The minimum recorded contact between any participant pair was 311 duty cycles. There is a chance the data might be biased due to our measurement of contacts with only discoverable non-participant devices for practical and ethical reasons. Our data only indicates that a device was discoverable at the specific point in time; it does not imply that the non-participant device was discoverable for the other time slots in the study. Nevertheless, even excluding devices seen at most in 5 different timeslots, the heavy tail remains, and the overall nature of the contacts remains the same.

Having established the extent of contact, it is logical to examine the duration of those contacts. This is often reported as CCDFs of contact duration [6, 11] corresponding to the amount of time nodes are likely to spend together. Figure 3 shows contact duration CCDF, for contacts between participants and contacts between all mobile nodes. Similar analysis of two other contact datasets, which captured contacts only between participants are provided for reference as well [14, 11].

Each of the contact duration CCDFs follows the general trend of a power law followed by an exponential roll-off, although the roll-off is less pronounced in the SHED1- All Mobile Devices. An interesting aspect of the CCDFs is the divergence of the all contacts line from all three of the other datasets. This suggests an impact of participant selection bias on the contact dynamics of the network. The Haggle dataset was captured over several days at an academic conference. The Flunet and SHED1 datasets were gathered at the same university, but 18 months apart.

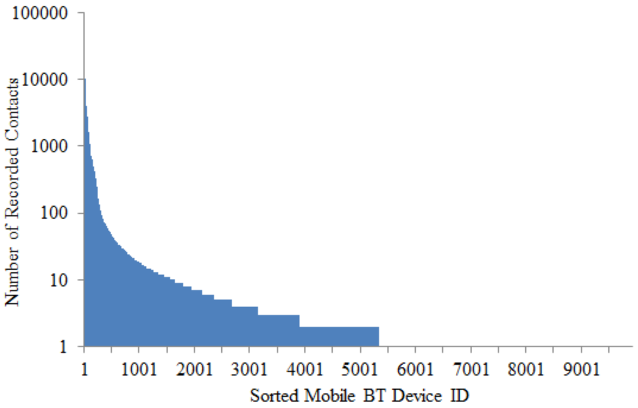

Figure 2: Number of contacts recorded for each mobile node

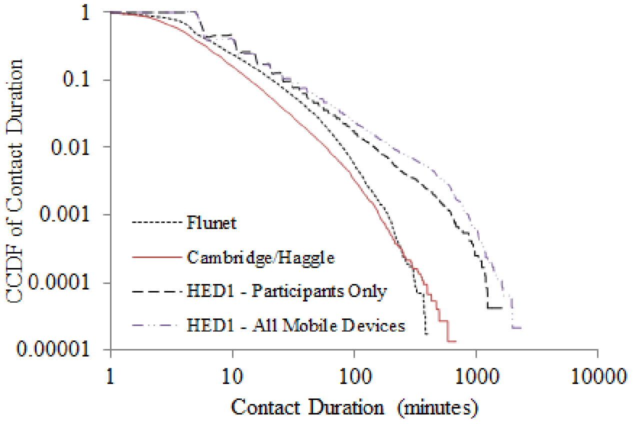

Figure 3. CCDF of BT contact durations. CCDF for contacts from two other studies are also provided for comparison.

Another fundamental parameter is the overall reliability of the data. Data reliability is much more difficult to measure in the previous data collection methods such as Flunet [11] and Haggle [14] because the data was acquired using a simple sensor system intended for industrial use. However, we can leverage the battery data and auto-synchronizing clock to determine how reliable each participant was, even on a daily basis. Figure 4 show the amount of time the phone was on with battery (and more likely with the participant) with red, on while connected to a computer (likely proximate to the participant most of the time if at work) with green, or plugged into a wall (less likely to be close to the participant) with blue. The sum of these parameters is the total on time of the phone and a proxy for participant compliance.

Overall compliance was moderate, averaging 54% for all battery states. The highest compliance participant reported 85% of the possible data over the study period, and the participant with lowest compliance reported approximately 30% of the possible data. Another notable exception was participant 11, who left his phone plugged into a computer most of the time, even at night, suggesting that the phone was a better proxy for his desk than himself. Nevertheless, the scope and richness of the data obtained are more than sufficient for the initial analysis presented here, as a demonstration of principle, and the additional richness provided by orthogonal sensor measurements of compliance provides us with a greater degree of certainty as to what the data is describing.

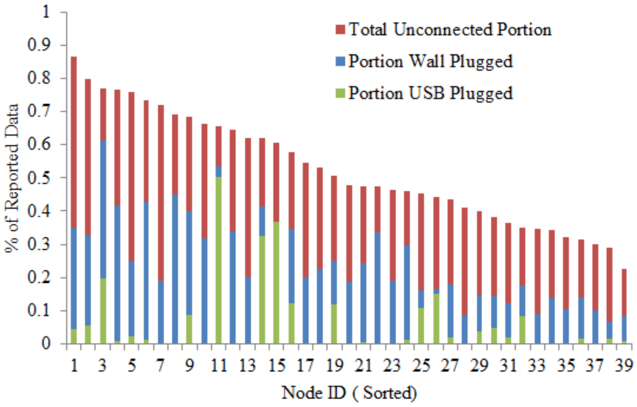

Figure 4. Participants' compliance, grouped by phone's plugged status.

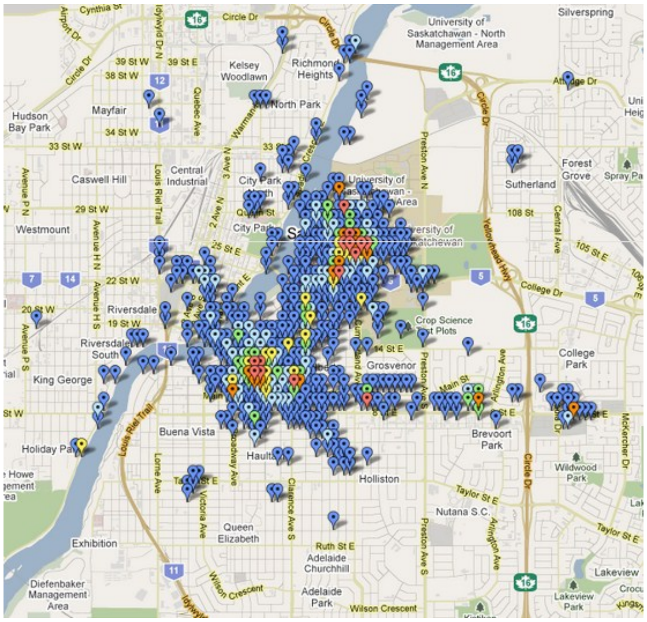

Figure 5. Mobility pattern of a participant during the study. Red shows higher and blue shows lower number of samples.

In addition to WiFi-based locations, GPS records also can represent the location information in the SHED1 dataset, particularly in outdoor environments where the density of WiFi routers are lower. Figure 5 shows the density of GPS records from a sample participant during the experiment. Colors closer to red indicate higher number of samples, and colors closer to blue are related lower number of samples. Although a few visits are recorded to most areas of the city, two primary locations, the commute path, and some favorite shopping places are easily identifiable, allowing us to infer the participant's mobility pattern.

4.2 Network analysis

To explore relationships connecting people and places within the study, we performed a variety of network analyses, each accompanied by network visualizations.

Figure 6 shows a network reflecting the relationship over work-week (Mon. to Fri.) between participants and non-participants (cellphone or smart-phone mobile Bluetooth devices). Within this network, a participant was considered connected to another person (a Bluetooth device involved in the study or not) if and only if there was some timeslot during that work-week in which the participant detected that device with a signal strength of at least -80 dB. The network is characterized by a strong core of participants, and "fans" of their external or stochastic contacts. While this is an intriguing phenomenon, it could be a sampling artifact. (If the entire city were telemetered or if the study was run for a longer time, the fans might become webs). However, it does suggest that there are heterogeneous risks for endogenous infection, an interesting topic for further study.

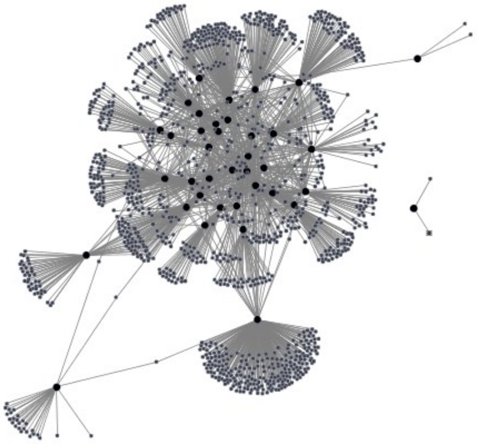

Figure 6. Network involving participants (black) and mobile Bluetooth devices (grey)

Figure 7 shows a visualization of a subset of the network where circles represent WiFi locations over the entire study. Within the figure, the circle associated with each place is sized such that the area is proportional to the number distinct non-study mobile devices seen per hour at that location, and the color associated with a given node is progressively brighter according to the cumulative number of distinct non-study mobile devices seen at that location throughout the study. As examined in the simulation (Section 4.3), the WiFi locations in the networks depicted here are of importance to spread of infection not only as the context for person-to-person transmission of infection, but also as pathogen reservoirs. For example, a place node with high temporal rates of seeing new individuals could be a highly-travelled location in which environmental reservoirs could be built up and maintained by traffic. A place node with a large cumulative number of observed non-study mobile devices suggests that participants may have remained at that location for a substantial amount of time, exposing them to higher cumulative risk infection from other individuals or environmental reservoirs. The capacity to collect information on contacts with and locations of those outside the participant population can lower degree of bias imposed by participant selection and help to cross-validate observations of behavior.

Figure 8 depicts a different form of WiFi device network, one which depicts only WiFi devices over a given interval of time. A pair of WiFi devices is considered connected in the network over that interval if there was at least one time slot during that interval in which a particular participant detected *both* devices with RSSI strength levels of -80 or stronger. Nodes colored in blue are known through their SSID to be affiliated with one of 4 standard University of Saskatchewan networks. The centrally located subset of the network shown in the figure includes primarily nodes within the city of Saskatoon, but others as well. For example, the isolated component located towards the top of Figure 8 appears to consist of a variety of WiFi devices located in Edmonton, Alberta – a city approximately 5 hours driving distance from Saskatoon, which was visited by at least one participant during the study. The capacity to understand place-

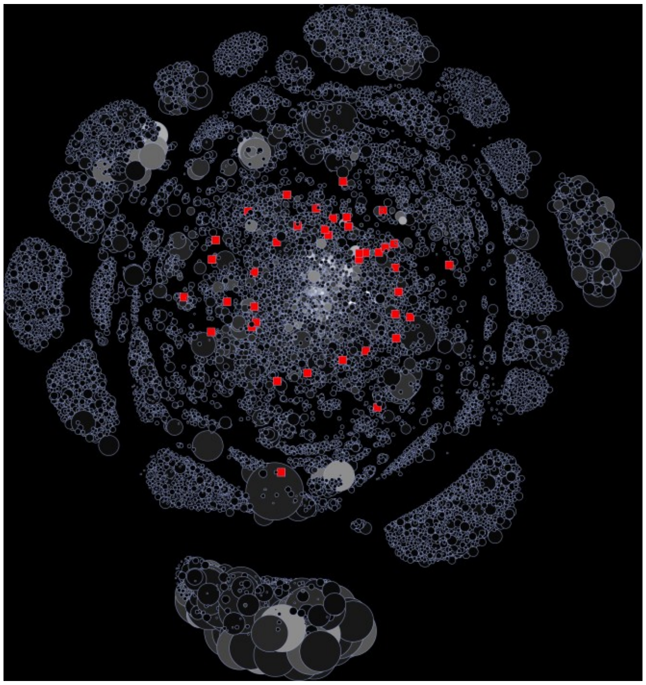

Figure 7. Network of person-place (distinct devices) contacts (squares represent participants, and circles represent places)

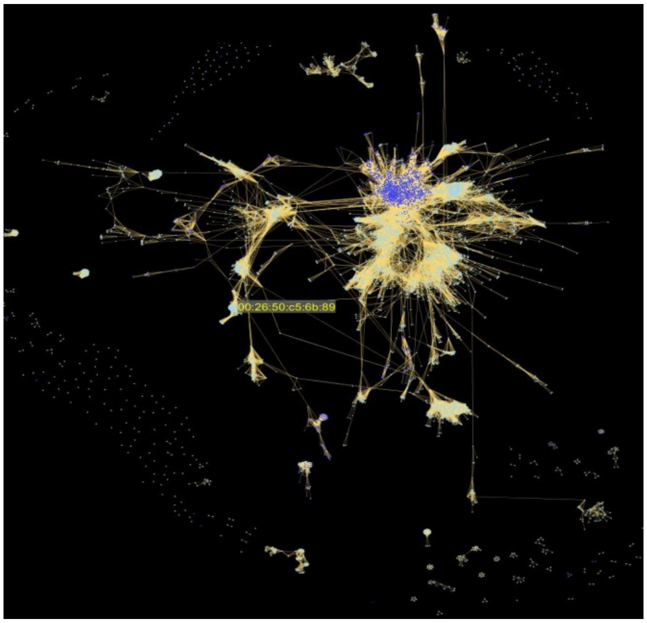

Figure 8. Structure of proximity network of WiFi devices

place connectivity raises important potential for richer epidemiologic analyses, such as those explored in Section 4.3.

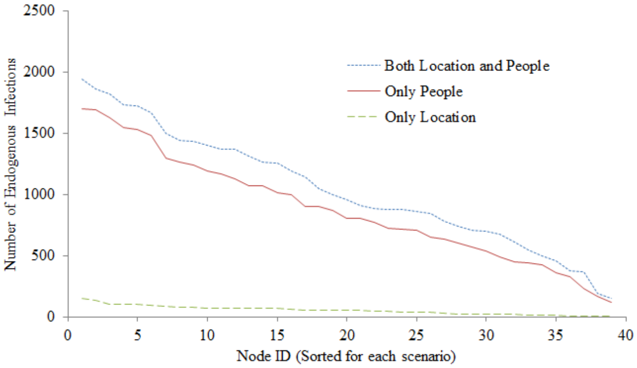

Figure 9. Number of endogenous infections per node for each scenario

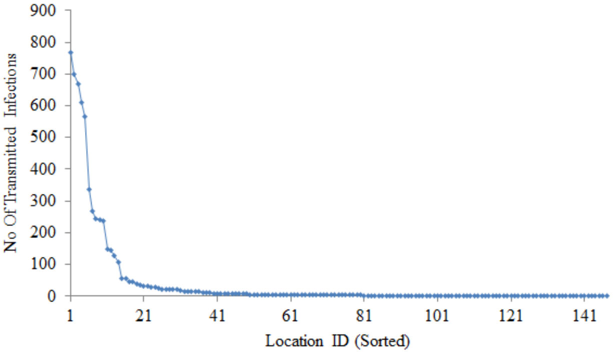

Figure 10. Number of transmitted infections per location.

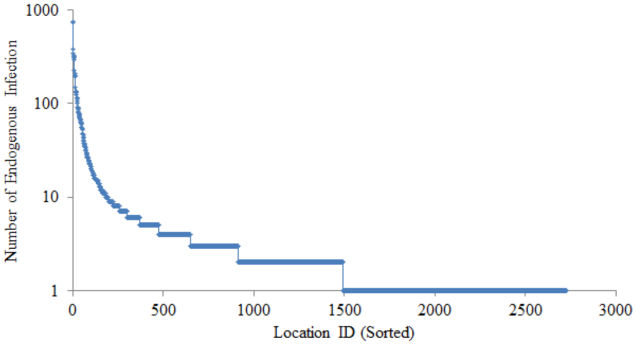

Figure 11. Number of person-person infection transmissions happened at each location where infection occurred.

4.3 Simulation Results

To demonstrate how the behavioral patterns recorded in the dataset can be used in pathogen transmission models, we focused on participants' locations and contacts data in SHED1 dataset, and used them in an infection transmission model. We studied the effect of location and proximal contacts, both separately and combined, in three different scenarios. Each scenario simulated for 100,000 Monte-Carlo realizations. In each realization, the agents followed the same contact sequence and location movement, while the stochastics associated with infection progression, duration, and transmission changed.

The attack rate for the first scenario, where infection transmission could happen both via people and environmental reservoirs, was 0.0115. This metric for the second scenario, where infection only could transfer via people, was 0.0098 and for the third scenario, where the infection only could happen through the reservoirs, was 0.0016. The substantially lower attack rate value in third scenario partly could be due to the short life-cycle of the pathogens in the environment (12 duty cycles, or 1 hour), and the exponential decay that occurs during that time. Clearly the assumptions underlying the model will affect the output; different relative hazards for person-person and person-place will result in different simulated health outcomes, but the parameters here are broadly consistent with a flu-like illness and serve as a proof-of-concept for the technique.

Figure 9 shows the number of endogenous infections received by each participant, during all realizations for each scenario. This graph also emphasizes the role of person-to-person contacts in pathogen transmission in comparison with shared location, as both scenarios with person-person infection transmission yielded considerably higher number of endogenous infections.

We used 34,048 unique locations based on scanned WiFi-Routers for the simulation, but not all the locations have equal importance in transmitting the infection. In all 200,000 realizations where place-person transmission was possible (first and third scenarios), only 147 locations infected at least one susceptible. Figure 10 shows the number of transmitted infections in each of these 147 locations. As it can be seen, a few locations infected more than 500 susceptibles (aggregated over all realizations), while the majority of locations caused less than 50 infections.

To understand the role of location in person-person infection transmission, we measured the number of endogenous infections which happened due to proximity of two people at each location. Simulation results showed that at least one person-person infection happened in 8% of the locations. Figure 11 shows the number of person-person transmissions occuring at each location, omitting those locations with no transmission. A behavior similar to Figure 10 can be seen here as well. A considerable number of infection transmissions happened at a small set of locations, while less than 10 transmission happened in a majority of the locations.

In aggregate, our simulation results demonstrate that place plays an important role in disease transmission even of short-lived environmental pathogens. However, this role is not primarily through the transmissions of pathogens via environmental reservoirs, but through common locals where transmission can take place. Identifying places with either a high degree of mixing or longer contact durations [26] could help reduce the spread of disease by prioritizing target areas for public health information resources.

5. DISCUSSION

The work we have described covers aspects of data collection, analysis and simulation. We are the first to our knowledge to collect multivariate data of this scope for health modeling and analysis, and the first to apply agent-based techniques and micro-contact data to the spread of pathogens from both agent-agent transmissions, and agent-environment transmissions. This paper makes specific contributions to each area, but perhaps more importantly demonstrates an overall integrated approach covering all aspects of the health-centered dynamic contact network analysis, from the software design of the tool through to agent-based analysis. We feel that this vertically integrated approach is appropriate for the study of contact network dynamics.

Both the analysis and the simulations focused on the relationship between people and places, and the associated impact on infectious disease. Our analysis showed that people had contact with a subset of common places and associates, but also had transient contact with a staggering number of other people and places. Our simulations indicate that people were more likely to receive infections from people or places in which they more commonly reside and where others more commonly reside, echoing the notion of strength from [26].

While the analysis and simulation work reported here have value unto themselves, the largest contribution of the paper is the overall process of orthogonal behavioral and health data collection, simulation and analysis. We have only begun to tap the potential of the dataset we have already collected. We have not addressed measures of activity in this paper, and have only mentioned the potential of GPS data. This data can be leveraged to provide greater precision in data description, by cross-referencing the WiFi-based localization described here with GPS data, as 90% of the unique locations described have associated GPS positions within the same duty cycle. This same data could also be leveraged against GIS databases to provide detailed investigations of situated activity and dining habits or to examine how residents of specific neighborhoods utilize services within and outside of their neighborhoods. We would contend that the primary impact of this work is the process itself as it has such far-reaching impacts into many aspects of epidemiology.

While the overall approach described here has profound implications to health research, the data collection and simulations only constitute a first step. The number of participants and duration of the study introduce clear selection bias, and do not capture longer-term changes in contact dynamics, occurring, for example, over the course of different seasons. These shortcomings of numbers and duration are common to other similar studies [26, 11, 6, 14], and relate to a large extent to the infancy of the methodology. The technical limitations to extending our system to a much larger study population are surmountable, but the logistics of marshaling thousands of participants for longitudinal study would require substantial infrastructure and organizational resources.

A duty cycle-based data collection system can only provide samples of reality, and always run the risk of missing data. As aggressive as our data collection was, we are still only capturing Bluetooth contacts 20% of the time and WiFi only a fraction of that. We based these numbers on a compromise between sensitivity and battery life. Additionally, as noted earlier in the paper, non-participant devices only observed once or twice may have been proximate to participants at other times but not longer in discoverable mode.

The simulation studies we performed demonstrated a powerful combination of agent-based Monte Carlo techniques and contact dynamics data, but the examples we provided were highly stylized version of flu, and lack ground truth through diagnostics or survey data as a consequence of our ethics approval. Additionally, even with the high number of locations we have access to in the data, it is certainly not of sufficient resolution to identify individual surfaces within a space. Our spatial probability of infection is then a joint probability of the probability that the agent will come into whatever surface has been contaminated and the probability of infection from the environmental pathogen itself.

Despite these shortcomings, our technique has delivered interesting insight into the role of contact dynamics between people and places, and an overall health informatics approach which has the potential to fundamentally alter the relationship between data and simulation.

6. FUTURE WORK

6.1 Data Collection

The system we have designed employs sensors which are commonly embedded in smartphones. In the future we wish to add spot surveys, communications monitoring and better directionality sensing to more fully exploit the available smartphone capabilities. While capturing sound and images is also possible using standard smartphone sensors, these modalities are even more ethically fraught, and therefore likely best avoided unless greater benefit can be demonstrated. Smartphones can be more than just sensor nodes; their local communications capabilities and significant processor and memory capacity can be leveraged with secondary sensors to capture additional medically relevant data such as blood pressure, blood glucose level or even weight [31]. Finally, we and others, should expand the scope of population under analysis from academic institutions to medical institutions and high-risk subgroups of the public at large to better understand infections and mitigate against the population bias noted in our introduction and results.

6.2 Data Analysis

The analysis we have presented in this paper has only scratched the surface of the data we already have in hand. Within this paper, we have omitted discussion of two entire sensor modalities with potential for greatly informing an understanding of detailed behavioral and environmental drivers for health patterns: acceleration and GPS position. When coupled with geographic information systems (GIS) these data streams can provide powerful insight into the impact of place on activity and by extension to chronic diseases such as obesity and diabetes. We can also leverage the significant number of orthogonal data streams at our disposal to validate the quality of the data beyond battery state analysis. For example, we could examine accelerometer records to determine if the phone was left on a level surface for a significant amount of time while not plugged in, to seek to isolate those records where the phone was unplugged but not on the participant's person. Finally, we can use the data outside of the medical milieu. For example, GPS records and connectivity data can be analyzed to give powerful insight into human mobility and connectivity in the information age.

6.3 Simulation Studies

The simulation studies we have described here provide an interesting new methodology for leveraging detailed contact data over time and place. However, we used a highly stylized disease model as a proof of concept, and additional work should be conducted to better understand the sensitivity to disease behavior and parameters to understand how dynamic network models depart from more traditional aggregate models or network-embedded agent-based models. In a similar vein, we seek to analyze the impact of temporal aggregation on agent-based models to examine the degree of temporal resolution required to yield health insight, and at what temporal scales the dynamic network resembles a classic aggregate or population level model. Finally, we should look at how detailed contact data can be replicated through simulation over participants and through time. While [26] has made an initial attempt at this leveraging, it is unclear whether their or other approaches have sufficient empirical or mathematical validity.

7. CONCLUSIONS

This paper presents the design, deployment and analysis - directly and through associated epidemiological models - of a smartphone-based human contact and mobility data collection system. We have demonstrated the feasibility of the system for data collection, identified interesting aspects of human contact dynamics with people and places and integrated these findings and data with agent-based simulation models, with associated contributions in all aspects of our design and analysis. This work constitutes a foundational demonstration of the future of ecologically valid epidemiological data collection, and an important first step in the understanding of human contact dynamics and health. In the future we intend to push the field forward along the three fronts of data collection, analysis and modeling to achieve greater understanding of contagion spread through human mobility and contact.

Acknowledgements: We would like to acknowledge NSERC for funding this work.

8. REFERENCES

[1] Anderson, R. M. and May, R. M. 1991. Infectious diseases of humans : dynamics and control. *Oxford; New York: Oxford University Press*

[2] Brauer, F., van den Driessche, P., and Wu, J. 2008. Spatial Structure: Partial Differential Equations Models *Mathematical Epidemiology* 1945 (2008), 191-203.

[3] Chunhua, O. and Jianhong, W. 2006. Spatial Spread of Rabies Revisited: Influence of Age Dependent Diffusion on Nonlinear Dynamics. SIAM J. Appl. Math. 67:1 (Nov. 2006) 138–163.

[4] Codeco, C. 2001. Endemic and epidemic dynamics of cholera: the role of the aquatic reservoir. BMC Infectious Diseases, 1:1 (Feb. 2001), 1.

[5] Dibble, C., Wendel, S., and Carle, K. 2007. Simulating pandemic influenza risks of US cities. *Proceedings of the 39th Winter Simulation Conference*, 1548-1550.

[6] Eagle, N. and Pentland, A. 2006. Reality mining: sensing complex social systems. Personal and Ubiquitous Computing, Volume 10 Issue 4, 255-268.

[7] Eames K. T. D. and Keeling M. J. 2003. Contact tracing and disease control. Proceedings of the Royal Society B: Biological Sciences, 270 (Dec. 2003), 2565-2571.

[8] Fall, K. 2003. A delay-tolerant network architecture for challenged internets, Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications, August 25-29, 2003, Karlsruhe, Germany, 27-34.

[9] FluWatch, Public Health Agency of Canada, viewed 29 June 2011 <http://origin.phac-aspc.gc.ca/fluwatch/09-10/w34_10/index-eng.php>.

[10] Hagenaars, T. J., Donnelly, C. A., Ferguson, N. M., and Anderson, R. M. 2000. The transmission dynamics of the aetiological agent of scrapie in a sheep flock. Mathematical Biosciences, 168:2 (Dec. 2000), 117-135.

[11] Hashemian, M., Stanley, K., and Osgood, N. 2010. Flunet: Automated tracking of contacts during flu season. In Proc. of The 6th Intl. workshop on Wireless Network Measurements, (Avignon, France, May 31, 2010), 348-353.

[12] Hashemian, M., Stanley, K. G., 2011. Effective Utilization of Place as a Resource in Pocket Switched Networks, to be appeared in 36th IEEE Conference on Local Computer Networks, (Bonn, Germany, October 4-7, 2011).

[13] Hethcote, H. W. and Yorke J. A. 1984. Gonorrhea transmission dynamics and control. Springer Lecture Notes in Biomathematics. Berlin, Springer.

[14] Hui, P., Chaintreau, A., Scott, J., Gass, R., Crowcroft, J. and Diot, C. 2005. Pocket switched networks and human mobility in conference environments, In Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking, 244-251, August 26, 2005, Philadelphia, USA.

[15] Ionides, E. L., Breto, C., and King, A. A. 2006. Inference for nonlinear dynamical systems. Proceedings of the National Academy of Sciences, 103:49 (Dec. 2006), 18438-18443.

[16] Kermack, W. O. and A. G. McKendrick. 1927. A Contribution to the Mathematical Theory of Epidemics. *Proc. R. Soc. Lond. A* 115: 772 (Aug. 1927), 700-721.

[17] Klovdahl, A., Graviss, E., Yaganehdoost, A., Ross, M., Wanger, A., Adams, G., et al. (2001). Networks and tuberculosis: an undetected community outbrea involving public places. Soc. Sci. Med., 52 (Mar. 2001), 681-694.

[18] Latkin, C., Mandell, W., Vlahov, D., Oziemkowska, M., and Celentano, D. 1996. People and places: behavioral settings and personal network characteristics as correlates of needle sharing. J. Acquir. Immune Defic. Syndr. Hum. Retrovirol., 13 (Nov. 1996), 273-280.

[19] Lee, K., Hong, S., Kim, S. J., Rhee, I., and Chong, S. 2009. SLAW: A Mobility Model for Human Walks. In Proceedings of INFOCOM, (Rio de Janeiro, Brazil, April 19-25, 2009), 855-863. DOI=10.1109/INFCOM.2009.5061995

[20] Madan, A., Cebrian, M., Lazer, D. and Pentland, A. 2010. Social Sensing for Epidemiological Behavior Change, In Proceedings of the 12th ACM Intl. conference on Ubiquitous computing(Copenhagen, Denmark, Sept. 26-29, 2010), 291-300.

[21] McBryde, E. S. and McElwain, D. L. S. 2006. A Mathematical Model Investigating the Impact of an Environmental Reservoir on the Prevalence and Control of Vancomycin-Resistant Enterococci. Journal of Infectious Diseases, 193:10 (May 2006), 1473-1474.

[22] Miller, M. W., Hobbs, N. T., and Tavener, S. J. 2006. Dynamics of prion disease transmission in mule deer. Ecological Applications, 16:6 (Dec. 2006), 2208-2214.

[23] Morris, M. and Kretzschmar, M. 1995 Concurrent Partnerships and transmission dynamics in networks. Social Networks, 17:3-4 (Oct. 1995), 299-318.

[24] Read, J. M., Eames, K. T. D., and Edmunds, W. J. 2008. Dynamic social networks and the implications for the spread of infectious disease. Journal of The Royal Society Interface, 5 (Sept. 2008), 1001-1007.

[25] Rohani, P., Breban, R., Stallknecht, D. E., and Drake, J. M. 2009. Environmental transmission of low pathogenicity avian influenza viruses and its implications for pathogen invasion. In Proceedings of the National Academy of Sciences, 106:25, 10365-10369.

[26] Salathe, M., Kazandjieva, M., Lee, J. W., Levis, P., Feldman, M. W., and Jones, J. H. 2010. A High-Resolution Human Contact Network for Infectious Disease Transmission. In Proc. of National Academy of Science, 107:51 (Dec. 2010).

[27] Schneeberger, A., Mercer, C. H., Gregson, S. A. J., Ferguson, N. M., Nyamukapa, C. A., Anderson, R. M., et al. 2004. Scale-Free Networks and Sexually Transmitted Diseases: A Description of Observed Patterns of Sexual Contacts in Britain and Zimbabwe. Sexually transmitted diseases, 31:6 (Jun. 2004), 380-387.

[28] Small T. and Hass A. 2005. Resource and Performance trade-offs in delay-tolerant wireless networks. In Proc. of the 2005 ACM SIGCOMM workshop on Delay-Tolerant Networking, (Philadelphia, USA, August 22-26, 2005), 260-267.

[29] Smieszek, T. 2009. A mechanistic model of infection: Why duration and intensity of contacts should be included in models of disease spread. Theoretical Biology and Medical Modelling (Nov. 2009), 6-25.

[30] Song, C., Qu, Z., Blumm, N., and Barabási, A. 2010. Limits of Predictability in Human Mobility. Science 327(5968), 1018–1021, DOI=10.1126/science.1177170.

[31] Stanley, K.G. and Osgood N.D. 2011. The Potential of Sensor-Based Monitoring as a Health Care, Health Promotion, and Research Tool. *Editorial in Annals of Family Medicine*. vol. 9, 296-298.

[32] Tuite, A. R., Greer, A. L., Whelan, M., Winter, A., Lee, B., Yan, P., Wu, J., Moghadas, S., Buckeridge, D., Pourbohloul, B. and Fisman, D. N. 2010. Estimated epidemiologic parameters and morbidity associated with pandemic H1N1 influenza. CMAJ, 182:2 (Dec. 2009), 131–136.

[33] Viboud, C., Bjørnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., and Grenfell, B. T. 2006. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. *Science*, 312:5772 (Apr. 2006) 447-451.

[34] Weber, T. P. and Stilianakis, N. I. 2008. Inactivation of influenza A viruses in the environment and modes of transmission: A critical review. Journal of Infection, 57:5 (Oct. 2008), 361-373.

[35] Wylie, J. L., Shah, L., and Jolly, A. 2007. Incorporating geographic settings into a social network analysis of injection drug use and bloodborne pathogen prevalence. Health & Place, 13:3 (Sept. 2007), 617-628.