FLUNET: Automated Tracking of Contacts During Flu Season

Mohammad S. Hashemian, Kevin G. Stanley, Nathaniel Osgood

FLUNET: Automated Tracking of Contacts During Flu Season

Mohammad S. Hashemian, Kevin G. Stanley, Nathaniel Osgood
Department of Computer Science
University of Saskatchewan
Saskatoon, Canada
m.hashemian@usask.ca, {kstanley, osgood}@cs.usask.ca

Abstract— **By analyzing people's contact patterns over time, it is possible to build efficient delay tolerant networking (DTN) algorithms and derive important data for parameterizing and calibrating epidemiological models. Significant research has been performed in the automated acquisition of contact patterns using mobile devices such as Zigbee motes or Bluetooth-enabled cellular phones. However, the limited number of studies described to date do not capture the breadth of human experience or specifically include the acquisition of health related information. In this paper we present Flunet, a mobile contact-tracking network deployed in a Canadian university environment during flu season. Flunet tracked contact patterns of 36 participants and their proximity to 11 stationary nodes using MicaZ motes over a period of three months. Participants filled out weekly surveys on their state of health. This study is distinct from others because we incorporate health information and the impact of sub-zero temperatures on mobility patterns. This paper presents a preliminary analysis of the data set, primarily from a DTN perspective. We present fundamental attributes of the dataset, the efficiency of routing for single pass and flooding-based algorithms and a preliminary look at the relationship between network characteristics and health status.**

Keywords-component; Sensor Networks, Contact Measurement, Delay Tolerant Network, Epidemiological Modeling

I. INTRODUCTION

Automated contact tracing between human agents has direct applications to health and networking research. In networking research, contact traces can provide detailed maps of contact duration and inter-contact separation, which can be directly applied to the design and validation of delay tolerant network routing algorithms [1]. In health research, contact patterns provide insight into critical factors underlying the spread of contagions, such as for sexually transmitted infections [2] like HIV, and for airborne illnesses such as influenza or tuberculosis [3]. The resulting network structure and dynamics can be used directly in agent-based and network models or can be processed to provide mixing matrices for population level models.

Several attempts have been made to derive contact patterns in human environments. Traditional contact tracing has relied on self-report, which is labor-intensive and time-consuming [4]. Individuals subject to contact tracing are sometimes reluctant to report contacts for reasons of confidentiality. Even given a willingness to offer contact histories, individuals can have limited ability to recall the occurrence and timing of a contact, particularly for those of shorter duration. Partly as a result of these limitations, researchers working with self-report contact data have found that the incorporation of geographic place into social network analysis can offer significant epidemiological insight [3].

Data sets, particularly those publicly available on the CRAWDAD repository, have been obtained from automated contact tracing, either directly using Wi-Fi, Bluetooth or Zigbee devices, or indirectly using time-stamped GPS coordinates. Example datasets include university students and staff [5], conference attendees [6], university wireless usage [7], rollerblade tours [8] and GPS traces at theme parks [9]. Researchers have also attempted to use secondary measures to estimate contact patterns by inferring them from published or recorded schedules. Examples include student attendance in classes based on anonymized schedules [10], vehicular networks such from bus schedules [11], or subway transit records [12].

While these datasets have enhanced our understanding of inter-human contact patterns and allowed the construction of more realistic synthetic algorithms for creating mobility models [9], they still do not capture the breadth of human endeavor, or include medical data which would make the contact records more applicable to epidemiological modeling. To fill this gap we present Flunet, a contact tracing experiment conducted at the University of Saskatchewan.

Flunet tracked 36 participants and 11 fixed nodes over the course of three months during a Canadian winter. Participants provided health information through weekly surveys, which reported symptoms characteristic of influenza-like illnesses. Flunet is distinct from other data sources for the following reasons:

- Medical Data: We collected health data from each of the participants in weekly surveys, allowing us to examine changes in contact dynamics for sick and healthy individuals.
- Duration: We have collected data for an entire flu season, encompassing over 1000 person-days of

This work was partially sponsored by the National Science and Engineering Research Council (NSERC) of Canada, Discovery Grant and RTI programs.

records, which is significantly more than other Zigbee-based experiments.

- Temperature: The mean temperature during our experiment was -12º C [13]. Such low temperatures both contribute to mobility patterns and facilitate the spread of influenza.

This paper presents a preliminary analysis of the dataset from both a DTN routing and epidemiological perspective. Section 2 describes our experimental setup. Section 3A presents a preliminary contact data analysis, section 3B verifies that classic DTN routing algorithms function as expected, and section 3C examines the health information provided in the participant surveys. A brief discussion is provided in section 4, future work in section 5 and conclusions in section 6.

II. EXPERIMENT SETUP

We were interested in the contacts between participants, the amount of time they spend at public places and how this might correlate with health conditions. To collect this data, we used contact tracing using MicaZ motes and weekly health surveys.

A. Contact Measurement

We selected 36 participants and asked them to carry a wireless module (MicaZ by Crossbow®) for 3 months, starting Nov. 9th, 2009 and ending Feb. 9th 2010. The participants were from 7 graduate labs in the Computer Science Department, departmental staff and undergraduate students. We also placed 11 stationary nodes in public places through the campus to measure the amount of time each person spent in specific public areas. Stationary locations were chosen by experimenters in high traffic, public locations.

Out of 11 stationary nodes, 3 were connected to a networked PC and acted as data sinks. The mobile nodes buffered the recorded contact information and offloaded the buffered data accumulated since the last offload time. To collect all the data at a central location, these base stations were connected to a central MySQL database which periodically uploaded collected data from the network to the main database. Contact with a base station triggered clock synchronization by setting the mobile node's internal clock to the value received from the server. When two mobile nodes established a new contact, they synchronize their clocks to the node which had more recently visited the server. This algorithm caused the synchronized time to propagate through the network, mitigating clock drift in mobile nodes with infrequent server contact. Half of the mobile nodes were equipped with MTS310 sensor boards to periodically measure the temperature of their surroundings.

To probe adjacent nodes, each node broadcasts a "HELLO" message every 30 seconds (4-second interval for stationary nodes) with random drift of +/-2 s to prevent packet collision. If a node currently in contact failed to receive 4 consecutive HELLO messages from its partner, it labeled the partner as departed and adds a new contact record to its internal buffer. Each contact record includes: control flags, adjacent node ID, contact start time, contact end time, distance (discretized to 'Close', 'Medium', and 'Far' based on received signal strength indicator (RSSI) value), and temperature (if applicable).

B. Surveys

Health information was collected via weekly surveys, which also requested participants estimate their total contact time with other members of the study to provide data on the difference between automated and self reported data collection. Participants were also asked to fill a one-time demographic/background survey which included questions on demographic variables, average amount of time spent with people and on campus, flu shot and H1N1 vaccination information, and their attitudes to the wireless module.

C. Maintaining the Integrity of the Specifications

The wireless module failures can be divided into two main categories: technical problems and human carelessness. The primary technical failure mode was a failure in the MicaZ receiver amplifier. The node could not receive HELLO packets from other nodes, but other nodes could record the faulty node's existence. We suspect this problem was due to electrostatic discharge (ESD) damage to the receiving amplifier, due to the prototype design of MicaZ motes and the dry, cold conditions which dominate Saskatchewan winters. Thirty-two nodes experienced receiver failure during the study. Defective nodes were replaced after failure detection. We have mitigated the impact of this failure node by defining contact between a pair of nodes as the union of their contact records. Battery failure or displacement, sensor board displacement caused by shaking the module, and storage memory failure were other technical problems which occurred rarely during the study.

Human error was dominated by participants forgetting to carry the module, and forgetting to replace the battery periodically. When the participant forgot to carry the module, either the module was left on or off. In the former case the module recorded additional erroneous data and in the latter case we lost potential contact information. We tried to minimize compliance issues by sending reminders, mentioning in the weekly surveys, and directly reminding the more egregious offenders. During the study we observed that as time progresses, people cared less about the mote, and more often forgot to change the battery or to keep the device on. Voluntarily turning the node off, and forgetting to switch the device on after turning it off are other examples of human carelessness which rarely affected data collection.

D. Simulation

We can consider the dataset as a delay tolerant network (DTN) with semi-predictable connectivity. To analyze data from this point of view, we implemented a DTN in Network Simulator 3 (ns-3) using collected data from all 47 nodes (mobile and stationary). We applied two fundamental DTN routing algorithms on the simulated network: 1) *Direct Message Passing* to measure the upper bound of end-to-end delivery time and the lower bound of energy consumption, and 2) *Epidemic Routing (Flooding)* [14] to measure the

lower bound of delivery time and the upper bound of energy consumption.

In simulation, we discretized the time into 30-second time slots, and computed the connectivity pattern at each time slot using the dataset. The result for each node was imported as its contact pattern during the simulation. For both algorithms, we generated 12,500 packets in the network during the simulation, using a uniform random variable with 0.1% generation rate; therefore each node generated approximately 3 messages per day. *Buffer Size* was set to 1000 packets for all the experiments and no *Packet TTL* was specified. The packet size was fixed to 29 bytes (the default packet size in TOS). The Flooding algorithm did not employ any anti-packet methods [15]. The carrier drops the packet as soon as delivery to the destination, but might receive the packet again in the future from other nodes.

III. RESULTS

Participants in this study consist of 75% males and 25% females, while 47% of them were in contact with other participants just on campus, and 53% were in contact off campus as well. 14% of the participants received a regular flu shot, and 39% of them had H1N1 vaccination. In addition, 83% of participants never smoke, while 6% smoke occasionally, and 11% smoke every day.

A. Dataset Characteristics

For the first set of results from the collected dataset, we focused on the data and its characteristics. The average duration of contacts at different hours of the day is interesting from DTN perspective. As shown in Fig. 1, contacts which happened in the morning between 7 AM to 9 AM are longer than other times of the day and the contact durations get shorter during afternoon. This can be explained by considering the effect of contacts between staff members during working hours at their offices, while the shorter contacts happen in afternoon between graduate students. During midnight to 6 AM, the number of contacts is much smaller and at shorter duration.

Fig. 2 shows the complementary cumulative distribution function (CCDF) of contact duration. In addition to CCDF for all the collected data, we also removed contact durations more than 10 hours (0.03% of total reported contacts) from dataset because we assumed contact of this duration was due to motes abandoned near each other. Removing this section of data shows a considerable difference in the result plot. Actual data of this duration would have required that participants use the washroom together. Although the graph shows a small fluctuation between 300 and 400 minutes, it is difficult to separate the cause of the departure from the curve, because while it is likely due to abandoned nodes behaving like stationary nodes, it could also be caused by people changing their behavior when in continuous proximity.

Fig. 3 shows contact duration and number of reported contacts for all mobile nodes during the experiment. This graph is plotted for 'Close', 'Medium', and 'Total' reported contacts, based on the RSSI value. All of the contact proximities show the same trend in data: there are fewer nodes which have long contact duration and a large number

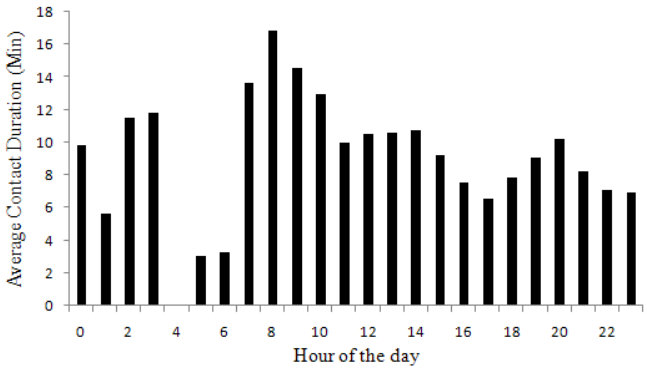

Figure 1. Average contact duration at different hours of the day

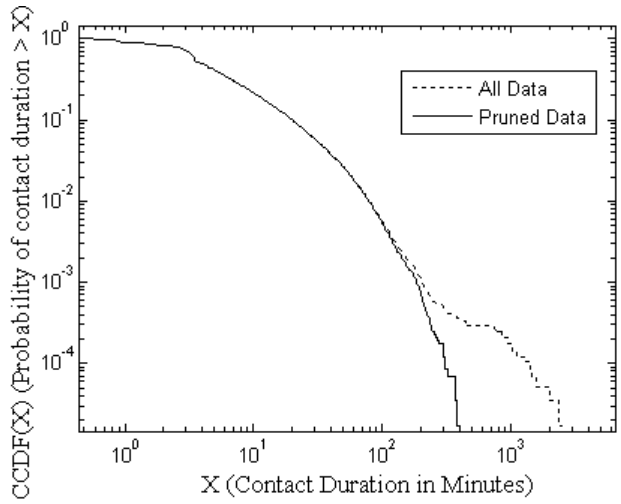

Figure 2. Complementary Cumulative Distribution Function (CCDF) of contact duration

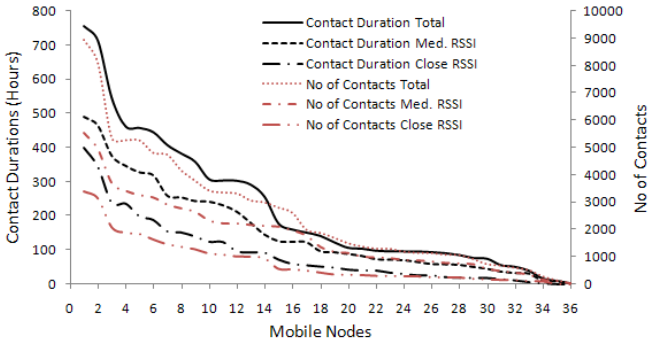

Figure 3. 'Close', 'Medium', and 'Total' contact duration (Left) and reported contacts (right) for mobile nodes

of reported contacts, while there are many nodes with small to medium contact duration and fewer contact records.

Fig. 4 shows duration of reported contacts from two selected participants with all other nodes in network, sorted by duration of contact. Out of 36 participants, 3 of the plots were similar to 'Sample 2', which is indicative of people with high centrality. Even though those people do not have extended contact durations with other people, they have seen all other participants at least once during the study. The plot for the other 33 people was similar to 'Sample 1' which shows a power law behavior in their contact pattern [16].

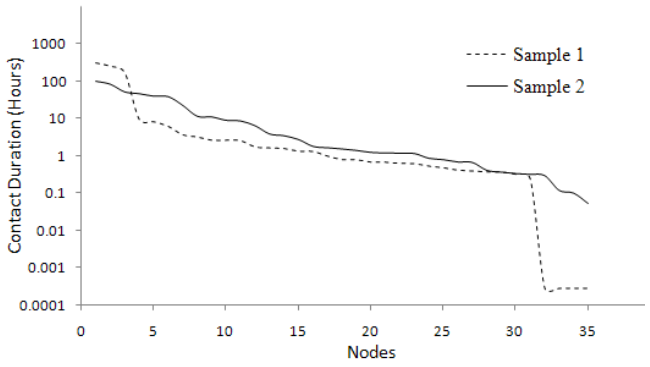

Figure 4. Contact duration between two sample participants and other nodes during the experiment

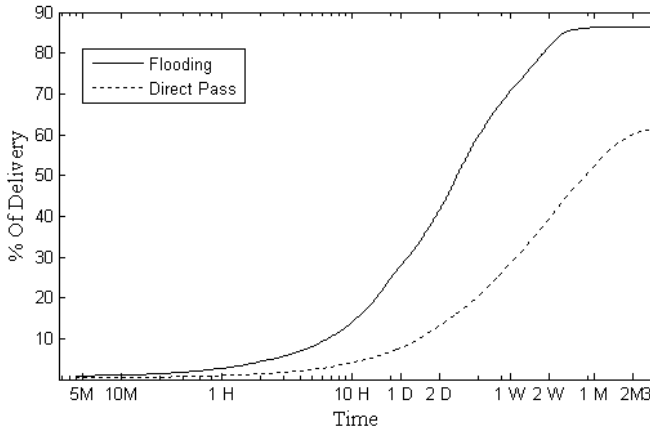

Figure 5. Packet delivery ratio for Flooding and Direct Pass algorithms

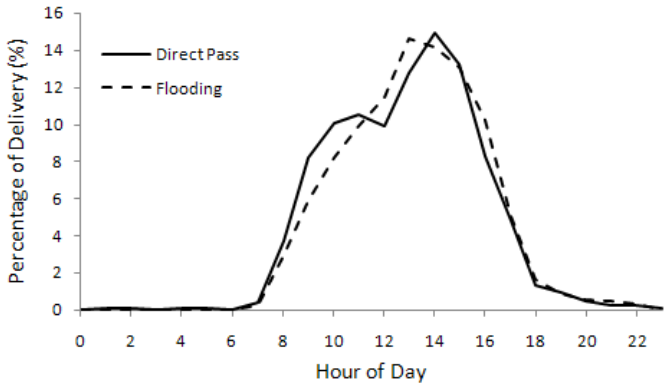

Figure 6. Delivery ratio at different hours of day

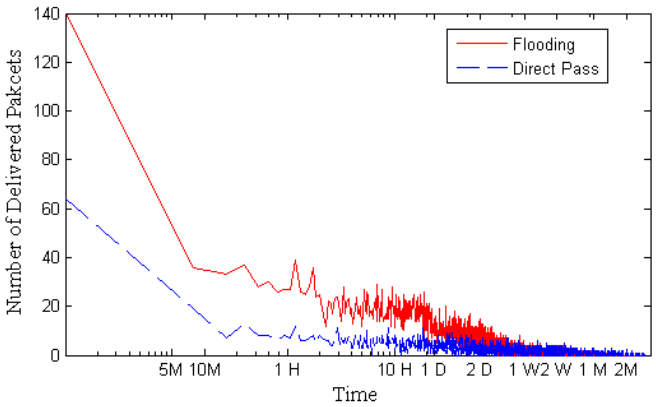

Figure 7. End-to-End delivery time histogram for Direct Pass and Flooding algorithms

This is reasonable specifically for people who worked exclusively in a lab, as they spent a considerable amount of time close to their lab mates, but they did not have contacts with as many other people in the study.

B. DTN Results

The collected data can represent a DTN network with unpredictable and opportunistic connectivity [1]. We used collected data for all 47 nodes to simulate a DTN network and implemented two basic routing algorithms: *Direct Pass*, and *Flooding*. Fig. 5 shows packet delivery ratio for both algorithms, with infinite TTL and buffer size equal to 1000 packets. With this buffer capacity, Direct Pass does not face buffer overflow, while in Flooding each node faces buffer overflow during the simulation. The buffer overflow ratio in flooding depends on the popularity of the node in the network. If the node has large number of contacts, it receives more packets and fills its buffer faster, and therefore potentially enters buffer overflow. If the node has fewer contacts, it receives fewer packets, and faces limited buffer overflow.

As it can be seen in Fig. 5, delivery ratio in Direct Pass constantly increases with time, while delivery ratio in Flooding does not have a notable change after two weeks, where it reaches to a saturation point at less than 100% delivery. This non-ideal delivery is partly due to the packets which are generated in the nodes near the end of simulation, when insufficient time remains for delivery, and partially due to dropped packets caused by buffer overflow.

Because the dataset represents a human environment, we suspected that most of the message deliveries happened during the day. Fig. 6 validates this hypothesis. As shown in this figure, the number of messages delivered between Midnight and 6 AM is negligible. Packet delivery increases at 7 AM and reaches its peak at 2 PM, and it decreases during afternoon and evening. This peak can change in different human environments. In this study, majority of participants were students with their own work schedules, and the minority were staff with fixed working hours. However, these schedules tended to overlap in the afternoon, increasing the delivery probability.

Fig. 7 shows histogram for end-to-end delivery time for the flooding and direct pass algorithms, discretized using 8 minute bins. As the histogram shows, a portion of generated packets can be delivered quickly, likely because the source and destination are in close proximity (e.g. the same laboratory) when the packet is generated, or one of the nodes is highly connected. The chance of delivering the packet decreases with time, and in Flooding histogram it approaches zero after two weeks. The number of delivered packets at each bin of the Direct Pass plot is smaller than its Flooding

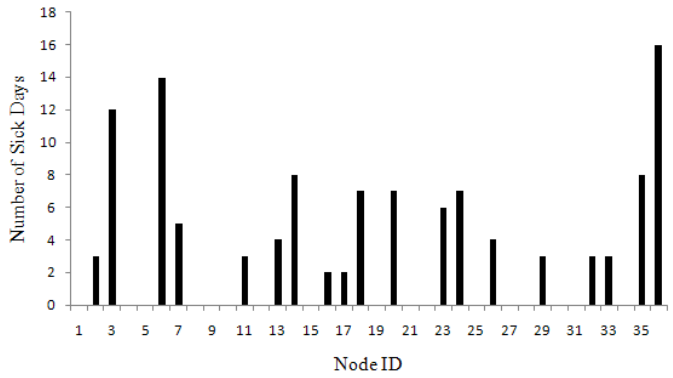

Figure 8. Total sick days for each participant

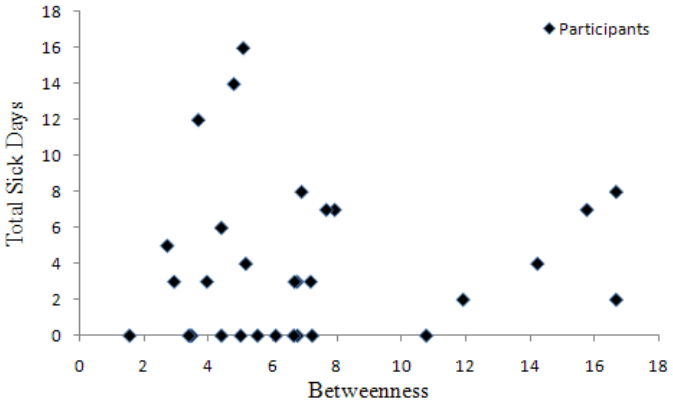

Figure 9. Betweenness vs. Total sick days

counterpart, but falls more gradually leaving a non-zero chance of delivery even after 2 weeks. This difference is as expected, because Flooding provides faster throughput than Direct Pass, but at the risk of buffer overflow.

C. *Dataset and Health*

We recorded information on participants' health during the study using weekly surveys. Compliance with the survey was sporadic, and we were forced to remove 5 participants due to insufficient data. Fig. 8 shows total number of sick days reported by each participant during the study. This data can be used to analyze the relation between peoples' contact patterns and their likelihood of being ill. We felt that participants were logically grouped into those who did not report a sick day, those that did reported less than 5 sick days, and those with greater than 5 sick days. While we recognize that processes of viral infection in humans are incompletely captured with aggregate statistics, we sought to determine if there was any correlation between the report of illness and participants' contact patterns in an attempt to validate the automated approach and to inform both future analysis of this data and the collection of new data. Descriptive statistics are shown in Table I.

While there were differences in contact duration for those that reported less than 5 days of illness and those with 5 or more days, there are no apparent differences between those who reported more than 5 sick days and those who reported none. While this may indicate some kind of correlation in our

TABLE I. DESCRIPTIVE STATISTICS ABOUT PARTICIPANTS WITH DIFFERENT SICK DAYS

	Never	***< 5 Days***	***≥ 5 Days***
No. of People	12	9	10
Mean Duration – Close (Hour)	1.42	1.08	1.42
Mean Duration – Any (Hour)	3.23	2.52	3.19
Mean Contacts	36.37	29.18	32.2
Male/Female	75%/25%	89%/11%	60%/40%
Flushot	0%	11%	10%
Mean Age	29.8	28.7	31.2
Std Dev Age	7.3	5.6	10.9

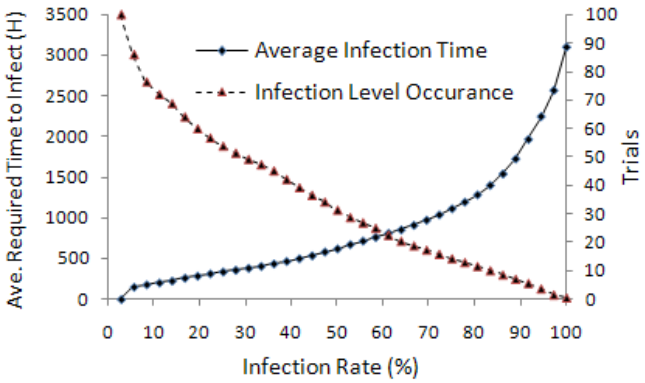

Figure 10: Time required to infect a given percentage of the population (left) and percentage of trials where at least the given percentage of the population was infected (right)

dataset, further research will be required to tease out the underlying causes, or eliminate the correlation. We suspect that some participants who reported no sick days did not fill out the surveys properly, underreporting their sickness causing a misclassification. However, these human factor issues can be difficult to isolate in data, and may require additional experiments to resolve. The trends noted are also influenced by important confounders, as older participants were more strongly represented in the more than 5 days category, and older participants in the study tended to be office staff, with consistent and prolonged contact patterns rather than students with more sporadic schedules.

To further visualize if sick days was correlated to network structure we plotted betweenness versus reported sick days as a scatter plot of participants in Fig. 9. Like the table, the results are ambiguous. More detailed epidemiological study will be required to further investigate this finding to determine if it is an effect or artifact.

We wanted to make a preliminary investigation of the utility of the dataset for agent-based epidemiological modeling. We assumed the virus spread from an infected node to an adjacent node, if two nodes stay in contact for at least 5 minutes with a "Close" categorization of RSSI value. This represents a 100% infection rate if the conditions are met, characteristic of a worst-case virulence human pathogen or a novel computer virus [17]. A packet representing the

virus was introduced into the system, and its propagation monitored. Fig. 10 shows the result of 9500 trials at different simulation times. Some trials infected the whole community, but many infected a small fraction, if the primary infection point was poorly connected or the packet was generated near the end of simulation. The time required to infect a given percentage of the population is shown on the left axis, and the number of trials where that percentage was infected is shown on the right axis.

IV. Discussion

We have presented a new dataset incorporating automated contact monitoring using MicaZ motes and health survey data. Our contact pattern measures and DTN performance are similar to those in previously reported analyses [5]. We had expected to see some distinctions between the contact patterns we observed, and the contact patterns observed in other studies due to the inclement weather typical of winters in Saskatchewan. However, our results are similar to those in studies like Reality Mining [5]. We expect that this similarity is due to work habits of North American Computer Science and Engineering graduate students, who spend the vast majority of their walking hours in indoor laboratory environments regardless of the weather. However, we are hopeful that the weather will aid in future analysis of the role of the stationary nodes because the winter weather strongly motivates students to utilize indoor routes between buildings, passing several nodes in public areas.

Although our preliminary analysis of the combined contact and medical data hinted at correlations, the simple analysis tools we employed were unable to unearth any conclusive findings. However, we have identified potential interactions, meriting additional research. The analysis of the data, and of confounders such as age, employment status and location caused us to note several potential pitfalls of directly mapping contact patterns to health data.

V. Future Work

This paper reports a preliminary analysis of our dataset, which itself results from a pilot exploration of the collection of simultaneous health and contact data. Our future work will focus on three areas: employing the data we have collected for the development and evaluation of DTN routing algorithms, analyzing our data from an epidemiological perspective to gain insight into the potential for automated data collection to contribute to infectious disease modeling, and as a starting point for longer and more detailed future studies employing more compelling, robust and comprehensive wireless sensor systems.

VI. Conclusion

We have presented the preliminary analysis of a new study conducted at the University of Saskatchewan covering 36 participants over a single flu season. Our data offers information both on contact patterns and analysis of participant health. We have found the general character of the data obtained in good agreement with similar studies from the point of view of the contact patterns themselves and the behavior of classic DTN routing algorithms using the aforementioned contact data. Our preliminary analysis of the health data suggests possible associations between contact rates and risk of infection, but significant additional work will be required to validate this finding with statistical rigor. We plan to follow this research with more analysis, new algorithms and additional studies.

References

[1] S. Jain, K. Fall, and R. Patra. Routing in a delay tolerant network. In *ACM SIGCOMM*, 2004.

[2] A.M. Jolly, S.Q. Muth, J.L. Wylie, and J.J. Potterat. Sexual Networks and Sexually Transmitted Infections: A Tale of Two Cities. Journal of Urban Health:Bulletin of NY Academy of Medicine 2001;78:433-445.

[3] A. Al-Azem, Social network analysis in tuberculosis control among the Aboriginal population of Manitoba [Thesis (PhD)]: University of Manitoba, Fall 2006; 2006.

[4] M. Morris, International Union for the Scientific Study of Population. Network epidemiology: a handbook for survey design and data collection. Oxford; New York: Oxford University Press; 2004.

[5] N. Eagle and A. Pentland, Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing V10*,May 2006, 255–268.

[6] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pockets Switched Networks and Human Mobility in Conference Environments. Proc. of ACM SIGCOMM'05, p 244–251, Aug. 2005.

[7] D. Kotz and K. Essien. Analysis of a Campus-wide Wireless Network. In *Proceedings of the Eighth Annual International Conference on Mobile Computing and Networking*, pages107–118, September 2002.

[8] P.-U. Tournoux, J. Leguay, F. Benbadis, V. Conan, M. D. de Amorim, and J. Whitbeck. The accordion phenomenon: Analysis, characterization, and impact on dtn routing. *Proc. IEEE Infocom*, 2009

[9] K. Lee, S. Hong, S. Kim, I. Rhee, and S. Chong. Slaw: A mobility model for human walks. In IEEE INFOCOM 2009, 2009.

[10] V. Srinivasan, M. Motani, and W. T. Ooi. Analysis and Implications of Student Contact Patterns Derived from Campus Schedules. In Proc. of ACM MobiCom, 2006.

[11] J. LeBrun and C. Chuah. Bluetooth Content Distribution Stations on Public Transit. In MobiShare '06: Proceedings of the 1st Workshop on Decentralized Resource Sharing in Mobile Computing and Networking, pages 63-65, NY, USA, 2006. ACM.

[12] L. Mcnamara, C. Mascolo, and L. Capra, Media sharing based on collocation prediction in urban transport, Proc. of ACM MobiCom 2008

[13] http://saskatoon.weatherstats.ca/charts/temperature-1year.html

[14] A. Vahdat, D. Becker, "Epidemic Routing for Partially-Connected Ad Hoc Networks", Duke Tech Report CS-2000-06, 2000

[15] T. Small and Z. Haas. Resource and Performance Tradeoffs in Delay-Tolerant Wireless Networks. Proc. ACM WDTN, P.260–267, Aug. 2005

[16] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Pocket Switched Networks: Real-World Mobility and its Consequences for Opportunistic Forwarding. Technical Report UCAM-CL-TR-617, University of Cambridge, Computer Laboratory, February 2005

[17] J. Su, K.K.W. Chan, A.G. Miklas, K. Po, A. Akhavan, S. Saroiu, E. de Lara, and A. Goel, A preliminary investigation of worm infections in a bluetooth environment. In: 4th Workshop of Recurring Malcode (WORM), Fairfax, VA, USA, 2006.