iEpi: An End to End Solution for Collecting, Conditioning and Utilizing Epidemiologically Relevant Data

Mohammad Hashemian[1], Dylan Knowles[1], Jonathan Calver[1], Weicheng Qian[1], Michael C. Bullock[1], Scott Bell[2], Regan L. Mandryk[1], Nathaniel D. Osgood[1,3], Kevin G. Stanley[1]
Dept. of Computer Science[1], Dept. of Geography and Planning[2], School of Public Health and Epidemiology[3]
University of Saskatchewan
Saskatoon, Canada

firstname.lastname@usask.ca

ABSTRACT
Smartphones have the potential to revolutionize health monitoring and delivery. Significant attention has been given to personal health devices and systems to help individuals and medical practitioners monitor health and treatment compliance. The data collected from these systems also has significant value to public health workers and epidemiologists. However, requirements for backend analysis and supplemental data differ between personal and public health applications. In this paper we describe iEpi, an end-to-end system for collecting, analyzing, and utilizing contextual microdata through smartphones for epidemiological and public health applications.

Categories and Subject Descriptors
J.3 LIFE AND MEDICAL SCIENCES

General Terms
Measurement, Design, Experimentation.

Keywords
Smartphone data collection; Epidemiology; Simulation.

1. INTRODUCTION
The advent of affordable smartphones has created a sufficiently sophisticated platform for the acquisition of individual medical data. Significant attention has been focused on the utilization of this information for personal health or direct medical telemetry. These applications seek to provide users with greater control over their health or existing conditions and provide medical practitioners with detailed information on patient status. Such monitoring tasks are intended to provide disaggregate information and have focused on extending the number of federated devices through middleware, signal analysis, and visualization interfaces for individuals or professionals.

A largely unexploited side effect of the advent of medically-motivated smartphone-based monitoring is the utility of aggregate data. The contextual and health information collected constitute a unique high-fidelity window on population health. Contact patterns between and among individuals can provide insight into the spread of contagious disease social norms and behavioral risk factors. Population level activity metrics, when combined with location data, can provide city-planners with an unprecedented view of the effect of the built environment on health behaviors. Coupled with environmental sensors (e.g. air quality monitors) or Geographic Information Systems (GIS), such location data can be used to highlight exposure patterns to environmental contaminants such as pollutants. While the front-end of the system for data collection and aggregation is similar to the individual health context, the back-end analysis requirements and supplemental data needs are substantially different. While initial research into the collection [1], analysis [2], or simulation [3] has been reported, previous research has not provided a coherent, consistent, and self-contained system for epidemiologists and public health researchers.

To satisfy this need, we present iEpi, an end-to-end system for the collection, visualization, and simulation of epidemiologically-relevant micro-contextual data. Although other systems exist [4, 5, 6], iEpi is distinguished by its focus on epidemiological data, its use of localization, contextual and on-device survey data, and its seamless integration with agent-based dynamic disease models. As a platform, iEpi provides a valuable tool for epidemiologists, public health workers, and facility managers to collect, visualize, and analyze micro-context data.

2. THE iEpi SYSTEM
Unlike body sensor network systems which focus on federating medically relevant sensors for use by a single individual, iEpi integrates across participants and through time to provide population-level insight. iEpi consists of several components: smartphone-based modules for data collection, server-side architectures for opportunistically collecting and recording data, aggregate post-processing capabilities, and procedures for linking the collected data to agent-based simulations. Currently, iEpi is primarily focused on detecting the frequency, duration, and proximity of human-human contact, as well as the location and activity level of participants. These parameters drive contagious disease and impact social determinants of health

MobileHealth'12, June 11, 2012, Hilton Head, South Carolina, USA.

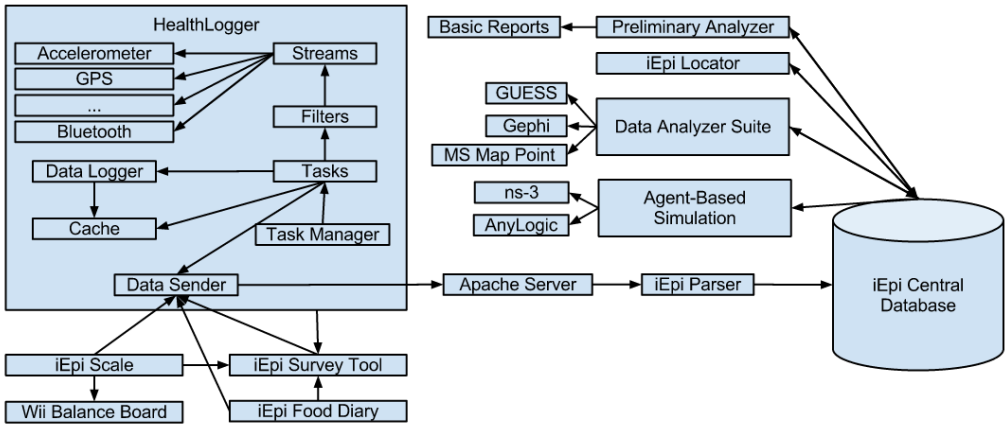

Figure 1: iEpi system architecture

[7]. The architecture of iEpi is depicted in Figure 1, and detailed description of each component follows.

2.1 Data Collection

The first iEpi component focuses on data collection. The data collection system is based on a set of Android programs and peripherals that permit the collection of sensor, context, and survey data from participants.

2.1.1 HealthLogger

The primary data collection component of iEpi is the autonomous data collection system, *HealthLogger*. HealthLogger is a background service that runs on Android phones to collect data from sensors at specified intervals. It is composed of five major modules:

1. Tasks & Task Manager: HealthLogger is a system of tasks which work in parallel to carry out specific operations such as updating the system, transmitting data, or reading from sensors. Tasks are scheduled for continuous or periodic work. The former repeats the assigned operation continuously, while periodic tasks are repeated every *Duty Cycle* for the period of *Burst Length*. The task manager also provides a central point of access to the many services used by tasks.

2. Streams & Filters: Streams provide a standardized interface to the various Android sensor APIs and other data sources. Streams can be simple or composite. Simple streams read from Android APIs and return data in a standard format; composite streams obtain data from other streams, process the data, and produce new data. The behavior of streams can be altered using filters to exclude or manipulate data. For instance, a MAC filter can exclude Bluetooth data from non-participant devices.

3. Data Logger & Data Cache: The data logger provides a central hub for data recording collected by HealthLogger's tasks. The logger stores recently collected data in an on-device database, which can then be processed. The data logger works in tandem with the data cache. The cache is used to add "memory" to HealthLogger by providing access to the most recent data. Other components on the phone that require real-time decisions – such as triggers for issuing context-relevant surveys – can access the cached data. For privacy reasons, all data stored in the databases are encrypted; however, cached data is not encrypted due to its short-lived nature.

4. Data Sender: The data sender acts as a general-purpose file uploader. It is used by HealthLogger and other iEpi components (as needed) to transmit data to the server.

5. "iEpian" (iEpi Configuration Language): Different behaviors of the HealthLogger can be configured through configuration files, which consist of commands given in an intuitive language, supporting iEpi's use by researchers with limited programming experience. The configuration file specifies the required sensors to be sampled (via Tasks) and the frequency of such sampling, using terms such as:

```
sample gps continuously

sample bluetooth_proximity every 5 minutes
for 10 seconds
```

Configuration files also define the behavior of HealthLogger's control tasks, such as where and how often to upload the data. Reflecting the need to adjust data collection and survey delivery in the course of a study, HealthLogger is designed to allow the configuration file updates on the fly.

2.1.2 iEpiScale

HealthLogger can be used to collect detailed data salient for public health, but many variables cannot be accurately measured or inferred from smartphone data alone. For instance, the growing obesity epidemic is a foremost public health concern, and individual's weight trajectories are of considerable interest. *iEpiScale* provides a simple Bluetooth-mediated approach for federating a scale (implemented using a Wii Balance Board) with HealthLogger on-device telemetry. The recorded weights are time-stamped and uploaded using HealthLogger's Data Sender. After each use, participants can optionally receive a short survey via iEpi's Survey Generator. The answers can represent the individual's perspective regarding their

recorded weight and possible future changes in their behavior in order to achieve a weight goal.

2.1.3 Food Diary

In addition to physical activity, weight change, and weight measurement, individuals' daily eating habits and participation in diet programs also shape their risk of obesity in important ways. The second interactive component of iEpi is responsible for collecting eating habits and diet reports. Such reports provide information on the motivation for and frequency of food consumption, and the nature of the food consumed. *PhotoFoodDiary* consists of two major components: a smartphone application for collecting and labeling food images; and a web-based interface for annotating and recording collected data. The smart-phone application allows people to take pictures of their food, answer short relevant surveys, and append notes. The pictures are time-stamped and location and can be browsed and tagged later by the user, either on the phone or on a personal computer via the website.

2.1.4 Contextual Surveys

While data automatically collected through different components of iEpi provides considerable insight on human behavior, participants' responses and their opinions in specific situations can be valuable in contextualizing the data collected by HealthLogger, iEpiScale, and iEpi Food Diary. The iEpi Survey Tool supports such functionality. iEpi Survey Tool is an Android application that receives survey requests from different components of iEpi and issues a survey to the user based on an XML file specifying survey content and structure. Question types include multiple choice, freeform text, Likert scale, and yes/no queries. Based on responses to questions, branching, varying the text prompts, and conditionally looping questions are supported.

2.2 Server Collection and Repository

The data collected by different components of iEpi is stored as encrypted compressed ASCII files in a buffer and is periodically uploaded to the server by the HealthLogger Data Sender. An Apache web server receives all files and places them in a server directory, where they are queued to be parsed and pushed into the central database for future analysis. A Java parser periodically scans the "new file" directory, decrypts and parses each file, commits to the database, and archives successfully parsed files. The central database uses an MS SQL Server instance, which contains indexed data tables for each of the expected data, either from sensors, scales, or survey responses. The database periodically receives data from the parser and makes them available for immediate analysis. To improve compliance and bug detection in the system, a separate tool called the *Compliance Report Generator* monitors the database, plots the percentage of time each phone was actively collecting data, and makes such reports available to the experimenters for review.

2.3 Post-processing

The central database receives and stores the data from different iEpi components. Such raw data can be analyzed and interpreted for a wide range of purposes, such as wireless routing [8], or health modeling [2].

2.3.1 SQL Queries

By default, HealthLogger is configured to collect sensor readings at each duty cycle for a specific burst length. In the first post-processing step, the data are aggregated by duty cycle and participant, and inserted into a separate table. Each record of this aggregated table represents the state of one phone during one duty cycle. Based on the aggregated data table, sets of pre-configured SQL commands can be used to generate aggregate measures of relevant parameters such as compliance, activity (from accelerometer readings), and contact counts. Descriptive statistics – such as the distributions of contacts or contact durations – can similarly be generated directly from preconfigured SQL queries.

2.3.2 Localization

Although GPS provides the coordinates of a participant's location, it requires line-of-sight to GPS satellites to be effective. This reduces indoor reliability, which is further amplified in large institutional settings such as universities and hospitals due to the predominance of indoor spaces. To address this lack of indoor localization, we developed *iEpi-Locator* based on the Sask Enhanced Positioning System (SaskEPS) [9], which employs receive signal strength (RSSI) measurements, calibrated access point (AP) locations, and trilateration to estimate the locations of the participants while indoors during post-hoc analysis (SaskEPS performs with GPS-like accuracy in the study area used here).

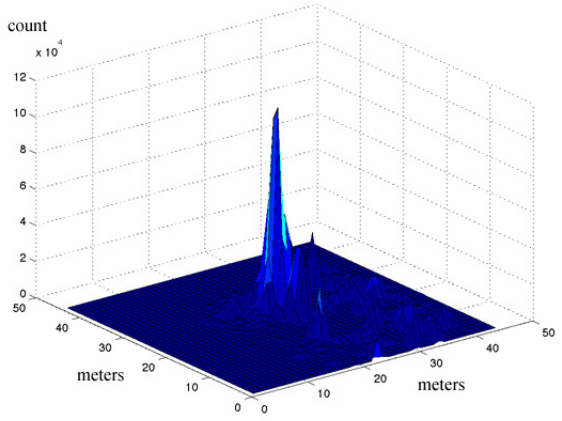

Figure 2: Surface plot of indoor localization

The RSSI value from all WiFi records for a given participant in a given duty cycle are converted to meters and organized into a tuple by the signal's origin AP (AP, distance). Each set of unique three-tuples are trilaterated to derive a location. The calculated location estimates are averaged into one position as the location of the person for that duty cycle.

The output of iEpi-Locator is subsequently linked to a Matlab script that produces visualizations of the localization. These include heatmaps and surface plots of participant locations throughout the experiment, which can be used to identify areas which might harbor environmental contagions or for further analysis in site planning. Figure 2 shows a surface plot of the localization of one participant: the horizontal axes indicate the normalized location in meters while the height of the vertical axis indicates the number of readings.

2.4 Data Analysis Suite

iEpi data can be used to represent patterns relevant to epidemiology or public health. Researchers in the social and health sciences are unlikely to have sophisticated computer skills. The Data Analysis Suite is designed to allow researchers to access iEpi data through a simpler interface and automate common analysis tasks. This .NET software facilitates data extraction from the iEpi database and helps create network diagrams, aggregate graphical representations, and maps. The suite can interact with external components, such as MapPoint and zedgraph.

The Data Analysis Suite's basic action is to populate tables using SQL queries. By exposing the syntax of the query to the user we ensure that experienced users are able to create custom queries, while novice users can employ built-in parameterized queries. Also, to facilitate collaboration, the suite allows experienced users to create and share queries with others. Users can also include metadata on the query and its output.

Through the automation of multiple queries and table joins, adjacency information and properties for nodes and edges can be rendered as GDF files to be linked to Gephi [10] or GUESS [11] for network visualization. GUESS provides the user with an interface to study network structure. The toolkit employs zedgraph to synchronize axes between graphs and facilitate drag and drop interaction. The interfaces allow user to select columns from a table resulting from a query and plot line, scatter, bar, and pie charts. To facilitate geographical plotting, the suite is integrated with MS MapPoint. Users can specify which columns of a table represent Eastings and Northings (in UTM coordinates) together with specific properties of the location to generate MapPoint GPX file.

2.5 Integration with Simulation

The iEpi system can record minute-resolution data from each individual. This data can be used in dynamic models to simulate the evolution of a system based on data under different conditions and study the intervention impact.

Two model classes are commonly employed in health simulation: Aggregate and agent-based models. Aggregate models – such as classic SIR – capture disease dynamics through stocks and flows. It is straightforward to employ the aggregate measures described in the previous section to more accurately estimate infection likelihoods and mixing matrices. However, the real advantage of contact micro-dynamics is realized in agent-based simulations.

While population level simulation tracks states using counts of people, agent-based simulation tracks individuals and models their internal states. Agent-based models are generally more faithful representations of the actual process of disease, but can be computationally expensive for large populations. Agent-based models commonly use a graph representing the probability of infection between any two nodes. To leverage the capacity of dynamic networks we employ a "Groundhog Day" methodology [2] (after the Bill Murray film) where Monte Carlo ensembles are simulated over the same temporal sequence of contact networks, capturing the variability due to disease parameters and interventions.

We have successfully integrated two simulation systems into the overall iEpi architecture, one for Computer Scientists and one oriented towards public health workers. Our first system employs ns-3, an open-source network simulation system, normally intended for communications research. Disease parameters are encoded using C++. Connection graphs are generated as flat files using SQL queries read by ns-3 during execution. The result is a binary that executes quickly and is easily batched for Monte Carlo execution on computing clusters. However, using this system requires a sophisticated computing knowledge uncommon amongst public health workers.

To support integration with mainstream epidemiological tools, we have also provided integration with the popular commercial software package AnyLogic. This system interfaces directly with the database and performs prioritized prefetching of relevant contact parameters, which can be seamlessly integrated with standard agent-based models created using AnyLogic's graphical programing language. By hiding the database structure and contact dynamics from the user, they can interact with the system as if they were simulating the disease with a standard static infection probability graphs.

3. RESULTS

We have deployed iEpi in two internal pilot studies to evaluate the technology and utility of the captured data. The first, SHED1 [2], was conducted over 5 weeks in April and May of 2011, but only used the on-board smartphone sensors, post-processing, and ns-3 simulation system. We are currently collecting the SHED2 dataset which employs

all of the capabilities of iEpi described here and is slated to run from February to May of 2012. In this paper, we present preliminary results from the first month of the SHED2 dataset. iEpi is also being evaluated in pilot deployments at partner institutions.

3.1 Examples from SHED2

While excerpts from the SHED1 database have already been published [2], SHED2 has yet to be disclosed in any form. Illustrative results meant to demonstrate the utility of iEpi are provided here. At the time of writing, the SHED2 data collection exercise has been running for 34 days with 38 participants. Figure 3 provides participants' compliance during first 34 days of the experiment.

Bluetooth contact data can be aggregated and fed into a population level simulation or the underlying contact patterns directly fed into a Groundhog Day simulation [2]. We used data collected during first 34 days of SHED2 in a flu-like disease simulation similar to [2]. The model was implemented in ns-3 and simulated 50,000 times. Figure 4 shows the sorted number of endogenous infections observed per participant over all realizations.

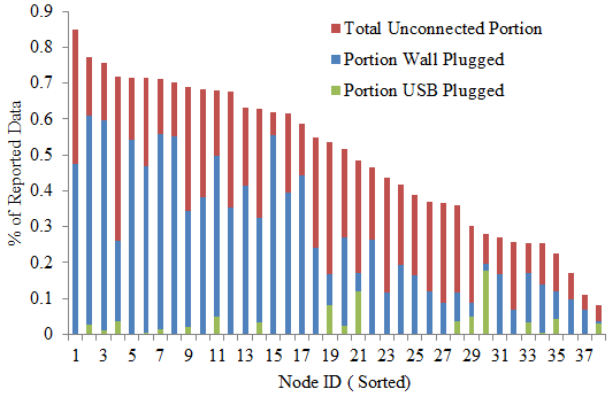

Figure 3: SHED2 compliance rate analysis

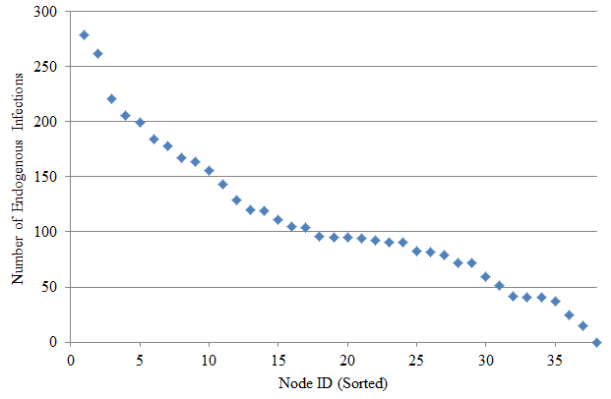

Figure 4: Endogenous infections in SHED2 flu simulation

4. DISCUSSION

While the full import of securing access to epidemiological information collected by iEpi remains to be seen, it affords several clear opportunities. The growing use of rich individual-level data in techniques such as Social Network Analysis [12] and Agent-based modeling elevate the importance of social context. The recognized utility of complementing social networks with place [13] and the centrality of observed patterns in informing the design of Agent-Based simulation models – has elevated the need for rich, individual-level longitudinal data. We anticipate that Social Network Analysis and Agent-Based modeling will benefit from iEpi's data collection and analysis suites.

While iEpi currently delivers value, the system exhibits some important limitations. Power constraints impose a stiff upper limit on the volume of data that can be collected and achievable temporal resolution. iEpi has yet to experience the challenges associated with deployments with hundreds or thousands of users. Handling the high load on the server for such studies may require server farms along with optimized and scalable databases. Important challenges with ensuring informed consent must also be addressed. While iEpi allows for specification of sensing and survey parameters in the course of a survey, contextual triggers for surveys are currently "hard-coded" and are not amenable to modification during a study. Finally, iEpi is primarily suitable for probing discrete populations such as workplaces or schools, due to the fidelity of the Bluetooth proximity measures, essentially introducing population bias to gain spatial contact resolution.

5. RELATED WORK

Recording and analyzing human behavioral patterns is an important component of many scientific disciplines. In mathematical epidemiology, human infection transmission models rely heavily on population heterogeneity and network structure to predict outbreak emergence and progression [5, 14]. Traditionally, surveys have served as the primary means of collecting human behavior data [15].

Different approaches have leveraged pervasive technologies to collect a wide range of behavioral data. Isella et al. [4] used RFID to record close encounters between patients in a hospital's pediatric ward. Similar technology was used by Cattuto et al. [16] to record minute by minute contacts between individuals at different social gatherings. Similarly, various other sensor modules have been used to collect contact proximity and rough location data [17, 5]. Because both RFIDs and sensor modules offer little value outside the purposes of the study, participants have little intrinsic motivation to carry such devices. By contrast, smartphones couple location and proximity sensors in the phone together with rich additional user-oriented functionalities. Several applications have been designed to support periodical data collection from available sensors on smartphones [18, 6].

Some automated data collection projects have applied the resulting data to model pathogen transmission [17], or to study the underlying network structure in communities and its role in communicable disease outbreaks [5]. Other work

has taken advantage of the available processing power of smartphones to provide real-time data processing, such as analyzing sensor input about conditions in the local environment and adapting accordingly [19], or providing participants with feedback that can be used in decision making [20]. Other work has also looked at using cellular call records to infer human proximity, which provides a much larger population, but sacrifices accuracy [21].

6. CONCLUSION

In this paper we have presented iEpi, a novel end-to-end solution for the gathering, analysis, and presentation of epidemiologically relevant data using a combination of smartphone data collection, networking infrastructure, and integrated existing and custom analysis suites. The primary contribution of iEpi is its focus on integrating and utilizing automatically collected data for public rather than individual health purposes. In the future, we plan to deploy iEpi in conjunction with public health partners to examine a number of questions related to the study of contagious, chronic, and social maladies, while continuing to expand the scope, robustness and scalability of the overall system.

7. ACKNOWLEDGMENTS

We would like to acknowledge NSERC, the GRAND-NCE, and GEOIDE-NCE for funding, and HPC Training Facilities for computing time.

8. REFERENCES

1. Salathé, M., Kazandjieva, M., Lee, J. W., et. al. 2010. A high-resolution human contact network for infectious disease transmission. In Proceedings of the National Academy of Sciences (Dec. 2010).
2. Hashemian, M. S., Stanley, K. G., Knowles, D. L., et. al. 2012. Human network data collection in the wild: the epidemiological utility of micro-contact and location data. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (Jan. 2012), Florida, USA.
3. Hashemian, M.S. and Stanley, K.G., and Osgood, N. D. 2012. Leveraging H1N1 Infection Transmission Modeling with Proximity Sensor Microdata. To appear in BMC Medical Informatics and Decision Making.
4. Isella, L., Romano, M., Barrat, A., et al. 2011. Close Encounters in a Pediatric Ward: Measuring Face-to-Face Proximity and Mixing Patterns with Wearable Sensors. PLoS ONE 6, 2 (Feb. 2011).
5. Hethcote, H. W., and Yorke, J. A. 1984. Gonorrhea transmission dynamics and control. Springer Lecture Notes in Biomathematics, 5. Berlin, Springer.
6. Mun, M., Reddy, S., Shilton, K., et. al. 2009. PEIR, the personal environmental impact report, as a platform for participatory sensing systems research. In Proceedings of the 7th international conference on Mobile systems, applications, and services (Jun. 2009), Kraków, Poland.
7. Wilkinson, R. and Marmot, M. 2003. Social determinants of health: the solid facts (2nd ed.). World Health Organisation Library, (2003), Copenhagen.
8. Hashemian, M.S. and Stanley, K.G. 2011. Effective utilization of place as a resource in pocket switched networks. IEEE 36th Conference on Local Computer Networks (Oct. 2011), 247-250.
9. Bell, S., Jung, W. R., Krishnakumar, V. 2010. WiFi-based enhanced positioning systems: accuracy through mapping, calibration, and classification. Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness, (Nov. 2010), San Jose, California.
10. Bastian, M., Heymann, S., and Jacomy, M. 2009. Gephi: An open source software for exploring and manipulating networks. In International AAAI Conference on Weblogs and Social Media (2009).
11. Adar, E. 2006. GUESS: A Language and Interface for Graph Exploration. ACM Conference on Human Factors in Computing Systems (Apr. 2006).
12. Valente, T. W. 2010. Social Networks and Health: Models, Methods, and Applications, Oxford University Press, New York, USA.
13. Jolly, A. M., Muth, S. Q., Wylie, J. L., Potterat, J. J. 2001. Sexual Networks and Sexually Transmitted Infections: A Tale of Two Cities. J. of Urban Health: Bulletin of the NY Academy of Med., 78 (2001) 433-445.
14. 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.
15. Mikolajczyk, R. T., Akmatov, M. K., Rastin, S., Kretzschmar, M. 2008. Social contacts of school children and the transmission of respiratory-spread pathogens. Epidemiol. Infect. 136, 6 (Jun. 2008), 813-822.
16. Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J. F., and Vespignani, A. 2010. Dynamics of person-to-person interactions from distributed RFID sensor networks. PLoS ONE 5, 7 (Jul. 2010).
17. Hashemian, M. S., Stanley, K. G., and Osgood, N. 2010. FLUNET: Automated Tracking of Contacts During Flu Season. WiOpt'10: Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (May 2010), 557-562. 7
18. Joki, A., Burke, J. A., and Estrin, D. 2007. Campaignr-a framework for participatory data collection on mobile phones. Technical report, CENS, UCLA, (2007).
19. Lu, H., Yang, J., Liu, Z., Lane, N. D., Choudhury, T., and Campbell, A. T. 2010. The Jigsaw continuous sensing engine for mobile phone applications. In Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems, (Nov. 2010), Zürich, Switzerland.
20. Consolvo, S., McDonald, D. W., et. al. 2008. Activity sensing in the wild: a field trial of ubifit garden. In Proceeding of the 26th annual SIGCHI conference on Human factors in computing systems, (Apr. 2008), Florence, Italy.
21. Lane, D.; Miluzzo, E;, Hong, L.; Peebles, D.; Choudhury, T.; Campbell, A.T., 2010 A Survey of Mobile Phone Sensing, IEEE Communications Magazine, September 2010, pp. 140-150.