PEARL: Probing Entity Aggregation in Real Life

Xingwu Liu, Yunfei Bai, Chunlin Huang, Xiaoyan Wang, Dongbo Bu
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
xingwuliu@gmail.com, baiyunfei@software.ict.ac.cn, {huangchunlin,wangxiaoyan,dbu}@ict.ac.cn

***Abstract*— The classical approach for predicting tubercles bacillus (TB) prevalence falls in the partial differential equation (PDE) framework, which is essentially equal to the assumption of uniform random interaction network among people. The assumption, however, conflicts with common knowledge that some people interacts with many partners while others interact with limited partners. To accurately capture the interaction patterns among people, a mobile-based system called PEARL is proposed in this study. PEARL utilizes the characteristic that Bluetooth device has an effective range of ~10 meters, which is also the infectious distance of TB. Experimental results on several volunteers suggest that the interaction pattern roughly conforms to scale-free distribution, which helps improve prediction of prevalence of TB in China.**

Keywords-human interaction pattern, Bluetooth, volunteer computing, PEARL, mobile

I. Introduction

Respiratory infectious diseases such as tubercles bacillus (TB) and H1N1 are harmful to human health and productivity; thus, it is a pressing and tough challenge to predict and inhibit the spread of these diseases. The most widely used method is based on modeling their spread in terms of partial differential equations (PDEs). However, we found that the PDE approach essentially relies on the assumption that the human interaction network [1] is a uniformly random graph, i.e. each pair of persons has the same probability of interaction. Unfortunately, the assumption obviously does not hold in practice: a salesman might interact with up to 100 people every day while a programmer does less than 10. Then what on earth is the human interaction network like? Some researchers consider it scale-free. It is plausible, but there is no direct, convincing evidence since we don't have a global view of the network. So, this paper presents an approach to exploring human interaction pattern in real life. Our approach, called PEARL, uses Bluetooth to detect interactions among mobile users, which can approximate human interaction pattern in real life, especially when considering spread of respiratory infectious diseases.

The contribution of this paper lies in three aspects.

First, it presents the novel idea of capture human interaction pattern through mobile software. The pure software approach, in the flavor of volunteer computing, makes possible the goal of low cost and high participation. And it hopefully heralds a new computing paradigm in which many computationally weak, resource-limited computing elements cooperate through ad hoc, local interaction. This paradigm, called Sea Computing, seems suitable for the Internet of Things (IoT)[1].

Second, it exemplifies how to exploit real life human interaction data, using TB prediction as a case study. The experiments show that with the knowledge of real life human interaction pattern, we achieve more accurate prediction of TB in China than any other prediction method does. In fact, there are a great number of scenarios in which such data may be helpful, for example, public security, commerce advertisement, ands so on.

Third, a novel approach is proposed to capture the burst, the prevalence, and possible spread routes of flu-like deceases. Infectious disease prediction research depends not only on human interaction pattern, but also on the disease-specific spread parameters. Currently, there is little data for reliably determining such parameters of flu-like deceases. Our PEARL+ promises to provide large amount of supporting data in a timely, low cost fashion.

The rest of this paper is organized as follows. Section II presents related work. Section III elaborates on PEARL and PEARL+. Section IV presents our experimental results. Section V concludes this paper.

II. Related work

The most closely related work is [2], where the interaction pattern among the teachers and students in a middle school was detected using specially designed sensors. A sensor can detect other sensors within the sensing radius, and report such events to a datacenter. The teachers and students are required to take the sensors along with them all the time. So, finally their interaction pattern can be found by analyzing the sensory data. The experiment lasted only 1 day, and only one school participated. Furthermore, it relies on specially designed sensors, which leads to high cost and low participation.

In [3], citizen interaction pattern in a city is detected by public cameras distributed throughout the city. A camera identifies people in its view and records that the people appearing in the same view have an interaction. Citizen

[1] The human interaction network consists of vertices standing for people and edges indicating their ends have an interaction. Two persons are considered in interaction if they stay in infection distance from each other.

interaction pattern can be obtained by combining the records from all the cameras. The main concern is how to efficiently identify people precisely enough. For example, if someone, say Alice, appears at block A in the morning and at block B in the afternoon, the system should identify Alice from the camera views both at A and at B, in order to get accurate human interaction pattern. This is a demanding task facing current technology if there are a big number of cameras. Considering that people generally spend most of their time indoors, the public-camera-based approach misses much important information, since it is conceivable that human indoor interaction follows a disparate paradigm from the outdoor one.

The work in [4] is essentially like that in [3]. Reference [4] relies on volunteers to find out how wild animals are distributed in a range. Volunteers take photos of the wild animals they meet, and take down the GPS coordinates of where the animals appear. Then they fill in a questionnaire with the collected information. All such information can be used to figure out where the animals live and how they migrate. This is typical of the paradigm of participatory sensing [10]. The main concern is that the volunteers are deeply involved, in the sense that they have to pay attention to project-specific affairs that are not their own business. For example, they should take time to label the photos and fill in questionnaires.

Reference [5] presents a vibration-sensor-based method to detect earthquake. In [5], a vibration sensor with USB interface is plugged into a volunteer's desktop. Upon sensing vibration, a report is sent to a central server which is responsible for collecting and analyzing the data. When there are enough such desktops in an area, the noise can be statistically filtered, boiling down to an accurate decision on whether there is an earthquake. However, the project relies on special-purpose sensor and each one costs up to tens of dollars. Hence it is hard to be widely deployed.

Reference [9] present the Social Network Enabled Flu Trends framework, which monitors messages posted on Twitter with a mention of flu indicators to track and predict the emergence and spread of an influenza epidemic in a population. This method can be used to evaluate the recent situation of the epidemic, which provides advice for **effective interventions. However, it focuses on social networks in cyber space, being little informative on human real-life interaction pattern or the real-world spread model of the** epidemic.

And there are many other Bluetooth based applications, such as [6], but they are not designed for uncover human interaction pattern or for disease spread prediction. Our work follows the style of volunteer computing[7][8], which has been developing very fast in recent years. Our server is deployed on CAS@home[8]. Compared with the related work, PEARL is purely software-based, so it enjoys low cost and can be easily disseminated. It aims to capture human real-life interaction pattern whose value is far beyond the prediction of respiratory infectious diseases. PEARL+ is designed for discovering how an influenza epidemic spread in real world, which helps establish an accurate spread model of influenza epidemic.

III. METHOD

A. PEARL system

The rationale behind PEARL lies in two facts: First, the critical infection distance of diseases such as TB is about 10 meters, which is exactly the discovery radius of Bluetooth. Second, a barrier such as a wall can greatly lower the risk of infection, exactly as it decreases the Bluetooth discoverability. Hence Bluetooth is an ideal functionality to approximate the infectious interaction among people: two persons, with Bluetooth functionality turned on at their mobiles, are considered to have an interaction (in the sense of infection of respiratory diseases) if and only if the Bluetooth devices are mutually discoverable.

Technically, we follow the volunteer computing paradigm; the volunteer community can be considered as a sample of human: the interaction pattern of volunteers can well represent that of human. Every volunteer is required to install on his Bluetooth mobile the PEARL client, which keeps his/her Bluetooth always discoverable and reports the Bluetooth MAC addresses of other volunteer mobiles that has been discovered by Bluetooth.

One may question whether the volunteer community can approximate the human society in terms of real-life interaction. The answer is yes according to the sampling lemma in [11], if there are as many volunteers as enough. We can use a big graph G to represent human interaction: each vertex is a distinct person, and there is an edge between vertices v_1 and v_2 if and only if v_1 and v_2 interacted. The set S of volunteers can be regarded as a uniform sample of the whole human society. Let $G[S]$ stand for the subgraph of G induced by S. One can easily check that $G[S]$ is exactly the Bluetooth interaction network among the volunteers. Let $k=|S|$, the number of volunteers. The sampling lemma in [11] claims that with probability at least $1-2^{-k}$, the cut distance of G and $G[S]$, is no more than $10/(\log k)^{1/2}$. It seems that the upper bound is not small enough, for k even of the order of million. Note that this is a universal upper bound. It is conceivable that for a specific, naturally formed network G, the bounded may be considerably lowered. As a result, when there are many volunteers, we are confident to use the volunteer interaction pattern to approximate the interaction pattern of the whole human society.

The architecture of PEARL is illustrated in the following Figure 1. A PEARL client is running on every volunteer's mobile (named *host* of the client). The client keeps the host's Bluetooth always discoverable. It periodically scans for Bluetooth devices, records the Bluetooth MAC address of the discovered devices, and sends the records to the server via email. The server is responsible for processing all the collected data, discarding the invalid records "*A* meets *B* …" if either *B* is not a volunteer mobile or *B* does not report "*B* meets *A* …". The valid records are used to construct the interaction network. There are many ways to construct the network. For example, the edges can be defined on hourly basis, daily basis, or even weekly basis, in order to capture how the interaction pattern evolves temporally on different time basis.

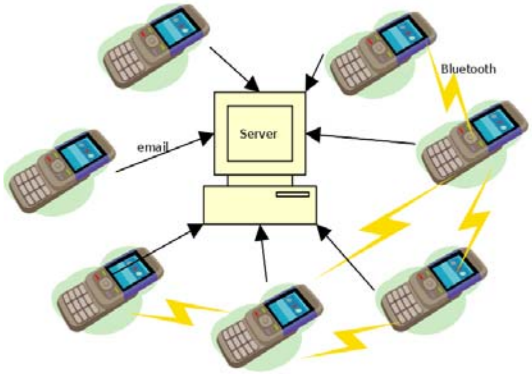

Figure 1. Architecture of PEARL

B. *PEARL+ system*

While PEARL uncovers human interaction pattern which is supposed to be related to the spread of respiratory infectious diseases (disease spread pattern), it has nothing to do with the disease spread pattern in itself. Using it to predict the spread of any diseases such as TB and H1N1 respectively, the results should be same, which is of course unreasonable. One may argue that it is a solution to take into account distinct spreading model of each disease. The question arise: what is the real-world spreading model of a certain respiratory infectious disease? The question is still open, and the prediction in practice can only rely on fictitious disease spread models. So, this paper presents PEARL+, taking the first step towards finding disease-specific spread pattern.

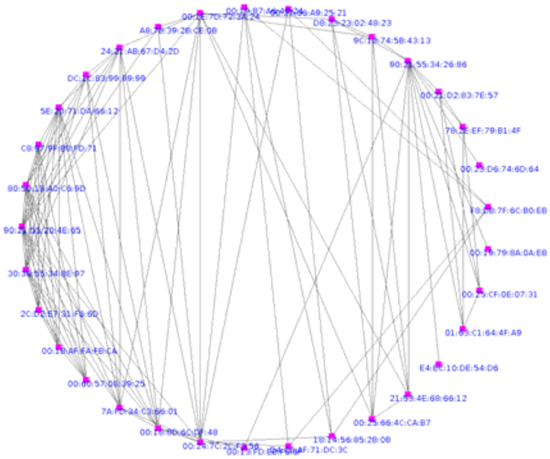

Figure 2. The interaction pattern acquired on 2011/04/08 (Friday).

The key idea of PEARL+ is to sensing the spread of a certain disease, say H1N1, by exploiting mobile sensors in addition to Bluetooth. Smart phones are widely used, and they are rich of various sensors such as voice sensor, vibration sensor, distance sensor, and so on. These sensors can be used to detect the symptom of H1N1. For example, voice sensor and vibration sensor together can reliably detect coughing and sneezing, which are typical of H1N1. In implementation, PEARL+ extends PEARL by adding sensor-featured codes to PEARL clients.

PEARL+ shares the architecture with PEARL. The difference lies in that its client extends that of PEARL by adding sensor-featured components. A preliminary version has been implemented for H1N1. In this implementation, the voice sensor identifies a cough or a sneeze by voice feature. Since voice sensor alone can't tell whether the sound come from the host or a nearby person, vibration sensor is also employed to identify a cough or a sneeze by vibration feature. Combining the two sensors, cough or sneeze can be rather reliably detected. To relieve the burden of mobile phones, the PEARL+ client filters sensory data only in coarse granularity: it extracts and reports suspicious signals that in some sense are similar to a cough or sneeze, letting the final, exact decision be made by the server.

IV. EXPERIMENTAL RESULTS

Currently we have dozens of volunteer mobiles within the Institute of Computing Technology, Chinese Academy of Sciences. The interaction network among the volunteers and their partners are shown in Figure 2 and 3. The figures suggest the following observations: First, on Saturday, the volunteers report interaction information of more partners than that in Friday. A reasonable explanation is that a volunteer interacts with more partners in weekend. Second, both figures suggest a non-uniform distribution of interactions among people, i.e. there are hubs in the interaction network, say the Bluetooth device with MAC address C8:97:9F:88:FD:71 in Figure 2. Third, both figures imply that the interaction network conforms to a scale-free distribution.

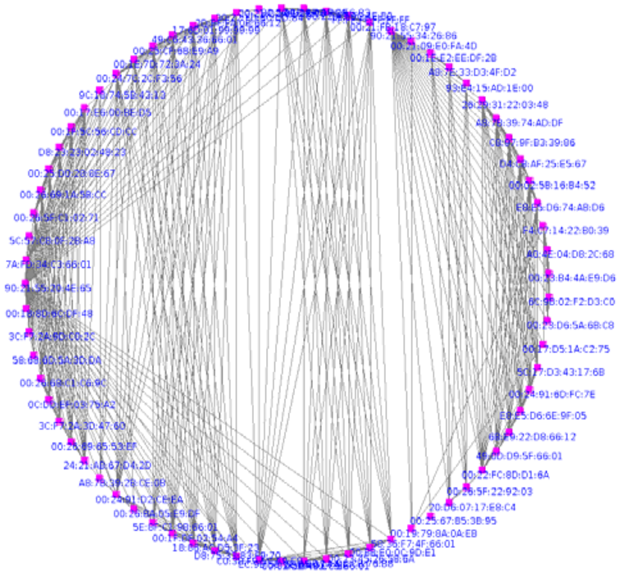

Figure 3. The interaction pattern acquired on 2011/04/09 (Saturday).

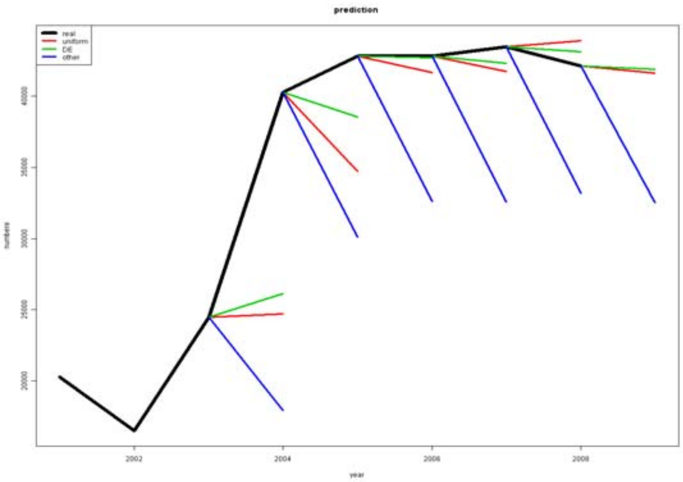

Figure 4. Comparison of prediction accuracy using different methods

We further predict the prevalence of TB in Sichuan province in China using the scale-free interaction network model. Specifically, we first construct interaction network among a population, and then simulate the spreading of TB in this network. We also conducted comparison of scale-free network and classical uniform interaction network. Experimental results suggest that the scale-free model (green line) turned out the closest to the real data (black line).

We have implemented PEARL for Symbian and Android platform. The binary code for each platform is no more than 2MB. Typically the daily data flow over mobile network is about 50KB. As to energy consumption, PEARL accounts for about 15% of the total battery consumption for typical users. Such high energy consumption is mainly due to the high frequency of Bluetooth scans. Since the currently-deployed PEARL is β-version, we set it to scan peers every five minutes, in order to test its performance in heavy load. In the next version, it will scan every ten minutes, and hopefully the battery consumption will decrease to less than 5%.

V. Conclusion and Discussion

We proposed PEARL to approximately capture human interaction pattern, in order to help predict and inhibit the spread of respiratory infectious diseases. One can see that human interaction pattern is in fact unrelated to disease spreading pattern, so PEARL+ is proposed to simultaneously capture both human interaction pattern and disease-specific spreading pattern. In this way, we can find out how these two patterns interplay, and most importantly provide the first real life disease spread model. We hope that our work can also attract attention to volunteer sensing.

Compared with PEARL which doesn't require any cost in hardware, the work in [2] and [5] needs specially designed sensors, resulting in high cost, so the experiment is hard to scale up. Unlike PEARL+, the work in [2] neglects disease-specific spreading information.

In addition, the approach in [3] doesn't collect indoor interaction data. And it relies on accurate, fast image processing, which is not available presently.

Acknowledgment

This work is partially supported by National Great Research Program of China (N2008ZX10003-009), Natural Science Foundation of China (30800168), and Major National Science and Technology Special Projects of China (2010ZX03004-003-03).

References

[1] L. Atzori, A. Iera , G. Morabito. *The Internet of Things: A survey*. Computer Networks. 54(15: 2787-2805, 2010

[2] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, J. H. Jones. *A High-Resolution Human Contact Network For Infectious Disease Transmission*. Proc Natl Acad Sci USA. 107(51):22020-5, 2010

[3] S. Eubank, H. Guclu, V. Kuma, M. Marathe, A. Srinivasan, Z. Toroczkai, N. Wang. *Modelling Disease Outbreaks In Realistic Urban Social Networks*. Nature. 429:180–184, 2004

[4] D. Aanensen, D. Huntley, E. Feil, F al-Own, B. Spratt. *EpiCollect: Linking Smart Phones To Web Applications For Epidemiology, Ecology And Community Data Collection*. PLoS One. 4(9):e6968, 2009

[5] http://qcn.stanford.edu

[6] N. Eagle, A. Pentland, D. Lazer. *Inferring Friendship Network Structure By Using Mobile Phone Data*. Proc Natl Acad Sci USA. 106(36):15274-8, 2009

[7] http://boinc.berkeley.edu

[8] http://casathome.ihep.ac.cn

[9] H. Achrekar, A. Gandhe, R. Lazarus, S. H. Yu, B. Liu.*Predicting Flu Trends using Twitter Data.* In Proc. of the 1st Internaltional Workshop of Cyber-Physical Networking Systems. April 2011, pp. 713-718

[10] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, M. B. Srivastava. *Participatory Sensing*. In ACM Sensys World Sensor Web Workshop. Oct., 2006

[11] C. Borgs, J. T. Chayes, L. Lovasz, V. T. Sos, and K. Vesztergombi. *Convergent Graph Sequences I: Subgraph Frequencies, Metric Properties, And Testing*. Advances in Math. 219(6):1801-1851,2008