2009 International Conference on Computational Science and Engineering

The Importance of Data Collection for Modelling Contact Networks

Eiko Yoneki
University of Cambridge Computer Laboratory
Cambridge, United Kingdom
eiko.yoneki@cl.cam.ac.uk

***Abstract*—The recently developed small wireless devices ranging from sensor boards to mobile phones provide a timely opportunity to gather unique data sets on complex human interactions, which in turn will support rich and meaningful modelling of the underlying networks. We are convinced that this approach will be fruitful and effective for tackling various issues such as infectious disease spread and beyond. Human social dynamics are far more complex than the current simplified models in network theory, and the data driven modelling with large-scale experimental results is essential for understanding and building systems that exploit real networks. An important issue to be addressed for taking empirical and heuristic approach is understanding the characteristics of data such as data collection methods, limitations, scale of noise, and so forth. This aspect has been many times neglected during inference of data. We express the importance of data driven approach in this paper.**

I. Introduction

The recent emergence of wireless technology (e.g. mobile phones and sensors) makes it possible to collect real world data on human proximity. Capturing human interactions with wireless sensors will allow us to understand complex patterns of human activities. For example, in one experiment people will carry mobile phones that record dynamic information about other devices nearby. A post-facto analysis of this data will yield valuable insight into how communities are formed, how much time people spend together, and how frequently they meet; such data exhibits complex network-like structures that are similar to social and biological networks. The analysis can also identify specific individuals who act as coalescing hubs of the network at different points in space and time, and who influence data flow. We have demonstrated series of experimental results [4][10][16].

In the 21st century, the epidemic spread of infectious diseases is still a serious threat, killing 15 million people every year. Beyond the advance of modern drugs, analyses of epidemic models by computer simulation have been useful for understanding the dynamics of disease spread. Advanced modelling provides predictions of future epidemics to build an early outbreak warning system. Accurate and reliable models of human social interactions are key input to such computer simulation. How does the epidemic diffusion occur, and how does the epidemic reflect the movement of people in the real world? Daily human interaction in real world is complex and the quantitative understanding of human dynamics is a difficult task and has not been explored at any depth. Real world networks are far more complex than the simplified models that derived from physics and graph theory. We must take advantage of real world data that reflects human interactions, which can be used to construct human contact networks and will provide valuable parameters for understanding and modelling such networks.

Building an effective and reliable human proximity detection system raises various issues. Particularly, optimal exploitation of technologies available across the hardware and software is necessary. Current detection mechanisms using WiFi access points or Bluetooth expect high failure, communication protocol limitation and complex statistics. Without in-depth understanding of the data collection mechanism, modelling networks will not be reliable. For example, the symmetry of edge detection is extremely low according to our experiments using Bluetooth. This indicates missing edges from the device detection leading inaccurate clustering coefficient calculation. With such noisy data, how deep can we infer contact networks? We need to understand at least the scale of missing edges. However, such important information is entirely missing in current research efforts.

Understanding epidemic spread often requires not only human interaction information but also sufficient data which enables to reconstruct physical environments. We envision large scale and detailed spatial world models including stationary objects (e.g. streets, landmarks, and mobile objects) that may be useful for modelling. Furthermore a model may be augmented by virtual objects to associate real world objects such as web links or events from social network services. We have done some initial work on such world modelling [13].

The rest of this paper is structured as follows. We describe background of complex networks in Section 2, introduce examples of human contact trace data sets in Section 3, and then describe the proximity detection mechanism by Bluetooth communication in Section 4. In Section 5, we discuss complexity of data collections followed by brief discussion of ethical/privacy issues in Section 6. Finally, we conclude in Section 7.

II. Epidemic in Complex Networks

In theoretical physics, Erdos and Renyi have shown the concept of phase transition in random networks [7], while Watts

978-0-7695-3823-5/09 $26.00 © 2009 IEEE
DOI 10.1109/CSE.2009.332

and Strogatz have shown small world networks for social relationships [15]. Barabasi described scale-free networks demonstrating the existence of hub nodes [2]. Universal rules govern the structure of all networks, whether they are social, technological or biological. Human society promotes cooperation, which is based on spatial and social relationships and can evolve as a consequence of social viscosity. In dynamic physical networks, how do people form communities and how does community structure affect epidemic spread in a population? How do hub nodes and weak links influence temporal or spatial effects and how does it affect the transmission characteristics of diseases? How do the community-like topology of interpersonal connections and its hierarchical nature yield a multi-level structure? Current models in network theory are too simplified, and answering these questions will require more in-depth understanding of real world networks. Multiple experiments and large-scale experimental data will be needed both for modelling and building systems.

Human proximity networks display extremely dynamic topology on a spatial and temporal scale. We refer to such networks as ‘time-dependent networks’. Generalisation of the measurements of complex networks is a recent active research topic. However, modelling dynamic temporal and spatial series of sub-networks (e.g. trees or motives) in time-dependent networks in a discrete form is a future challenge. Several researchers have worked on a predictive model for epidemics such as an influenza pandemic [8]. Such models require precise information of mobility, interaction, and behavioural assumption of the population. On the other hand, interactions between individuals are assumed to follow random encounters. Thus, human connectivity information including spatial and temporal information has yet to be incorporated in such models for improving the predictions.

Apart from confirming previously known results, such as that degree distributions with high variance of occurrence of high-degree individuals can be associated with an accelerated course of the epidemic [3]. Many other network characteristics (e.g. population size, geographical location) can be uncovered. Clustering will be an important factor to drive the epidemic, and looking into causal contact patterns of the epidemics will give additional insight. The patterns of interactions between individuals are key to understanding how infectious diseases spread, and each interaction may not be described in a binary form. Only considering monogamous pair relationships may be problematic, and consideration of the strength and regularity of connections will be necessary.

III. Real World Human Connectivity Traces

The Reality Mining project [6] collected proximity, location and activity information, with nearby nodes being discovered through periodic Bluetooth scans and location information from cell tower IDs. Several other groups have performed similar studies. Most of these [6] [1] [12] use Bluetooth to measure device connectivity, while others [9] rely on WiFi. The duration of experiments varies from 2 days to over one year, and the numbers of participants vary. We have analysed various traces from the Crawdad database [5] and several examples are listed below, and Table I summarises the configuration of them.

Experimental data set	MIT	UCSD	CAM	INFC06
Device	Phone	PDA	iMote	iMote
Network type	Bluetooth	WiFi	Bluetooth	Bluetooth
Duration (days)	246	77	11	3
Granularity (seconds)	300	600	120	120
Number of Experimental Devices	97	274	36	78

Table I
Characteristics of the experiments

MIT: in the MIT Reality Mining project [6], 100 smart phones were deployed to students and staff at MIT over a period of 9 months. These phones were running software that logged contacts.

UCSD: in the UCSD Wireless Topology Discovery [14], approximately 300 wireless PDAs running Windows CE were used to collect WiFi access point information periodically for 11 weeks.

CAM: in the Cambridge Haggle project [1], 40 iMotes were deployed to 1st year and 2nd year undergraduate students for 11 days. iMotes are sensor boards equipped Bluetooth for detecting proximity devices.

INFC06: 78 iMotes were deployed at the Infocom 2006 conference for 4 days [4].

IV. Proximity Detection with Bluetooth

Bluetooth is a low-power open standard for Personal Area Networks (PANs) and has gained its popularity due to its emphasis on short-range, low-power and easy integration into devices. The platform used in the Haggle experiments is the Intel Mote ISN100-BA (known as the ‘iMote’). The iMote runs TinyOS and is equipped with an ARM7TDMI processor operating at 12MHz, with 64kB of SRAM, 512kB of flash storage, and a multi-colored LED, and a Bluetooth 1.1 radio. The specifications lists the radio range to be 30 meters.

It is a complex task to collect accurate connectivity traces using Bluetooth communication, as the device discovery protocol may limit detection of all the devices nearby. Bluetooth uses a special physical channel for devices to discover each other. A device becomes discoverable by entering the inquiry substate where it can respond to inquiry requests. The inquiry scan substate is used to discover other devices. The discovering device iterates (hops) through all possible inquiry scan physical channel frequencies in a pseudo-random fashion. For each frequency, it sends an inquiry request and listens for responses. Therefore, a Bluetooth device cannot scan for other devices and be discoverable at the same time. Bluetooth inquiry can only happen in 1.28 second intervals. An interval of $4 \times 1.28 = 5.12$ seconds gives a more than 90% chance of finding a device. However, there is no data available when there are many devices and many human bodies around. The Bluetooth standard [12], recommends being in the inquiry scan substate for 10.24 seconds in order to collect all responses in an error-free environment. The power consumption of Bluetooth also limits the scanning interval, if devices have limited recharging capability. The iMote connectivity traces in Haggle use a scanning interval of approximately 2 minutes, while the Reality Mining project uses 5 minutes. The ratio of devices with Bluetooth enabled to the total number of devices is around only an average 15% of population.

Bluetooth for proximity detection is widely available and a lot of people carry a Bluetooth enabled mobile phone with them. Thus, it is possible to detect a certain amount of peoples phones without handing a special device to each of them, which makes Bluetooth appealing for experiments involving a large quantity of people. The range of Bluetooth varies between 10m and 100m, depending on the device class. In mobile phones, the range is usually 10m. We have observed the devices can be detected in 20m range if there is no obstacles, while if there is any obstacles such as a thick wall it limits to 5m range.

V. Complexity of connectivity data collection

We haves developed a range of ways for detecting and recording spatial proximity [4][10][16]. These include small custom built battery-powered sensor devices (i.e. Intel iMote) and mobile phones. In each case, software has been developed to record contacts with other devices. Each device is uniquely identifiable, so a network of contacts, which includes information about which devices interact, can be built. Furthermore, the duration of interactions are logged (both the duration of a single interaction and the cumulative duration of all interactions over the study period) to enable weights to be assigned to links in this contact network. Trials have taken place in a range of settings - from academic conferences to the streets of a city. Although developed for the purpose of designing wireless data-forwarding protocols, these methods are clearly directly relevant to social network epidemiology and provide an easy-to-use means of measuring weighted social networks. The technology has proved robust with reliable data collection at initial level. However, reliable network modelling requires further massive and precise experimental data.

In [17], we have shown the distribution of the inter-meeting time, where meeting indicates interaction among several nodes using K-Clique based meeting detection and inter-meeting is time between meetings. We immediately note that the bulk of inter-meetings times are within 24 hrs. However, the distribution does not appear very power-law in its nature, except perhaps for the early head of the distribution. While the interaction times between nodes is not power-law, the duration of meetings does appear to be. We can explain this discrepancy by noting that meetings involve many nodes and that counting these as pairs of interactions leads to weighting the meeting duration by the (often large) number of pairs, thus skewing the distribution. Furthermore, the Bluetooth detection data is noisy, which can make estimating the contiguous duration of an interaction unreliable, with the probability of a break due to noise growing with the length of the meeting. This thus also skews pairwise interactions to be shorter than they actually are. Unless Bluetooth detection gives high accuracy result, using the clustering methods, which can be better tolerated with noise is necessary for more reliably estimating the meeting duration.

CRAWDAD [5] provides an archive for a large amount of such human connectivity/mobility data. The current average data is small-scale with limited device detection capability. For example, the Bluetooth scanning interval is > 5 minutes, which may miss many devices in a busy street and may not provide sufficient information for critical analyses such as prevention of disease epidemics. From the collected traces, the reflectivity value (i.e. when A sees B, the probability that B sees A) is extremely low, even when setting the single time unit size to 10-15 minutes. This does not lead to accurate transitivity values for evaluating the network clustering. There have been various trace data archived, but each trace is collected in different ways. Thus, we need to look into the level of accuracy required for the model, so that data collection can be organised accordingly. This process is completely missing in current research.

It is desirable to explore not only Bluetooth communication based proximity detection, but explore various methods including 802.15/ZigBee boards. Existing trace data typically lacks geographical information. We are experimenting GPS equipped mobile phones, small computers, and embedded Linux boards to design tracking and localisation mechanisms in an efficient and inexpensive way. Software that detects proximate devices for the phone will be developed based on our previous work [12], extending to GPS tracking and capturing image/sound as necessary to record contexts in the environments. The proximity networks represent pair relationships, proximity based modularity, and social interactions, while online based interactions such as email, instant messaging, and social network services (e.g. Facebook, LinkedIn, Orkut) represent another type of social interactions. We collect data from online social networks to be used for network analysis, including correlation between two types of social networks.

VI. Ethical/Privacy/Anonymity issues

We am well aware of ethical and privacy issues for the collected data, and the data must never be used to identify individuals. The collected data will be anonymised before analysis. Software developed for mobile phones may involve collaboration between ad hoc groups of members. When new encounters occur, there are complex issues in knowing what entities to trust. Based on predefined trust, recommendations, risk evaluation and experience from past interactions, an entity may derive new trust metrics to use as the basis for authorisation policies for access control. This raises serious concerns about privacy, surveillance and freedom of action. While providing location information can clearly be a one-way system where the location providing tools do not track who is receiving, once a device receives information, its location is potentially available to others.

VII. Conclusions

Data-driven modelling of human interaction dynamics is described in this paper, where experimental measurements are followed by mathematical modelling. We emphasise that real-world data needs to drive modelling. The derived network models need to be accurate and parameterised with data on human interaction patterns, modularity, and details of time dependent activity and it is important to understand data collection methods, limitations, scale of noise for data collections. How the data is measured would influence how deep we are able to infer network properties. The derived models can be used by many applications, for example, determination of epidemic spread and construction of synthetic networks. Data collection requires careful attention to the ethic and privacy issues that will have to be addressed.

In data analysis, the focus on on modelling structure and dynamics could be set from the following aspects: 1) Physical proximity networks represent pair relationships, proximity based modularity, and social interactions including studying online based interactions such as email, instant messaging, and social network services (e.g. Facebook, Orkut). 2) Real social networks are not random as they exhibit modularity. 3) Many real world networks are weighted, but little analysis has been done in this area [11]. Topology is closely related to edge weights, which influences how the modularity of networks is formed. The connectivity traces can be represented by weighted graphs in which the weight of an edge represents, for example, the contact duration and frequency for the two end vertices. 4) Human proximity demonstrates the topology changes for every time unit. Thus, existing network measurement metrics for static networks are difficult to apply. Centrality measurements give insight into the roles and tasks of nodes in a network including degree, betweenness, and closeness centralities. 5) Analysing the structural properties of growing networks is important. In each time unit, several nodes appear or disappear, and each node selects or deselects possible counter parts from existing networks. Identifying such dynamics from empirical trace defines the form of network evolution, where high dynamics indicates significant network transition.

Acknowledgment

This research is funded in part by the EU grants for the Haggle project, IST-4-027918, and the SOCIALNETS project, 217141.

References

[1] Haggle Project, http://www.haggleproject.org, 2008.

[2] A. Barabsi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[3] M. Barthlemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani. Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. *Physical Review Letters*, 92:178701, 2004.

[4] A. Chaintreau et al. Impact of human mobility on the design of opportunistic forwarding algorithms. In *Proc. INFOCOM*, April 2006.

[5] D. College. A community resource for archiving wireless data at dartmouth, http://crawdad.cs.dartmouth.edu/index.php, 2007.

[6] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, V10(4):255–268, May 2006.

[7] P. Erdos and A. Rnyi. On Random Graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[8] S. Eubank, H. Guclu, V. Kumar, M. Marathe, a. Srinivasan, z. Toroczkal, and N. Wang. Modelling disease outbreaks in ralistic urban social networks. *Nature*, 429, 2004.

[9] T. Henderson et al. The changing usage of a mature campus-wide wireless network. In *Proc. Mobicom*, 2004.

[10] P. Hui, J. Crowcroft, and E. Yoneki. BUBBLE Rap: Social Based Forwarding in Delay Tolerant Networks. In *MobiHoc*, 2008.

[11] M. Newman. Analysis of weighted networks. *Physical Review E*, 70:056131, 2004.

[12] T. Nicolai, E. Yoneki, N. Behrens, and H. Kenn. Exploring social context with the wireless rope. In *Proc. Workshop MONET: LNCS 4277*, 2006.

[13] A. Peddemors and E. Yoneki. Decentralized Probabilistic World Modeling with Cooperative Sensing. In *KiVS Workshop on Global Sensor Networks,*, 2009.

[14] UCSD. Wireless topology discovery project, http://sysnet.ucsd.edu/wtd/wtd.html, 2004.

[15] D. J. Watts. *Small Worlds – The Dynamics of Networks between Order and Randomneess*. Princeton University Press, Princeton, New Jersey, 1999.

[16] E. Yoneki and J. Crowcroft. Wireless Epidemic Spread in Dynamic Human Networks. *Bio-Inspired Computing and Communication*, LNCS(5151), 2008.

[17] E. Yoneki, D. Greenfield, and J. Crowcroft. Dynamics of Inter-Meeting Time in Human Contact Networks. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2009.