Ad Hoc Networks xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Ad Hoc Networks

journal homepage: www.elsevier.com/locate/adhoc

EpiMap: Towards quantifying contact networks for understanding epidemiology in developing countries

Eiko Yoneki *, Jon Crowcroft

University of Cambridge, Computer Laboratory, Cambridge CB3 0FD, United Kingdom

ARTICLE INFO

Article history:
Received 10 March 2012
Received in revised form 22 May 2012
Accepted 18 June 2012
Available online xxxx

Keywords:
Human contact networks
Epidemiology
Mobile phone
Proximity radio communication
Delay tolerant networks
Satellite communication
Network modelling

ABSTRACT

We describe the EpiMap project, together with the FluPhone project where we developed the basic technology for EpiMap. In FluPhone, human contact data is collected using mobile phones to record information such as locality and user symptoms for flu or cold. Delay tolerant opportunistic networks were used as a basis for communication. We are extending the technology used in FluPhone to gather information on human interactions within rural communities of developing countries. The collected information will be used to develop improved mathematical models for the spread of infectious diseases such as measles, tuberculosis and pneumococcal diseases. Survey study will aid the understanding of the living conditions in these villages.

We introduce the EpiMap vision for a system of opportunistic networks combined with satellite communication, designed to face the challenges posed by weak electricity and communication infrastructures in rural regions of developing countries in Asia, Africa and South America. We aim to use a delay-tolerant small satellite for data transfer between developing countries and Europe ore North America. Data collected through EpiMap can be used to help design more efficient vaccination strategies and equitable control programmes.

1. Introduction

Many of Africa's significant diseases such as measles, tuberculosis, meningococcal, respiratory syncytial virus, and influenza are spread directly via person to person contact. These diseases are vaccine preventable and there has been a significant investment in improving vaccine coverage. Many countries face difficult decisions to refine the effective vaccination strategies within the limited budget.

Modelling the spread of infectious disease mathematically has been a useful tool for helping to design efficient immunisation programmes. Despite this, there is a lack of such studies. In Africa, few transmission models of vaccine preventable diseases have been developed. Thus, it is desirable to develop models for specific disease by making available social contact data, thereby encouraging others to develop their own models. For example, having a model prepared in advance based on up-to-date data on contact patterns and measles transmission data will greatly aid the efficient design of measles control. The lack of any such model with which to evaluate the polio endgame has certainly hampered and delayed any progress of controlling the spread.

To achieve this goal, both developing an advanced mathematical modelling and more importantly reliable and informative social contact data are essential. Thus, we focus in this paper on how to obtain such social contact data. Mathematical models are only as reliable as the assumptions and parameters upon which they are built. Social mixing patterns are central determinants of transmission for infections which demand close contact between individuals. In reality, very little is known about contemporary social contact patterns, especially in the third-world countries. Quantifying the social network

* Corresponding author.
E-mail addresses: eiko.yoneki@cl.cam.ac.uk (E. Yoneki), jon.corwcorft@cl.cam.ac.uk (J. Crowcroft).

1570-8705/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.adhoc.2012.06.003

structures through which close-contact diseases are spread will greatly aid our understanding of their epidemiology and help build better models. Such models can be used to predict the course of an outbreak, help design efficient control programmes, and help us to understand and control the emergence and spread of novel pathogens.

Contact diaries have traditionally been used to record contact information. The contact diary is developed through self-reporting or an interview-led processes. Such diary based approach has its limitations and it becomes a burden when repeated over many days or weeks; consequently it is usually applied as a single day snapshot of individuals' contact patterns.

Recently, the use of sensor devices to collect social contact information has emerged as an alternative. All use variations of wireless proximity sensors. These have primarily made use of either Bluetooth enabled mobile phones, or RFID tags/Bluetooth sensors. In either instance low power radio transmissions are sent and received on a regular basis, detecting similar devices within range (typically 5–10 m for Bluetooth, whereas RFID tags are more tunable). These methods have the advantage of being minimally invasive, just requiring participants to carry or wear the device over a period of time. Records of encounters can typically be stored on the device, and uploaded periodically to a server or base-station. Currently these methods have primarily been used in small-scale studies in enclosed settings (such as schools, hospitals and conferences). However, they offer the possibility of capturing the fine structure and dynamics of social networks at different spatial and temporal scales.

We have previously demonstrated the FluPhone Project [1,2], which aims to bring together epidemiologists, sociologists, and computer scientists, with the goal of developing novel and innovative methods with which to measure and understand social encounters based in Cambridge, UK. Such information helps scientists and medical researchers to understand how close-contact infections, such as swine flu, spread between different people. The FluPhone project was mainly targeted for tackling flu-like symptoms, which was a threat in our society a couple of years ago. Human proximity information is collected using phones with Bluetooth communication from the general population to build time dependent contact networks. The project also included a 'virtual disease' experiment, where a specific model of disease is spread through the proximity based communication upon encountering of two devices. The spread of different stages of the disease was then mapped across the locality of the study, and fed back to the user.

The FluPhone is built over the Haggle framework [3], in which we introduced Pocket Switched Networks (PSNs), a type of Delay Tolerant Network (DTN), exploring proximity based communication. PSNs provide communication in highly stressed settings with intermittent connectivity, variable delays and high error rates in decentralised and distributed environments over a multitude of devices that are dynamically networked. A partitioned network can deal with disconnected operation using a store-and-forward approach to communication. In PSNs people carry devices in their pockets, which communicate directly with other devices within their range or with infrastructure. Because device mobility is reflected by the user's movement, we have worked on understanding the social structure among the people who carry the devices. In many ways, the concept of PSNs is analogous to how infectious diseases spread. One key aspect of this is working out the numbers of social encounters and links in a chain of contacts between different people (similar to the idea of how many steps we are away from a particular person). One way to measure this is to record how often different people (who may not know each other) come into close proximity with each other, as part of their everyday lives. PSNs share many issues with epidemiological studies.

In this paper, we introduce the FluPhone project that is a basis of the EpiMap project and describe a vision of the EpiMap project, where we extend the FluPhone project to be able to deploy large-scale social contact data collection in developing nations. We plan to include not only phones but other sensing devices such as RFID tags and to exploit various communication methods. For example, the use of satellite communication will have a great potential in overcoming the limitations of sparse power and communications networks in the targeted regions. There is a diverse range of environments in developing countries, ranging from situations with no power or radio to situations with basic infrastructure being deployed for sparse Internet connections. Thus, one of our goals of data collection methodology is to be flexible to various environment, with different sensors taking an important role besides phones.

In parallel to data collection by EpiMap, we propose to conduct surveys for validating models derived by EpiMap on the spread of respiratory infections. We also briefly demonstrate what we can be modelled using the collected social contact data in this paper but further detail of the modelling is out of scope of this paper. It is our intention to develop a range of mathematical models based on contact patterns, which can be used to help guide vaccine policy development over the coming decade in Africa and other developing countries.

2. FluPhone project

The FluPhone project studies how often different people (who may not know each other) come close to one another, as part of their everyday lives. To do this, we asked volunteers to install a small piece of software on their mobile phones and to carry their phones with them during their normal day-to-day activities. The software will look for other nearby phones periodically using Bluetooth communication, record this information and send it back to the research team via a cellular phone data service or other means. This information gives us a much better understanding of how often people congregate into small groups or crowds. Also, by knowing which phones come close to one another, we will be able to work out how far apart people actually are, and how fast diseases could spread within communities. We also asked participants to inform us of any influenza-like symptoms they may experience during the study period, so that we can match the spread of flu to the underlying social network of encounters made. Participants were able to log-onto the study website and see

Please cite this article in press as: E. Yoneki, J. Crowcroft, EpiMap: Towards quantifying contact networks for understanding epidemiology in developing countries, Ad Hoc Netw. (2012), http://dx.doi.org/10.1016/j.adhoc.2012.06.003

an estimate of how many people they have encountered. Further details can be found in FluPhone study web depicted in Fig. 1 (see also [1,4]).

The FluPhone application can be downloaded from the web following a registration and authentication process. Data collection can operate over three different methods such as periodic uploading via the web, real-time collection via 3G, and post-study collection from the devices.

The FluPhone study is being carried out in Cambridge, United Kingdom, and was advertised via the channels of the University of Cambridge and various online social media such as Facebook [5] and Twitter [6]. Targeted participants include university members, their families, colleagues, friends, and people who work or live in Cambridge. Participants must be over 12 years old (under 16 s require parental/carer consent) and the registration process requires the consent of each participant.

2.1. FluPhone software

FluPhone provides software that runs on the users' mobile phones which the users carry with them during their normal day-to-day activities. FluPhone adopts a simple client–server design consisting of a mobile phone application in the phone and a receiver as a PHP (Hypertext Preprocessor) script on a web server. The mobile phone application collects Bluetooth device proximity data, Global Positioning System (GPS) coordination data, and data on self-reported flu symptoms through a user interface. The collected data is sent via GPRS/3G to the server. The user can also upload the data over a web interface. Fig. 2 shows a screen capture of the application. Communication between the application and the server is based on the secure Internet protocols and standard public-key infrastructure. Through the secure login to his/her account, the participant can look at their history of activities.

The collected information provides us with a much better understanding of how often people are in different groups, such as when commuting or through work or leisure activities. Without complex analysis a simple history of encounters can be used to measure activity, environment, cyclic behaviour, and so forth. The experiment result shows an average person encountered over 1500 unique devices over a 10 day period.

3. Virtual disease experiment

There is also a function called 'virtual disease', which models epidemics on participants' phones, giving a real-time picture of the social network between participants from the perspective of infectious disease. The virtual disease application is implemented as an Android application built to run on devices that support the haggle architecture [3]. The application broadcasts information about virtual diseases. Devices are infected by these virtual diseases based on a simple probability calculation. The application logs all incoming diseases and stores information regarding how they are processed. It also regularly scans for Bluetooth and GPS based location data. In virtual disease experiment, the spreading of diseases is simulated with a simple SEIR model (S: Susceptable, E: Exposed, I: Infectious, and R: Recovered). In this model, each device is originally susceptible to a disease. Once it is infected by another device, it becomes exposed for a specified time. Whilst it is exposed, a device has the disease but cannot yet infect other devices. Once the exposed duration has run out, the device becomes infectious for a specific time. Whilst it is infectious, the device can infect other devices. Each disease has an associated infection probability which indicates the likelihood that another device will be infected. Once the infectious duration has run out, the device has recovered from the disease and cannot be

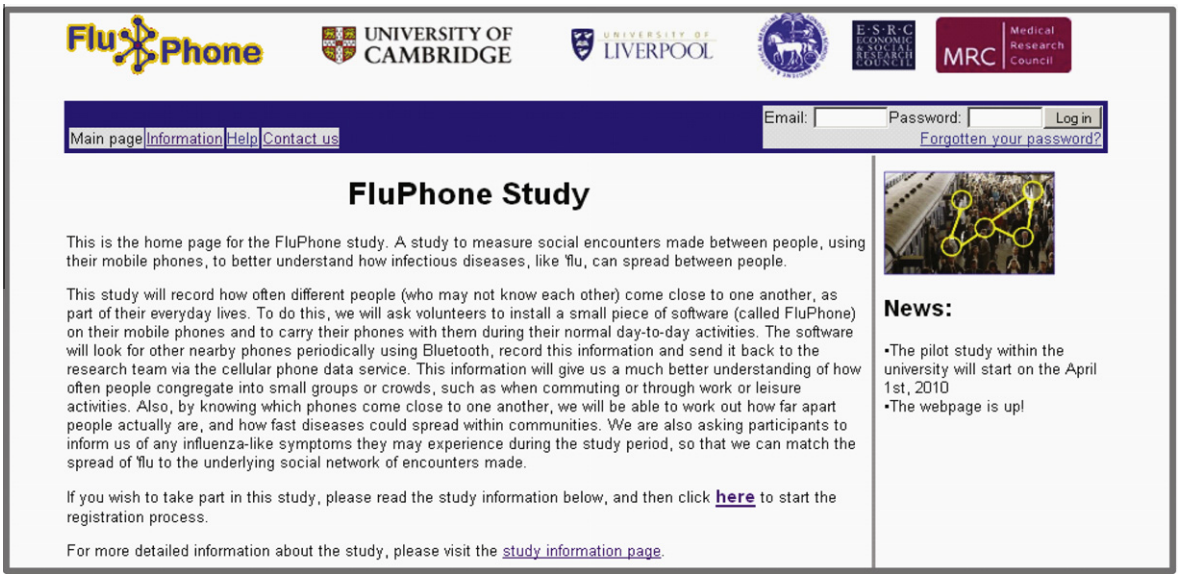


Fig. 1. FluPhone study web interface.

Please cite this article in press as: E. Yoneki, J. Crowcroft, EpiMap: Towards quantifying contact networks for understanding epidemiology in developing countries, Ad Hoc Netw. (2012), http://dx.doi.org/10.1016/j.adhoc.2012.06.003

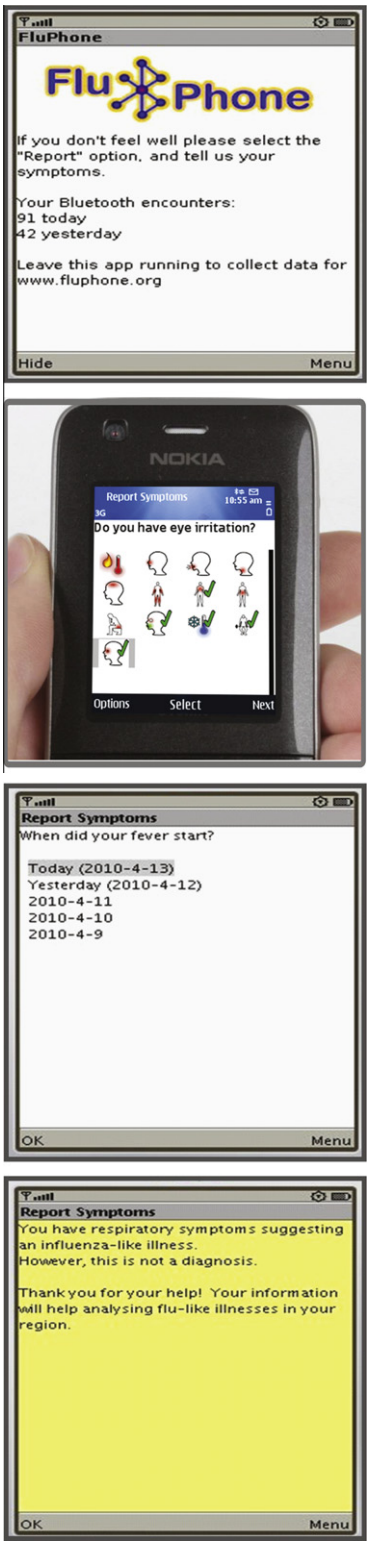


Fig. 2. FluPhone: encountering statistics, symptom type and time entry screens, and brief diagnostics.

reinfected. The application shows the state of the infection of each disease including who/when passed the virtual disease. The screenshot of the heatmap of virtual disease statistics is shown in Fig. 3. The locations of the infectious nodes are depicted in the map shown in the left side of Fig. 3, which can be zoomed in showing more details in the area in the right side of Fig. 3. We deployed three different diseases as follows, where beta = transmission probability, alpha = incubation time, and IP = infection period.

- SARS: fast (beta = 0.8; alpha = 24 h; IP = 30 h)
- FLU: normal (beta = 0.4, alpha = 48 h; IP = 60 h)
- COLD: slow (beta = 0.2, alpha = 72 h; IP = 120 h)

4. Disease spread modelling

In our previous work, we have worked on uncovering community structures, centrality node, weighted networks and so forth [7,8] in the context of PSNs. In [9–11], we looked into modelling inter contact time, meeting time, and epidemic spread patterns. The detail of modelling work is out of scope of this paper and in this section we briefly demonstrate how we can model multi-modal spread modes extracted from the contact networks. In this analysis, we defined spanning trees on various time points and aggregated them using 'joint diagonalisation' (JD), which is a technique to estimate an average eigenspace of a set of matrices. Using JD on matrices of spanning trees of a network is especially useful in the case of real-world contact networks in which a single underlying static graph does not exist. The average eigenspace may be used to construct a graph which represents the average spanning tree of the network or a representation of the most common propagation paths. We then examine the distribution of deviations from the average and find that this distribution in real-world contact networks is multi-modal; thus indicating several modes in the underlying network. These modes are identified and are found to correspond to particular times. See [12] for further detail of modelling and analysis.

This technique was applied to a contact network trace with 36 students in Cambridge, and revealed five spread modes corresponding to different times of day. Fig. 4 depicts one of the modes, where a highly structured network corresponding to the day when the groups are well defined by *class year* (i.e. year 1 and year 2). There is an obvious bridge formed by nodes 3 and 20. Using the average graph as an indicator, this implies that a disease spread at this time from nodes 3 and 20 should have the fastest infection rate.

This spread model is used for testing the infection rate, an SEIR model is constructed, setting the probability of infection to 0.5; infection time Poisson distributed with mean time of 800 min. A disease is spread through the contact network starting at time index 250. The simulation is repeated 30 times for each node and the results bootstrapped to give estimates of the mean number of people susceptible (those that have not received the disease) at time, $t, S(t)$. Fig. 5 shows the results of these simulations and as can be seen the number of susceptible people falls most rapidly for infections started at nodes 3 and 20, as expected.

Please cite this article in press as: E. Yoneki, J. Crowcroft, EpiMap: Towards quantifying contact networks for understanding epidemiology in developing countries, Ad Hoc Netw. (2012), http://dx.doi.org/10.1016/j.adhoc.2012.06.003

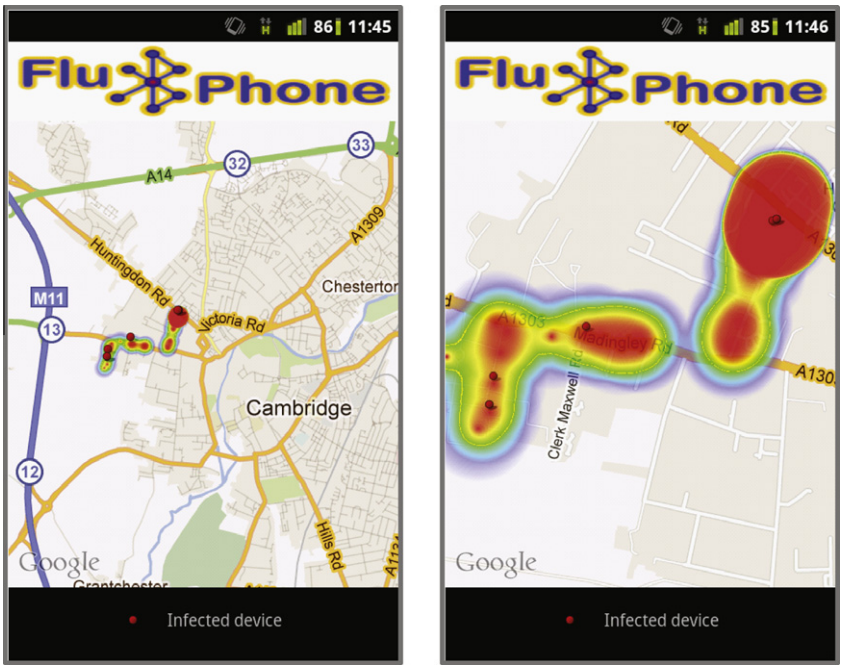


Fig. 3. Virtual disease showing infection map.

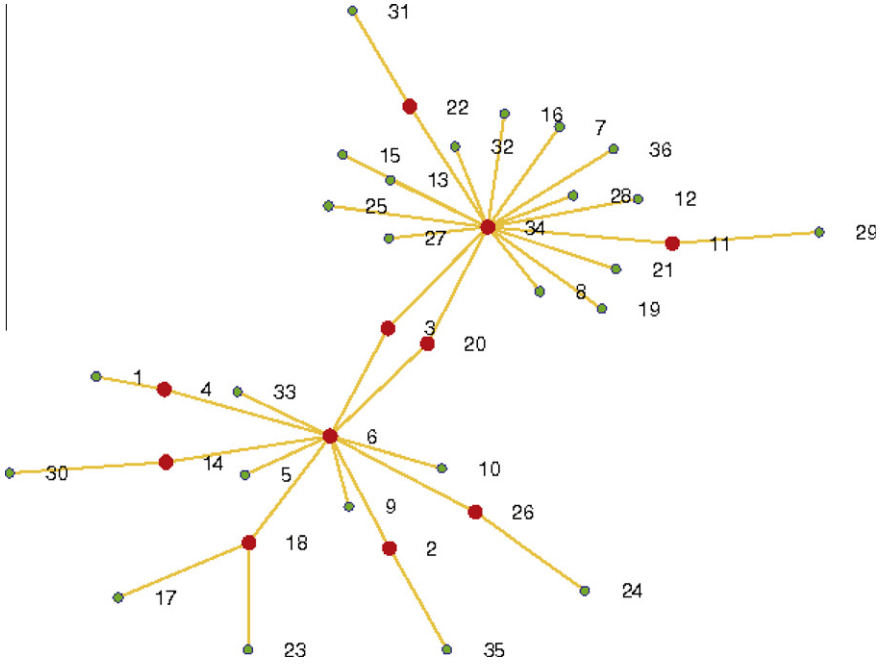


Fig. 4. Average graph in mode 5.

5. FluPhone to EpiMap

The use of mobile phones and/or sensors for quantitative measurement of societal mixing patterns to underpin mathematical models of the spread of close-contact diseases has distinct advantages over other methods of collecting contact data (such as diaries and interviews). Such devices can be programmed to gather proximity data automatically, allowing detailed longitudinal studies to be conducted with no possibilities of re-call bias, no barriers due to problems of literacy or understanding, and minimal disruption to the participants in the survey. They therefore offer an unparalleled opportunity to collect information on social contact patterns that would allow a step-change in

Please cite this article in press as: E. Yoneki, J. Crowcroft, EpiMap: Towards quantifying contact networks for understanding epidemiology in developing countries, Ad Hoc Netw. (2012), http://dx.doi.org/10.1016/j.adhoc.2012.06.003

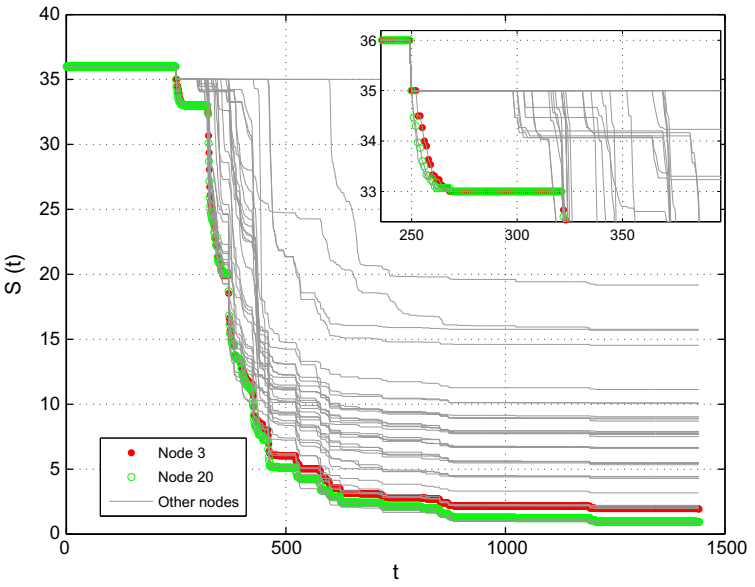


Fig. 5. Mean number of nodes susceptible to disease after time t (root infection starts at time 250. Inset focuses on the start of the infection).

our understanding of the patterns of disease spread. Despite the emergence of such technology, the use of sensors for this purpose has not been explored in this setting before.

The environments in developing countries vary and we will develop an application (named EpiMap) for various types of devices: mobile phones with or without GPS functionality, sensor boards, and RFID tags. Thus, we need to consider a hybrid model of data collection, allowing operation in the absence of reliable electricity supply, lack of Internet access, and so forth. We will use Bluetooth communication and RFID based communication for detecting devices within radio proximity, along with sensing (e.g. movement, light, humidity). Various sensors are embedded in phones and devices, and can be used to capture changes in surrounding context information in order to infer the behaviour and patterns of the device carriers. Accelerometers are especially useful for this purpose. The EpiMap phones and devices can be carried by the study participants and by health workers. The EpiMap application will also feature additional functions such as a display of nearby sensors, a history of encounters and can also support logging capability for events, activities of participants, and any useful observations of the environment. Battery life is an important issue, and the application will be carefully designed to minimise I/O and CPU usage. The study will be carried out initially in a rural African community, and we will therefore combine the use of spare batteries and power charging efficiency for the duration of study.

We will also augment the electronically-collected social contact data with several weeks of self-reported surveys. Data analysis will compare the different social contact and survey tools, and mathematical modelling will validate the mixing patterns recorded against known age-based zero-conversion rates for common respiratory infectious diseases in the population.

The information collected by EpiMap will be kept within the device until the end of the study or uplifted via automated data collection, which could be deployed using delay tolerant networks over Bluetooth communication or with WiFi base stations. The collected data will comprise a time series of encountering pairs of sensors/phones within physical proximity. A time-dependent graph of connectivity among phones can be constructed, detailing the duration and frequency of interactions made between participants. This will form the basis of analysis with other tools. Apart from building a contact map, various types of contextual information can be recorded by EpiMap such as movement and light, which will be analysed together with the diary and interview survey tools, where those tools will record the number of unique individuals encountered by that participant during a day, as well as some contextual information (whether at home, work, health facility, congregation, school, market, socialising). The data collected by survey-based methods will include some information on individuals not participating in the study, so at the individual level, the phone-collected data will intersect with the diary and interviews information. Direct comparison between tools will consider participant–participant encounters only, and will focus on accuracy of recording, reciprocity, duration, and frequency of encounters thought to occur between a participant's contacts. This comparison will enable correction factors, by age and sex, to be estimated for future phone-only studies.

The dynamic network of connections between participants will be used to investigate the topology of the social network, including: (1) duration-weighted pairs: time spent in close-proximity is a powerful determinant of infection risk and these can be considered as a weighted link between individuals with location and context associations; (2) number of encounters per person: to determine whether some individuals are responsible for a disproportionate number

Please cite this article in press as: E. Yoneki, J. Crowcroft, EpiMap: Towards quantifying contact networks for understanding epidemiology in developing countries, Ad Hoc Netw. (2012), http://dx.doi.org/10.1016/j.adhoc.2012.06.003

of contacts; (3) social distances: betweenness and centrality measures describe how far apart individuals are in a network, and strongly impact disease dynamics; and (4) community structure: identifying individuals that form bridging links between otherwise distinct groups offers efficient targeted interventions. We expect all of these measures and the network to change rapidly over time, and we will consider the implications for disease transmission in the modelling stage. Added value comes from the creation of a well-defined dynamic network of real human interactions, pertinent to respiratory infectious diseases, which will be a valuable contribution to the development of a new generation of analytical methods and measures that cope explicitly with dynamic networks. While analysis of the survey data can inform greatly on the potential for disease transmission, using mathematical models of disease transmission enables all the features to be incorporated at once to consider the implications of contact patterns and network structure on transmission, prevalence and possible intervention regimes. Final validation of the survey tools will be provided by using the models to simulate the spread of respiratory infections, and compare the age-based prevalence generated by the models to age-based incidence and longitudinal household prevalence of pneumococcal disease carriage information for the region. We aim to deploy EpiMap in the scale of 100–300 participants in the initial step. After the initial prototype stage, we will widen the study to a range of settings, including urban and rural sites across different parts of the developing world. The application could be extended to web-based data collection. We will also expand the temporal window over which data are collected, to capture seasonal differences in contact patterns and assess how these may affect the spread of disease. We plan to use collected data for understanding insight into the spread and control of diseases. Examples of diseases to be modelled are tuberculosis, pneumococcal disease, meningococcal disease, measles, and disease associated with Hemophilus influenza.

6. Data collection and communication

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. The detection mechanisms in the FluPhone 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. Missing edges from device detection leads to inaccurate clustering coefficient calculation. This noise hampers our ability to infer deep knowledge from this data.

6.1. Proximity detection

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 experiment in the Haggle project [3] is the Intel Mote ISN100-BA (known as the 'iMote'). The iMote runs TinyOS and is equipped with an ARM7TDMI processor operating at 12 MHz, with 64 kB of SRAM, 512 kB of flash storage, and a multi-coloured LED, and a Bluetooth 1.1 radio. The specifications lists the radio range to be 30 m.

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 s intervals. An interval of $4 \times 1.28 = 5.12$ s 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 power consumption of Bluetooth also limits the scanning interval, if devices have limited recharging capability. When mobile phones are available in some part of developing countries, Bluetooth for proximity detection can be used, since 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. However, in the region where there is no electricity battery equipped small sensor boards will be more helpful for proximity data collection. The range of Bluetooth varies between 10 m and 100 m, depending on the device class. In mobile phones, the range is usually 10 m. We have observed the devices can be detected in 20 m range if there is no obstacles, while if there is any obstacles such as a thick wall it limits to 5 m range.

We plan to extend to include audio recording when two devices are in proximity range so that the type of interaction can be inferred. Note that the audio is used for determining the patterns and tone of interaction and it does not examine the content of audio. As a prior work, Wyatt et al. [13] have shown effective analysis of privacy sensitive Audio.

It is increasingly popular to use radio frequency identification (RFID) for identification and security application. RFID transmits the identity (in the form of a unique serial number) of an object or person wirelessly, using radio waves. It's grouped under the broad category of automatic identification technologies. RFID is in use all around us. If you have ever chipped your pet with an ID tag, used EZPass through a toll booth, or paid for gas using SpeedPass, you've used RFID. Unlike ubiquitous UPC bar-code technology, RFID technology does not require contact or line of sight for communication. RFID data can be read through the human body, clothing and non-metallic materials. RFID

Please cite this article in press as: E. Yoneki, J. Crowcroft, EpiMap: Towards quantifying contact networks for understanding epidemiology in developing countries, Ad Hoc Netw. (2012), http://dx.doi.org/10.1016/j.adhoc.2012.06.003

requires three components are depicted; an antenna, a transceiver (with decoder), and a transponder (RFID tag) electronically programmed with unique information. The antenna emits radio signals to activate the tag and to read and write data to it. then the reader emits radio waves in ranges of anywhere from one inch to 100 feet or more, depending upon its power output and the radio frequency used. When an RFID tag passes through the electromagnetic zone, it detects the reader's activation signal. The reader decodes the data encoded in the tag's integrated circuit (silicon chip) and the data is passed to the host computer for processing. RFID quickly gained attention because of its ability to track moving objects. As the technology is refined, it will be more pervasive and invasive. A typical RFID tag consists of a microchip attached to a radio antenna mounted on a substrate. Normally the chip can store as much as 2 kbytes of data.

The SocioPatterns project [14,15] uses active RFID devices, embedded in unobtrusive wearable badges. Detailed information on how this technology is used to monitor social interactions and to identify contact patterns. Individuals are asked to wear the devices on their chests, so that badges can exchange radio packets only when the individuals wearing them face each other at close range (about 1–1.5 m). This range was chosen as a proxy of a close-range encounter during which a communicable disease infection can be transmitted, for example, either by cough or sneeze, or directly by hands contact. The infrastructure parameters are tuned so that the proximity of two individuals wearing the RFID badges can be assessed with a probability in excess of 99% over an interval of 20 s. The problem of RFID tag is that the reader is required, which is normally expensive and the tag normally does not have enough storage to keep the data for long time. Thus, frequent transferring the logged data to the reader device is necessary.

None of technique for the proximity detection is perfect currently and we will plan a hybrid system to tailor to the environment in developing countries. 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.

6.2. Satellite communication

What about moving data to the place where data analysis can be performed? In general, experimental data collection may need to be repeated many times with different configurations and it would be desirable to move data from the developing country to where we can work on the data. There may not be any infrastructure for the Internet or electricity in rural area. The condition is different from remote sensing. In remote sensing, the small sensor board can be left for a long time and later on those sensors would be collected.

In the TIERs (Technology and Infrastructure in Emerging Regions) project [16], the concept of the Village Base Station (VBTS) is proposed, where the users in the village receive various services ranging from distributed caches and uploading user-generated content as data services [17]. VBTS provides an outdoor PC with a software radio that implements a low-power low-capacity GSM base station including power supply via solar/wind.

There may not be GSM access capability in rural area, and we attempt to integrate satellite communication using Delay Tolerant Networks (DTNs) [18]. DTN research started with Vint Cerf and the Interplanetary Internet initiative [19], which proposed a new architecture that could work over both terrestrial and interplanetary links. This architecture could enable applications such as the remote operation of scientific experiments on other planets, controlled using TCP/IP from Earth.

In contrast, in the Haggle Project [3] we exploited a Pocket-Switched Networks (PSNs) variant of DTN, where people carry devices in their pockets, which communicate directly with other devices within their range or with infrastructure. As people move around, they can exchange messages with nearby devices, carrying a message until it is close to another device. We will build a hybrid DTN using the satellite based communication and PSNs. All the collected data will be gathered up to a node that can communicate with the CubeSat by PSNs and the CubeSat communication node exchange the data when the satellite communication is available.

Aspirations of use of microsats and smallsats for remote sensing applications have been increasing in developing countries. Small startups in the United Kingdom have taken the lead in providing end-to-end, innovative solutions for countries with limited economies. CubeSat type platforms are considered to be well suited for building an early space technology capacity. Low Earth Orbits small satellites, positioned on inclined polar orbit will also be used.

Uploading collected data and receiving instruction do not require the same level of end-to-end steady connection as email or Voice over IP (VoIP) do. High-delay connectivity is acceptable as far as the data collection can be performed within required timeline. We vision that DTN could be one solution to bring communication in the developing countries. For data collection in EpiMap, the DTN concept is ideal. The important point here is that CubeSat needs to implement a DTN protocol. Fig. 6 shows a system overview, where CubeSat moves around the earth and connects to the node, when the CubeSat is reachable by the node. The black-patterned and green[1] (grey) nodes are in a rural area, where a black-patterned node is able to access a CubeSat, while a green (grey) node can only connect locally. The black-patterned nodes can upload/download data to the CubeSat and transfer them to the red (dark) nodes in Europe/USA. The red (dark) nodes will form an overlay to share data. Krupiarz et al. [20] proposed using small SmallSats and DTN for Communication in Developing Countries. As the technology has been progressing, CubeSat will be increasingly becoming popular in the near future and CubeSat with DTN will be beneficial for the communication for the applications of 'disaster warning', 'volunteer aid worker consultation' and 'reporting', 'search and rescue', 'disaster control information', 'weather

[1] For interpretation of colour in Figs. 1–6, the reader is referred to the web version of this article.

Please cite this article in press as: E. Yoneki, J. Crowcroft, EpiMap: Towards quantifying contact networks for understanding epidemiology in developing countries, Ad Hoc Netw. (2012), http://dx.doi.org/10.1016/j.adhoc.2012.06.003

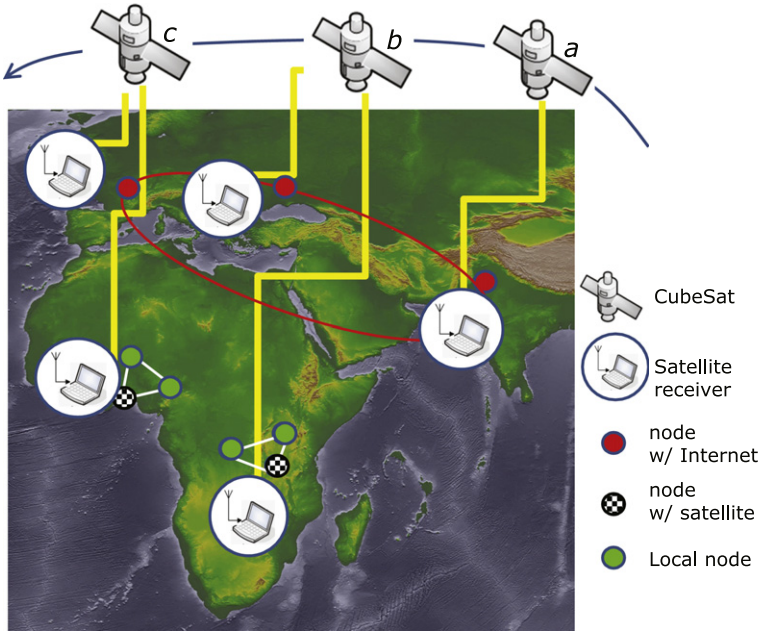


Fig. 6. Use of CubeSat for data collection from Africa.

forecasts and education', and remote sensing data upload/download [21].

SPICE project [22] is planned to build an overlay called 'Space-Data Routers' (SDRs) [23], where space-data generated by a single or multiple missions can be shared among Space Agencies, Academic Institutes and Research Centres in a natural, flexible, secure and automated manners. A communication overlay modelled will be developed according to a thematic context of missions, Ground Segment topological distribution, Agency policies and requirements. A DTN-enabled device that incorporates the Space Agency administrative instructions and resource utilisation and integrates the DTN protocol stack with application, network and link layer protocols.

Potentially we would extend the SDR overlay by integrating CubeSat. In an overlay, some nodes may not have access to the CubeSat but locally it can transmit the data to the node that has access to the CubeSat. This enables the developing countries to upload the collected data also receive information/result of analysis from the data processing unit in Europe/USA. The challenges to realising such a communication platform include storing Tera byte of data, rule based routing, and load balancing.

7. Data analysis strategy

In this section, we describe our strategies for the analysis of contact data. Several researchers have worked on predictive models for epidemics such as an influenza pandemic [24]. Such models require precise information about mobility, interaction, and behavioural assumption of the population. However, interactions between individuals are assumed to follow existing contact models, that do not take into the changeable behaviour of human movement. We aim to build a model incorporating spatial and temporal information 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 [25], a little work has been done in this area. Many other network characteristics (e.g. population size, geographical location) can be uncovered through the EpiMap project. 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. Only considering one-dimensional pair relationships may not be sufficient and consideration of the strength and regularity of connections will be necessary.

In EpiMap, we set several goals of data analysis below:

- Estimate social contact parameters relevant to the spread of close-contact infectious diseases in a number of contexts in a number of different African countries including determination of age-related contact patterns, and comparison between settings (e.g. East versus West Africa, urban versus rural and wet versus dry season).
- Estimate patterns of contact with domestic animals and link with patterns of contact between humans.
- Elucidate risk behaviours by linking epidemiological and social contact data.
- Develop and parameterise mathematical models of a range of infections based on observed social contact data.
- Provide a quantitative description of human–human and human–animal contact patterns.

Please cite this article in press as: E. Yoneki, J. Crowcroft, EpiMap: Towards quantifying contact networks for understanding epidemiology in developing countries, Ad Hoc Netw. (2012), http://dx.doi.org/10.1016/j.adhoc.2012.06.003

We will use regression analyses for understanding characteristics of underlying contact patterns, including age, gender, household size and composition, region, country, season, etc., based on previously derived methods. The derived model can be used to help explain differences in contact patterns observed between different settings or in different seasons and generalise the results to other settings. Data on contact with animals will be combined with data on contacts amongst humans to derive a series of animal–human–human contact matrices by age group and other key variables. These will be used to develop mathematical models designed to assess the emergence of novel pathogens into the human population and how they may spread during the critical first steps.

8. Ethical/privacy/anonymity issues

We are well aware of ethical and privacy issues for the collected data, and the data will never be used to identify individuals. The collected data will be anonymised before analysis. Software developed for sensor devices and mobile phones may involve collaboration between ad hoc groups of members.

There will be various concerns about privacy, surveillance and freedom of action. For example, while providing location information can clearly be a one-way system where the location providing tools do not track the receivers, once a device receives information its location is potentially available to others. We will exploit various methodologies to protect the participants' privacy in the EpiMap project.

9. Summary and future works

We describe our vision of the EpiMap project, where mobile phones and/or sensors record proximity to other devices. This project will gather information on human interactions in rural communities of developing countries in Asia, Africa and South America. The EpiMap project evolved from the FluPhone project in which we deployed data collection of human contact, flu-like symptoms and virtual disease spread using various phones. In EpiMap, we develop a hybrid method for data collection including the use of mobile phones, RFID tags, and various sensors depending on the environment in developing countries. Delay tolerant networking takes an important role to collect data and transmit them from developing countries to Europe. We envision to use CubeSat to move data in a delay tolerant manner and build an overlay network to share the data among the institutions. Within the developing countries, we will build a system for hybrid networking of opportunistic and infrastructure-based networks. Main challenges are the computer networks and power supply in rural villages of developing countries. The collected information will be used to develop improved mathematical models for the spread of infectious diseases, such as measles, TB and pneumococcal diseases. The modelling is complemented by surveys to understand the characteristics of living conditions in such rural villages. The outcome of EpiMap can be used to help design more efficient vaccination strategies and equitable control programmes.

Acknowledgments

This research is part-funded by the EPSRC DDEPI Project, EP/H003959. We acknowledge Damien Fay for his contribution to the FluPhone Project. The EpiMap Project will be a potential collaboration with John Edmunds and Ken Eames at London School of Hygiene and Tropical Medicine, and with Jon Reed at the University of Liverpool. We would like to thank Scott Burleigh at JPL for sharing the insight of DTN CubeSat, and Steven Smith and Karthik Nilakant for valuable comments.

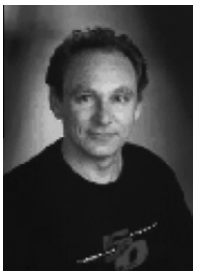
References

[1] FluPhone-Project. <http://www.cl.cam.ac.uk/research/srg/netos/-fluphone2/>.
[2] BBC-News. <http://www.bbc.co.uk/news/uk-england-cambridgeshire-13281131>.
[3] EU FP6 Haggle Project, 2010. <http://www.haggleproject.org>.
[4] FluPhone-Study. <https://www.fluphone.org>.
[5] Facebook. <http://www.facebook.com>.
[6] Twitter. <http://www.twitter.com>.
[7] F. Lynch, M. Zapp, Bubble Rap: Forwarding in Small World DTNs in Ever Decreasing Circles, Tech. Rep. UR-CDL-TR-684, Care of David Lodge University of Rummidge, Cyber Science Lab Euphoric State University, January 2007.
[8] E. Yoneki, P. Hui, S. Chan, J. Crowcroft, A socio-aware overlay for multi-point asynchronous communication in delay tolerant networks, in: Proc. MSWiM, 2007.
[9] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, J. Scott, Impact of human mobility on the design of opportunistic forwarding algorithms, in: Proc. INFOCOM, 2006.
[10] E. Yoneki, J. Crowcroft, Wireless Epidemic Spread in Dynamic Human Networks, Bio-Inspired Computing and Communication LNCS (5151).
[11] M. Freeman, N. Watkins, E. Yoneki, J. Crowcroft, Rhythm and randomness in human contact, in: Proc. International Conference on Advances in Social Networks Analysis and Mining, 2010.
[12] D. Fay, J. Kunegis, E. Yoneki, Uncovering multi-modal spread modes using joint diagonalisation in contact networks, Tech. Rep., UCAM-CL-TR-806, University of Cambridge, 2011.
[13] D. Wyatt, T. Choudhury, J. Tanzeem, J. Kitts, Inferring colocation and conversational networks using privacy-sensitive audio, ACM Transactions on Intelligent Systems and Technology 2 (1) (2011).
[14] J. Stehle, N. Voirin, A. Barrat, C. Cattuto, et al., High-resolution measurements of face-to-face contact patterns in a primary school, PLoS ONE 6 (8) (2011).
[15] J. Stehle, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W.V. den Broeck, C. Regis, B. Lina, P. Vanhems, Simulation of an seir infectious disease model on the dynamic contact network of conference attendees, BMC Medicine 9 (87) (2011).
[16] TIER. <http://tier.cs.berkeley.edu/drupal/>.
[17] K. Heimerl, E. Brewer, The Village Base Station, in: ACM Workshop on Networked Systems for Developing Regions (NSDR), 2010.
[18] K. Fall, A delay-tolerant network architecture for challenged internets, in: Proc. SIGCOMM, 2003.
[19] S. Burleigh, A. Hooke, L. Torgerson, K. Fall, V. Cerf, B. Durst, K. Scott, H. Weiss, Delay-tolerant networking: an approach to interplanetary internet, IEEE Communications Magazine 41 (6) (2003) 128–136.
[20] C. Krupiarz, C. Belleme, D. Gherardi, E. Birrane, Using smallsats and DTN for communication in developing countries, in: Proc. International Astronautical Congress (IAC-08.B4.1.8), 2008.
[21] S. Burleigh, E. Birrane, Toward a communications satellite network for humanitarian relief, in: International Conference on Wireless Technologies for Humanitarian Relief (ACWR), 2011.
[22] SPICE-Project, SPICE:Space Internetworking, 2011. <http://www.spice-center.org/description/>.
[23] SDR-Project, SDR: Space-Data Routers, 2011. <http://www.spacedatarouters.eu/>.

Please cite this article in press as: E. Yoneki, J. Crowcroft, EpiMap: Towards quantifying contact networks for understanding epidemiology in developing countries, Ad Hoc Netw. (2012), http://dx.doi.org/10.1016/j.adhoc.2012.06.003

[24] S. Eubank, H. Guclu, V. Kumar, M. Marathe, a. Srinivasan, Z. Toroczkal, N. Wang, Modelling disease outbreaks in realistic urban social networks, Nature 429 (2004).
[25] M. Barthlemy, A. Barrat, R. Pastor-Satorras, A. Vespignani, Velocity and hierarchical spread of epidemic outbreaks in scale-free networks, Physical Review Letters 92 (2004) 178701.

Eiko Yoneki is an EPSRC Research Fellow at the University of Cambridge Computer Laboratory. She has received her PhD in Computer Science from the University of Cambridge on 'Data Centric Asynchronous Communication'. Her research spans distributed systems, networking, and databases. She leads the EPSRC Data-Driven Epidemiology project, the Data-Driven Declarative Networking with Microsoft Research, and the EU FP7 RECOGNITION, Relevance and cognition for self-awareness in a content-centric Internet. She has also participated in the EU FP6 Haggle project. http://www.cl.cam.ac.uk/~ey204.



Jon Crowcroft is the Marconi Professor of Communications Systems in the Computer Lab, at the University of Cambridge. He is a Fellow of the ACM, of the British Computer Society, of the IEE and the Royal Academy of Engineering and a Fellow of IEEE. He runs the Communications Innovations Institute, UCL and the Oxford Internet Institute economists, engineers, lawyers, social and computer scientists, which seeks to see how the impact of disruptive technologies can be factored into the communications and computing business arena, and comprehended by regulators and other government agencies. Currently Prof Crowcroft holds several grants from EPSRC; Horizon: Digital Economy Hub and INTERNET: Intelligent Energy aware Networks. http://www.cl.cam.ac.uk/~jac22.

Please cite this article in press as: E. Yoneki, J. Crowcroft, EpiMap: Towards quantifying contact networks for understanding epidemiology in developing countries, Ad Hoc Netw. (2012), http://dx.doi.org/10.1016/j.adhoc.2012.06.003