Your Friends Have More Friends Than You Do: Identifying Influential Mobile Users Through Random Walks

Bo Han
Department of Computer Science
University of Maryland
College Park, MD 20742, USA
bohan@cs.umd.edu

Aravind Srinivasan
Department of Computer Science and
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA
srin@cs.umd.edu

ABSTRACT

In this paper, we study the problem of identifying influential users in mobile social networks. Traditional approaches find these users through centralized algorithms on either friendship or social-contact graphs of all users. However, the computational complexity of these algorithms is known to be very high, making them unsuitable for large-scale networks. We propose a *lightweight* and *distributed* protocol, `iWander`, to identify influential users through fixed-length *random walks*. To the best of our knowledge, we are the first to design a distributed protocol on smartphones that leverages random walks for identifying influential mobile users, although this technique has been used in other areas.

The most attractive feature of `iWander` is its *extremely low* message overhead, which lends itself well to mobile applications. We evaluate the performance of `iWander` for two applications, targeted immunization of infectious diseases and target-set selection for information dissemination. Through extensive simulation studies using a real-world mobility trace, we demonstrate that targeted immunization using `iWander` achieves a comparable performance with a degree-based immunization policy that vaccinates users with large number of contacts first, while consuming only less than 1% of this policy's message overhead. We also show that target-set selection based on `iWander` outperforms the random and degree-based target-set selections for information dissemination in several scenarios.

Categories and Subject Descriptors

C.2.1 [**Computer Communication Network**]: Network Architecture and Design—*Wireless communication*

General Terms

Design, Performance

Keywords

Influential mobile users, random walks, distributed protocol, disease control, information dissemination.

1. INTRODUCTION

Mobile social networks, under the merging of social networks that link humans and Internet that connects computers [18], have emerged as a new frontier in the mobile computing research community. Mobile social networking is social networking where mobile users interact, communicate and connect with each other using their wireless devices. There have been several novel mobile social applications developed recently (e.g., Micro-Blog [14], PeopleNet [21], and SociableSense [29]). Meanwhile, more and more native mobile social networking services, such as Loopt and Foursquare, have been created.

Not all mobile users are equal in terms of mobility. Some of them, such as salespeople, may travel to many places during a day, while others, such as graduate students, may stay in their office for most of the working time. When considering the problem of information dissemination in mobile networks, if we employ these active salespeople as the initial physical carriers, they may be able to further propagate information to a much larger fraction of mobile users, compared with selecting initial carriers randomly. This is exactly the rationale behind the influence maximization problem of information diffusion in traditional social networks [10, 17]. Similarly, if we monitor these critical individuals, we may be able to detect the outbreaks of infectious diseases much earlier, for example, during the flu season [4].

In this paper, we address the following question: *how do we identify influential users in mobile social networks through distributed solutions with low message overhead?* Influential users are individuals with high centrality in their social-contact graphs. In our previous work [16], we propose a heuristic algorithm to select influential mobile users for information dissemination, which is an extension of the greedy algorithm by Kempe, Kleinberg, and Tardos [17]. Nguyen et al. [23] propose to find these users through the detection of overlapping community structures in dynamic networks. However, these solutions are all *centralized* and require the complete social-contact graphs.

There are two major challenges when finding these critical mobile users. First, given the large size of mobile social networks, the proposed solutions must be distributed. Besides the drawback of requiring complete contact graphs, centralized schemes are known to have high computational com-

MobiHoc'12, June 11–14, 2012, Hilton Head Island, SC, USA.

plexity, especially on large social graphs. For example, as reported by Chen et al. [3], finding a small set of nodes with high centrality in a graph with 15,000 vertices could take days on a modern server machine. Second, because these distributed protocols usually run on battery-supported mobile devices, such as smartphones[1], we need to control their communication overhead, as data transmission is the major source of smartphone energy consumption.

Our approach is motivated by the "friendship paradox" [13] that "*your friends have more friends than you do*" and leverages random walks to identify critical users. The reason behind this paradox is that people with larger numbers of friends may have a high probability of being observed among one's friend circle. Thus, the friends of randomly selected individuals may have higher centrality in friendship graphs than average. Although the original proof in Feld [13] is for the static friendship graph of traditional social networks, we can easily extend it for the dynamic contact graph of mobile social networks.

This paper makes the following contributions.

- We design a distributed and lightweight protocol, called `iWander`, to identify critical individuals in mobile social networks (Section 3). With the design principle of decreasing message overhead, `iWander` can significantly reduce the energy consumption on smartphones. We assume that everyone in the examined network has a smartphone that runs `iWander` in the background. The key idea behind `iWander` is to perform fixed-length random walks periodically by a small group of smartphones and estimate the centrality of individuals through their random-walk counters (i.e., the number of times their smartphones are visited by random walks). To verify the feasibility of `iWander`, we also implement a proof-of-concept prototype on Nokia N900 smartphones.
- We present a targeted immunization policy based on the centrality information provided by `iWander` to contain the spread of infectious diseases (Section 4). We evaluate the performance of our proposed random-walk based immunization through extensive simulation studies. The simulation results from a real-world mobility trace show that random-walk based immunization always outperforms random immunization and performs very close to degree-based immunization with less than 1% of its message overhead. The results also demonstrate that selecting monitors based on `iWander` can offer early outbreak detection of infectious diseases.
- We show how to benefit from `iWander` for information dissemination in mobile social networks (Section 5). Specifically, we study the target-set selection problem which chooses target users based on the random-walk counters of mobile users provided by `iWander`. Surprisingly, we find that differently from targeted immunization, if we choose all target users with high centrality, the resultant scheme only outperforms random selection for small target sets. We also propose another enhanced scheme that chooses both influential and non-influential users into the target set. Our simulation results verify that this enhanced scheme outperforms random selection for large target sets.

2. RELATED WORK

In this section, we review related work about identifying influential users in social networks, applications of random walks, infectious disease control for public health and information dissemination in mobile networks.

2.1 Identifying Influential Users

2.1.1 Traditional Social Networks

Identifying influential users has been extensively studied for information diffusion in traditional social networks [10, 17, 30]. Domingos and Richardson [10, 30] were the first to introduce a fundamental algorithmic problem of information diffusion: what is the initial target set of k users, if we want to maximize the propagation of information in a social network? Kempe et al. [17] prove that the information dissemination function of this influence maximization problem is submodular for several classes of models. They also propose a greedy algorithm that outperforms heuristics based on node centrality and distance centrality. Although the greedy algorithm of Kempe et al. [17] achieves the best known result so far with a provable approximation ratio of $(1 - 1/e)$, it is computationally expensive [3]. To solve this problem, Chen et al. [3] propose an improvement to reduce the algorithm's running time.

2.1.2 Mobile Networks

The problem of influence maximization has also been extended to mobile networks. Previously, we have studied the target-set selection problem for information delivery as the first step toward bootstrapping mobile data offloading [16]. In particular, we investigate how to select a target set with only k users among all subscribed users, such that we can maximize the number of users that receive the delivered information through mobile-to-mobile opportunistic communications. We propose a heuristic algorithm to select the target set by exploring the regularity of human mobility. Nguyen et al. [23] propose to select critical nodes through overlapping community detection in dynamic networks. They present a framework to adaptively update the community structure based on history information. They also show that this framework can improve the performance of forwarding protocols in delay-tolerant networks and schemes to contain worms in online social networks.

2.1.3 Targeted immunization

Targeted immunization has been proposed to eradicate infections for scale-free complex networks [8], by considering the heterogeneous connectivity properties of these networks. The idea is to cure the highly connected nodes, the hubs, to restore the epidemic threshold in the diffusion process. Christakis and Fowler [4] propose a mechanism for detecting contagious outbreaks. They recruited 390 Harvard College students to participate in an experiment and asked them to nominate up to three friends. Their work demonstrates that by monitoring only the friends of these randomly selected students they can provide an early detection of flu by up to 13.9 days. Christley et al. [5] evaluate the performance of several network centrality measures for identifying high-risk individuals, including degree, shortest-path betweenness and random-walk betweenness. Their results show that degree performs very close to other network measures in predicting risk of infection.

[1] We focus on smartphones, the most popular mobile devices, in this paper.

Remark: All the above approaches for various problems, ranging from influence maximization to targeted immunization, are based on *centralized* solutions. Motivated by the friendship paradox, we leverage smartphones to perform random walks among mobile users during their contacts and design a distributed lightweight protocol to identify the most influential individuals.

2.2 Random Walks

The term random walk was first introduced by Karl Pearson [27]. We are interested in random walks on graphs, where a walker starts from a source node to a destination node and for each step of this travel, the next node to visit is selected uniformly at random from the neighbor-set of the current node.

Random walks have been integrated into centrality measurement of social science. For instance, Newman [22] proposes the random-walk betweenness centrality, a relaxation of the shortest-path betweenness. This measure defines how often a node in a graph is visited by a random walker between *all* possible node pairs. Noh and Rieger [24] introduce the random-walk closeness centrality metric, which measures how fast a node can receive a random-walk message from other nodes in the network.

Random walks have also been widely explored in other fields, such as computer science, economics, biology and psychology, for various purposes. For example, Braginsky and Estrin [1] route queries on a random walk to sensor nodes around which a particular event occurs. Yu et al. [35] propose SybilGuard which uses a special kind of random walk, where every node chooses the next hop based on a pre-computed random permutation, to limit the bad effect of sybil attacks on peer-to-peer systems.

2.3 Infectious Disease Control

Public-health researchers have developed tools to study the spreading patterns of infectious diseases and mitigate the effects of epidemics. Eubank et al. [12] model the outbreaks of infectious disease in urban social networks. They find that the contact network is a small-world graph, but the locations graph is scale-free which enables efficient outbreak detection that places sensors in the hubs of the graph. They evaluate the performance of several vaccination strategies for smallpox using a realistic large-scale simulation framework. The results show that it is possible to contain outbreaks through a combination of targeted vaccination and early detection.

Recently, researchers started to measure human contact networks for infectious disease transmission using mobile devices, such as sensor motes or RFID badges. Salathé et al. [31] measure the close proximity interactions among 788 individuals at an American high school during a typical day. Through trace-driven simulation studies, they show that targeted immunization using the contact-network data is more effective than random immunization. Stehlé et al. [33] report a similar study in a primary school in French which measured face-to-face proximity of 6-12 years children and teachers. Based on the measurement results, they provide several public-health implications of infectious diseases, for example, closing selected classes instead of the whole school.

In this paper, we propose `iWander` to make it feasible to identify influential individuals for targeted immunization in a distributed way, through fixed-length random walks.

2.4 Mobile Information Dissemination

Information dissemination is an important application of mobile networks. Papadopouli and Schulzrinne [25] propose 7DS, a peer-to-peer information dissemination system, to increase data availability for mobile users with intermittent connectivity. With 7DS, mobile devices query data from neighboring peers when they fail to access Internet with their own connections. Small and Haas [32] propose a networking model, called the Shared Wireless Infostation Model (SWIM), which allows information to travel within a network using mobile users as physical carriers. They demonstrate the effectiveness of SWIM using a practical information system of radio-tagged whales.

McNamara et al. [20] propose a scheme to choose the best sources (peers who can remain co-located long enough to complete data transfer) for content sharing among co-located mobile users in urban transport. Using three different experimental traces, Zyba et al. [36] study fundamental properties of human interactions that may affect the performance of information dissemination in mobile ad-hoc networks. They find that the efficiency of content distribution depends on not only the devices' social status, but also the number and density of devices.

In this paper, we propose to identify critical mobile users for facilitating information dissemination through distributed random walks.

3. THE RANDOM WALKS PROTOCOL

In this section, we present the detail of `iWander` design and its proof-of-concept prototype implementation.

3.1 The Protocol

We propose to leverage *random walks* to design a distributed protocol, `iWander`, for identifying influential users in mobile social networks. The intuition is that if we periodically initialize random walks from a small group of smartphones, influential mobile users may be visited by these random walks more frequently than average.

The proposed `iWander` protocol works as follows. Every ΔT hours, `iWander` generates a tiny probing message with a given probability q on each smartphone. The message contains *only* a pre-configured time-to-live (TTL) field L. During the contacts of a smartphone with its peers, if it has a probing message in its local queue, it sends this message to another randomly selected peer. When a smartphone receives a probing message, it decreases L in the message by 1, and then stores it in its local queue, waiting for the opportunity to forward the message to other peers. A probing message with $T = 0$ will be finally discarded. `iWander` maintains a random-walk counter on each smartphone, initialized to zero, to record how many times it has received the probing messages (i.e., visited by these random walks).

After collecting the random-walk counters from all users recorded by their smartphones, we can determine the set of k critical users from the head of the user list sorted by these counters. The reason is that based on the friendship paradox, influential users have high probabilities to be visited by random walks and thus own large random-walk counters. Differently from the random-walk betweenness metric proposed by Newman [22], `iWander` applies *fixed-length* instead of *all-pairs* random walks for two reasons. First, in practice, it is difficult for a mobile user to know every other user and thus specify random-walk destinations. Second, the message

	discovery	idle
Bluetooth	253.05 (5.51)	16.54 (1.11)
WiFi	836.65 (8.98)	791.02 (5.23)

Table 1: The power level of Bluetooth and WiFi on Nokia N900 during discovery and idle modes (in mW).

overhead of all-pairs random walks may be much higher than fixed-length random walks, which makes them unsuitable for battery-powered smartphones.

The update and reset of random-walk counters are determined by the upper layer applications. In practice, they may reset these counters periodically, for example, at midnight (12:00 AM) of every day. They can also apply an exponential moving average to update these counters by assigning a higher weight to recent counters.

In summary, the performance of iWander relies on three parameters: q – the probability that a smartphone generates a probing message (i.e., the fraction of mobile users that initialize random walks), L – the length of random walks (i.e., the number of mobile users visited by a single random walk), and ΔT – the frequency of generating new random-walk probing messages. It is important to understand their impact on the performance of iWander, because they determine both the quality of identified influential users and the number of probing messages spreading over the network.

To reduce energy consumption on smartphones, we prefer short random walks with only a few steps. "Static" versions of social-contact networks are often very dense and "expander-like" (see, e.g., Eubank et al. [12]). In such highly-mixing networks, it is well-known that a random walk of length $O(\log n)$ or less, where n is the number of nodes in the network, suffices to come very close to the stationary distribution of the random walk (in which each vertex has a probability proportional to its degree). Our networks are inherently mobile and thus not static, but their static snapshots will likely be expander-like. The mobile networks will also likely mix well, serving to explain intriguing results such as those of Grossglauser and Tse [15]. Thus, the short random walks that we take will likely come quite close to sampling vertices approximately according to their degrees.

In Section 4.2, we show how the length of random walks L affects the performance of iWander through trace-driven simulation studies. We also evaluate the performance of iWander with different probabilities (q) and frequencies (ΔT) of the generation of random-walk probing messages. We leave the theoretical analysis of the optimal values for these parameters as our future work.

3.2 Proof of Concept

To demonstrate the feasibility of iWander, we implement a prototype in C language on Nokia N900 smartphones. We choose Bluetooth as the underlying communication protocol for iWander due to its low energy consumption. We measured the power of discovery and idle modes of Bluetooth and WiFi devices and summarize the average results and standard deviations for 10 runs in Table 1, which shows that in Bluetooth discovery mode the power of N900 is less than 1/3 of WiFi discovery. Moreover, when the Bluetooth device is in idle mode, the power of N900 is negligible. The reason for high power of WiFi idle mode is that to enable device discovery, a WiFi device needs to run in ad-hoc mode and sends out Beacon messages periodically. Given that the power of WiFi idle mode is also higher than that of Bluetooth discovery mode, no matter what the duration of device discovery is, the energy consumption of WiFi discovery will be higher than that of Bluetooth discovery.

Due to the simplicity of iWander design, its prototype implementation using the BlueZ[2] protocol stack has less than 300 lines of code and the size of the compiled file is only around 32 kB, which means that we can easily deploy it on a variety of mobile devices. Unfortunately, it is hard to evaluate the performance of iWander in practice because it is difficult to recruit a large number of participants. In the next two sections, we present two applications of iWander, targeted immunization of infectious diseases and target-set selection for information dissemination, and evaluate their performance through trace-driven simulation studies using a real-world mobility trace.

4. CONTROLLING INFECTIOUS DISEASES

In this section, we demonstrate how to leverage the critical individuals identified by iWander to control infectious diseases and perform early outbreak detection.

4.1 Random-Walk Based Immunization

Smartphones and Internet-related technologies have recently been used to collect data pertaining to the behavior of individuals for various purposes, including disease control and health care. For example, the FluPhone[3] study collects information on social encounters in Cambridge, UK using mobile phones, with the goal of helping medical researchers to better understand the propagation of close-contact infections. Pollak et al. [28] design a mobile phone based game to motivate children to practice healthy eating habits. Moreover, Cook et al. [7] propose Google Flu Trends which uses aggregated Google search queries to provide near-real time estimates of the level of flu in 121 cities of the US.

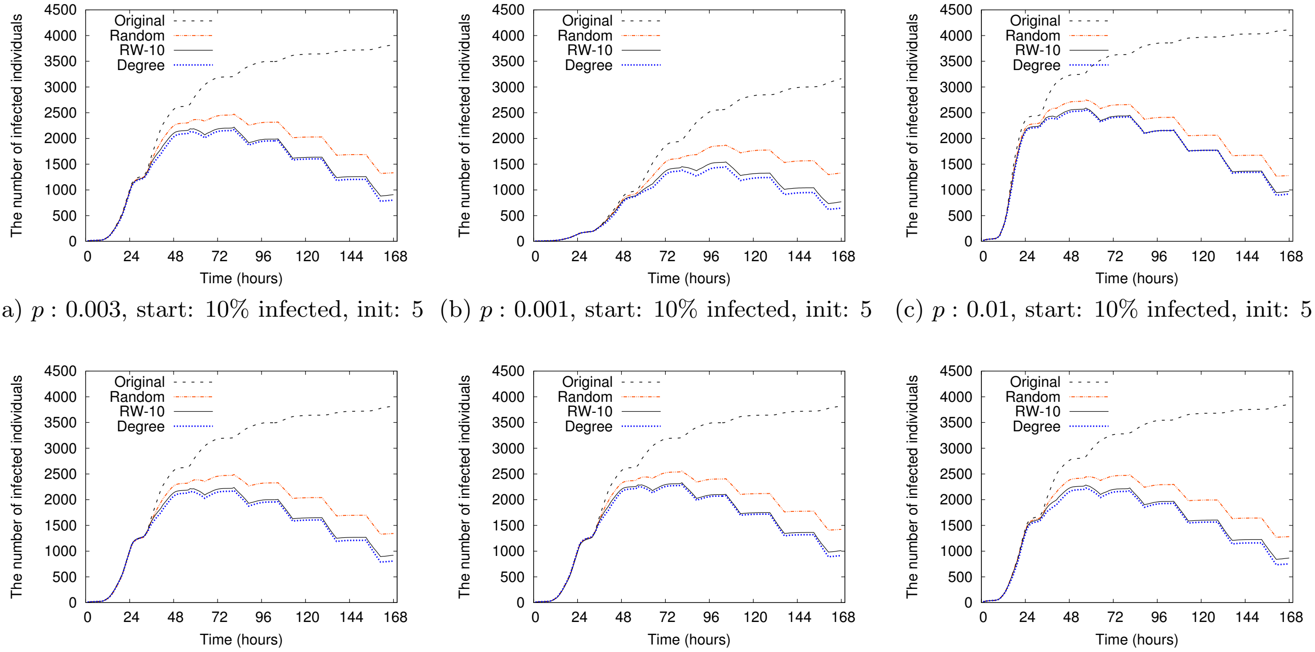
We propose to perform targeted immunization of infectious diseases based on the random-walk counters maintained by iWander. For example, during the flu season, iWander can periodically report these counters on the smartphones of college students to the university health center. The medical staff can then vaccinate students with high random-walk counters first to contain the spread of flu. We can also use these counters to detect the outbreaks of infectious diseases, where the medical staff monitor the health condition of students with high counters instead of randomly selected students.

The centralized collection of random-walk counters is required by this specific application and the target-set selection for mobile information dissemination in Section 5. For other applications, such as distribution of self-generated content among users, it is possible to extend iWander and design a fully distributed protocol to compute and disseminate these counters among mobile users, for example, by leveraging diffusing computations [9].

There are several differences between our proposed targeted immunization scheme and those in the literature, for example, by Christakis and Fowler [4] and Christley et al. [5]. First, our scheme can benefit from the social contacts detected directly by smartphones, instead of using the estimation through friendship graphs generated from surveys [4].

[2] The default Bluetooth protocol stack of most Linux distributions, http://www.bluez.org/

[3] https://www.fluphone.org/



(a) $p : 0.003$, start: 10% infected, init: 5 (b) $p : 0.001$, start: 10% infected, init: 5 (c) $p : 0.01$, start: 10% infected, init: 5

(d) $p : 0.003$, start: 24 hours, init: 5 (e) $p : 0.003$, start: 30% infected, init: 5 (f) $p : 0.003$, start: 10% infected, init: 10

Figure 1: Comparison of the evolution of infected individuals for three immunization policies, random, degree-based, and random-walk-based, with different infection probabilities, immunization start conditions, and initial infections.

Second, our scheme can reflect the dynamics of social contacts in a timely way and avoid the computation-extensive centralized data analysis. Finally, our fixed-length random-walk metric is an extension of the general all-pairs random-walk betweenness centrality [22] and the one-step diffusion-style estimation of node centrality [4], and its low message overhead makes it amenable to run on smartphones.

4.2 Performance Evaluation

We evaluate the performance of `iWander` for infectious disease control through extensive trace-driven simulations.

4.2.1 Simulation Setup

We implement a simulator in C based on the SIR model, to simulate the spread of infectious diseases. Each individual can be in one of three states: susceptible, infectious, and recovered. Initially, all individuals are in the susceptible state. At the beginning of the simulation, we randomly select a small group of individuals and set their status to be infectious. Transmission of disease occurs from an infectious to a susceptible individual with a probability of p per 60-second contact. Thus, the probability of disease transmission from an infectious individual to a susceptible individual, co-located for t seconds, is $1-(1-p)^{\lfloor t/60 \rfloor}$. Finally, an infectious individual is recovered from the disease if he or she is vaccinated.

To simulate the social contacts of individuals, we use a real-world mobility trace, the Dartmouth data set [19], which records at WiFi access points the association and disassociation events of wireless devices. We use a one-week trace of this data set, from 2004-03-01 to 2004-03-07, which includes 4522 devices. As in many previous studies that use this kind of data set, for example in Zyba et al. [36], we consider that the owners of wireless devices are in "social contacts" if their devices are associated with the same access point. We note that although the Dartmouth data set is based on WiFi association data, the user mobility derived from it is for general purpose and has been widely used in the literature [2, 6, 34, 36].

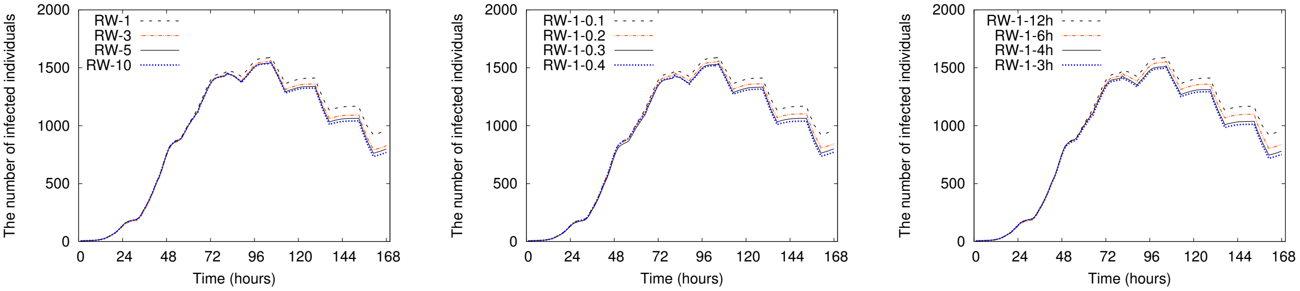
The main reason we chose the Dartmouth data set is that it involves a large number of mobile users, although this data set has its own limitations. For example, the user mobility derived from WiFi association events may not be complete (only around WiFi APs). There are some other publicly available data sets, such as the Reality Mining data set of mobile phone users [11] and the Cabspotting traces of San Francisco's taxi cabs[4]. However, compared to them which either is too small (e.g., the Reality Mining data set with only less than 100 users) or cannot represent the human mobility (e.g., the traces of cabs), we believe the Dartmouth data set is more suitable for our purpose.

For all figures presented in this section, we run the simulation 1,000 times to get average values and standard deviations. For the sake of clarity, we plot standard deviations only in Figure 3 of message overhead.

4.2.2 Targeted Immunization

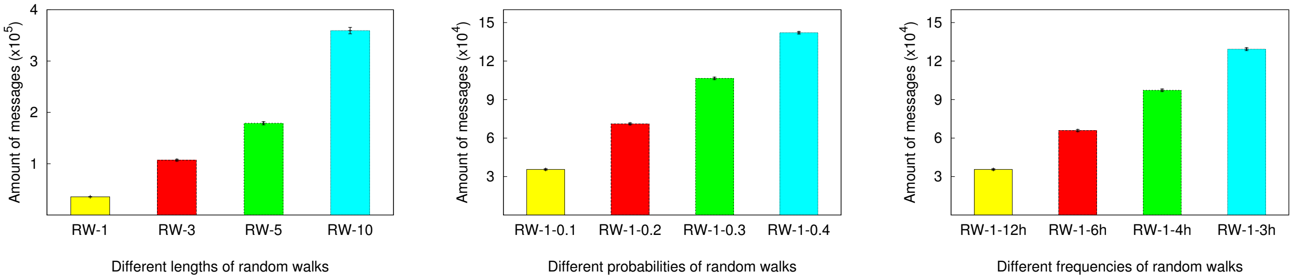
We compare the performance of random-walk based immunization with random immunization, `Random`, and degree-based immunization, `Degree`. With `Random`, the medical staff vaccinate college students randomly. Using `Degree`, the smartphone attached with a student performs device discovery every 60 seconds to record the number of smartphones it has contacts with (i.e., node degree in the aggregated social-contact graphs). Then the medical staff vaccinate students with large number of contacts first. During random-walk

[4] `http://cabspotting.org/`



(a) different lengths of random walks (b) different prob. of random walks (c) different frequencies of random walks

Figure 2: Comparison of random-walk based immunizations with different lengths, probabilities and frequencies.



(a) q: 0.1, ΔT: 12 hours (b) ΔT: 12 hours (c) q: 0.1

Figure 3: Comparison of the amount of messages for different lengths, probabilities and frequencies of random walks. The number of messages for the degree-based scheme is $1.26{\times}10^8$.

based immunization, `iWander` also performs device discovery every 60 seconds only when the message queues on smartphones are not empty. Finally, we assume that vaccinations happen only during the day time, from 9:00AM to 5:00PM, and that on average 60 students are vaccinated every hour.

There are two reasons why we chose degree-based immunization for comparison. First, Christley et al. [5] report that for the networks they examined, degree performs at least as good as other network centrality metrics, such as shortest-path or random-walk betweenness, in predicting risk of infection. Second, it can be easily implemented in a distributed way. For example, Pásztor et al. [26] propose a selective reprogramming mechanism for sensor networks, which determines target sensor nodes using the results of distributed community detection based on node degrees.

For the random-walk based and degree-based immunizations, we update the medical staff with the latest random-walk counters and the number of contacts of all students every 12 hours. Smartphones can send this information to a centralized server through cellular networks. This message overhead should be low, because it contains only a number and two bytes should be enough for the most of the cases. During the immunizations, the medical staff use the most recent information to get a sorted list of all students and then select from this list the student to be vaccinated for the next minute.

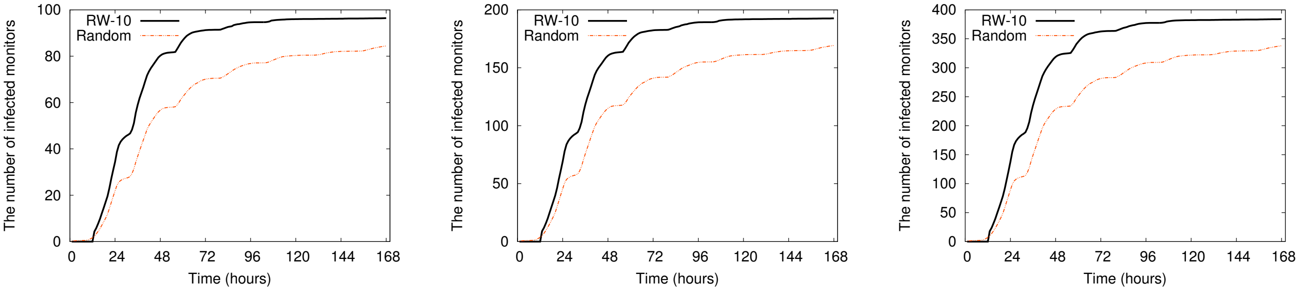
We plot the evolution of the number of infected individuals during the one-week simulated period in Figure 1 for various immunization policies, with different infection probabilities, immunization start conditions, and initial infections. During the outbreak of an infectious disease, we assume that the medical staff start immunizations under two conditions: (1). they have an estimation of the percent of infected individuals and start immunizations after a certain percentage of students are infected; (2). the medical staff start immunizations after a certain amount of time, say 24 hours.

In Figure 1, `Original` plots the curves without immunization as the baseline. As we can see from these subfigures, the number of infected individuals increases much more slowly from the midnight till the morning, compared with other periods in a day, mainly because college students move less frequently during that time period. It is true especially for the first 2 or 3 days, when a large number of students get infected. In all figures of this paper, `RW-`n plots the curves for random walks with n steps.

Among these 6 subfigures, Figures 1a, 1b, and 1c plot the number of infected individuals with different infection probabilities, 0.003, 0.001 and 0.01, 5 initial infections and immunizations after 10% of students are infected. Figures 1d and 1e plot the cases for immunizations after 24 hours and 30% of infections with 0.003 infection probability and 5 initial infections. Finally, Figure 1f plots the case with 0.003 infection probability, 10 initial infections and immunizations after 10% infections. In all these 6 subfigures, `RW-10` performs very close to `Degree` and they all outperform `Random`. Compared to `Random`, the improvement of `RW-10` ranges from 14.10% (Figure 1c) to 25.36% (Figure 1b).

4.2.3 Effects of Various Random-Walk Parameters

We also evaluate the performance of random-walk based immunization with different lengths, probabilities and frequencies of random walks, and plot the simulation results



(a) 100 monitors (b) 200 monitors (c) 400 monitors

Figure 4: Comparison of early detection of outbreaks with randomly selected monitors and those selected using RW-10.

in Figures 2a, 2b and 2c. All the curves in Figure 2 show the number of infected individuals under random-walk based immunization with 0.001 infection probability, 5 initial infections and immunizations after 10% infections. As we can see from these 3 subfigures, we can improve the performance of random-walk based immunization when increasing the length of random walks from 1 to 10, increasing the probability from 0.1 to 0.4, or increasing the frequency from once every 12 hours to 3 hours. However, we achieve these improvements at the expense of higher message overhead.

We plot the message overhead of iWander with different lengths, probabilities and frequencies of random walks in Figures 3a, 3b and 3c. There are three types of messages, probing request and probing response messages for device discovery, and random-walk probing messages for iWander. In all these subfigures, the baseline is iWander with 1-step random walks and smartphones generate random-walk messages with probability 0.1 every 12 hours. For Degree, all messages are generated during device discovery and the total number of messages is 1.26×10^8 for the simulated period. The amount of messages generated by iWander is extremely low, less than 1% of Degree (3.6×10^5 for RW-10 in Figure 3a and less than 1.5×10^5 for RW-1-0.4 in Figure 3b and RW-1-3h in Figure 3c).

4.2.4 Early Detection of Outbreaks

We can also benefit from iWander for early outbreak detection, which is important to control the spread of infectious diseases [4, 12]. We investigate how to choose a subset of students whose health conditions are monitored to provide early detection, similar to the approach in Christakis and Fowler [4]. Motivated by the observation that monitoring a sample of individuals with high centrality in social-contact networks could allow early detection of contagious outbreaks before they happen in the whole population [4], we propose to choose monitors based on the random-walk counters maintained by iWander.

We plot the evolution of the number of infected monitors chosen randomly and based on iWander in Figures 4a, 4b and 4c with 100, 200, and 400 monitors. In this scenario, the infection probability is 0.003 and there are 5 initial infections. Smartphones generate random-walk messages with probability 0.1 every hour. The medical staff choose a group of monitors based on the random-walk counters reported at the noon of 2004-03-01. These subfigures confirm that iWander does offer early outbreak detection, compared with the random selection scheme. For example, if we draw the conclusion that an outbreak is occurring when 60% of the monitors are infected, we can detect the outbreak around 21 hours earlier.

5. FACILITATING INFORMATION DISSEMINATION

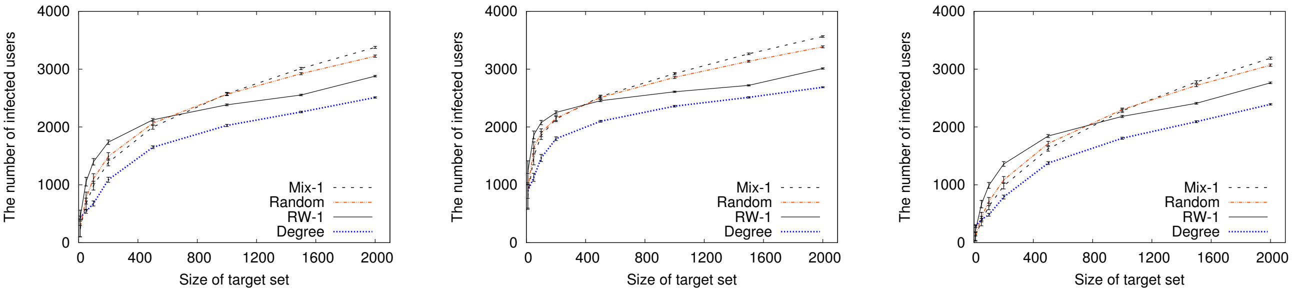
In this section, we illustrate how to benefit from iWander for target-set selection of information dissemination.

5.1 Target-Set Selection Using Random Walks

Motivated by the importance of influence maximization in traditional social networks, in our previous work we study the target-set selection problem for information dissemination in mobile social networks [16]. We leverage opportunistic communications and social participation to facilitate information dissemination and thus reduce the amount of data traffic over 3G networks. We also propose a centralized heuristic algorithm based on the regularity of human mobility, which requires the complete social-contact graph of a given time period and shares the same computational inefficiency as the original greedy algorithm by Kempe, Kleinberg, and Tardos [17].

In this paper, we leverage the random-walk counters of iWander to select target users without requiring global network structure and thus design a distributed solution for the target-set selection problem. Smartphones attached with mobile users run iWander in the background and periodically report their random-walk counters to a centralized server of information service providers. The providers then sort all users based on these counters and then choose the top-k users into the target set. In this scenario mobile users not in the target set can also help to propagate information once they receive it from either target users or others.

The process of information dissemination in mobile social networks is mainly determined by user behaviors. Usually, smartphones can start the exchange of information after they know each other through periodic device discovery. A key concept in the target-set selection problem is the *information dissemination probability* and it is defined as the probability p that information propagates among mobile users after each device discovery. The value of p may be affected by several factors, including status of mobile users and their privacy concerns. Mobile users with high levels of privacy concerns or those who are very busy with their work may have a low probability to involve in information dissemination process. Similar to the transmission of in-



(a) p: 0.01

(b) p: 0.05

(c) p: 0.005

Figure 5: Comparison of the number of infected users for four target-set selection schemes with different values of p.

fectious diseases, given the value of p, the probability that two mobile users with a 60-second device discovery interval can exchange information during a t-second contact is $1-(1-p)^{\lfloor t/60 \rfloor}$.

We note that the purpose of target-set selection for mobile information dissemination is different from targeted immunization, although the usage of random-walk counters is similar in these two applications. For targeted immunization, we want to vaccinate all influential individuals as early as possible. For target-set selection, if two influential mobile users spend most of their time together, a good choice may be selecting only one of them into the target set. Moreover, as we will show in Section 5.2.2, adding non-influential users into the target set can increase the number of infected users for large target sets.

5.2 Performance Evaluation

We develop another trace-driven simulator also in C, using the same Dartmouth data set [19], to evaluate the performance of random-walk based target-set selection. In this simulator, we assume that the underlying wireless communication is reliable. We have measured the performance of Bluetooth-based opportunistic communications on Nokia N900 smartphones, such as the device discovery probability [16]. We are currently working on a packet-level simulator to take into account the low layer issues, including the failure of random-walk probing messages and the transmission of data packets in information dissemination.

5.2.1 Simulation Setup

The simulator first generates the contacts trace of mobile users under the same assumption that they are in contacts if their wireless devices are associated with the same access point. It then replays the contact events for the given information dissemination period, from 6:00PM to 10:00PM on 2004-03-01.[5] Based on the pre-configured information dissemination probability, the simulator determines randomly whether a user can receive information from peers after each device discovery. We also call the users that can receive information before delivery deadline *infected users*. Usually, information providers will send information to uninfected users at the end of dissemination period, to guarantee that every user can finally receive the delivered information [16].

We compare the performance of random-walk based target-set selection, `RW-1`, with random selection, `Random`, and the degree-based selection, `Degree`. The interval of device discovery is 60 seconds, which means that smartphones have the chance to start the exchange of information every 60 seconds. Similar to degree-based immunization, `Degree` also uses the number of other smartphones that a smartphone has contacts with as the metric to select target users. For `RW-1`, smartphones generate 1-step random-walk messages of `iWander` with probability 0.1 every hour. `RW-1` and `Degree` choose target users based on the updated random-walk counters and the number of contacts of smartphones at the beginning of information dissemination period.

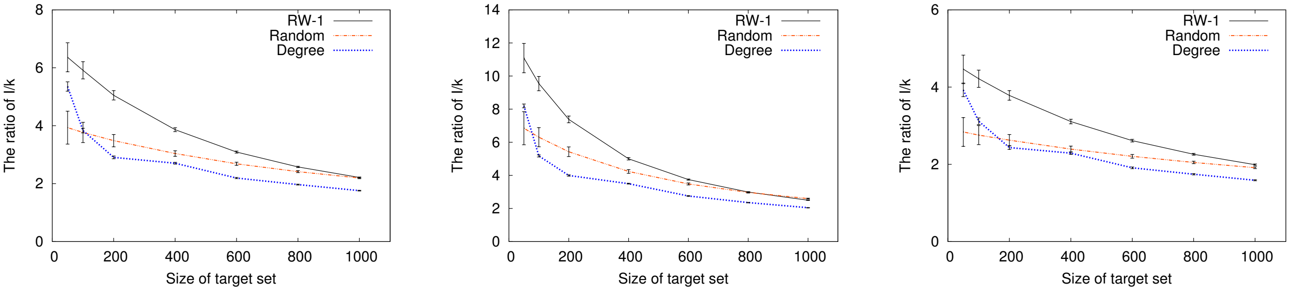
5.2.2 The Number of Infected Users

We plot the number of infected users I for `RW-1`, `Random` and `Degree` in Figure 5. Suppose the number of subscribed users is n, the amount of reduced mobile data traffic will be $n-(k+(n-I)) = I-k$ [16]. We run the simulation 1,000 times and report the average values with standard deviations. The information dissemination probability p is 0.01, 0.05 and 0.005 for Figures 5a, 5b and 5c. We vary the size of target set from 10 to 2,000. As we can see from these subfigures, `RW-1` and `Random` outperform `Degree` when the size of target set is larger than 10. `RW-1` performs better than `Random` for small target sets. For example, for a target set with 50 users, `RW-1` can deliver information to 51% more users than `Random` (667 vs. 441) when p is 0.005. The improvement is 37% when p is 0.01 (1054 vs. 772) and 14% when p is 0.05 (1863 vs. 1639).

The performance of `RW-1` becomes worse than `Random` when the size of target set is larger than 1,000. One of the possible reasons is that non-influential users (i.e., users with low centrality in social-contact networks) also play an important role in information dissemination. These users are called vagabonds in Zyba et al. [36], which demonstrates that under certain circumstances the effectiveness of information dissemination in mobile social networks predominantly depends on the number of vagabonds. When the size of target set is large, `Random` has a higher probability to select more vagabonds into the target set, who may have very little chance to receive information before delivery deadline. However, `Degree` and `RW-1` select only mobile users with high centrality into the target set and ignore these vagabonds.

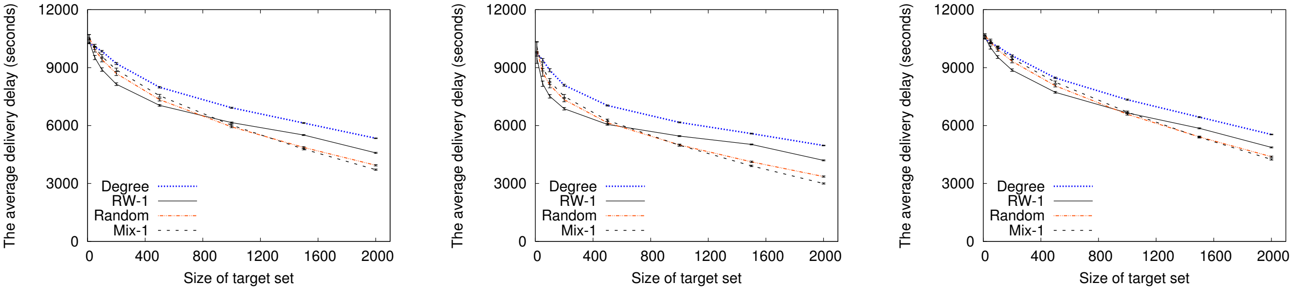
To verify this possible reason, we modify `RW-1` by selecting 90% of target users with low centrality from the end of the user list sorted by random-walk counters. We call this en-

[5] We have also evaluated other information dissemination periods with different durations and got similar results with those presented in this paper.



(a) p: 0.01 (b) p: 0.05 (c) p: 0.005

Figure 6: Comparison of the ratio between the number of infected users I and the size of target set k for three target-set selection schemes with different values of p. Only target users can propagate information to others.



(a) p: 0.01 (b) p: 0.05 (c) p: 0.005

Figure 7: Comparison of delivery delay for 4 target-set selection schemes with different values of p.

hanced scheme Mix-1, which also uses 1-step random walks. The three subfigures in Figure 5 show clearly that Mix-1 outperforms Random for large target sets. We tried other different percentages of non-influential target users and these variations perform very close to each other.

We also evaluate the performance of these schemes for another scenario where only target users are willing to propagate information to others. We show the results of only RW-1, Random, and Degree with k ranging from 50 to 1,000 in Figure 6 for clarity. These subfigures plot the ratio between the number of infected users, I and k for different target-set sizes. In this uncooperative scenario, RW-1 performs much better than Random and Degree for small target sets. For example, when $p = 0.05$ and $k = 100$, the improvement of this ratio is 51% and 85% compared with Random, and Degree. However, for the cooperative scenario, the improvement under the same condition in Figure 5 is only 9% (Random) and 42% (Degree). For large target sets, Random performs very close to RW-1 because in these cases Random has more chances to select influential mobile users into a target set.

Differently from targeted immunization, increasing the values of q, L, or ΔT has limited impact on the performance of random-walk based target-set selection. We omit these results due to the limited space.

5.2.3 Delivery Delay

We finally compare the delivery delay of these four target-set selection schemes for the cooperative scenario. We set the delivery delay of target users to be 0 and the users who cannot receive information before delivery deadline to be 10,800 seconds, the same as the duration of information dissemination period. We plot the delivery delay for different information dissemination probabilities in Figure 7. Similarly to the observation from Figure 5, RW-1 outperforms Random for small target sets and Mix-1 outperforms Random for large target sets, in terms of delivery delay. Moreover, they all perform better than Degree when the size of target set is larger than 50.

In summary, when information service providers can deliver information directly to only a small number of users, we should use the pure random-walk based target-set selection policy. However, the enhanced scheme that mixes both influential and non-influential users into the target set is preferable when it is possible to deliver information to a large number of users directly.

6. CONCLUSION

In this paper, we propose a lightweight and distributed protocol, named iWander, to identify influential mobile users who have high centrality in their social-contact networks. iWander leverages fixed-length random walks and runs in the background of smartphones attached to mobile users. It estimates the centrality of individuals based on the number of times their smartphones are visited by random walks. We evaluate the performance of iWander using trace-driven simulations for two applications, targeted immunization of infectious diseases and target-set selection for information dissemination.

Our simulation results show that the proposed random-walk based immunization outperforms random immunization and performs very close to degree-based immunization, but generating only less than 1% of its message overhead. For the information dissemination application, the proposed random-walk based target-set selection performs better than random selection for small size of target set and another proposed scheme that chooses also users with low centrality into the target set outperforms random selection when the size of target set is large.

We are exploring the design space of device discovery to further reduce the message overhead of `iWander`. We also plan to evaluate its performance using other real-world human-contact traces [31].

7. ACKNOWLEDGEMENT

We thank the anonymous reviewers and our shepherd Konstantinos Pelechrinis for their insightful comments. We thank Madhav V. Marathe for valuable inputs. We also thank Pan Hui for the equipment support of several experiments. Aravind Srinivasan and Bo Han were supported in part by US National Science Foundation (NSF) ITR Award CNS-0426683, NSF Award CNS-0626636, and NSF Award CNS 1010789.

8. REFERENCES

[1] D. Braginsky and D. Estrin. Rumor Routing Algorithm For Sensor Networks. In *Proceedings of MobiCom 2002*, pages 22–31, Sept. 2002.

[2] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of Human Mobility on Opportunistic Forwarding Algorithms. *IEEE Transactions on Mobile Computing*, 6(6):606–620, June 2007.

[3] W. Chen, Y. Wang, and S. Yang. Efficient Influence Maximization in Social Networks. In *Proceedings of SIGKDD 2009*, pages 199–207, June-July 2009.

[4] N. A. Christakis and J. H. Fowler. Social Network Sensors for Early Detection of Contagious Outbreaks. *PLoS ONE*, 5(9):e12948, Sept. 2010.

[5] R. M. Christley, G. L. Pinchbeck, R. G. Bowers, D. Clancy, N. P. French, R. Bennett, and J. Turner. Infection in Social Networks: Using Network Analysis to Identify High-Risk Individuals. *American Journal of Epidemiology*, 162(10):1024–1031, Nov. 2005.

[6] V. Conan, J. Leguay, and T. Friedman. Fixed Point Opportunistic Routing in Delay Tolerant Networks. *IEEE Journal on Selected Areas in Communications*, 26(5):773–782, June 2008.

[7] S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi. Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *PLoS ONE*, 6(8):e23610, Aug. 2011.

[8] Z. Dezső and A.-L. Barabási. Halting viruses in scale-free networks. *Physical Review E*, 65(5):055103, May 2002.

[9] E. W. Dijkstra and C. S. Scholten. Termination Detection for Diffusing Computations. *Information Processing Letters*, 11(1):1–4, Aug. 1980.

[10] P. Domingos and M. Richardson. Mining the Network Value of Customers. In *Proceedings of SIGKDD 2001*, pages 57–66, Aug. 2001.

[11] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, Sept. 2009.

[12] S. Eubank, H. Guclu, V. S. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling Disease Outbreaks in Realistic Urban Social Networks. *Nature*, 429(6988):180–184, May 2004.

[13] S. L. Feld. Why Your Friends Have More Friends Than You Do. *American Journal of Sociology*, 96(6):1464–1477, May 1991.

[14] S. Gaonkar, J. Li, R. R. Choudhury, L. Cox, and A. Schmidt. Micro-Blog: Sharing and Querying Content Through Mobile Phones and Social Participation. In *Proceedings of MobiSys 2008*, pages 174–186, June 2008.

[15] M. Grossglauser and D. N. C. Tse. Mobility Increases the Capacity of Ad Hoc Wireless Networks. *IEEE/ACM Transactions on Networking*, 10(4):477–486, Aug. 2002.

[16] B. Han, P. Hui, V. S. A. Kumar, M. V. Marathe, J. Shao, and A. Srinivasan. Mobile Data Offloading through Opportunistic Communications and Social Participation. *IEEE Transactions on Mobile Computing*, 11(5):821–834, May 2012.

[17] D. Kempe, J. Kleinberg, and Éva Tardos. Maximizing the Spread of Influence through a Social Network. In *Proceedings of SIGKDD 2003*, pages 137–146, Aug. 2003.

[18] J. Kleinberg. The Convergence of Social and Technological Networks. *Communications of the ACM*, 51(11):66–72, Nov. 2008.

[19] D. Kotz, T. Henderson, I. Abyzov, and J. Yeo. CRAWDAD trace dartmouth/campus/movement/01_04 (v. 2005-03-08). Downloaded from `http://crawdad.cs.dartmouth.edu/dartmouth/campus/movement/01_04`, Mar. 2005.

[20] L. McNamara, C. Mascolo, and L. Capra. Media Sharing based on Colocation Prediction in Urban Transport. In *Proceedings of MobiCom 2008*, pages 58–69, Sept. 2008.

[21] M. Motani, V. Srinivasan, and P. S. Nuggehalli. PeopleNet: Engineering A Wireless Virtual Social Network. In *Proceedings of MobiCom 2005*, pages 243–257, Aug.-Sept. 2005.

[22] M. E. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54, Jan. 2005.

[23] N. P. Nguyen, T. N. Dinh, S. Tokala, and M. T. Thai. Overlapping Communities in Dynamic Networks: Their Detection and Mobile Applications. In *Proceedings of MobiCom 2011*, pages 85–95, Sept. 2011.

[24] J. D. Noh and H. Rieger. Random Walks on Complex Networks. *Physical Review Letters*, 92(11):118701, Mar. 2004.

[25] M. Papadopouli and H. Schulzrinne. Effects of Power Conservation, Wireless Coverage and Cooperation on Data Dissemination among Mobile Devices. In *Proceedings of MobiHoc 2001*, pages 117–127, Oct. 2001.

[26] B. Pásztor, L. Mottola, C. Mascolo, G. P. Picco, S. Ellwood, and D. Macdonald. Selective Reprogramming of Mobile Sensor Networks through Social Community Detection. In *Proceedings of EWSN 2010*, pages 178–193, Feb. 2010.

[27] K. Pearson. The Problem of the Random Walk. *Nature*, 72(1865):294, July 1905.

[28] J. Pollak, G. Gay, S. Byrne, E. Wagner, D. Retelny, and L. Humphreys. It's Time to Eat! Using Mobile Games to Promote Healthy Eating. *IEEE Pervasive Computing*, 9(3):21–27, July-Sept. 2010.

[29] K. K. Rachuri, C. Mascolo, M. Musolesi, and P. J. Rentfrow. SociableSense: Exploring the Trade-offs of Adaptive Sampling and Computation Offloading for Social Sensing. In *Proceedings of MobiCom 2011*, pages 73–84, Sept. 2011.

[30] M. Richardson and P. Domingos. Mining Knowledge-Sharing Sites for Viral Marketing. In *Proceedings of SIGKDD 2002*, pages 61–70, July 2002.

[31] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51):22020–22025, Dec. 2010.

[32] T. Small and Z. J. Haas. The Shared Wireless Infostation Model - A New Ad Hoc Networking Paradigm (or Where there is a Whale, there is a Way). In *Proceedings of MobiHoc 2003*, pages 233–244, June 2003.

[33] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. V. den Broeck, C. Régis, B. Lina, and P. Vanhems. High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School. *PLoS ONE*, 6(8):e23176, Aug. 2011.

[34] J. Yoon, B. D. Noble, M. Liu, and M. Kim. Building Realistic Mobility Models from Coarse-Grained Traces. In *Proceedings of MobiSys 2006*, pages 177–190, June 2006.

[35] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. SybilGuard: Defending Against Sybil Attacks via Social Networks. In *Proceedings of SIGCOMM 2006*, pages 267–278, Sept. 2006.

[36] G. Zyba, G. M. Voelker, S. Ioannidis, and C. Diot. Dissemination in Opportunistic Mobile Ad-hoc Networks: the Power of the Crowd. In *Proceedings of INFOCOM 2011*, pages 1179–1187, Apr. 2011.