Your Friends Have More Friends Than You Do: Identifying Influential Mobile Users Through Random-Walk Sampling

Bo Han, Jian Li and Aravind Srinivasan *Fellow, IEEE*

***Abstract*—In this paper, we investigate the problem of identifying influential users in mobile social networks. Influential users are individuals with high centrality in their social-contact graphs. Traditional approaches find these users through centralized algorithms. 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-walk sampling*. We prove that random-walk sampling with $O(\log n)$ steps, where n is the number of nodes in a graph, comes quite close to sampling vertices approximately according to their degrees. To the best of our knowledge, we are the first to design a distributed protocol on mobile devices that leverages random walks for identifying influential users, although this technique has been used in other areas.**

The most attractive feature of `iWander` is its *extremely low* control-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 generating only less than 1% of this policy's control messages. We also show that target-set selection based on `iWander` outperforms the random and degree-based selections for information dissemination in several scenarios.

***Index Terms*—Influential mobile users, centrality, random walks, disease control, information dissemination.**

I. Introduction

Mobile social networks, under the merging of social networks that link humans and the Internet that connects computers, 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 (e.g., PeopleNet [21], and SociableSense [28]). The performance of mobile social networks depends heavily on human mobility, which can increase their capacity through opportunistic communications and packet forwarding [11].

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 [7], [15]. 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 [3].

In this paper, we address the following question: *how do we identify influential users in mobile social networks through distributed solutions with low control-message overhead?* In our previous work [12], we proposed a heuristic algorithm to select influential mobile users for information dissemination, which is an extension of the greedy algorithm of Kempe, Kleinberg, and Tardos [15]. Recently, 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 of mobile users.

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 complexity, especially on large social graphs. For example, as reported by Chen et al. [2], 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, we need to control their communication overhead, as data transmission is the major source of energy consumption on mobile devices.

Our approach is motivated by the "friendship paradox" [9] that "*your friends have more friends than you do*" and leverages random-walk probe messages to sample mobile users and thus to identify critical users. This paradox illustrates "the phenomenon that most people have fewer friends than their friends have", on average. The reason behind it is that due

Manuscript received June 20, 2012; revised March 18, 2012.

Bo Han is with AT&T Labs – Research, 180 Park Avenue, Florham Park, NJ 07932, USA (e-mail: bohan@research.att.com). This work was done when he was a graduate student at the University of Maryland.

Jian Li is with the Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China (e-mail: lijian83@mail.tsinghua.edu.cn).

Aravind Srinivasan is with the Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA (e-mail: srin@cs.umd.edu).

to sampling bias people with a large number of friends may have a high probability of being observed among one's friend circle. Thus, the friends of *randomly* selected individuals, which are sampled by our scheme, may have higher centrality in friendship graphs than *average*. Although the original proof in Feld [9] 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 III). We assume that everyone in the examined network has a mobile device that runs `iWander` in the background. The key idea behind `iWander` is to sample users through random-walk probe messages generated periodically by mobile devices and estimate the centrality of individuals through their random-walk counters (i.e., the number of times their mobile devices are visited by random-walk probe messages). We prove that for static graphs that are "expander-like" (see, e.g., Eubank et al. [8]), the nodes with high random-walk counters are very likely to be those with high degrees. Our networks are inherently mobile and thus not static, but their static snapshots will likely be expander-like. Mobile networks will also likely mix well, serving to explain intriguing results such as those of Grossglauser and Tse [11].
- We present a targeted immunization policy based on the centrality information provided by `iWander` to contain the spread of infectious diseases (Section IV). We evaluate the performance of our proposed random-walk based immunization through extensive simulation studies using a real-world mobility trace. The simulation results show that random-walk based immunization always outperforms random immunization and performs very close to degree-based immunization with less than 1% of its control-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 V). 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 performs better than random selection only 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.

II. Related Work

We review related work on identifying influential individuals in various networks and applications of random walks in this section.

A. Identifying Influential Users

1) Traditional Social Networks: Identifying influential users has been extensively studied for information diffusion in traditional social networks [7], [15], [30]. Domingos and Richardson [7], [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. [15] prove that the information dissemination function of this influence maximization problem is submodular for the independent cascade model and the linear threshold model. They also propose a centralized greedy algorithm that outperforms heuristics based on node centrality and distance centrality. To solve the computational inefficiency of the centralized algorithms, Chen et al. [2] propose an improvement to reduce the algorithm's running time.

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 [12]. 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. Nguyen et al. [23] propose to select critical nodes through overlapping community detection in dynamic networks and nodes in more communities have higher priority in scenarios, such as message forwarding. They present a framework to adaptively update the community structure based on history information.

3) Targeted immunization: Targeted immunization has been proposed to eradicate infections for scale-free complex networks, by considering the heterogeneous connectivity properties of these networks. Christakis and Fowler [3] propose a mechanism for detecting contagious outbreaks. 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 at Harvard College. Christley et al. [4] evaluate the performance of network centrality measures for identifying high-risk individuals, including degree, shortest-path betweenness and random-walk betweenness. They show that degree performs very close to other network measures in predicting risk of infection. Lelarge [18] proposes a percolated threshold model for random networks. He also compares the performance of the degree based vaccination and the acquaintance vaccination [5] under security attacks.

Remark: All the above approaches for various problems, ranging from influence maximization to targeted immunization, are based on *centralized* solutions, except the acquaintance vaccination. We use random-walk probe messages generated by mobile devices to sample users during their contacts and design a distributed protocol to identify the most influential individuals. The acquaintance vaccination can be viewed as a special case of our approach which samples only the one-hop neighbors. Moreover, we demonstrate the effectiveness of our scheme for not only infectious disease control but also mobile content dissemination.

B. Random Walks

The term random walk was first introduced by Karl Pearson [26]. 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 random walkers 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.

Based on random walks, there are efficient sampling methods in peer-to-peer networks [32], online social networks [10], and other complex networks [29]. Stutzbach et al. [32] propose the Metropolized Random Walk with Backtracking (MRWB) to provide unbiased samples of representative peer properties in realistic unstructured P2P systems. Gjoka et al. [10] demonstrate that the Metropolis-Hastings random walk and a re-weighted random walk perform better than Breadth-First-Search (BFS) for obtaining an unbiased sample of Facebook users. Ribeiro and Towsley [29] propose the Frontier sampling method which uses multiple dependent random walkers to solve a known problem that traps a random walker inside a local neighborhood when the graphs are disconnected or loosely connected.

Random walks have also been widely explored in other fields, such as computer security, social science, economics, biology and psychology, for various purposes. For example, Xie et al. [33] propose to perform random moonwalks to identify the origins of a warm attack, under the assumption that the complete communication graph among hosts is available. Yu et al. [34] 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.

Differently from the above work, we employ random walks to design a distributed sampling scheme which can estimate the centrality of individuals. Also, our approach with low control message overhead is suitable for mobile applications. We refer interested readers to a preliminary version of this paper which appeared in MOBIHOC 2012 [14] for literature reviews of infectious disease control in public health and information dissemination in mobile networks.

III. The Random Walks Protocol

In this section, we present the details of `iWander` design, offer its theoretical analysis on static graphs, and discuss its proof-of-concept prototype implementation.

A. 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-walk probe messages from a small group of mobile devices, influential users may be visited by these probe messages more frequently than average.

The proposed `iWander` protocol works as follows. Every ΔT hours, `iWander` generates a tiny probe message with a given probability q on each mobile device and saves it in the device's local queue. The message contains *only* a pre-configured hop-limit field L. During the contacts of a mobile device with its peers, if it has a probe message in its queue, it sends this message to another uniformly and randomly selected peer. When a mobile device receives a probe 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 probe message with $L = 0$ will be finally discarded. `iWander` maintains a random-walk counter on each mobile device, initialized to zero, to record how many times it has received the probe messages (i.e., visited by these random-walk messages).

After collecting the random-walk counters from all users recorded by their mobile devices, 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 the random-walk destination of probe messages. Second, the message overhead of all-pairs random walks may be much higher than fixed-length random walks, which makes them unsuitable for battery-powered mobile devices.

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 mobile device generates a random-walk probe message, L – the length of random walks performed by probe messages (i.e., the number of mobile users visited by a single probe message), and ΔT – the frequency of generating new random-walk probe messages. It is important to understand the impact of these three parameters on the performance of `iWander`, because they determine both the quality of identified influential users and the number of probe messages spreading over the network.

B. Theoretical Analysis

We analyze the parameter L of our protocol on static graphs. To reduce energy consumption on mobile devices, we prefer short random walks with only a few steps. "Static" versions of social-contact networks are often very dense and expander-like. In such highly-mixing networks, we prove that a random walk of length $O(\log n)$, 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). Thus, the short random walks that we take will likely come quite close to sampling vertices approximately according to their degrees, because the static snapshots of dynamic mobile networks will likely be expander-like.

Let n be the number of nodes and m be the number of edges in the graph $G(V, E)$, which refers to a static version of the dynamic graph. Let $d(v)$ be the degree of vertex v and $d(S) = \sum_{v \in S} d(v)$ for any $S \subseteq V$. Suppose A is the adjacency matrix of G and D is the diagonal matrix $\mathrm{diag}(\frac{1}{d(v_1)}, \ldots, \frac{1}{d(v_n)})$. Suppose $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n$ are the eigenvalues of the symmetric matrix $N = D^{1/2} A D^{1/2}$. We assume the graph is an *expander graph*, which means the mixing rate $\lambda = \min(|\lambda_2|, |\lambda_n|)$ of G is a constant less than 1 [19]. There are several other definitions of expander graphs, such as *vertex expansion* or *edge expansion*, and they are more or less equivalent to each other.

Initially, we choose αn nodes to generate random-walk probe messages where α is a positive number between 0 and 1. After all random-walk probe messages run L steps, we select βn vertices with the highest *random-walk counters*. Denote this set by $\widehat{S}$ and the set of βn vertices with the highest *degrees* by S^*. Here, α is essentially the same as q (the probability to generate random-walk probe messages) and β is an input parameter whose value depends on the upper layer applications. We show that with high probability, the total degree of the chosen set $\widehat{S}$ is close to that of the optimal set S^*. Let the sum of all counters be $M = \alpha n L$.

Theorem 1: For any constants $\alpha, \beta, \epsilon > 0$ and sufficiently large n, after $L = \frac{216(3-\ln\beta)}{\epsilon^2 \alpha \beta} \log_{\frac{1}{\lambda}} n = O(\log n)$ steps, we have that

$$\Pr[d(\widehat{S}) \geq (1-\epsilon) d(S^*)] \geq 1 - \exp(-\Omega(n))$$

Before proving the main theorem, we present the following well-known facts.

Lemma 1: The stationary distribution π of a random walk is proportional to the degree distribution of the graph, i.e., $\pi(v) = \frac{d(v)}{2m}$, where m is the number of edges of the graph.

Lemma 2: (See e.g., [19]) Consider several independent random walks starting at arbitrary nodes. Let $P_{i,t}(v)$ be the probability that the ith random walk visits v at time t and let $P_{i,t}(S) = \sum_{v \in S} P_{i,t}(v)$. The stationary distribution of the random walk is π. We have that

$$|P_{i,t}(S) - \pi(S)| \leq \sqrt{d(S)} \lambda^t.$$

It is well known that the stationary distribution of a random walk is proportional to the degree distribution of the graph. More specifically, $\pi(v) = \frac{d(v)}{2m}$, where m is the number of edges of the graph.

We also need the following forms of the Chernoff bound [20].

Lemma 3: Suppose we have n independent random variables $X_1, X_2, \ldots, X_n$ distributed over $[0, T]$ and let $X = \sum_{i=1}^{n} X_i$, the following bounds hold:

1) For $0 < \delta \leq 1$, $\Pr[X \notin (1 \pm \delta)\mathbb{E}[X]] \leq 2\exp\left(-\mathbb{E}[X]\delta^2/3T\right)$, where $(1 \pm \delta)\mathbb{E}[X]$ denotes the interval $[(1-\delta)\mathbb{E}[X], (1+\delta)\mathbb{E}[X]]$.
2) For $\delta > 1$, $\Pr[X \geq (1+\delta)\mathbb{E}[X]] \leq \exp\left(-(1+\delta)\log(1+\delta)\mathbb{E}[X]/4T\right)$.
3) For any $t > 0$, $\Pr[X \geq \mathbb{E}[X] + t] \leq \exp\left(-2t^2/nT^2\right)$.

First, we show that with high probability, the actual sum of counters over the nodes in S^* is close to $M\pi(S^*)$.

Lemma 4: Let $C_t(v)$ be the random-walk counter of v at time t and $C_t(S) = \sum_{v \in S} C_t(v)$. We have that

$$\Pr\big[\, C_L(S^*) \in [(1 \pm \frac{\epsilon}{2}) M\pi(S^*)]\big] \geq 1 - \exp(-\Omega(n)) \quad (1)$$

Proof: Consider a particular vertex $v \in V$. Let $I_{i,t}(v)$ be the indicator random variable that the ith random walk visits v at time t. We can easily see from our proposed protocol that $C_t(v) = \sum_{i=1}^{\alpha n} \sum_{t'=1}^{t} I_{i,t'}(v)$. By linearity of expectation,

$$\mathbb{E}[C_t(v)] = \sum_{i=1}^{\alpha n} \sum_{t'=1}^{t} \mathbb{E}[I_{i,t'}(v)] = \sum_{i} \sum_{t' \leq t} P_{i,t'}(v)$$

Consider a random walk i. By Lemma 2, we can see that for any $S \subseteq V$,

$$\begin{aligned}
\big|\sum_{t' \leq L} P_{i,t'}(S) - \pi(S) \cdot L\,\big| &\leq \sum_{t' \leq L} |P_{i,t'}(S) - \pi(S)| \\
&= \sum_{t' \leq \log_{\frac{1}{\lambda}} n} |P_{i,t'}(S) - \pi(S)| \\
&\quad + \sum_{\log_{\frac{1}{\lambda}} n \leq t' \leq L} |P_{i,t'}(S) - \pi(S)| \\
&\leq \log_{\frac{1}{\lambda}} n + \sqrt{d(S)} \frac{\lambda^{\log_{\frac{1}{\lambda}} n}}{1-\lambda} \leq 2\log_{\frac{1}{\lambda}} n
\end{aligned}$$

Therefore, we obtain that for any $S \subseteq V$,

$$\begin{aligned}
|\mathbb{E}[C_L(S)] - M\pi(S)| &\leq 2\alpha n \log_{\frac{1}{\lambda}} n \leq \frac{\epsilon}{4} \alpha n L \beta \\
&= \frac{\epsilon}{4}\, \beta M \leq \frac{\epsilon}{4} M\pi(S^*)
\end{aligned}$$

In particular, we have that

$$\mathbb{E}[C_L(S^*)] \in [(1 \pm \frac{\epsilon}{4}) M\pi(S^*)]. \quad (2)$$

Since all random walks are independent of each other, using the Chernoff bound, we can get that

$$\begin{aligned}
&\Pr\big[\, |C_L(S^*) \in [(1 \pm \frac{\epsilon}{5})\mathbb{E}[C_L(S^*)]\big] \\
&\geq 1 - 2\exp(-\frac{\mathbb{E}[C_L(S^*)]\epsilon^2}{75L}) \\
&= 1 - \exp(-\Omega(n)) \quad (3)
\end{aligned}$$

Combining (2) and (3), we get

$$\begin{aligned}
&\Pr\big[\, C_L(S^*) \in [(1 \pm \frac{\epsilon}{2}) M\pi(S^*)]\big] \\
&\geq 1 - \exp(-\Omega(n)) \quad (4)
\end{aligned}$$

■

Next, we show that with high probability, $C_L(\widehat{S})$, the actual sum of counters over the nodes in the chosen set $\widehat{S}$, is at most $(1 + \epsilon/2) M\pi(\widehat{S})$. Note that the proof of this part is more involved than Lemma 4, since $\widehat{S}$ is a random set (so we can not directly apply the Chernoff bound on this set). Now, we provide the high level idea of the proof. We first outline a

natural idea that does not work, then sketch how to remedy this idea. Let $X_i(S) = \sum_{t=1}^{L} \sum_{v \in S} I_{i,t}(v)$. We use $\mathcal{E}(S)$ to denote the event that $\sum_i X_i(S) > \mathbb{E}[\sum_i X_i(S)] + \epsilon\beta M/2$. Notice that $C_L(\widehat{S}) \geq \beta M$, if we can prove $\mathcal{E}(S)$ does not happen for all S (with $|S| = \beta n$) with high probability, we are done. Consider a fixed set S. Since $\{X_i(S)\}_i$ are independent, we can use the third Chernoff bound to show that $\Pr[\mathcal{E}(S)] \leq \exp(-\frac{\epsilon^2\beta^2M^2}{2\alpha n L^2}) = \exp(-\epsilon^2\alpha\beta^2 n/2)$. However, the number of sets S with $|S| = \beta n$ is $\binom{n}{\beta n} > \exp(\epsilon^2\alpha\beta^2 n/2)$. So the union bound does not give a meaningful answer. To remedy this, we attempt to break $X_i(S)$ into two parts $X_{i,\mathsf{odd}}(S)$ and $X_{i,\mathsf{even}}(S)$. Each part is a sum of $K/2$ *almost independent* random variables where K is a constant to be fixed later. Hence, $C_L(S)$ can be represented as the sum of $K\alpha n$ random variables with smaller ranges (instead of αn random variables), which can result in a sharper concentration bound for $\mathcal{E}(S)$.

To make the above argument formal, we need the notions of variational distance and coupling (see e.g., [20]). The variational distance of two discrete random variables (or probability distributions) X and Y (with the same support supp) is defined to be $\Delta(X,Y) = \frac{1}{2}\sum_{x \in \mathsf{supp}} |\Pr(X = x) - \Pr(Y = x)|$. A *coupling* between two random variables X and Y is a pair of (correlated) random variables (X', Y') such that the marginal distribution of X' has distribution μ, and the marginal distribution of Y' has distribution μ. It is well known that $\Delta(X,Y) = \min_{(X',Y') \text{ is a coupling}} \Pr[X' \neq Y']$ [20]. The following lemma follows from the standard coupling argument.

Lemma 5: We are given two sequences of random variables $X = \{X_i\}_{1 \leq i \leq k}$ and $Y = \{Y_i\}_{1 \leq i \leq k}$. Let supp denote the common support of $X_1, \ldots, X_k$. Suppose X is Markovian in the sense that $\Pr(X_i|X_1, \ldots, X_{i-1}) = \Pr(X_i|X_{i-1})$ for all i. Suppose that Y is Markovian as well. The following statement holds:

(a) $\Delta(X,Y) \leq \sum_{i=2}^{k} \max_{x \in \mathsf{supp}} \Delta(X_i|x, Y_i|x) + \Delta(X_1, Y_1)$ where $X_i|x$ denotes the random variable with distribution $\Pr(X_i \mid X_{i-1} = x)$.

(b) Suppose both X and Y are trajectories (sequences of consecutive states) of the same Markov chain. Then, $\Delta(X,Y) = \Delta(X_1, Y_1)$.

Proof: We provide a proof for completeness. Both statements can be seen from the following standard way to couple X and Y (i.e., to construct a coupling (X', Y') between X and Y). In particular, we specify how (X', Y') is generated. First, we know there is a coupling (X_1', Y_1') between X_1 and Y_1 so that $\Pr[X_1' \neq Y_1'] = \Delta(X_1, Y_1)$. When $X_1' = Y_1' = x$ for some $x \in \mathsf{supp}$, we couple X_2 and Y_2 together in an optimal way, i.e., $\Pr[X_2' \neq Y_2' \mid X_1' = Y_1' = x] = \Delta(X_2|x, Y_2|x)$. If $X_1' \neq Y_1'$, we couple X_2 and Y_2 arbitrarily. In general, when $X_{i-1}' = Y_{i-1}' = x$ for some $x \in \mathsf{supp}$, we couple X_i and Y_i together in an optimal way. Otherwise, we couple X_i and Y_i arbitrarily. We can see the coupling (X', Y') constructed in this way is such that $\Pr[X' \neq Y'] \leq \Pr[X_1' \neq Y_1'] + \sum_{i=2}^{k} \Pr[X_i' \neq Y_i' \mid X_{i-1}' = Y_{i-1}'] \Pr[X_{i-1}' = Y_{i-1}']]$. Combining with the fact that $\Pr[X_i' \neq Y_i' \mid X_{i-1}' = Y_{i-1}'] \leq \max_{x \in \mathsf{supp}} \Delta(X_i|x, Y_i|x)$, the first statement follows.

For the second statement, because X and Y are from the same Markov chain, if $X_1' = Y_1'$, we can couple the rest X_is and Y_is perfectly (i.e., $\Pr[X_i' = Y_i' \mid X_1' = Y_1'] = 1$ for all $i \geq 2$) since the distributions of the subsequent transitions of both X and Y are exactly the same. Hence $\Pr[X' \neq Y'] = \Pr[X_1' \neq Y_1'] = \Delta(X_1, Y_1)$, from which we can see $\Delta(X,Y) \leq \Delta(X_1, Y_1)$. It is also straightforward to see that $\Delta(X,Y) \geq \Delta(X_1, Y_1)$. Thus, the second statement holds. ∎

Lemma 6: Let $\widehat{S}$ be the chosen set. We have that

$$\Pr[M\pi(\widehat{S}) > (1 - \epsilon/2)C_L(\widehat{S})] \geq 1 - \exp(-\Omega(n)).$$

Proof: We divide the whole process into $K = \frac{36}{\epsilon^2\alpha\beta}(3 - \ln\beta)$ stages. The jth stage consists of the steps $(j-1)T + 1, (j-1)T + 2, \ldots, jT$ where $T = 6\log_{1/\lambda} n$.

We let $P_{i,t}^u(v)$ to denote the probability that the ith random walk visits v at time step t conditioning on the event that the ith random walk visits u at time step $t - T$. Note that Lemma 2 does not depend on the starting node. Therefore, we have that for any i, any node u and any time step t after the first stage,

$$|P_{i,t}^u(v) - \pi(v)| \leq \sqrt{n}\lambda^t \leq \sqrt{n}\lambda^T < \frac{1}{n^5}. \tag{5}$$

Now, consider the ith random work and a particular node v. Recall that $I_{i,t}(v)$ is the indicator random variable that the ith random walk visits v at time t. Let $X_{i,j}(v) = \sum_{t \in \text{stage } j} I_{i,t}(v)$, i.e., the number of times the ith random walk visits v during stage j. From (5), $X_{i,3}(v)$, $X_{i,5}(v)$, $X_{i,7}(v)$, $\ldots$ are *almost* independent of each other. Hence, the sum $\sum_i \sum_{j=3,5,\ldots} X_{i,j}(v)$ consists of $\alpha n K/2$ almost independent random variables, which would result in a much sharper bound than summing αn independent random variables.

We need to define some auxiliary random variables. Let $Y(v)$ be the number of times that a random walk visits v during T steps if the initial node is selected according to the stationary distribution. Let $Y_{i,j}(v)$ be an independent random variable with the same distribution as $Y(v)$ for each $1 \leq i \leq \alpha n$ and $1 \leq j \leq T$. By Lemma 5(b) and noticing $X_{i,j}(v)$ is a function of the trajectory of the ith random walk during stage j[1], we can see that for any $j > 1$,

$$\begin{aligned}\Delta(X_{i,j}(v), Y_{i,j}(v)) &= \Delta(X_{i,j}(v), Y(v)) \\ &\leq |P_{i,(j-1)T+1}(v) - \pi(v)| \leq \frac{1}{n^5}.\end{aligned} \tag{6}$$

Let $X_{i,j}^u(v)$ denote the number of times the ith random walk visits v during stage j conditioning on the event that the position of the ith walk visits u at the end of stage $j-2$. Let $\Delta_j = \max_u \Delta(X_{i,j}^u(v), Y_j(v))$. From (5), we can also see that

$$\Delta_j \leq \sqrt{n}\lambda^T \leq \frac{1}{n^5}.$$

Let $X_{i,\mathsf{odd}}(v) = \sum_{j=3,5,\ldots} X_{i,j}(v)$ and $Y_{i,\mathsf{odd}}(v) = \sum_{j=3,5,\ldots} Y_{i,j}(v)$. Note that we do not include the first stage in $X_{i,\mathsf{odd}}(v)$ and $Y_{i,\mathsf{odd}}$ for now. From Lemma 5(a), we have that

$$\begin{aligned}&\Delta(X_{i,\mathsf{odd}}(v), Y_{i,\mathsf{odd}}(v)) \\ &\quad\leq \Delta(X_{i,3}(v), Y_{i,3}(v)) + \sum_{j=5,7,9,\ldots} \Delta_j < \frac{1}{n^4}.\end{aligned}$$

[1]We also use the simple fact that $\Delta(f(X), f(Y)) \leq \Delta(X,Y)$ for any function f.

Consider an arbitrary fixed set S

$$\Delta(X_{i,\mathsf{odd}}(S), Y_{i,\mathsf{odd}}(S)) \leq \sum_{v\in S} \Delta(X_{i,\mathsf{odd}}(v), Y_{i,\mathsf{odd}}(v)) < \frac{1}{n^3}$$

Therefore, there is a coupling $(X'_{i,\mathsf{odd}}(S), Y'_{i,\mathsf{odd}}(S))$ between $X_{i,\mathsf{odd}}(S)$ and $Y_{i,\mathsf{odd}}(S)$ such that

$$\Pr[X'_{i,\mathsf{odd}}(S) \neq Y'_{i,\mathsf{odd}}(S)] \leq \frac{1}{n^3}.$$

Let $Z'_{i,\mathsf{odd}}(S) = X'_{i,\mathsf{odd}}(S) - Y'_{i,\mathsf{odd}}(S)$. It is easy to see that

$$\mathbb{E}[Z'_{i,\mathsf{odd}}(S)] \leq \Pr[X'_{i,\mathsf{odd}}(S) \neq Y'_{i,\mathsf{odd}}(S)] \cdot T < \frac{1}{n^2}.$$

Since $\{Z'_{i,\mathsf{odd}}(S)\}_i$ are independent random variables, we can apply the second Chernoff bound:

$$\begin{aligned}
&\Pr[\sum_{i=1}^{\alpha n} Z'_{i,\mathsf{odd}}(S) \geq \frac{1}{12}\epsilon\beta M] \\
&\leq \Pr[\sum_{i=1}^{\alpha n} Z'_{i,\mathsf{odd}}(S) \geq \mathbb{E}[Z'_{i,\mathsf{odd}}(S)] + \frac{1}{13}\epsilon\beta M] \\
&\leq \exp\left(-\frac{1}{4}B\cdot\log B\cdot\frac{\mathbb{E}[Z'_{i,\mathsf{odd}}(S)]}{T}\right) \\
&= \exp(-\Omega(n\log n)), \qquad (7)
\end{aligned}$$

where $B = 1 + \frac{\epsilon\alpha\beta nL}{13\mathbb{E}[Z'_{i,\mathsf{odd}}(S)]}$. Since $\{Y_{i,j}(S)\}_{i,j}$ are independent random variables, again we get from the third Chernoff bound that:

$$\begin{aligned}
&\Pr[\sum_{i=1}^{\alpha n}\sum_{j=3,5,\ldots} Y_{i,j}(S) \geq \frac{M\pi(S)}{2} + \frac{\epsilon\beta M}{12}] \\
&\leq \exp(-4(\frac{\epsilon\beta M}{12T})^2/\alpha nK) = \exp(-\frac{1}{36}K\epsilon^2\beta^2\alpha n). \\
&= \exp(-\beta n(3-\ln\beta)) \qquad (8)
\end{aligned}$$

Hence, we can see that

$$\begin{aligned}
&\Pr[\sum_{i=1}^{\alpha n} X_{i,\mathsf{odd}}(S) \geq \frac{1}{2}M\pi(S) + \frac{1}{6}\epsilon\beta M] \\
&= \Pr[\sum_{i=1}^{\alpha n} X'_{i,\mathsf{odd}}(S) \geq \frac{1}{2}M\pi(S) + \frac{1}{6}\epsilon\beta M] \\
&\leq \Pr[\sum_{i=1}^{\alpha n} Y'_{i,\mathsf{odd}}(S) \geq \frac{1}{2}M\pi(S) + \frac{1}{12}\epsilon\beta M] \\
&\quad + \Pr[\sum_{i=1}^{\alpha n} Z'_{i,\mathsf{odd}}(S) \geq \frac{1}{12}\epsilon\beta M] \\
&\leq \exp(-\beta n(3-\ln\beta) + \exp(-\Omega(n\log n)) \qquad (9) \\
&\leq \exp(-\beta n(2-\ln\beta)).
\end{aligned}$$

By union bound, the probability that there is some set S with $|S| = \beta n$ such that $\sum_{i=1}^{\alpha n} X_{i,\mathsf{odd}} \geq \frac{1}{2}M\pi(S) + \frac{1}{6}\epsilon\beta M$ is at most

$$\begin{aligned}
&\binom{n}{\beta n}\cdot\exp(-\beta n(2-\ln\beta)) \leq (\frac{ne}{\beta n})^{\beta n}\exp(-\beta n(2-\ln\beta)) \\
&\leq \exp(\beta n(1-\ln\beta) - \beta n(2-\ln\beta)) = \exp(-\Omega(n))
\end{aligned}$$

Using the same argument, we can get that the probability that there is some set S with $|S| = \beta n$ such that $\sum_{i=1}^{\alpha n} X_{i,\mathsf{even}}(S) \geq \frac{1}{2}M\pi(S) + \frac{1}{6}\beta M$ is at most $\exp(-\Omega(n))$. Hence, the probability that there is no set S with $|S| = \beta n$ such that $\sum_{i=1}^{\alpha n}\sum_{j=2}^{K} X_{i,j}(S) \geq M\pi(S) + \frac{1}{3}\epsilon\beta M$ is at most $\exp(-\Omega(n))$. We notice that $X_{i,1}(S)$ is at most $\beta nT = \beta nL/K \leq \frac{\epsilon}{100}\beta M$. Therefore, the probability that for every $S \subset V$ with $|S| = \beta n$ such that $C_L(S) = \sum_{i=1}^{\alpha n}\sum_{j=1}^{K} X_{i,j}(S) < M\pi(S) + \frac{1}{2}\epsilon\beta M$ is at least $1 - \exp(-\Omega(n))$.

Finally, noticing that $C_L(\widehat{S}) \geq \beta M$, we conclude that $\Pr[M\pi(\widehat{S}) > (1-\epsilon/2)C_L(\widehat{S})] \geq 1 - \exp(-\Omega(n))$. ■

Now, everything is ready to prove the main theorem of the section. Assuming the events in both Lemma 4 and Lemma 6 happen (with probability $1 - \exp(-\Omega(n))$), we have

$$M\pi(\widehat{S}) > (1-\frac{\epsilon}{2})C_L(\widehat{S}) \geq (1-\frac{\epsilon}{2})C_L(S^*) \geq (1-\frac{\epsilon}{2})^2 M\pi(S^*).$$

Thus $\Pr[\pi(\widehat{S}) > (1-\epsilon)\pi(S^*)] \geq 1 - \exp(-\Omega(n))$. Note that $\pi(S) = d(S)/2m$ for any $S \subseteq V$ and we complete the proof.

IV. Controlling Infectious Diseases

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

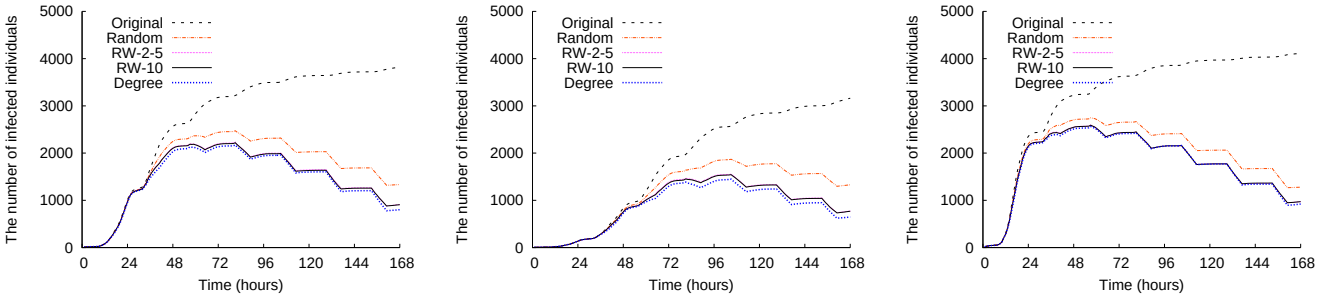
A. Random-Walk Based Immunization

Mobile devices 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 (https://www.fluphone.org/) 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. [27] design a mobile phone based game to motivate children to practice healthy eating habits.

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 V. 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 [6].

There are several differences between our proposed targeted immunization scheme and those in the literature, for example, by Christakis and Fowler [3] and Christley et al. [4]. First, our scheme can benefit from the social contacts detected directly by mobile devices, instead of using the estimation



(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

Fig. 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.

through friendship graphs generated from surveys [3]. 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 [3], and its low control message overhead makes it amenable to be run on mobile devices.

B. Performance Evaluation

We evaluate the performance of `iWander` for infectious disease control through extensive trace-driven simulations. Although we have implemented a prototype of `iWander` in C language on Nokia N900 smartphones with less than 300 lines of code, it is hard to evaluate it performance in practice with large number of participants.

1) Simulation Setup: We implement a simulator in C based on the SIR model [16], to simulate the spread of infectious diseases. Our simulation is based on a probabilistic temporal graph model, in which diseases propagate among individuals with a certain probability and the infections depend on the temporal nature of contact events. 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 [17], 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. [35], 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 [1], [35].

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 Haggle data set of mobile users [1] and the Cabspotting traces of San Francisco's taxi cabs (http://cabspotting.org/). However, some of them is too small (e.g., the Haggle data set with only less than 100 users) and others 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. We chose to not plot the standard deviation for the sake of clarity. The standard deviations are small, for example, usually less than 100 after 80 hours in Figure 1a.

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 mobile device attached with a student performs device discovery every 60 seconds to record the number of other devices it has contacted 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 based immunization, `iWander` also performs device discovery every 60 seconds only when the message queues on mobile devices 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. [4] 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. [25] 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. Mobile devices 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 hour.

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 percentage 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 generating a single random-walk probe message from a given mobile device with n steps, and `RW-`m`-`n for generating m probe messages from a mobile device with n steps.

In 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. Figure 1f plots the case with 0.003 infection probability, 10 initial infections and immunizations after 10% infections.

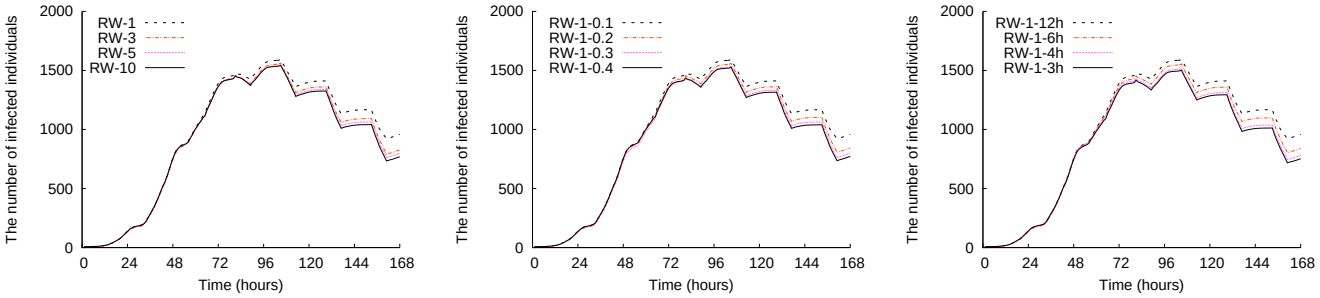
As we can see from these 6 subfigures, `RW-10` performs very close to `Degree` and they all outperform `Random`. Compared to `Random`, the improvement of `RW-10` ranges from 14.10% (Figure 1c) to 25.36% (Figure 1b). On average `RW-2-5` generates the same amount of random-walk probe messages as `RW-10`, and it performs very close to (slightly worse than) `RW-10` because probe messages with longer steps have more chances to visit influential users.

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 performed by probe messages, and plot the simulation results 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 control message overhead of `iWander` with different lengths, probabilities and frequencies of random walks in Figures 3a, 3b and 3c. There are three types of control messages, probe request and probe response messages for device discovery, and random-walk probe messages for `iWander`. In all these subfigures, the baseline is `iWander` with 1-step random walks and mobile devices generate random-walk messages with probability 0.1 every 12 hours.

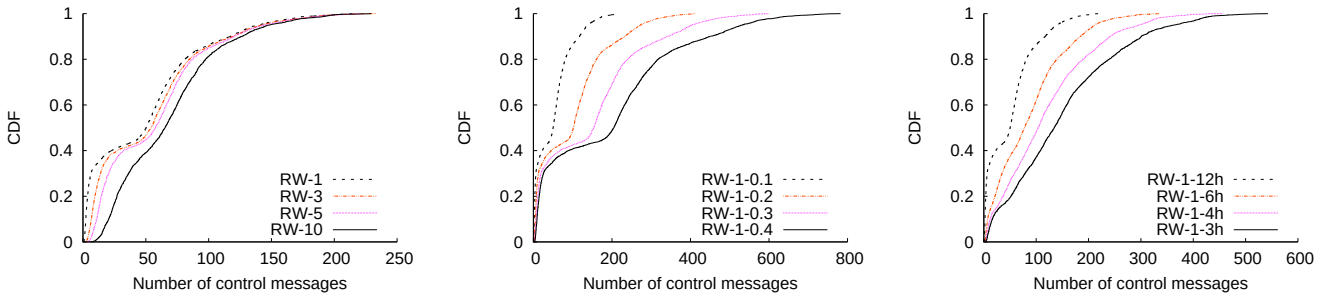
We note that in practice the actual message overhead of device discovery depends on the underlying communication technology. Bluetooth and WiFi are the two most commonly available technologies on smartphones that we can utilize for device discovery. When using Bluetooth, a device will send out inquiry messages periodically to actively discover its peers. WiFi devices can also periodically send out Beacon messages to announce the existence of a network and to facilitate device discovery. In this paper, we simulate the device discovery for Bluetooth because it is more energy efficient than WiFi [13].

We plot the CDF of the amount of one-day per-user control messages transmitted by mobile devices on 2004-03-01. As we can see from these subfigures, around 50% of mobile devices generate less than 200 control messages when using `iWander`. For `Degree`, all messages are transmitted during device discovery and the number of per-user control messages ranges from 1,441 to 25,608 for the simulated period. The amount of one-day per-user control messages transmitted by `iWander` is extremely low, less than 800 for all cases. An interesting observation from these three subfigures is that there



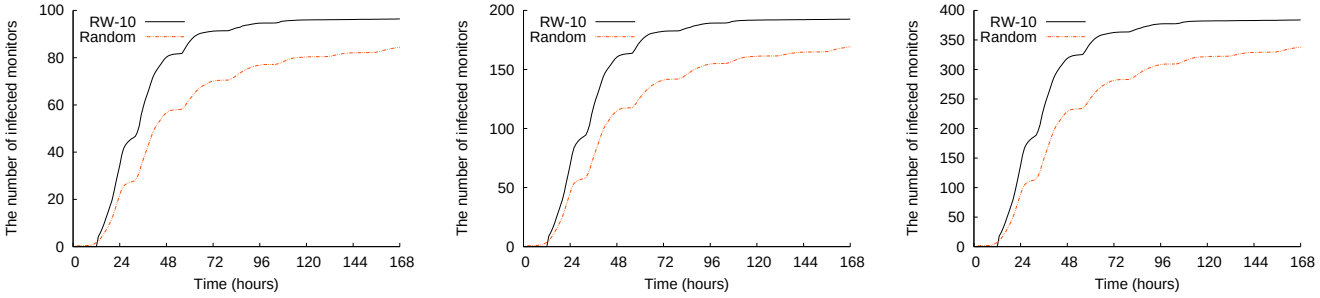
(a) different lengths of random walks (b) different prob. of random walks (c) different frequencies of random walks

Fig. 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

Fig. 3: Comparison of the amount of per-user control messages for different lengths, probabilities and frequencies of random walks. The number of per-user control messages for the degree-based scheme ranges from 1,441 to 25,608.



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

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

are two kinds of mobile devices: active (high mobility and transmitting a large number of control messages) and inactive. In Section V, we harness this observation to improve the performance of mobile information dissemination.

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 [3], [8]. 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 [3]. 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 [3], 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. Mobile devices generate random-walk probe 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.

V. Facilitating Information Dissemination

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

A. 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 [12]. We employ opportunistic communications and social participation to facilitate information dissemination and thus reduce the amount of data traffic in 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 [15].

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. Mobile devices attached with 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 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, mobile devices 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 infectious 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, as we will show in Section V-B2, adding non-influential users into the target set can increase the number of infected users for large target sets.

B. Performance Evaluation

We develop another trace-driven simulator also in C, using the same Dartmouth data set [17], 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 [12]. We are currently working on a packet-level simulator to take into account the low layer issues, including the failure of random-walk probe messages and the transmission of data packets in information dissemination.

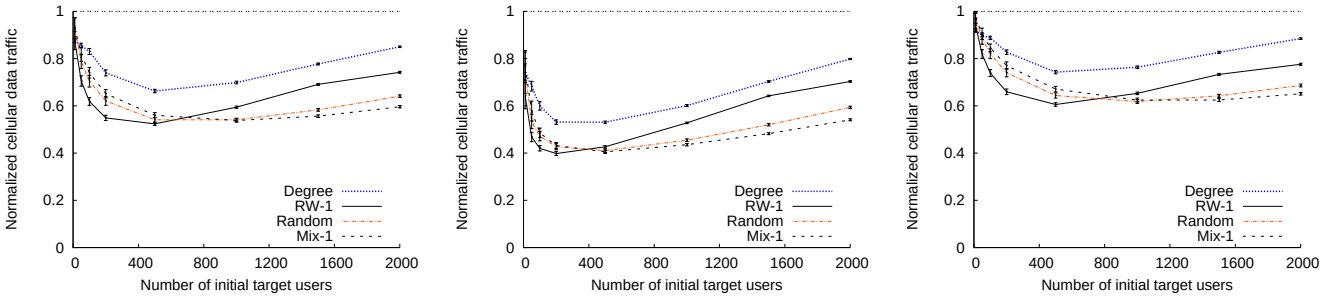
1) Simulation Setup: The simulator first generates the contacts trace of mobile users under the same assumption that they are in contacts if their wireless devices are associated with the same access point. It then replays the contact events for the given information dissemination period, from 12:00PM to 15:00PM on 2004-03-01.[2] 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 [12].

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 mobile devices have the chance to start the exchange of information every 60 seconds. Similar to degree-based immunization, `Degree` also uses the number of other devices that a mobile device has contacted with as the metric to select target users. For `RW-1`, mobile devices generate 1-step random-walk probe 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 mobile devices at the beginning of information dissemination period.

2) The Amount of Cellular Data Traffic: We plot the normalized amount of cellular data traffic for `RW-1`, `Random` and `Degree` in Figure 5. In these subfigures, the y-axis value is normalized over the amount of cellular data traffic of a baseline scheme, in which information service providers send content to every user through cellular unicast delivery. 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). Thus, `RW-1` can reduce more cellular data traffic than `Random`.

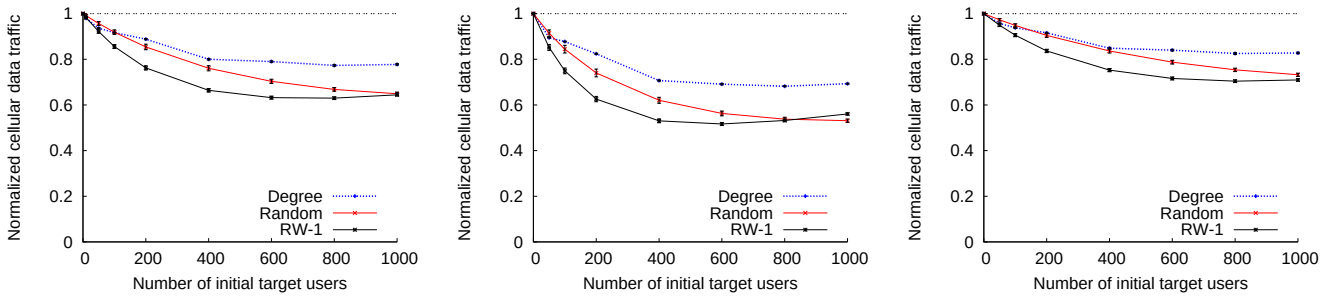
The performance of `RW-1` becomes worse than `Random` for large target sets. One of the possible reasons is that

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



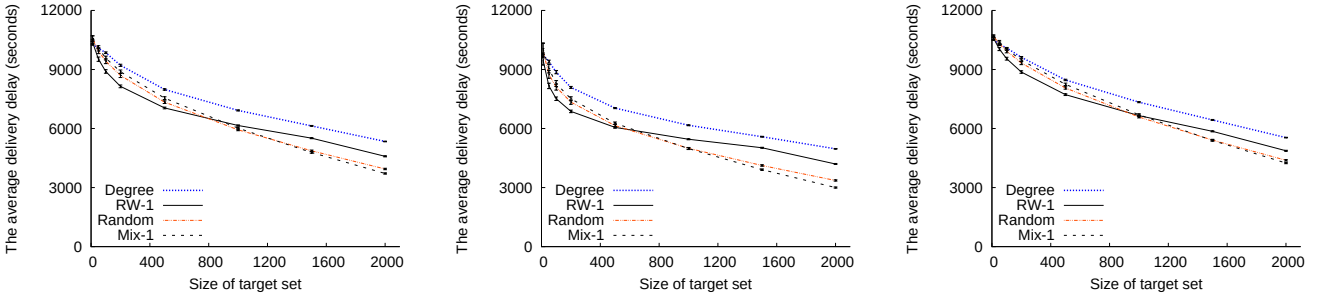
(a) p: 0.01 (b) p: 0.05 (c) p: 0.005

Fig. 5: Comparison of the normalized cellular data traffic for four target-set selection schemes with different values of p.



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

Fig. 6: Comparison of the normalized cellular data traffic 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

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

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. [35], 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 enhanced 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 also plot the normalized cellular data traffic during the information dissemination. In this uncooperative scenario, `RW-1` performs much better than `Random` and `Degree`. For example, for a target set with 600

users, RW−1 can reduce the amount of cellular data traffic by 48.34% when $p = 0.05$, compared with the baseline scheme. The percentage of reduction is 36.81% when $p = 0.01$ and 28.40% when $p = 0.005$. For large target sets, Random performs slightly better than 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.

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 performs better than 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.

VI. 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 mobile devices attached to users. It estimates the centrality of individuals based on the number of times their mobile devices are visited by random-walk probe messages. We prove that for expander-like static graphs the proposed random-walk sampling is very close to sampling vertices according to their degrees.

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 control 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].

VII. Acknowledgement

We thank the anonymous reviewers for their insightful comments. We thank Bobby Bhattacharjee and Madhav V. Marathe for valuable inputs. 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. Jian Li is supported in part by the National Natural Science Foundation of China Grant 61202009, 61033001, 61061130540 and 61073174, and the National Basic Research Program of China Grant 2011CBA00300 and 2011CBA00301.

References

[1] 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.

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

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

[4] 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.

[5] R. Cohen, S. Havlin, and D. ben Avraham. Efficient Immunization Strategies for Computer Networks and Populations. *Physical Review Letters*, 91(24):247901, Dec. 2003.

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

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

[8] 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.

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

[10] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *Proceedings of INFOCOM 2010*, pages 1–9, Mar. 2010.

[11] 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.

[12] 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.

[13] B. Han and A. Srinivasan. eDiscovery: Energy Efficient Device Discovery for Mobile Opportunistic Communications. In *Proceedings of ICNP 2012*, pages 1–10, Oct.-Nov. 2012.

[14] B. Han and A. Srinivasan. Your Friends Have More Friends Than You Do: Identifying Influential Mobile Users Through Random Walks. In *Proceedings of MOBIHOC 2012*, pages 5–14, June 2012.

[15] 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.

[16] W. O. Kermack and A. G. McKendrick. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society of London (Series A)*, 115(772):700–721, Aug. 1927.

[17] 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.

[18] M. Lelarge. Efficient Control of Epidemics over Random Networks. In *Proceedings of SIGMETRICS 2009*, pages 1–12, June 2009.

[19] L. Lovász. Random Walks on Graphs: A Survey. *Combinatorics, Paul Erdős is Eighty*, 2(1):1–46, 1993.

[20] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.
[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] 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.
[26] K. Pearson. The Problem of the Random Walk. *Nature*, 72(1865):294, July 1905.
[27] 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.
[28] 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.
[29] B. Ribeiro and D. Towsley. Estimating and Sampling Graphs with Multidimensional Random Walks. In *Proceedings of IMC 2010*, pages 390–403, Nov. 2010.
[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] D. Stutzbach, R. Rejaie, N. Dufeld, S. Sen, and W. Willinger. On Unbiased Sampling for Unstructured Peer-to-Peer Networks. In *Proceedings of IMC 2006*, pages 27–39, Oct. 2006.
[33] Y. Xie, V. Sekar, D. A. Maltz, M. K. Reiter, and H. Zhang. Worm Origin Identification Using Random Moonwalks. In *Proceedings of the 2005 IEEE Symposium on Security and Privacy*, pages 242–256, May 2005.
[34] 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.
[35] 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.

Bo Han received the Bachelor's degree in Computer Science and Technology from Tsinghua University in 2000, the M.Phil. degree in Computer Science from City University of Hong Kong in 2006 and the Ph.D. degree in Computer Science from the University of Maryland in 2012. He is currently a senior member of technical staff at AT&T Labs Research. He interned at AT&T Labs Research in the summers of 2007, 2008 and 2009, Deutsche Telekom Laboratories for the summer of 2010, and HP Labs during the summer of 2011. His research interests are in the areas of wireless networking and mobile computing, with a focus on developing simple yet efficient and elegant solutions for real-world networking and systems problems.

Jian Li is an Assistant Professor at the Institute for Interdisciplinary Information Sciences, Tsinghua University. He got his BSc degree from Sun Yat-sen (Zhongshan) University, China, MSc degree in Computer Science from Fudan University, China and PhD degree in the University of Maryland, USA. His research interests lie in the areas of algorithms, databases and wireless sensor networks. He co-authored several research papers that have been published in major computer science conferences and journals. He received the best paper awards at VLDB 2009 and ESA 2010.

Aravind Srinivasan (Fellow, IEEE) is a Professor (Dept. of Computer Science and Institute for Advanced Computer Studies) at the University of Maryland, College Park. He received his degrees from Cornell University (Ph.D.) and the Indian Institute of Technology, Madras (B.Tech.). His research interests are in randomized algorithms, networking, social networks, combinatorial optimization, and related areas. He has published several papers in these areas, in journals including Nature, Journal of the ACM, IEEE/ACM Transactions on Networking, and SIAM Journal on Computing. He is an editor of four journals, and has served on the program committees of various conferences.