

# Multigraph Sampling of Online Social Networks

Minas Gjoka, Carter T. Butts, Maciej Kurant, and Athina Markopoulou, *IEEE Member*

**Abstract**—State-of-the-art techniques for probability sampling of users of online social networks (OSNs) are based on random walks on a single social relation (typically friendship). While powerful, these methods rely on the social graph being fully connected. Furthermore, the mixing time of the sampling process strongly depends on the characteristics of this graph.

In this paper, we observe that there often exist other relations between OSN users, such as membership in the same group or participation in the same event. We propose to exploit the graphs these relations induce, by performing a random walk on their *union multigraph*. We design a computationally efficient way to perform multigraph sampling by randomly selecting the graph on which to walk at each iteration. We demonstrate the benefits of our approach through (i) simulation in synthetic graphs, and (ii) measurements of Last.fm— an Internet website for music with social networking features. More specifically, we show that multigraph sampling can obtain a representative sample and faster convergence, even when the individual graphs fail, *i.e.*, are disconnected or highly clustered.

**Index Terms**—Sampling methods, Social network services, Last.fm, Random walks, Multigraph, Graph sampling.

## I. INTRODUCTION

THE POPULARITY of Online Social Networks (OSNs) has skyrocketed within the past decade, with the most popular having at present hundreds of millions of users (a number that continues to grow apace). This success has inspired a number of measurement and characterization studies, as well as studies of the interaction between OSN structure and systems design, and of user behavior within OSNs. Despite their attractions, the large size and access limitations of most OSN services (*e.g.*, API query limits, treatment of user data as proprietary) make it difficult or impossible to obtain a complete census of user accounts and/or topology. Sampling methods are thus essential for practical estimation of OSN properties. While sampling can, in principle, allow precise inference from a relatively small number of observations, this depends critically on the ability to draw a sample with known statistical properties. The lack of a sampling frame (*i.e.*, a complete list of users, from which individuals can be directly sampled) for most OSNs makes principled sampling especially difficult; recent work in this area has thus focused on sampling methods that evade this limitation.

Manuscript received 15 December 2010; revised 2 June 2011. This work has been supported by the following grants: NSF CDI award 1028394, AFOSR FA9550-10-1-0310, AFOSR MURI Prime Award FA9550-09-0643; SNF grant PBELP2-130871, Switzerland; and DOD ONR award N00014-08-1-1015.

The authors are with the California Institute for Telecommunications and Information Technology (CalIT2) at the University of California, Irvine, CA 92697-2800 (e-mail: {mgjoka, buttsc, mkurant, athina}@uci.edu). Carter T. Butts is also with the Sociology Department at UC Irvine. Athina Markopoulou is with the EECS Department at UC Irvine and with the Center for Pervasive Communications and Computing (CPC).

Digital Object Identifier 10.1109/JSAC.2011.111012.

Key to current sampling schemes is the fact that OSN users are, by definition, connected to one another via some relation, referred to here as the “social graph.” Specifically, samples of OSN users can be obtained by crawling the OSN social graph, obviating the need for a sampling frame. An early family of crawling techniques followed BFS/Snowball-type approaches, where nodes of a graph reachable from an initial seed are explored exhaustively [1]–[3]. It is now well-known that these techniques produce biased samples with poor statistical properties when the full graph is not covered [4]–[6]. A more recent body of work employs systematic random walks on the social graph, and can achieve an asymptotic probability sample of users by online or a posteriori correction for the (known) bias induced by the crawling process [6,7]. While random walk sampling can be very effective, its success is ultimately dependent on the connectivity of the underlying social graph. More specifically, random walks can yield a representative sample of users only if the social graph is fully connected. Furthermore, the speed with which the random walks converge to the target distribution strongly depends on characteristics of the graph, *e.g.*, clustering.

In this paper, we start from the observation that in OSNs, there are often multiple relations connecting the nodes. For example, users may be linked not only by direct social ties, but also by being members of the same group, participating in the same event, or using the same application. Moreover, many systems allow all neighbors in such relations to be enumerated (either through scraping or API calls). In other words, there often exist multiple, crawlable relation graphs—including but not limited to ties like “friendship”—defined on the same set of nodes. We propose to exploit such multiple-relation (*i.e.*, multiplex) graphs by giving a crawler more edges to choose from, compared to a crawler restricted to one relation only (typically the social graph). For example, we might be able to discover users that have no direct social ties, which is impossible by crawling the social graph alone.

There are many ways one can exploit multiplex graphs. A naive approach would be to run many crawlers, one on each individual relation graph, and then combine the collected samples. However, this technique yields biased samples if any individual relation graph is fragmented, and fails to exploit opportunities for convergence acceleration by mixing across relations. A better approach is to combine all individual relation graphs into a single *union* (simple) *graph*: the resulting union graph is frequently connected even if its constituent graphs are not. Moreover, the union graph may also be less tightly clustered than its constituents, helping a crawler to converge faster than on the individual graphs. However, walking on the union graph requires, at every step, the enumeration of all neighbors in all relations, which can

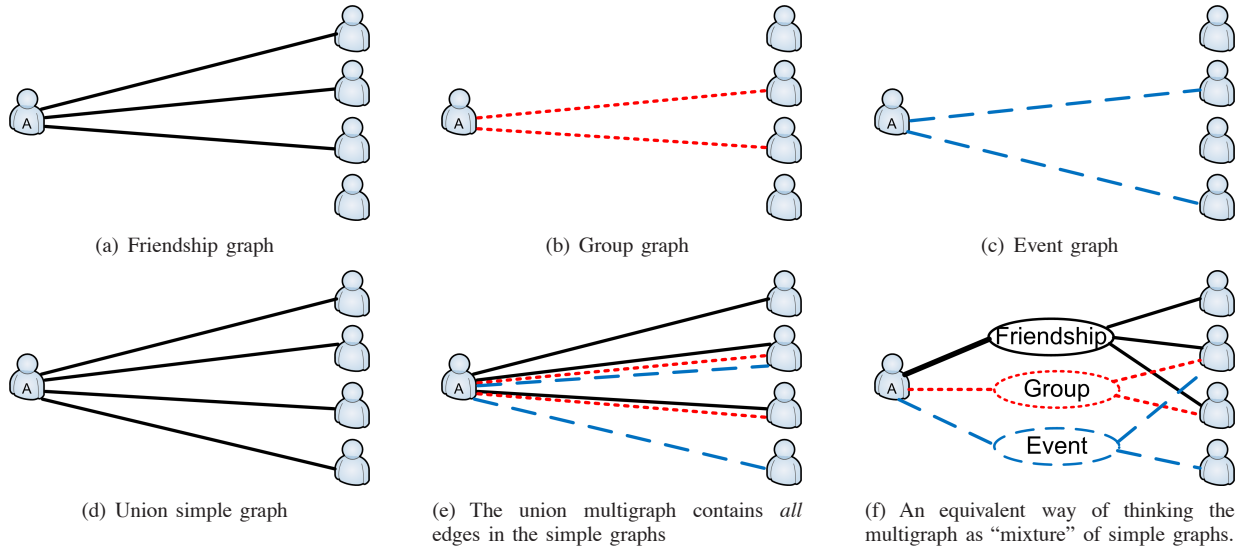


Fig. 1. Multigraph sampling illustration. (a-c) Graphs for three different relation  $G_i$ : Friendship, Group and Event. (d) Union (simple) graph, as presented in Definition 1. (e) Union multigraph, as presented in Definition 2. Node  $A$  has degrees  $d_1(A)=3$ ,  $d_2(A)=2$  and  $d_3(A)=2$  in the Friendship, Group and Event graphs, respectively. Its total degree in the union multigraph is  $d(A)=7$ . (f) An alternative view of the union multigraph. The neighbor selection Algorithm 1 first selects a graph  $G_i$  with probability  $\frac{d_i(A)}{d(A)}$ . Next, it picks a random neighbor in the selected graph  $G_i$ , i.e., with probability  $\frac{1}{d_i(A)}$ .

be costly in time and bandwidth. Instead we propose a third, cost-efficient approach.

We propose a novel two-stage algorithm that walks on the *union multigraph*. Our multigraph sampling first selects the relation on which to walk and then enumerates the neighbors with regards to that relation only, which makes it, in practice, even more efficient than union graph sampling. We prove that this algorithm achieves convergence to the proper equilibrium distribution when the union multigraph is connected. We also demonstrate the benefits of multigraph sampling in two settings: (i) by simulation of synthetic random graphs; and (ii) by measurements of Last.fm - an Internet website for music with social networking features. We chose Last.fm as an example of a network that is highly fragmented with respect to the social graph as well as other relations. We show that multigraph sampling can obtain a representative sample when each individual graph is disconnected. Along the way, we also give practical guidelines on how to efficiently implement multigraph sampling for OSNs more generally.

The structure of the rest of the paper is as follows. Section II describes our sampling methodology. Section III evaluates our methodology on synthetic graphs. Section IV applies our methodology to sample Last.fm and provides practical recommendations. Section V discusses related work. Finally, Section VI concludes the paper.

## II. SAMPLING METHODOLOGY

### A. Terminology and Definitions

We consider different sets of edges  $\mathcal{E} = \{E_1, \dots, E_Q\}$  on a common set of users  $V$ . Each  $E_i$  captures a symmetric relation between users, such as friendship or group co-membership.  $(V, E_i)$  thus defines an undirected graph  $G_i$  on  $V$ . We make no assumptions of connectivity or other special properties of each  $G_i$ . Fig 1(a-c) shows an example of  $Q = 3$  different relations and relation graphs  $G_i$  defined on the same 5 nodes.

Fig. 2(a-e) shows  $Q=5$  such graphs defined on a set of 50 nodes.

Consider set of graphs  $G_i = (V, E_i)$ ,  $i = 1, \dots, Q$ , defined on a common node set  $V$ .  $\mathcal{E}$  can be used to construct several types of combined structures on  $V$ . We will employ the following two such structures:

*Definition 1:* The *union (simple) graph*  $G' = (V, E')$  of  $G_1, \dots, G_Q$  is defined as the graph on  $V$ , whose edges are given by the set  $E' = \cup_{i=1}^Q E_i$ . ■

*Definition 2:* The *union multigraph*  $G = (V, E)$  of  $G_1, \dots, G_Q$  is defined as the multigraph on  $V$ , whose edges are given by the multiset  $E = \uplus_{i=1}^Q E_i$ . ■

Note that the union multigraph  $G$  can contain multiple edges between a pair of nodes, while the union graph  $G'$  contains only one (or no) edge. Every multigraph  $G$  can be reduced to the union graph  $G'$  by merging together multiple edges between two nodes into one. Our focus, in this paper, is on the union multigraph, also referred to as simply the *multigraph*, because it allows us to more efficiently implement sampling on multiple relations. However, we also use the union graph as a helpful conceptual tool.

### B. Some False Starts

We seek to draw a sample of the nodes in  $V$ , so that the draws are (at least approximately) independent and the sampling probability of each node is known up to a constant of proportionality. There are several ways to achieve this goal using multiple graphs. We discuss some of them below.

1) *Naive Multiple Graph Sampling:* A naive way is to run many random walks, one per each individual graph  $G_i$ , and to combine the collected samples. However, if a particular  $G_i$  is disconnected (as are all five graphs in Fig. 2(a-e)), a walk on  $G_i$  is restricted to the connected component around its starting point and thus never converges to the desired target distribution. This results in an asymptotically biased sample, dependent on the initial seeds.

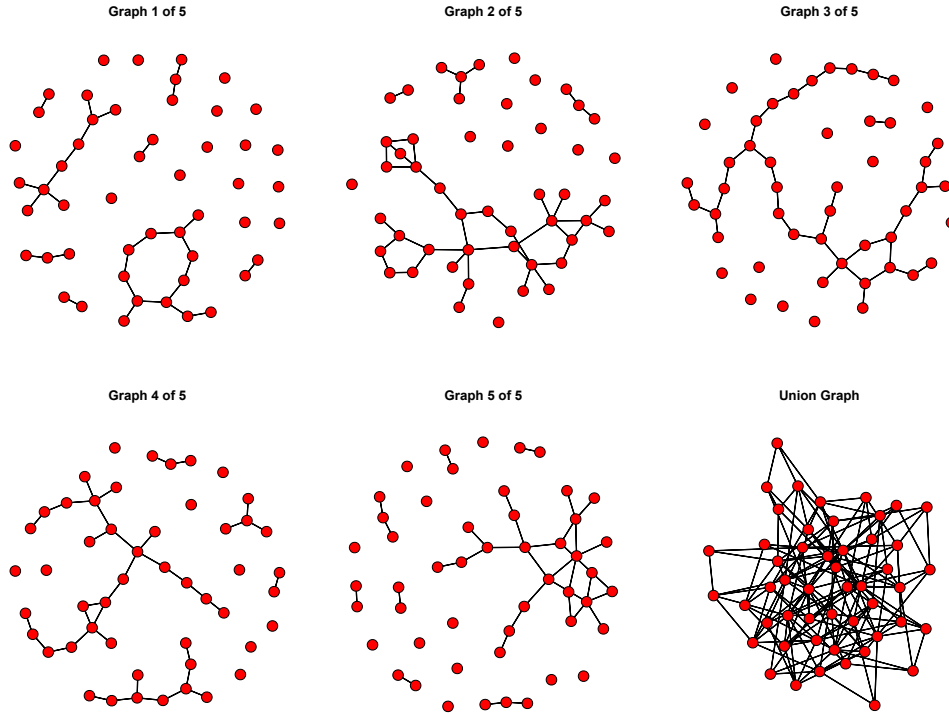


Fig. 2. Example of multiple graphs vs. their union. Five draws from a random  $(N, p)$  graph with  $N = 50$  and expected degree 1.5 are depicted in (a)-(e). Each simple graph is disconnected, while their union graph  $G'$ , depicted in (f), is fully connected. The union multigraph  $G$  (not shown here) is also connected: it has - possibly multiple - edges between the exact same pairs of nodes as  $G'$ .

2) *Union Simple Graph Sampling*: A much better approach is to perform a random walk on the union graph  $G'$ . We show examples of union graphs in Fig 1(d) and Fig. 2(f). Note that although in both cases the individual graphs are disconnected, their union graphs are well-connected, which allows for a quick convergence of the random walk.

A potential practical difficulty with a random walk on the union graph  $G'$  is that, at each step, computing the neighborhood union can be quite expensive: it requires the enumeration of all edges adjacent to the current vertex  $v$ , in each relation graph  $G_i$ . This may be very costly, depending on  $v$ 's neighborhood size (which can be large in heavily clustered relations, such as group co-membership), query costs, and the number of relations  $Q$ .

C. *Union Multigraph Sampling*

We can address the enumeration problem of the union graph  $G'$  by considering the *union multigraph*; see Definition 2 and example illustrated in Fig. 1(e). We employ a random walk that moves from one vertex to another by selection of random edges on the multigraph. A naive implementation of such a random walk still requires the enumeration of all neighbors of the current node  $v$ . Instead, we propose to use the following two-stage neighbor-selection procedure described in Algorithm 1 and depicted in Fig.1(f), which requires enumeration of  $v$ 's neighborhood for only a single graph.

Denote by  $d_i(v)$  the degree of node  $v$  in graph  $G_i$ , and by  $d(v) = \sum_{i=1}^Q d_i(v)$  its total degree in the union multigraph. First, we select a graph  $G_i$  with probability  $\frac{d_i(v)}{d(v)}$ . Second, we pick uniformly at random an edge of  $v$  within the selected  $G_i$  (i.e., with prob.  $\frac{1}{d_i(v)}$ ), and we follow this edge to  $v$ 's neighbor. This procedure is equivalent to selecting an edge of  $v$  uniformly at random in the union multigraph, because  $\frac{d_i(v)}{d(v)} \cdot \frac{1}{d_i(v)} = \frac{1}{d(v)}$ .

Note that in Step 3, Algorithm 1 requires only the values of degrees  $d_i(v)$  of all relation graphs. Only in Step 4 of Algorithm 1 does one enumerate all neighboring edges in the selected  $G_i$ . Because the degree information  $d_i(v)$  is usually much cheaper to obtain (e.g., via simple low-bandwidth API calls) than enumerating all  $d_i(v)$  edges, Algorithm 1 has the potential to save much bandwidth compared to the union simple graph sampling (which enumerates all neighboring edges in all relations). This benefit is amplified when higher numbers of relations ( $Q$ ) are used. Algorithm 1 may also be helpful in certain offline applications involving surveys and human respondents (e.g., RDS [8]), in which selection of random neighbors is possible but enumeration is not.

Algorithm 1 leads to the following equilibrium distribution:

*Proposition 2.1:* If  $G$  is connected and contains at least one triangle, then Algorithm 1 leads to equilibrium distribution  $\pi(v) = \frac{d(v)}{\sum_{u \in V} d(u)}$ .

*Proof:* Let  $d(v, u) = d(u, v)$  be the number of edges between nodes  $v$  and  $u$  in the union multigraph  $G$ . The sampling process of Algorithm 1 is a Markov chain on  $V$

**Algorithm 1** Multigraph Sampling Algorithm**Require:**  $v_0 \in V$ , simple graphs  $G_i, i = 1 \dots Q$ 

- 1: Initialize  $v \leftarrow v_0$ .
- 2: **while** NOT CONVERGED **do**
- 3:   Select graph  $G_i$  with probability  $\frac{d_i(v)}{d(v)}$
- 4:   Select uniformly at random a neighbor  $v'$  of  $v$  in  $G_i$
- 5:    $v \leftarrow v'$
- 6: **end while**
- 7: **return** all sampled nodes  $v$  and their degrees  $d(v)$ .

with transition probabilities  $P_{vu} = \frac{d(v,u)}{d(v)}$ ,  $u, v \in V$ . So long as  $G$  is finite and connected, this random walk is irreducible and positive recurrent. The presence of a triangle within  $G$  further guarantees aperiodicity.

A Markov chain of this type is equivalent to a random walk on an undirected weighted graph with edge weights  $w(v, u) = d(v, u)$ . A random walk on weighted graph is known to have the unique equilibrium distribution  $\pi(v) = \frac{w(v)}{\sum_{u \in V} w(u)}$ , where  $w(v) = \sum_u w(v, u)$  (e.g., see [9, Example 4.32]), and the proof follows immediately by substitution. ■

**D. Practical Issues**

Various practical issues need to be addressed when implementing these ideas in practice. For completeness, we briefly repeat some good practices here, and we refer the interested reader to our parallel work on this issue [6,10].

1) *Choice of crawling technique:* There are many ways to crawl a multigraph, e.g., by using various random walks or graph traversal techniques. In [10], we showed that a simple random walk with correction for unequal sample weights, also called Re-Weighted Random Walk (RWRW), is more efficient than competitors such as Metropolis-Hastings Random Walks. Therefore, throughout this paper, we employ RWRW described below.

2) *Re-Weighted Random Walk:* Much like a classic random walk on a simple graph, a random walk on multigraph  $G$  is inherently biased towards high-degree nodes. Indeed, per Proposition 2.1, the probability of sampling a node  $v$  is proportional to its degree  $d(v)$  in  $G$ . [11]–[13] show how to apply the Hansen-Hurwitz estimator [14] to correct for this bias. Let  $x(v)$  be an arbitrary function defined on graph nodes  $V$ , with mean  $\bar{x} = \frac{1}{|V|} \sum_{v \in V} x(v)$ . Then

$$\hat{\bar{x}} = \frac{\sum_{v \in S} x(v)/d(v)}{\sum_{v \in S} 1/d(v)} \quad (1)$$

is an unbiased and consistent estimator of  $\bar{x}$ . By default, we use this reweighting procedure throughout the paper (referring to the combination of random walks with post-hoc reweighting as the RWRW method).

3) *Multiple Walks and Convergence Diagnostics:* In previous work [6,10], we recommended the use of multiple, simultaneous random walks to reduce the chance of obtaining samples that overweight non-representative regions of the graph. We also recommended the use of formal convergence diagnostics to assess sample quality in an online fashion, which help to determine when a set of walks is in approximate equilibrium, and hence when it is safe to stop sampling. Use of

both multiple walks and convergence diagnostics are critical to effective sampling of OSNs, as our sample case (Section IV) illustrates.

In this paper, we use three convergence diagnostics, following [6,10]. First, we track the running means for various scalar parameters of interest as a function of the number of iterations. Second, we use the Geweke [15] diagnostic within each random walk, which verifies that mean values for scalar parameters at the beginning of the walk (here the first 10% of samples) does not differ significantly from the corresponding mean at the end of the walk (here the last 50%). Third, we use the Gelman-Rubin [16] diagnostic to verify convergence across walks, by ensuring that the parameter variance between walks matches the variance within walks.

**III. EVALUATION IN SYNTHETIC GRAPHS**

In this section, we use synthetic graphs to demonstrate two key benefits of the multigraph approach, namely (i) improved connectivity of the union multigraph, even when the underlying individual graphs are disconnected, and (ii) improved mixing time, even when the individual graphs are highly clustered. The former is necessary for the random walk to converge. The latter determines the speed of convergence.

**Erdos-Renyi graphs.** In Example II-A and Fig. 2, we noted that even sparse, highly fragmented graphs can have well-connected unions. In Fig. 3, we generalize this example and quantify the benefit of the multigraph approach. We consider here a collection  $G_1, \dots, G_Q$  of  $Q$  Erdos-Renyi random graphs  $(N, p)$  with  $N=1000$  nodes and  $p=1/1000$ , i.e., with the expected number of edges  $|E|=500$  each. We then look at properties of their multigraph  $G$  with increasing numbers of simple graphs  $Q$ .

In order to characterize the connectivity of  $G$ , we define  $f_{LCC}$  as the fraction of nodes that belong to the largest connected component in  $G$ . For  $Q=1$  we have  $f_{LCC} \simeq 0.15$ , which means that each simple ER graph is heavily fragmented. Indeed, at least 999 edges are necessary for connectivity. However, as  $Q$  increases,  $f_{LCC}$  increases. With a relatively small number of simple graphs, say for  $Q=6$ , we get  $f_{LCC} \simeq 1$ , which means that the multigraph is fully connected with high probability. In other words, combining several simple graphs into a multigraph allows us to reach (and sample) many nodes otherwise unreachable.

Note that this example illustrates a more general phenomenon. Given  $Q$  independent random graphs with  $N$  nodes each and with expected densities  $p_1, \dots, p_Q$ , the probability that an edge  $\{u, v\}$  belongs to their union graph  $G'$  is  $p^* = 1 - \prod_{i=1}^Q (1 - p_i)$ . For  $p_i$  approximately equal, this approaches 1 exponentially fast in  $Q$ . Asymptotically, the union graph will be almost surely connected where  $(N-1)p^* > \ln N$  [17, pp413–417], in which case the union multigraph is also trivially connected. Thus, intuitively, a relatively small number of sparse graphs are needed for the union to exceed its connectivity threshold.

In order to characterize the mixing time, we plot the second eigenvalue  $\lambda_2$  of the transition matrix of the random walk (on the LCC).  $\lambda_2$  is well-known to relate to the mixing time of the associated Markov chain [18]: the smaller  $\lambda_2$ , the faster the

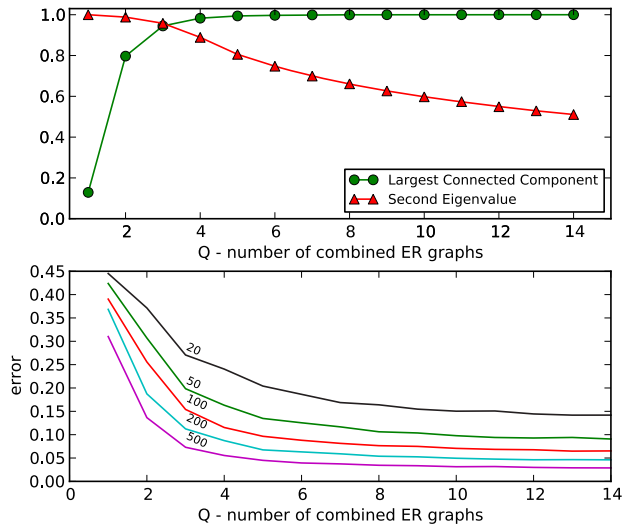
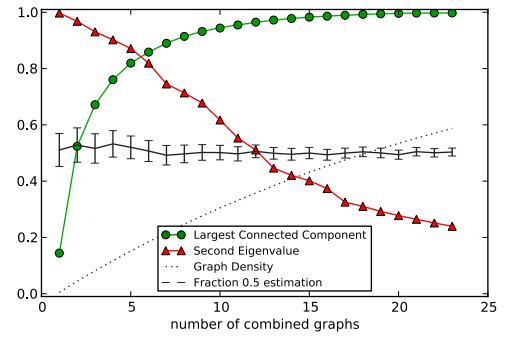


Fig. 3. Multigraph that combines from several Erdos-Renyi graphs. We generate a collection  $G_1, \dots, G_Q$  of  $Q$  random Erdos-Renyi (ER) graphs with  $|V| = 1000$  nodes and expected  $|E| = 500$  edges each. (top) We show two properties of multigraph  $G$  as a function of  $Q$ . (1) Largest Connected Component (LCC) fraction ( $f_{LCC}$ ) is the fraction of nodes that belong to the largest connected component in  $G$ . (2) The second eigenvalue of the transition matrix of random walk on the LCC is related to the mixing time. (bottom) We also label a fraction  $f = 0.5$  of nodes within LCC and run random walks of lengths 20...500 to estimate  $f$ . We show the estimation error (measured in the standard deviation) as a function of  $Q$  (x axis) and walk length (different curves).

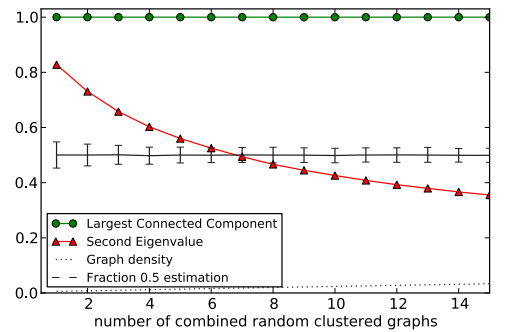
convergence. In Fig. 3(top), we observe that  $\lambda_2$  significantly drops with growing  $Q$ . (However, note that adding a new edge to an existing graph does not always guarantee the decrease of  $\lambda_2$ . It is possible to design examples where  $\lambda_2$  increases, although they are rare.)

To further illustrate the connection between  $\lambda_2$  and the speed of convergence, we conducted an experiment with a simple practical goal: apply random walk to estimate the size of an exogenously defined “community” in the network. We labeled as “community members” a fraction  $f = 0.5$  of nodes within LCC (these nodes were selected with the help of a randomly initiated BFS to better imitate a community). Next, we ran 100 random walks of lengths 20...500 within this LCC, and we used them to estimate  $f$ . In Fig. 3 (bottom), we show the standard error of this estimator, as a function of  $Q$  (x axis) and walk length (different curves). This error decreases not only with the walk length, but also with the number  $Q$  of combined graphs. This means that by using the multigraph sampling approach we improve the quality of our estimates. Alternatively, we may think of it as a way to decrease the sampling cost. For example, in Fig. 3, a random walk of length 500 for  $Q=3$  (i.e., when LCC is already close to 1) is equivalent to a walk of length 100 for  $Q \simeq 8$ , which results in a five-fold reduction of the sampling cost.

**ER Graph Plus Random Cliques.** One may argue that ER graphs are not good models for capturing real-life relations. Indeed, in practice, many relations are highly clustered; e.g., a friend of my friend is likely to be my friend. In an extreme case, all members of some community may form a clique. This is quite common in OSNs, where we are often able to browse all members of a group, or all participants of an event.



(a) Combination of one ER graph ( $|V| = 200$  nodes and  $|E| = 100$  edges) with a set of  $k-1$  cliques (of size 40 randomly chosen nodes each).



(b) Combination of multiple regular random graphs with clustering [19]. We set the parameters such that each of  $|V| = 1000$  nodes has degree equal to 2, and each edge participates in exactly one triangle.

Fig. 4. Multigraphs resulting from a combination of various graphs.

Interestingly, the multigraph technique is efficient also under the presence of cliques. In Fig. 4(a), we consider one ER graph, combined with an increasing number of random cliques. We plot the same three metrics as in Fig 3, and we obtain qualitatively similar results. This robustness is a benefit of the multigraph approach.

**Random Graphs with Clustering.** Finally, in Fig 4(b), we consider a combination of random graphs with clustering [19]. The results confirm our previous observations.

#### IV. MULTIGRAPH SAMPLING OF LAST.FM

In this section, we apply multigraph sampling to `Last.fm`—a music-oriented OSN that allows users to create communities of interest that include both listeners and artists. `Last.fm` is built around an Internet radio service that compiles a preference profile for each listener and recommends users with similar tastes. In June 2010, `Last.fm` was reported to have around 30 million users and was ranked in the top 400 websites in Alexa. We chose `Last.fm` to demonstrate our approach because it provides an example of a popular OSN that is fragmented with respect to the social graph (referred to on the site as “friendship”) as well as other relations. For example, many `Last.fm` users mainly listen to music and do not use the social networking features, which makes it difficult to reach them through crawling the friendship graph; likewise, users with similar music tastes may form clusters that are disconnected from other users with very similar music tastes. This intuition was confirmed by our empirical observations.

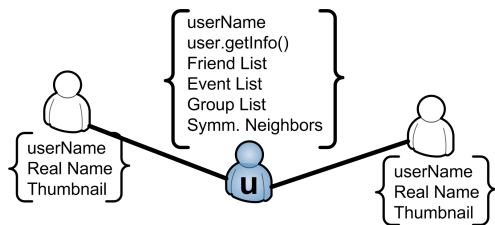


Fig. 5. Information collected for a sampled user  $u$ . (a) `userName` and `user.getInfo()`: Each user is uniquely identified by her `userName`. The API call `user.getInfo` returns: real Name, `userID`, country, age, gender, subscriber, playcount, number of playlists, bootstrap, thumbnail, and user registration time. (b) Friends list: List of mutually declared friendships. (c) Event list. List of past and future events that the user indicates she will attend. We store the `eventID` and number of attendees. (d) Group list. List of groups of which the user is a member. We store the group name and group size. (e) Symmetric neighbors. List of mutual neighbors.

Despite these challenges, we show that multigraph sampling is able to obtain a fairly representative sample in this case, while single graph sampling on any specific relation fails.

#### A. Crawling Last.fm

We sample Last.fm via random walks on several individual relations as well as on their union multigraph. Fig 5 shows the information collected for each sampled user.

1) *Walking on Relations*: We consider the following relations between two users:

- *Friends*: This refers to mutually declared friendship between two users.
- *Groups*: Users with something in common are allowed to start a group. Membership in the same group connects all involved users.
- *Events*: Last.fm allows users to post information on concerts or festivals. Attendees can declare their intention to participate. Attendance in the same event connects all involved users.
- *Neighbors*: Last.fm matches each user with up to 50 similar neighbors based on common activity, membership, and taste. The details of neighbor selection are proprietary. We symmetrize this directed relation by considering only mutual neighbors as adjacent.

First, we collect a sample of users by a random walk on the graph for each individual relation, that is Friends, Groups, Events, and Neighbors. Then, we consider sets of relations (namely: Friends-Events, Friends-Events-Groups, and Friends-Events-Groups-Neighbors) and we perform a random walk on the corresponding union multigraph. In the rest of the section, we refer to random walks on different simple graphs or multigraphs as *crawl types*.

2) *Uniform Sample of userIDs (UNI)*: Last.fm usernames uniquely identify users in the API and HTML interface. However, internally, Last.fm associates each username with a `userID`, presumably used to store user information in the internal database. We discovered that it is possible to obtain usernames from their `userIDs`, a fact that allowed us to obtain a uniform, “ground truth” sample of the user population. Examination of registration and ID information indicates that Last.fm allocates `userIDs` in an increasing order. Fig. 6 shows the exact

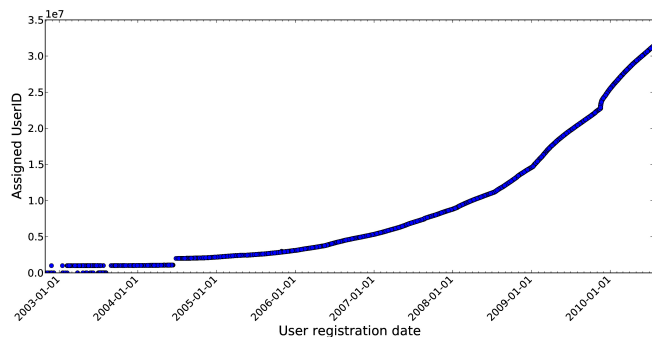


Fig. 6. Last.fm assigns `userIDs` in increasing order after 2005: `userID` vs registration time.

registration date and the assigned `userID` for each sampled user in our crawls obtained through exploration. With the exception of the first  $\sim 2M$  users (registered in the first 2 years of the service), for every  $userID_1 > userID_2$  we have  $registration\_time(userID_1) > registration\_time(userID_2)$ . We also believe that `userIDs` are assigned sequentially because we rarely observe non-existent `userIDs` after the  $\sim 2,000,000$  threshold. We conjecture that the few non-existent `userIDs` after this threshold are closed or banned accounts. At the beginning of the crawl, we found no indication of user accounts with IDs above  $\sim 31,200,000$ . Just before the crawls, we registered new users that were assigned user IDs slightly higher than the latter value.

Using the `userID` mechanism, we obtained a reference sample of Last.fm users by uniform rejection sampling [20]. Specifically, each user was sampled by repeatedly drawing uniform integers between 0 and 35 million (*i.e.*, the maximum observed ID plus a  $\sim 4$  million “safety” range) and querying the `userID` space. Integers not corresponding to a valid `userID` were discarded, with the process being repeated until a match was obtained. IDs obtained in this way are uniformly sampled from the space of user accounts, irrespective of how IDs are actually allocated within the address space [21]. We employ this procedure to obtain a sample of 500K users, referred to here as “UNI.” We note that the same method has been recently used in [6], as well as in [22]. The latter also examined population growth and active vs. inactive users, which are out of the scope of this paper.

Although UNI sampling currently solves the problem of uniform node sampling in Last.fm and is a valuable asset for this study, it is not a general solution for sampling OSNs. Such an operation is not generally supported by OSNs. Furthermore, the `userID` space must not be sparse for this operation to be efficient. In the Last.fm case, the small `userID` space makes this possible at the time of this writing; however, a simple increase of the `userID` space to 48 or 64 bits would render the technique infeasible. In summary, we were able to obtain a uniform sampling of `userIDs` and use it as a baseline for evaluating the sampling methods of interest against the target distribution.

3) *Estimating Last.fm population size*: In addition to the UNI sample presented in Table I, we obtained a second UNI sample of the same size one week later. We then applied the

TABLE I  
SUMMARY OF COLLECTED DATASETS IN JULY 2010. THE PERCENTAGE OF USERS KEPT IS DETERMINED FROM CONVERGENCE DIAGNOSTICS. AVERAGES SHOWN ARE AFTER CONVERGENCE AND RE-WEIGHTING WHICH CORRECTS SAMPLING BIAS.

Crawltype	Friends	Events	Groups	Neighbors	Friends-Events	Friends-Events-Groups	Friends-Events-Groups-Neighbors	UNI
# Total Users	$5 \times 50K$	$5 \times 50K$	$5 \times 50K$	$5 \times 50K$	$5 \times 50K$	$5 \times 50K$	$5 \times 50K$	500K
% Unique users	71.0%	58.5%	74.3%	53.1%	59.4%	75.5%	75.6%	99.1%
# Users kept	245K	245K	245K	245K	200K	187K	200K	500K
Crawling period	07/13-07/16	07/13-07/18	07/13-07/17	07/13-07/17	07/13-07/18	07/13-07/18	07/13-07/21	07/13-07/16
Avg # friends	10.7	18.0	15.8	12.2	9.8	6.8	6.6	1.2
Avg # groups	2.40	4.71	5.22	2.90	2.47	0.71	0.67	0.30
Avg # events (past/future)	2.44/0.17	7.49/0.56	3.96/0.28	2.94/0.27	2.30/0.17	0.74/0.05	0.73/0.04	0.28/0.02

capture-recapture method [23] to estimate the Last.fm user population during the period of our crawling. According to this method, the population size is estimated to be :

$$P_{\text{Last.fm}} = \frac{N_{\text{UNI1}} \times N_{\text{UNI2}}}{R} = 28.5M,$$

where  $N_{\text{UNI1}} = N_{\text{UNI2}} = 500K$  and  $R$  is the number of valid common userIDs sampled during the first and second UNI samples. This estimation is consistent with our observations of the maximum userID space and close to the reported size of Last.fm on various Internet websites.

4) *Topology Change During Sampling Process*: While substantial change could in theory affect the estimation process, Last.fm evolves very little during the duration of our crawls. The increasing-order userID assignment allows us to infer the maximum growth of Last.fm during this period, which we estimate at 25K/day on average. Therefore, with a population increase of 0.09%/day, the user growth during our crawls (2-7 days) is calculated to range between 0.18% – 0.63% per crawl type, which is quite small. Furthermore, the comparison between the two UNI samples revealed almost identical distributions for the properties studied here, as shown in Fig 7. Therefore, in the rest of the paper, we assume that any changes in the Last.fm network during the crawling period can be ignored. This is unlike the context of dynamic graphs, where considering the dynamics is essential, *e.g.*, see [7,24,25].

5) *Efficient Multigraph Sampling in Last.fm* : To collect data from Last.fm, we use a combination of API calls and data scraping. Consider that we are sampling user  $u$ . For efficient implementation of multigraph sampling we proceed in two stages, as shown in Fig 1(f).

In the first stage, we discover the graphs of user  $u$ , and  $u$ 's degrees in them. In our study, we use the API calls `user.getfriends`, `user.getneighbors`, `user.getpastevents`, and `user.getevents` to collect the list of friends, neighbors, past events, and future events respectively. Due to a lack of an API call that lists the groups of a user, we use data scraping to collect the list of groups and corresponding size for each group. We treat each individual group and event as a different graph in the multigraph. We also consider the set of friends and neighbors to comprise the friends and neighbors graph respectively in the multigraph. At the end of the first stage, we select one of the graphs  $G_i$  in accordance with Algorithm 1 in Section II.

We should note that at the end of the first stage, we have not enumerated any user from any of the groups and events graphs. Each of these graphs is quite large (up to tens of thousands

of users) and depending on the user, there are many groups or events per user (up to thousands). On the other hand, we have enumerated users of friends and neighbors since knowledge of neighborhood size is equivalent to enumeration<sup>1</sup> for these graphs. Overall, our two stage approach saves us bandwidth and time by avoiding the enumeration of users for graphs that we are not going to sample from at each iteration.

In the second stage, we pick uniformly at random one of the nodes from the graph  $G_i$ , selected at the end of the first stage. If the graph  $G_i$  is a graph of a group or an event, we need to carefully implement this action to be efficient. More specifically, we do not need to enumerate all group members or event attendants from a group or event graph. Instead, we can take advantage of the pages functionality that OSNs often provide and only fetch the page that corresponds to the user selected uniformly at random. In our study, to fetch group members we use the API call `group.getmembers`, which returns 50 users per page. To fetch event attendants we use data scraping<sup>2</sup>, which also returns 50 users per HTML page.

6) *Data Collection*: We used a cluster of machines to execute all crawl types under comparison simultaneously. For each crawl type, we run  $|V_0| = 5$  different independent walks. The starting points for the five walks, in each crawl type, are randomly selected users identified by the web site as listeners of songs from each of five different music genres: country, hip hop, jazz, pop and rock. This set was chosen to provide an overdispersed seed set, while not relying on any special-purpose methods (*e.g.*, UNI sampling). To ensure that differences in outcomes do not result from choice of seeds, the same seed users are used for all crawl types. We let each independent crawl continue until we determine convergence per walk and per crawl, using online diagnostics as introduced in [6] and described in Section II-D3. Eventually, we collected exactly 50K samples for each random walk crawl type. Finally, we collect a UNI sample of 500K users.

7) *Summary of Collected Datasets*: Table I summarizes the collected datasets. Each crawl type contains  $5 \times 50K = 250K$  users. We observe that there is a large number of repetitions in the random walks of each crawl type, ranging from 25% (in Friends-Events-Groups-Neighbors) to 47% (in Neighbors).

<sup>1</sup>There might be workarounds to enumerating friends but they are not necessarily more efficient. For example, we could extract the number of friends by data scraping. In general, in another setting we could do away with any kind of enumeration in the first stage.

<sup>2</sup>We prefer data scraping to the API call `event.getattendees` because the API call i) is not pagged ii) does not return users that marked "maybe" for the event iii) is very slow for large events.

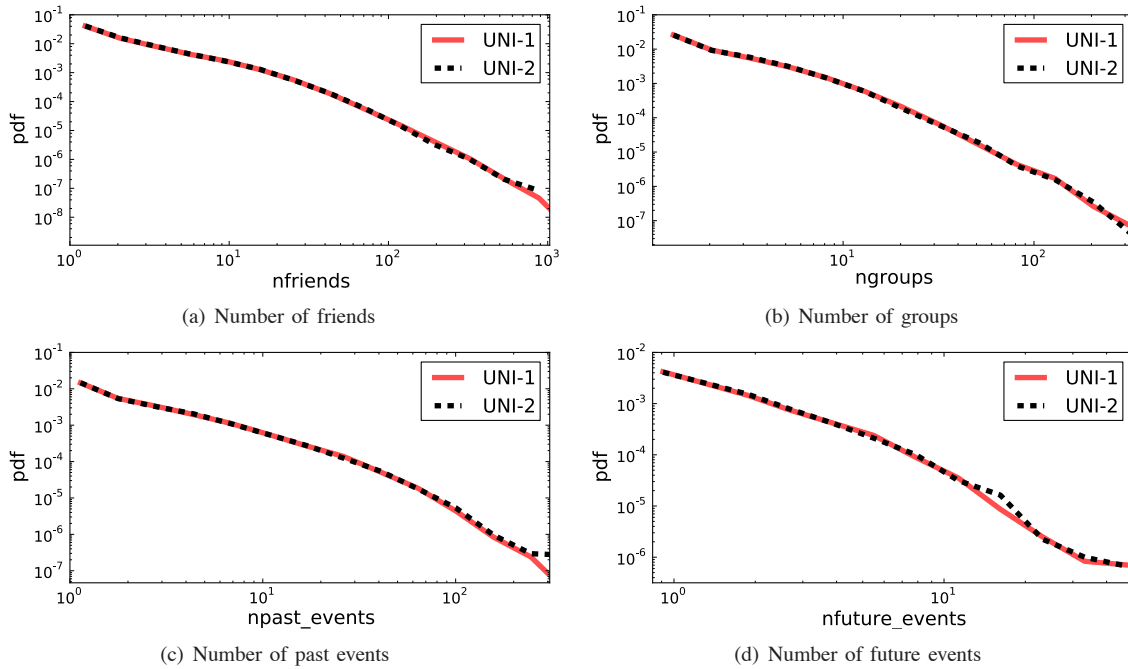


Fig. 7. Probability distribution function (PDF) of UNI samples obtained one week apart from each other. Results are binned.

TABLE II

PERCENTAGE OF TIME A PARTICULAR GRAPH (EDGES CORRESPONDING TO THIS GRAPH) IS USED DURING THE CRAWL BY ALGORITHM 1

Crawl Type	Friends	Events	Groups	Neighbors
	Graph	Graphs	Graphs	Graphs
Friends	100%	0%	0%	0%
Events	0%	100%	0%	0%
Groups	0%	0%	100%	0%
Neighbors	0%	0%	0%	100%
Friends-Events	2.2%	97.8%	0%	0%
Friends-Events-Groups	0.3%	5.4%	94.3%	0%
Friends-Events-Groups-Neighbors	0.3%	5.5%	94.2%	0.02%

This appears to stem from the high levels of clustering observed in the individual networks. It is also interesting to note that the crawling on the multigraph Friends-Events-Groups-Neighbors is able to reach more unique nodes than any of the single graph crawls.

Table II shows the fraction of Markov chain transitions using each individual relation. The results for the single-graph crawl types Friends, Events, Groups, and Neighbors are as expected: they use their own edges 100% of the time and other relations' 0%. Besides that, we see that Events relations dominate Friends when they are combined in a multigraph, and Groups dominate Friends, Events, and Neighbors when combined with them. This occurs because many groups and events are quite large (hundreds or thousands of users), leading participants to have very high relationship-specific degree for purposes of Algorithm 1, and thus for the Group or Event relations to be chosen more frequently than low-degree relations like Friends. In the crawl types obtained through a random walk, the highest overlap of users is observed between Groups and Friends-Events-Groups-Neighbors (66K) while the lowest is between Neighbors and Friends-Events (5K). It is noteworthy that despite the dominance of Groups and the high overlap between Groups and Friends-Events-Groups-

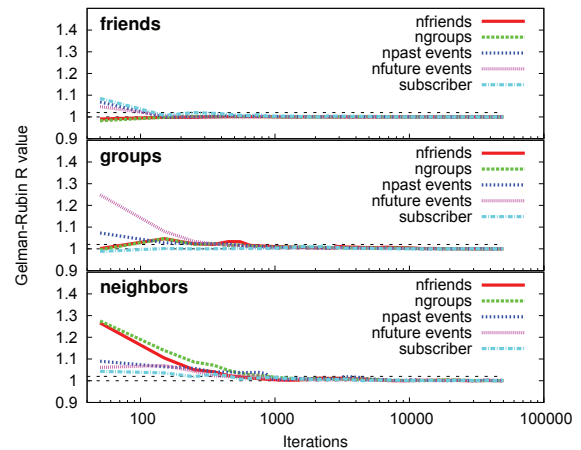


Fig. 8. Convergence diagnostic tests w.r.t. to four different user properties (“nfriends”: number of friends, “ngroups”: number of groups, “npast\_events”: number of past events, “nfuture\_events”: number of future events, and “subscriber”) and three different crawl types (Friends, Groups, Neighbors).

Neighbors, the aggregates for these two crawl types in Table I lead to very different samples of users.

### B. Evaluation Results

1) *Convergence: Burn-in.* To determine the burn-in for each crawl type in Table I, we run the Geweke diagnostic separately on each of its 5 chains, and the Gelman-Rubin diagnostic across all 5 chains at once, for several different properties of interest. The Geweke diagnostic shows that first-order convergence is achieved within each walk after approximately 500 iterations at maximum. For the single relation crawl types (Friends, Events, Groups, and Neighbors), the Gelman-Rubin diagnostic indicates that convergence is attained within 1000 iterations per walk (target value for R below 1.02 and close to 1) as shown in Fig. 8.

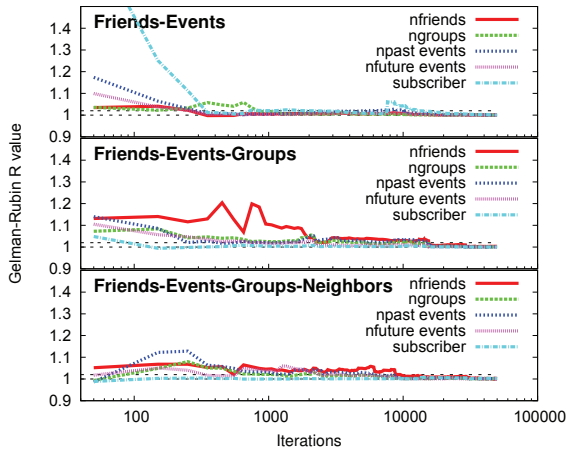


Fig. 9. Convergence diagnostic tests w.r.t. to four different user properties (“nfriends”: number of friends, “ngroups”: number of groups, “npast\_events”: number of past events, “nfuture\_events”: number of future events, and “subscriber”) and three different crawl types (Friends-Events, Friends-Events-Groups, Friends-Events-Groups-Neighbors).

On the other hand, multigraph crawl types take longer to reach equilibrium. Fig. 9 presents the Gelman-Rubin (R) score for three multigraph crawl types (namely Friends-Events, Friends-Events-Groups, Friends-Events-Groups-Neighbors) and five user properties (namely number of friends, number of groups, number of past/future events, and subscriber - a binary value which indicates whether the user is a paid customer). We observe that it takes 10K, 12.5K and 10K samples for each crawl type correspondingly to converge. However, as we show next, they include more isolated users and better reflect the ground truth, while the single graph sampling methods fail to do so. This underscores an important point regarding convergence diagnostics: while useful for determining whether a random walk sample approximates its equilibrium distribution, they cannot reliably identify cases in which the equilibrium itself is biased (*e.g.*, due to non-connectivity). For the rest of the analysis, we discard the number of samples each crawl type needed to converge.

*Total Running Time.* Before we analyze the collected datasets, we verify that the remaining walk samples, after discarding burn-in, have reached their stationary distribution. Table I contains the “Number of users kept” for each crawl type. We use the convergence diagnostics on the remaining samples to assess convergence formally. The results are qualitatively similar to the burn-in determination section. We also perform visual inspection of the running means in Fig 10 for four different properties, which reveals that the estimation of the average for each property stabilizes within 2-4k samples per walk (or 10k-20k over all 5 walks).

2) *Discovering Isolated Components:* As noted above, part of our motivation for sampling Last.fm using multigraph methods stems from its status as a fragmented network with a rich multigraph structure. In particular, we expected that large parts of the user base would not be reachable from the largest connected component in any one graph. Such users could consist of either isolated individuals or highly clustered sets of users lacking ties to rest of the network. We here call *isolate*, any user that has degree 0 in a particular graph

TABLE III  
PERCENTAGE OF SAMPLED NODES THAT ARE ISOLATES (HAVE DEGREE 0) W.R.T. TO A PARTICULAR (MULTI)GRAPH.

Crawl Type	Friends	Future	Past	Groups
	Isolates	Events	Events	Isolates
Friends	0%	93.7%	73.2%	60.4%
Events	19.2%	78.2%	4.5%	41.7%
Groups	21.2%	89.9%	62.0%	0.0%
Neighbors	40.4%	89.5%	71.2%	62.4%
Friends-Events	6.2%	93.5%	69.9%	61.6%
Friends-Events-Groups	5.5%	98.15%	88.1%	85.3%
Friends-Events-Groups-Neighbors	7.4%	98.3%	86.7%	86.3%
UNI	87.9%	99.2%	96.1%	93.8%

relation. Walk-based sampling on that particular graph relation has no way of reaching those isolates, but a combination of graphs might be able to reach them, assuming that a typical user participates in different ways in the network (*e.g.*, a user with no friends may still belong to a group or attend an event).

In Table III, we report the percentage of nodes in each crawl type that are estimated to be isolates, and compare this percentage to the UNI sample. Observe that there is an extremely high percentage of isolate users in any single graph: *e.g.*, UNI samples are 88% isolates in the Friends relation, 96-99% isolates in the Events relation, and 93.8% isolates in the Groups relation. Such isolates are not necessarily inactive: for instance, 59% of users without friends have either a positive playcount or playlist value, which means that they have played music (or recorded their offline playlists) in Last.fm, and hence are or have been active users of the site. This confirms our expectation that Last.fm is indeed a fragmented graph.

More importantly, Table III allows us to assess how well different crawl types estimate the % of users that are isolates with respect to a particular relation or set of relations. We observe that the multigraph that includes all relations (Friends-Events-Groups-Neighbors) leads to the best estimate of the ground truth (UNI sample - shown in the last row). The only exception is the friends isolates, where the single graph Neighbors gives a better estimate of the percentage of isolates over all other crawl types. The multigraph crawl type Friends-Events-Groups-Neighbors uses the Neighbors relation only 0.02% of of the time, and thus does not benefit as much as might be expected (though see below). A weighted random walk that put more emphasis on this relation (or use of a relation that is less sparse) could potentially improve performance in this respect.

### 3) Comparing Samples to Ground Truth:

*I. Comparing to UNI.* In Table III, we saw that multigraph sampling was able to better approximate the percentage of isolates in the population. Here we consider other user properties (namely number of friends, past events, and groups a user belongs to, and whether he or she is a subscriber to Last.fm). In Fig 10, we plot the sample mean value for four user properties across iteration number, for all crawl types and for the ground truth (UNI). One can see that crawling on a single graph, *i.e.*, Friends, Events, Groups, or Neighbors alone, leads to poor estimates. This is prefigured by the previous results, as single graph crawls undersample individuals such as isolates on their corresponding relation,

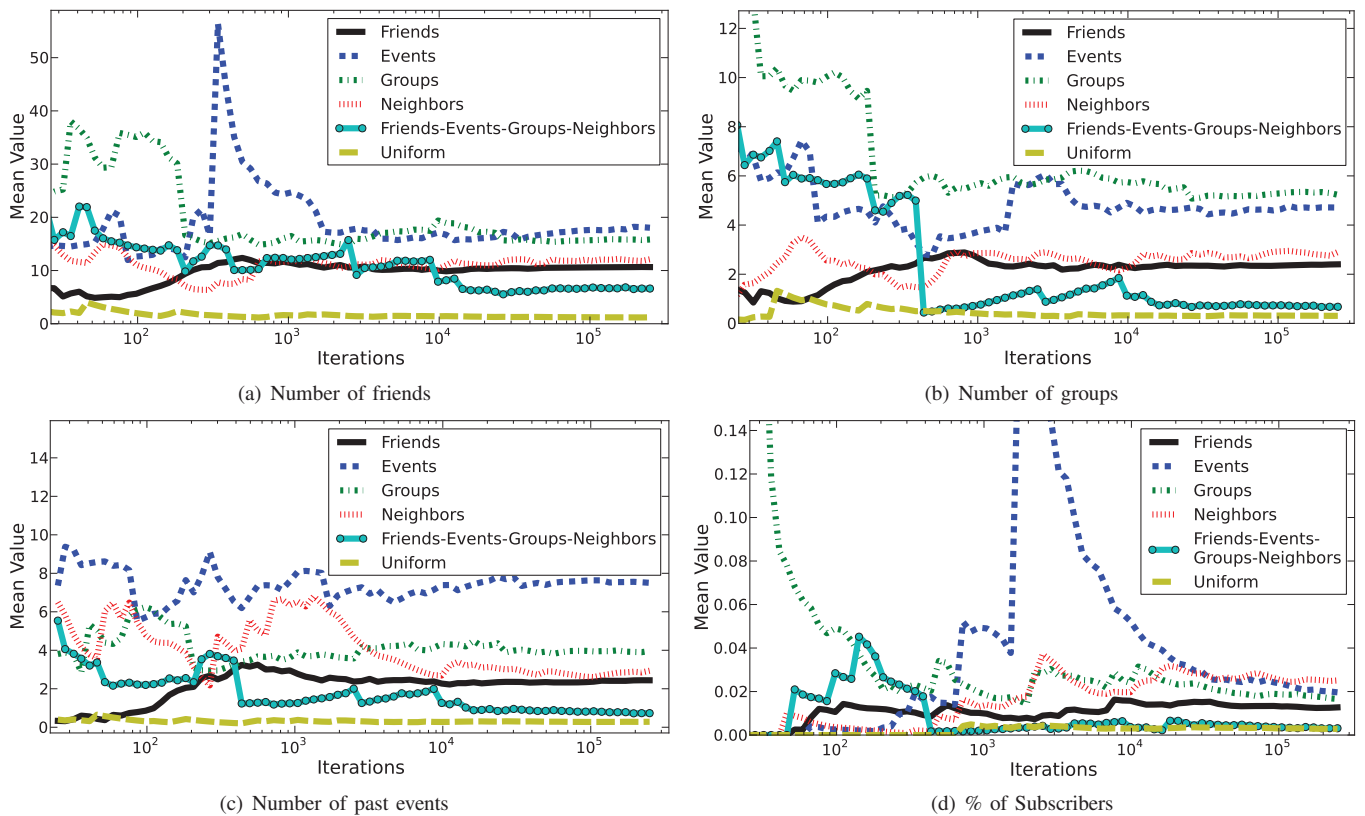


Fig. 10. Single graph vs multigraph sampling. Sample mean over the number of iterations, for four user properties (number of friends, number of groups, number of past events, % subscribers), as estimated by different crawl types.

who form a large portion of the population. We also notice that Events and Groups alone consistently overestimate the averages, as these tend to cover the most active portion of the user base. However combining them together with other relations helps considerably. The multigraph that utilizes all relations, Friends-Events-Groups-Neighbors, is the closest to the truth. For example, it approximates very closely the avg number of groups and % of paid subscribers (Figs 10(b) 10(d)).

In Fig 11, we plot the probability distributions for four user properties of interest. Again, the crawl type Friends-Events-Groups-Neighbors is closest to the ground truth, in terms of shape of the distribution and vertical distance to it. Nevertheless, we observe that in both the probability distribution and the running mean plots, there is sometimes a gap from UNI, which is caused by the imperfect approximation of the % of isolates. That is the reason that the gap is the largest for the number of friends property (Fig 10(a), 11(a)).

*II. Comparing to Weekly Charts.* Finally, we compare the estimates obtained by different crawl types, to a different source of the ground truth - the weekly charts posted by Last.fm. This is useful as an example of how one can (at least approximately) validate the representativeness of a random walk sample in the absence of a known uniform reference sample.

Last.fm reports on its website weekly music charts and statistics, generated automatically from user activity. To mention a few examples, “Weekly Top Artists” and “Weekly Top Tracks” as well as “Top Tags“, “Loved Tracks“ are reported.

Each chart is based on the actual number of people listening to the track, album or artist recorded either through an Audio-scrobber plug-in (a free tracking service provided by the site) or the Last.fm radio stream. To validate the performance of multigraph sampling, we estimate the charts of “Weekly Top Artists” and “Weekly Top Tracks” from our sample of users for each of the crawl types in Table I, and we compare it to the published charts for the week July 04-July 11 2010, *i.e.*, the week just before the crawling started. To generate the charts from our user samples, we utilize API functions that allow us to fetch the exact list of artists and tracks that a user listened during a given date range. Fig. 12 shows the observed artist/track popularity rank and the percentage of listeners for the top 420 tracks/artists (the maximum available) from the Last.fm Charts, with the estimated ranks and percentage of listeners for the same tracks/artists in each crawl type. As can be seen, the rank curve estimated from the multigraph Friends-Events-Groups-Neighbors tracks quite well the actual rank curve. Additionally, the curve that corresponds to the UNI sample is virtually lying on top of the “Last.fm Charts” line. On the other hand, the single graph crawl types Friends, Events, Groups, and Neighbors are quite far from actual charts. Here, as elsewhere, combining multiple relations gets us much closer to the truth than would reliance on a single graph.

## V. RELATED WORK

Early graph exploration methods that were used to measure OSNs were based on BFS and snowball sampling [1]–[3]. These methods have been shown to have a generally unknown

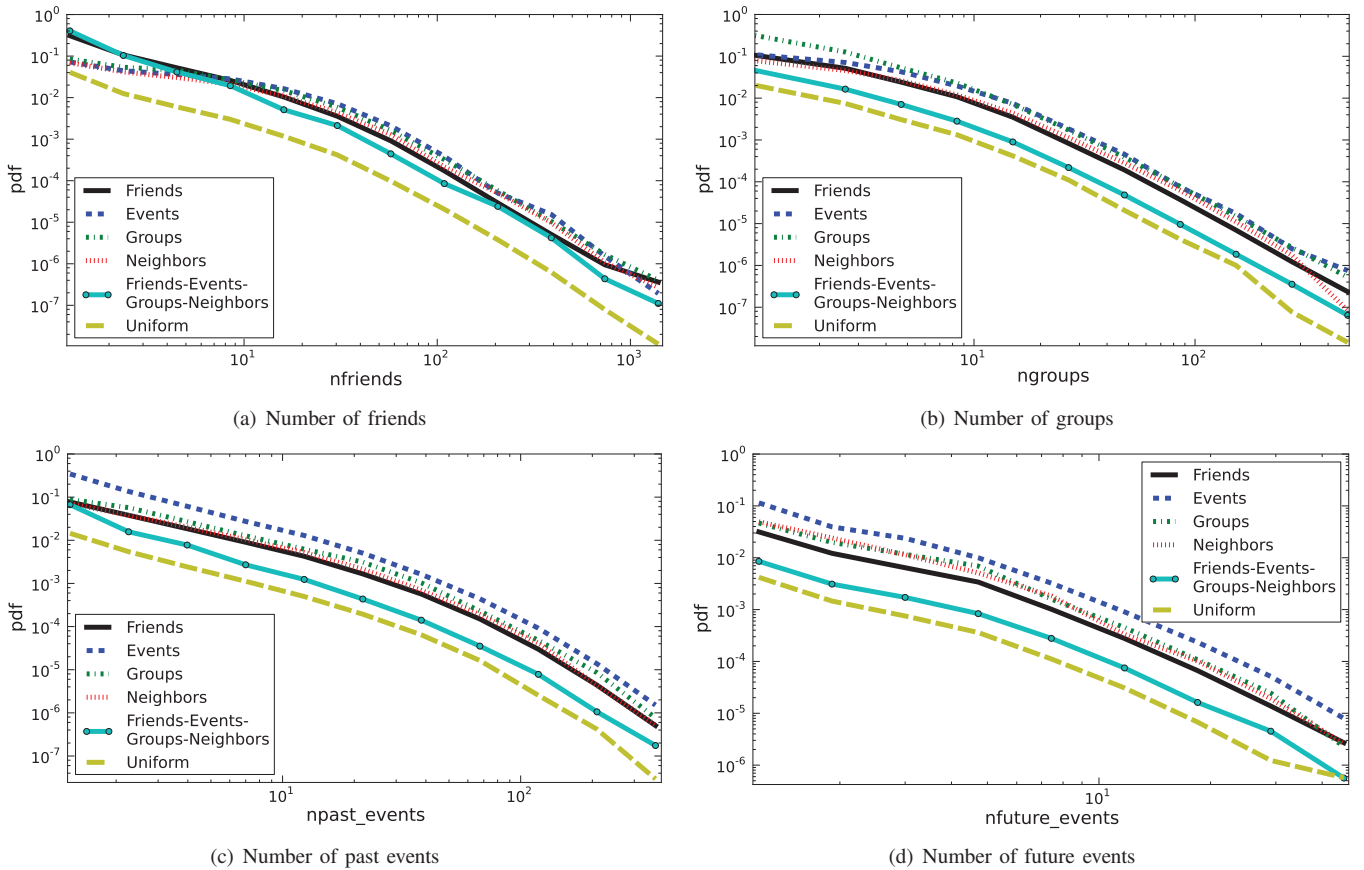


Fig. 11. Single graph vs multigraph sampling. Probability distribution function (pdf) for three user properties (number of friends, number of groups, number of past events), as estimated by different crawl types.

bias towards high degree nodes when far from completion. In our recent and ongoing work, we attempt to correct for this bias [5,26]; however, BFS is out of the scope of this paper. Recent work in [6,7,27] used random walks (where the bias is known) to sample users in OSNs, namely Friendster, Twitter and Facebook. Random walks have also been used to sample peer-to-peer networks [28]–[30] and other large graphs [31].

Design of random walk techniques to improve mixing include [18,32]–[34]. Boyd *et al.* [18] pose the problem of finding the fastest mixing Markov Chain on a known graph as an optimization problem. However, in our case such an exact optimization is not possible since we are exploring an unknown graph. Ribeiro *et al.* [32] introduce Frontier sampling and explore multiple dependent random walks to improve sampling in disconnected or loosely connected sub-graphs. Multigraph sampling has the same goal but instead achieves it by exploring the social graph using multiple relations. Therefore, Frontier sampling is an orthogonal idea, which can potentially be combined with multigraph sampling for additional benefits. Multigraph sampling is also remotely related to techniques in the MCMC literature (*e.g.*, Metropolis-coupled MCMC or simulated tempering [33]) that seek to improve Markov chain convergence by mixing states across multiple chains with distinct stationary distributions. In [34,35] Thompson *et al.* introduce a family of adaptive cluster sampling (ACS) schemes, which are designed to explore nodes that satisfy some condition of interest; although random walk

sampling is distinct from cluster sampling, the former does fit more broadly within the area of adaptive designs.

As noted in Section IV-A4, we consider that the network of interest remains static during the duration of the crawl. We confirmed that this is a good approximation in the case of Last.fm by comparing two snapshots taken one week apart. Therefore, in this work, we do not consider dynamics, which are essential in other contexts [7,24,25].

Recent data collection studies of Last.fm include: [36], which develops a track recommendation system using social tags and friendship between users, [37], which examines user similarity to predict social links, and [38], which explores the meaning of friendship in Last.fm through survey sampling. We emphasize that the importance of a representative sample is crucial to the usefulness of such datasets.

In our previous work [6] and its extended version [10], we proposed a framework for crawling a single graph. In the implementation part of this paper, we adopt some of the practical recommendations of that work (*e.g.*, the use of the RWRW as the preferred crawling technique, the use of online convergence diagnostics, etc). However, our focus here is on comparing multigraph sampling vs. single graph sampling, and on demonstrating its utility on fragmented networks such as Last.fm. To the best of our knowledge, our work is the first to explore sampling OSNs on a combination of multiple relations.

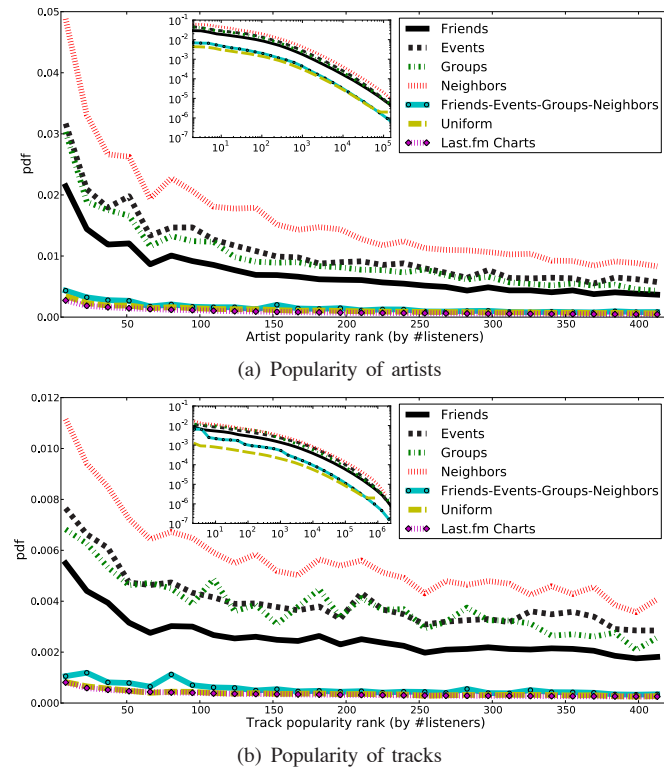


Fig. 12. Weekly Charts for the week 07/04-07/11. Artists/tracks are retrieved from “Last.fm Charts” and remain the same for all crawl types. Data is linearly binned (30 points). Inset: Artist/track popularity rank and percentage of listeners for all artists/tracks encountered in each crawl type.

## VI. CONCLUSION

In this paper, we have introduced *multigraph sampling* - a novel technique for random walk sampling of OSNs using multiple underlying relations. Multigraph sampling generates probability samples in the same manner as conventional random walk methods, but is more robust to poor connectivity and clustering within individual relations. As we demonstrate using the Last.fm service, multigraph methods can give reasonable approximations to uniform sampling even where the overwhelming majority of users in each underlying relation are isolates, thus making single-graph methods fail. Our experiments with synthetic graphs also suggest that multigraph sampling can improve the coverage and the convergence time for partitioned or highly clustered networks. Given these advantages, we believe multigraph sampling to be a useful addition to the growing suite of methods for sampling OSNs.

The focus of this paper was on (i) demonstrating the utility of multigraph sampling compared to single graph sampling and (ii) on the design of a two-stage efficient algorithm that implements the idea.

Open questions include the selection of a few -out of many candidate- relations to use when sampling, so as to optimize the multigraph sampler performance. Intuitively, we expect that negatively correlated relations will prove most effective. A related question is the weighting of the different relations for the same purpose. Gaining intuition into these problems will be particularly helpful in designing optimal OSN sampling schemes and is a direction for future work.

## REFERENCES

- [1] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, “Analysis of topological characteristics of huge online social networking services,” in *Proc. 16th Int. Conf. on World Wide Web*, Banff, Alberta, Canada, 2007, pp. 835–844.
- [2] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, “Measurement and analysis of online social networks,” in *Proc. 7th ACM SIGCOMM Conf. on Internet measurement*, San Diego, CA, 2007, pp. 29–42.
- [3] C. Wilson, B. Boe, A. Sala, K. Puttaswamy, and B. Zhao, “User interactions in social networks and their implications,” in *Proc. 4th ACM European Conf. on Computer systems*, Nuremberg, Germany, 2009, pp. 205–218.
- [4] S. H. Lee, P.-J. Kim, and H. Jeong, “Statistical properties of sampled networks,” *Physical Review E*, vol. 73, p. 16102, 2006.
- [5] M. Kurant, A. Markopoulou, and P. Thiran, “On the bias of BFS (Breadth First Search),” in *Proc. 22nd Int. Teletraffic Congr., also in arXiv:1004.1729*, 2010.
- [6] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, “Walking in Facebook: A Case Study of Unbiased Sampling of OSNs,” in *Proc. IEEE INFOCOM*, San Diego, CA, 2010.
- [7] A. H. Rasti, M. Torkjazi, R. Rejaie, and D. Stutzbach, “Evaluating Sampling Techniques for Large Dynamic Graphs,” *Univ. Oregon, Tech. Rep. CIS-TR-08-01*, Sep. 2008.
- [8] D. D. Heckathorn, “Respondent-Driven Sampling II: Deriving Valid Estimates from Chain-Referral Samples of Hidden Populations,” *Social Problems*, vol. 49, pp. 11–34, 2002.
- [9] S. Ross, *Introduction to probability models*. Academic Press, 2003, vol. 8.
- [10] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, “Practical Recommendations on Sampling OSN Users by Crawling the Social Graph,” *To appear in IEEE J. Sel. Areas Commun. on Measurement of Internet Topologies*, 2011.
- [11] D. D. Heckathorn, “Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations,” *Social Problems*, vol. 44, pp. 174–199, 1997.
- [12] M. Salganik and D. D. Heckathorn, “Sampling and estimation in hidden populations using respondent-driven sampling,” *Sociological Methodology*, vol. 34, no. 1, pp. 193–240, 2004.
- [13] M. Newman, “Ego-centered networks and the ripple effect,” *Social Networks*, vol. 25, pp. 83–95, 2003.
- [14] M. Hansen and W. Hurwitz, “On the Theory of Sampling from Finite Populations,” *Annals of Mathematical Statistics*, vol. 14, no. 3, 1943.
- [15] J. Geweke, “Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments,” in *Bayesian Statistics*, 1992, pp. 169–193.
- [16] A. Gelman and D. Rubin, “Inference from iterative simulation using multiple sequences,” in *Statistical science*, vol. 7, no. 4, 1992, pp. 457–472.
- [17] D. West, *Introduction to graph theory*. Prentice Hall, 1996.
- [18] S. Boyd, P. Diaconis, and L. Xiao, “Fastest mixing Markov chain on a graph,” *SIAM review*, vol. 46, no. 4, pp. 667–689, 2004.
- [19] M. Newman, “Random graphs with clustering,” *Physical review letters*, vol. 103, no. 5, p. 58701, 2009.
- [20] A. Leon-Garcia, *Probability, statistics, and random processes for electrical engineering*. Pearson/Prentice Hall, 2008.
- [21] M. Gjoka, C. T. Butts, M. Kurant, and A. Markopoulou, “Multigraph Sampling of Online Social Networks,” *To appear in IEEE J. Sel. Areas Commun. on Measurement of Internet Topologies*, 2011.
- [22] R. Rejaie, M. Torkjazi, M. Valafar, and W. Willinger, “Sizing up online social networks,” *IEEE Network*, vol. 24, no. 5, pp. 32–37, 2010.
- [23] G. Seber, “The estimation of animal abundance and related parameters,” *New York*, 1982.
- [24] W. Willinger, R. Rejaie, M. Torkjazi, M. Valafar, and M. Maggioni, “OSN Research: Time to face the real challenges,” in *Proc. 2nd Workshop on Hot Topics in Measurement & Modeling of Computer Systems*, Seattle, WA, 2009.
- [25] U. Acer, P. Drineas, and A. Abouzeid, “Random walks in time-graphs,” in *Proc. 2nd Int. Workshop on Mobile Opportunistic Networking*, 2010, pp. 93–100.
- [26] M. Kurant, A. Markopoulou, and P. Thiran, “Towards Unbiased BFS Sampling,” *To appear in IEEE J. Sel. Areas Commun. on Measurement of Internet Topologies*, 2011.
- [27] B. Krishnamurthy, P. Gill, and M. Arlitt, “A few chirps about twitter,” in *Proc. 1st workshop on Online social networks*, Seattle, WA, 2008, pp. 19–24.

- [28] C. Gkantsidis, M. Mihail, and A. Saberi, "Random walks in peer-to-peer networks," in *Proc. IEEE INFOCOM*, Hong Kong, China, 2004.
- [29] A. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach, "Respondent-driven sampling for characterizing unstructured overlays," in *Proc. IEEE INFOCOM Mini-conference*, Rio de Janeiro, Brazil, 2009, pp. 2701–2705.
- [30] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger, "On unbiased sampling for unstructured peer-to-peer networks," in *Proc. 6th ACM SIGCOMM Conf. on Internet measurement*, Rio de Janeiro, Brazil, 2006.
- [31] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proc. 12th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, Philadelphia, PA, 2006, pp. 631–636.
- [32] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *Proc. 10th ACM SIGCOMM Conf. on Internet measurement*, Melbourne, Australia, 2010.
- [33] W. Gilks and G. Roberts, "Strategies for improving MCMC," *Markov chain Monte Carlo in practice*, pp. 89–114, 1996.
- [34] S. Thompson, "Adaptive cluster sampling," *Journal of the American Statistical Association*, vol. 85, no. 412, pp. 1050–1059, 1990.
- [35] —, "Stratified Adaptive Cluster Sampling," *Biometrika*, vol. 78, no. 2, pp. 389–397, 1991.
- [36] I. Konstas, V. Stathopoulos, and J. Jose, "On social networks and collaborative recommendation," in *Proc. 32nd Int. ACM SIGIR Conf. on Research and development in information retrieval*, Boston, MA, 2009, pp. 195–202.
- [37] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer, "Folks in folksonomies: social link prediction from shared metadata," in *Proc. 3rd ACM Int. Conf. on Web search and data mining*, New York, NY, 2010, pp. 271–280.
- [38] N. Baym and A. Ledbetter, "Tunes that bind? Predicting friendship strength in a music-based social network," *Information, Communication and Society*, vol. 12, no. 3, 2009.



**Carter T. Butts** received his B.S. degree from Duke University in 1996, and his Ph.D. from Carnegie Mellon University in 2002. In 2002 he joined the Department of Sociology at the University of California, Irvine, where he is currently associate professor; he holds appointments in the Departments of Sociology and Statistics, and is a member of the Institute for Mathematical Behavioral Sciences and the California Institute for Telecommunications and Information Technology. He currently serves as an area editor for Computational and Mathematical Organization Theory, and serves on the editorial boards of the Journal of Mathematical Sociology and Sociological Methodology. He holds several paper awards from the American Sociological Association, was the 2010 recipient of the Leo A. Goodman early career award from the ASA Section on Methodology, and was the 2008 recipient of the Linton C. Freeman Award from the International Network for Social Network Analysis. His current research interests include the structure of spatially embedded large-scale interpersonal networks; models for informant accuracy, network inference, and graph comparison; representation and modeling of intertemporal relational data; and measurement and modeling of online social networks.



**Maciej Kurant** received a M.Sc. degree from Gdansk University of Technology, Poland, in 2002, and a Ph.D. degree from EPFL, Lausanne, Switzerland, in 2009. Currently, he is a postdoc at University of California, Irvine. His main areas of research interest include sampling and inference from large-scale networks (such as Internet topologies or the human brain), multipath routing with FEC, and survivability in WDM networks. He is also a founder of AcronymCreator.net - a tool that helps creating new, meaningful acronyms.



**Athina Markopoulou** (SM '98, M'02) is an assistant professor in the EECS Dept. at the University of California, Irvine. She received the Diploma degree in Electrical and Computer Engineering from the National Technical University of Athens, Greece, in 1996, and the M.S. and Ph.D. degrees, both in Electrical Engineering, from Stanford University in 1998 and 2003, respectively. She has been a postdoctoral fellow at Sprint Labs (2003) and at Stanford University (2004–2005), and a member of the technical staff at Arastra Inc. (2005). Her research interests include network coding, network measurements and security, media streaming and online social networks. She received the NSF CAREER award in 2008.



**Minas Gjoka** received his B.S. (2005) degree in Computer Science at the Athens University of Economics and Business, Greece, and his M.S. (2008) and Ph.D. (2010) degrees in Networked Systems at the University of California, Irvine. He is currently a postdoc at the University of California, Irvine. His research interests are in the general areas of networking and distributed systems, with emphasis on online social networks, peer-to-peer systems, and network measurement.