Unbiased Sampling over Online Social Networks

Mojtaba Torkjazi
Department of Computer & Information Science
University of Oregon
moji@cs.uoreogn.edu

ABSTRACT

During recent years, Online Social Networks (OSNs) have evolved in many ways and attracted millions of users. The dramatic increase in the popularity of OSNs has encouraged network researchers to examine their connectivity structure. The majority of empirical studies for characterizing OSN connectivity graphs have analyzed snapshots of the system taken in different times. These snapshots are collected by measurements that crawl OSN connectivity graphs. However, OSN owners are often unwilling to expose their user information due to privacy concerns. On the other hand, because of large population and dynamics of OSNs, the task of crawling may result in an incomplete or distorted snapshot of the system.

These challenges have clearly heightened the urgency for developing efficient and accurate graph sampling techniques. Although a couple of techniques, such as Metropolized Random Walk (MRW) and Respondent-Driven Sampling (RDS), have been proposed for P2P systems, little is known about their accuracy over connectivity graph of OSNs. In this paper, we focus on MRW and RDS sampling techniques and thoroughly investigate their performance in sampling OSN systems. Our main findings can be summarized as follows: *(i)* both techniques are sensitive to graph structure, but RDS exhibits better performance; *(ii)* heterogeneous degree distribution and high number of *unbalanced* edges in OSN graphs are the main factors for poor performance of MRW over such graphs; *(iii)* RDS is unable to properly sample low degree nodes which are hard to reach; and *(iv)* OSN graphs composes of a dense core in the middle and a large number of partitions hanging from the core. High internal and low external connectivity of this core make it almost infeasible for sampling techniques to explore all regions of OSN graphs.

1. INTRODUCTION

Online Social Networks are one of the new trends in the Internet applications. Typically, an OSN is a web-based service that allows individuals to *(i)* build a public or private profile within a bounded system, *(ii)* specify a list of users with whom they have a connection, and *(iii)* view and explore their list of connections and those made by others within the system. A connection between two users implies some kind of relationship or communication such as friendship or interaction (*e.g.*, commenting in a blog, tagging a photo, subscribing in one's videos). However, we use "friendship" as a general term to refer to these connections. This can be presented by a graph, called a *connectivity graph* or *friendship graph*, where nodes represent users and edges represent friendship between users. User attributes such as name, location, age, education, and shared content (*e.g.*, video, photo album, blog) can be stored as meta data in the nodes of the *connectivity graph*.

The popularity of OSNs has significantly increased in recent years. There is a wide variety of incentives for average computer users to join these websites. First, OSNs establish an effective framework for social activities. Through OSNs, not only a user can get in touch with her friends, but also she has a chance to meet new people with common interests. Second, most OSNs embed interesting features, such as videos in YouTube, photos in Flickr, and applications in Facebook. Third, it is very easy to use and navigate through OSNs with only a little knowledge of the Internet.

Increasing popularity of OSNs has motivated network research community for characterizing these systems. Characterizing OSNs will provide a valuable insight about the user participation that would be beneficial for understanding user behaviors, quantifying the traffic associated with OSNs, realizing performance bottlenecks of the current systems, and leveraging their characterizations in designing new protocols and applications [23, 26, 36].

Except in a couple of studies [1, 20], OSN owners are often unwilling to expose their user information to third parties due to privacy concerns. Therefore, the most common approach to empirically study such systems is by analyzing the snapshots of the systems taken in different times. Such snapshots are typically captured by a crawler that queries a set of known nodes to learn about their neighbors and progressively discovers the connectivity structure of the graph. Crawling provides global view of the system to characterize properties of the con-

nectivity graph and its evolution over time. However, capturing accurate snapshot of the system is a challenging task in characterizing large OSNs. Limited rate of crawling[1] along with the OSNs large population make this task almost impractical. As a result, one would either capture a complete snapshot which is distorted due to the dynamics of the system during crawl time, or rely on a partial snapshot which is likely to be biased towards certain group of nodes [1, 5, 25]. In addition, since friendship in most of OSNs is not essentially mutual (*i.e.*, user A is in user B's friend list, but not vice versa), OSNs connectivity graphs are basically directed, and therefore, some nodes in the graph may not be reachable depending on the starting node of the crawl [25].

Graph sampling [15,18,28,30,31] is the promising approach to estimate a particular node property across the whole graph based on that property of collected samples. Towards this end, several techniques have been proposed for P2P systems. Stutzbach *et al.* [31] and Rasti *et al.* [28] introduce Metropolized Random Walk (MRW) and Respondent-Driven Sampling (RDS) techniques, respectively. Both techniques are implemented over a widely-deployed P2P network, namely Gnutella, and show desirable results. However, these two techniques perform inefficiently on connectivity graphs of OSNs which exhibit different connectivity structure from P2P networks [25]. Also, due to very large population of OSNs, in comparison with P2P systems, graph sampling techniques become more important, and hence, having a promising technique for sampling over these graphs is even more important.

In this study, we focus on the connectivity graphs of three popular OSNs, namely Flickr, LiveJournal, and YouTube. We consider *Metropolized Random walk (MRW)* and *Respondent-Driven Sampling (RDS)* and examine their accuracy in estimating node degree (*i.e.*, number of friends) as the main property of nodes. The main contributions of this paper are as follows: *(i)* we examine MRW and RDS over two synthetic graphs and connectivity graphs of OSNs; *(ii)* we identify the main structural properties of OSN graphs that cause poor performance of MRW and RDS over OSNs; and finally *(iii)* we present a high-level structural view for OSN graphs.

Our main findings can be summarized as follows: *(i)* both techniques are sensitive to graph structure, but RDS exhibits better performance; *(ii)* heterogeneous degree distribution and high number of *unbalanced* edges in OSN graphs are the main factors for poor performance of MRW over such graphs; *(iii)* RDS is unable to properly sample low degree nodes which are hard to reach; and *(iv)* OSN graphs composes of a dense core in the middle and a large number of partitions hanging from the core. High internal and low external connectivity of this core make it almost infeasible for sampling technique to explore all regions of OSN graphs.

The rest of this paper is organized as follows: Section 2 reviews some related works in this area. In Section 3, overviews of both MRW and RDS techniques are presented, followed by presenting technical isssues in Section 4. Then in Section 5, we evaluate performance of both techniques on OSNs and discuss their limitations. Section 6 explores structural properties of OSN graphs to provide insights on the root causes of lower performance of sampling techniques. Finally, Section 7 concludes our study and summarizes our future work.

2. RELATED WORK

In recent years, online social networks have received significant attention and many studies have been conducted on this topic. In this section, we first discuss different classes of related work on OSN characterization, and then present an overview of the studies on graph sampling techniques.

2.1 Connectivity Graph and its Evolution

One of the earliest studies on characterizing connectivity graphs of OSNs is done by Ahn *et al.* [1], in which they compared the structures of three online social networking services: Cyworld, MySpace, and Orkut. They used Snowball Sampling to collect data from such systems, which has been shown to overestimate node degree [3]. The authors obtained data directly from Cyworld operators which enables them to conduct a study on evolution of the Cyworld. Another large-scale measurement study and analysis of the structure of multiple online social networks is performed by Mislove *et al.* [25]. They examined data gathered from four popular online social networks, namely Flickr, LiveJournal, Orkut, and YouTube. Using 58 parallel crawlers and Breadth First Search (BFS) technique, they collected the largest weakly connected component (WCC) of these graphs.

The evolution of Flickr and Yahoo! 360 has been studied recently to examine how these systems change over time [20]. Another study on the growth of Flickr is reported in [24], in which the authors focused on the dynamics of the friendship graph in Flickr. Creation rate of the new links and an evolution model are presented in this paper. Leskovec *et al.* [21] presents a more detailed study of network evolution by analyzing four large online social networks, namely Flickr, LinkedIn, Answers, and Delicious, with full temporal information about node and edge arrivals. Their evaluation is done by using maximum-likelihood principle, and compar-

[1]Although some OSNs facilitate capturing the information by providing APIs, limitations on the rate of queries sent to the server (*e.g.*, 10 queries/second in Flickr and 100 queries/hour for Twitter) put burdens on the crawling speed.

ing the graphs generated by their model with the real graphs. They also show that their microscopic model maintains macroscopic properties of the graphs.

In our recent work [33], we examined the evolution of user population in MySpace. While majority of empirical studies on OSNs have focused on the growth of these systems, we measured the pattern of decline in user population and their activity in MySpace.

2.2 User Interactions

Characterizing user interaction in online social networks is a new topic of research in this domain. The first attempt to analyze user interactions was by Chun *et al.* [8]. They analyzed structural characteristics of the activity network in Cyworld and compared it with its connectivity graph to find similar patterns. In another study [17], Krishnamurthy *et al.* present a detailed characterization of the user behavior in Twitter, an application that allows users to send short messages.

User interactions in Flickr are investigated by Valafar *et al.* [34]. They present a measurement study of the Flickr showing that a very small fraction of users in the main component of the friendship graph is responsible for the vast majority of user interactions in Flickr.

Finally, Cha *et al.* [6] explain how user activities disseminate in Flickr. They based their work on the hierarchy and order of addition of a photo as favorite in Flickr and comparison of that with the structure of the friendship graph.

2.3 OSN Applications

Characterizing online social networks can be leveraged in designing new protocols and applications. For example, Mislove *et al.* [23] analyzed the differences between the Web and social networking systems in terms of the mechanisms they use to publish and locate useful information. They examined the potential for using online social networks to enhance Internet search.

In an earlier study [36], Yu *et al.* present an interesting solution to Sybil attacks by incorporating the notion of friendship and trust of social networks into P2P systems. The fundamental observation is that in such a system, attackers will have only a small set of links to the honest portion of the network.

Trusted relationship is also used to thwart unwanted communication [26]. They introduced a system, Ostra, which bounds the total amount of unwanted communication a user can produce based on the number of trust relationships the user has, and relies on the fact that it is difficult for a user to create arbitrarily many trust relationships.

2.4 Large-Scale Graph Measurements

Large-scale graph measurement has been extensively the topic of research in recent years. Kleinberg *et al.* [16] and Broder *et al.* [5] study the Web graph and try to characterize its structure and evolution over time. They also proposed new algorithms for search mechanism and community detection on the Web. In the context of P2P systems, Stutzbach *et al.* [32], present a detailed characterization of P2P overlay topologies and their dynamics focusing on Gnutella network.

Graph sampling techniques have introduced as a scalable and promising approach for large-scale graph characterizations. These techniques have been used to extract information about graphs (e.g., selecting representative sub-graphs from a large, intractable graph) while maintaining properties of the original structure [15, 18, 19, 30]. A couple of sampling methods have been proposed for using in P2P networks [9, 10, 28, 31]. Among them, we are interested in Metropolized Random Walk (MRW) introduced by Stutzbach *et al.* [31], and Respondent-Driven Sampling (RDS) by Rasti *et al.* [28]. We will fully discuss these two techniques in Section 3. For a detailed review of related works on graph sampling techniques, see [31].

3. SAMPLING UNDIRECTED GRAPHS

Online social networks can be represented by a *connectivity graph* (G), in which nodes (V) represent users and edges (E) show relationship between users. Furthermore, user properties such as name, age, location, and number of friends can be stored as attributes in the nodes.

The objective of sampling techniques is to select a minimal fraction of nodes as samples in order to estimate the distribution of a node property with certain level of accuracy. When the global view of the graph is not available, the only promising approach is selecting nodes by using random walkers. A random walker starts from a starting node and progressively selects a random neighbor of the current node as its next step.

In this section, we first introduce basic definitions of Markov chain and how Markov chain is used to model random walks, and then discuss MRW and RDS sampling techniques in Section 3.2 and 3.3, respectively.

3.1 Markov Chain and Random Walk

A stochastic process with a set of states is called a Markov chain if future states of the process are independent of the past states and only depend on the current state. Markov chain is a widely used model for random walks on graphs which will be discussed later in this section. Each Markov chain with N states is described by a *transition probability matrix* $P_{N\times N}$, in which P_{ij} is the probability of moving from state i to state j. This matrix satisfies

- $0 \leq P_{ij} \leq 1;\ i,j = 0, ..., n-1.$
- $\sum_{j=0}^{n-1} P_{ij} = 1;\ i = 0, ..., n-1.$

Let $\pi(t)$ be the probability distribution of the states at time t, then the probability distribution at time $t+1$ is $\pi(t+1) = \pi(t)P$ and consequently we have $\pi(t) = \pi(0)P^t$. A distribution π is called *stationary distribution* if it satisfies $\pi = \pi P$. If a chain has a stationary distribution it means that it will converge to a stationary distribution after sufficient amount of steps and regardless of its starting state.

Markov chains are used to model random walks on graphs. Consider a graph $G = (N, E)$ with $|N|$ nodes and $|E|$ edges. A random walk starts from an initial node (state) and at each step it makes a transition between two adjacent nodes u and v using edge (u, v). Each edge in the graph is assigned with a probability which is considered by random walk on each transition. Obviously, the sum of probabilities of edges connecting to a node must be one. It is easily calculated that stationary distribution for a random walk on a connected undirected graph is $\pi_i = \frac{deg(i)}{2|E|}$, where $deg(i)$ is the degree of node i.

3.2 Metropolized Random Walk

A regular random walk works as follows: it starts from a node A, and chooses a node B among its neighbors with uniform probability. So the transition probability matrix $P_{N\times N}$ is:

$$P_{ij} = \begin{cases} \frac{1}{\text{degree}(i)} & j \text{ is a neighbor of } i, \\ 0 & \text{otherwise} \end{cases}$$

Using regular random walk for sampling over a graph yields to biased samples. As shown in Section 3.1, the probability of sampling a node is proportional to its degree, which means that nodes with higher degrees are more probable to be sampled by the random walk. In order to have unbiased samples, the probability of selecting a node should be uniform across the graph; *i.e.*, $\pi(i) = \frac{1}{|V|}$, where $|V|$ is the number of nodes.

The Metropolis-Hasting technique [7, 13, 22, 28] provides a way to modify the next-hop selection to produce any desired stationary distribution, $\pi(j)$. In [31], Stutzbach *et al.* choose the following equation for the next-hop selection to achieve a uniform distribution:

$$Q(i,j) = \begin{cases} P_{ij} \min\left(\frac{\text{degree}(i)}{\text{degree}(j)}, 1\right) & \text{if } i \neq j, \\ 1 - \sum_{k\neq i} Q(i,k) & \text{if } i = j \end{cases}$$

Therefore, at each step of random walk, MRW algorithm performs as follows:

- Select a neighbor j of node i uniformly at random.
- Generate a random number, p, between 0 and 1.
- If $p \leq \frac{deg(i)}{deg(j)}$, j is the next step.
- Otherwise, stay at i as the next step.

Basically, the bias toward high degree nodes is removed by reducing the probability of transitioning to such nodes at each step of Metropolized Random Walk.

3.3 Respondent Driven Sampling

Respondent-Driven Sampling (RDS) is a development of Snowball Sampling (SBS), a technique used in social sciences to make estimation of important population parameters in "hidden" populations[2] (*e.g.*, drug users). In SBS, the number of samples grows like a rolling snowball as it iteratively recruits more samples from referrals of individuals in the current sample list.

It has been shown that Snowball Sampling techniques are subject to biases toward respondents with higher number of referrals [3, 11]. In [14, 29, 35], RDS is introduced as a variant of SBS which incorporates a mathematical model into Snowball Sampling that re-weights the samples to compensate for the fact that samples were collected non-uniformly.

Random walk on a graph could be interpreted as a special case of RDS in which each respondent recruits exactly one individual. This in turn can be recast as a Monte Carlo Markov Chain (MCMC) problem [29]. Similar to social sciences, we are interested in estimating node attributes based on node degrees seen during the random walk.

We wish to estimate the distribution of a node property X; specifically, consider any partition $\{R_1, \ldots, R_m\}$ of the range of possible values of X. We partition the node set V accordingly into groups of nodes $\{V_1, \ldots, V_m\}$, *i.e.*, $V_i = \{v \in V : X(v) \in R_i\}$. A simple example is when X is positive integer value and we group by value: $V_i = \{v \in V : X(v) = i\}$.

The RDS approach is to estimate the proportion p_i of nodes that are in group i from observed node degree and group memberships of nodes traversed in the random walk. Specifically, let the walk comprise of n steps, visiting a set of nodes $T = \{t_1, t_2, \ldots, t_n\}$ (note that nodes may be visited more than once). Let $T_i = T \cap V_i$ denote the visited nodes that lie in group i. It is well known that the stationary distribution of a random walk on a connected graph with node set V is $\pi(v) = \text{degree}(v)/\sum_{u\in V} \text{degree}(u)$ where $\text{degree}(v)$ is the degree of the node v. Hence, for any node property X, the Hansen-Hurwitz [12] estimator $\hat{S}(X) := n^{-1} \sum_{v\in T} \frac{X(v)}{\pi(v)}$ is an unbiased and consistent estimator of the sum $S(X) := \sum_{v\in V} X(v)$ when T is drawn from a stationary random walk, i.e., one that evolves from an initial node that is randomly selected according to the stationary distribution. Consider the special case; when $X = I_{V_i}$ is the indicator of a node being in group

[2]Heckathorn in [14] defines a population as "hidden" if it has the following two characteristics: first, population size is unknown; and second, individuals of this population refuse to cooperate in order to protect their privacy.

i; *i.e.*, $I_{V_i}(v) = 1$ if $v \in V_i$ and 0 otherwise, then $\hat{S}(I_{V_i})$ estimates the total number of nodes in V_i. When $X = 1$ then $\hat{S}(1)$ estimates the total number of nodes $|V|$ in the graph. Thus we can estimate the proportion p_i by

$$\hat{p}_i = \frac{\hat{S}(I_{V_i})}{\hat{S}(1)} = \frac{\sum_{v \in T_i} \frac{1}{\text{degree}(v)}}{\sum_{u \in T} \frac{1}{\text{degree}(u)}}$$

$\hat{p}_i$ is consistent—it converges to the true value p_i—as the number n of visited nodes grows. The RDS estimator can be recognized as an importance sampling estimator weighted by the stationary distribution π, applied to the MCMC of the random walk on the vertex set V.

Although both MRW and RDS use random walks to collect samples, there is a slight difference in the way they produce unbiased results. The random walk in MRW is modified so that the probability of selecting a node is uniform. On the other hand, RDS first selects samples using regular random walk, and then post-processes the samples to have unbiased results.

4. PRACTICAL ISSUES

Directed Graphs: The focus of our stusy is on undirected graphs. In directed graphs, the probability of a random walker to land in a node is not only related to its in-degree, but also is a function of its neighbors' in-degree. Therefore, calculating the "stationary distribution" is complicated and needs global view of the graph, see [4, 27] for more details.

Main Node Property: Moreover, we only consider those node properties that may interact with the random walk; *i.e.*, the degree of an individual node in the graph determines the probability that a node is visited by a random walker.

Methodology: We simulate MRW and RDS over snapshots of three OSNs, namely Flickr, LiveJournal, and YouTube, taken by a recent study [25]. Simulations offer opportunities for evaluating the accuracy of these techniques. Moreover, by comparing our results with earlier studies [28, 31], we identify structural properties of the graph that closely correlate with the accuracy of sampling techniques.

Performance Metric: To validate MRW and RDS sampling techniques, we use the Kolmogorov-Smirnov test, KS, to quantify the distance between the estimated cumulative distribution function (CDF) of a desired property from collected samples and the CDF from all nodes. Suppose $\hat{F}(x)$ is the estimated cumulative distribution function and $F(x)$ is the true cumulative distribution function. The KS statistic is formally defined as follows:

$$KS = \sup_x |\hat{F}(x) - F(x)|$$

where $sup(S)$ is the supremum[3] of set S. A value of $KS = \epsilon$ means that the error of the estimation is at least ϵ.

5. EVALUATION

In this section, we examine how connectivity structure of a graph affects the performance of RDS and MRW sampling techniques. For this purpose, we use two synthetic graphs and three OSN snapshots which are introduced below:

Random graphs (RA): A random graph obtained by starting with a set of n nodes and adding m edges between them at random.

Barabási-Albert graphs (BA): The "scale-free" graphs of the preferential attachment proposed by Barabási *et al.* [2] to generate graphs with power-law degree distributions.

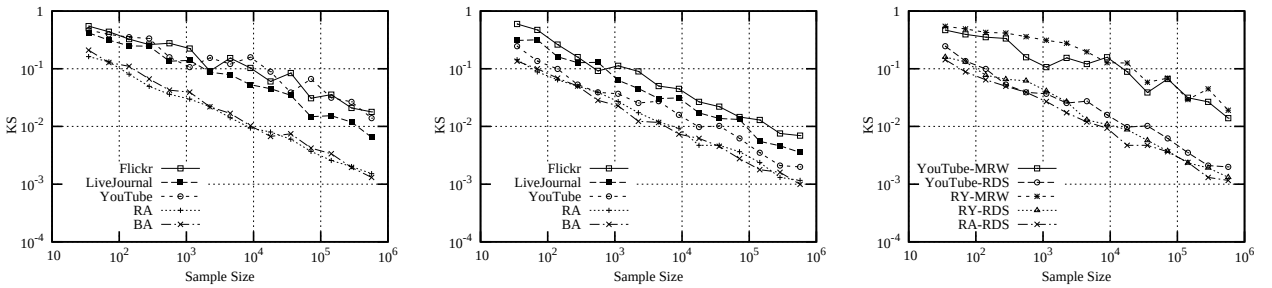
OSN snapshots: We use snapshots of three popular OSNs, namely Flickr, LiveJournal, and YouTube, taken by Mislove *et al.* [25]. Although their data sets are not perfect, we believe that they are the best available snapshots of real OSNs. These data sets[4] contain all of the user-to-user links captured by their crawlers at the time. These links are directed since friendship in Flickr, LiveJournal, and YouTube is not reciprocal. For the purpose of our study, we assume that all of the links are bidirectional; *i.e.*, for each link (x, y) in the data set we added (y, x) link to the connectivity graph. We make this modification because our focus in this study is on undirected graphs. This assumption has no significant effect on the structure of these graphs as [25] reveals that more than 60% of the links are symmetric in these OSNs. Moreover, to have a connected component, we removed all isolated islands of friendship graph.

5.1 Performance of Sampling Techniques

To minimize the effect of randomness in different experiments, we run 10 parallel samplers with a relatively long walk length of 500K hops, and calculate the corresponding KS errors. Figures 1(a) and 1(b) show the average KS error for degree distribution as a function of sample size for MRW and RDS techniques, respectively. Following two main points can be concluded from Figure 1. *(i)* The performance of RDS improves with higher number of samples; *i.e.*, larger walk lengths lead to lower KS errors. Although the rate of this improvement is almost similar across all graph structures; for a fixed sample size, RDS performs worse in OSN graphs. *(ii)* Comparing Figure 1(a) and Figure 1(b) reveals that MRW exhibits lower performance than RDS. Especially, not only performance of MRW is significantly worse over

[3]Supremum of S is the least element of S that is greater than or equal to each element of S.

[4]`http://socialnetworks.mpi-sws.org/data-imc2007.html`



(a) MRW (b) RDS (c) RDS & MRW in Randomized YouTube

Figure 1: Performance of RDS and MRW in estimating node degree distribution over different graph types.

Graph Type	#Nodes	#Edges	Density	Avg. CC	Diameter
RA	10,000	50,000	5	0.0009	7
BA	10,000	50,000	5	0.0065	5
Flickr (A)	1,624,992	15,476,835	9.5	0.1892	27
Flickr (C)	280,562	12,210,537	43.5	0.2625	13
Flickr (G)	404,343	514,992	1.3	0.1510	> 40
LiveJournal (A)	5,189,809	48,688,097	9.4	0.2749	20
LiveJournal (C)	2,807,323	41,230,255	14.7	0.2938	13
LiveJournal (G)	62,269	84,334	1.4	0.1981	> 290
YouTube (A)	1,134,890	2,987,624	2.6	0.0808	21
YouTube (C)	306,482	1,707,249	5.6	0.1264	12
YouTube (G)	70,066	76,346	1.1	0.0641	> 80

Table 1: High Level Statistics. A: All Nodes, C: Core, G: Giant Partition, CC: Clustering Coefficient.

OSN graphs, but also the rate of improvement in such graphs is much slower.

5.2 Identifying Limiting Factors

The question is **"Which structural properties of OSN graphs degrade performance of MRW and RDS?"** To answer this question, we first examine the macro level properties of such graphs and then focus on their micro level properties.

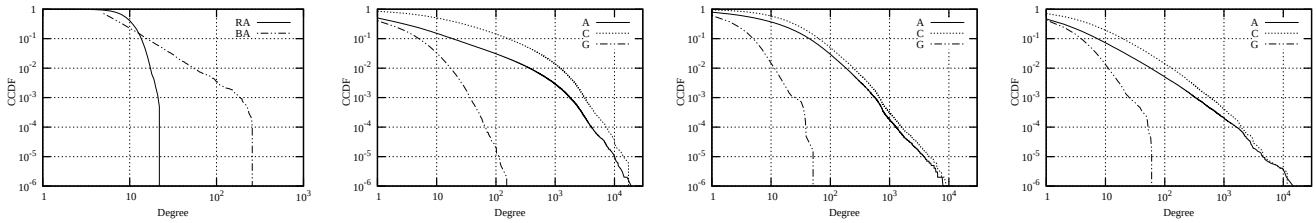
Table 1 presents high-level statistics of our graphs. OSN graph sizes vary by almost a factor of five, while the number of edges varies by one order of magnitude. Other metrics such as density and clustering coefficient are also different across OSN graphs. However, despite these differences, Mislove *et al.* show that these graphs share similar structural properties [25]. Structural properties of connectivity graph of social networks are totally different from those of random graphs. For example, OSN graphs show high level of local clustering [25]. As we can see in Table 1, the clustering coefficients of social networks are several orders of magnitude larger than RA and BA graphs.

Figure 2 depicts node degree distributions for OSN graphs. From one perspective, performance of MRW and RDS can be attributed to very heterogeneous degree distribution, similar to what we have in online social networks. In order to examine this hypothesis, we shuffled (randomly rewired) all edges in YouTube graph to generate a random graph, labeled as RY, without changing the degree of individual nodes. Figure 1(c) plots performance of MRW and RDS over RY. While edge shuffling does not affect performance of MRW, it improves performance of RDS[5]. This result suggests that heterogeneous degree distribution is not the sole underlying cause of poor performance of RDS over OSN graphs, but probably one of the prominent factors to degrade performance of MRW.

5.2.1 Limiting Factors of MRW

To further explore poor performance of MRW on graphs with heterogeneous degree distributions, we calculate the average percentage of unique visited nodes for sample size of 100K nodes, Table 3. Although RDS visits the same number of unique nodes, MRW visits much lower percentage in OSN graphs. The observed overall lower performance of MRW can be attributed to the following phenomenon. In OSN graphs, we have several clusters of low degree nodes that are connected through high degree nodes [25]. The only way for an MRW walker to leave a cluster is via a much higher degree node that resides outside this cluster. *i.e.*, the walker has to traverse an edge from a low degree node within

[5]We obtained the same results over LiveJournal and Flickr after edge shuffling.



(a) RA and BA (b) Flickr (c) LiveJournal (d) YouTube

Figure 2: Node degree distribution. A: All Nodes, C: Core, G: Giant partition.

Graph Type	% of *unbalanced* edges			% of *self-loops* and its breakdown on *unbalanced* edges			
	$ratio > 100$	$ratio > 500$	$ratio > 1000$	total	$ratio > 100$	$ratio > 500$	$ratio > 1000$
RA	-	-	-	17.6%	-	-	-
BA	-	-	-	40.6%	-	-	-
Flickr	4.9%	1.2%	0.6%	75.1%	21.1%	8.2%	5.4%
LiveJournal	1.2%	0.2%	0.1%	60.8%	7.7%	1.2%	0.3%
YouTube	20.0%	8.5%	5.2%	75.7%	35.4%	20.5%	14.0%

Table 2: Percentage of unbalanced edges and self-loops on them. Balance ratio of an edge (x, y) is $degree(x)/degree(y)$.

such a cluster to a much higher degree node outside this cluster (We will call these edges *unbalanced* edges in the rest of this paper.). As described in Section 3.2, for the MRW technique, the probability of moving along such an edge is proportional to the ratio of the (low) degree of the node within the cluster and the (very high) degree of the node outside this cluster. Therefore, when a walker ends up in one of these clusters, it will keep looping among its low degree nodes, thus collecting redundant samples which in turn degrades the accuracy of sampling.

Graph Type	MRW	RDS
RA	69.67%	86.30%
BA	69.32%	84.55%
Flickr	10.24%	82.04%
LiveJournal	29.03%	86.37%
YouTube	7.82%	69.53%

Table 3: Avg. percentage of unique samples for sample size of 100K

Table 2 shows the percentage of *unbalanced* edges in different graph types along with the average percentage of self-loops[6] that MRW walkers take because of selecting such edges. Percentage of *unbalanced* edges and self-loops on these edges in LiveJournal is less than the other two OSNs. This is in agreement with our results in Figure 1(a) where MRW shows better performance over LiveJournal. Our results so far suggest that the high number of *unbalanced* edges in OSN graphs, which is a side-effect of their heterogeneous degree distribution, yields to more error in estimation of node properties in MRW technique.

[6]The state in which MRW walker stays in the same node since the probability of transitioning to a neighbor is not large enough.

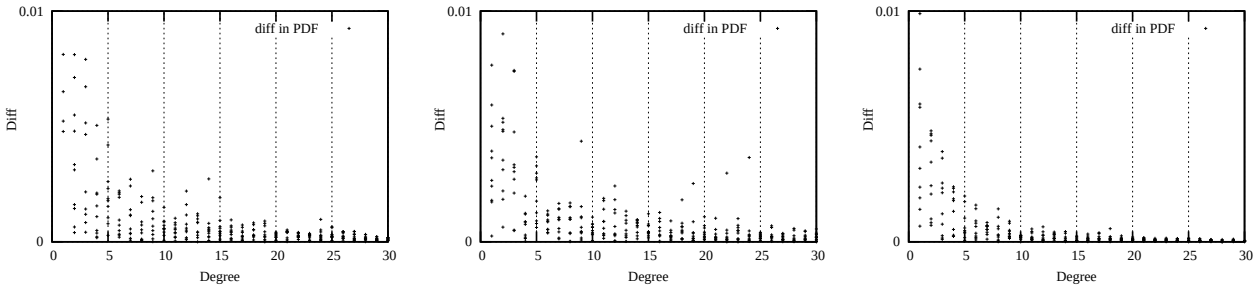
5.2.2 Limiting Factors of RDS

The performance of MRW is affected by *unbalanced* edges in OSN graphs. However, as we can see in Figure 1(b), even a regular random walk, used in RDS, shows low performance. Therefore, we need to investigate those structural properties which affect the behavior of regular random walkers in such graphs.

Figure 3, in a finer grain level than Figure 1, presents those nodes that are not properly sampled by RDS. Each dot shows the difference between Probability Distribution Function of node degree, d, in reference graph and samples; i.e, $|PDF_{OSN}(d) - PDF_{RDS}(d)|$. We can see in Figure 3 that RDS technique, most of the time, is inaccurate in sampling of low degree nodes (*i.e.*, degrees less than five) in OSN graphs. We are interested to find why RDS is unable to sample them properly.

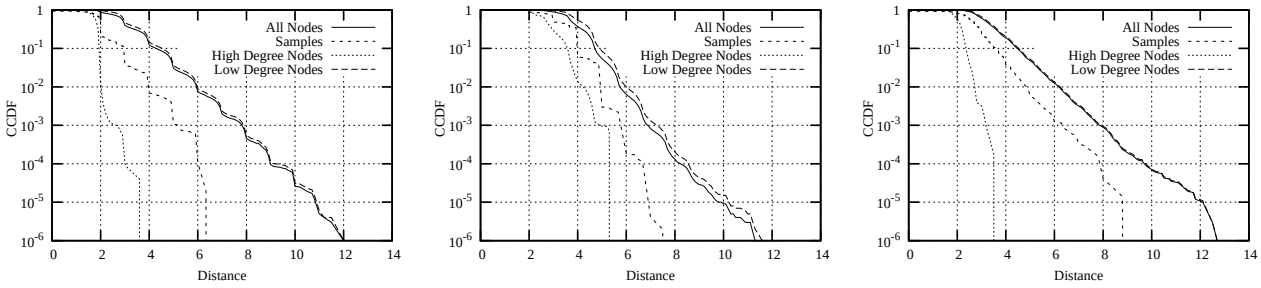
For this purpose, we flood (*i.e.*, BFS crawling) from the 10 highest degree nodes, as we believe these nodes are more "central" [25] in the graph, *i.e.*, average shortest path from these nodes to other nodes is shorter. Then, we measure the distance of each node in the graph as the mean shortest path to the starting nodes of our flooding. We also run 10 random walkers with a very long walk length of 10M hops, started from the same starting nodes.

Figure 4 shows the distance distribution of nodes visited by random walkers and also for all nodes in the graph. As we can see in this figure, even a very long random walker cannot discover all parts of the graph. Although we have some nodes with distance of more than 10 in OSN graphs, walkers do not go farther than



(a) Flickr (b) LiveJournal (c) YouTube

Figure 3: RDS error in estimating node degree distribution over OSN graphs; walk length = 100K, repeat = 10.



(a) Flickr (b) LiveJournal (c) YouTube

Figure 4: Distance distribution of nodes.

nine hops. Even for a distance of four, the gap between CCDF values of all nodes and samples is 0.11, 0.31, and 0.15 for Flickr, LiveJournal, and YouTube, respectively. This means that random walkers do not properly sample nodes with distance of greater than four.

In order to find what type of nodes are located at far distances, we plot distance distribution for both low degree nodes and high degree nodes in Figure 4. We roughly define a node as low degree if it's degree is less than 10, and high degree if it's degree is more than 50[7]. Interestingly, all high degree nodes have distances less than four (except for 1% in LiveJournal), while low degree nodes are all over the graph. It is obvious in Figure 4 that in all three OSNs, the line for distance distribution of samples is above the line for high degree nodes and below the line for low degree nodes. It means that random walkers go far enough to sample all nodes with high degree, but not far enough to sample nodes with low degree. This confirms our earlier observations in Figure 3 that low degree nodes are the main contributors of the KS error.

[7]We make this assumption based on Figure 2. While portion of low degree nodes is 15%, 37%, and 7%, the portion of high degree nodes is 5%, 9%, and 1% in Flickr, LiveJournal, and YouTube, respectively.

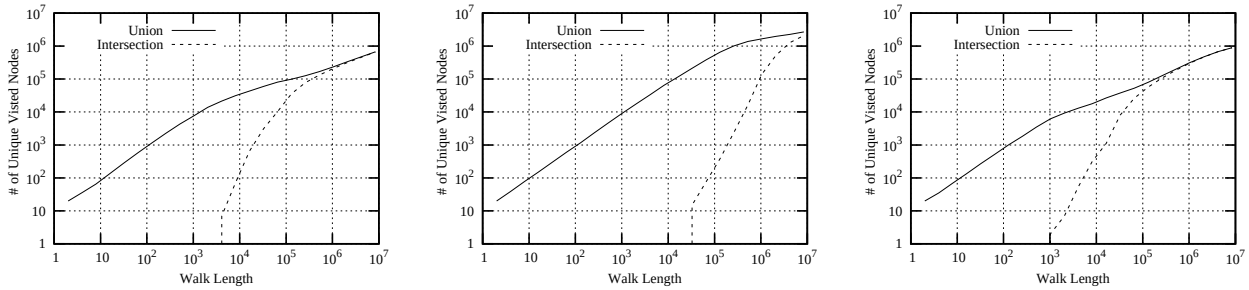
6. OSN GRAPH STRUCTURE

Our results are in agreement with [25] in which Mislove *et al.* show that OSN graphs have a dense "core" of very well-connected high degree nodes in the middle. This core plays a "central" role in connectivity of the graph and removing nodes from the core will yield to graph partitioning. Several other partitions are hanging from the core and forming the *shell* of the graph. Each of these partitions are connected to the core through a relatively small number of *gateway* edges, see Figure 6. Also, external connectivity of the core is less than its internal connectivity, while this is reverse for the shell (see Section 6.1 for more details on external and internal connectivity). Because of this fact, random walkers are absorbed to the core and tend to remain there. This, in turn, results in improperly sampling of nodes in the shell.

In this section, we first focus on the core in Section 6.1, and then we move on characterizing the shell in Section 6.2.

6.1 Characterzing Core Component

We loosely define a core of a network as any (minimal) set of nodes that satisfies three properties: *First*, its internal connectivity must be much larger than its external connectivity [32]. *Second*, the core must be



(a) Flickr (b) LiveJournal (c) YouTube

Figure 5: Number of unique nodes visited by 10 random walkers with length of 10M started from 10 random nodes.

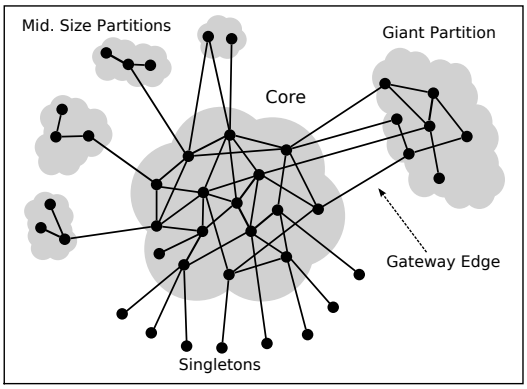


Figure 6: High level view of OSN graph structure: A core in the middle and several partitions hanging from the core via gateway edges.

connected with a relatively small diameter. *Third*, the core is essential for the connectivity of the graph and its removal will break the graph into many node islands.

This definition along with our results in previous sections imply that a random walker is more likely to visit the nodes in the core than other parts of the network. We use this indication to identify the core of these graphs by running 10 long random walkers from 10 random starting points. Each walker explores a region of the graph and we expect that nodes located in the core will be discovered by all walkers. Therefore, the intersection set of the visited nodes by all walkers grows with longer walk lengths as walkers visit more nodes. We consider this set as the core, once its size equals the size of union set of visited nodes (*i.e.*, those nodes that have been visited at least by a walker). This is a reasonable approach since when random walkers do not explore new regions, it means that they keep hitting nodes in the core.

Figure 5 illustrates our methodology for core detection. We can see six *knee points* in Figure 5 where there is an obvious slow down in growth of both union and intersection sets. This is due to the fact that after some number of steps, union and intersection sets become saturated and since then proceed with a slow growth. Interestingly, in all three OSNs, when intersection sets become saturated, their size is very close to the size of union sets. When union and intersection sets are almost equal, it implies that our walkers are trapped in the core, and subsequently, we can consider the intersection set as the core of the graph. Although this set of nodes is loosely specified as the core, we claim that it satisfies our criteria for the core definition. In this section, we focus on the first two properties of the core (based on our definition at the beginning of this section) and leave the third property for Section 6.2.

Internal Connectivity: To quantify the connectivity between the core and the shell, we examined whether nodes inside the core have a higher tendency to connect to each other rather than nodes outside the core [32]. For this purpose, we calculate the ratio (R) of internal edges to the total number of edges for four sets of nodes and present it in Table 4. First, we divide the graph to two sets, core and shell. As we see, OSN graphs show different R values for core and shell which means they have different internal (and external) connectivity. In other words, while nodes in the core have more inclination to connect to each other (more than 79% internal edges in contrast with less than 21% external edges), nodes in the shell behave reversely. These unbalanced values of R verify our approach to identify the core.

High degree nodes in the core of OSN graphs are responsible for large values of R. However, there is also another factor which contributes to large internal connectivity of the core of such graphs. Most of high degree nodes residing in the core are very well-connected to each other. To focus on the latter issue, we randomly rewire the edges in OSN graphs to destroy the well-formed connection between high degree nodes of the core. We then calculate the R values for the yielded graphs. C_{rnd} column in Table 4 denotes R values for randomized OSN graphs and we can see an obvious decrease in internal connectivity of their core. In conclu-

Graph Type	Internal Connectivity [32]					
	C	C_{rnd}	S	G	C_1	C_2
Flickr	0.935	0.853	0.521	0.642	0.497	0.503
LiveJournal	0.930	0.883	0.285	0.439	0.500	0.499
YouTube	0.786	0.628	0.427	0.716	0.497	0.504

Table 4: Internal Connectivity. C: Core, C_{rnd}: randomized graph, G: Giant Partition, S: Shell, C_1: Core Subset 1, C_2: Core Subset 2.

Graph Type	#Total	#Giant & Size	#Singletons	#Mid. Size	#Trees
Flickr	509,770	1 & 404,343	428,678	81,091	66,812
LiveJuornal	1,510,696	1 & 62,269	1,268,216	242,479	201,476
YouTube	500,810	1 & 70,066	423,328	77,481	71,294

Table 5: High Level Statistics of Partitions

sion, both high degree nodes and their internal connection are responsible for large R values for the core of OSN graphs.

Moreover, as another confirmation, we divided the core to two random sets of nodes with the same number of nodes and calculated R values for them. As we can see in Table 4, the core is homogeneous in terms of internal connectivity since for all three OSNs, *(i)* neither of the sets show any tendency towards connecting to either internal or external nodes ($R = 0.5$), and *(ii)* both sets present similar connectivity behavior (same values of R).

Small Diameter: For the second property of the core, we calculated its diameter and compared it with that of the entire graph. The results are presented in Table 1. Relatively smaller diameter of the core is mainly due to the fact that density and portion of high degree nodes inside the core are larger than those of the entire graph, see Figure 2 for degree distribution of the core.

Our study for core detection of OSN graphs indicates a denser core for Flickr and a looser core for YouTube. The core of Flickr shows more internal connectivity, and much more density, see Tables 1 and 4. For Flickr, only 17% of the nodes reside in its core, while this value for LiveJournal and YouTube is 54% and 22%, respectively. On the other side, density and internal connectivity of YouTube core are lower than the other two.

6.2 Characterizing Shell Component

Removing the core of OSN graphs breaks it into so many disconnected partitions. Table 5 presents a high level statistics of these partitions. Interestingly, we found three types of partitions: singletons, middle-sized partitions, and giant partition.

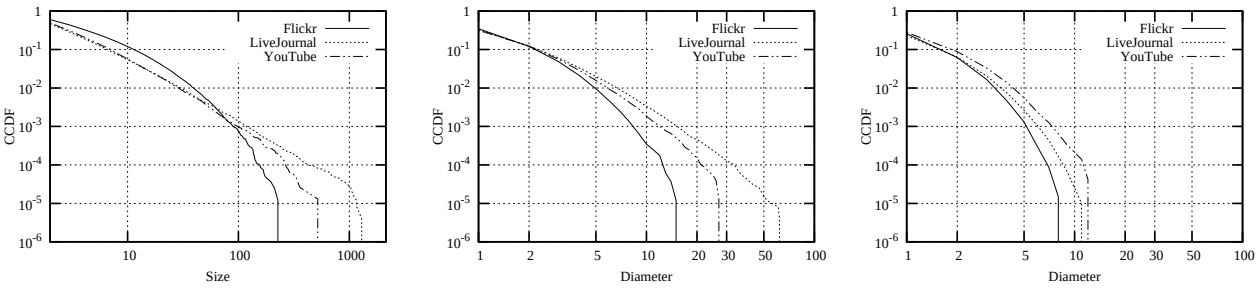
Singletons and Middle-Sized Partitions: More than 80% of partitions are singletons; *i.e.*, single nodes hanging from the core. Middle-sized partitions are islands of nodes with a population less than 2000 nodes. Figure 7(a) plots power-law population distribution of these partitions. The majority of middle-sized partitions have small number of nodes (90% less than 10 nodes), and a few of them have higher populations. Figure 7(b) shows the diameter of middle-sized partitions. We can see some partitions with much larger diameter than the entire graph, especially in LiveJournal. As we mentioned in Section 6.1, the core plays a central role in the OSN graph connectivity. Since many shortest paths between pairs of nodes go through the core, removing the core will increase the average shortest path length between nodes of a partition, and subsequently, increase its diameter. Since diameter distribution of middle-sized partitions for LiveJournal is above Flickr, and that of Flickr is above YouTube in Figure 7(b), we can derive that connectivity graph of LiveJournal is more stretched than that of Flickr, and connectivity graph of Flickr is more stretched than that of YouTube.

Trees: More than 80% of middle-sized partitions are forming a tree. Figure 7(c) plots diameter distribution for trees. Although more than 90% of these trees have a diameter of less than two, there are some trees with diameters greater than five. Larger diameters for tree-like partitions makes it less probable for random walkers to reach leaves of the trees in order to sample them.

Giant Partition: Finally, the shells of these OSNs have a giant partition with a several orders of magnitude larger population than other partitions. High level statistics for the giant partition is presented in Tables 1 and 4. Structural properties of this partition is significantly different from the core. In contrast with the core, it has lower density, clustering coefficient, and internal connectivity and dramatically larger diameter. Also, it shows different degree distribution as we can see in Figure 2, which differs from power-law degree distribution of the core and the entire graph.

7. CONCLUSION AND FUTURE WORK

In this paper, we articulated an extensive analysis of RDS and MRW sampling techniques over connectivity



(a) Size of Middle-Size Partitions (b) Diameter of Middle-Size Partitions (c) Diameter of Tree-Like Partitions

Figure 7: Size and diameter distribution of partitions. Giant partition and singletons are removed.

graphs of online social networks. This study provides essential insights for the behavior of MRW and RDS over OSNs and investigates the underlying causes of their poor performance on such graphs.

We found that in OSN graphs, as a result of their heterogeneous degree distribution, most of low degree nodes connect to nodes with much higher degrees; and thus, we have a large number of *unbalanced* edges in such graphs. This phenomenon degrades performance of MRW in estimating node properties.

Moreover, the existence of a very dense core in the middle of OSN graphs with high internal connectivity and low external connectivity, along with a sparse and deep shell around the core, do not allow random walkers to properly explore all regions in such graphs. This, in turn, will result in over-sampling of the core and under-sampling of the nodes in the shell, and subsequently, larger errors in estimation of node properties of the entire graph.

We are continuing to focus exclusively on improving MRW and RDS sampling techniques to perform efficiently and in a *topology-agnostic* way in all graph structures. Once we achieve the desirable level of accuracy, we can furnish OSN researchers with our tool to be empirically used for unbiased data collection from popular OSNs. We believe this is a venue of high demand in OSN research community since in each characterization study based on measurement, representative data is the king.

While unbiased sampling over undirected graphs has been largely explored during these years, no significant study has been done on directed graphs. In directed graphs, the probability of a random walker to land in a node is not only related to its in-degree, but also is a function of its neighbors' in-degree. Therefore, calculating the "stationary distribution" is complicated and needs global view of the graph, see [4, 27] for more details. The main question here is whether it is possible to address challenges of directed graphs in an extension of MRW and RDS.

Another direction for characterizing OSNs is looking into their dynamic aspects. We have recently embarked on analyzing and modeling macro-level dynamic behavior of a couple of popular systems, such as MySpace [33], in terms of their change and evolution over time. For this purpose, access to promising sampling techniques which can deal with unknown nature of dynamics and churn in OSNs is required.

8. REFERENCES

[1] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of Topological Characteristics of Huge Online Social Networking Services. In *International World Wide Web Conference (WWW)*, 2007.

[2] A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286, 1999.

[3] L. Becchetti, C. Castillo, D. Donato, and A. Fazzone. A Comparison of Sampling Techniques for Web Graph Characterization. In *Workshop on Link Analysis (LinkKDD)*, 2006.

[4] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Computer Networks and ISDN Systems*, 1998.

[5] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph Structures in the Web: Experiments and Models. In *International World Wide Web Conference (WWW)*, 2000.

[6] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi. Characterizing Social Cascades in Flickr. In *1st ACM SIGCOMM Workshop on Social Networks (WOSN)*, 2008.

[7] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4), 1995.

[8] H. Chun, H. Kwak, Y. ho Eom, Y.-Y. Ahn, S. Moon, and H. Jeong. Comparison of Online Social Relations in terms of Volume vs. Interaction: A case Study of Cyworld. In *Internet Measurement Conference (IMC)*, 2008.

[9] C. Cooper, M. Dyer, and C. Greenhill. Sampling Regular Graphs and a P2P Network. In *Symposium on Discrete Algorithms*, 2005.

[10] C. Gkantsidis, M. Mihail, and A. Saberi. Random Walks in Peer-to-Peer Networks. In *INFOCOM*, 2004.

[11] L. A. Goodman. Snowball Sampling. *Annals of Mathematics Statistics*, 32, 1961.

[12] M. H. Hansen and W. N. Hurwitz. On the Theory of Sampling from Finite Populations. *Annals of Mathematical Statistics*, 14(4), 1943.

[13] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57, 1970.

[14] D. D. Heckathorn. Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems*, 44(2), 1997.

[15] M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On Near-Uniform URL Sampling. In *International World Wide Web Conference*, 2001.

[16] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tompkins. The Web as a Graph: Measurements, Models, and Methods. In *International World Wide Web Conference (WWW)*, 1999.

[17] B. Krishnamurthy, P. Gill, and M. Arlitt. A few Chirps about Twitter. In *1st ACM SIGCOMM Workshop on Social Networks (WOSN)*, 2008.

[18] V. Krishnamurthy, M. Faloutsos, M. Chrobak, L. Lao, J.-H. Cui, and A. G. Percus. Reducing Large Internet Topologies for Faster Simulations. In *IFIP Networking*, 2005.

[19] V. Krishnamurthy, J. Sun, M. Faloutsos, and S. Tauro. Sampling Internet Topologies: How Small Can We Go? In *International Conference on Internet Computing*, 2003.

[20] R. Kumar, J. Novak, and A. Tomkins. Structure and the Evolution of Online Social Networks. In *Knowledge Discovery and Data Mining Conference (KDD)*. Yahoo! Research, 2006.

[21] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic Evolution of Social Networks. In *Knowledge Discovery and Data Mining Conference (KDD)*, 2008.

[22] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21, 1953.

[23] A. Mislove, K. P. Gummadi, and P. Druschel. Exploiting Social Networks for Internet Search. In *5th Workshop on Hot Topics in Network (HotNets-V)*, 2006.

[24] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the Flickr Social Network. In *1st ACM SIGCOMM Workshop on Social Networks (WOSN)*, 2008.

[25] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Internet Measurement Conference (IMC)*, 2007.

[26] A. Mislove, A. Post, P. Druschel, and K. P. Gummadi. Ostra: Leveraging Trust to Thwart Unwanted Communication. In *Networked Systems Design and Implementation (NSDI)*, 2008.

[27] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999. Previous number = SIDL-WP-1999-0120.

[28] A. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach. Respondent-driven Sampling for Characterizing Unstructured Overlays. In *IEEE INFOCOM Mini-conference*, 2009.

[29] M. J. Salganik and D. D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34, 2004.

[30] M. P. H. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences*, 102(12), 2005.

[31] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On Unbiased Sampling for Unstructured Peer-to-Peer Networks. In *Internet Measurement Conference (IMC)*, 2006.

[32] D. Stutzbach, R. Rejaie, and S. Sen. Characterizing Unstructured Overlay Topologies in Modern P2P File-Sharing Systems. In *Internet Measurement Conference (IMC)*, 2005.

[33] M. Torkjazi, R. Rejaie, and W. Willinger. Hot Today, Gone Tomorrow: On the Migration of MySpace Users. In *2nd ACM SIGCOMM Workshop on Social Networks (WOSN)*, 2009.

[34] M. Valafar, R. Rejaie, and W. Willinger. Beyond Friendship Graphs: A Study of User Interactions in Flickr. In *2nd ACM SIGCOMM Workshop on Social Networks (WOSN)*, 2009.

[35] E. Volz and D. D. Heckathorn. Probability Based Estimation Theory for Respondent Driven Sampling. *Journal of Official Statistics*, 24(1), 2008.

[36] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. SybilGuard: Defending against Sybil Attacks via Social Networks. In *Proceedings of ACM SIGCOMM*, volume 36. ACM Press, 2006.