Practical Characterization of Large Networks Using Neighborhood Information

Pinghui Wang[1], Junzhou Zhao[2], Bruno Ribeiro[3], John C.S. Lui[1],
Don Towsley[3], and Xiaohong Guan[2]
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
[2]MOE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, China
[3]Department of Computer Science, University of Massachusetts Amherst, MA, USA
{phwang, jzzhao, xhguan}@sei.xjtu.edu.cn, cslui@cse.cuhk.edu.hk,
{towsley, ribeiro}@cs.umass.edu

ABSTRACT

Characterizing large online social networks (OSNs) through node querying is a challenging task. OSNs often impose severe constraints on the query rate, hence limiting the sample size to a small fraction of the total network. Various ad-hoc subgraph sampling methods have been proposed, but many of them give biased estimates and no theoretical basis on the accuracy. In this work, we focus on developing sampling methods for OSNs where querying a node also reveals partial structural information about its neighbors. Our methods are optimized for NoSQL graph databases (if the database can be accessed directly), or utilize Web API available on most major OSNs for graph sampling. We show that our sampling method has provable convergence guarantees on being an unbiased estimator, and it is more accurate than current state-of-the-art methods. We characterize metrics such as node label density estimation and edge label density estimation, two of the most fundamental network characteristics from which other network characteristics can be derived. We evaluate our methods on-the-fly over several live networks using their native APIs. Our simulation studies over a variety of offline datasets show that by including neighborhood information, our method drastically (4-fold) reduces the number of samples required to achieve the same estimation accuracy of state-of-the-art methods.

1. INTRODUCTION

The literature on sampling large networks is vast and rich. Various techniques have been proposed for subgraph sampling and characterization of large networks [12, 17, 21] (refer to Ahmed et al. [3] for a good survey). These techniques, however, often lack provable guarantees. This means that after sampling a fraction of a large network, one has no guarantees whether the metrics obtained are to be trusted. Fortunately, researchers have recently made a push towards network characterization through sampling with provable properties and accuracy guarantees.

Techniques adapted to sample networks stored at NoSQL graph databases or accessible from Web APIs (e.g. available on Facebook, Foursquare, Pinterest, among others) must refrain from randomly sampling too many nodes and all together avoid sampling edges, either due to caching inefficiencies or limitations in the API. In practice, most online social networks (OSNs), including those we present in this study, do not provide random sampling primitives. Practitioners perform random sampling by guessing user IDs in the user ID space, which, if sparsely populated, imposes a large number of query misses until a valid user is found. In this context, techniques that heavily rely on random sampling, such as Dasgupta et al. [7], suffers from low query rate. Dasgupta et al. partially compensates the low query rate through the use of neighborhood information present in the node query reply of a number of major OSNs (e.g. Foursquare, Pinterest, Sina microblog). Similarly, graph streaming techniques, such as Ahmed et al. [3], are also not well adapted to this environment as they require visiting all edges, which is prohibitively expensive in a large network with millions or even billions of edges.

Recently, great focus has been placed on developing techniques that use specially constructed "*crawlers*" to query the network and to provide asymptotically unbiased estimates of a handful of network characteristics [10, 27]. Chief among these techniques are random walks, which provide provable accuracy and convergence guarantees (see Ribeiro and Towsley [28] and Avrechenkov et al. [4]). Random walks present a number of desirable properties that are useful to characterize large networks; (1) they require either few or no independently sampled nodes and produce asymptotically unbiased estimates and accuracy guarantees under mild conditions for a large family of directed* and undirected networks, even when the network has multiple disconnected components, as long as some limited amount of random sampling is available [4, 27, 28], (2) use crawling to collect samples (which effectively implements importance sampling on node degrees), and can achieve relatively high query rates on NoSQL graph databases or using Web APIs, and (3) does not require any advance knowledge of the network, such as its size or topology. However, existing random walk (RW) techniques do not take advantage of the extra neighborhood information, despite the fact that neighborhood information is readily available in many OSNs at (practically) no sampling cost (obtained from the node query reply). Including such extra information in RW-based estimator while retaining unbiased guarantees is challenging due to different types of biases involved in the sampling process.

Contributions: In this work, we consider the generalization of RW sampling and combine current state-of-the-art estimators to include

*In directed networks where querying a node retrieves both the incoming and outgoing edges of that node.

neighborhood information. Our estimator drastically reduces (by 4-fold) the number of samples required to achieve the same estimation accuracy. Examples of OSNs that provide neighborhood information are found everywhere, e.g. Pinterest [24], Foursquare [8], Sina microblog [32], and Xiami [35]. Our generalization allows us to include neighboring information in the estimation of a variety of network characteristics from nodes sampled using a random walk-based technique called Frontier Sampling [27]. We also implement our method to sample networks in the wild, and discover that the degree distribution of Foursquare *does not* exhibit a heavy tail, and, by using adapted versions of state-of-the-art algorithms we also estimate that the average distance between users to be 5.8, which is between values of average distances observed for Twitter (4.1) and MSN messenger network (6.6) [16, 18].

This paper is organized as follows. Several basic sampling techniques are summarized in Section 2. In Sections 3, we present the methodology of using neighborhood information to estimate node label density. In Section 4, we propose methods using neighborhood information to estimate edge label density. The performance evaluation and testing results are presented in Section 7. Section 8 presents applications of our methods on Foursquare and Pinterest websites. Section 9 summarizes related work. Section 10 concludes.

2. PRELIMINARIES

In this section we introduce *Frontier Sampling* (FS), a generalization of random walk sampling methods [27]. For ease of presentation, we assume undirected, connected, and non-bipartite networks. Unless we state otherwise, denote by $G_d = (V, E_d)$ the directed graph under study, and $G = (V, E)$ the undirected graph generated by ignoring the direction of edges in G_d.

The above assumptions are not too restrictive for the following reasons: For *directed networks*, our method can be trivially adapted to include directed OSNs such as Twitter and Flickr, which provide direct Web API access to the incoming and outgoing edges of a node (such that our crawler can traverse on an undirected version of the network). Sometimes, there is a cost associated with obtaining the incoming and outgoing edges of nodes with large in or out degrees (e.g., Flickr provides only up to one hundred incoming or outgoing edges per query). For *connected networks*, our previous work [4] consider random walk sampling and show how to augment disconnected graphs using randomly sampled nodes into connected graphs without changing the properties of the estimators. Note that PageRank-style jumps are not suited for the task as they create unknown biases in the estimators, see [4]. A trivial adaptation of the above argument can be used to show that FS retains its properties on disconnected or bipartite networks if a limited amount of random sampling is available (e.g. one hundred sampled nodes in a network with millions of nodes).

Our accuracy guarantees follow directly from our results in Ribeiro and Towsley [28], which provides provable guarantees on the mean squared error (MSE) accuracy of the degree distribution estimates given by random walk sampling as a function of the number of samples and the first nontrivial eigenvalue of the Laplacian of G.

2.1 Frontier Sampling (FS)

Frontier Sampling [27] is a fully distributed sampling algorithm that performs m independent RWs on G. If $m = 1$, FS behaves exactly like a RW. When $m > 1$, compared to a single RW, FS can be more robust to the problems that arise from the walker getting trapped at a loosely connected component of G. The k-th FS walker starts at node $s_0^{(k)}$, $k = 1, \ldots, m$. Each FS walker has a predefined budget T (we explain how T is chosen at the end of this section). Denote by $\mathcal{N}(u)$ the set of neighbors of any node $u \in V$, and by $d_u = |\mathcal{N}(u)|$ the degree of u. At each step an FS walker at node u moves to a randomly node from $\mathcal{N}(u)$, deducting from the budget T a random quantity $X \sim \text{Exp}(d_u)$, an exponentially distributed random variable with mean $1/d_u$. FS stops when T becomes negative. If G is a connected and non-bipartite graph, the probability that a node v is sampled by FS converges to the following distribution

$$\pi_v^{\text{FS}} = \frac{d_v}{2|E|}, \quad v \in V.$$

FS can also be used to sample edges randomly, as the probability of traversing an edge $(u, v) \in E$ converges to the uniform distribution [27], that is

$$\rho_{u,v} = \frac{1}{|E|}, \quad (u, v) \in E.$$

The choice of budget T is often defined as the average number of nodes that one wishes to sample, n, divided by the number of FS walkers m times the average degree, $\bar{d}$. In practice, one does not need to know $\bar{d}$ as T may be increased dynamically on-the-fly. Because we can adjust T on-the-fly, in what follows we take the liberty to assume that FS samples exactly n nodes.

We merge all n samples collected by the FS walkers into a single stream $(s_1, \ldots, s_n)$ in any order. Let s_i be the i-th node sampled by FS, $i \geq 1$. Let "a.s." denote "almost sure" convergence, i.e., that the event of interest happens with probability one. Then,

LEMMA 1 (THEOREM 4.1 [27]). *For any function $\phi(v) : V \to \mathbb{R}$, where $\sum_{\forall v \in V} \phi(v) < \infty$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \phi(s_i) \xrightarrow{a.s.} \sum_{\forall v \in V} \phi(v) \pi_v^{\text{FS}}.$$

An important property of FS is that seeding $m \gg 1$ walkers with i.i.d. nodes sampled uniformly at random (UNI) is equivalent to initializing walkers in steady state [27, Theorem 5.4]. In practice, $m = 100$ initial UNI samples nodes, $(s_0^{(1)}, \ldots, s_0^{(100)})$, are enough to initialize the FS walkers close to their joint steady state (thus, requiring only 100 "expensive" UNI sampled nodes). This property of FS, sampling nodes according to their degree but being initialized in steady state using UNI, is the result of a Markov chain trick used to uniformize continuous time Markov chains into discrete transition probability matrices. For more details see Ribeiro and Towsley [27].

In practice FS also works on disconnected graph as long as the initial choice of nodes is chosen from UNI and m is large. However, for disconnected graphs no convergence property can be shown as the different FS walkers do not mix. In the absence of UNI samples, FS behaves much like a single RW and it is advised that m should be kept reasonably small. Recent results in Ribeiro and Towsley [28] show provable guarantees of accuracy of RW-based methods:

LEMMA 2 (THEOREM III.1 [28]). *Let $w_1, \ldots, w_n$ be a set of nodes sampled independently proportionally to their degree, and $s_1, \ldots, s_n$ is a sequence of RW sampled nodes, then*

$$\textit{MSE}\left(\frac{1}{n}\sum_{i=1}^{n}\phi(s_i)\right) \leq \frac{1}{1-\alpha}\textit{MSE}\left(\frac{1}{n}\sum_{i=1}^{n}\phi(w_i)\right), \ \forall k, \quad (1)$$

where α is the first nontrivial eigenvalue of the Laplacian, $\mathcal{L}$, of G. The bound is tight [28].

The bound in Lemma 2 shows that the increase in MSE of RW (FS with $m = 1$) sampling is at most $1/(1-\alpha)$ times larger than the

MSE of independently sampling nodes according to their degree (importance sampling proportional to the degree). The value of α goes to zero as the graph gets more connected, reducing the gap between MSEs of RW and i.i.d. importance sampling.

The value of α can also be "artificially" decreased using the RW with restarts (RWRST) [4], which augments the graph with edges of small weights. RWRSTs are not to be confused with PageRank [23] as the two Markov chains have remarkably different statistical properties. Moreover, empirically FS achieves similar fast mixing if seed nodes are chosen uniformly (UNI) and $n \gg m$ [27]. Not surprisingly, FS behavior is remarkably similar to RWRST. A RWRST is a RW that at node $v \in V$ chooses to jump to a randomly chosen node with probability $h/(d_v + h)$ or select a neighbor of v with probability $d/(d_v + h)$, where $h > 0$ is a parameter of the algorithm. A RWRST stopped at the $m + 1$ restart can be emulated by m independent RWs that at node $v \in V$ stop with probability $h/(d_v + h)$. Using results in Avrachenkov et al. [4, Theorem 2.1] it is easy to show that the MSE of a RWRST over a d-regular bounded by multiplying the right hand side of (1) by $(1 - \alpha)/(1 - \alpha d/(d + h))$.[†] Moreover, for any fixed number of sampled nodes n, the value of h is a random variable that increases with m, implying that the latter MSE bound should decrease with m. While this result is particular to d-regular graphs, these are likely to hold for a large class of graphs. The similarity between FS and the m simulated RWRST indicates that the FS MSE likely decreases with m, as long as $m \ll n$ and FS seed nodes are UNI sampled. Unfortunately, a formal proof eluded us, as analyzing the FS Markov chain is more challenging than the analysis of RWRST in Avrachenkov et al. [4] . We leave this analysis as future work.

3. NODE LABEL DENSITY ESTIMATION

In what follows we propose methods for estimating node label density. Define $L(v)$ to be the node label of node v under study, with range $\boldsymbol{L} = \{l_1, ..., l_K\}$. Denote by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ the node label density, where θ_k $(1 \le k \le K)$ is the fraction of nodes with label l_k. For example, when $L(v)$ is defined as the degree of node v, then $\boldsymbol{\theta}$ is the node degree distribution of G. If $L(v)$ denote the gender of node (or user) v, then $\boldsymbol{\theta}$ is the gender distribution of the OSN under study.

3.1 Simple Estimators of Node Densities

To estimate $\boldsymbol{\theta}$ based on sampled nodes $\{s_i\}_{i=1,\ldots,n}$, the stationary distribution of sampling methods (e.g. UNI, RW, and FS) $\boldsymbol{\pi} = (\pi_v : v \in V)$ is needed to correct the bias induced by the underlying sampling method. For $v \in V$, we have $\pi_v = \frac{1}{|V|}$ for UNI, and $\pi_v = \frac{d_v}{2|E|}$ for RW and FS. Since the values of $|V|$ and $|E|$ are usually unknown, unbiasing the error is not straightforward. Instead, one may use a non-normalized stationary distribution $\hat{\boldsymbol{\pi}} = (\hat{\pi}_v : v \in V)$ to reweight sampled nodes s_i $(1 \le i \le n)$, where $\hat{\pi}_v$ is computed as

$$\hat{\pi}_v \propto \begin{cases} 1 & \text{for UNI,} \\ d_v & \text{for RW and FS,} \end{cases} \tag{2}$$

Let $\mathbf{1}(\mathbf{P})$ be the indicator function that equals one when predicate $\mathbf{P}$ is true, and zero otherwise, θ_k is estimated as follows

$$\hat{\theta}_k = \frac{1}{C} \sum_{i=1}^{n} \frac{\mathbf{1}(L(s_i) = l_k)}{\hat{\pi}_{s_i}}, \quad 1 \le k \le K, \tag{3}$$

[†]Note that in a d-regular graph one should think that ϕ is a meaningful density function over the nodes, as estimating the degree distribution on a graph that only has degree d is meaningless.

where $C = \sum_{i=1}^{n} \hat{\pi}_{s_i}^{-1}$. Ribeiro and Towsley [27] shows that $\hat{\theta}_k$ $(1 \le k \le K)$ is an asymptotically unbiased estimate of θ_k.

3.2 Estimators Using Neighborhood Information of Sampled Nodes

When the degrees and the node labels of sampled nodes' neighbors are available, we propose the following estimator utilizing this free neighborhood information

$$\breve{\theta}_k = \frac{1}{\breve{C}} \sum_{i=1}^{n} \sum_{w \in \mathcal{N}(s_i)} \frac{\mathbf{1}(L(w) = l_k)}{\hat{\pi}_{s_i} d_w}, \quad 1 \le k \le K, \tag{4}$$

where $\breve{C} = \sum_{i=1}^{n} \sum_{w \in \mathcal{N}(s_i)} \hat{\pi}_{s_i}^{-1} d_w^{-1}$. The above estimator is similar to one proposed in Dasgupta et al. [7]. However the estimator in Dasgupta et al. [7] requires $|V|$ to be known in advance, which is usually not available. Moreover, Dasgupta et al. [7] focuses on designing independent node sampling methods (e.g. UNI, independent weighted node sampling), which we argued has a low query rate. Whereas we focus on crawling methods such as RW and FS. For each node $v \in V$, Eq. (2) shows that $\pi_v / \hat{\pi}_v$ has the same value, denoted as C_π. In what follows we analyze the accuracy of estimator $\breve{\theta}_k$ $(1 \le k \le K)$.

THEOREM 1. *$\breve{\theta}_k$ $(1 \le k \le K)$ is an asymptotically unbiased estimate of θ_k.*

Proof. Applying Lemma 1, we have

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{w \in \mathcal{N}(s_i)} \frac{\mathbf{1}(L(w) = l_k)}{\hat{\pi}_{s_i} d_w} \right] \\ &\xrightarrow{a.s.} \sum_{v \in V} \left(\pi_v \sum_{w \in \mathcal{N}(v)} \frac{\mathbf{1}(L(w) = l_k)}{\hat{\pi}_v d_w} \right) \\ &= C_\pi \sum_{w \in V} \sum_{v \in \mathcal{N}(w)} \frac{\mathbf{1}(L(w) = l_k)}{d_w} \\ &= C_\pi \sum_{w \in V} \mathbf{1}(L(w) = l_k) = C_\pi |V| \theta_k. \end{aligned}$$

Similarly we prove that $\lim_{n\to\infty} \breve{C}/n \xrightarrow{a.s.} C_\pi |V|$. Therefore we have $\lim_{n\to\infty} \breve{\theta}_k \xrightarrow{a.s.} \theta_k$. □

We can easily find that neighbors of sampled nodes are biased to nodes with high degrees even for UNI. Therefore, $\breve{\theta}_k$ is estimator based on biased samples. Ribeiro and Towsley [27] shows that UNI has smaller mean square error (MSE) for small degree nodes than biased sampling methods such as RW. It is consistent with our results in Section 7, which show that $\breve{\theta}_k$ may exhibit larger MSE than $\hat{\theta}_k$ defined in (3). Thus, we present the following mixture estimator for θ_k

$$\hat{\theta}_k^{\text{mix}} = \alpha_k \hat{\theta}_k + (1 - \alpha_k) \breve{\theta}_k, \quad 1 \le k \le K, \tag{5}$$

where parameter α_k lies between zero and one, and is used to determine the relative importance of two estimates $\hat{\theta}_k$ and $\breve{\theta}_k$. Suppose that $\hat{\theta}_k$ and $\breve{\theta}_k$ are independent. Then $\hat{\theta}_k^{\text{mix}}$ has the smallest variance when $\alpha_k = \frac{\text{Var}(\breve{\theta}_k)}{\text{Var}(\hat{\theta}_k) + \text{Var}(\breve{\theta}_k)}$.

In what follows we propose estimators of node label density $\boldsymbol{\theta}$ using the available neighborhood information of sampled nodes for directed OSNs such as Pinterest, Sina microblog, and Xiami, where a node has knowledge of in-degrees (the number of followers) and out-degrees (the number of following) of its incoming neighbors

and outgoing neighbors. For a node $v \in V$, denote by $d_v^{(\mathrm{I})}$ its in-degree, $d_v^{(\mathrm{O})}$ its out-degree, $\mathcal{N}^{(\mathrm{I})}(v) = \{u : (u,v) \in E_d\}$ the set of its followers, and $\mathcal{N}^{(\mathrm{O})}(v) = \{u : (v,u) \in E_d\}$ the set of its following. Define $\psi(u,v) = 0$ when v is not a neighbor of u, $\psi(u,v) = 2$ when v is an out-going and incoming neighbor of u, and otherwise $\psi(u,v) = 1$. Let $\mathcal{N}(v) = \mathcal{N}^{(\mathrm{I})}(v) \cup \mathcal{N}^{(\mathrm{O})}(v)$. Using the properties of sampled nodes' neighbors, we estimate θ_k as follows

$$\breve{\theta}_k^* = \frac{1}{\breve{C}_d} \sum_{i=1}^{n} \sum_{w \in \mathcal{N}(s_i)} \frac{\psi(s_i,w)\mathbf{1}(L(w)=l_k)}{\hat{\pi}_{s_i}(d_w^{(\mathrm{I})} + d_w^{(\mathrm{O})})}, \quad 1 \le k \le K,$$

where $\breve{C}_d = \sum_{i=1}^{n} \sum_{w \in \mathcal{N}(s_i)} \psi(s_i,w) \hat{\pi}_{s_i}^{-1} (d_w^{(\mathrm{I})} + d_w^{(\mathrm{O})})^{-1}$.

THEOREM 2. *$\breve{\theta}_k^*$ $(1 \le k \le K)$ is an asymptotically unbiased estimate of θ_k.*

Proof. Applying Lemma 1, we have

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{w \in \mathcal{N}(s_i)} \frac{\psi(s_i,w)\mathbf{1}(L(w)=l_k)}{\hat{\pi}_{s_i}(d_w^{(\mathrm{I})} + d_w^{(\mathrm{O})})} \right] \\
&\xrightarrow{a.s.} \sum_{v \in V} \left(\pi_v \sum_{w \in \mathcal{N}(v)} \frac{\psi(v,w)\mathbf{1}(L(w)=l_k)}{\hat{\pi}_v(d_w^{(\mathrm{I})} + d_w^{(\mathrm{O})})} \right) \\
&= C_\pi \sum_{v \in V} \sum_{w \in \mathcal{N}(v)} \frac{\psi(v,w)\mathbf{1}(L(w)=l_k)}{d_w^{(\mathrm{I})} + d_w^{(\mathrm{O})}} \\
&= C_\pi \sum_{w \in V} \sum_{v \in \mathcal{N}(w)} \frac{\psi(w,v)\mathbf{1}(L(w)=l_k)}{d_w^{(\mathrm{I})} + d_w^{(\mathrm{O})}} \\
&= C_\pi \sum_{w \in V} \mathbf{1}(L(w)=l_k) = C_\pi |V| \theta_k.
\end{aligned}$$

The third equation holds because $\sum_{v \in \mathcal{N}(w)} \psi(w,v) = d_w^{(\mathrm{I})} + d_w^{(\mathrm{O})}$. Similarly we proof that $\lim_{n\to\infty} \breve{C}^*/n \xrightarrow{a.s.} C_\pi |V|$. Therefore we have $\lim_{n\to\infty} \breve{\theta}_k^* \xrightarrow{a.s.} \theta_k$. □

Next, we propose methods for graphs such as Citeseerx website, where we can obtain a node's neighbors' out-degrees when we sample a node. However in-degrees of sampled nodes' neighbors are not available. Then we estimate node label density θ_k $(1 \le k \le K)$ based on sampled nodes and their out-going neighbors, that is

$$\breve{\theta}_k^{(\mathrm{O})} = \frac{1}{\breve{C}_d^*} \sum_{i=1}^{n} \left(\frac{\gamma \mathbf{1}(L(s_i)=l_k)}{\hat{\pi}_{s_i}(d_{s_i}^{(\mathrm{I})} + \gamma)} + \sum_{w \in \mathcal{N}^{(\mathrm{O})}(s_i)} \frac{\mathbf{1}(L(w)=l_k)}{\hat{\pi}_{s_i}(d_w^{(\mathrm{I})} + \gamma)} \right)$$

where $\breve{C}_d^* = \sum_{i=1}^{n} \left(\frac{\gamma}{\hat{\pi}_{s_i}(d_{s_i}^{(\mathrm{I})} + \gamma)} + \sum_{w \in \mathcal{N}^{(\mathrm{O})}(s_i)} \frac{1}{\hat{\pi}_{s_i}(d_w^{(\mathrm{I})} + \gamma)} \right)$, and $\gamma > 0$.

THEOREM 3. *$\breve{\theta}_k^{(O)}$ $(1 \le k \le K)$ is an asymptotically unbiased estimate of θ_k.*

Proof. Denote by $V_0^{(\mathrm{I})}$ the set of nodes in V whose in-degrees are larger than 0. Clearly only nodes in $V_0^{(\mathrm{I})}$ can appear in a node's out-going neighbor list. Applying Lemma 1, we have

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\gamma \mathbf{1}(L(s_i)=l_k)}{\hat{\pi}_{s_i}(d_{s_i}^{(\mathrm{I})} + \gamma)} + \sum_{w \in \mathcal{N}^{(\mathrm{O})}(s_i)} \frac{\mathbf{1}(L(w)=l_k)}{\hat{\pi}_{s_i}(d_w^{(\mathrm{I})} + \gamma)} \right] \\
&\xrightarrow{a.s.} \sum_{v \in V} \pi_v \left(\frac{\gamma \mathbf{1}(L(v)=l_k)}{\hat{\pi}_v(d_v^{(\mathrm{I})} + \gamma)} + \sum_{w \in \mathcal{N}^{(\mathrm{O})}(v)} \frac{\mathbf{1}(L(w)=l_k)}{\hat{\pi}_v(d_w^{(\mathrm{I})} + \gamma)} \right) \\
&= C_\pi \left(\sum_{v \in V} \frac{\gamma \mathbf{1}(L(v)=l_k)}{d_v^{(\mathrm{I})} + \gamma} + \sum_{v \in V \setminus V_0^{(\mathrm{I})}} \frac{d_v^{(\mathrm{I})} \mathbf{1}(L(v)=l_k)}{d_v^{(\mathrm{I})} + \gamma} \right) \\
&= C_\pi \left(\sum_{v \in V_0^{(\mathrm{I})}} \mathbf{1}(L(v)=l_k) + \sum_{v \in V \setminus V_0^{(\mathrm{I})}} \mathbf{1}(L(v)=l_k) \right) \\
&= C_\pi \sum_{w \in V} \mathbf{1}(L(w)=l_k) = C_\pi |V| \theta_k.
\end{aligned}$$

The first equation holds because

$$\begin{aligned}
&\sum_{v \in V} \sum_{w \in \mathcal{N}^{(\mathrm{O})}(v)} \frac{\mathbf{1}(L(w)=l_k)}{d_w^{(\mathrm{I})} + \gamma} \\
&= \sum_{w \in V \setminus V_0^{(\mathrm{I})}} \sum_{v \in \mathcal{N}^{(\mathrm{O})}(w)} \frac{\mathbf{1}(L(w)=l_k)}{d_w^{(\mathrm{I})} + \gamma} \\
&= \sum_{w \in V \setminus V_0^{(\mathrm{I})}} \frac{d_w^{(\mathrm{I})} \mathbf{1}(L(w)=l_k)}{d_w^{(\mathrm{I})} + \gamma}.
\end{aligned}$$

Similarly we proof that $\lim_{n\to\infty} X/n \xrightarrow{a.s.} C_\pi |V|$. Therefore we have $\lim_{n\to\infty} \breve{\theta}_k^{(\mathrm{O})} \xrightarrow{a.s.} \theta_k$. □

Similar to the mixture estimator (5), $\breve{\theta}_k^*$ and $\breve{\theta}_k^{(\mathrm{O})}$ can be combined with $\hat{\theta}_k$ to estimate θ_k more accurately.

4. EDGE LABEL DENSITY ESTIMATION

Define $L(u,v)$ to be the label of edge (u,v), with range $\boldsymbol{L'} = \{l'_1, ..., l'_{K'}\}$. Denote the edge label density by $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{K'})$, where τ_k $(1 \le k \le K')$ is the fraction of edges with label l'_k. For undirected graph G, we let edge label function $L(u,v) = L(v,u)$. For example, when define $L(u,v) = (\min\{d_u, d_v\}, \max\{d_u, d_v\})$ for edge (u,v) in undirected graph G, the pair of degrees of nodes u and v, $\boldsymbol{\tau}$ is the joint node degree distribution. Note that the labels of edges (u,v) and (v,u) in directed graph G_d may not be the same. In this section we propose methods for estimating $\boldsymbol{\tau}$ for undirected graphs and directed graphs respectively.

4.1 Simple Estimators of Edges Densities

Based on edges $\{(u_i, v_i)\}_{i=1,\ldots,n}$ sampled by RW and FS, [27] estimates τ_k for an undirected graph G as follows

$$\hat{\tau}_k = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(L'(u_i,v_i) = l'_k), \quad 1 \le k \le K'. \tag{6}$$

It shows that $\hat{\tau}_k$ $(1 \le k \le K')$ is an asymptotically unbiased estimate of τ_k for undirected graphs. Similarly, in this paper we estimate $\boldsymbol{\tau}$ of directed graph G_d as follows

$$\begin{aligned}
\hat{\tau}_k^* = \frac{1}{H_d} \sum_{i=1}^{n} \big(&\mathbf{1}(L'(u_i,v_i) = l'_k)\mathbf{1}((u_i,v_i) \in E_d) \\
&+ \mathbf{1}(L'(v_i,u_i) = l'_k)\mathbf{1}((v_i,u_i) \in E_d) \big).
\end{aligned}$$

where $H_d = \sum_{i=1}^{n} \mathbf{1}((u_i, v_i) \in E_d) + \mathbf{1}((v_i, u_i) \in E_d)$. We can easily prove $\hat{\tau}_k^*$ $(1 \le k \le K')$ is an asymptotically unbiased estimate of τ_k for directed graph G_d.

4.2 Estimators Using Neighborhood Information of Sampled Nodes

In this paper we assume that we can obtain the labels of all (incoming and outgoing) edges of a node when we query a node from G (G_d). Using the neighborhood information of sampled nodes s_i $(1 \le i \le n)$ obtained by UNI, RW and FS, we estimate τ_k of G as follows

$$\breve{\tau}_k = \frac{1}{\breve{H}} \sum_{i=1}^{n} \sum_{w \in \mathcal{N}(s_i)} \frac{\mathbf{1}(L'(s_i, w) = l'_k)}{\hat{\pi}_{s_i}}, \quad 1 \le k \le K', \quad (7)$$

where $\breve{H} = \sum_{i=1}^{n} \sum_{w \in \mathcal{N}(s_i)} \hat{\pi}_{s_i}^{-1}$. Then we have

THEOREM 4. *$\breve{\tau}_k$ $(1 \le k \le K')$ is an asymptotically unbiased estimate of τ_k for undirected graphs.*

Proof. Applying Lemma 1, we have

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{w \in \mathcal{N}(s_i)} \frac{\mathbf{1}(L'(s_i, w) = l'_k)}{\hat{\pi}_{s_i}} \\
&\xrightarrow{a.s.} \sum_{v \in V} \left(\pi_v \sum_{w \in \mathcal{N}(v)} \frac{\mathbf{1}(L'(v, w) = l'_k)}{\hat{\pi}_v} \right) \\
&= C_\pi \sum_{v \in V} \sum_{w \in \mathcal{N}(v)} \mathbf{1}(L'(v, w) = l'_k) = 2C_\pi |E| \tau_k.
\end{aligned}$$

Similarly we have $\lim_{i\to\infty} \mathrm{E}[\breve{H}/n] \to 2C_\pi|E|$. Therefore we have $\lim_{n\to\infty} \breve{\tau}_k \xrightarrow{a.s.} \tau_k$. □

Utilizing the free neighborhood information of sampled nodes s_i $(1 \le i \le n)$, we estimate τ_k of G_d as follows

$$\breve{\tau}_k^* = \frac{1}{\breve{H}_d} \sum_{i=1}^{n} \sum_{w \in \mathcal{N}(s_i)} \left(\frac{\mathbf{1}(L'(s_i, w) = l'_k)\mathbf{1}((s_i, w) \in E_d)}{\hat{\pi}_{s_i}} + \frac{\mathbf{1}(L'(w, s_i) = l'_k)\mathbf{1}((w, s_i) \in E_d)}{\hat{\pi}_{s_i}} \right)$$

where $\breve{H}_d = \sum_{i=1}^{n} \sum_{w \in \mathcal{N}(s_i)} \frac{\mathbf{1}((s_i,w)\in E_d) + \mathbf{1}((w,s_i)\in E_d)}{\hat{\pi}_{s_i}}$. Similar to Theorem 4, we have

THEOREM 5. *$\breve{\tau}_k^*$ $(1 \le k \le K')$ is an asymptotically unbiased estimate of τ_k for directed graphs.*

In summary, $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \ldots, \hat{\tau}_{K'})$ and $\hat{\boldsymbol{\tau}}^* = (\hat{\tau}_1^*, \ldots, \hat{\tau}_{K'}^*)$ computed as described above form asymptotically unbiased estimates of $\boldsymbol{\tau}$ for undirected and directed graphs respectively. When properties of sampled nodes' neighbors are available, we utilize all edge labels observed from this neighborhood information, and provide asymptotically unbiased estimates $\breve{\boldsymbol{\tau}} = (\breve{\tau}_1, \ldots, \breve{\tau}_{K'})$ and $\breve{\boldsymbol{\tau}}^* = (\breve{\tau}_1^*, \ldots, \breve{\tau}_{K'}^*)$ of $\boldsymbol{\tau}$ for undirected and directed graphes respectively.

5. HIGH DEGREE NODE DETECTION

In this section, we study the problem of detecting the N nodes with the largest degrees in undirected graph $G = (V, E)$. Let S be the set of nodes sampled by methods such as RW. Previous methods use the N nodes in S with the largest degrees to estimate high degree nodes [2, 5]. In [5], weighted RW (WRW) is used to detect high degree nodes. WRW can be viewed as a RW over a weighted graph, where each edge $(u, v) \in E$ has a positive weight $w(u, v) = w(v, u)$ [6]. At each step, WRW selects the next-hop node v at random among the neighbors of the current node u with probability proportional to weight $w(u, v)$. WRW (with well defined edge weights) and RW are efficient for detecting high degree nodes, since they are biased to sample high degree nodes [5, 27]. Note that the WRW proposed in [5] sets weight $w(u, v) = (d_u d_v)^\beta$ for each edge $(u, v) \in E$ for detecting top-N high degree nodes, which indicates that at each step their WRW need to obtain degrees of current sampled node's neighbors. However their description does not account for the cost of retrieving this information. In [20], a new method, expansion sampling (XS), is proposed for detecting high degree nodes. Denote by $\mathcal{N}(S)$ the neighborhood of S, where $\mathcal{N}(S)$ consists of nodes in $V \setminus S$ that are neighbors of nodes in S, that is $\mathcal{N}(S) = \{u : u \in V \setminus S, \exists v \in S, (u, v) \in E\}$. Starting from a random node s, and $S = \{s\}$, XS adds the node in $\mathcal{N}(S)$ which has the largest number of neighbors in $V \setminus (\mathcal{N}(S) \cup S)$ to S, and repeats this process. For a node $u \in \mathcal{N}(S)$, denote by $d_u^{(S)}$ the number of edges between u and nodes in S, and $d_u^{(\mathcal{N}(S))}$ the number of edges between u and nodes in $\mathcal{N}(S)$. Then, the number of its neighbors in $V \setminus (\mathcal{N}(S) \cup S)$ equals $d_u - d_u^S - d_u^{(\mathcal{N}(S))}$. From knowledge of edges of nodes in S, we know d_u and d_u^S. However $d_u^{(\mathcal{N}(S))}$ cannot be obtained based on available information of S and $\mathcal{N}(S)$. In order to identify the node in $\mathcal{N}(S)$ which has the largest number of neighbors in $V \setminus (\mathcal{N}(S) \cup S)$, it is necessary to crawl all nodes in $\mathcal{N}(S)$. The original description of XS [20] does not account for this cost. On the other hand when each node has knowledge of its neighbors' degrees. It is possible to identify the node in $\mathcal{N}(S)$ that has the largest number of neighbors in $V \setminus S$. Therefore we propose a Modified XS (MXS) method, which adds this node to S at each step. Finally we output the N nodes with the largest degrees in $\mathcal{N}(S) \cup S$ as the final results. The above methods can be easily modified to identify N nodes with the largest in-degrees or out-degrees in directed graph such as Sina microblog, Tencent microblog, and Xiami, where a node has knowledge of its neighbors' out-degrees and in-degrees. Here at each step MXS adds the node with the largest sum of in-degree and out-degree in $\mathcal{N}(S)$ to S.

6. SHORT PATH DISCOVERY

In this section, we study the problem of performing topology discovery and message routing with incomplete topological information, which is important for applications such as discovery of short paths between OSN users and routing algorithms (e.g. Bubble Rap [13]) for delivering messages between users using a social network. Formally the problem is: Two nodes u and v are looking for short paths on undirected graph G. Ribeiro et al. [26] find that a RW has the ability to observe a large fraction of the edges by visiting a relatively small number of nodes on power law graphs. Here an edge is observed when at least one of its endpoints is visited by the RW. They propose a RW based short path discovery algorithm works as follows: Two RWs are started from u and v separately. Each RW takes B steps. Let S be the set of nodes sampled by two RWs. Finally They use the shortest path in observed graph $G^* = (V^*, E^*)$ for routing between u and v, where $V^* = S \cup \mathcal{N}(S)$ and E^* consists of edges in E which have at least one endpoint contained by S. From Section 5, we know that WRW and MXS can efficiently find high degree nodes and observe more edges based on neighborhood information. We propose two new methods, which perform a WRW and MXS starting from two initial nodes u and v respectively. We finally use the shortest path

in graph $G^* = (V^*, E^*)$ observed by WRW or MXS for routing between u and v.

7. DATA EVALUATION

We perform our experiments on a variety of real world networks that are summarized in Table 1. Xiami is a popular website devoted to music streaming and music recommendations. Similar to Twitter, Xiami builds a social network based on follower and following relationships. Flickr and YouTube are popular photo sharing and video sharing websites. In these websites, a user can subscribe to other user updates such as blogs and photos. These networks can be represented by direct graphs, with nodes representing users and a directed edge from u to v represents that user u subscribes to user v. Epinions is a who-trusts-whom OSN providing general consumer reviews, where a directed edge from u to v represents that user u trusts user v. Slashdot is a technology-related news website for its specific user community, where a directed edge from u to v represents that user u tags user v as a friend or foe. In the following experiments, we evaluate our methods in comparisons with previous methods based on the largest connected component (LCC) of these graphs under the same sampling budget B, where B is defined as the number of sampled nodes.

Table 1: Overview of graph datasets used in our simulations."directed-edges" refers to the number of directed edges in a directed graph, "edges" refers to the number of edges in an undirected graph, and "LCC" refers to the largest connected component of a given graph.

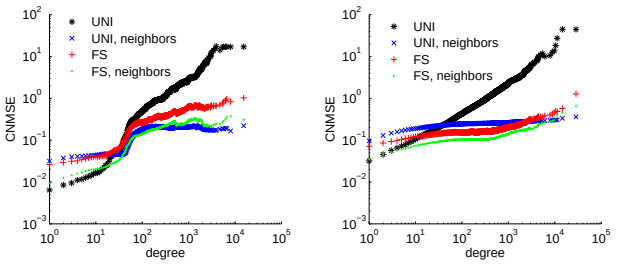
Graph	LCC		
	nodes	edges	directed-edges
Xiami [34]	1,748,010	16,015,779	16,568,449
YouTube [22]	1,134,890	2,987,624	4,942,035
Flickr [22]	1,624,992	15,476,835	22,477,014
Soc-Epinions [30]	75,877	405,739	405,739
Soc-Slashdot [19]	77,360	469,180	828,161

7.1 Node Label Density Estimation

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ be the (in-) degree distribution, where θ_k $(1 \le k \le K)$ is the fraction of nodes with (in-) degree k. In our study, we estimate both θ_k and $\xi_k = \sum_{i=k+1}^{K} \theta_i$, the CCDF (complementary cumulative distribution function) of $\boldsymbol{\theta}$, which is the statistic of choice when it comes to display (in-degree) degree distributions. For estimator $\hat{\theta}_k$, we define the normalized root mean square error (NMSE) as $\text{NMSE}(\hat{\theta}_k) = \sqrt{\text{E}[(\hat{\theta}_k - \theta_k)^2]}/\theta_k$, $k = 1, 2, \ldots$. In the following experiments, we use 1,000 independent runs to estimate $\text{E}[(\hat{\theta}_k - \theta_k)^2]$. Similarly, we define the NMSE of the CCDF of $\boldsymbol{\theta}$, which we denote as the CNMSE to avoid confusion with the NMSE of $\boldsymbol{\theta}$.

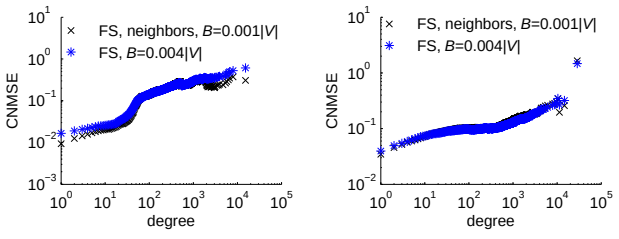
Fig. 1 shows the CNMSEs of estimates of degree distribution $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, where sampling budget $B = 0.001|V|$. Fig. 1 shows that the degree estimates produced by UNI and FS using neighbor information almost have the same accuracy. For FS, the degree distribution estimate greatly improves when neighbor information is used, which is almost *twice* as accurate than previous FS without using neighbor information for Xiami. [27, 29] show that NMSEs are roughly proportional to $1/\sqrt{B}$. It indicates that FS using neighbor information is *four times* more time efficient than the previous FS, which is consistent with our results shown in Fig. 2. For UNI method, the degree distribution estimator based on using the neighbor information of sampled nodes exhibits larger errors than the estimator given by sampled nodes for small degrees (degrees smaller than 20 for Xiami, 30 for YouTube). For the degree distribution estimator given by neighbors of sampled nodes, we can see that FS is more accurate than UNI for most degrees.

For directed graphs, Fig. 3 shows results for the in-degree distribution estimates. When in-degrees and out-degrees of sampled nodes are available, the in-degree distribution estimator given by neighbors of sampled nodes *outperforms* the estimator given by sampled nodes for FS method. For small in-degrees (3 for Xiami, 18 for YouTube), the in-degree distribution estimator given by neighbors of sampled nodes exhibits larger errors than the estimator given by sampled nodes for UNI method. Meanwhile, the results show that we can also give an accurate in-degree distribution estimate given by out-going neighbors of sampled nodes, which is a little less accurate than the estimate obtained by all neighbors' information. Fig. 4 shows the results of the mixture estimator in (5). We observe that the mixture estimator *outperforms* the estimator based on sampled nodes and the estimator based on neighbors of sampled nodes. Let c denote the cost of UNI, the average number of IDs queried until one valid ID is obtained. For example, Flickr has a random node sampling cost of $c = 77$ [29]. Here we set the cost of crawling methods FS and RW as 1. Next we compare with performance of crawling methods with social sampling (SS), a node sampling method proposed by Dasgupta et al. [7]. Here SS is equivalent to the estimator given by neighbors of nodes sampled by UNI. Fig. 5 shows that SS exhibits larger errors as c increases. When sampling cost $c = 10$, FS and RW are much more accurate than SS under the same sampling budget. Meanwhile we can see that FS exhibits smaller errors than RW.



(a) Xiami (b) YouTube

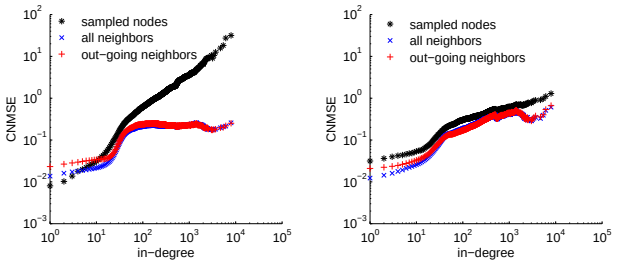
Figure 1: Results of degree distribution estimations for undirected graphs, $B = 0.001|V|$.



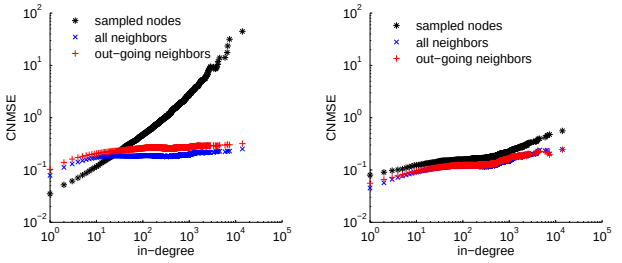
(a) Xiami (b) YouTube

Figure 2: To achieve the same MSE, regular FS requires at least $4\times$ the number of the samples of FS with neighbor information.

7.2 Edge Label Density Estimation

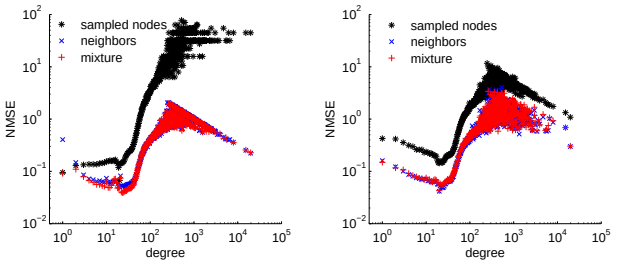


(a) Xiami, UNI (b) Xiami, FS

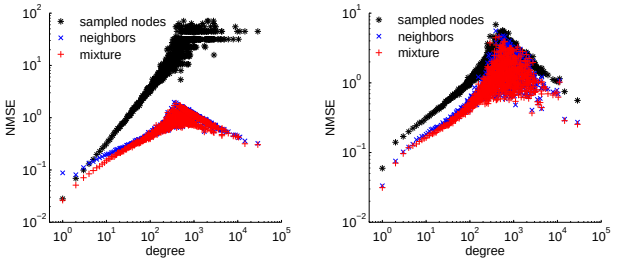


(c) YouTube, UNI (d) YouTube, FS

Figure 3: Results of in-degree distribution estimations for directed graphs, $B = 0.001|V|$.



(a) Xiami, UNI (b) Xiami, RW



(c) YouTube, UNI (d) YouTube, RW

Figure 4: Results of degree distribution estimations for the mixture estimator, $B = 0.001|V|$.

We evaluate the performance of our methods for estimating the joint degree distribution $\boldsymbol{\phi} = (\phi(i,j) : i \geq j > 0)$ for undirected graph G, where $\phi(i,j)$ is the fraction of edges consisting of two nodes with degree i and j separately. For two-dimensional distribution $\boldsymbol{\phi}$, we define δ as

$$\delta = \sqrt{\sum_{i \geq j > 0} (\hat{\phi}(i,j) - \phi(i,j))^2} \ ,$$

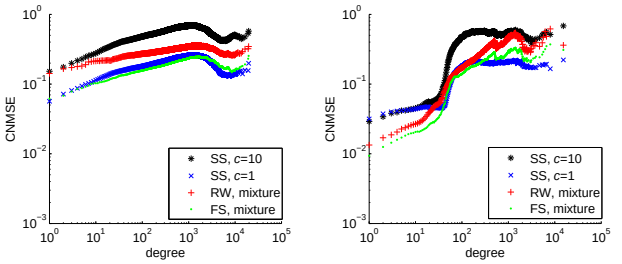
which is a metric that measures the error of its estimate $\hat{\boldsymbol{\phi}}$.

Fig. 6 shows the complementary cumulative distribution function (CCDF) of δ for 1,000 independent estimates, where the sampling budget is $B = 0.001|V|$. It shows that RW and UNI using sampled nodes' neighborhood information are more accurate. All estimates have errors larger than 0.1 when we have no knowledge of sampled nodes' degrees. More than 85% of estimates have errors smaller than 0.1 when sampled nodes' degrees are available.

Let us illustrate how to apply the edge label density estimation. Consider the directed graph of Xiami, 53.8% of users are male (M), 37.5% are female (F), and 8.7% are unknown (U). A directed edge (u, v) can be classified into the following nine types when the edge label is defined as $u.gender \rightarrow v.gender$: 1) M→M, 2) M→F, 3) M→U, 4) F→M, 5) F→F, 6) F→U, 7) U→M, 8) U→F, 9) U→U. Fig. 7 shows edge density $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_9)$, where τ_i $(1 \leq i \leq 9)$ is the fraction of type i edges. We can easily find that the fraction of edges with a certain edge type approximately equals the product of the fractions of nodes with its two endpoints' genders. This indicates that users' following behaviors in Xiami are not directly related with gender. Fig. 8 shows results for estimating $\boldsymbol{\tau}$. Similarly we can find that RW and UNI using sampled nodes' neighborhood information exhibits small errors, and are two times more accurate than the simple RW method.

7.3 High Degree Node Detection

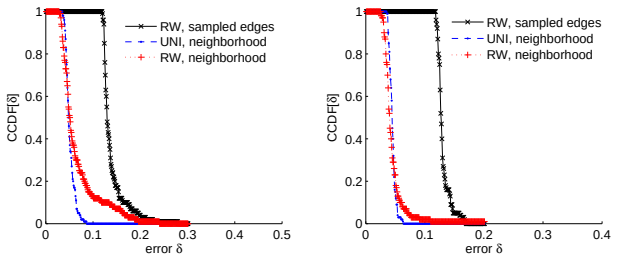
Fig.9 shows the results of previous methods for detecting top-100 high degree nodes, where the edge weight function is defined



(a) Flickr (b) Xiami

Figure 5: Results of degree distribution estimations for different node sampling cost c, $B = 0.001|V|$.

as $w(u,v) = (d_u d_v)^\beta$ for WRW. For previous methods without free neighborhood information of sampled nodes, we assume that XS and WRW both must retrieve degree information of a neighbor of sampled nodes with the same cost of sampling a node. Fig. 9 shows that all of RW, WRW, and XS need to sample more than 10% of nodes to obtain an accurate result for detecting top-100 degree nodes with the largest degrees. Fig. 10 shows the results of RW, WRW, and MXS using free neighborhood information of sampled nodes. A total of 1,000 runs are used to produce the averages seen in the graph. It shows that RW, WRW, MXS using neighborhood information are much more efficient than previous methods, and MXS outperforms RW and WRW. We observe that MXS detects almost 90% of the top-100 high degree nodes based on a very small fraction of sampled nodes, $B = 10^{-5} \times |V|$. Meanwhile, we compare MXS and XS based on the assumption that XS can be implemented at no cost of looking up sampled nodes' neighbor's neighbor information, and find they has little difference. We omit the details here. Similarly Figs. 11 and 12 show that MXS is much more efficient than the other two methods for detecting top-100 high out-degree nodes and top-100 high in-degree nodes for directed graphs, where each node has the knowledge of its neighbors'



(a) Soc-Epinions (b) Soc-Slashdot

Figure 6: CCDF of errors of joint degree distribution estimates, $B = 0.001|V|$.

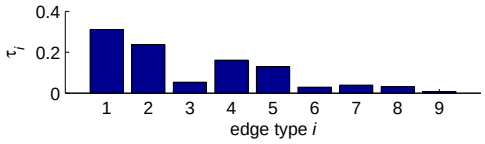
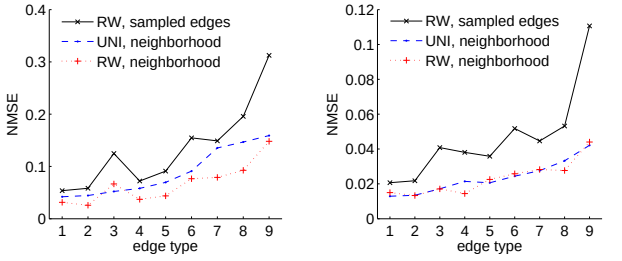


Figure 7: (Xiami) Density of edges with different types. Type 1: M→M, 2: M→F, 3: M→U, 4: F→M, 5: F→F, 6: F→U, 7: U→M, 8: U→F, 9: U→U.

out-degrees and in-degrees. The edge weight function of WRW is defined as $w(u,v) = (d_u^{(O)} + d_u^{(I)})^\beta (d_v^{(O)} + d_v^{(I)})^\beta$.

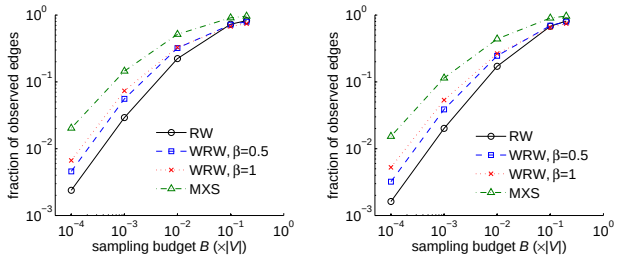
7.4 Short Path Discovery

Fig. 13 shows that MXS observes many more edges than WRW and RW under the same sampling budget B, where edge weight function is defined as $w(u,v) = (d_u d_v)^\beta$ for WRW. Note that when $\beta = 0$, WRW is the same as RW. As β increases, we can see that WRW collects more edges. In what follows we evaluate our MXS and WRW based short path discovery methods compared with the previous RW based method in [26]. For two nodes with distance $d < \infty$ in G, let d^* be the length of the short path observed by sampling methods. When there is no path observed for them, we denote $d^* = \infty$, and a failure is reported. For all $d^* < \infty$, we use $E[d^* - d]$ as a metric to measure the performance of detecting the shortest paths. Figs. 15-16 show results for 10,000 node pairs generated randomly, where the sampling budget is set as $B = 20$. Fig. 14 shows the fractions of sampled node pairs with given distances (length of shortest pathes in original graphes) for Soc-Slashdot and Soc-Epinions. Fig. 15 shows the fractions of short path discovery failures as a function of the distance. Y axis shows the fraction of failures for node pairs with a given distance. We can see that RW and WRW generate a large number of failures especially for node pairs with a long distance. However there is almost no failure for our new method MXS. Moreover Fig. 16 shows that MXS and WRW usually discover shorter paths in comparison with RW.



(a) $B = 0.001|V|$ (b) $B = 0.01|V|$

Figure 8: (Xiami) NMSE of edge gender type density estimates.



(a) Soc-Epinions (b) Soc-Slashdot

Figure 13: Fractions of observed edges.

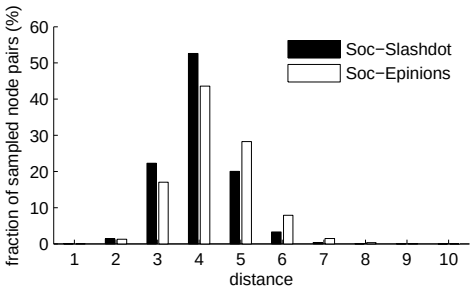
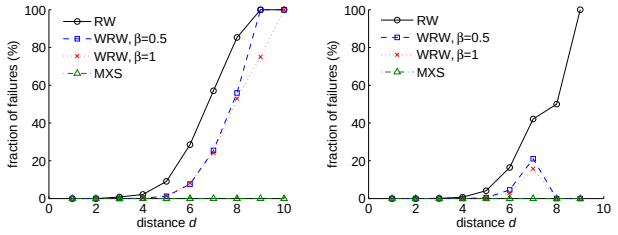


Figure 14: (Soc-Slashdot and Soc-Epinions) Fractions of sampled node pairs with given distances.

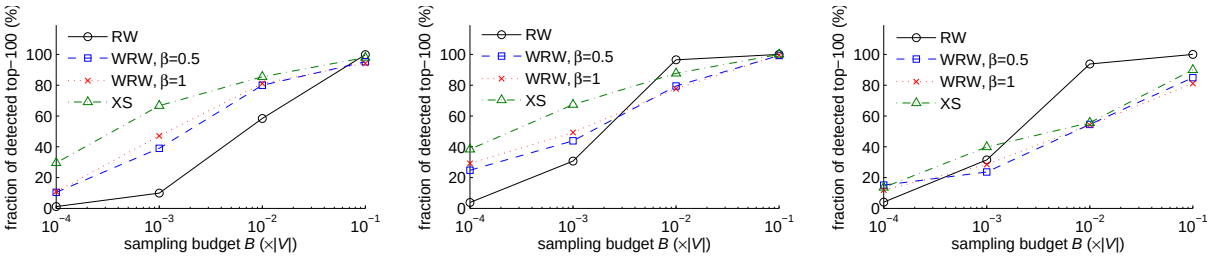


(a) Soc-Epinions (b) Soc-Slashdot

Figure 15: Fractions short path discovery failures, $B = 20$.

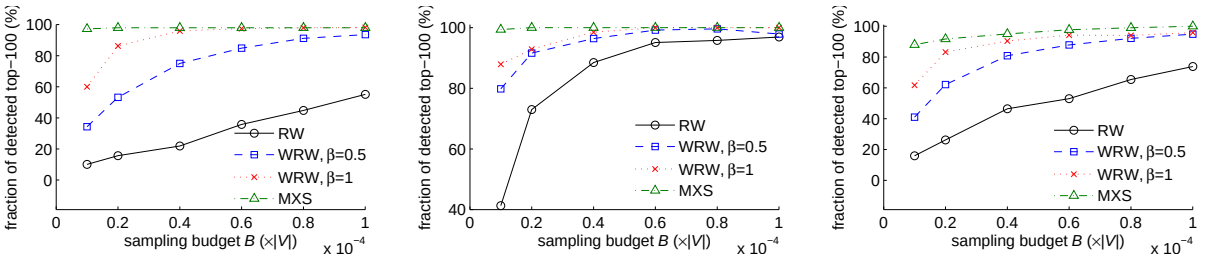
8. APPLICATIONS

Foursquare is a location-based OSN, which provides web and mobile services for users to explore interested places and leave tips or comments, and share their check in histories to their friends. As of December 2012, it has over 25 million active users [1]. In Foursquare when we visit a node, we can also obtain its friends' locations (living places) and degrees. In what follows, we use our methods which take advantage of this neighborhood information to characterize the Foursquare graph topology. Base on 1.9×10^5 users we sampled and their neighborhood information, we estimate the node degree distribution, joint degree distribution, and location



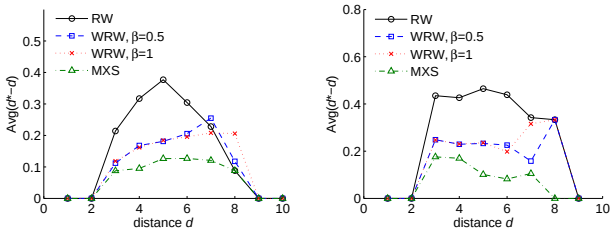
(a) Xiami (b) Flickr (c) YouTube

Figure 9: (Previous methods) Results of top-100 high degree node detection.



(a) Xiami (b) Flickr (c) YouTube

Figure 10: (Our methods, using neighbor information of sampled nodes) Results of top-100 high degree node detection.



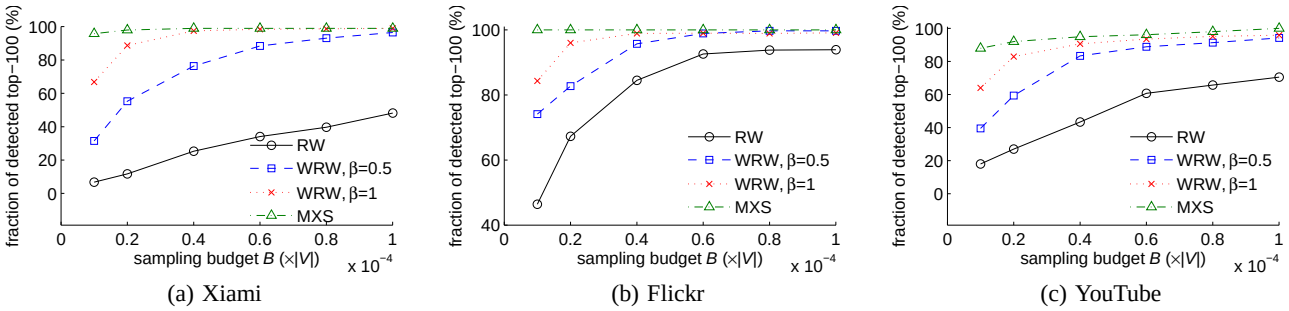
(a) Soc-Epinions (b) Soc-Slashdot

Figure 16: NMSE of short path lengthes, $B = 20$.

distribution. To obtain the optimal parameter α_k $(1 \leq k \leq K)$ of our mixture estimator in (5), we first split sampled nodes into 100 subsets with the same size. The variance of $\hat{\theta}_k$ in Equation (3), the estimator using sampled nodes, is computed based on its estimations obtained from 100 node subsets. Similarly, we estimate the variance of $\breve{\theta}_k$ in Equation (4), the estimator only using neighborhood information. We then set $\alpha_k = \frac{\mathrm{Var}(\breve{\theta}_k)}{\mathrm{Var}(\hat{\theta}_k)+\mathrm{Var}(\breve{\theta}_k)}$. Fig. 17 (a) shows results of estimating degree distribution using our mixture estimator based on all sampled nodes. Average degree is estimated as 21.2. We can see that the degree distribution of Foursquare *does not* exhibit a heavy tail, which has a sharp drop starting from degree 1,000. This may be caused by the policy set by Foursqure which limits users to have only 1,000 friends [9]. From our observed edges, we find that 53.8% of edges have no node with degree larger than 100. Fig. 17 (c) show results for estimating joint degree distribution $\boldsymbol{\phi} = (\phi(i,j) : i \geq j > 0)$, where $\phi(i,j)$ is the fraction of edges consisting of two nodes with degree i and j separately. This result is quite interesting for it shows that friends tend to have similar degrees. Fig. 17 (b) shows Foursquare users' location distribution. Note that users could provide any text strings to Foursquare as their locations. This induces different granularity of users' locations. For example, users from Katsushika-ku in Tokyo, Japan may reveal their locations as "Katsushika-ku, Tokyo, Japan", "Katsushika-ku, Tokyo", or "Tokyo, Japan". For simplicity, we split sampled users' location strings by comma, and classify two locations into a same group when they have at least one similar substring. Hence, there location strings in the above example will be clustered together and labeled by the most frequency part, say "Tokyo". Due to the limited space, we only show results for top 20 popular locations. We can see that the most popular location in Foursquare is New York state (NY), which accounts for nearly 10% of uses. The second popular location is Indonesia, which accounts for 8%.

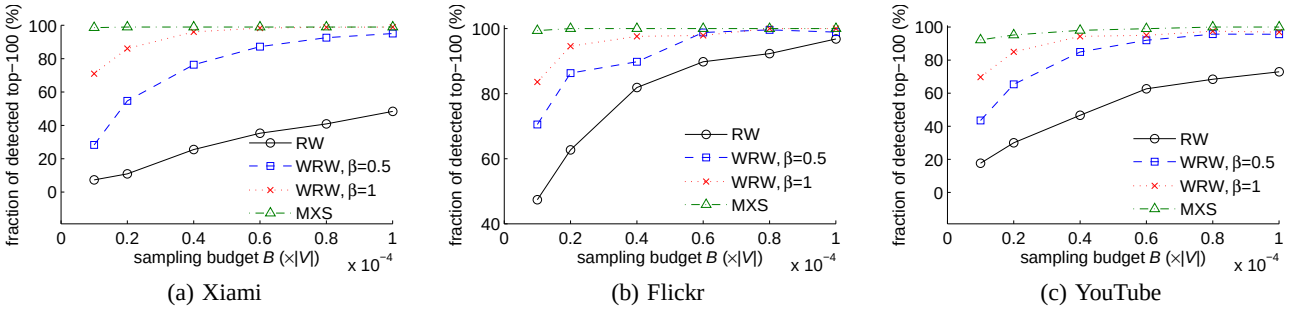
We randomly sample 20 nodes with degrees not smaller than one, and apply a MXS starting from each random node, where the sampling budget B is 1,000. For the top 100 nodes with largest degrees in all sampled nodes and their neighbors given by 20 MXSes, we show their frequencies detected by MXSes in Fig. 17 (d), where the y-axis is defined as the fraction of MXSes successfully detected the i-th high degree node, $1 \leq i \leq 100$. We can see that most of these high degree nodes are detected by 90% of MXSes, which indicates that they may be very close to the ground truth. These high degree nodes have degrees in the range [1061, 1083], which are shown in Fig. 17(e). The top three high degrees are 1396, 1200, and 1185. Fig. 17(f) shows the length distribution of discovered short pathes between 20 initially sampled nodes. Pathes for all 190 node pairs are successfully discovered. Pathes with lengths 5, 6, and 7 account for 93%. The average length is 5.8.

Pinterest is a photo sharing OSN, which allows users to create boards (theme-based image collections) and pin/repin (collect) images onto their boards from other Pinterest users' image boards and external websites. Similar to Twitter, Pinterest users can follow other users if they have similar tastes. In Pinterest when we visit a node we can also obtain its friends' in-degree (number of followers) and out-degree (number of following). We collected 10,000 users using a RW. The maximum in-degree and out-degree we discovered is 13,331,207, and 61,338. The CCDFs of in-degree and



(a) Xiami (b) Flickr (c) YouTube

Figure 11: (Our methods, using neighborhood information of sampled nodes) Results of top-100 high out-degree node detection.



(a) Xiami (b) Flickr (c) YouTube

Figure 12: (Our methods, using neighborhood information of sampled nodes) Results of top-100 high in-degree node detection.

out-degree distributions are shown as Fig. 18.

9. RELATED WORK

Few network sampling methods use neighborhood information to provide accurate estimates that have convergence guarantees. The work closest to ours is Dasgupta et al. [7]. Dasgupta et al. randomly samples nodes (either uniformly or with a known bias) and then uses neighborhood information to improve its unbiased estimator. However, randomly sampling nodes is practical only if performed uniformly (in our scenarios, rejection sampling to bias the samples makes little sense) and suffers from low query rate in NoSQL graph databases and Web APIs. Dasgupta et al. partially compensates the low query rate through the use of neighborhood information present in the node query reply of a number of major OSNs. Moreover, their estimators require knowledge of degrees of sampled nodes' neighbors, which incurs extra query costs when applied to OSNs such as Pinterest and Sina microblog that do not provide free neighbor degree information.

Kurant et al. [14] designs a RW-based method that uses a weighted RW to perform stratified sampling on social networks. These weights are computed using neighborhood information. Kurant et al. uses their technique on Facebook and show that their stratified sampling technique achieves higher estimation accuracy than other methods. However, the neighborhood information in their method is limited to helping find random walk weights and not used in the estimator. Interestingly, our estimator can be easily combined with the weighted random walk in [14] to improve its accuracy.

Maiya and Berger-Wolf [21] empirically investigates the performance of a number of subgraph sampling methods (e.g., breadth-first search, random walks, etc.) and their performance in respect to various topological properties (e.g., degree, clustering coefficient). Maiya and Berger-Wolf, however, does not use neighborhood information to improve the estimators or provide convergence guarantees. The literature also shows a variety of subgraph sampling works without convergence or accuracy guarantees [12, 17], which have been empirically tested over a variety of networks. The above works [12, 17, 21] also consider subgraph sampling techniques that can preserve other metrics, such as the eigenvalues of the original network [17], but without accuracy guarantees.

Breadth-First-Search (BFS) introduces a large bias towards high degree nodes, and it is difficult to remove these biases in general, although it can be ameliorated if the network in question is almost random [15]. Random walk (RW) is biased to sample high degree nodes, however its bias is known and can be easily corrected [27]. Random walks in the form of Respondent Driven Sampling (RDS) [11, 31] has been used to estimate population densities using snowball samples of sociological studies. RDS was developed for small social networks with hidden links while our method considers large online social networks without hidden links.

The Metropolis-Hasting RW (MHRW) [33] modifies the RW procedure, aimed at sampling nodes with equal probability. However, in Ribeiro and Towsley [28] we prove that MHRW degree distribution estimates perform poorly in comparison to RWs, more markedly for large degree nodes whose error grows proportionally to the degree value. Empirically, the accuracy of RW and MHRW has beeen compared in [10, 25] and, as predicted by our theoretical results, RW is consistently more accurate than MHRW.

10. CONCLUSIONS AND FUTURE WORK

In this paper, we study the problem of estimating characteristics for graphs where nodes have knowledge of their neighbors' properties. This feature is actually quite common in many OSNs, such as Pinterest [24], Foursquare [8], Sina microblog [32], and Xiami [35]. We propose efficient network characteristic (degree and edge density distributions) estimators from sampling which have show provable convergence and accuracy guarantees. Our method is tailored to NoSQL graph databases (e.g. Neo4j) and the type of Web API present in major social network websites such as Face-

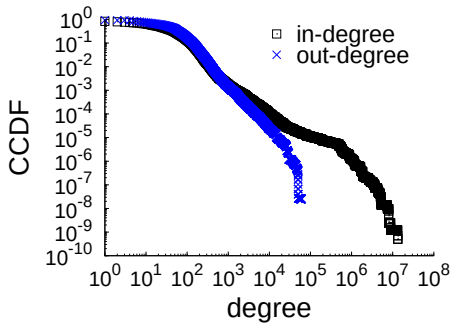
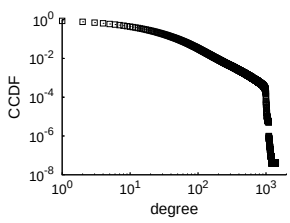
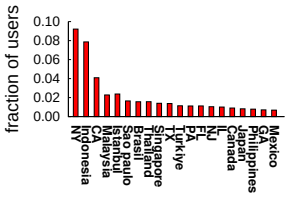


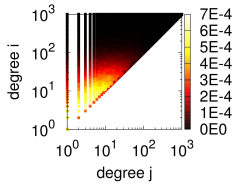
Figure 18: (Pinterest) CCDFs of estimated in-degree and out-degree distributions. Estimated average in-degree and out-degree are 60.2 and 61.1, respectively.



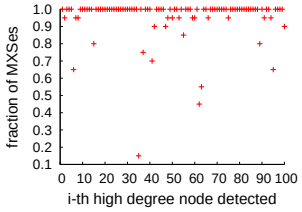
(a) Degree distribution



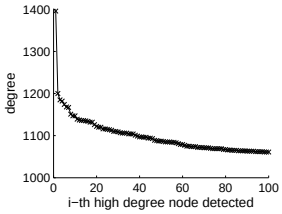
(b) Locations distribution



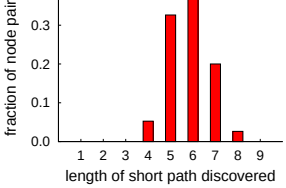
(c) Joint degree distribution $\phi = (\phi(i,j) : i \geq j > 0)$



(d) High degree node detection



(e) Degrees of top-100 high degree nodes



(f) Short path discovery

Figure 17: Foursquare structure statistics.

book, Google+, Twitter, Pinterest [24], and Foursquare [8]. We can also adapt known techniques to detect high-degree nodes and short path discovery between nodes . Our experimental results show that our estimator drastically reduces (by 4-fold) the number of samples required to achieve the same estimation accuracy. Our generalization allows us to include neighboring information in the estimation of a variety of network characteristics from nodes sampled using a random walk-based technique called Frontier Sampling [27]. As future work, we plan to replace Lemma 2 with a bound that, for $m > 1$, considers all samples of FS.

Acknowledgments

This work was supported by the NSF grant CNS-1065133 and ARL Cooperative Agreement W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied of the NSF, ARL, or the U.S. Government. This work was also supported in part by the NSFC funding 60921003 and 863 Program 2012AA011003 of China.

11. REFERENCES

[1] `https://foursquare.com/about/`.

[2] Online estimating the k central nodes of a network. In *Proceedings of IEEE Network Science Workshop 2011*, June 2011.

[3] N. K. Ahmed, J. Neville, and R. Kompella. Network sampling: From static to streaming graphs. *arXiv preprint arXiv:1211.3412*, 2012.

[4] K. Avrachenkov, B. Ribeiro, and D. Towsley. Improving random walk estimation accuracy with uniform restarts. In *The 7th Workshop on Algorithms and Models for the Web Graph*, pages 98–109, December 2010.

[5] C. Cooper, T. Radzik, and Y. Siantos. A fast algorithm to find all high degree vertices in power law graphs. In *Proceedings of WWW 2012 LSNA workshop*, pages 1007–1016, April 2012.

[6] D. Coppersmith, P. Doyle, P. Raghavan, and M. Snir. Random walks on weighted graphs, and applications to on-line algorithms (extended). *Journal of the ACM*, pages 369–378, 1993.

[7] A. Dasgupta, R. Kumar, and D. Sivakumar. Social sampling. In *Proceedings of ACM SIGKDD 2012*, pages 235–243, August 2012.

[8] Foursquare. `http://www.foursquare.com`, 2012.

[9] Foursquare limit. `http://www.quora.com/Foursquare/What-are-some-of-th` 2012.

[10] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in facebook: A case study of unbiased sampling of OSNs. In *Proceedings of IEEE INFOCOM 2010*, pages 2498–2506, April 2010.

[11] D. D. Heckathorn. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49(1):11–34, 2002.

[12] C. Hubler, H.-P. Kriegel, K. Borgwardt, and Z. Ghahramani. Metropolis algorithms for representative subgraph sampling. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 283–292. IEEE, 2008.

[13] P. Hui, J. Crowcroft, and E. Yoneki. Bubble rap: Social-based forwarding in delay tolerant networks. In *Proceedings of ACM MobiHoc 2008*, pages 241–250, May 2008.

[14] M. Kurant, M. Gjoka, C. T. Butts, and A. Markopoulou. Walking on a graph with a magnifying glass: stratified sampling via weighted random walks. In *SIGMETRICS*, volume 39, pages 241–252, 2011.

[15] M. Kurant, A. Markopoulou, and P. Thiran. Towards unbiased bfs sampling. *IEEE Journal on Selected Areas in Communications*, 29(9):1799–1809, September 2011.

[16] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of WWW 2010*, pages 591–600, April 2010.

[17] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of ACM SIGKDD 2006*, pages 631–636, June 2006.

[18] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of WWW 2008*, pages 915–924, April 2008.

[19] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[20] A. S. Maiya and T. Y. Berger-Wolf. Online sampling of high centrality individuals in social networks. In *Proceedings of the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2010)*, pages 91–98, June 2010.

[21] A. S. Maiya and T. Y. Berger-Wolf. Benefits of bias: towards better characterization of network sampling. In *SIGKDD*, pages 105–113, 2011.

[22] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of ACM SIGCOMM Internet Measurement Conference 2007*, pages 29–42, October 2007.

[23] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. Technical report, Stanford InfoLab, 1999.

[24] Pinterest. `http://www.pinterest.com`, 2013.

[25] A. H. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach. Respondent-driven sampling for characterizing unstructured overlays. In *Proceedings of IEEE INFOCOM Mini-conference 2009*, April 2009.

[26] B. Ribeiro, P. Bash, and D. Towsley. Multiple random walks to uncover short paths in power law networks. In *Proceedings of IEEE Infocom NetSciCom Workshop 2012*, pages 1–6, April 2012.

[27] B. Ribeiro and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In *Proceedings of ACM SIGCOMM Internet Measurement Conference 2010*, pages 390–403, November 2010.

[28] B. Ribeiro and D. Towsley. On the estimation accuracy of degree distributions from graph sampling. In *IEEE Conference on Decision and Control*, 2012.

[29] B. Ribeiro, P. Wang, F. Murai, and D. Towsley. Sampling directed graphs with random walks. In *Proceedings of IEEE INFOCOM 2012*, pages 1692–1700, April 2012.

[30] M. Richardson, R. Agrawal, and P. Domingos. Trust management for the semantic web. In *Proceedings of the 2nd International Semantic Web Conference*, pages 351–368, October 2003.

[31] M. J. Salganik and D. D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34:193–239, 2004.

[32] Sina microblog. `http://weibo.com`, 2012.

[33] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Transactions on Networking*, 17(2):377–390, April 2009.

[34] P. Wang, J. Zhao, X. Guan, J. C. Lui, and D. Towsley. Sampling contents distributed over graphs. Technical Report TR-1201, Xi'an Jiaotong University, 2012.

[35] Xiami. `http://www.xiami.com`, 2012.