Sampling Node Pairs Over Large Graphs

Pinghui Wang [#1], Junzhou Zhao [*2], John C.S. Lui [#3], Don Towsley [+4], Xiaohong Guan [*$5]

[#]*Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong*
[1]phwang@sei.xjtu.edu.cn [3]cslui@cse.cuhk.edu.hk

[*]*MOE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an, China*
[2]jzzhao@sei.xjtu.edu.cn

[+]*Department of Computer Science, University of Massachusetts Amherst, MA, US*
[4]towsley@cs.umass.edu

[$]*Department of Automation and NLIST Lab, Tsinghua University, Beijing, China*
[5]xhguan@xjtu.edu.cn

***Abstract*— Characterizing user pair relationships is important for applications such as friend recommendation and interest targeting in online social networks (OSNs). Due to the large scale nature of such networks, it is infeasible to enumerate all user pairs and so sampling is used. In this paper, we show that it is a great challenge even for OSN service providers to characterize user pair relationships even when they posses the complete graph topology. The reason is that when sampling techniques (i.e., uniform vertex sampling (UVS) and random walk (RW)) are naively applied, they can introduce large biases, in particular, for estimating similarity distribution of user pairs with constraints such as existence of mutual neighbors, which is important for applications such as identifying network homophily. Estimating statistics of user pairs is more challenging in the absence of the complete topology information, since an unbiased sampling technique such as UVS is usually not allowed, and exploring the OSN graph topology is expensive. To address these challenges, we present asymptotically unbiased sampling methods to characterize user pair properties based on UVS and RW techniques respectively. We carry out an evaluation of our methods to show their accuracy and efficiency. Finally, we apply our methods to two Chinese OSNs, Doudan and Xiami, and discover significant homophily is present in these two networks.**

I. Introduction

Online social networks (OSNs) such as Facebook and Twitter have become extremely popular within the last few years. Billions of network users currently spend 22% of all their online time on OSNs on average, and this surpasses the average time spent on email [39]. Meanwhile, OSNs have greatly changed people's network activities. OSNs help people to keep in touch with old friends and meet new friends with common interests. They provide individuals online private spaces and multiple ways to interact using chat, messaging, email, video, voice chat, file sharing, blogging, discussion groups and so on. Characterizing user pair properties is of fundamental importance and has the following important applications

- *Network homophily detection.* Homophily refers to the tendency of users to connect to others with common interests. Singla et al. [34] show that significant homophily is present in the MSN Messenger network. That is, users who chat with each other are more likely to share interests in terms of their Web search topics, and personal characteristics such as their ages and locations. Similar findings hold for users who never talk to each other but do have at least one friend in common. For a user in these networks with significant homophily, we can infer her unstated (private) personal characteristics and give the user valuable recommendations based on her neighbors' characteristics and interests.
- *Distance distribution measurement.* The distance between two nodes A and B is measured by their shortest path length in an OSN. Characterizing the distance distribution measurement is necessary for calculating the average distance among pairs and the effective diameter (the 90th percentile of all distances), which are fundamental statistics for understanding the nature and evolution of the network. For example, the famous six degree of separation shows that any two people could be connected on average within six hops from each other [24], which indicates that human society is a small world type network.

In this paper, we design efficient methods to characterize node pairs in network. In particular, we not only characterize all pairs (contained in set $\mathbf{S}$) but also connected pairs (contained in set $\mathbf{S}^{(1)}$), and pairs that share a neighbor (contained in set $\mathbf{S}^{(2)}$), where set $\mathbf{S}$ consists of all node pairs in G, subset $\mathbf{S}^{(1)}$ consists of pairs of connected nodes, and subset $\mathbf{S}^{(2)}$ consists of pairs of nodes with at least one common neighbor. Methods for characterizing node pairs in three sets can be easily applied to problem such as measuring homophily or distance distribution measurement. For example, we can estimate the underlying distance distribution of G based on sampling random node pairs uniformly from $\mathbf{S}$. By comparing the interest similarity of user pairs in sets $\mathbf{S}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$, we can infer whether users are connected and clustered based on their common interests. Due to the large sizes of these networks, exhaustive enumeration of all node pairs is computational prohibitive. Existing sampling techniques such as uniform vertex sampling (UVS) and random walk (RW) cannot be directly applied. A naive application of sampling

techniques can generate large biases in estimated statistics. For example, one might propose the following approach for sampling a node pair $[u, v]$ from $\mathbf{S}^{(2)}$. It first samples a node x from graph G using UVS. Then u and v are set to two random neighbors of x. It is a simple way to sample two random nodes u and v with at least one neighbor. However in what follows we show this sampling method does not sample node pairs uniformly, and removing sampling bias is costly. Given that x is sampled, each pair of neighbors of x is selected with the same probability $\frac{2}{d_x(d_x-1)}$, where d_x is the number of its neighbors. Denote $\mathbf{M}(u, v)$ as the set of common neighbors of u and v. Then we find that node pair $[u, v]$ is sampled with probability proportional to $\sum_{x\in\mathbf{M}(u,v)} \frac{1}{d_x(d_x-1)}$. which is related not only with the number of common neighbor of u and v, but also with the degree of each common neighbor of u and v. We can easily find that it is costly to correct the bias for sampling pair $[u, v]$ since one needs to query nodes u, v and all common neighbors of u and v.

When UVS is not feasible (either because we do not have the full graph topology, or generation cost of random node is too expensive) and exploring the OSN graph topology is resource limited and expensive, it is much more challenging to estimate node pair statistics. To address the above issues, we systematically study the problem of sampling node pairs in a large graph, and present sampling methods for estimating characteristics of node pairs in sets $\mathbf{S}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$ that can be applied to settings where the graph topology may or may not be known. To sample node pairs in $\mathbf{S}^{(2)}$, we propose two methods: weighted vertex sampling (WVS) and neighborhood random walk (NRW), and develop corresponding unbiased estimators for measuring node pairs' statistics. Compared to WVS, NRW does not require the use of UVS, and it can be viewed as a regular RW over a new graph $\hat{G}$, where a node in $\hat{G}$ is an edge in original graph G, and an edge in $\hat{G}$ consists of two edges in G with a common node. Simulation results show that our methods are efficient. Finally we apply our methods to OSNs Doudan and Xiami, and find that users tend to connect to others with common interests.

This paper is organized as follows. The problem is formulated in Section II. Section III and Section IV present node pair sampling methods for ones with or without the complete graph topology respectively. Performance evaluation and testing results are presented in Section V. Section VI presents real applications on Xiami and Douban websites. Section VII summarizes related work. Section VIII concludes.

II. Problem Formulation

Let $G = (V, E)$ be an undirected graph, where V is the set of nodes and E the set of edges. G contains no self-loops. In what follows, (u, v) denotes an edge in G, and $[u, v]$ a node pair in G. Note that $(u, v) \neq (v, u)$ and $[u, v] \neq [v, u]$. We present sampling methods to measure characteristics of node pairs in the following sets:

- **whole set** $\mathbf{S} = \{[u, v] : u, v \in V \text{ and } u \neq v\}$
- **one-hop subset** $\mathbf{S}^{(1)} = \{[u, v] : (u, v) \in E\}$
- **two-hop subset** $\mathbf{S}^{(2)} = \{[u, v] : u \neq v, u, v \in V, \exists x \in V, (u, x) \in E \text{ and } (v, x) \in E\}$
- **one to two-hop subset** $\mathbf{S}^{(2+)} = \mathbf{S}^{(2)} \cup \mathbf{S}^{(1)}$.

We easily find that the intersection of $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ to be non-empty, and $\mathbf{S}^{(2)}$ may not contain each edge $(u, v) \in E$ since u and v need not have any mutual neighbors. $\mathbf{S}^{(1)}$ consists of all pairs of nodes whose distance is exactly one, and $\mathbf{S}^{(2+)}$ consists of all pairs of nodes with distance no greater than two. Define function $F : V \times V \to \mathbb{R}$. For node pair $[u, v]$, $F(u, v)$ defines the value of the pair's property under study, e.g., the number of mutual neighbors of u and v. Note that $F(u, v)$ needs not equal to $F(v, u)$, e.g., $F(u, v)$ is defined as the number of neighbors of u excluding the common neighbors of u and v. Let $\{a_1, \ldots, a_K\}$ be the range of $F(u, v)$. We propose sampling methods to estimate the node pair distributions $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$, $\boldsymbol{\omega}^{(1)} = (\omega_1^{(1)}, \ldots, \omega_K^{(1)})$, $\boldsymbol{\omega}^{(2)} = (\omega_1^{(2)}, \ldots, \omega_K^{(2)})$, and $\boldsymbol{\omega}^{(2+)} = (\omega_1^{(2+)}, \ldots, \omega_K^{(2+)})$, where ω_k, $\omega_k^{(1)}$, $\omega_k^{(2)}$, and $\omega_k^{(2+)}$ $(1 \leq k \leq K)$ are the fractions of pairs $[u, v]$ in sets $\mathbf{S}$, $\mathbf{S}^{(1)}$, $\mathbf{S}^{(2)}$, and $\mathbf{S}^{(2+)}$ respectively with $F(u, v) = a_k$. Define $\mathbf{S}^{(1-)} = \mathbf{S}^{(1)} \backslash \mathbf{S}^{(2)}$. For each element $[u, v] \in \mathbf{S}^{(1-)}$, u and v are connected but do not have any mutual neighbor. Similarly define $\boldsymbol{\omega}^{(1-)} = (\omega_1^{(1-)}, \ldots, \omega_K^{(1-)})$, where $\omega_k^{(1-)}$ $(1 \leq k \leq K)$ is the fraction of pairs $[u, v]$ with $F(u, v) = a_k$ in set $\mathbf{S}^{(1-)}$. Let $\alpha = \frac{|\mathbf{S}^{(1-)}|}{|\mathbf{S}^{(1)}|}$ and $\beta = \frac{|\mathbf{S}^{(1)}|}{|\mathbf{S}^{(2)}|}$. Then we have

$$\omega_k^{(2+)} = \frac{|\mathbf{S}^{(1-)}|\omega_k^{(1-)} + |\mathbf{S}^{(2)}|\omega_k^{(2)}}{|\mathbf{S}^{(1-)}| + |\mathbf{S}^{(2)}|} = \frac{\alpha\beta\omega_k^{(1-)} + \omega_k^{(2)}}{\alpha\beta + 1}. \quad (1)$$

This $\boldsymbol{\omega}^{(2+)}$ can be obtained from α, β, $\omega_k^{(1-)}$, and $\omega_k^{(2)}$, where α and $\omega_k^{(1-)}$ can be calculated based on node pairs in set $\mathbf{S}^{(1)}$, and β and $\omega_k^{(2)}$ can be calculated based on node pairs in set $\mathbf{S}^{(2)}$. Since $\omega_k^{(2+)}$ is very close to $\omega_k^{(2)}$ for most OSN graphs with very small α and β, therefore we focus on designing methods for estimating characteristics of node pairs in $\mathbf{S}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$ in the following sections.

Since $|\mathbf{S}^{(1)}| = 2|E|$, $|\mathbf{S}| = |V|(|V| - 1)$ and $|\mathbf{S}^{(2)}|$ is usually much larger than $|V|$, sampling is unavoidable for estimating $\boldsymbol{\omega}$ and $\boldsymbol{\omega}^{(2+)}$ even for a moderate size graph with several hundred thousands of nodes. In the following two sections, we propose sampling methods based on two common sampling techniques: *Uniform Vertex Sampling* (UVS) and *Random Walk* (RW) respectively.

III. Node Pair Sampling Methods Based on UVS

In this section, we present sampling methods based on UVS to estimate node pair characteristics for $\mathbf{S}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$.

A. Basic Sampling Operations and Their Cost

Suppose that we can sample nodes from graph G using UVS with replacement. For example, for the Xiami OSN, there is a numeric ID associated with each node. The ID values of nodes are sequentially assigned. Then one can perform UVS by sample IDs randomly from the ID space with replacement. We assume this computation complexity is $O(1)$.

In what follows, we present methods for sampling nodes from V with any desired stationary distribution $\boldsymbol{\pi}=(\pi_v : v \in V)$, which is important for sampling node pairs as we will show later. First, we present an *independent weighted vertex sampling* (IWVS) method. Denote I_v as the ID of node v. Then we assign a weight to each node v as follows

$$W[I_v] = \sum_{u \in V \text{ and } I_u \leq I_v} \pi_u.$$

Let $\mathbf{v}_i$ be the node with ID $i \in \{1, \ldots, |V|\}$. At each step, we generate a random number τ drawn uniformly from range (0,1). Then sample node $\mathbf{v}_i$ whose ID i satisfies $W[i] \leq \tau < W[i+1]$. $\mathbf{v}_i$ can be efficiently identified using binary search, and its computational complexity is $O(\log |V|)$.

Note that when π_v depends on the graph topology, say the degree of v, we need the complete graph toplogy in advance to build the vector W. Often, the complete graph topology is not be available. Therefore, we propose a way to modify UVS using the Metropolis-Hasting technique [4], [9], [23]. This method does not require the complete graph topology, and reduces the memory space used for storing array W and extra computation for looking up the ID of a sampled node at each step. UVS can be modeled as a Markov chain with transition matrix $P = [P_{u,v}]$, $u, v \in V$, where $P_{u,v} = \frac{1}{|V|}$ is defined as the probability that node v is selected as the next sampled node given that the current node sampled is u. To generate a sequence of random samples from a desired stationary distribution $\boldsymbol{\pi}$, the Metropolis-Hastings technique is a Markov chain Monte Carlo method based on modifying the transition matrix of UVS as

$$P^{\star}_{u,v} = \begin{cases} P_{u,v} \min\left(\frac{\pi_v P_{v,u}}{\pi_u P_{u,v}}, 1\right) & \text{if } v \neq u, \\ 1 - \sum_{w \neq u} P^{\star}_{u,w} & \text{if } v = u. \end{cases} \tag{2}$$

It provides a way to alter the next node selection to produce any desired stationary distribution $\boldsymbol{\pi}$. Metropolis-Hastings based weighted vertex sampling (MHWVS) with target distribution $\boldsymbol{\pi}$ works as follows: at each step, MHWVS selects a node v using UVS and then accepts the move with probability $\min\left(\frac{\pi_v}{\pi_u}, 1\right)$. Otherwise, MHWVS remains at u. The computational complexity of sampling a node by MHWVS is $O(1)$.

B. Sampling Node Pairs From $\mathbf{S}$ *and* $\mathbf{S}^{(1)}$

To sample a node pair $[u, v]$ randomly from $\mathbf{S}$, we can use UVS to select two different nodes u and v randomly. $\mathbf{1}(\mathbb{P})$ defines the indicator function that equals one when predicate $\mathbb{P}$ is true, and zero otherwise. Based on sampled pairs $\{[u_i, v_i]\}_{i=1,\ldots,n}$, the fraction ω_k $(1 \leq k \leq K)$ is estimated as follows

$$\hat{\omega}_k = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(F(u_i, v_i) = a_k). \tag{3}$$

Each node pair $[u_i, v_i]$ is sampled with the same probability $\frac{1}{|V|(|V|-1)}$, the expectation of $\mathbf{1}(F(u_i, v_i) = a_k)$ is

$$E\left[\mathbf{1}(F(u_i, v_i) = a_k)\right] = \sum_{[u,v] \in \mathbf{S}} \frac{\mathbf{1}(F(u, v) = a_k)}{|V|(|V|-1)} = \omega_k,$$

and the variance is

$$Var\left[\mathbf{1}(F(u_i, v_i) = a_k)\right] = \sum_{[u,v] \in \mathbf{S}} \frac{\mathbf{1}^2(F(u, v) = a_k)}{|V|(|V|-1)} - \omega_k^2 = \omega_k - \omega_k^2.$$

Then we have

$$E[\hat{\omega}_k] = \omega_k, \quad \text{and} \quad Var[\hat{\omega}_k] = \frac{\omega_k - \omega_k^2}{n}.$$

Denote by d_x the degree of node $x \in V$. To sample node pairs from $\mathbf{S}^{(1)}$, we randomly select a node u according to probability distribution $(\pi_x^{(1)} : x \in V)$ using vertex sampling method IWVS or MHWVS, where $\pi_x^{(1)}$ is defined as

$$\pi_x^{(1)} = \frac{d_x}{2|E|}.$$

Then select a neighbor v at random. It is easy to see that node pair $[u, v]$ in $\mathbf{S}^{(1)}$ is sampled uniformly. Based on sampled pairs $\{[u_i, v_i]\}_{i=1,\ldots,n}$, we estimate $\omega_k^{(1)}$ $(1 \leq k \leq K)$ as follows

$$\hat{\omega}_k^{(1)} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(F(u_i, v_i) = a_k). \tag{4}$$

When IWVS is used to sample nodes, we can show that each (u_i, v_i) $i = 1, \ldots, n$ is an edge sampled uniformly and independently from graph G. Similar to the derivation of $\hat{\omega}_k$, we have

$$E[\hat{\omega}_k^{(1)}] = \omega_k^{(1)}, \quad \text{and} \quad Var[\hat{\omega}_k^{(1)}] = \frac{\omega_k^{(1)} - (\omega_k^{(1)})^2}{n}.$$

C. Sampling Node Pairs From $\mathbf{S}^{(2)}$

To sample node pair randomly from $\mathbf{S}^{(2)}$, we first randomly select a node $x \in V$ with degree greater than two according to probability distribution $(\pi_x^{(2)} : x \in V)$, where $\pi_x^{(2)}$ is defined as

$$\pi_x^{(2)} = \frac{d_x(d_x - 1)}{M},$$

where $M = \sum_{y \in V} d_y(d_y - 1)$. Then we generate a node pair $[u, v]$ by sampling two different neighbors u and v of x at random. There are $d_x(d_x - 1)$ node pairs consisting of two different neighbors of x, therefore each one of these node pairs is sampled with the same probability $\frac{1}{M}$. Define $m(u, v)$ to be the number of mutual neighbors of u and v. Then a node pair $[u, v]$ in $\mathbf{S}^{(2)}$ is sampled with probability

$$\pi_{[u,v]}^{(2)} = \frac{m(u, v)}{M}. \tag{5}$$

Based on sampled pairs $\{[u_i, v_i]\}_{i=1,\ldots,n}$, we estimate $\omega_k^{(2)}$ $(1 \leq k \leq K)$ as follows

$$\hat{\omega}_k^{(2)} = \frac{1}{H} \sum_{i=1}^{n} \frac{\mathbf{1}(F(u_i, v_i) = a_k)}{m(u_i, v_i)}, \tag{6}$$

where $H = \sum_{i=1}^{n} \frac{1}{m(u_i, v_i)}$. Let $\bar{m} = \frac{M}{|\mathbf{S}^{(2)}|}$ denote the average number of mutual neighbors of node pairs in $\mathbf{S}^{(2)}$. The accuracy of $\hat{\omega}_k^{(2)}$ can be stated by the following theorem.

Theorem *1:* $\hat{\omega}_k^{(2)}$ $(k = 1, \ldots, K)$ is an asymptotically unbiased estimator of $\omega_k^{(2)}$. When $\{[u_i, v_i]\}_{i=1,\ldots,n}$ are sampled independently using IWVS, we have

$$P\left(|\hat{\omega}_k^{(2)} - \omega_k^{(2)}| \leq \frac{2\epsilon\omega_k^{(2)}}{1-\epsilon}\right) > 1 - \frac{1}{n\epsilon^2}\left(\frac{\bar{m}}{\omega_k^{(2)}} + \bar{m} - 2\right)$$

where $0 < \epsilon < 1$.

Proof. We have the following equation for each $i = 1, \ldots, n$, and $k = 1, \ldots, K$

$$\begin{aligned} E\left[\frac{\mathbf{1}(F(u_i, v_i) = a_k)}{m(u_i, v_i)}\right] &= \sum_{[u,v]\in\mathbf{S}^{(2)}} \pi_{[u,v]}^{(2)} \times \frac{\mathbf{1}(F(u,v) = a_k)}{m(u,v)} \\ &= \sum_{[u,v]\in\mathbf{S}^{(2)}} \frac{\mathbf{1}(F(u,v) = a_k)}{M} \\ &= \frac{\omega_k^{(2)}}{\bar{m}}. \end{aligned} \tag{7}$$

The second equation holds because (5), and the last equation holds because $\sum_{[u,v]\in\mathbf{S}^{(2)}} \mathbf{1}(F(u,v) = a_k) = |\mathbf{S}^{(2)}|\omega_k^{(2)}$, and $\bar{m} = \frac{M}{|\mathbf{S}^{(2)}|}$. Meanwhile,

$$\begin{aligned} &Var\left[\frac{\mathbf{1}(F(u_i, v_i) = a_k)}{m(u_i, v_i)}\right] \\ &= \sum_{[u,v]\in\mathbf{S}^{(2)}} \pi_{u,v}^{(2)} \times \frac{\mathbf{1}(F(u,v) = a_k)}{m^2(u,v)} - \frac{(\omega_k^{(2)})^2}{\bar{m}^2} \\ &= \sum_{[u,v]\in\mathbf{S}^{(2)}} \frac{\mathbf{1}(F(u,v) = a_k)}{Mm(u,v)} - \frac{(\omega_k^{(2)})^2}{\bar{m}^2}. \end{aligned} \tag{8}$$

Similar to (7) and (8), we have

$$E\left[\frac{1}{m(u_i, v_i)}\right] = \frac{1}{\bar{m}}, \tag{9}$$

and

$$Var\left[\frac{1}{m(u_i, v_i)}\right] = \sum_{[u,v]\in\mathbf{S}^{(2)}} \frac{1}{Mm(u,v)} - \frac{1}{\bar{m}^2}. \tag{10}$$

Denote

$$\xi_{k,1} = \frac{\bar{m}}{n}\sum_{i=1}^{n} \frac{\mathbf{1}(F(u_i, v_i) = a_k)}{m(u_i, v_i)}, \quad \text{and} \quad \xi_{k,2} = \frac{\bar{m}H}{n}.$$

Then, from (7) and (9) we have

$$E[\xi_{k,1}] = \omega_k^{(2)}, \quad \text{and} \quad E[\xi_{k,2}] = 1.$$

Application of the law of large numbers yields $\lim_{n\to\infty}\xi_{k,1} \xrightarrow{a.s.} \omega_k^{(2)}$ and $\lim_{n\to\infty}\xi_{k,2} \xrightarrow{a.s.} 1$, where "$a.s.$" denotes "almost sure" converge, i.e., the event happens with probability one. Therefore $\hat{\omega}_k^{(2)}$ is asymptotically unbiased because $\lim_{n\to\infty}\hat{\omega}_k^{(2)} = \lim_{n\to\infty}\frac{\xi_{k,1}}{\xi_{k,2}} \xrightarrow{a.s.} \omega_k^{(2)}$.

Since IWVS samples node pairs independently, we have the following equation from (8)

$$\begin{aligned} Var[\xi_{k,1}] &= \frac{\bar{m}^2}{n}\sum_{[u,v]\in\mathbf{S}^{(2)}} \frac{\mathbf{1}(F(u,v) = a_k)}{Mm(u,v)} - \frac{(\omega_k^{(2)})^2}{\bar{m}^2} \\ &= \frac{1}{n}\left(\frac{\bar{m}}{|\mathbf{S}^{(2)}|}\sum_{[u,v]\in\mathbf{S}^{(2)}} \frac{\mathbf{1}(F(u,v) = a_k)}{m(u,v)} - (\omega_k^{(2)})^2\right) \\ &\leq \frac{1}{n}\left(\bar{m}\omega_k^{(2)} - (\omega_k^{(2)})^2\right). \end{aligned}$$

The last inequality holds because of $\sum_{[u,v]\in\mathbf{S}^{(2)}}\mathbf{1}(F(u,v) = a_k) = |\mathbf{S}^{(2)}|\omega_k^{(2)}$ and $m(u,v) \geq 1$. Similarly, from (10) we have

$$\begin{aligned} Var[\xi_{k,2}] &= \frac{\bar{m}^2}{n}\sum_{[u,v]\in\mathbf{S}^{(2)}} \frac{1}{Mm(u,v)} - \frac{1}{\bar{m}^2} \\ &= \frac{1}{n}\left(\frac{\bar{m}}{|\mathbf{S}^{(2)}|}\sum_{[u,v]\in\mathbf{S}^{(2)}} \frac{1}{m(u,v)} - 1\right) \\ &\leq \frac{1}{n}(\bar{m} - 1). \end{aligned}$$

Using Chebyshev's inequality, we obtain

$$P\left(|\xi_{k,1} - \omega_k^{(2)}| < \epsilon\omega_k^{(2)}\right) \geq 1 - \frac{1}{n\epsilon^2}\left(\frac{\bar{m}}{\omega_k^{(2)}} - 1\right),$$

and

$$P\left(|\xi_{k,2} - 1| < \epsilon\right) \geq 1 - \frac{1}{n\epsilon^2}(\bar{m} - 1).$$

Therefore, we have the following inequalities

$$\frac{1-\epsilon}{1+\epsilon}\omega_k^{(2)} \leq \hat{\omega}_k^{(2)} \leq \frac{1+\epsilon}{1-\epsilon}\omega_k^{(2)}$$

with probability at least $1 - \frac{1}{n\epsilon^2}\left(\frac{\bar{m}}{\omega_k^{(2)}} + \bar{m} - 2\right)$. □

In summary, the complexities of all methods presented in this Section are shown in Table I. We see that MHWVS is more computation and memory efficient than IWVS.

TABLE I

COMPUTATIONAL AND SPACE COMPLEXITIES OF THE NODE PAIR SAMPLING METHODS BASED VERTEX SAMPLING TECHNIQUES.

	each sampling step: operations	**weighted sampling method: (operations, memory)**
$\mathbf{S}$	$O(1)$	not required
$\mathbf{S}^{(1)}$, IWVS	$O(\log\|V\|)$	$(O(\|V\|), O(\|V\|))$
$\mathbf{S}^{(1)}$, MHWVS	$O(1)$	$(O(1), O(1))$
$\mathbf{S}^{(2)}$, IWVS	$O(\log\|V\|)$	$(O(\|V\|), O(\|V\|))$
$\mathbf{S}^{(2)}$, MHWVS	$O(1)$	$(O(1), O(1))$

IV. Node Pair Sampling Methods Based on RW

In what follows, we assume that UVS is not feasible (either because we do not have the full graph topology, or generation cost of random ID is too expensive), and that graph G is connected. Instead, we study the use of a random walk as a node pair sampling technique. Random walks (RWs) have been extensively studied in the graph theory literature [21]. From an initial node, a RW selects a neighbor of the current node at random as the next-hop node. The walker moves to this neighbor and samples its information. Denote by $\boldsymbol{\pi} = (\pi_v : v \in V)$ the stationary distribution of RW, where $\pi_v = \frac{d_v}{2|E|}$. For a connected and non-bipartite graph G, the probability of being at node $v \in V$ converges to π_v [21]. Therefore, one can view this as a *non-uniform vertex sampling* algorithm: at each step, a node is selected from V according to the probability distribution $\boldsymbol{\pi}$. Note that RW is biased towards large degree nodes. However its bias is known and can be corrected [10], [32]. Compared to UVS, RW exhibits smaller estimation errors for characteristics associated with high degree nodes.

A. Sampling Node Pairs From $\mathbf{S}$ and $\mathbf{S}^{(1)}$

We use two independent RWs to sample node pairs $\{[u_i, v_i]\}_{i=1,\ldots,n}$ randomly from $\mathbf{S}$, where u_i and v_i are nodes sampled from graph G by these two RW at each step i separately. This sampling method can be viewed as a regular RW over $G^{(2)} = (V^{(2)}, E^{(2)})$, where $V^{(2)} = \{[u, v] : u, v \in V\}$ and $E^{(2)} = \{([u, v], [x, y]) : (u, x), (v, y) \in E\}$. It is clear that a node (node pair) $[u, v]$ in graph $G^{(2)}$ has $d_u d_v$ neighbors. When G is a connected and non-bipartite graph, we can easily show that graph $G^{(2)}$ is also connected and non-bipartite. Then an RW over $G^{(2)}$ exhibits a stationary distribution $\boldsymbol{\pi}_S = (\pi_{[u,v]} : u, v \in V)$, with

$$\pi_{[u,v]} = \frac{d_u d_v}{4|E|^2}, \quad u, v \in V.$$

Note that this RW may sample $[u, u]$ with stationary probability $\sum_{u \in V} \pi_{[u,u]} = \frac{\sum_{u \in V} d_u^2}{4|E|^2}$. Finally we estimate ω_k $(1 \le k \le K)$ as follows

$$\hat{\omega}_k^{\star} = \frac{1}{J} \sum_{i=1}^{n} \frac{\mathbf{1}(F(u_i, v_i) = a_k)\mathbf{1}(u_i \neq v_i)}{d_{u_i} d_{v_i}},$$

where $J = \sum_{i=1}^{n} \frac{\mathbf{1}(u_i \neq v_i)}{d_{u_i} d_{v_i}}$.

Theorem *2:* When graph G is connected and non-bipartite, $\hat{\omega}_k^{(\star)}$ $(k = 1, \ldots, K)$ is an asymptotically unbiased estimator of ω_k.

The proof can be found in [37].

To estimate characteristics of node pairs in $\mathbf{S}^{(1)}$, we sample node pars $\{[u_i, v_i]\}_{i=1,\ldots,n}$ based on a RW on G, where u_i and v_i are nodes sampled by RW at steps i and $i+1$ separately. We can easily show that (u_i, v_i) is an edge in G. The probabilities of a RW sampling edges are equal when the RW reaches steady state [28], we estimate $\omega_k^{(1)}$ $(1 \le k \le K)$ as follows

$$\hat{\omega}_k^{(1\star)} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(F(u_i, v_i) = a_k).$$

Theorem *3:* When graph G is connected and non-bipartite, $\hat{\omega}_k^{(1\star)}$ $(k = 1, \ldots, K)$ is an asymptotically unbiased estimator of $\omega_k^{(1)}$.

The proof can be found [37].

B. Sampling Node Pairs From $\mathbf{S}^{(2)}$

We present a new method named neighborhood random walk (NRW) to sample node pairs randomly from $\mathbf{S}^{(2)}$. It can be viewed as a regular RW over graph $\hat{G} = (\hat{V}, \hat{E})$, with node set $\hat{V} = \{(u, v) : (u, v) \in E\}$, edge set $\hat{E} = \{((u, v), (u, v')) : (u, v) \in E, (u, v') \in E, v \neq v'\}$. Let (u, v) be the initial edge for a NRW. Denote by $N_{(u,v)}$ the set of edges connected to u or v excluding edge (u, v). Clearly $|N_{(u,v)}| = d_u + d_v - 2$. Then NRW selects a random edge from $N_{(u,v)}$ as the next sampled edge. Formally, the NRW can be modeled as a Markov chain with transition matrix $P^{\text{NRW}} = [P^{\text{NRW}}_{e,e'}]$, where $e = (u, v)$ and $e' = (u', v')$ are edges in E, and $P^{\text{NRW}}_{e,e'}$ is defined as the probability that edge e' is selected as the next-hop edge given that its current edge e. $P^{\text{NRW}}_{e,e'}$ is computed as

$$P^{\text{NRW}}_{(u,v),(u',v')} = \begin{cases} \frac{1}{d_u + d_v - 2} & \text{if } (u', v') \in N_{(u,v)}, (u, v) \in E \\ 0 & \text{otherwise.} \end{cases}$$

We can easily show that a node (u, v) (an edge in G) in graph $\hat{G}$ connects to $d_u + d_v - 1$ nodes in $\hat{G}$, its degree in $\hat{G}$ is $d_u + d_v - 1$. Meanwhile $\hat{G}$ has $|\hat{E}| = M/2$ edges, where $M = \sum_{y \in V} d_y(d_y - 1)$. Then we have

Theorem *4:* When graph G is connected and non-bipartite, the NRW exhibits a stationary distribution $\boldsymbol{\pi}_E = (\pi_{(u,v)} : (u, v) \in E)$, where $\pi_{(u,v)}$ is

$$\pi_{(u,v)} = \frac{d_u + d_v - 2}{M}, \quad (u, v) \in E.$$

Proof. Suppose that the NRW is currently at edge (u, v) with probability distribution $\pi_{(u,v)}$, then edge $(u', v') \in E$ is selected with probability $p_{(u',v')}$ computed as

$$\begin{aligned} p_{(u',v')} &= \sum_{(u,v) \in N_{(u',v')}} \pi_{(u,v)} P^{\text{NRW}}_{(u,v),(u',v')} \\ &= \frac{|N_{(u',v')}|}{M} \\ &= \pi_{(u',v')} \end{aligned}$$

Therefore $\boldsymbol{\pi}_E$ is the stationary distribution of a Markov chain with transition matrix P^{NRW}. When G is connected and non-bipartite, $\hat{G}$ is also connected and non-bipartite. node (u, v) (an edge in G) in graph $\hat{G}$ connects to $d_u + d_v - 1$ nodes in $\hat{G}$, its degree in $\hat{G}$ is $d_u + d_v - 1$, and NRW can be viewed as a regular RW on graph $\hat{G}$, therefore the probability of NRW being at an edge $(u, v) \in E$ converges to $\boldsymbol{\pi}_E$ from [21]. □

The pseudo-code for the NRW based node pair sampling algorithm is depicted in **Algorithm 1**. Let (x_i, y_i) and s_i be i-th $(i \ge 0)$ visited edge and node. For each step i, the next visited edge (x_{i+1}, y_{i+1}) is randomly selected from $N_{(x_i,y_i)}$, which has exactly one common node with current edge (x_i, y_i). By excluding this common node, we obtain

two distinct nodes u and v in these two edges and output node pair $[u, v]$ or $[v, u]$ with equal probability. Each edge $((w, u), (w, v))$ in graph $\hat{G}$ can generate a node pair consisting of two distinct nodes u and v by excluding the common node w, therefore node pair $[u, v]$ can be generated by $m(u, v)$ different edges in $\hat{G}$, where $m(u, v)$ is the number of common neighbors of nodes u and v in original graph G. NRW can be viewed as a regular RW over graph $\hat{G}$, and it samples edges randomly from $\hat{G}$ with the same probability [21], therefore a node pair (u, v) is sampled by NRW with a stationary probability $\frac{m(u,v)}{M}$. Based on sampled pairs $\{[u_i, v_i]\}_{i=1,\ldots,n}$, we estimate $\omega_k^{(2)}$ $(1 \le k \le K)$ as follows

$$\hat{\omega}_k^{(2*)} = \frac{1}{H} \sum_{i=1}^{n} \frac{\mathbf{1}(F(u_i, v_i) = a_k)}{m(u_i, v_i)},$$

where $H = \sum_{i=1}^{n} \frac{1}{m(u_i,v_i)}$.

Theorem *5:* When graph G is connected and non-bipartite, $\hat{\omega}_k^{(2\star)}$ $(k = 1, \ldots, K)$ is an asymptotically unbiased estimator of $\omega_k^{(2)}$.

The proof can be found [37].

V. Data Evaluation

Our simulation experiments are performed over a variety of real world graphs, which are summarized in Table II. Wikipedia is a free encyclopedia written collaboratively by volunteers. Each registered user has a talk page, which the user and other users can edit in order to communicate and discuss updates to various articles on Wikipedia. Nodes in the network represent Wikipedia users and a directed edge from node u to node v represents that user u voted on user v. Gnutella is a peer-to-peer file sharing network. Nodes represent hosts in the Gnutella network topology and edges represent connections between the Gnutella hosts. Another network is Epinions, a general consumer review website. Users build a who-trust-whom online social network, where a directed edge from nodes u to v represents that u trusts v. Slashdot is a technology-related news website where a node represents a user and a directed edge from nodes u to v represents that u tags v as a friend or foe. We test our sampling methods on their corresponding undirected graphs which were generated by ignoring the directions of edges.

Define the normalized mean square error as

$$NMSE(\hat{\omega}_k) = \frac{\sqrt{E[(\hat{\omega}_k - \omega_k)^2]}}{\omega_k}, k = 1, 2, \ldots$$

We use this metric to measure the relative error of the estimate $\hat{\omega}_k$ with respect to its true value ω_k. Because we use relative error, when $\hat{\omega}_k$ is small, we consider values as large as $NMSE(\hat{\omega}_k) = 1$ to be acceptable.

A. Distance distribution

We evaluate the performance of UVS for estimating $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$, the distance distribution of node pairs in $\mathbf{S}$, where K is the graph diameter, and graphs used are the largest connected component (LCC) of Wiki-vote and the LCC of

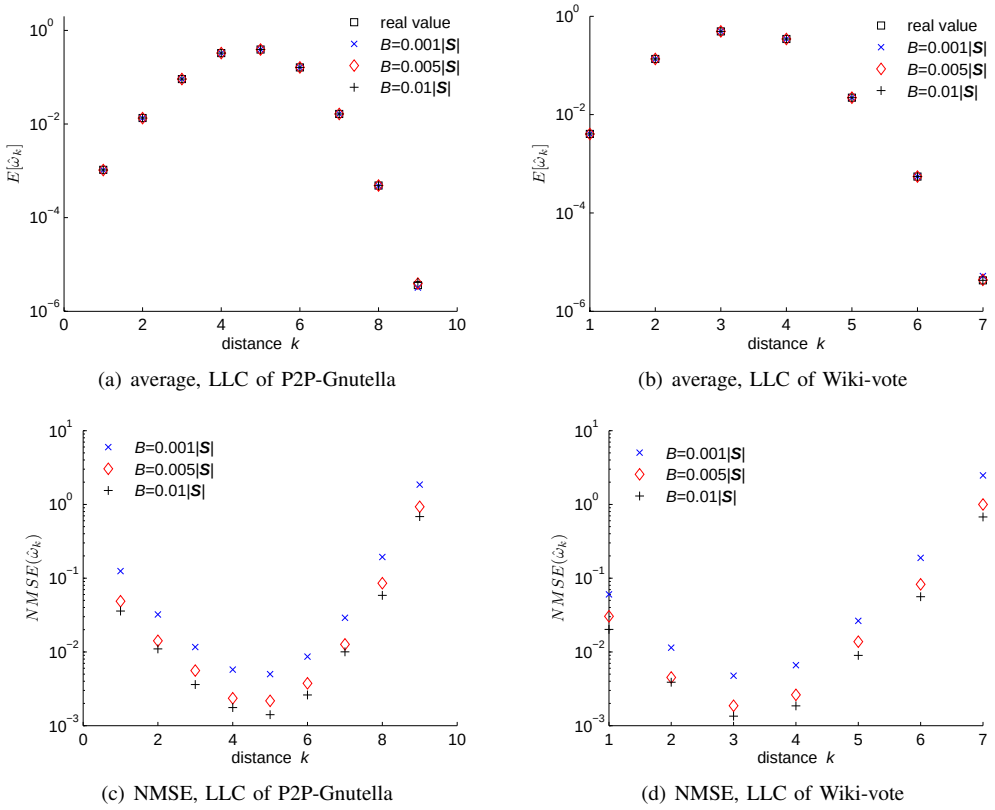
Algorithm 1: NRW pseudo-code.

```
/* n is the sampling budget, (x_0, y_0) is the
   initial edge, and (x_i, y_i) and s_i are visited
   edge and node at step i.                          */
input : n and (x_0, y_0) ∈ E
output: node pairs [u_1, v_1], [u_2, v_2], ..., [u_n, v_n]
i ← 0;
while i <= n do
    /* U(0,1) is a uniform (0,1) random sample      */
    Generate p ← U(0,1);
    /* d_x is the degree of node x in G             */
    if p < (d_{x_i} − 1)/(d_{x_i} + d_{y_i} − 2) then
        /* randomNeighbor(x, y) returns a node
           selected randomly from neighbors of
           node x excluding node y                   */
        s_i ← randomNeighbor(x_i, y_i);
        x_{i+1} ← x_i and y_{i+1} ← s_i;
        /* u and v are nodes in sequentially
           visited edges (x_i, y_i) and (x_{i+1}, y_{i+1})
           excluding their common node               */
        u ← y_i and v ← s_i;
    else
        s_i ← randomNeighbor(y_i, x_i);
        x_{i+1} ← s_i and y_{i+1} ← y_i;
        u ← x_i and v ← s_i;
    end
    Generate q ← U(0,1);
    if q < 0.5 then
        u_{i+1} ← u and v_{i+1} ← v;
    else
        u_{i+1} ← v and v_{i+1} ← u;
    end
    i ← i + 1;
end
```

P2P-Gnutella. Fig. 1 presents $NMSE(\hat{\omega}_k)$ $(1 \le k \le K)$ for sampling budgets $B = \{0.001|\mathbf{S}|, 0.005|\mathbf{S}|, 0.01|\mathbf{S}|\}$. When $B \ge 0.05|\mathbf{S}|$, the $NMSE(\omega_k)$ is always smaller than one. We observe that the error of sampling B pairs from $\mathbf{S}$ is roughly proportional to $1/\sqrt{B}$. For instance, in Fig. 1 we see that an order of magnitude increase in B roughly decreases the error by $1/\sqrt{10}$.

B. Mutual neighbor count distribution

The number of mutual neighbors for a pair of nodes is usually used as a metric to indicate the strength of their relationship [33]. Define $\omega_k^{(1)}$ and $\omega_k^{(2)}$ as the fraction of node pairs with $k \ge 1$ mutual neighbors in $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ respectively. Fig. 2 shows the complementary cumulative distribution function (CCDF) of $\boldsymbol{\omega}^{(1)}$ and $\boldsymbol{\omega}^{(2)}$ for graphs soc-Epinions and soc-Slashdot. The size of $\mathbf{S}^{(2)}$ is 7.34×10^7 and 9.49×10^7 for soc-Epinions and soc-Slashdot respectively. The statistics for LCC of soc-Epinions and LCC of soc-Slashdot



(a) average, LLC of P2P-Gnutella

(b) average, LLC of Wiki-vote

(c) NMSE, LLC of P2P-Gnutella

(d) NMSE, LLC of Wiki-vote

Fig. 1. Average and NMSE of distance distribution estimates.

TABLE II

OVERVIEW OF DIRECTED GRAPH DATASETS USED IN OUR SIMULATIONS.

Graph	Entire Graph		LCC	
	nodes	edges	nodes	edges
Wiki-vote [17], [18]	7,115	103,689	7,066	103,663
P2P-Gnutella [31]	10,876	39,994	10,876	39,994
soc-Epinions [30]	75,879	508,837	75,877	508,836
soc-Slashdot [20]	77,360	905,468	77,360	90,5468

"LCC" refers to the largest connected component in the undirected graph generated by ignoring the directions of edges.

are similar.

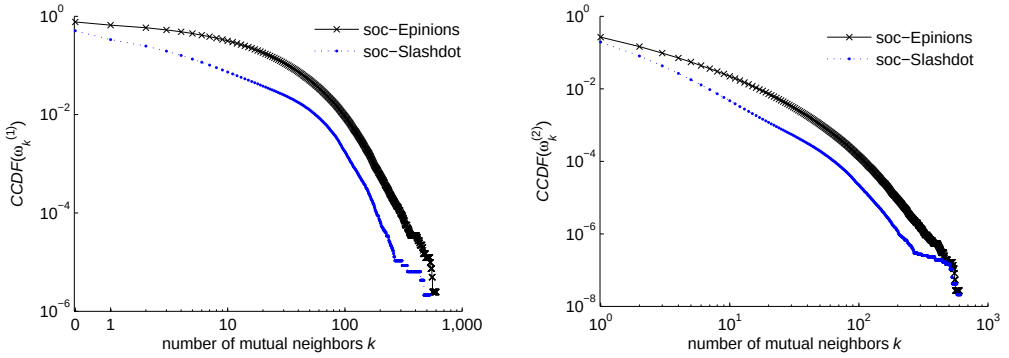
For set $\mathbf{S}^{(1)}$, we evaluate the performance of sampling methods: IWVS and MHWVS presented in Section III-B, and RW presented in Section IV-A using graph soc-Epinions. Fig. 3 presents $NMSE(\hat{\omega}_k^{(1)})$ for sampling budgets $B = \{0.001|\mathbf{S}^{(1)}|, 0.005|\mathbf{S}^{(1)}|, 0.01|\mathbf{S}^{(1)}|\}$. We find that the error of sampling B pairs from $\mathbf{S}^{(1)}$ is roughly proportional to $1/\sqrt{B}$ for each method. Fig. 4 compares the NMSE of the three sampling methods with the same sampling budget. It shows that RW and IWVS are slightly more accurate than MHWVS, and RW almost has the same accuracy of IWVS. The results for graph soc-Slashdot are similar and can be found in [37].

For set $\mathbf{S}^{(2)}$, we evaluate the performance of the following methods: IWVS and MHWVS presented in Section III-C, and NRW presented in Section IV-B using graph soc-Slashdot. Fig. 5 presents $NMSE(\hat{\omega}_k^{(2)})$ for sampling budgets $B = \{0.001|\mathbf{S}^{(2)}|, 0.005|\mathbf{S}^{(2)}|, 0.01|\mathbf{S}^{(2)}|\}$. When $B > 0.05|\mathbf{S}^{(2)}|$, all $NMSE(\omega_k^{(2)})$ are smaller than one for each sampling methods. Fig. 6 compares the NMSE of three sampling methods under the same sampling budget. It shows that NRW and IWVS have much smaller errors than MHWVS, and NRW almost exhibit the same accuracy as IWVS. The results for graph soc-Epinions are similar and can be found in [37].

C. Similarity distribution

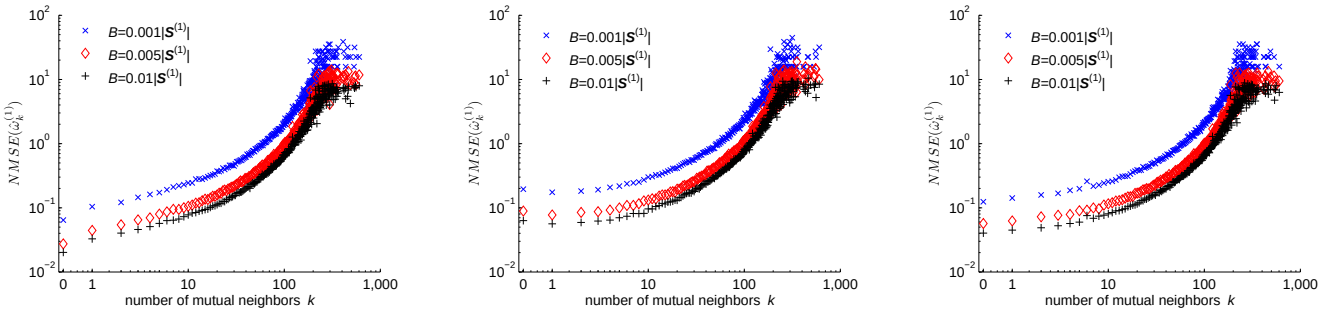
It is hard to obtain all users' interests in a real large OSN due to resource limits. Using publicly available graph topologies, we manually generate interests and distribute them over these graphs, and use them as benchmark datasets for our simulation experiments. We use the following interest distribution scheme (IDS) to distribute interests over a graph: To distribute an interest possessed by k different nodes, it first selects a random node v that can reach at least $k-1$ different nodes, where two nodes are reachable if there is at least one path between them in the undirected graph. Then we distribute this interest to node v and the closest $k-1$ nodes connected to v.

Define the truncated Pareto distribution as $\theta_k = \frac{\alpha}{\gamma k^{\alpha+1}}$, $k = 1, \ldots, W$, where $\alpha > 0$ and $\gamma = \sum_{k=1}^{W} \frac{\alpha}{k^{\alpha+1}}$. In the following experiments we generate 10^5 distinct interests, and for each interest the number of nodes possessed it is random variable selected from set $\{1, \ldots, W\}$ according to the truncated Pareto distribution with parameter $\alpha = 1$ and $W = 10^3$. The graphs used are the LCC of P2P-Gnutella and LCC of Wiki-vote, where the size of $\mathbf{S}^{(2)}$ is 2.69×10^5 and 3.46×10^6 separately.



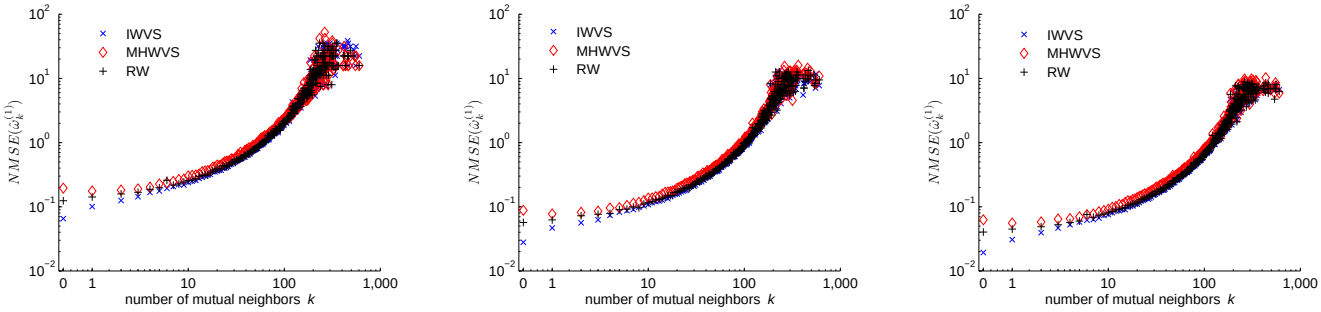
(a) set $\mathbf{S}^{(1)}$ (b) set $\mathbf{S}^{(2)}$

Fig. 2. (soc-Epinions and soc-Slashdot) CCDF of the distributions of node pairs by the mutual neighbor count.



(a) (soc-Epinions) IWVS (b) (soc-Epinions) MHWVS (c) (LCC of soc-Epinions) RW

Fig. 3. NMSE of distribution estimates of pairs in $\mathbf{S}^{(1)}$ by the mutual neighbor count for different sampling budget B.



(a) $B = 0.001|\mathbf{S}^{(1)}|$ (b) $B = 0.005|\mathbf{S}^{(1)}|$ (c) $B = 0.01|\mathbf{S}^{(1)}|$

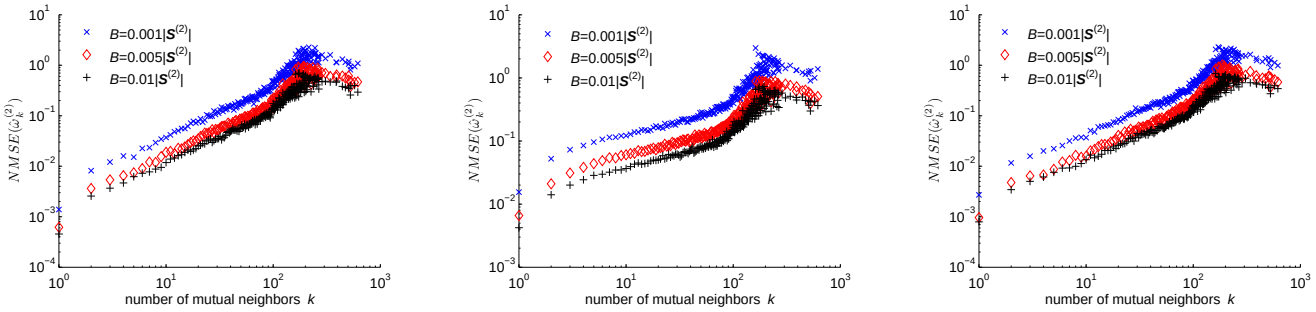
Fig. 4. (LCC of soc-Epinions) Compared NMSE of distribution estimates of pairs in $\mathbf{S}^{(1)}$ by the mutual neighbor count for different methods.

Define ω_k, $\omega_k^{(1)}$ and $\omega_k^{(2)}$ as the fraction of node pairs with $k \geq 1$ common interests in $\mathbf{S}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$ separately. Fig. 7 shows the CCDF of $\boldsymbol{\omega}$, $\boldsymbol{\omega}^{(1)}$, and $\boldsymbol{\omega}^{(2)}$ finally generated.

Fig. 8 shows the $NMSE$ of sampling methods for set $\mathbf{S}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$ under the same number of sampled pairs. For set $\mathbf{S}$, results show that UVS is more accurate for estimating ω_k with small k, and RW is more accurate for estimating ω_k with large k. This is because of RW is biased to sample high degree nodes, and IDS generates more interests for high degree nodes than nodes with small degrees. It is similar to the observation for estimating degree distribution using RW and UVS [28]. For set $\mathbf{S}^{(1)}$, we find that IWVS, MHWVS, and RW almost have the same accuracy. For set $\mathbf{S}^{(2)}$, the results show that IWVS has the smallest errors for estimating $\omega_k^{(2)}$ with small k, and MHWVS performs worst. In [37], we also show simulation results for other IDS distribution schemes such as independent cascade model [8].

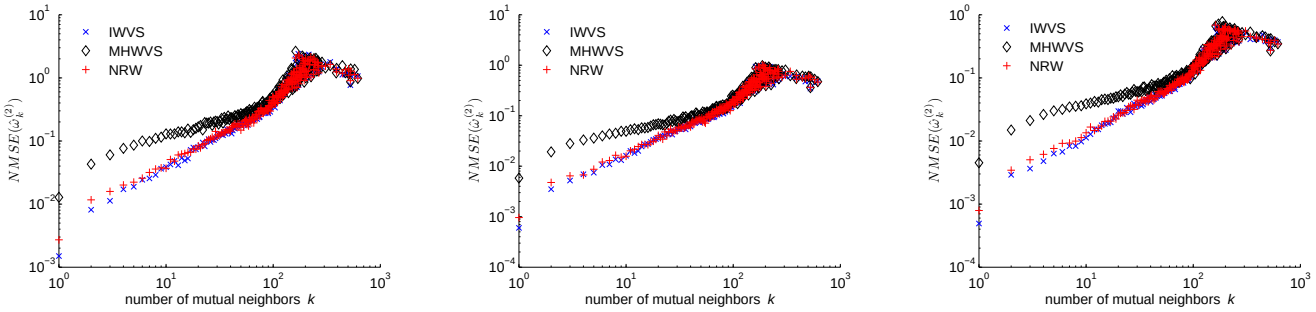
VI. Applications

In this section, we conduct real experiments on two popular Chinese OSNs: Douban and Xiami. Douban mainly provides an exchange platform for reviews and recommendations on movies, books, and music albums. It has approximately 6 million registered users as of 2009 [40]. Each user of Douban



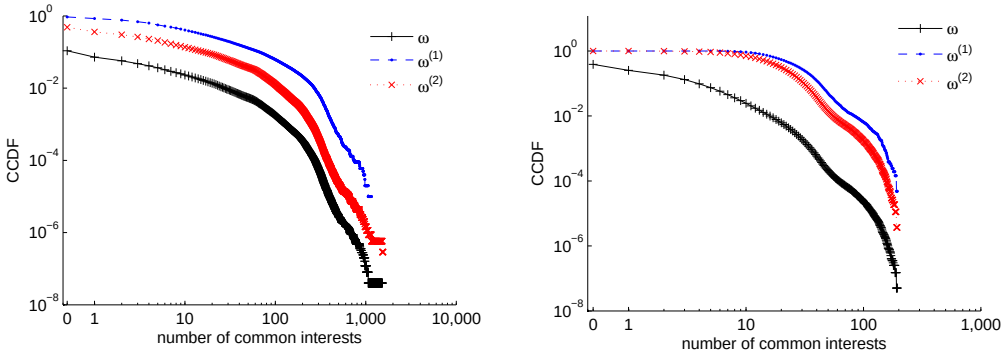
(a) (soc-Slashdot) IWVS (b) (soc-Slashdot) MHWVS (c) (LCC of soc-Epinions) NRW

Fig. 5. NMSE of distribution estimates of pairs in $\mathbf{S}^{(2)}$ by the mutual neighbor count for different sampling budget B.



(a) $B = 0.001|\mathbf{S}^{(2)}|$ (b) $B = 0.005|\mathbf{S}^{(2)}|$ (c) $B = 0.01|\mathbf{S}^{(2)}|$

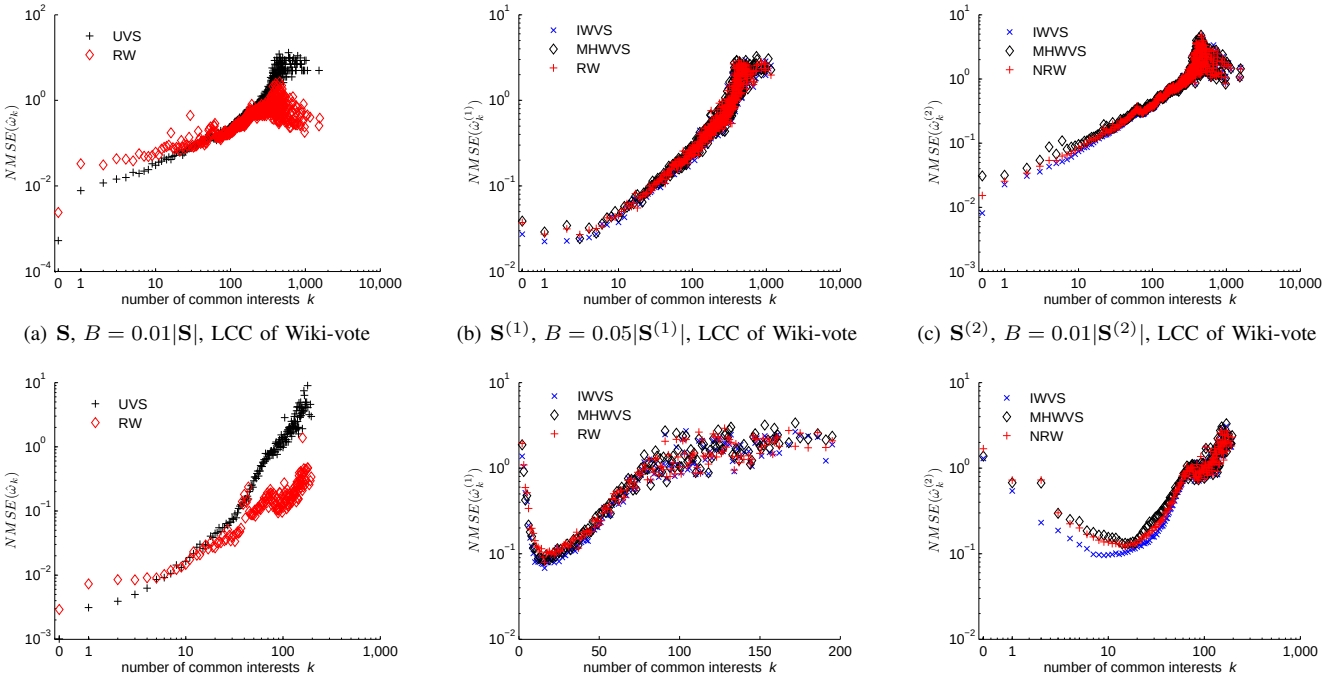
Fig. 6. (LCC of soc-Slashdot) Compared NMSE of distribution estimates of pairs in $\mathbf{S}^{(2)}$ by the mutual neighbor count for different methods.



(a) LCC of Wiki-vote (b) LCC of P2P-Gnutella

Fig. 7. Distribution of pairs in $\mathbf{S}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$ by the common interest count separately.

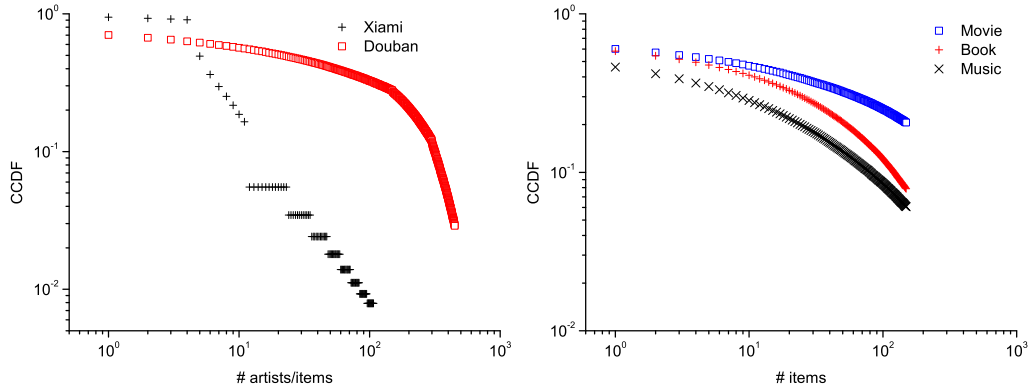
maintains three lists for books, movies and music albums respectively. Xiami is a popular website devoted for music streaming service and music recommendation, and has approximately 1.7 million users as of 2011 [36]. Each user of Xiami maintains a list of his/her favorite artists. Fig. 9 shows statistics of users' interests in Xiami and Douban, which is measured based on 101,401 unique Douban users and 524,283 unique Xiami users sampled by a RW. On average, a Xiami user is interested in 8.76 artists, and a Douban user in 96.03 items consisting of 46.26 movies, 29.43 books, and 20.34 music albums. To measure interest similarities of users in Xiami and Douban, we collected 171,860 Xiami user pairs and 50,700 Douban user pairs from set $\mathbf{S}$, 105,736 Xiami user pairs and 85,631 Douban user pairs from set $\mathbf{S}^{(1)}$ using the UVS based methods presented in Section IV, and 359,522 Xiami user pairs and 96,361 Douban user pairs from set $\mathbf{S}^{(2)}$ by our new RW based method NRW. As shown in Fig. 10, we observe that user pairs in $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ have much more common interests than user pairs in $\mathbf{S}$, and user pairs in $\mathbf{S}^{(2)}$ have a fewer common interests than user pairs in $\mathbf{S}^{(1)}$. This is also true for three different kinds of interests, movies, books, music albums in Douban, which is shown in Fig. 11. This indicates that users in Xiami and Douban tend to connect to ones with the similar interests.



(a) $\mathbf{S}$, $B = 0.01|\mathbf{S}|$, LCC of Wiki-vote (b) $\mathbf{S}^{(1)}$, $B = 0.05|\mathbf{S}^{(1)}|$, LCC of Wiki-vote (c) $\mathbf{S}^{(2)}$, $B = 0.01|\mathbf{S}^{(2)}|$, LCC of Wiki-vote

(d) $\mathbf{S}$, $B = 0.01|\mathbf{S}|$, LCC of P2P-Gnutella (e) $\mathbf{S}^{(1)}$, $B = 0.05|\mathbf{S}^{(1)}|$, LCC of P2P-Gnutella (f) $\mathbf{S}^{(2)}$, $B = 0.01|\mathbf{S}^{(2)}|$, LCC of P2P-Gnutella

Fig. 8. Compared NMSE of distribution estimates of pairs in $\mathbf{S}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$ by the common interest count for different methods.



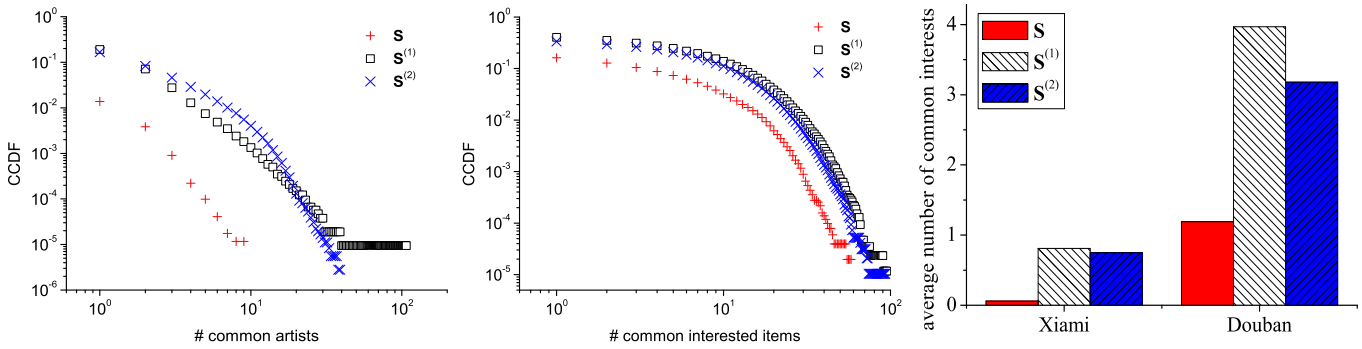
(a) Xiami and Douban, distribution of users by the number of interested artists/items.

(b) Douban, distribution of users by the number of interested movies, books, music albums.

Fig. 9. (Douban and Xiami) Statistics of users' interests

VII. **Related Work**

Let us provide a brief summary on related work. Singla et al. [34] reveal that significant homophily is present in the MSN Messenger network based on the study of user pairs' similarities in terms of their Web search topics, and personal characteristics such as their ages and locations. Similar results also are found in [16]. There are also works on measuring the distance statistics of user pairs in OSNs [15], [16], [19]. Leskovec et al. [19] show that the effective diameter for a range of real networks gradually decreases as the network grows, which contradicts the basic assumption of existing network evolution models. Previous graph sampling work focuses on designing accurate and efficient methods for measuring graph characteristics, such as node degree distribution [6], [27]–[29], [35] and the topology of nodes' groups [12]. These sampling methods have been widely applied to characterize complex networks, such as P2P networks [7], [22], [27], [35], and OSNs [5], [6], [11], [25], [38]. We summarize previous graph sampling work as follows: Breadth-First-Search (BFS) introduces bias towards high-degree nodes that is unknown and difficult to remove in general graphs [1], [13], [14]. Random walk (RW) is biased to sample high degree nodes, however its bias is known and can be corrected for [10], [32]. Compared to uniform vertex sampling (UVS), a random walk has smaller estimation errors for the characteristics of high degree nodes, especially for networks where UVS is

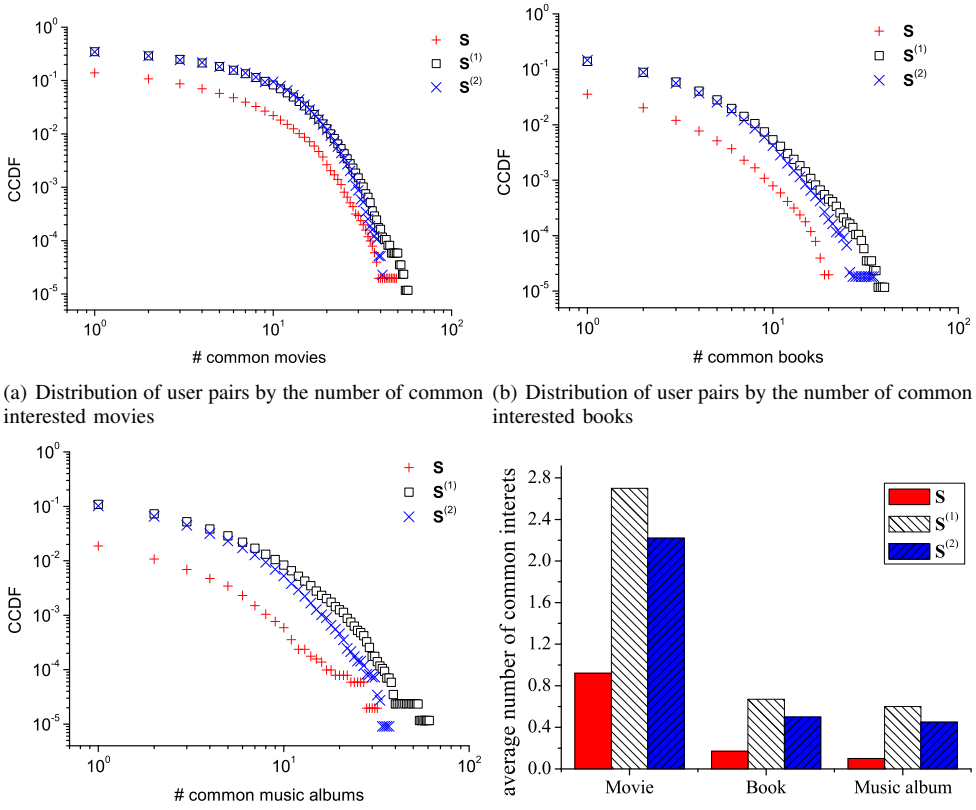


(a) Xiami, distribution of user pairs by the number of common interested artists

(b) Douban, distribution of user pairs by the number of common interested items

(c) Xiami and Douban, average number of common interests

Fig. 10. (Xiami and Douban) Statistics of users' common interests



(a) Distribution of user pairs by the number of common interested movies

(b) Distribution of user pairs by the number of common interested books

(c) Distribution of user pairs by the number of common interested music albums

(d) Average number of common interested movies, books, and music albums

Fig. 11. Statistics of users' common interested movies, books, and music albums in Douban

costly (e.g., Flickr, Facebook, and MySpace) [28]. Compared to RW that reweights sampled values to obtain an unbiased estimate of graph characterizes, The Metropolis-Hasting RW (MHRW) [6], [35], [41] modifies the random walk procedure using the Metropolis-Hasting technique, which aims to sample each node uniformly. The accuracy of RW and MHRW is compared in [6], [27], and in a variety of experiments estimates obtained by RW are shown to be consistently more accurate than or equal to that of MHRW. The mixing time of RW determines the efficiency of the sampling, and it is found practically much larger than commonly believed [26] for many OSNs. There are a lots of work to decrease mixing time [2], [3], [5], [11], [28]. To the best of our knowledge, this paper is the first to study and provide a sound theoretical analysis of the problem of sampling node pairs with constraints in the graph.

VIII. Conclusions

In this work we systemically study the problem of estimating characteristics of node pairs in sets $\mathbf{S}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$ for

ones with/witout the complete graph topology. We propose two kinds of sampling methods based on uniform vertex sampling and random walk techniques, and prove that they are asymptotical unbiased estimators. Our simulation results show that random walk based methods and uniform vertex based methods almost have the similar accuracy, especially for the sampling methods for $\mathbf{S}^{(1)}$. Finally we apply our methods to Doudan and Xiami OSNs, and discover that there is a strong tendency for users to connect to others with common interests.

REFERENCES

[1] Dimitris Achlioptas, David Kempe, Aaron Clauset, and Cristopher Moore. On the bias of traceroute sampling or, power-law degree distributions in regular graphs. In *Symposium on Theory of Computing 2005*, pages 694–703, May 2005.

[2] Konstantin Avrachenkov, Bruno Ribeiro, and Don 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.

[3] Stephen Boyd, Persi Diaconis, and Lin Xiao. Fastest mixing markov chain on a graph. *SIAM Review*, 46(4):667–689, December 2004.

[4] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, November 1995.

[5] Minas Gjoka, Carter T. Butts, Maciej Kurant, and Athina Markopoulou. Multigraph sampling of online social networks. *IEEE Journal on Selected Areas in Communications*, 29(9):1893–1905, September 2011.

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

[7] Christos Gkantsidis, Milena Mihail, and Amin Saberi. Random walks in peer-to-peer networks: algorithms and evaluation. *Performance Evaluation*, 63(3):241–263, March 2006.

[8] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001.

[9] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

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

[11] Maciej Kurant, Minas Gjoka, Carter T. Butts, and Athina Markopoulou. Walking on a graph with a magnifying glass: Stratified sampling via weighted random walks. In *Proceedings of ACM SIGMETRICS 2011*, pages 281–292, June 2011.

[12] Maciej Kurant, Minas Gjoka, Yan Wang, Zack W. Almquist, Carter T. Butts, and Athina Markopoulou. Coarse-grained topology estimation via graph sampling. Technical Report arXiv:1105.5488, 2011.

[13] Maciej Kurant, Athina Markopoulou, and Patrick Thiran. On the bias of bfs (breadth first search) and of other graph sampling techniques. In *Proceedings of International Teletraffic Congress 2010*, September 2010.

[14] Maciej Kurant, Athina Markopoulou, and Patrick Thiran. Towards unbiased bfs sampling. *IEEE Journal on Selected Areas in Communications*, 29(9):1799–1809, September 2011.

[15] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of WWW 2010*, pages 591–600, April 2010.

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

[17] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of WWW 2010*, pages 641–650, April 2010.

[18] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the 28th ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1361–1370, April 2010.

[19] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of ACM SIGKDD 2005*, pages 177–187, June 2005.

[20] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael 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.

[21] L. Lovász. Random walks on graphs: a survey. *Combinatorics*, 2:1–46, 1993.

[22] Laurent Massoulié, Erwan Le Merrer, Anne-Marie Kermarrec, and Ayalvadi Ganesh. Peer counting and sampling in overlay networks: random walk methods. In *Proceedings of the PODC 2006*, pages 123–132, July 2006.

[23] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equations of state calculations by fast computing machines. *IEEE Journal on Selected Areas in Communications*, 21(6):1087–1092, June 2011.

[24] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[25] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of ACM SIGCOMM Internet Measurement Conference 2007*, pages 29–42, October 2007.

[26] Abedelaziz Mohaisen, Aaram Yun, and Yongdae Kim. Measuring the mixing time of social graphs. In *Proceedings of ACM SIGCOMM Internet Measurement Conference 2010*, pages 390–403, November 2010.

[27] Amir H. Rasti, Mojtaba Torkjazi, Reza Rejaie, Nick Duffield, Walter Willinger, and Daniel Stutzbach. Respondent-driven sampling for characterizing unstructured overlays. In *Proceedings of IEEE INFOCOM Mini-conference 2009*, April 2009.

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

[29] Bruno Ribeiro, Pinghui Wang, Fabricio Murai, and Don Towsley. Sampling directed graphs with random walks. In *Proceedings of IEEE INFOCOM 2012*, pages 1692–1700, April 2012.

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

[31] Matei Ripeanu, Ian T. Foster, and Adriana Iamnitchi. Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. *IEEE Internet Computing Journal*, 6(1):50–57, 2002.

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

[33] Xiaolin Shi, Lada A. Adamic, and Martin J. Strauss. Networks of strong ties. *Physica A: Statistical Mechanics and its Applications*, 378(1):33–47, May 2007.

[34] Parag Singla and Matthew Richardson. Yes, there is a correlation - from social networks to personal behavior on the web. In *Proceedings of WWW 2008*, pages 655–664, April 2008.

[35] Daniel Stutzbach, Rea Rejaie, Nick Duffield, Subhabrata Sen, and Walter Willinger. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Transactions on Networking*, 17(2):377–390, April 2009.

[36] Pinghui Wang, Junzhou Zhao, Xiaohong Guan, John C.S. Lui, and Don Towsley. Sampling contents distributed over graphs. Technical Report TR-1201, Xi'an Jiaotong University, 2012.

[37] Pinghui Wang, Junzhou Zhao, John C.S. Lui, Don Towsley, and Xiaohong Guan. Sampling node pairs over graphs, available at http://www.cse.cuhk.edu.hk/%7ephwang/samplingpairreport.pdf. Technical report, The Chinese University of Hong Kong, 2012.

[38] Yong yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of WWW 2007*, pages 835–844, May 2007.

[39] Nielsen statistics, june 2010, (`http://www.cs.cornell.edu/projects/kddcup/datasets.html`).

[40] Junzhou Zhao, John C. S. Lui, Don Towsley, Xiaohong Guan, and Yadong Zhou. Empirical analysis of the evolution of follower network: A case study on douban. In *Proceedings of IEEE INFOCOM NetSciCom 2011*, pages 941–946, April 2011.

[41] Ming Zhong and Kai Shen. Random walk based node sampling in self-organizing networks. *ACM SIGOPS Operating Systems Review*, 40(3):49–55, July 2006.