Unbiased Characterization of Node Pairs Over Large Graphs

Pinghui Wang, Noah's Ark lab, Huawei, Hong Kong
Junzhou Zhao, Xi'an Jiaotong University, China
John C.S. Lui, The Chinese University of Hong Kong, Hong Kong
Don Towsley, University of Massachusetts Amherst, US
Xiaohong Guan, Xi'an Jiaotong University, China

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 possess 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 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 three OSNs: Foursquare, Douban and Xiami, and discover significant homophily is present in these networks.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Miscellaneous

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Social network, graph sampling, random walks, homophily

ACM Reference Format:
Pinghui Wang, Junzhou Zhao, John C.S. Lui, Don Towsley, and Xiaohong Guan, 2014. Unbiased Characterization of Node Pairs Over Large Graphs. ACM Trans. Knowl. Discov. Data. V, N, Article A (January YYYY), 25 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

Online social networks (OSNs) such as Facebook and Twitter have become extremely popular within the last few years. OSNs have greatly changed people's network activities. They help peo-

This work was supported by the NSF grant CNS-1065133, ARL Cooperative Agreement W911NF-09-2-0053, and ARO under MURI W911NF-08-1-0233. 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 National Natural Science Foundation of China (61103240, 61103241, 61221063, 91118005, 61221063, U1301254), 863 High Tech Development Plan (2012AA011003), 111 International Collaboration Program of China, the Application Foundation Research Program of SuZhou (SYG201311) and the Prospective Research Project on Future Networks of Jiangsu Future Networks Innovation Institute.

Author's addresses: Pinghui Wang, Units 525-530, Core Building 2, Hong Kong Science Park Shatin, Hong Kong, Email: wang.pinghui@huawei.com. Junzhou Zhao, PO Box 1088, No.28, Xianning West Road, Xi'an, Shaanxi, 710049, China, Email: jzzhao@sei.xjtu.edu.cn. John C.S. Lui, Room 111, Ho Sin Hang Engineering Building, The Chinese University of Hong Kong, Shatin, Hong Kong, Email: cslui@cse.cuhk.edu.hk. Don Towsley, Department of Computer Science, University of Massachusetts, Amherst, MA, 01003, USA, Email: towsley@cs.umass.edu. Xiaohong Guan, PO Box 1088, No.28, Xianning West Road, Xi'an, Shaanxi, 710049, China, Email: xhguan@xjtu.edu.cn.

An early version of this paper appeared in the IEEE ICDE 2013 conference [Wang et al. 2013].

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

ple 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. [Singla and Richardson 2008] 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 [Milgram 1967], 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 the set $\mathbf{S}$) but also connected pairs (contained in the set $\mathbf{S}^{(1)}$), and pairs that share a neighbor (contained in the set $\mathbf{S}^{(2)}$), where $\mathbf{S}$ consists of all node pairs in G, $\mathbf{S}^{(1)}$ consists of pairs of connected nodes, and $\mathbf{S}^{(2)}$ consists of pairs of nodes with at least one common neighbor. Methods for characterizing node pairs in these three sets can be easily applied to problems 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 $\mathbf{S}$, $\mathbf{S}^{(1)}$, 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) can be directly applied to characterizing node pairs in $\mathbf{S}$. However, UVS might not be publicly available for OSNs. Moreover, it is a challenge to characterize node pairs in $\mathbf{S}^{(2)}$. 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 the 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 the node pair $[u, v]$ since one needs to query nodes u, v and all common neighbors of u and v.

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 $\mathbf{S}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$. Our major contributions can be summarized as follows:

1) When UVS is available, we propose a weighted vertex sampling (WVS) method to sample node pairs in $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$, and develop corresponding unbiased estimators for measuring node

pairs' statistics. Our WVS method can be applied to settings where the graph topology may or may not be known.

2) 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 pairs' statistics. Besides estimating statistics of node pairs in $\mathbf{S}$ and $\mathbf{S}^{(1)}$ by using the regular random walk (RW) sampling method, we propose a neighborhood random walk (NRW) method to characterize node pairs in $\mathbf{S}^{(2)}$. 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 the original graph G, and an edge in $\hat{G}$ consists of two edges in G with a common node.

This paper is organized as follows. The problem is formulated in Section 2. Section 3 and Section 4 present node pair sampling methods for ones with or without the complete graph topology respectively. Performance evaluation and testing results are presented in Section 5. Section 6 presents real applications on Foursquare[1], Xiami[2] and Douban[3] websites. Section 7 summarizes related work. Section 8 concludes.

2. 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 $[u, v] \neq [v, u]$. 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 $\mathbf{S}^{(1)}$ 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. Note that $\mathbf{S}^{(2)}$ may not contain each pair of nodes $[u, v]$ in $\mathbf{S}^{(1)}$ because u and v need not have any mutual neighbors. For a node pair $[u, v]$, let function $F(u, v)$ define 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., $F(u, v)$ could be 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 node pairs $[u, v]$ with $F(u, v) = a_k$ in $\mathbf{S}$, $\mathbf{S}^{(1)}$, $\mathbf{S}^{(2)}$, and $\mathbf{S}^{(2+)}$ respectively. 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, we define $\boldsymbol{\omega}^{(1-)} = (\omega_1^{(1-)}, \ldots, \omega_K^{(1-)})$, where $\omega_k^{(1-)}$ $(1 \leq k \leq K)$ is the fraction of node 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

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

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 the node pairs in $\mathbf{S}^{(1)}$, and β and $\omega_k^{(2)}$ can be calculated based on the node pairs in $\mathbf{S}^{(2)}$. Since

[1] www.foursquare.com
[2] www.xiami.com
[3] www.douban.com

$\omega_k^{(2+)}$ 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 the 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 $\boldsymbol{\omega}$, $\boldsymbol{\omega}^{(1)}$, 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 UVS and RW respectively.

3. NODE PAIR SAMPLING BASED ON UVS

To the best of our knowledge, previous work such as [Singla and Richardson 2008] directly uses sampling methods such as UVS and uniform edge sampling (UES) to characterize the node pairs in $\mathbf{S}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$. Clearly, it is easy to generate a random node pair in $\mathbf{S}$ through sampling two different nodes from V by UVS, and generate a random node pair in $\mathbf{S}^{(1)}$ by sampling an edge from E by UES. However, in practice UES is not publicly available for most OSNs, and it is difficult to uniformly sample node pairs from $\mathbf{S}^{(2)}$ and accurately compute corresponding statistics by UVS and UES. To solve these problems, in this section we present our sampling methods based on UVS to estimate statistics of the node pairs in $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$.

3.1. Basic Sampling Operations and Their Cost

Suppose that we can sample nodes from the graph G using UVS with replacement. For example, there is a numeric ID associated with each node for OSNs such as Foursquare. Then one can perform UVS by sampling IDs randomly from the ID space with replacement. This computation complexity is $O(1)$ when the ID values of nodes are sequentially assigned.

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. For simplicity, we denote $V = \{\mathbf{v}_1, \dots, \mathbf{v}_{|V|}\}$, where $\mathbf{v}_i$ is the node with ID $i \in \{1, \dots, |V|\}$. Then IWVS assigns a weight $W[i]$ to each node $\mathbf{v}_i$, where $W[i]$ is defined as

$$W[i] = \sum_{1 \le j \le i} \pi_{\mathbf{v}_j}.$$

At each step, IWVS generates a random number τ drawn uniformly from the range (0,1), and then samples the node $\mathbf{v}_i$ whose ID i satisfies $W[i] \le \tau < W[i+1]$. Then we easily find that the probability of sampling $\mathbf{v}_j$ is $\pi_{\mathbf{v}_j}$. $\mathbf{v}_i$ 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 [Chib and Greenberg 1995; Hastings 1970; Metropolis et al. 2011]. This method does not require the complete graph topology, and reduces the memory space used for storing the 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 a 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

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

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 $O(1)$.

3.2. Sampling Node Pairs From S and $\mathrm{S}^{(1)}$

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

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

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

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

and the variance is

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

Then we have

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

Denote by d_u the degree of a node $u \in V$. To sample a node pair from $\mathbf{S}^{(1)}$, we select a random node u according to the probability distribution $(\pi_u^{(1)} : u \in V)$ using IWVS or MHWVS, where $\pi_u^{(1)}$ is defined as

$$\pi_u^{(1)} = \frac{d_u}{2|E|}.$$

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

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

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 the graph G. Similar to the derivation of $\hat{\omega}_k$, we have

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

3.3. Sampling Node Pairs From $\mathrm{S}^{(2)}$

To sample a node pair from $\mathbf{S}^{(2)}$ at random, we first select a random node $x \in V$ with degree greater than two according to the 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 $\frac{1}{M}$. Denote $m(u, v)$ as the number of mutual neighbors of u and v. Then a node pair $[u, v]$ in $\mathbf{S}^{(2)}$ is sampled with probability

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

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

$$\hat{\omega}^{(2)}_k = \frac{1}{H} \sum_{i=1}^{n} \frac{\mathbf{1}(F(u_i, v_i) = a_k)}{m(u_i, v_i)}, \tag{3}$$

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 the node pairs in $\mathbf{S}^{(2)}$. The accuracy of $\hat{\omega}^{(2)}_k$ can be stated by the following theorem.

THEOREM 3.1. *$\hat{\omega}^{(2)}_k$ $(1 \leq k \leq K)$ is an unbiased estimator of $\omega^{(2)}_k$. When $[u_i, v_i]$ $(1 \leq i \leq n)$ are sampled independently using IWVS, we have*

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

where $0 < \epsilon < 1$.

The proofs of all theorems in this paper are given in Appendix.

4. NODE PAIR SAMPLING 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 the graph G is connected. Instead, we study the use of a random walk (RW) as a node pair sampling technique. RWs have been extensively studied in the graph theory literature [Lovász 1993]. 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 RW being at node v converges to π_v [Lovász 1993]. 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 [Heckathorn 2002; Salganik and Heckathorn 2004]. Compared to UVS, RW exhibits smaller estimation errors for characteristics associated with high degree nodes.

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

We use two independent RWs to sample node pairs $[u_i, v_i]$ $(1 \leq i \leq n)$ randomly from $\mathbf{S}$, where u_i and v_i are nodes sampled from the graph G by these two RWs at the i-th step repectively. 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 $G^{(2)}$ has $d_u d_v$ neighbors. When G is a connected and non-bipartite graph, we can easily show that $G^{(2)}$ is also connected and non-bipartite. Then a 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 4.1. *When the graph G is connected and non-bipartite, $\hat{\omega}_k^{(\star)}$ $(1 \le k \le K)$ is a consistent estimator of ω_k, i.e., $\lim_{n\to\infty} \hat{\omega}_k^{(\star)} \xrightarrow{a.s.} \omega_k$, where "a.s." denotes "almost sure" converge, i.e., the event happens with probability one.*

To estimate statistics of node pairs in $\mathbf{S}^{(1)}$, we sample node pars $[u_i, v_i]$ $(1 \le i \le n)$ by applying a RW over G, where u_i and v_i are nodes sampled by the 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 [Ribeiro and Towsley 2010], 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 4.2. *When the graph G is connected and non-bipartite, $\hat{\omega}_k^{(1\star)}$ $(1 \le k \le K)$ is a consistent estimator of $\omega_k^{(1)}$, i.e., $\lim_{n\to\infty} \hat{\omega}_k^{(1\star)} \xrightarrow{a.s.} \omega_k^{(1)}$.*

4.2. 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 a graph $\hat{G} = (\hat{V}, \hat{E})$, with node set $\hat{V} = \{(u, v) : (u, v) \in E\}$, edge set $\hat{E} = \{((u, v), (u, v')) : (u, v) \in E, (u, v') \in E, v \neq v'\}$. $\hat{G}$ is similar to the line graph proposed in [Kang et al. 2011], which is used for calculating node centralities. 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 the 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 $P^{\text{NRW}} = [P_{e,e'}^{\text{NRW}}]$, where $e = (u, v)$ and $e' = (u', v')$ are edges in E, and $P_{e,e'}^{\text{NRW}}$ is defined as the probability that e' is selected as the next-hop edge given that its current edge e. $P_{e,e'}^{\text{NRW}}$ is computed as $P_{(u,v),(u',v')}^{\text{NRW}} = \frac{1}{d_u+d_v-2}$ if $(u', v') \in N_{(u,v)}$ and $(u, v) \in E$, otherwise $P_{(u,v),(u',v')}^{\text{NRW}} = 0$. We can easily show that a node (u, v) (an edge in G) in $\hat{G}$ connects to $d_u + d_v - 2$ nodes in $\hat{G}$, its degree in $\hat{G}$ is $d_u + d_v - 2$. Meanwhile $\hat{G}$ has $|\hat{E}| = M/2$ edges, where $M = \sum_{y\in V} d_y(d_y - 1)$. Then we have

THEOREM 4.3. *When the graph G is connected and non-bipartite, NRW exhibits a stationary distribution $\boldsymbol{\pi}_E = (\pi_{(u,v)} : (u, v) \in E)$, where $\pi_{(u,v)}$ is*

$$\pi_{(u,v)} = \frac{d_u + d_v - 2}{M}, \quad (u, v) \in E. \tag{4}$$

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 the i-th $(i \ge 0)$ visited edge and node. For each step i, the next visited edge (x_{i+1}, y_{i+1}) is selected from $N_{(x_i,y_i)}$ at random, which has exactly one common node with current edge (x_i, y_i). Clearly the common node is x_i with probability $\frac{d_{x_i}-1}{d_{x_i}+d_{y_i}-2}$, or y_i with probability $\frac{d_{y_i}-1}{d_{x_i}+d_{y_i}-2}$. By excluding the 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 the 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 u and v in the 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 [Lovász 1993], therefore a node pair (u, v) is sampled by NRW with a stationary probability $\frac{m(u,v)}{M}$. Based on sampled pairs $[u_i, v_i]$ $(1 \le i \le n)$, we estimate $\omega_k^{(2)}$ $(1 \le k \le K)$ as follows

$$\hat{\omega}_k^{(2*)} = \frac{1}{H} \sum_{i=1}^{n} \frac{\mathbf{1}(F(u_i, v_i) = a_k)}{m(u_i, v_i)}, \tag{5}$$

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

THEOREM 4.4. *When the graph G is connected and non-bipartite, $\hat{\omega}_k^{(2\star)}$ $(1 \le k \le K)$ is a consistent estimator of $\omega_k^{(2)}$, i.e., $\lim_{n\to\infty} \hat{\omega}_k^{(2\star)} \xrightarrow{a.s.} \omega_k^{(2)}$.*

Next, we propose a better sampling method than NRW. [Lee et al. 2012] reveals that duplicated samples generated by a RW will cause estimation errors. Clearly the NRW might generate duplicated edge samples. To reduce the error of $\hat{\omega}_k^{(2*)}$ induced by the temporal correlation over random samples generated by a NRW, we present an avoid backtracking NRW (ABNRW) method, which uses the avoid backtracking method proposed in [Lee et al. 2012]. The ABNRW works as follows: At each step $i \ge 2$, it first computes $|N_{(u,v)}| = d_u + d_v - 2$. If $|N_{(u,v)}| \ge 2$, it selects an edge randomly from $N_{(u,v)} \backslash \{(x_{i-1}, y_{i-1})\}$ as the next sampled edge (x_{i+1}, y_{i+1}). Otherwise the ABNRW moves to the previous sampled edge, i.e., $(x_{i+1}, y_{i+1}) = (x_{i-1}, y_{i-1})$. From [Lee et al. 2012], we can easily show that the ABNRW samples “nodes” (i.e., edges in the graph G) in the graph $\hat{G}$ with the stationary distribution $\boldsymbol{\pi}_E = (\pi_{(u,v)} : (u, v) \in E)$ defined in (4). This is the same as NRW. Then ABNRW generates node pairs based on sampled edges, which is the same as NRW. The pseudo-code for the ABNRW based node pair sampling algorithm is depicted in Algorithm 2. Finally we use the same estimator (5) to measure the concentration $\omega_k^{(2)}$ $(1 \le k \le K)$ based on node pairs sampled by ABNRW. It is important to point out that the estimator given in (5) is also consistent for ABNRW.

5. DATA EVALUATION

Our simulation experiments are performed over a variety of real world graphs, which are summarized in Table I. 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 nodes u to v represents that u voted on 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.

We introduce the error metric used to compare the different sampling methods. Mean square error (MSE) is a common measure to quantify the error of an estimate $\hat{\omega}$ with respect to its true value $\omega > 0$. It is defined as $\text{MSE}(\hat{\omega}) = \text{E}[(\hat{\omega} - \omega)^2] = \text{var}(\hat{\omega}) + (\text{E}[\hat{\omega}] - \omega)^2$. We can see that $\text{MSE}(\hat{\omega})$ decomposes into a sum of the variance and bias of the estimator $\hat{\omega}$, both quantities are important and need to be as small as possible to achieve good estimation performance. When $\hat{\omega}$ is an

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

unbiased estimator of ω, then $\text{MSE}(\hat{\omega}) = \text{var}(\hat{\omega})$. In our experiments, we study the normalized root mean square error (NRMSE) to measure the relative error of the estimator $\hat{\omega}_k$ of ω_k, $k = 1, 2, \ldots$. $\text{NRMSE}(\hat{\omega}_k)$ is defined as:

$$\text{NRMSE}(\hat{\omega}_k) = \frac{\sqrt{\text{MSE}(\hat{\omega}_k)}}{\omega_k}, \qquad k = 1, 2, \ldots.$$

When $\hat{\omega}_k$ is an unbiased estimator of ω_k, then $\text{NRMSE}(\hat{\omega}_k)$ is equivalent to the normalized standard error of $\hat{\omega}_k$, i.e., $\text{NRMSE}(\hat{\omega}_k) = \sqrt{\text{var}(\hat{\omega}_k)}/\omega_k$. Note that our metric uses the relative error. Thus, when ω_k is small, we consider values as large as $\text{NRMSE}(\hat{\omega}_k) = 1$ to be acceptable. In all our experiments, we average the estimates and calculate their NRMSEs over 1,000 runs.

5.1. Distance Distribution

We evaluate the performance of UVS for estimating $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$, the distance distribution of the 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 P2P-Gnutella. Fig. 1 presents $\text{NRMSE}(\hat{\omega}_k)$ $(1 \leq k \leq K)$ for sampling budgets $B = \{0.001|\mathbf{S}|, 0.005|\mathbf{S}|, 0.01|\mathbf{S}|\}$. When $B \geq 0.05|\mathbf{S}|$, the $\text{NRMSE}(\hat{\omega}_k)$ is always smaller than one. On average, the ratios of $\text{NRMSE}(\hat{\omega}_k)|_{B=0.005|\mathbf{S}|}$ to $\text{NRMSE}(\hat{\omega}_k)|_{B=0.001|\mathbf{S}|}$ are 0.436 and 0.440 for Wiki-vote and P2P-Gnutella graphs respectively, the ratios of $\text{NRMSE}(\hat{\omega}_k)|_{B=0.01|\mathbf{S}|}$ to $\text{NRMSE}(\hat{\omega}_k)|_{B=0.005|\mathbf{S}|}$ are 0.709 and 0.719 for

Algorithm 2: ABNRW pseudo-code.

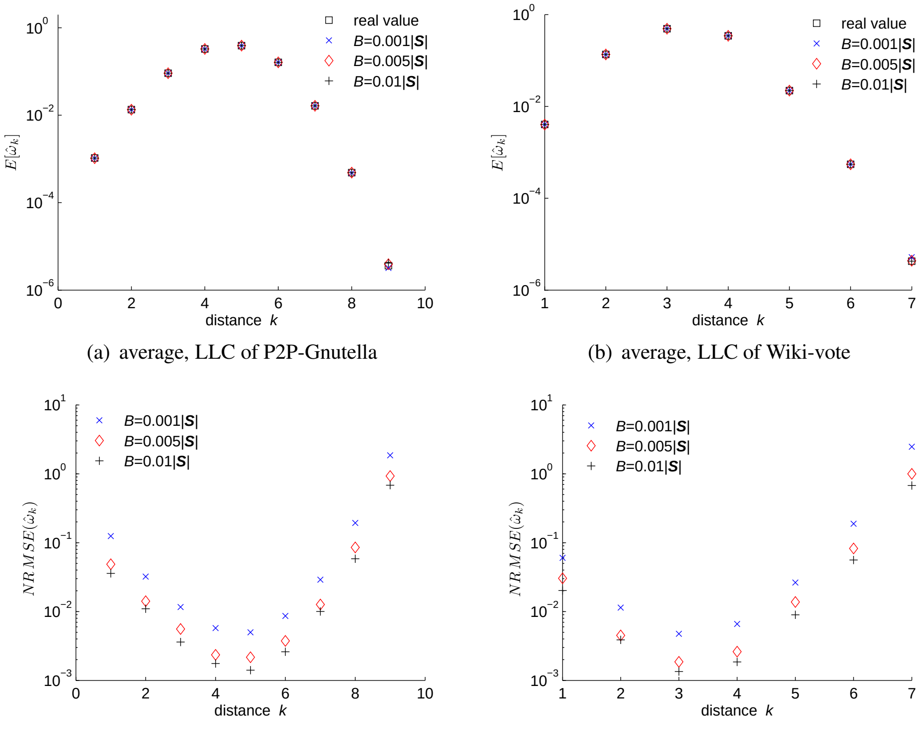
```
/* $n$ is the sampling budget, $(x_0, y_0)$ is the initial edge, and $(x_i, y_i)$ and $s_i$ are the
   visited edge and node at the $i$-th step.                                                       */
input  : $n$ and $(x_0, y_0) \in E$
output: node pairs $[u_1, v_1], [u_2, v_2], \ldots, [u_n, v_n]$

$i \leftarrow 0$;
while $i <= n$ do
    /* $d_x$ is the degree of a node $x$ in $G$                                                    */
    if $d_{x_i} + d_{y_i} = 3$ then
        $x_{i+1} \leftarrow x_i$ and $y_{i+1} \leftarrow y_i$;
    else
        /* $c$ records the common node between two edges $(x_{i-1}, y_{i-1})$ and $(x_i, y_i)$. $w$
           records the node differing $(x_{i-1}, y_{i-1})$ from $(x_i, y_i)$. $c$ and $w$ are set to
           null when $i = 0$.                                                                      */
        $c \leftarrow \{x_{i-1}, y_{i-1}\} \cap \{x_i, y_i\}$;
        $w \leftarrow \{x_{i-1}, y_{i-1}\} - \{x_i, y_i\}$;
        /* $U(0,1)$ is a uniform $(0,1)$ random sample.                                            */
        Generate $p \leftarrow U(0,1)$;
        if $p < \frac{d_{x_i} - 1 - \mathbf{1}(c = x_i)}{d_{x_i} + d_{y_i} - 3}$ then
            /* randomNeighbor$(x, Y)$ returns a node selected randomly from the
               neighbors of the node $x$ excluding nodes in the set $Y$.                            */
            if $c = x_i$ then
                $s_i \leftarrow$ randomNeighbor$(x_i, \{y_i, w\})$;
            else
                $s_i \leftarrow$ randomNeighbor$(x_i, \{y_i\})$;
            end
            $x_{i+1} \leftarrow x_i$ and $y_{i+1} \leftarrow s_i$;
            /* $u$ and $v$ are the nodes in two sequentially visited edges $(x_i, y_i)$ and
               $(x_{i+1}, y_{i+1})$ excluding their common node.                                    */
            $u \leftarrow y_i$ and $v \leftarrow s_i$;
        else
            if $c = y_i$ then
                $s_i \leftarrow$ randomNeighbor$(y_i, \{x_i, w\})$;
            else
                $s_i \leftarrow$ randomNeighbor$(y_i, \{x_i\})$;
            end
            $x_{i+1} \leftarrow s_i$ and $y_{i+1} \leftarrow y_i$;
            $u \leftarrow x_i$ and $v \leftarrow s_i$;
        end
    end
    Generate $q \leftarrow U(0,1)$;
    if $q < 0.5$ then
        $u_{i+1} \leftarrow u$ and $v_{i+1} \leftarrow v$;
    else
        $u_{i+1} \leftarrow v$ and $v_{i+1} \leftarrow u$;
    end
    $i \leftarrow i + 1$;
end
```

Table I. Overview of Graph Datasets Used in Our Simulations. ”LCC” refers to the largest connected component in the undirected graph generated by ignoring the directions of edges.

Graph	Entire Graph		LCC	
	nodes	edges	nodes	edges
Wiki-vote [Leskovec et al. 2010a; 2010b]	7,115	103,689	7,066	103,663
P2P-Gnutella [Ripeanu et al. 2002]	6,301	20,777	6,299	20,776
soc-Epinions [Richardson et al. 2003]	75,879	508,837	75,877	508,836
soc-Slashdot [Leskovec et al. 2009]	77,360	905,468	77,360	905,468

Wiki-vote and P2P-Gnutella graphs respectively, and the ratios of $\text{NRMSE}(\hat{\omega}_k)|_{B=0.01|\mathbf{S}|}$ to $\text{NRMSE}(\hat{\omega}_k)|_{B=0.001|\mathbf{S}|}$ are 0.307 and 0.316 for Wiki-vote and P2P-Gnutella graphs respectively. From these results, we observe that the error of sampling B node pairs from $\mathbf{S}$ is roughly proportional to $1/\sqrt{B}$.



(a) average, LLC of P2P-Gnutella

(b) average, LLC of Wiki-vote

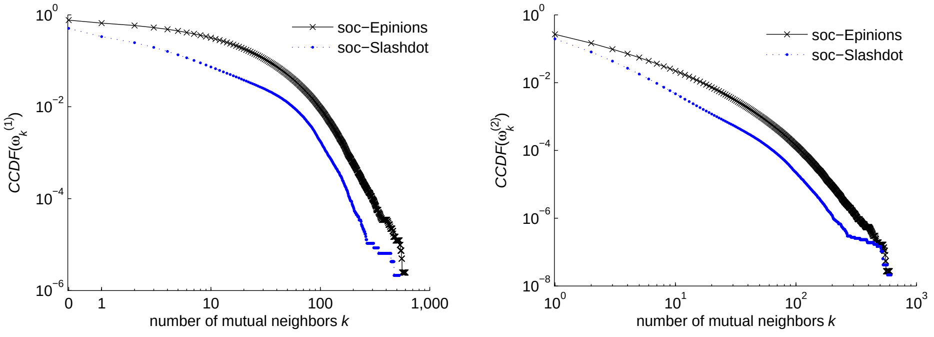
(c) NRMSE, LLC of P2P-Gnutella

(d) NRMSE, LLC of Wiki-vote

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

5.2. 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 [Shi et al. 2007]. Define $\omega_k^{(1)}$ and $\omega_k^{(2)}$ as the fraction of node pairs with $k \geq 1$ 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 the graphs soc-Epinions and soc-Slashdot. The sizes of $\mathbf{S}^{(2)}$ are 7.34×10^7 and 9.49×10^7 for soc-Epinions and soc-Slashdot respectively. The statistics for the LCCs of soc-Epinions and soc-Slashdot are similar.



(a) $\mathbf{S}^{(1)}$

(b) $\mathbf{S}^{(2)}$

Fig. 2. (soc-Epinions and soc-Slashdot) CCDF of the distributions of the node pairs in $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ by the mutual neighbor count.

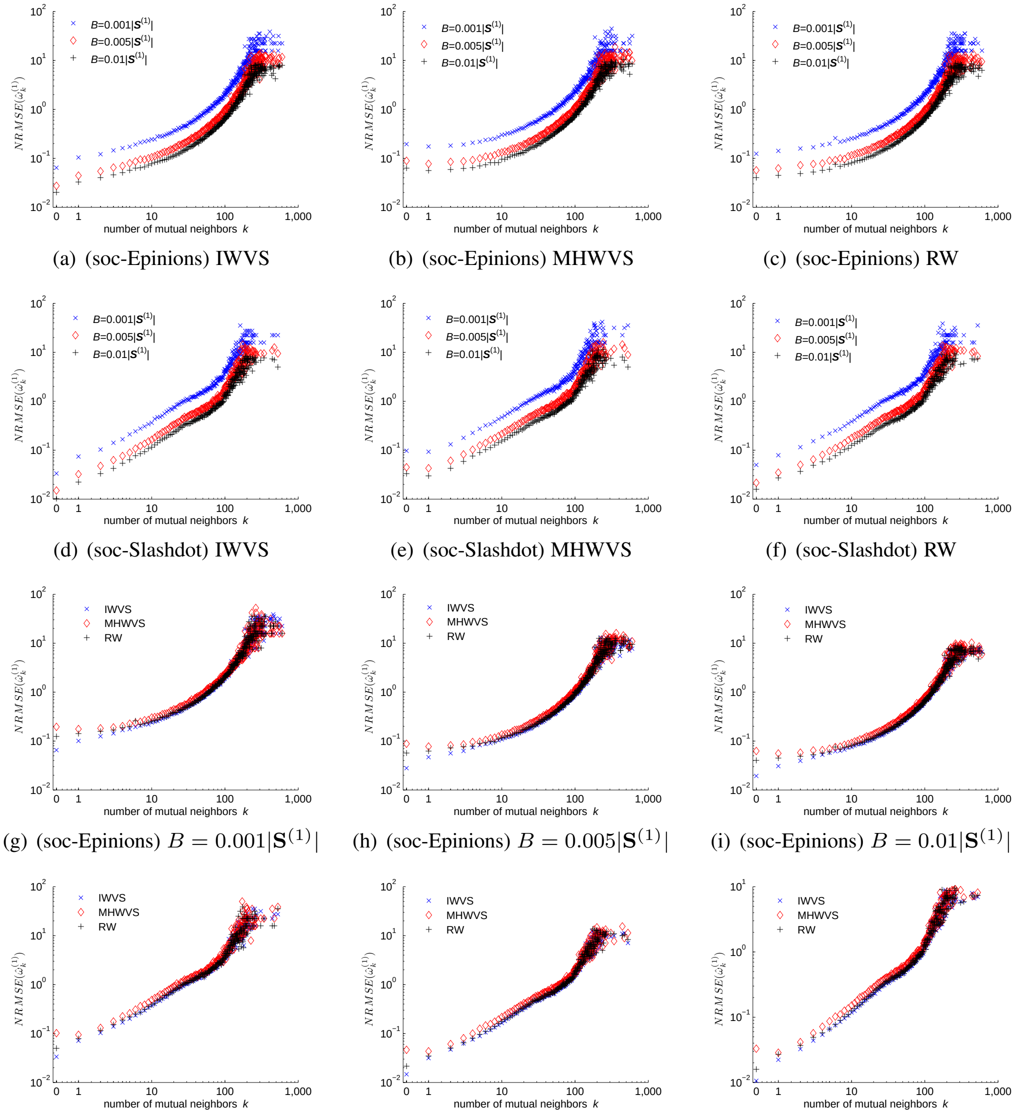
For $\mathbf{S}^{(1)}$, we evaluate the performance of sampling methods: IWVS and MHWVS presented in Section 3.2, and RW presented in Section 4.1 using the graphs soc-Epinions and soc-Slashdot. Figs. 3 (a)–(f) present NRMSE($\hat{\omega}_k^{(1)}$) for sampling budgets $B = \{0.001|\mathbf{S}^{(1)}|, 0.005|\mathbf{S}^{(1)}|, 0.01|\mathbf{S}^{(1)}|\}$. We find that the error of sampling B node pairs from $\mathbf{S}^{(1)}$ is roughly proportional to $1/\sqrt{B}$ for each method. Figs. 3 (g)–(l) compare the NRMSEs 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.

For $\mathbf{S}^{(2)}$, we evaluate the performance of the following methods: IWVS and MHWVS presented in Section 3.3, and NRW presented in Section 4.2 using the graphs soc-Epinions and soc-Slashdot. Figs. 4 (a)–(f) present NRMSE($\hat{\omega}_k^{(2)}$) 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 NRMSE($\omega_k^{(2)}$) are smaller than one for each sampling method. Figs. 4 (g)–(l) compare the NRMSEs of three sampling methods under the same sampling budget. It shows that 1) NRW, ABNRW, and IWVS have much smaller errors than MHWVS; 2) ABNRW is more accurate than NRW; 3) ABNRW almost exhibits the same accuracy as IWVS.

5.3. 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 schemes (IDSs) to distribute interests over a graph:

- IDS I: It distributes each interest to a node independently selected from the graph at random.
- IDS II: 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 the node v and the closest $k-1$ nodes connected to v.
- IDS III: It distributes interests over undirected graphs using independent cascade model [Goldenberg et al. 2001]. First distribute an interest i to a random selected node v. Then distribute i to the other nodes from v iteratively. When a new node first receives interest i, it is given a single chance to distribute i to each of its neighbors currently without i with a probability p_S.



(a) (soc-Epinions) IWVS

(b) (soc-Epinions) MHWVS

(c) (soc-Epinions) RW

(d) (soc-Slashdot) IWVS

(e) (soc-Slashdot) MHWVS

(f) (soc-Slashdot) RW

(g) (soc-Epinions) $B = 0.001|\mathbf{S}^{(1)}|$

(h) (soc-Epinions) $B = 0.005|\mathbf{S}^{(1)}|$

(i) (soc-Epinions) $B = 0.01|\mathbf{S}^{(1)}|$

(j) (soc-Slashdot) $B = 0.001|\mathbf{S}^{(1)}|$

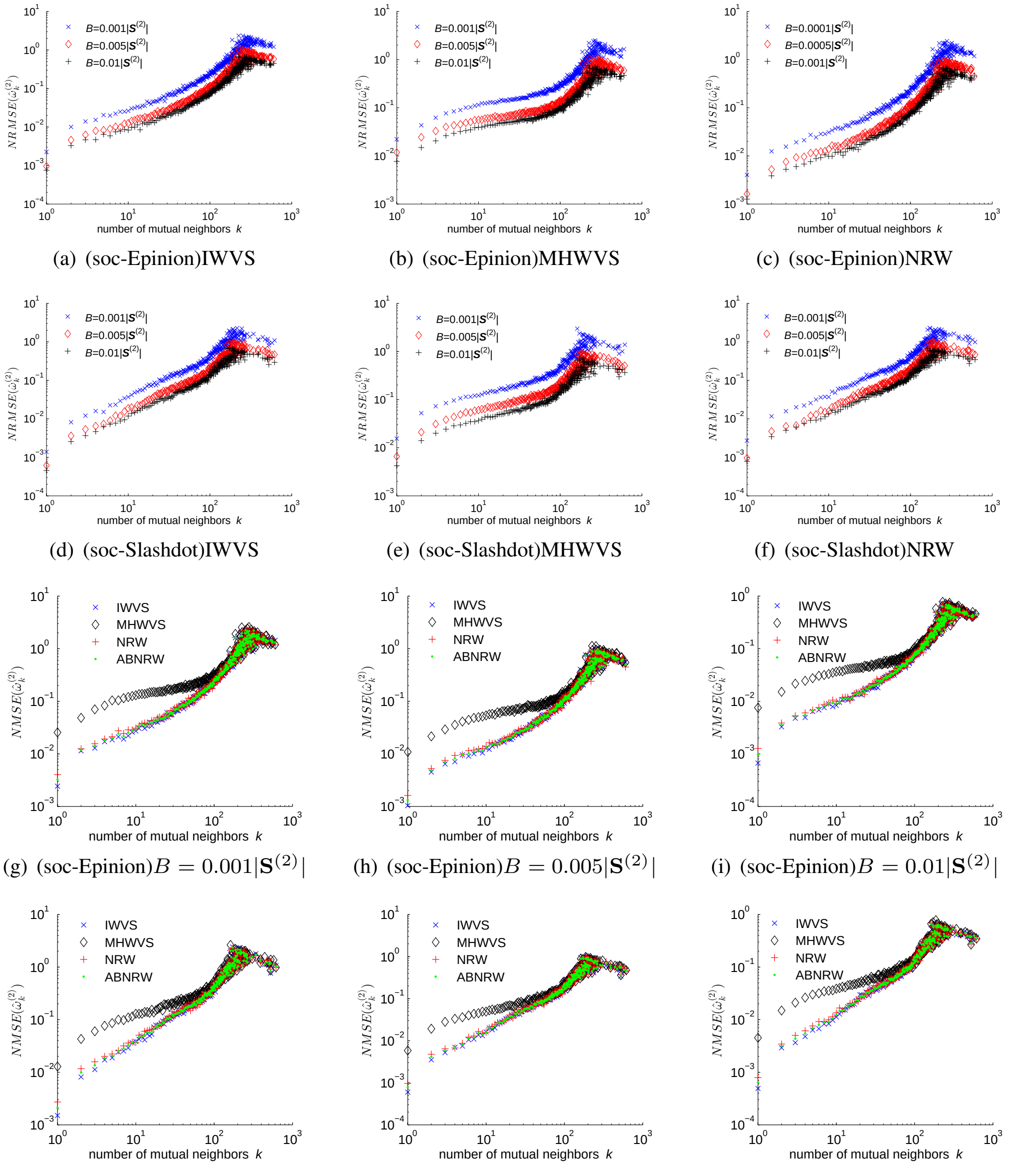
(k) (soc-Slashdot) $B = 0.005|\mathbf{S}^{(1)}|$

(l) (soc-Slashdot) $B = 0.01|\mathbf{S}^{(1)}|$

Fig. 3. (soc-Epinions and soc-Slashdot) NRMSE of distribution estimates of the node pairs in $\mathbf{S}^{(1)}$ by the mutual neighbor count.

IDS I models the scenario where interests are distributed independently with the graph topology. Unlike IDS I, IDS II and IDS III are used to model the scenario where interests are spreading over undirected and directed graphs respectively.

Define the truncated Pareto distribution as $\theta_k = \frac{\alpha}{\gamma k^{\alpha+1}}$, $k = 1, \dots, W$, where $\alpha > 0$ and $\gamma = \sum_{k=1}^{W} \frac{\alpha}{k^{\alpha+1}}$. In the following experiments, we generate 10^5 distinct interests for IDS I and IDS II, and for each interest the number of nodes possessed it is a random variable selected from $\{1, \dots, W\}$ according to the truncated Pareto distribution with $\alpha = 1$ and $W = 10^3$. For CDS III, we generate 10^4 distinct interests and set $p_S = 0.01$. The graphs used are the LCCs of P2P-



(a) (soc-Epinion)IWVS (b) (soc-Epinion)MHWVS (c) (soc-Epinion)NRW

(d) (soc-Slashdot)IWVS (e) (soc-Slashdot)MHWVS (f) (soc-Slashdot)NRW

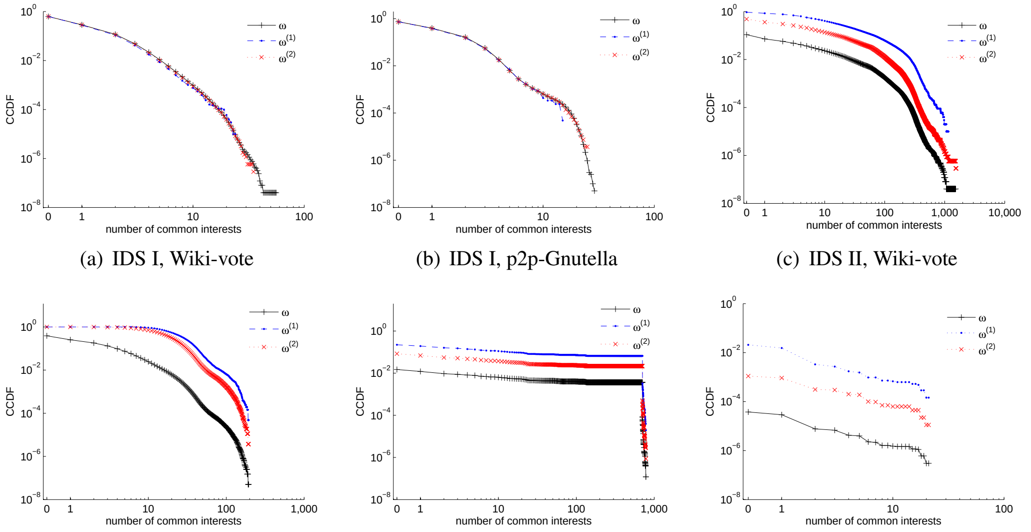
(g) (soc-Epinion)$B = 0.001|\mathbf{S}^{(2)}|$ (h) (soc-Epinion)$B = 0.005|\mathbf{S}^{(2)}|$ (i) (soc-Epinion)$B = 0.01|\mathbf{S}^{(2)}|$

(j) (soc-Slashdot)$B = 0.001|\mathbf{S}^{(2)}|$ (k) (soc-Slashdot)$B = 0.005|\mathbf{S}^{(2)}|$ (l) (soc-Slashdot)$B = 0.01|\mathbf{S}^{(2)}|$

Fig. 4. (soc-Epinion and soc-Slashdot) NRMSE of distribution estimates of the node pairs in $\mathbf{S}^{(2)}$ by the mutual neighbor count.

Gnutella and Wiki-vote, where the sizes of $\mathbf{S}^{(2)}$ are 2.69×10^5 and 3.46×10^6 respectively. Define ω_k, $\omega_k^{(1)}$ and $\omega_k^{(2)}$ as the fraction of the node pairs with $k \geq 1$ common interests in $\mathbf{S}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$ respectively. Fig. 5 shows the CCDFs of $\boldsymbol{\omega}$, $\boldsymbol{\omega}^{(1)}$, and $\boldsymbol{\omega}^{(2)}$ generated by our simulations.

Figure 6 shows NRMSEs of sampling methods for the set $\mathbf{S}$ under the same number of sampled pairs $B = 0.01|\mathbf{S}|$. Results for IDS II 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 RW is biased to



(a) IDS I, Wiki-vote (b) IDS I, p2p-Gnutella (c) IDS II, Wiki-vote

(d) IDS II, p2p-Gnutella (e) IDS III, Wiki-vote (f) IDS III, p2p-Gnutella

Fig. 5. Distributions of the node pairs in $\mathbf{S}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$ by the common interest count.

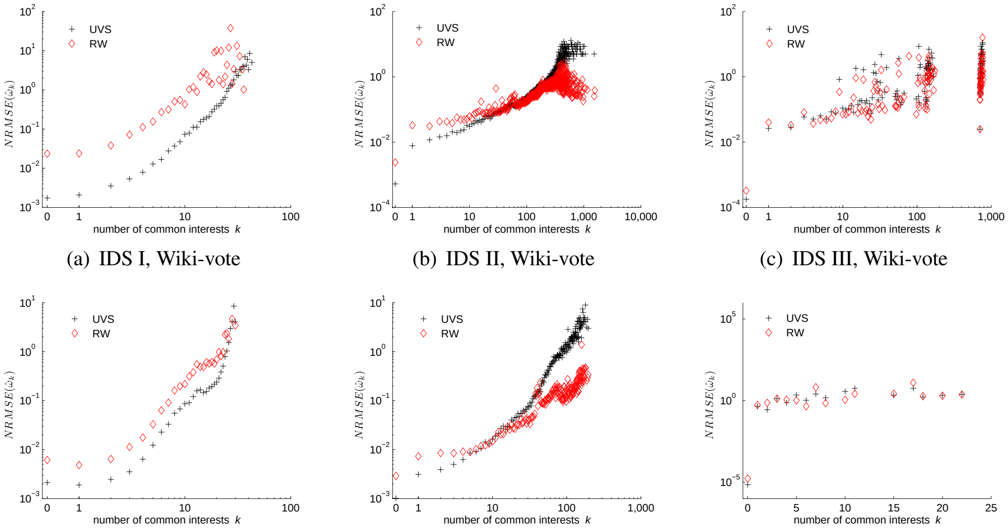
sample high degree nodes, and IDS II generates more interests for high degree nodes than nodes with small degrees. It is similar to the observation for estimating degree distribution using RW and UVS [Ribeiro and Towsley 2010]. Figs. 7 shows NRMSEs of sampling methods for the set $\mathbf{S}^{(1)}$ under the same number of sampled pairs $B = 0.05|\mathbf{S}^{(1)}|$. We find that IWVS, MHWVS, and RW almost have the same accuracy. Figs. 8 shows NRMSEs of sampling methods for the set $\mathbf{S}^{(2)}$ under the same number of sampled pairs $B = 0.01|\mathbf{S}^{(2)}|$. We can see that IWVS, MHWVS, and ABNRW almost have the same accuracy for IDS I and IDS III. For IDS II, our results show that IWVS has the smallest errors for estimating $\omega_k^{(2)}$ with small k, ABNRW has much smaller errors for estimating $\omega_k^{(2)}$ with small k than MHWVS for the graph Wiki-vote, and ABNRW almost exhibits the same accuracy as MHWVS for the graph P2P-Gnutella.

6. APPLICATIONS

In this section, we conduct real experiments on Foursquare and two popular Chinese OSNs: Douban and Xiami.

6.1. Real Experiments on Foursquare

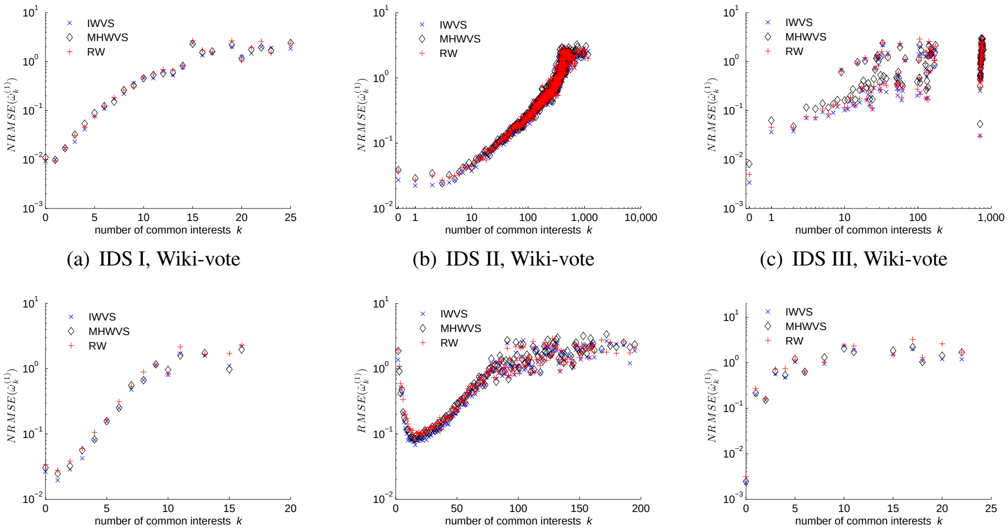
We sampled $22,247$ users and $22,500$ edges from Foursquare using RWs and a NRW respectively. For a sampled Foursquare user, we collected his/her friends, home location (e.g., city and country), and tips on venues (i.e., places like coffee shops, restaurants, shopping malls). Based on these samples, we compare the similarities of locations and interested venues of a pair of users selected from $\mathbf{S}^{(1)}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$. The result is shown in Figure 9. Compared to a pair of users selected from $\mathbf{S}$, the probability of two friends living in the same city or country is 31 times larger, and the probability of two friends checking in the same place is 113 times larger. Moreover, when two users share a common friend, the probability of they living in the same city/country increases to 17 times larger, and the probability of they checking in the same place increases to 64 times larger in comparison with a pair of users selected from $\mathbf{S}$.



(a) IDS I, Wiki-vote (b) IDS II, Wiki-vote (c) IDS III, Wiki-vote

(d) IDS I, p2p-Gnutella (e) IDS II, p2p-Gnutella (f) IDS III, p2p-Gnutella

Fig. 6. (Wiki-vote and p2p-Gnutella) Compared NRMSE of distribution estimates of the node pairs in $\mathbf{S}$ by the common interest count for different methods.



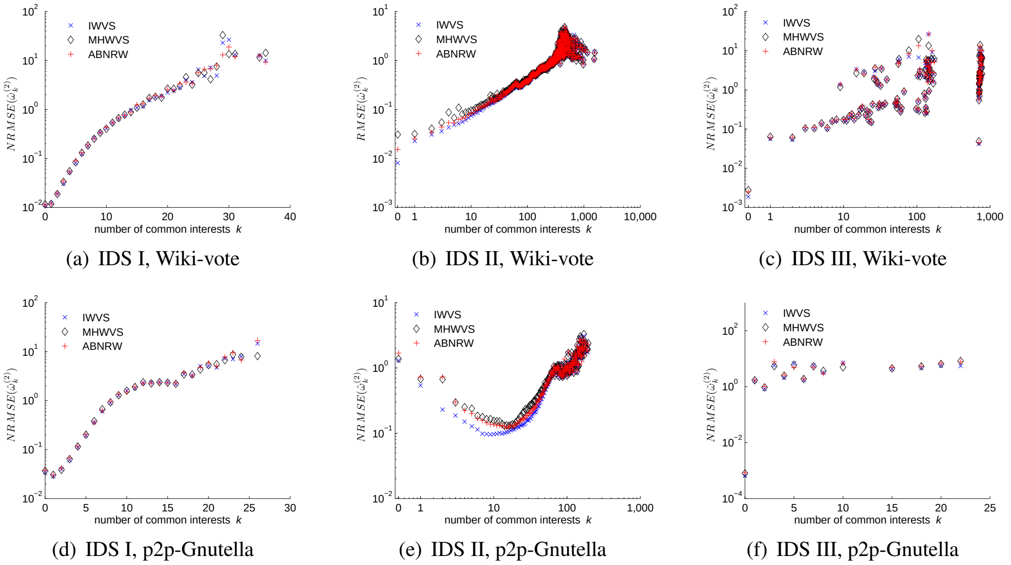
(a) IDS I, Wiki-vote (b) IDS II, Wiki-vote (c) IDS III, Wiki-vote

(d) IDS I, p2p-Gnutella (e) IDS II, p2p-Gnutella (f) IDS III, p2p-Gnutella

Fig. 7. (Wiki-vote and p2p-Gnutella) Compared NRMSE of distribution estimates of the node pairs in $\mathbf{S}^{(1)}$ by the common interest count for different methods.

6.2. Real experiments on 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 [Zhao et al. 2011]. Each user of Douban 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 [Wang et al. 2012]. Each user of Xiami maintains a list



(a) IDS I, Wiki-vote (b) IDS II, Wiki-vote (c) IDS III, Wiki-vote

(d) IDS I, p2p-Gnutella (e) IDS II, p2p-Gnutella (f) IDS III, p2p-Gnutella

Fig. 8. (Wiki-vote and p2p-Gnutella) Compared NRMSE of distribution estimates of the node pairs in $\mathbf{S}^{(2)}$ by the common interest count for different methods.

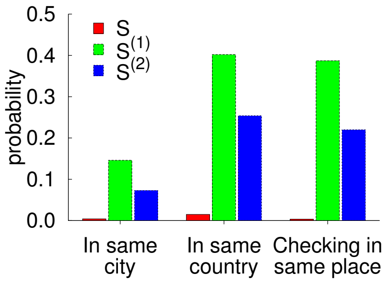
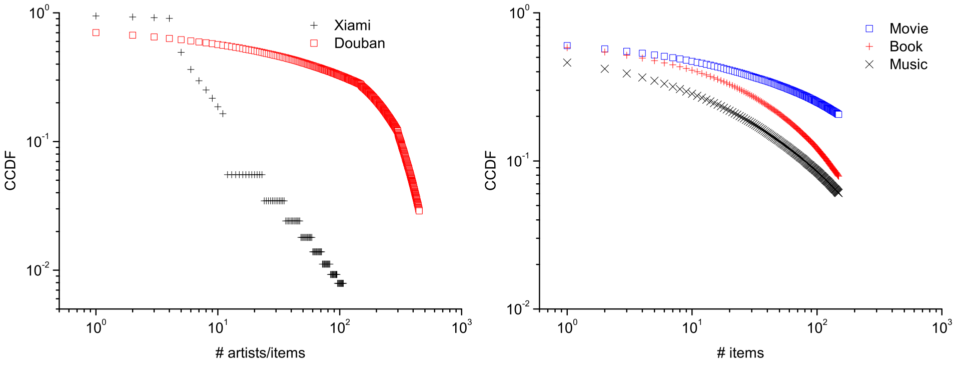


Fig. 9. The similarities of locations and interested venues of a pair of users selected from $\mathbf{S}^{(1)}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$.

of his/her favorite artists. Fig. 10 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 the set $\mathbf{S}$, 105,736 Xiami user pairs and 85,631 Douban user pairs from the set $\mathbf{S}^{(1)}$, and 359,522 Xiami user pairs and 96,361 Douban user pairs from the set $\mathbf{S}^{(2)}$. As shown in Fig. 11, 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. 12. This indicates that users in Xiami and Douban tend to connect to ones with the similar interests.

7. RELATED WORK

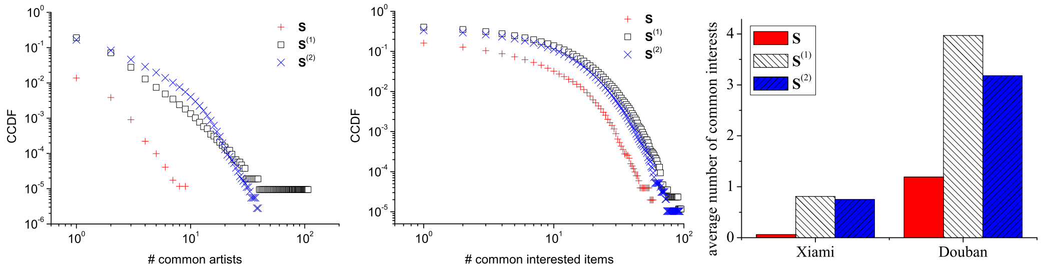
Let us provide a brief summary on related work. Singla et al. [Singla and Richardson 2008] reveal that significant homophily is present in the MSN Messenger network based on the study of user pairs' similarities in terms of their Web search topics, and personal characteristics such as their



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



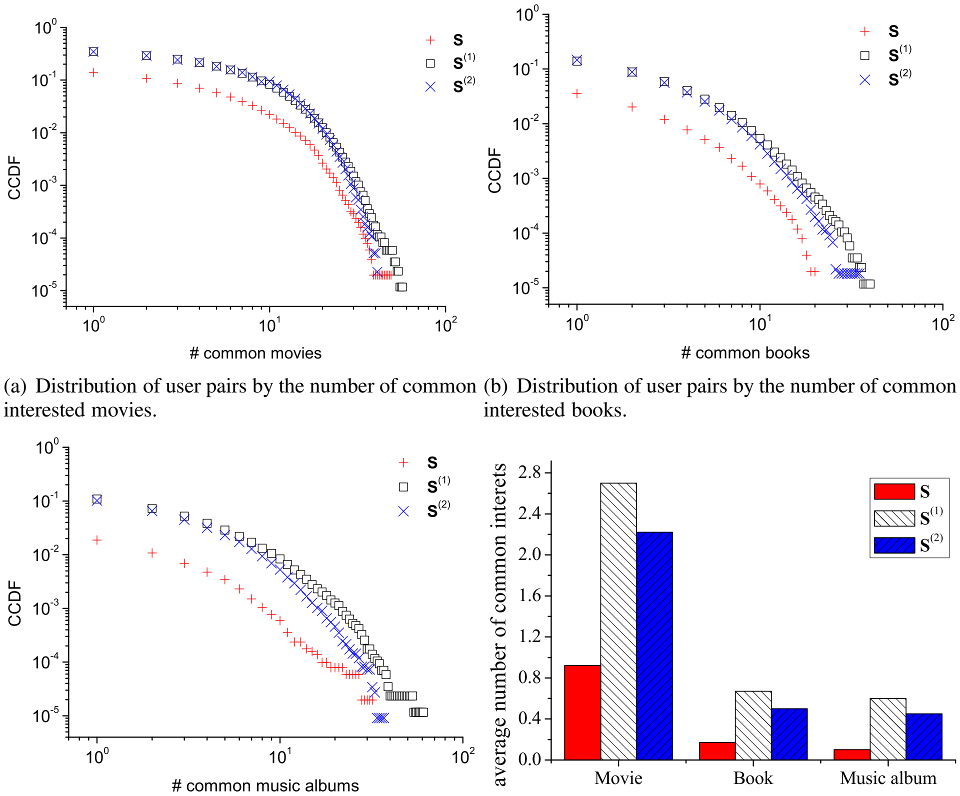
(a) Xiami, distribution of user pairs by the number of common interested artists.

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

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

Fig. 11. (Xiami and Douban) Statistics of users' common interests.

ages and locations. Similar results also are found in [Leskovec and Horvitz 2008]. There are also works on measuring the distance statistics of user pairs in OSNs [Leskovec et al. 2005; Leskovec and Horvitz 2008; Kwak et al. 2010]. Leskovec et al. [Leskovec et al. 2005] 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 the node degree distribution [Stutzbach et al. 2009; Rasti et al. 2009; Gjoka et al. 2010; Ribeiro and Towsley 2010; Ribeiro et al. 2012] and the topology of nodes' groups [Kurant et al. 2011b]. These sampling methods have been widely applied to characterize complex networks, such as P2P networks [Massoulié et al. 2006; Gkantsidis et al. 2006; Stutzbach et al. 2009; Rasti et al. 2009], and OSNs [Mislove et al. 2007; Ahn et al. 2007; Gjoka et al. 2010; Gjoka et al. 2011; Kurant et al. 2011a]. Leskovec and Faloutsos [Leskovec and Faloutsos 2006] conducted simulations on several real networks to study relations between characteristics of the original graph and the subgraph generated by different sampling methods. 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 [Achlioptas et al. 2005; Kurant et al. 2010; 2011]. RW is biased to sample high degree nodes, however its bias is known and can be corrected for [Heckathorn 2002; Salganik and Heckathorn 2004]. Compared to UVS, RW has smaller estimation errors for the characteristics of high degree nodes, especially for networks where UVS is costly (e.g., Flickr, Facebook, and MySpace) [Ribeiro and Towsley 2010]. Compared to RW that reweights sampled values



(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. 12. Statistics of users' common interested movies, books, and music albums in Douban.

to obtain an unbiased estimate of graph characterizes, Metropolis-Hasting RW (MHRW) [Zhong and Shen 2006; Stutzbach et al. 2009; Gjoka et al. 2010] modifies the RW procedure using the Metropolis-Hasting technique, which aims to sample each node uniformly. The accuracy of RW and MHRW is compared in [Rasti et al. 2009; Gjoka et al. 2010], 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 [Mohaisen et al. 2010] for many OSNs. There are a lot of work to decrease the mixing time [Boyd et al. 2004; Ribeiro and Towsley 2010; Avrachenkov et al. 2010; Kurant et al. 2011a; Gjoka et al. 2011]. 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.

8. CONCLUSIONS

In this work we systemically study the problem of estimating characteristics of the node pairs in $\mathbf{S}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$ for ones with/witout the complete graph topology. We propose two kinds of sampling methods based on UVS and RW techniques, and prove that they are consistent estimators. Our simulation results show that RW based methods and UVS based methods almost have the similar accuracy, especially for the sampling methods for $\mathbf{S}^{(1)}$. Finally we apply our methods to Foursquare, Douban and Xiami OSNs, and discover that there is a strong tendency for users to connect to others with common interests.

Appendix

LEMMA 8.1. *[Roberts and Rosenthal 2004; Jones 2004; Lee et al. 2012] Given that the undirected graph $G = (V, E)$ is connected and non-bipartite. Let u_i be the i-th node sampled by a RW, where $1 \leq i \leq n$ and n be the number of samples. Denote by $\boldsymbol{\pi} = (\pi_v, v \in V)$ the stationary distribution of the RW, where $\pi_v = \frac{d_v}{2|E|}$. Then, for any function $f(v) : V \to \mathbb{R}$, where $\sum_{\forall v \in V} f(v) < \infty$,*

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

LEMMA 8.2. *[Meyn and Tweedie 2009, Theorem 17.2.1] [Ribeiro and Towsley 2010] Given that the undirected graph $G = (V, E)$ is connected and non-bipartite. Let (u_i, v_i) be the i-th edge sampled by a RW, where $1 \leq i \leq n$ and n be the number of samples. Then, for any function $f(u, v) : V \times V \to \mathbb{R}$, where $\sum_{\forall (u,v) \in E} f(u, v) < \infty$,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(u_i, v_i) \xrightarrow{a.s.} \frac{1}{|E|}\sum_{\forall (u,v) \in E} f(u, v).$$

8.1. Proof of Theorem 3.1

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

$$\begin{aligned} \mathrm{E}\left[\frac{\mathbf{1}(F(u_i, v_i) = a_k)}{m(u_i, v_i)}\right] &= \sum_{[u,v]\in \mathbf{S}^{(2)}} \pi^{(2)}_{[u,v]} \times \frac{\mathbf{1}(F(u, v) = a_k)}{m(u, v)} \\ &= \sum_{[u,v]\in \mathbf{S}^{(2)}} \frac{\mathbf{1}(F(u, v) = a_k)}{M} \quad (6) \\ &= \frac{\omega^{(2)}_k}{\bar{m}}. \quad (7) \end{aligned}$$

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

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

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

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

and

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

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

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

Application of the law of large numbers yields $\lim_{n\to\infty} \xi_{k,1} \xrightarrow{a.s.} \omega_k^{(2)}$ and $\lim_{n\to\infty} \xi_{k,2} \xrightarrow{a.s.} 1$. Therefore we have $\lim_{n\to\infty} \hat{\omega}_k^{(2)} = \lim_{n\to\infty} \frac{\xi_{k,1}}{\xi_{k,2}} \xrightarrow{a.s.} \omega_k^{(2)}$.

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

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

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 $m(u,v) \ge 1$. Similarly, from (10) we have

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

Using Chebyshev's inequality, we obtain

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

and $P\left(|\xi_{k,2} - 1| \le \epsilon\right) \ge 1 - \frac{1}{n\epsilon^2}(\bar{m}-1)$. Then we have inequalities $(1-\epsilon)\omega_k^{(2)} \le \xi_{k,1} \le (1+\epsilon)\omega_k^{(2)}$ and $1-\epsilon \le \xi_{k,2} \le 1+\epsilon$ with probability $P\left(|\xi_{k,1} - \omega_k^{(2)}| \le \epsilon\omega_k^{(2)}\right) + P\left(|\xi_{k,2} - 1| \le \epsilon\right) - P\left(|\xi_{k,1} - \omega_k^{(2)}| \le \epsilon\omega_k^{(2)} \text{ or } |\xi_{k,2} - 1| \le \epsilon\right)$, which is not smaller than $1 - \frac{1}{n\epsilon^2}\left(\frac{\bar{m}}{\omega_k^{(2)}} + \bar{m} - 2\right)$. Therefore, we have the following inequalities

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

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

8.2. Proof of Theorem 4.1

This sampling method can be viewed as a regular RW over the graph $G^{(2)} = (V^{(2)}, E^{(2)})$, where $V^{(2)} = \{[u,v] : u,v \in V\}$ and $E^{(2)} = \{([u,v],[x,y]) : (u,x),(v,y) \in E\}$. When $G^{(2)}$ is connected and non-bipartite. Then RW performed over $G^{(2)}$ exhibits a stationary distribution $\boldsymbol{\pi}_S = (\pi_{[u,v]} : u,v \in V)$, where $\pi_{[u,v]}$ is computed as $\pi_{[u,v]} = \frac{d_u d_v}{4|E|^2}$, $u,v \in V$. From Lemma 8.1, we have

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{1}(F(u_i,v_i)=a_k)\mathbf{1}(u_i\neq v_i)}{d_{u_i}d_{v_i}} \\
&\xrightarrow{a.s.} \sum_{\forall(u,v)\in V^{(2)}} \frac{\mathbf{1}(F(u,v)=a_k)\mathbf{1}(u\neq v)}{d_u d_v}\pi_{[u,v]} \\
&= \frac{1}{4|E|^2}\sum_{\forall[u,v]\in\mathbf{S}} \mathbf{1}(F(u,v)=a_k) \\
&= \frac{|V|(|V|-1)}{4|E|^2}\omega_k
\end{aligned}$$

and

$$\lim_{n\to\infty}\frac{J}{n} \xrightarrow{a.s.} \sum_{\forall(u,v)\in V^{(2)}} \frac{\mathbf{1}(u\neq v)}{d_u d_v}\pi_{[u,v]} = \frac{|V|(|V|-1)}{4|E|^2}.$$

Therefore we have $\lim_{n\to\infty}\hat{\omega}_k^{(\star)} \xrightarrow{a.s.} \omega_k$.

8.3. Proof of Theorem 4.2

From Lemma 8.2, we have

$$\begin{aligned}
\hat{\omega}_k^{(1\star)} &\xrightarrow{a.s.} \frac{1}{|E|}\sum_{\forall(u,v)\in E} \mathbf{1}(F(u_i,v_i)=a_k) \\
&= \frac{1}{2|E|}\sum_{\forall(u,v)\in\mathbf{S}^{(1)}} \mathbf{1}(F(u,v)=a_k) \\
&= \omega_k^{(1)}.
\end{aligned}$$

8.4. Proof of Theorem 4.3

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

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

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. A node (u,v) (an edge in G) in graph $\hat{G}$ connects to d_u+d_v-2 nodes in $\hat{G}$, its degree in $\hat{G}$ is d_u+d_v-2, and NRW can be viewed as a regular RW on graph $\hat{G}$, therefore, from [Lovász 1993] we find that the probability of NRW being at an edge $(u,v)\in E$ converges to $\boldsymbol{\pi}_E$.

8.5. Proof of Theorem 4.4

NRW can be viewed as a regular RW over the graph $\hat{G}=(\hat{V},\hat{E})$, where the node set $\hat{V}=\{(u,v):(u,v)\in E\}$, the edge set $\hat{E}=\{((u,v),(u,v')):(u,v)\in E,(u,v')\in E, v\neq v'\}$. When $\hat{G}$ is connected and non-bipartite, NRW exhibits a stationary distribution $\boldsymbol{\pi}_E=(\pi_{(u,v)}:(u,v)\in E)$, where $\pi_{(u,v)}$ is $\pi_{(u,v)}=\frac{d_u+d_v-2}{M}$. For $\hat{e}=((u,u'),(v,v'))$ an edge in $\hat{E}$, where u and v are two different nodes, define

$$\Phi(\hat{e}) = \frac{\mathbf{1}(F(u,v)=a_k)+\mathbf{1}(F(v,u)=a_k)}{2m(u,v)}.$$

Let Q be a random variable with probability distribution $P(Q=0)=P(Q=1)=0.5$. Define

$$\Psi(\hat{e},Q)=\frac{\mathbf{1}(F(u,v)=a_k)}{m(u,v)}(1-Q)+\frac{\mathbf{1}(F(v,u)=a_k)}{m(u,v)}Q.$$

Denote by $\hat{e}_i=((x_{i-1},y_{i-1}),(x_i,y_i))$ $(1\le i\le n)$ edges in $\hat{G}$ sampled by NRW, where (x_{i-1},y_{i-1}) and (x_i,y_i) are edges in the original graph G visited by NRW at steps $i-1$ and i. Let u_i' and v_i' be the two distinct nodes in (x_{i-1},y_{i-1}) and (x_i,y_i) respectively. Let Q_i be a random variable with probability distribution $P(Q_i=0)=P(Q_i=1)=0.5$. Since NRW generates a node pair $[u_i,v_i]$ at step i as $[u_i,v_i]=[u_i',v_i']$ when $Q_i=0$, and $[u_i,v_i]=[v_i',u_i']$ when $Q_i=1$, we have

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

where $H=\sum_{i=1}^{n}\frac{1}{m(u_i,v_i)}$. For $\hat{e}=((u,u'),(v,v'))$ an edge in $\hat{E}$, where u and v are two different nodes, denote $\Gamma(\hat{e})=\{i:\hat{e}_i=\hat{e},1\le i\le n\}$. When $\hat{G}$ is connected and non-bipartite, we easily show that $\lim_{n\to\infty}|\Gamma(\hat{e})|=\infty$. Then,

$$\begin{aligned}&\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Psi(\hat{e}_i,Q_i)\\
&=\lim_{n\to\infty}\frac{1}{n}\sum_{\hat{e}\in\hat{E}}\sum_{i\in\Gamma(\hat{e})}\frac{\mathbf{1}(F(u,v)=a_k)}{m(u,v)}(1-Q_i)+\lim_{n\to\infty}\frac{1}{n}\sum_{\hat{e}\in\hat{E}}\sum_{i\in\Gamma(\hat{e})}\frac{\mathbf{1}(F(v,u)=a_k)}{m(u,v)}Q_i\\
&=\lim_{n\to\infty}\frac{1}{n}\sum_{\hat{e}\in\hat{E}}\frac{\mathbf{1}(F(u,v)=a_k)}{m(u,v)}\sum_{i\in\Gamma(\hat{e})}(1-Q_i)+\lim_{n\to\infty}\frac{1}{n}\sum_{\hat{e}\in\hat{E}}\frac{\mathbf{1}(F(v,u)=a_k)}{m(u,v)}\sum_{i\in\Gamma(\hat{e})}Q_i.\end{aligned}$$

Since random variables Q_i are drawn from $(0,1)$ uniformly and independently, application of the law of large numbers yields $\lim_{n\to\infty}\frac{\sum_{i\in\Gamma(\hat{e})}Q_i}{|\Gamma(\hat{e})|}=\lim_{n\to\infty}\frac{\sum_{i\in\Gamma(\hat{e})}(1-Q_i)}{|\Gamma(\hat{e})|}=\frac{1}{2}$. Then we have

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Psi(\hat{e}_i,Q_i)&=\lim_{n\to\infty}\frac{1}{n}\sum_{\hat{e}\in\hat{E}}|\Gamma(\hat{e})|\frac{\mathbf{1}(F(u,v)=a_k)+\mathbf{1}(F(v,u)=a_k)}{2m(u,v)}\\
&=\lim_{n\to\infty}\frac{1}{n}\sum_{\hat{e}\in\hat{E}}|\Gamma(\hat{e})|\Phi(\hat{e})\\
&=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Phi(\hat{e}_i).\end{aligned}$$

From Lemma 8.2, we have

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Phi(\hat{e}_i)\xrightarrow{a.s.}&\frac{1}{|\hat{E}|}\sum_{\forall(\hat{e})\in\hat{E}}\Phi(\hat{e})\\
&=\frac{2}{M}\sum_{\forall[u,v]\in\mathbf{S}^{(2)}}m(u,v)\cdot\frac{\mathbf{1}(F(u,v)=a_k)}{2m(u,v)}\\
&=\frac{1}{M}|\mathbf{S}^{(2)}|\omega_k^{(2)}.\end{aligned}$$

Therefore

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\Psi(\hat{e}_i,Q_i)\xrightarrow{a.s.}\frac{1}{M}|\mathbf{S}^{(2)}|\omega_k^{(2)}. \tag{11}$$

Similarly, from Lemma 8.2 we have

$$\begin{aligned}\lim_{n\to\infty} H &\xrightarrow{a.s.} \frac{1}{|\hat{E}|}\sum_{\forall u\neq v,((u,u'),(v,v'))\in\hat{E}}\frac{1}{m(u,v)}\\ &= \frac{1}{M}\sum_{\forall[u,v]\in\mathbf{S}^{(2)}} m(u,v)\times\frac{1}{m(u,v)}\\ &= \frac{1}{M}|\mathbf{S}^{(2)}|. \end{aligned}\tag{12}$$

From (11) and (12), we have $\lim_{n\to\infty}\hat{\omega}_k^{(2\star)}\xrightarrow{a.s.}\omega_k^{(2)}$.

Acknowledgement: The work of John C.S. Lui was supported in part by the RGC GRF 415013.

References

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

Yongyeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. 2007. Analysis of Topological Characteristics of Huge Online Social Networking Services. In *Proceedings of WWW 2007*. 835–844.

Konstantin Avrachenkov, Bruno Ribeiro, and Don Towsley. 2010. Improving Random Walk Estimation Accuracy with Uniform Restarts. In *The 7th Workshop on Algorithms and Models for the Web Graph*. 98–109.

Stephen Boyd, Persi Diaconis, and Lin Xiao. 2004. Fastest Mixing Markov Chain on A Graph. *SIAM Rev.* 46, 4 (December 2004), 667–689.

Siddhartha Chib and Edward Greenberg. 1995. Understanding the Metropolis-Hastings Algorithm. *The American Statistician* 49, 4 (November 1995), 327–335.

Minas Gjoka, Carter T. Butts, Maciej Kurant, and Athina Markopoulou. 2011. Multigraph Sampling of Online Social Networks. *IEEE Journal on Selected Areas in Communications* 29, 9 (September 2011), 1893–1905.

Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. 2010. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *Proceedings of IEEE INFOCOM 2010*. 2498–2506.

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

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

W. K. Hastings. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* 57, 1 (April 1970), 97–109.

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

Galin L. Jones. 2004. On the Markov chain central limit theorem. *Probability Surveys* 1 (2004), 299–320.

U Kang, Spiros Papadimitriou, Jimeng Sun, and Hanghang Tong. 2011. Centralities in Large Networks: Algorithms and Observations. In *Proceedings of SDM 2011*. 119–1306.

Maciej Kurant, Minas Gjoka, Carter T. Butts, and Athina Markopoulou. 2011a. Walking on a Graph with a Magnifying Glass: Stratified Sampling via Weighted Random Walks. In *Proceedings of ACM SIGMETRICS 2011*. 281–292.

Maciej Kurant, Minas Gjoka, Yan Wang, Zack W. Almquist, Carter T. Butts, and Athina Markopoulou. 2011b. *Coarse-Grained Topology Estimation via Graph Sampling*. Technical Report arXiv:1105.5488.

Maciej Kurant, Athina Markopoulou, and Patrick Thiran. 2010. On the bias of BFS (Breadth First Search) and of Other Graph Sampling Techniques. In *Proceedings of International Teletraffic Congress 2010*.

Maciej Kurant, Athina Markopoulou, and Patrick Thiran. 2011. Towards Unbiased BFS Sampling. *IEEE Journal on Selected Areas in Communications* 29, 9 (September 2011), 1799–1809.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a Social Network or a News Media?. In *Proceedings of WWW 2010*. 591–600.

Chul-Ho Lee, Xin Xu, and Do Young Eun. 2012. Beyond Random Walk and Metropolis-Hastings Samplers: Why You Should Not Backtrack for Unbiased Graph Sampling. In *Proceedings of ACM SIGMETRICS/Performance 2012*. 319–330.

Jure Leskovec and Christos Faloutsos. 2006. Sampling from large graphs. In *Proceedings of ACM SIGKDD 2006*. 631–636.

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

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010a. Predicting Positive and Negative Links in Online Social Networks. In *Proceedings of WWW 2010*. 641–650.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010b. Signed Networks in Social Media. In *Proceedings of the 28th ACM Conference on Human Factors in Computing Systems (CHI)*. 1361–1370.

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

Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. 2009. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. *Internet Mathematics* 6, 1 (2009), 29–123.

L. Lovász. 1993. Random walks on graphs: a survey. *Combinatorics* 2 (1993), 1–46. Issue Paul Erdös is Eighty.

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

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 2011. Equations of State Calculations by Fast Computing Machines. *IEEE Journal on Selected Areas in Communications* 21, 6 (June 2011), 1087–1092.

Sean Meyn and Richard L. Tweedie. 2009. *Markov Chains and Stochastic Stability*. Cambridge University Press.

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

Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2007. Measurement and Analysis of Online Social Networks. In *Proceedings of ACM SIGCOMM Internet Measurement Conference 2007*. 29–42.

Abedelaziz Mohaisen, Aaram Yun, and Yongdae Kim. 2010. Measuring the Mixing Time of Social Graphs. In *Proceedings of ACM SIGCOMM Internet Measurement Conference 2010*. 390–403.

Amir H. Rasti, Mojtaba Torkjazi, Reza Rejaie, Nick Duffield, Walter Willinger, and Daniel Stutzbach. 2009. Respondent-driven Sampling for Characterizing Unstructured Overlays. In *Proceedings of IEEE INFOCOM Mini-conference 2009*.

Bruno Ribeiro and Don Towsley. 2010. Estimating and Sampling Graphs with Multidimensional Random Walks. In *Proceedings of ACM SIGCOMM Internet Measurement Conference 2010*. 390–403.

Bruno Ribeiro, Pinghui Wang, Fabricio Murai, and Don Towsley. 2012. Sampling Directed Graphs with Random Walks. In *Proceedings of IEEE INFOCOM 2012*. 1692–1700.

Matthew Richardson, Rakesh Agrawal, and Pedro Domingos. 2003. Trust Management for the Semantic Web. In *Proceedings of the 2nd International Semantic Web Conference*. 351–368.

Matei Ripeanu, Ian T. Foster, and Adriana Iamnitchi. 2002. Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design. *IEEE Internet Computing Journal* 6, 1 (2002), 50–57.

Gareth O. Roberts and Jeffrey S. Rosenthal. 2004. General state space Markov chains and MCMC algorithms. *Probability Surveys* 1 (2004), 20–71.

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

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

Parag Singla and Matthew Richardson. 2008. Yes, There is a Correlation - From Social Networks to Personal Behavior on the Web. In *Proceedings of WWW 2008*. 655–664.

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

Pinghui Wang, Junzhou Zhao, John C.S. Lui, Don Towsley, and Xiaohong Guan. 2012. *Sampling Content Distributed Over Graphs, available at http://www.cse.cuhk.edu.hk/%7ecslui/samplingcontentreport.pdf*. Technical Report. The Chinese University of Hong Kong.

Pinghui Wang, Junzhou Zhao, John C. S. Lui, Don Towsley, and Xiaohong Guan. 2013. Sampling node pairs over large graphs. In *Proceedings of IEEE ICDE 2013*. 781–792.

Junzhou Zhao, John C. S. Lui, Don Towsley, Xiaohong Guan, and Yadong Zhou. 2011. Empirical Analysis of the Evolution of Follower Network: A Case Study on Douban.. In *Proceedings of IEEE INFOCOM NetSciCom 2011*. 941–946.

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