A Tale of Three Graphs: Sampling Design on Hybrid Social-Affiliation Networks

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

[#] *MOEKLINNS Lab, Xi'an Jiaotong University, China*
[1] jzzhao@sei.xjtu.edu.cn
[5] xhguan@sei.xjtu.edu.cn

[*] *Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong*
[2] cslui@cse.cuhk.edu.hk

[+] *School of Computer Science, University of Massachusetts at Amherst, USA*
[3] towsley@cs.umass.edu

[$] *Huawei Noah's Ark Lab, Hong Kong*
[4] wang.pinghui@huawei.com

Abstract—**Random walk-based graph sampling methods have become increasingly popular and important for characterizing large-scale complex networks. While powerful, they are known to exhibit problems when the graph is loosely connected, which slows down the convergence of a random walk and can result in poor estimation accuracy. In this work, we observe that many graphs under study, called *target graphs*, usually do not exist in isolation. In many situations, a target graph is often related to an *auxiliary graph* and an *affiliation graph*, and the target graph becomes better connected when viewed from these three graphs as a whole, or what we called a *hybrid social-affiliation network*. This viewpoint brings extra benefits to the graph sampling framework, e.g., when directly sampling a target graph is difficult or inefficient, we can efficiently sample it with the assistance of auxiliary and affiliation graphs. We propose three sampling methods on such a hybrid social-affiliation network to estimate target graph characteristics, and conduct extensive experiments on both synthetic and real datasets, to demonstrate the effectiveness of these new sampling methods.**

I. Introduction

Online social networks (OSNs) such as Facebook, Sina Weibo, and Twitter have attracted much attention in recent years because of their ever-increasing popularity and importance in our daily lives [1]–[3]. An OSN not only provides a platform for people to connect with their friends, but also provides an opportunity to study user characteristics, which are valuable in a variety of applications. For example, Twitter users' tweeting activities (e.g., number of tweets related to a movie) can be used to predict movie box-office revenues [4], and OSN users' mood characteristics can forecast stock market prices [5]. Therefore, measuring user characteristics in OSNs is an important task.

Exactly calculating user characteristics requires the complete OSN data. For third parties who do not possess the data, they can only rely on public APIs to crawl the OSN. However, commercial OSNs are typically unwilling to grant third parties full permission to access the data due to user privacy and business secrecy. They often impose barriers to limit third party's large-scale crawling [6], and restrict the rate of requesting APIs [7]. As a result, crawling the complete data of a large-scale OSN is practically impossible.

To address this challenge, sampling methods have been developed, i.e., a small fraction of OSN users are sampled and used to estimate user characteristics. In the literature, random walk-based graph sampling methods have gained popularity [8]–[15]. In random walk sampling, a walker is launched over a graph, which continuously moves from a node to one of its neighbors selected uniformly at random, to obtain a collection of node samples. These samples can yield unbiased estimates of graph characteristics [15,16].

Motivation. While random walk sampling is powerful, if a graph is loosely connected, e.g., consists of communities, it will suffer from *slow mixing* [17], i.e., requires a long "burn-in" period to reach steady state, which results in the need of a large number of samples in order to achieve good estimation accuracy. Recent studies have found that mixing times in many real-world networks are larger than expected [18]. To overcome the slow mixing problem, one approach is to incorporate *uniform node sampling* (UNI) into random walk sampling, which is also known as *random walk with jumps* (RWwJ) [9,12,14]. In UNI, nodes are independently sampled uniformly at random by querying randomly generated node IDs in a graph. For example, users in many OSNs have unique numeric IDs, and UNI can be conducted by generating random IDs in the user ID space and including those valid IDs into samples. RWwJ then leverages UNI to perform *jumps* on a graph. Specifically, at each step of RWwJ, the walker jumps with a probability determined by the node where it currently resides, to a node sampled by UNI. By incorporating UNI into random walk sampling, the walker can jump out of a community or disconnected component of a graph, and avoid being trapped, thereby reducing mixing times [9].

The main drawback of RWwJ is that UNI can be resource-intensive when the valid node ID space of a graph is very sparsely populated. For example, the fractions of valid IDs in MySpace and Flickr are only about 10% and 1.3% respectively [12], and as a result, one has to generate about 10 (or 77) random IDs to obtain a valid ID in MySpace (or Flickr). This problem can become even worse in some practical situations as illustrated by the following example.

Example 1. *A restaurant company wants to build a new chain store in one of two small candidate cities in China. A market surveyor is sent to evaluate the consumption abilities of inhabitants of the two cities. In China, Sina Weibo [19] is the most popular microblogging website and provides a check-in service*[1] *that enables people to share consumption information with their friends. For example, a Weibo user can share the location of a restaurant and photos of her dinner with her friends via a mobile application. Since most citizens use this check-in service to share their consumption information on Weibo, the surveyor decides to use Weibo as a platform to conduct his research. To measure the average consumption abilities of citizens in the two cities, he plans to uniformly sample two collections of Weibo users in the two cities respectively. It is known that every Weibo user ID consists of ten digits ranging from "1000000000" to "5058913818" (as of March 25, 2014). He generates random numbers in this range as test IDs and finds that about* 11% *of the test IDs are valid Weibo users. However, because the population sizes of the two cities are small, e.g., hundreds of thousands of citizens compared to hundreds of millions of Weibo users, the percentage of a valid user residing in the two cities is on the order of* 0.1%.

In the above example, the surveyor expects a test ID to fall into one of the two cities to obtain a valid Weibo user sample, but UNI becomes extremely inefficient because the probability that UNI obtains a valid sample equals $P(\textit{ID is valid}) \times P(\textit{ID falls into one city}) = 0.11 \times 0.001 \approx 10^{-4}$. This results in the need for a surveyor to try 10^4 times on average to obtain a single valid user residing in one of the two cities. Even worse, in some OSNs such as Pinterest [21], user IDs are represented by arbitrary-length strings, which makes UNI practically impossible. Without the ability to efficiently conduct UNI on a graph, RWwJ will also become inefficient. This raises the following question: how do we sample nodes efficiently on a graph when uniform node sampling is inefficient or impractical at all?

Present Work. In the previous example, the problem is how to effectively sample Weibo users in the two cities, and UNI is inefficient because of the sparsity of user ID space. Since directly sampling users is inefficient, we propose to sample users in an *indirect* manner. We notice that check-in information (i.e., which user checked in which place) shared by users often contain the venue information, e.g., the location of the restaurant (i.e., a latitude-longitude coordinate on a map) where the user lunched. Most such OSNs provide APIs for querying venues within an area specified by a rectangle region with south-west and north-east corners latitude-longitude coordinates given [22,23]. This function can be used to design efficient sampling methods for sampling venues in an area on a map [24]–[26]. Since we can easily sample venues within an area of interest, we are then able to *indirectly* sample Weibo users in an area by *relating users to venues through check-in relations that exist between them*. This will be more efficient than directly sampling users in an area. We present the detailed design of this sampling method in Section III and evaluate it in Section IV.

An important lesson learnt from Example 1 is that, when direct sampling of the *user space* is inefficient, we can switch to sample the *venue space*, and the relations that exist between the two sample spaces allow us to efficiently sample the user space. In general, we use *three graphs* to represent the two sample spaces and their relations. In Example 1, we consider a venue as another type of node besides user node, and build the following three graphs: (1) a *user graph* formed by users and their relations, (2) a *venue graph* formed by venues and their relations (the edge set is actually empty in Example 1), and (3) a *bipartite graph* formed by users, venues and their check-in relations. In Example 1, although directly sampling the user graph is very difficult or extremely inefficient, we can easily sample the venue graph, and use the bipartite graph to connect the two sample spaces to indirectly sample the user graph with efficiency.

Because the *affiliation relationship* between users and venues plays an important role in this method, we refer to the three graphs jointly as a *hybrid social-affiliation network*. The formal definition of hybrid social-affiliation network will be given in Section II, and the detailed design of the sampling methods on hybrid social-affiliation networks will be presented in Section III.

Contributions. We make three contributions in this work:

- We introduce the concept of hybrid social-affiliation network and formulate a sampling problem over it to characterize graphs. (Section II).
- We design three sampling methods over such a hybrid social-affiliation network. These methods allow us to efficiently sample a graph when it is difficult to sample the graph directly (Section III).
- We conduct extensive experiments to validate the proposed methods on both synthetic and real-world datasets (Section IV).

II. Problem Definition

We begin with introducing the graph characteristics of interest to us, and then formally define the hybrid social-affiliation network.

A. Graph Characteristics

We model an OSN as an undirected graph $G(\mathcal{U}, \mathcal{E})$, where $\mathcal{U}$ and $\mathcal{E}$ are sets of users and relations among users, respectively. Users in the graph are labeled. We let $\mathcal{L}$ be a set of user labels

[1] The Check-in service [20] in Weibo allows a registered user to "check in" at venues using a mobile application by selecting from a list of venues nearby the user, and the location of the user is usually based on GPS hardware in the mobile device. In addition, a user can attach a text description or several photos associated with the venue while checking in.

associated with the graph, and each user is mapped to a subset of labels he owns by a *characteristic function* $L\colon \mathcal{U} \mapsto 2^{\mathcal{L}}$. For example, if $\mathcal{L} = \{\text{male}, \text{female}\}$, then $L(u)$ represents the gender of user u.

In many applications, we are interested in estimating the fractions of users having some labels, e.g., the fraction of male/female customers buying a product. This can be represented by the *label distribution* $\{\theta_s\}_{s\subseteq\mathcal{L}}$, where θ_s is the fraction of users with labels s in graph G. That is

$$\theta_s = \frac{1}{n}\sum_{u\in\mathcal{U}} \mathbf{1}\left\{s \subseteq L(u)\right\},\ s \subseteq \mathcal{L},$$

where $n = |\mathcal{U}|$ is the graph size, $\mathbf{1}\{\cdot\}$ is the indicator function and $\mathbf{1}\{C\} = 1$ if condition C is true; otherwise $\mathbf{1}\{C\} = 0$.

With this definition of graph characteristics, **our objective** of this paper is to design an algorithm for collecting node samples within a budget B from graph G and for providing unbiased estimates of $\{\theta_s\}_{s\subseteq\mathcal{L}}$ with low statistical error.

B. Hybrid Social-Affiliation Network

Example 1 motivates us to introduce a *hybrid social-affiliation network* which can be used to design efficient estimators when direct sampling of graph G is difficult or inefficient. According to our previous analysis, a hybrid social-affiliation network consists of three graphs: $G(\mathcal{U}, \mathcal{E})$, $G'(\mathcal{V}, \mathcal{E}')$, and $G_b(\mathcal{U}, \mathcal{V}, \mathcal{E}_b)$, where $\mathcal{U}, \mathcal{V}$ are sets of nodes, and $\mathcal{E}, \mathcal{E}', \mathcal{E}_b$ are sets of edges. More specifically,

- $G(\mathcal{U}, \mathcal{E})$ is the *target graph*, whose characteristics $\{\theta_s\}_{s\subseteq\mathcal{L}}$ are of interest to us and need to be measured, e.g., the user social network in Example 1.
- $G'(\mathcal{V}, \mathcal{E}')$ is an *auxiliary graph*, which can be more efficiently sampled than the target graph, e.g., the venue graph (with $\mathcal{E}' = \emptyset$) in Example 1.
- $G_b(\mathcal{U}, \mathcal{V}, \mathcal{E}_b)$ is an *affiliation graph* [27, Chapter 8], which is a bipartite graph connecting nodes in the target and auxiliary graphs, e.g., the graph formed by users, venues and their check-in relations in Example 1.

An example of such a hybrid social-affiliation network is illustrated in Fig. 1, from which we observe that the disconnected target graph becomes better connected with the assistance of the auxiliary and affiliation graphs, so the target graph can be more efficiently sampled.

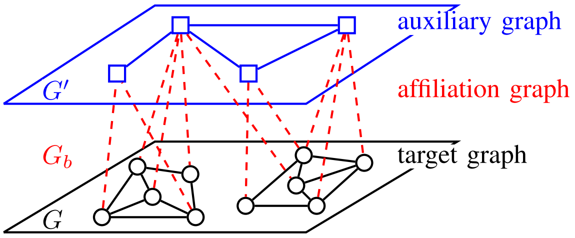


Fig. 1. **Illustration of a hybrid social-affiliation network.** Target graph together with auxiliary and affiliation graphs form a better connected graph than target graph itself, which improves sampling efficiency on target graph.

In addition to Example 1, many other graph measurement problems can be formulated as hybrid social-affiliation network sampling problems. To illustrate, we consider another example.

Example 2. *Mtime [28] is an online movie database in China, which comprises two types of accounts: users and actors. Users in Mtime can follow each other to form a social network. Movie actors can also form connections with each other if they cooperated in same movies. Moreover, a user can follow movie actors if he/she is a fan of the actor.*

In the above example, if we want to measure the characteristics of the graph formed by Mtime users, and direct sampling of this user graph is inefficient because the user graph is not well connected due to user interest differences, geographic constraints, etc. (A detailed analysis of the Mtime network can be found in Section IV.) However, we can build a hybrid social-affiliation network as follows:

- *Target graph* G consists of Mtime users and their following relations.
- *Auxiliary graph* G' consists of actors and their cooperation relations.
- *Affiliation graph* G_b consists of Mtime users, actors and the fan relations between them.

Unlike ordinary people, movie actors, especially pop stars, are more easily to form connections with each other because they have more opportunities to participate in the same events such as Oscar and Cannes. In other words, auxiliary graph is more likely to be well connected than target graph. We can leverage this feature to measure target graph characteristics more efficiently.

III. Sampling Design on Hybrid Social-Affiliation Networks

In this section, we design three methods for characterizing a target graph by sampling a hybrid social-affiliation network.

A. Indirectly Sampling Target Graph by Vertex Sampling on Auxiliary Graph (VS^A)

The first method is based on the assumption that vertex sampling, e.g., UNI, is easy to conduct on the auxiliary graph but not on the target graph, as is the case in Example 1. We present a sampling method VS$^{\text{A}}$ to indirectly sample the target graph under this setting. The basic idea of VS$^{\text{A}}$ is illustrated in Fig. 2.

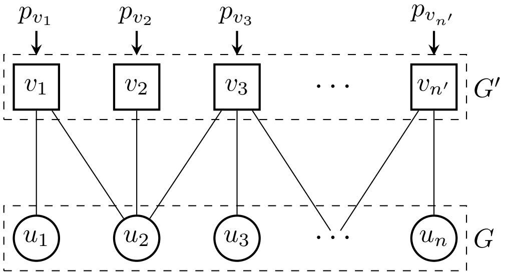


Fig. 2. **Illustration of VS$^{\text{A}}$.** Edges in target and auxiliary graphs are omitted.

In VS$^{\text{A}}$, we assume that a node $v \in \mathcal{V}$ can be sampled with probability p_v in auxiliary graph G'. For example, when graph G' supports uniform node sampling, then $p_v = 1/n', \forall v \in \mathcal{V}$, where $n' = |\mathcal{V}|$ is the size of graph G'. When a node $v \in \mathcal{V}$ is sampled, we collect all of its neighbors in the affiliation graph as samples. We will describe how these samples can yield unbiased estimates of target graph characteristics. The detailed design of VS$^{\text{A}}$ is described in the follows.

Sampling Design. Sampling design of VS$^{\text{A}}$ consists of the following two steps:

(i) Sample a collection of B' nodes with replacement in the auxiliary graph G'. Denote these samples as sequence $\mathcal{S}' = [y_1, \ldots, y_{B'}]$.

(ii) For each $v \in \mathcal{S}'$, let $\mathcal{U}_v \subseteq \mathcal{U}$ denote the set of neighbors of v in G_b. Include the pair $(v, \mathcal{U}_v)$ into sequence $\mathcal{S} = [(y_1, \mathcal{U}_{y_1}), \ldots, (y_{B'}, \mathcal{U}_{y_{B'}})]$.

Using sequence $\mathcal{S}$, VS$^{\text{A}}$ estimates target graph characteristics $\{\theta_s\}_{s \subseteq \mathcal{L}}$ according to following estimators.

Estimators. If in advance, we know the graph size of G is n, we can use the following estimator to estimate θ_s,

$$\hat{\theta}_s^{\text{VS}^{\text{A}}} = \frac{1}{nB'} \sum_{i=1}^{B'} \frac{1}{p_{y_i}} \sum_{u \in \mathcal{U}_{y_i}} \frac{\mathbf{1}\{s \subseteq L(u)\}}{d_u^{(b)}}, \tag{1}$$

where $d_u^{(b)}$ is the degree of node u in affiliation graph G_b.

When n is unknown, we introduce another estimator

$$\check{\theta}_s^{\text{VS}^{\text{A}}} = \frac{1}{Z} \sum_{i=1}^{B'} \frac{1}{p_{y_i}} \sum_{u \in \mathcal{U}_{y_i}} \frac{\mathbf{1}\{s \subseteq L(u)\}}{d_u^{(b)}}, \tag{2}$$

where

$$Z = \sum_{i=1}^{B'} \frac{1}{p_{y_i}} \sum_{u \in \mathcal{U}_{y_i}} \frac{1}{d_u^{(b)}}.$$

The following theorem guarantees the *unbiasedness* of the two estimators.

Theorem 1. *Estimator* (1) *is an unbiased estimator of θ_s. Estimator* (2) *is an asymptotically unbiased estimator of θ_s.*

Proof. To prove Estimator (1) is unbiased, we show that

$$\begin{aligned}\mathbb{E}\left[\hat{\theta}_s^{\text{VS}^{\text{A}}}\right] &= \frac{1}{nB'} \sum_{i=1}^{B'} \mathbb{E}\left[\frac{1}{p_{y_i}} \sum_{u \in \mathcal{U}_{y_i}} \frac{\mathbf{1}\{s \subseteq L(u)\}}{d_u^{(b)}}\right] \\ &= \frac{1}{n} \sum_{v \in \mathcal{V}} p_v \frac{1}{p_v} \sum_{u \in \mathcal{U}_v} \frac{\mathbf{1}\{s \subseteq L(u)\}}{d_u^{(b)}} \\ &= \frac{1}{n} \sum_{u \in \mathcal{U}} \mathbf{1}\{s \subseteq L(u)\} \\ &= \theta_s.\end{aligned}$$

The second equality holds because that $y_1, \ldots, y_{B'}$ are i.i.d random variables. The third equality holds because that when we merge two summations into one summation with respect to u, an item in the inner summation is added $d_u^{(b)}$ times for each u. Hence, $\hat{\theta}_s^{\text{VS}^{\text{A}}}$ is unbiased.

To prove that estimator (2) is asymptotically unbiased, we can use the ratio form of the law of large numbers in [16, Theorem 17.2.1 on P. 428]. Hence,

$$\lim_{B' \to \infty} \check{\theta}_s^{\text{VS}^{\text{A}}} = \frac{\mathbb{E}\left[nB'\hat{\theta}_s^{\text{VS}^{\text{A}}}\right]}{\mathbb{E}[Z]} = \theta_s,$$

where $\mathbb{E}[Z] = nB'$ can be proved in a similar way as of the proof of unbiasedness of $\hat{\theta}_s^{\text{VS}^{\text{A}}}$. □

Remark. It is important to know that VS$^{\text{A}}$ can provide unbiased estimates of target graph characteristics under the condition that *every node in the target graph is connected to nodes in the auxiliary graph*. Because VS$^{\text{A}}$ can only sample nodes in $\mathcal{U}$ satisfying $d_u^{(b)} > 0$ according to the design of VS$^{\text{A}}$. If a node u is not connected to any node in $\mathcal{V}$, u cannot be indirectly sampled by VS$^{\text{A}}$. In Example 1, since we are only interested in users who share their check-ins in Weibo, therefore Example 1 satisfies this condition.

B. Random Walk on Target Graph Incorporating with Vertex Sampling on Auxiliary Graph (RW$^{\text{T}}$VS$^{\text{A}}$)

In some situations, $d_u^{(b)} = 0$ for some $u \in \mathcal{U}$. For example, some Mtime users in Example 2 may not follow any movie actor, and these users cannot be sampled by VS$^{\text{A}}$ (and as a result, VS$^{\text{A}}$ can not provide unbiased estimates of Mtime user characteristics). To address this issue, we propose a second sampling method RW$^{\text{T}}$VS$^{\text{A}}$, which combines random walk sampling on the target graph with vertex sampling on the auxiliary graph.

The basic idea of RW$^{\text{T}}$VS$^{\text{A}}$ is that, we launch a random walk on the target graph, and at each step allow the walker to jump with a probability dependent on the node of which it currently resides. This is similar to RWwJ [9,12,14] on the target graph G, but with the major difference that in RW$^{\text{T}}$VS$^{\text{A}}$ the walker jumps to a node in G by jumping first to a node in G', and then randomly selecting one of its neighbors in G_b. We refer to this as an *indirect jump*, and show in experiments that indirect jumps in RW$^{\text{T}}$VS$^{\text{A}}$ bring similar benefits as the direct jumps in RWwJ. An additional advantage of using random walk on the target graph is that it better characterizes highly connected nodes than uniform node sampling as random walks are biased towards high degree nodes in G. We depict RW$^{\text{T}}$VS$^{\text{A}}$ in Fig. 3, where each node in G is connected to a virtual jumper node to conduct indirect jumps, through doing vertex sampling over auxiliary graph G'.

Similar to VS$^{\text{A}}$, we assume that a node v in G' can be sampled with probability p_v. In RW$^{\text{T}}$VS$^{\text{A}}$, we virtually connect each node $u \in \mathcal{U}$ to a jumper node j with edge (u, j). Each edge (u, j) is assigned a weight ω_u, and the walker residing at node u moves to j to perform an indirect jump with probability

$$p_{uj} = \frac{\omega_u}{d_u + \omega_u}.$$

To determine $\{\omega_u\}_{u \in \mathcal{U}}$, we note that in an indirect jump, the walker jumps from j to node u with probability

$$p_{ju} = q_u = \sum_{v \in \mathcal{V}_u} \frac{p_v}{d_v^{(b)}}, \tag{3}$$

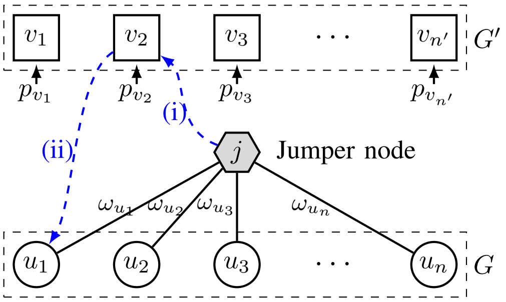


Fig. 3. **Illustration of RW$^{\text{T}}$VS$^{\text{A}}$ and indirect jump**. Each node u in G is virtually connected to a jumper node j with weight ω_u. An indirect jump is performed by: **(i)** randomly sampling a node v in G', and **(ii)** randomly choosing a neighbor of v in G_b as the target node to jump to.

where $\mathcal{V}_u \subseteq \mathcal{V}$ is the set of neighbors of u in the affiliation graph G_b, and $d_v^{(b)}$ is the degree of node v in G_b. By setting the weights to satisfy

$$\omega_u = \alpha q_u, \ u \in \mathcal{U}, \tag{4}$$

for any $\alpha \geq 0$, we ensure that the stationary probabilities of the random walk visiting $u \in \mathcal{U}$, and j are

$$\pi_u = \frac{d_u + \omega_u}{2|\mathcal{E}| + 2\alpha} \qquad \text{and} \qquad \pi_j = \frac{\alpha}{2|\mathcal{E}| + 2\alpha}.$$

Note that if $d_u^{(b)} = 0$, then $p_{uj} = p_{ju} = 0$, and the walker cannot jump from u (or to u). However, u can still be sampled by the random walk on G.

RW$^{\text{T}}$VS$^{\text{A}}$ exhibits similar properties as RWwJ. That is, when $\alpha = 0$, RW$^{\text{T}}$VS$^{\text{A}}$ becomes a simple random walk on the target graph, and when $\alpha \to \infty$, RW$^{\text{T}}$VS$^{\text{A}}$ is equivalent to VS$^{\text{A}}$.

Sampling Design. Suppose the random walk starts at node $x_1 \in \mathcal{U}$, and at step i the random walk is at node x_i. We calculate the probability q_{x_i} according to Eq. (3) and $\omega_{x_i} = \alpha q_{x_i}$. At step i, the walker jumps with probability $\omega_{x_i}/(d_{x_i} + \omega_{x_i})$; otherwise, the walker moves to a neighbor u of x_i chosen uniformly at random and $x_{i+1} = u$. An indirect jump is performed as follows:

(i) sample a node $v \in \mathcal{V}$ in the auxiliary graph with probability p_v.

(ii) sample a neighbor u of v uniformly at random in the affiliation graph, and let $x_{i+1} = u$.

Estimator. Based on the sample path $\mathcal{S} = (x_i, \ldots, x_B)$ collected by RW$^{\text{T}}$VS$^{\text{A}}$, an estimator for θ_s is

$$\hat{\theta}_s^{\text{RW}^\text{T}\text{VS}^\text{A}} = \frac{1}{Z} \sum_{i=1}^{B} \frac{\mathbf{1}\{s \subseteq L(x_i)\}}{d_{x_i} + \omega_{x_i}}, \tag{5}$$

where $Z = \sum_{i=1}^{B} 1/(d_{x_i} + \omega_{x_i})$.

Theorem 2. *Estimator* (5) *is asymptotically unbiased.*

Proof. First, we know that when RW$^{\text{T}}$VS$^{\text{A}}$ reaches steady state, each node u is sampled with probability

$$p_u = \frac{\pi_u}{1 - \pi_j} = \frac{d_u + \omega_u}{2|\mathcal{E}| + \alpha}. \tag{6}$$

Next, let $D \triangleq \sum_{i=1}^{B} \mathbf{1}\{s \subseteq L(x_i)\}/(d_{x_i} + \omega_{x_i})$, then

$$\begin{aligned} \mathbb{E}[D] &= \mathbb{E}\left[\sum_{i=1}^{B} \frac{\mathbf{1}\{s \subseteq L(x_i)\}}{d_{x_i} + \omega_{x_i}}\right], \\ &= B \sum_{u \in \mathcal{U}} p_u \frac{\mathbf{1}\{s \subseteq L(u)\}}{d_u + \omega_u} = \frac{Bn}{2|\mathcal{E}| + \alpha}\theta_s. \end{aligned}$$

Similarly, we can show that

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E}\left[\sum_{i=1}^{B} \frac{1}{d_{x_i} + \omega_{x_i}}\right], \\ &= B \sum_{u \in \mathcal{U}} p_u \frac{1}{d_u + \omega_u} = \frac{Bn}{2|\mathcal{E}| + \alpha}. \end{aligned}$$

Now, we invoke Theorem 17.2.1 in [16, P. 428], which is the ratio form of the law of large numbers, and we have

$$\lim_{B \to \infty} \hat{\theta}_s^{\text{RW}^\text{T}\text{VS}^\text{A}} = \frac{\mathbb{E}[D]}{\mathbb{E}[Z]} = \theta_s.$$

□

Remark. Note that RW$^{\text{T}}$VS$^{\text{A}}$ requires vertex sampling (e.g., UNI) on the auxiliary graph G'. If vertex sampling is also not allowed on G', RW$^{\text{T}}$VS$^{\text{A}}$ cannot be applied. However, one can replace the vertex sampling on G' by a random walk on G'. Unfortunately, this naive approach can perform very poorly when the auxiliary graph G' is not well connected, because a poorly connected graph can easily trap a simple random walk in a community. In what follows, we design a third method to address this challenge.

C. Random Walk on Target Graph Incorporating with Random Walk on Auxiliary Graph (RW^{T}RW^{A})

When both the target and auxiliary graphs do not support vertex sampling, neither VS$^{\text{A}}$ nor RW$^{\text{T}}$VS$^{\text{A}}$ can be applied. Therefore, we design the RW$^{\text{T}}$RW$^{\text{A}}$ method to address this setting. RW$^{\text{T}}$RW$^{\text{A}}$ consists of two parallel random walks on G and G' respectively. The two parallel random walks cooperate with each other, and can be considered as two RWwJs, as illustrated in Fig. 4. Different from RW$^{\text{T}}$VS$^{\text{A}}$, nodes in G and G' are both virtually connected to two jumper nodes j and j' to perform indirect jumps on G and G' respectively.

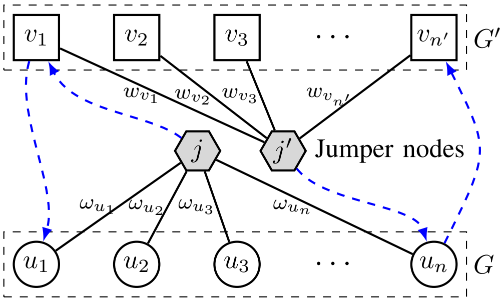


Fig. 4. **Illustration of RW$^{\text{T}}$RW$^{\text{A}}$ and indirect jumps**. Nodes in G and G' are virtually connected to two jumper nodes j and j' to perform indirect jumps, respectively. Indirect jumps are illustrated using dashed blue arcs.

The basic idea behind RW$^{\mathrm{T}}$RW$^{\mathrm{A}}$ is as follows. Suppose the two random walks are RW on G and RW' on G', and at step i, they reside at $x_i \in \mathcal{U}$ and $y_i \in \mathcal{V}$, respectively. If one random walk needs to jump at step i, say RW on G, then it jumps to a uniformly at random chosen neighbor of y_i in the affiliation graph, which is assigned to x_{i+1}. Similar jumping procedure also applies to RW' on G'. Hence, they are equivalent to two RWwJs, and both can avoid being trapped on G and G'.

We need to determine edge weights $\{\omega_u\}_{u\in\mathcal{U}}$ and $\{w_v\}_{v\in\mathcal{V}}$, which control the probability of jumping of the random walks on G and G', respectively. Obviously, the stationary distributions $\{\pi_u\}_{u\in\mathcal{U}}$ and $\{\pi_v\}_{v\in\mathcal{V}}$ of the two random walks on G and G' are related to these weights. Here we can leverage our previous analysis of indirect jumps in RW$^{\mathrm{T}}$VS$^{\mathrm{A}}$, and derive that, when parameters ω_u and w_v satisfy the following conditions

$$\omega_u = \alpha \sum_{v\in\mathcal{V}_u} \frac{\pi_v}{d_v^{(b)}},\ u \in \mathcal{U}, \tag{7}$$

$$w_v = \beta \sum_{u\in\mathcal{U}_v} \frac{\pi_u}{d_u^{(b)}},\ v \in \mathcal{V}, \tag{8}$$

for any $\alpha, \beta > 0$, the stationary distributions of the two random walks on G and G' (discarding states j and j') are

$$\pi_u = \frac{d_u + \omega_u}{2|\mathcal{E}| + \alpha},\ u \in \mathcal{U}, \tag{9}$$

$$\pi_v = \frac{d_v + w_v}{2|\mathcal{E}'| + \beta},\ v \in \mathcal{V}. \tag{10}$$

Arranging Eqs. (7)–(10) in matrix forms, we obtain

$$\pi_{\mathcal{U}} = \frac{d_{\mathcal{U}} + \omega_{\mathcal{U}}}{2|\mathcal{E}| + \alpha}, \qquad \pi_{\mathcal{V}} = \frac{d_{\mathcal{V}} + w_{\mathcal{V}}}{2|\mathcal{E}'| + \beta}, \tag{11}$$

$$\omega_{\mathcal{U}} = \alpha A D_{\mathcal{V}}^{-1} \pi_{\mathcal{V}}, \qquad w_{\mathcal{V}} = \beta A^T D_{\mathcal{U}}^{-1} \pi_{\mathcal{U}}, \tag{12}$$

where $A_{n\times n'}$ is the adjacency matrix of G_b, $\omega_{\mathcal{U}} = [\omega_u]^T_{u\in\mathcal{U}}$, $w_{\mathcal{V}} = [w_v]^T_{v\in\mathcal{V}}$, $\pi_{\mathcal{U}} = [\pi_u]^T_{u\in\mathcal{U}}$, $\pi_{\mathcal{V}} = [\pi_v]^T_{v\in\mathcal{V}}$, $d_{\mathcal{U}} = [d_u]^T_{u\in\mathcal{U}}$ and $d_{\mathcal{V}} = [d_v]^T_{v\in\mathcal{V}}$ are vectors, $D_{\mathcal{U}} = diag(d^{(b)}_{u_1}, \ldots, d^{(b)}_{u_n})$ and $D_{\mathcal{V}} = diag(d^{(b)}_{v_1}, \ldots, d^{(b)}_{v_{n'}})$ are diagonal matrices.

Equations (11)–(12) uniquely determine $\omega_{\mathcal{U}}$ and $w_{\mathcal{V}}$, i.e.,

$$\omega^*_{\mathcal{U}} = c(I - cc' A D_{\mathcal{V}}^{-1} A^T D_{\mathcal{U}}^{-1})^{-1} A D_{\mathcal{V}}^{-1}(d_{\mathcal{V}} + c' A^T D_{\mathcal{U}}^{-1} d_{\mathcal{U}}),$$
$$w^*_{\mathcal{V}} = c'(I - cc' A^T D_{\mathcal{U}}^{-1} A D_{\mathcal{V}}^{-1})^{-1} A^T D_{\mathcal{U}}^{-1}(d_{\mathcal{U}} + c A D_{\mathcal{V}}^{-1} d_{\mathcal{V}})$$

where $c = \alpha/(2|\mathcal{E}'| + \beta)$ and $c' = \beta/(2|\mathcal{E}| + \alpha)$ are constants.

The above results illustrate that, when α and β are given, $\omega_{\mathcal{U}}$ and $w_{\mathcal{V}}$ are uniquely determined. However, one needs complete knowledge of G, G' and G_b to determine their values. While, we are interested in sampling the target graph without having to know either G, G' or G_b in advance. In what follows, we design RW$^{\mathrm{T}}$RW$^{\mathrm{A}}$ in a way that only makes use of *local knowledge* of these graphs during the random walks.

First, we need to know what happens when $\omega_{\mathcal{U}} \neq \omega^*_{\mathcal{U}}$ or $w_{\mathcal{V}} \neq w^*_{\mathcal{V}}$. If $\omega_{\mathcal{U}}$ deviates from $\omega^*_{\mathcal{U}}$ (e.g., $\omega_{\mathcal{U}}$ is given and different from $\omega^*_{\mathcal{U}}$), we can still derive a "stationary distribution" $\{\pi_u\}$ for the random walk over target graph by Eq. (9). Because $\{w_v\}$ are related to $\{\pi_u\}$ according to Eq. (8), we can use local knowledge (by querying $\mathcal{U}_v$ from G_b) to obtain $w_v, \forall v$, and derive a "stationary distribution" $\{\pi_v\}$ for the random walk over auxiliary graph by Eq. (10). However, because $\omega_{\mathcal{U}} \neq \omega^*_{\mathcal{U}}$, Eq. (7) will not hold, and as a result, $\{\pi_u\}$ and $\{\pi_v\}$ will not be the stationary distributions.

To solve this contradiction, we notice that Eq. (7) actually relates to the indirect jumps on target graph G. The walker indirectly jumps to a node $u \in \mathcal{U}$ with probability $q_u \triangleq \omega_u/\alpha$ according to our analysis in RW$^{\mathrm{T}}$VS$^{\mathrm{A}}$ (see Eq. (3)). When $\omega_{\mathcal{U}}$ deviates from $\omega^*_{\mathcal{U}}$, we derive a different ω'_u by Eq. (7), which indicates the worker actually jumping to u with probability $q'_u \triangleq \omega'_u/\alpha \neq q_u$. The walker expects to jump to a node that follows distribution $\{q_u\}$, but it actually jumps to a node that follows distribution $\{q'_u\}$. This is another way to explain the contradiction. Fortunately, with this understanding, the contradiction can be easily solved by applying a Metropolis-Hastings (MH) sampler [29, Chapter 7], by considering $\{q_u\}$ as the *desired distribution* and $\{q'_u\}$ as the *proposal distribution*. We can use a MH sampler to build a Markov chain (referred as the MH chain) that generates samples with desired distribution $\{q_u\}$, and each time when the walker requires jumping, it jumps to a latest sample of MH chain. This guarantees that the walker jumps to u following distribution $\{q_u\}$, and ensures that $\{\pi_u\}$ and $\{\pi_v\}$ are still the stationary distributions of the random walks on target and auxiliary graphs.

Sampling Design. The complete sampling design of RW$^{\mathrm{T}}$RW$^{\mathrm{A}}$ comprises three parallel Markov chains as illustrated in Fig. 5, and we need to specify a desired distribution $\{q_u\}_{u\in\mathcal{U}}$ in advance, e.g., a uniform distribution.

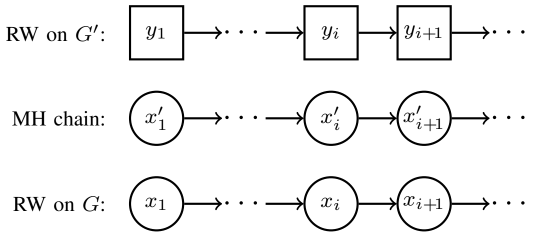


Fig. 5. Three parallel Markov chains in RW$^{\mathrm{T}}$RW$^{\mathrm{A}}$.

• *Random Walk on Auxiliary Graph G':* Suppose the random walk resides at node $y_i \in \mathcal{V}$ at step i. Then we can calculate w_{y_i} according to Eq. (8). At step $i+1$, the random walk executes one of the following two steps.

Jump: With probability $w_{y_i}/(d_{y_i} + w_{y_i})$, the walker jumps to a random neighbor $v \in \mathcal{V}$ of node x_i in G_b, and $y_{i+1} = v$;

Walk: Otherwise, the walker moves to a random neighbor $v \in \mathcal{V}$ of y_i in G', and $y_{i+1} = v$.

• *Metropolis-Hastings (MH) Chain:* Suppose the MH chain resides at node x'_i at step i. At step $i+1$, we randomly choose a neighbor $u \in \mathcal{U}$ of y_i in G_b. This is equivalent to sample a node $u \in \mathcal{U}$ with probability q'_u.

Acceptance: With probability r_i, we accept u and $x'_{i+1} = u$, where $r_i = \min\{1, (q_u q'_{x'_i})/(q_{x'_i} q'_u)\}$;

Rejection: Otherwise, we reject u and $x'_{i+1} = x'_i$.

• *Random Walk on Target Graph G:* Suppose the random walk resides at node $x_i \in \mathcal{U}$ at step i. $\omega_{x_i} = \alpha q_{x_i}$ where q_{x_i} is specified in advance. At step $i+1$, the walker executes one of the following two steps.

Jump: With probability $\omega_{x_i}/(d_{x_i} + \omega_{x_i})$, the walker jumps to x'_{i+1}, and $x_{i+1} = x'_{i+1}$;

Walk: Otherwise, the walker moves to a random neighbor $u \in \mathcal{U}$ of x_i in G, and $x_{i+1} = u$.

This sampling design ensures that we use only local knowledge of the three graphs to obtain a sample path $\mathcal{S} = (x_1, \ldots, x_B)$, which can yield unbiased estimates of target graph characteristics.

Estimator. We use the sample path $\mathcal{S} = (x_1, \ldots, x_B)$ generated by the random walk on G to design an estimator for θ_s as follows,

$$\hat{\theta}_s^{\mathrm{RW^T RW^A}} = \frac{1}{Z} \sum_{i=1}^{B} \frac{\mathbf{1}\{s \subseteq L(x_i)\}}{d_{x_i} + \omega_{x_i}}, \tag{13}$$

where $Z = \sum_{i=1}^{B} 1/(d_{x_i} + \omega_{x_i})$.

Theorem 3. *Estimator* (13) *is asymptotically unbiased.*

Proof. First we note that the random walk on target graph G has the same form of stationary distribution with RW$^{\mathrm{T}}$VS$^{\mathrm{A}}$ given by Eq. (6). (Note that we have removed jumper node j when calculating π_u in Eq. (9). Hence, π_u in Eq. (9) is equivalent to p_u in Eq. (6).) The remaining of the proof is similar to the proof of Theorem 2. □

IV. Experiments

In this section, we conduct experiments on both synthetic and real datasets to evaluate our proposed methods. Our goal is to demonstrate the unbiasedness of proposed estimators and study their estimation errors with respect to different factors such as sampling budget B (i.e., the number of sampled nodes in target graph) and parameter values, i.e., α and β.

For a graph, it is natural to consider node degrees as labels, i.e., $\mathcal{L} = \{0, 1, \ldots, M\}$ where M is the maximum degree in the graph. If we choose the characteristic function $L(u) = \{d_u\}$, and let $\theta_l \triangleq \theta_{\{l\}} = \frac{1}{n}\sum_{u \in \mathcal{U}} \mathbf{1}\{l \in L(u)\}, l \in \mathcal{L}$, then θ_l is the fraction of nodes with degree l in the graph, and $\{\theta_l\}_{l \in \mathcal{L}}$ therefore is the degree distribution (or PDF) of the graph. We can also consider choosing $L(u) = \{0, 1, \ldots, d_u - 1\}$, then θ_l is the fraction of nodes with degree larger than l in the graph, and $\{\theta_l\}_{l \in \mathcal{L}}$ is the complementary cumulative degree distribution (or CCDF) of the graph. To distinguish the two characteristic functions, in the following discussion, we will use θ_l to denote the parameter for PDF and Θ_l for CCDF.[2]

A. Experiments on Synthetic Data

In our first experiment, we examine the soundness of the proposed sampling methods using synthetic data.

Synthetic Data. We generate a hybrid social-affiliation network by connecting three Barabási-Albert (BA) graphs [31], namely G_1, G_2 and G_3. Each BA graph contains 100,000 nodes, and the three BA graphs have different average degrees: 4, 10 and 20 respectively. G_1 and G_3 are connected by one edge to form the target graph G. G_2 is the auxiliary graph G', and the affiliation graph G_b is formed by connecting nodes in G and G' according to the following two steps:

1) connect every node in G to a randomly selected node in G';
2) randomly choose 200,000 pairs of nodes in G and G' and connect them to form the remaining edges in G_b.

The first step ensures that every node in $\mathcal{U}$ satisfies $d_u^{(b)} > 0$ so that we can apply VS$^{\mathrm{A}}$ method on this dataset.

Results and Analysis. First we demonstrate that the proposed estimators $\breve{\theta}_l^{\mathrm{VS^A}}$, $\hat{\theta}_l^{\mathrm{RW^T VS^A}}$ and $\hat{\theta}_l^{\mathrm{RW^T RW^A}}$ are indeed asymptotically unbiased. To show this, we apply these sampling methods to estimate θ_2 and θ_{12}, i.e., the fraction of nodes with degree 2 and 12 in target graph, and compare their estimates to the ground truth for different sampling budgets B (from 0 to $0.01n$, where $n = |\mathcal{U}|$). The results are shown in Fig. 6. It is clear to see that when sampling budget B increases, all estimators converge to the ground truth. Hence, the proposed estimators are asymptotically unbiased.

Next, we study the estimation error of each estimator for estimating the PDF and CCDF of degree distribution. We choose the *normalized rooted mean squared error* (NRMSE) as a metric to evaluate the estimation error of an estimator, which is defined as follows

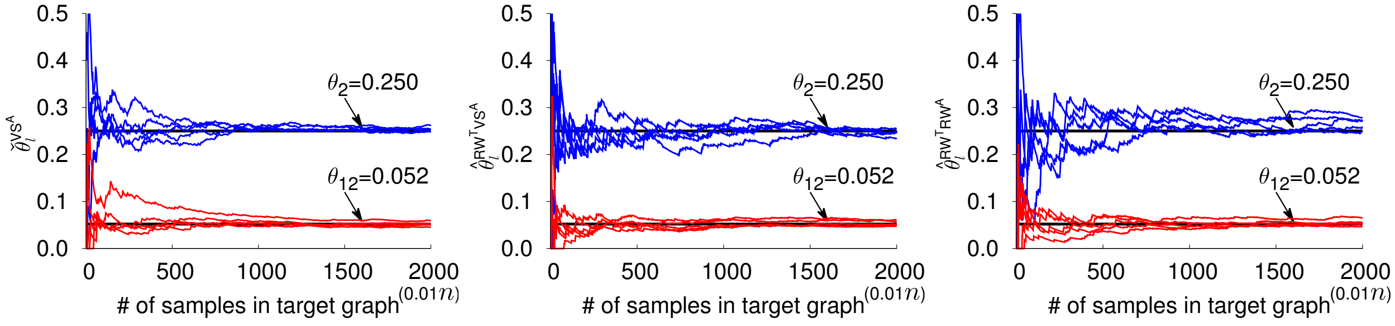
$$\mathrm{NRMSE}(\hat{\theta}_l) = \frac{\sqrt{\mathbb{E}\left[(\hat{\theta}_l - \theta_l)^2\right]}}{\theta_l}.$$

NRMSE measures the relative difference between an estimated value and a real value. The smaller the NRMSE, the better an estimator is. NRMSE can also be defined on $\hat{\Theta}_l$, which we omit here. To compare the NRMSE of different estimators, we fix the sampling budget B to be 1% of the number of nodes in the target graph, and calculate the averaged empirical NRMSE over $1,000$ runs. The results are shown in Fig. 7.

In Figure 7, we also show the NRMSE for a simple random walk (RW) estimator on the target graph. Because the target graph G has a bottleneck, i.e., its two components are connected by a single edge. RW can hardly converge within $B = 0.01n$ steps. Therefore, we observe that NRMSE for RW is almost the largest among all estimators. Comparing VS$^{\mathrm{A}}$ with RW in Figs. 7(a) and 7(d), we find that VS$^{\mathrm{A}}$ can provide smaller NRMSE for low degree nodes than RW. However, VS$^{\mathrm{A}}$ produces larger NRMSE for large degree nodes than RW. Therefore, VS$^{\mathrm{A}}$ can better estimate small degree nodes than large degree nodes in a graph.

The weakness of VS$^{\mathrm{A}}$ can be overcome by RW$^{\mathrm{T}}$VS$^{\mathrm{A}}$ and RW$^{\mathrm{T}}$RW$^{\mathrm{A}}$. From Figs. 7(b), 7(e) and 7(c), 7(f) we can see that when indirect jumps are incorporated into random walks in RW$^{\mathrm{T}}$VS$^{\mathrm{A}}$ and RW$^{\mathrm{T}}$RW$^{\mathrm{A}}$, NRMSE for large degree

[2] Usually, CCDF is the plot of choice when people show degree distribution. Therefore, in some experiments, we mainly show the results of CCDF in this paper, and results for PDF can be found in our technical report [30].

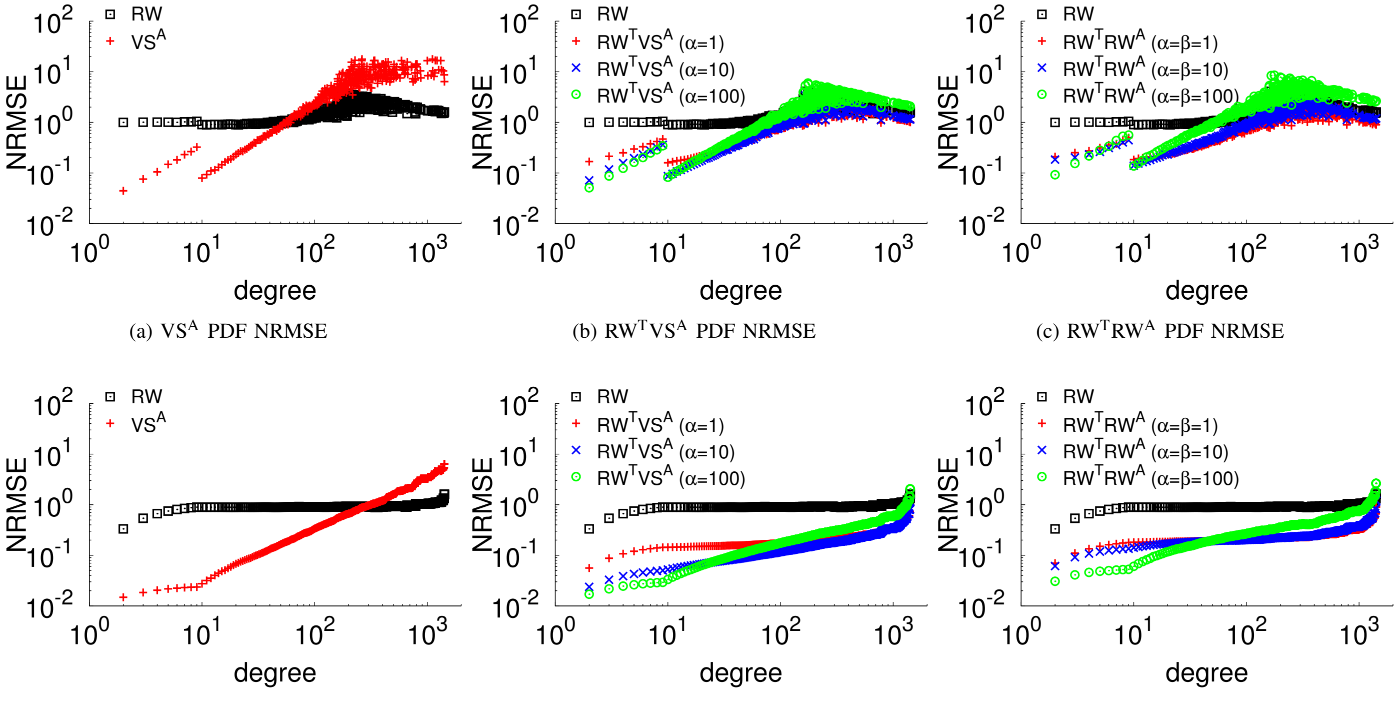


(a) $\breve{\theta}_l^{\mathrm{VS}^\mathrm{A}}$

(b) $\hat{\theta}_l^{\mathrm{RW}^\mathrm{T}\mathrm{VS}^\mathrm{A}}$ ($\alpha = 10$)

(c) $\hat{\theta}_l^{\mathrm{RW}^\mathrm{T}\mathrm{RW}^\mathrm{A}}$ ($\alpha = \beta = 10$)

Fig. 6. Asymptotic unbiasedness of the estimators ($l = 2, 12$).



(a) VS$^\mathrm{A}$ PDF NRMSE

(b) RW$^\mathrm{T}$VS$^\mathrm{A}$ PDF NRMSE

(c) RW$^\mathrm{T}$RW$^\mathrm{A}$ PDF NRMSE

(d) VS$^\mathrm{A}$ CCDF NRMSE

(e) RW$^\mathrm{T}$VS$^\mathrm{A}$ CCDF NRMSE

(f) RW$^\mathrm{T}$RW$^\mathrm{A}$ CCDF NRMSE

Fig. 7. NRMSE for different estimators. (Each result is averaged over $1,000$ runs, and $B = 0.01n$.)

nodes decreases, and NRMSE for small degree nodes remains smaller than RW. If we increase the probability of jumping at each step of random walk by increasing α and β, we observe that NRMSE for small degree nodes decreases, but NRMSE for large degree nodes increases. This behavior is similar to RWwJ [9,12] because we have declared that RW$^\mathrm{T}$VS$^\mathrm{A}$ and RW$^\mathrm{T}$RW$^\mathrm{A}$ are equivalent to RWwJs in their designs.

B. Experiments on LBSN Datasets

In the second experiment, we apply VS$^\mathrm{A}$ method on two real-world location-based social network (LBSN) datasets to solve the problem mentioned in Example 1, i.e., measure user characteristics within an area of interest.

LBSN Datasets. We obtain two public LBSN datasets from Brightkite and Gowalla [32]. Brightkite and Gowalla are once two popular LBSNs where users shared their locations by checking-in. Users in the two social networks are also connected by undirected friendship relations, which form two user social networks. The statistics of these two datasets are summarized in Table I.

Because we are only interested in users that have check-ins, VS$^\mathrm{A}$ can be applied to these two datasets. Suppose the we want to measure characteristics of users located around New York City (NYC), as specified by a rectangle region on a map: latitude range $40.4° \sim 41.4°$, longitude range $-74.3° \sim -73.3°$ (see Fig. 8). The goal is to estimate degree distribution of the users who checked in this region. As we explained in Introduction, directly sampling users is inefficient. Here, we apply the VS$^\mathrm{A}$ method along with a venue sampling method — Random Region Zoom-In (RRZI) [26] to sample users in NYC more efficiently.

Venue Sampling. RRZI [26] utilizes a venue query API provided by most LBSNs to sample venues on a map. The API requires a user to specify a rectangle region by provid-

TABLE I
SUMMARY OF TWO LBSN DATASETS.

dataset		Brightkite	Gowalla
G	network type	undirected	undirected
	# of users	$58,228$	$196,591$
	# of friendship edges	$214,078$	$950,327$
	# of users in LCC[1]	$56,739$	$196,591$
	# of edges in LCC	$212,945$	$950,327$
G' and G_b	# of venues	$772,966$	$1,280,969$
	# of users having check-ins	$51,406$	$107,092$
	# of check-ins	$4,491,143$	$6,442,890$
G' and G_b for NYC	# of venues in NYC[2]	$23,484$	$26,448$
	# of users checking in NYC	$4,257$	$7,399$
	# of check-ins in NYC	$33,656$	$113,423$

[1] The largest connected component.
[2] The New York City (Fig. 8).

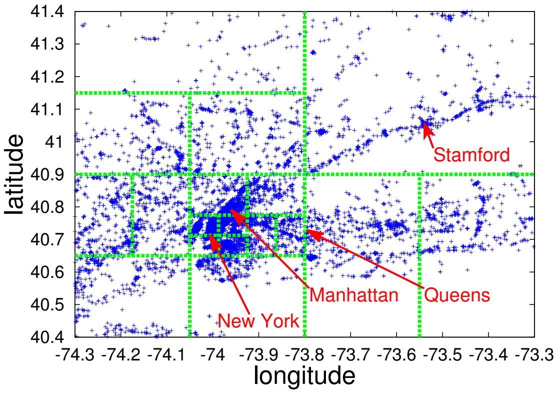


Fig. 8. Venue distribution in New York City and illustration of accessible subregions used by RRZI. Each subregion contains less than K venues.

ing the south-west and north-east corners latitude-longitude coordinates, and then the API returns a set of venues in this region. Usually, the API can only return at most K venues in a queried region. RRZI regularly zooms in the region until the subregion is fully *accessible*, i.e., the API returns less than K venues in the subregion. The zooming-in process is equivalent to dividing the region into many non-overlapping accessible subregions (as illustrated in Fig. 8), and each subregion is associated with a fixed probability related to the zooming-in strategy. This feature enables RRZI to provide samples of venues within an area of interest.

Results. Combining VS$^{\mathrm{A}}$ with RRZI, we conduct two experiments to indirectly sample users in NYC on Brightkite and Gowalla. We totally sample 5% of venues in NYC and calculate the degree distribution of users in NYC. The results are shown in Figs. 9 and 10.

From Figures 9(a) and 10(a), we observe that RRZI-VS$^{\mathrm{A}}$ method can provide good estimates of user characteristics in NYC. The estimates for low degree users are better than high degree users, and this is clear to see from the NRMSE plots in Figs. 9(b), 9(c) and 10(b), 10(c). This feature coincides with our previous analysis using synthetic data. From the NRMSE plots, we can also find an approximate law that a larger K, i.e., the maximum number of venues the API can return, reduces the estimation error of RRZI-VS$^{\mathrm{A}}$. However, it is not true for estimating large degree users on Gowalla in Fig. 10(c). In fact, a better way to reduce estimation error is to combine VS$^{\mathrm{A}}$ with other better venue sampling methods introduced in [24]–[26]. However, we omit this due to space limitation.

C. Experiments on Mtime Dataset

In the third experiment, we apply RW$^{\mathrm{T}}$VS$^{\mathrm{A}}$ and RW$^{\mathrm{T}}$RW$^{\mathrm{A}}$ on Mtime to measure Mtime user characteristics as we have introduced in Example 2.

Mtime Dataset. As we introduced in Example 2, users and actors in Mtime naturally form a hybrid social-affiliation network. To build a ground-truth dataset as the testbed of RW$^{\mathrm{T}}$VS$^{\mathrm{A}}$ and RW$^{\mathrm{T}}$RW$^{\mathrm{A}}$, we downloaded the complete Mtime network data by traversing user IDs ranging from 100000 to 10000000, and actor IDs ranging from 892000 to 2100000.

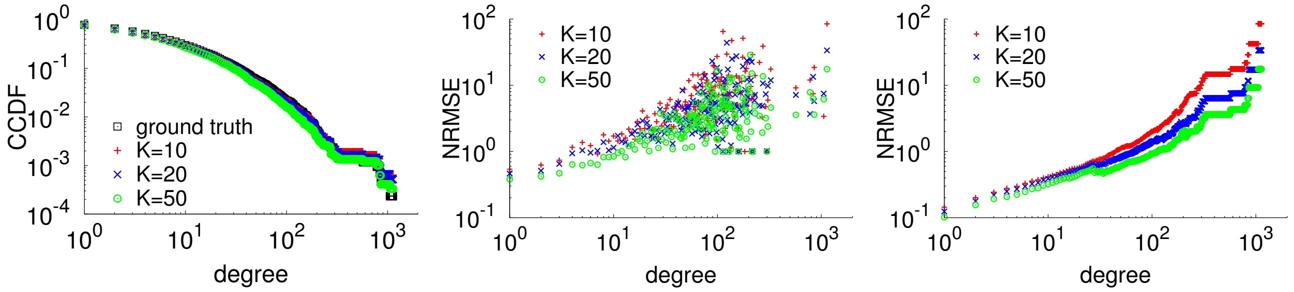
For each Mtime user, we collect the set of users he/she follows and users who follow him. This builds up a directed follower network among Mtime users. Each Mtime user maintains a list including a subset of movie actors he/she is interested in. This information is used to build up the fan-relations between users and actors. For each movie actor, we collect the movies he/she participated in, and if two actors participated in a same movie, we connect them. This builds up a cooperative network among actors. The complete Mtime dataset is summarized in Table II.

TABLE II
SUMMARY OF THE MTIME DATASET.

G	user follower network type	directed
	total users (isolated and non-isolated)[3]	$1,878,127$
	# of non-isolated users in follower network	$1,035,164$
	# of following relations	$14,861,383$
	# of users in LCC	$987,055$
	# of following relations in LCC	$14,791,482$
G'	actor cooperative network type	undirected
	total actors (isolated and non-isolated)	$1,123,340$
	# of non-isolated actors in cooperative network	$1,122,166$
	# of cooperative relations	$10,344,364$
	# of actors in LCC	$1,114,065$
	# of cooperative relations in LCC	$10,328,904$
G_b	# of fan relations	$225,558,343$
	# of users following actors	$1,419,339$
	# of isolated users following actors	$842,963$
	# of actors having fans	$441,413$
	# of isolated actors having fans	$1,174$
	# of isolated actors having only isolated fans	225
	# of isolated users following only isolated actors	393

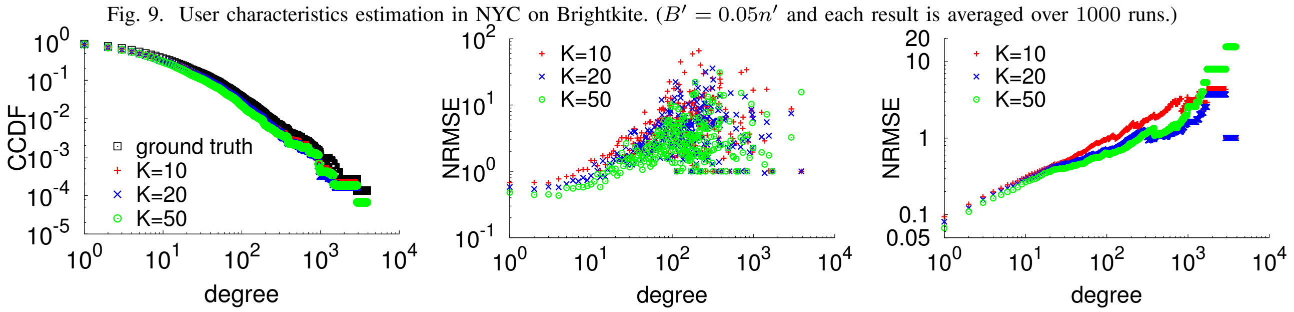
[3] An isolated node in a graph is a node with zero degree.

Analysis of the Dataset. First we provide some statistics of Mtime dataset. In Table II, we compare the first block with second block, which are related to target graph G and auxiliary graph G' respectively. We find that about 19% of the user IDs and 93% of the actor IDs are valid. This indicates that conducting UNI on the auxiliary graph is more efficient than conducting UNI on the target graph. Moreover, we find that more than 47% of the Mtime users are not in LCC, but the same number for actors is less than 0.1%. This indicates that the auxiliary graph is better connected than the target graph. Although a large fraction of users are isolated nodes in the



(a) RRZI-VS^A estimates (b) RRZI-VS^A PDF NRMSE (c) RRZI-VS^A CCDF NRMSE

Fig. 9. User characteristics estimation in NYC on Brightkite. ($B' = 0.05n'$ and each result is averaged over 1000 runs.)



(a) RRZI-VS^A estimates (b) RRZI-VS^A PDF NRMSE (c) RRZI-VS^A CCDF NRMSE

Fig. 10. User characteristics estimation in NYC on Gowalla. ($B' = 0.05n'$ and each result is averaged over 1000 runs.)

target graph, from the last block (regarding the affiliation graph G_b), we find that almost all the isolated users are connected to non-isolated actors (except a few hundreds of them). So the majority of isolated users are indirectly connected to other users through actors. This is illustrated in Fig. 11. The advantage of introducing the hybrid social-affiliation network is now clear for Mtime dataset, i.e., we can study a larger user space than simply the LCC of target graph (when UNI on target graph is inefficient or not allowed).

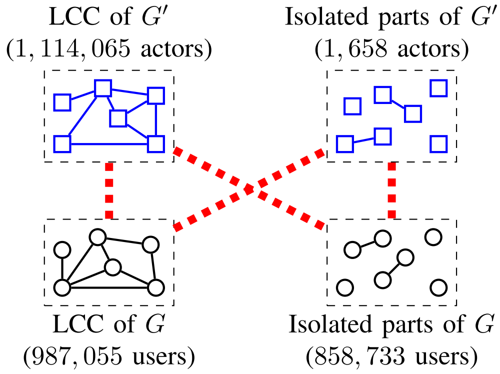


Fig. 11. The Mtime network components. Dashed red lines denote fan relations between actors and users.

Results. Using the Mtime dataset as a testbed, we demonstrate that RW^{T}VS^A and RW^{T}RW^A methods can provide good estimates of user characteristics. Although the user follower network is directed, we can build an undirected version of the target graph on-the-fly while sampling because a user's incoming and out-going neighbors are known once the user is queried [10,12]. Different from previous experiments, here the user labels can be in-degrees or out-degrees. Hence, we will evaluate the in-degree and out-degree distribution estimations of the two estimators, respectively.

Results of the RW^{T}VS^A method are depicted in Fig. 12. In Fig. 12(a) and (e), we show the in-degree and out-degree CCDF estimates. We can see that RW^{T}VS^A can provide unbiased estimates. From Fig. 12(b) and (f), we observe that when sampling budget increases, the NRMSE decreases for both in-degree and out-degree estimations. From Fig. 12(c) and (g), we observe that when more jumps are allowed by increasing α from 1 to 100, estimation accuracy also increases.

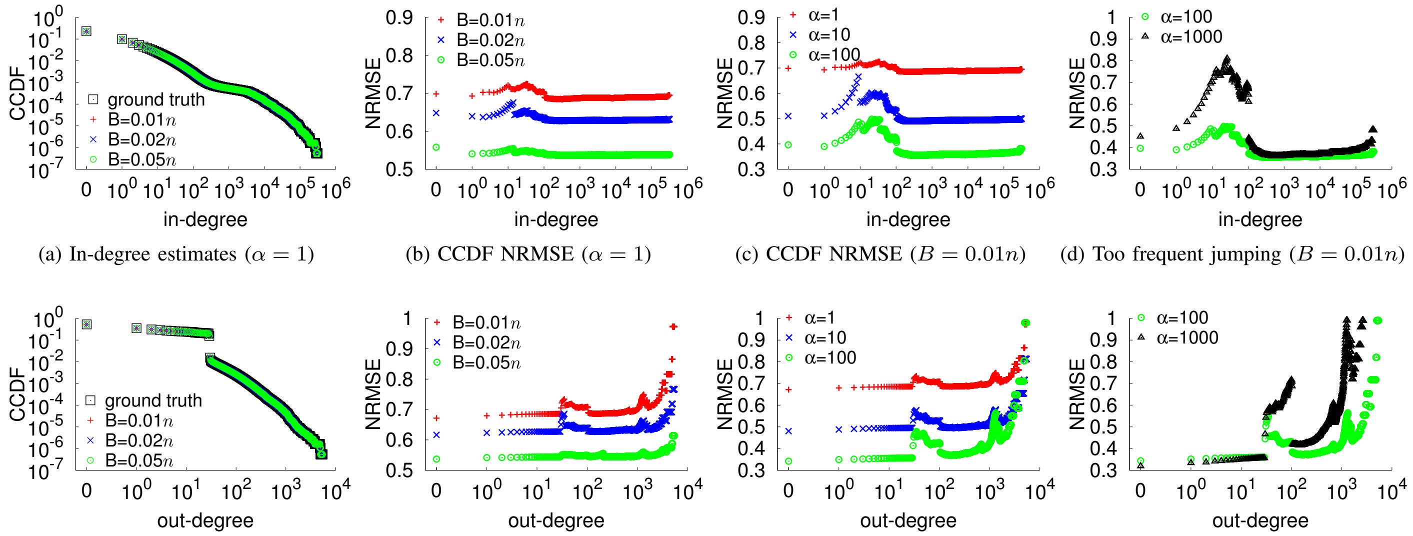
Results for RW^{T}RW^A are similar to the results of RW^{T}VS^A, and we show them in Fig. 13. First, from Fig. 13(a) and (e), we observe that RW^{T}RW^A can also well estimate the in-degree and out-degree distributions. Second, from Fig. 13(b) and (f), we can find that when sampling budget increases, the estimation accuracy increases significantly for both in-degree and out-degree estimations. Last, from Fig 13(c) and (g), we find that when more jumps are allowed (by increasing α and β), the NRMSE also decreases.

However, it is worth noting that α and β should not be too large for both RW^{T}VS^A and RW^{T}RW^A. Because we know that when $\alpha \to \infty$, RW^{T}VS^A becomes VS^A, which is biased on the Mtime dataset, and hence causes large NRMSE. Similar behavior happens to RW^{T}RW^A, too. We depict these observations in Figs. 12(d), 12(h) and 13(d), 13(h).

V. Related Work

We briefly review the related literature in this section.

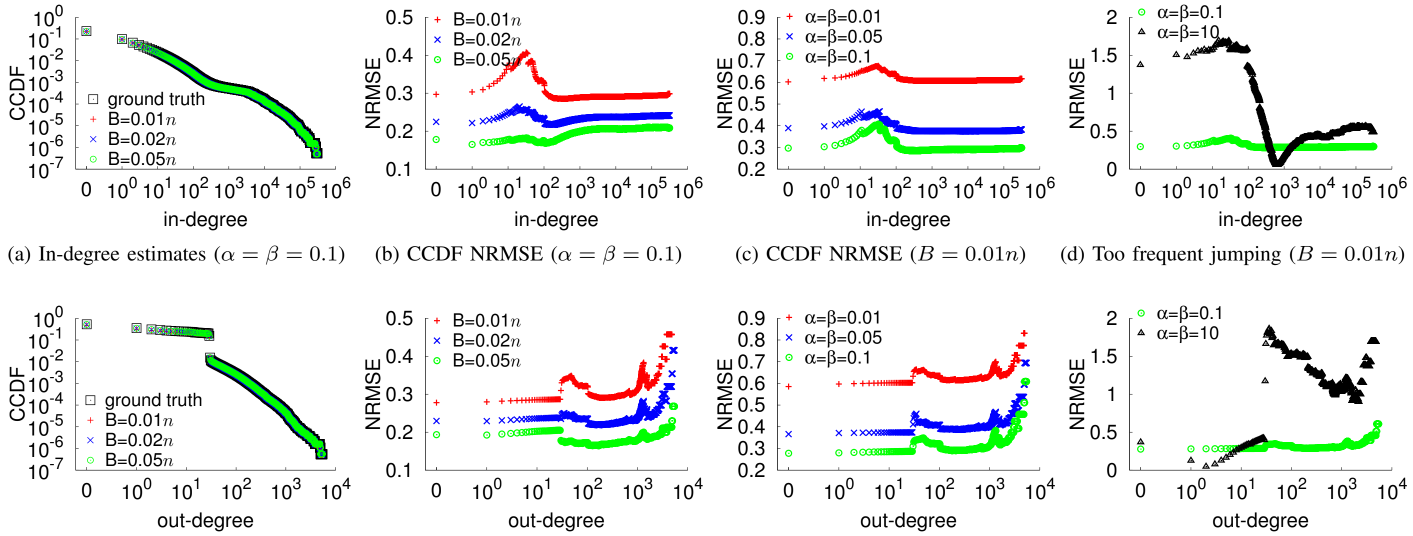
Sampling methods, especially random walk-based graph sampling methods, have been widely used to characterize large-scale complex networks. These applications include, but are not limited to, estimating peer statistics in peer-to-peer networks [8,33], uniformly sampling users from online social



(a) In-degree estimates ($\alpha = 1$) (b) CCDF NRMSE ($\alpha = 1$) (c) CCDF NRMSE ($B = 0.01n$) (d) Too frequent jumping ($B = 0.01n$)

(e) Out-degree estimates ($\alpha = 1$) (f) CCDF NRMSE ($\alpha = 1$) (g) CCDF NRMSE ($B = 0.01n$) (h) Too frequent jumping ($B = 0.01n$)

Fig. 12. $\text{RW}^{\text{T}}\text{VS}^{\text{A}}$ degree distribution estimates and NRMSE analysis. Each result is averaged over $10,000$ runs.



(a) In-degree estimates ($\alpha = \beta = 0.1$) (b) CCDF NRMSE ($\alpha = \beta = 0.1$) (c) CCDF NRMSE ($B = 0.01n$) (d) Too frequent jumping ($B = 0.01n$)

(e) Estimates ($\alpha = \beta = 0.1$) (f) CCDF NRMSE ($\alpha = \beta = 0.1$) (g) CCDF NRMSE ($B = 0.01n$) (h) Too frequent jumping ($B = 0.01n$)

Fig. 13. $\text{RW}^{\text{T}}\text{RW}^{\text{A}}$ degree distribution estimates and NRMSE analysis. Each result is averaged over $10,000$ runs.

networks [11,13,14,34], characterizing structure properties of large-scale networks [35]–[38], and measuring statistics of point-of-interests on maps [26]. The above literature is mostly concerned with sampling methods that seek to *directly* sample nodes (or samples) in target graphs (or sample spaces). However, direct sampling is not always efficient as we argued in this work.

When the target graph (or sample space) can not be directly sampled or direct sampling is inefficient, several methods based on *graph manipulation* have been proposed to improve sampling efficiency. For example, Gjoka et al. [39] study an approach to improve sampling efficiency through building a *multigraph* using different kinds of relations (i.e., edges) that exist on an OSN. A multigraph is better connected than any individual graph formed by only one kind of relations. Therefore, the random walk can converge fast on this multigraph. Zhou et al. [40] exploit several criteria to rewire the target graph on-the-fly to increase the graph conductance [17] and reduce mixing time of random walks. Our method differs from theirs that we do not manipulate target graphs. We study a new approach that utilizes an auxiliary graph and an affiliation graph to assist sampling on target graph indirectly.

Birnbaum and Sirken [41] designed a survey method for estimating the number of diagnosed cases of a rare disease in a population. Directly sampling patients of a rare disease from the huge human population is obviously inefficient, so they studied how to sample hospitals so as to sample patients indirectly. Their method motivates us to design the VS^A method. However, as we pointed out, VS^A method cannot sample nodes that are not connected to auxiliary graph, and we overcome this problem by designing $\text{RW}^\text{T}\text{VS}^\text{A}$ and $\text{RW}^\text{T}\text{RW}^\text{A}$ methods. Our work also complements existing sampling methods related to random walk with jumps [9,12,14] by removing the necessity of uniform node sampling on target graphs.

VI. Conclusion

When graphs become large in scale, sampling methods become necessary tools in the study of characterizing their properties. Among these sampling methods, random walk-based crawling methods have gained popularity. However, if the graph under study is not well connected, random walk-based graph sampling methods suffer from the slow mixing problem. In this work, we observe that a graph usually does not exist in isolation. Usually, the target graph is accompanied with an auxiliary graph and an affiliation graph, and they form a hybrid social-affiliation network together. We find that the target graph becomes better connected with the assistances of the other two graphs. This new viewpoint brings benefits to the graph sampling framework. We design three sampling methods to measure the target graph from this new viewpoint, and these methods are demonstrated to be effective on both synthetic and real datasets. Therefore, our method complements existing methods in the literature of graph sampling.

Acknowledgement

This research was sponsored in part by the ARL Cooperative Agreement W911NF-09-2-0053 of US; National Natural Science Foundation (61221063, U1301254, 61202392, 61103240, 61103241, 91118005), 863 High Tech Development Plan (2012AA011003) and 111 International Collaboration Program, of China, and the Hong Kong RGC GRF 415013.

References

[1] D. J. Watts, "The new science of networks," *Annual Review of Sociology*, vol. 30, no. 1, pp. 243–270, 2004.

[2] D. Lazer and et al., "Computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009.

[3] S. Aral and D. Walker, "Identifying influential and susceptible members of social networks," *Science*, vol. 337, no. 6092, pp. 337–341, 2012.

[4] S. Asur and B. A. Huberman, "Predicting the future with social media," in *WI-IAT*, 2010.

[5] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[6] M. Mondal, B. Viswanath, P. Druschel, K. P. Gummadi, A. Clement, A. Mislove, and A. Post, "Defending against large-scale crawls in online social networks," in *CoNEXT*, 2012.

[7] "Weibo API limits," http://open.weibo.com/wiki/Rate-limiting, 2014.

[8] L. Massoulié, E. L. Merrer, A.-M. Kermarrec, and A. Ganesh, "Peer counting and sampling in overlay networks: Random walk methods," in *PODC*, 2006.

[9] K. Avrachenkov, B. Ribeiro, and D. Towsley, "Improving random walk estimation accuracy with uniform restarts," in *WAW*, 2010.

[10] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *IMC*, 2010.

[11] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Practical recommendations on crawling online social networks," *JSAC*, vol. 29, no. 9, pp. 1872–1892, 2011.

[12] B. Ribeiro, P. Wang, F. Murai, and D. Towsley, "Sampling directed graphs with random walks," in *INFOCOM*, 2012.

[13] C.-H. Lee, X. Xu, and D. Y. Eun, "Beyond random walk and Metropolis-Hastings samplers: Why you should not backtrack for unbiased graph sampling," in *SIGMETRICS*, 2012.

[14] X. Xu, C.-H. Lee, and D. Y. Eun, "A general framework of hybrid graph sampling for complex network analysis," in *INFOCOM*, 2014.

[15] P. Wang, J. Zhao, J. C. S. Lui, D. Towsley, and X. Guan, "Sampling node pairs over graphs," in *ICDE*, 2013.

[16] S. Meyn and R. L. Tweedie, *Markov Chains and Statistic Stability*, 2nd ed. Cambridge University Press, 2009.

[17] A. Sinclair and M. Jerrum, "Approximate counting, uniform generation and rapidly mixing markov chains," *Information and Computation*, vol. 82, no. 1, pp. 93–133, 1989.

[18] A. Mohaisen, A. Yun, and Y. Kim, "Measuring the mixing time of social graphs," in *IMC*, 2010.

[19] "Sina Weibo," http://weibo.com, July 2014.

[20] "Weibo place," http://place.weibo.com, July 2014.

[21] "Pinterest," http://www.pinterest.com, July 2014.

[22] "Weibo search API," http://open.weibo.com/wiki/2/location/pois/search/by_area, July 2014.

[23] "Foursquare search API," https://developer.foursquare.com/docs/venues/search, July 2014.

[24] Y. Li, M. Steiner, L. Wang, Z.-L. Zhang, and J. Bao, "Dissecting Foursquare venue popularity via random region sampling," in *CoNEXT*, 2012.

[25] Y. Li, L. Wang, M. Steiner, J. Bao, and T. Zhu, "Region sampling and estimation of geosocial data with dynamic range calibration," in *ICDE*, 2014.

[26] P. Wang, W. He, and X. Liu, "An efficient sampling method for characterizing points of interests on maps," in *ICDE*, 2014.

[27] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[28] "Mtime," http://www.mtime.com, July 2014.

[29] C. P. Robert and G. Casella, *Monte Carlo Statistic Methods*, 2nd ed. Springer, 2004.

[30] "Technical report," http://nskeylab.xjtu.edu.cn/dataset/jzzhao/ICDE2015TR.pdf, Nov. 2014.

[31] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[32] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *KDD*, 2011.

[33] C. Gkantsidis, M. Mihail, and A. Saberi, "Random walks in peer-to-peer networks: Algorithms and evaluation," *Performance Evaluation*, vol. 63, no. 3, pp. 241–263, 2006.

[34] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in Facebook: A case study of unbiased sampling of OSNs," in *INFOCOM*, 2010.

[35] L. Katzir, E. Liberty, and O. Somekh, "Estimating sizes of social networks via biased sampling," in *WWW*, 2011.

[36] S. J. Hardiman and L. Katzir, "Estimating clustering coefficients and size of social networks via random walk," in *WWW*, 2013.

[37] C. Seshadhri, A. Pinar, and T. G. Kolda, "Triadic measures on graphs: The power of wedge sampling," in *SDM*, 2013.

[38] P. Wang, J. C. Lui, B. Ribeiro, D. Towsley, J. Zhao, and X. Guan, "Efficiently estimating motif statistics of large networks," *TKDD*, 2014.

[39] M. Gjoka, C. T. Butts, M. Kurant, and A. Markopoulou, "Multigraph sampling of online social networks," *JSAC*, vol. 29, no. 9, pp. 1893–1905, 2011.

[40] Z. Zhou, N. Zhang, Z. Gong, and G. Das, "Faster random walks by rewiring online social networks on-the-fly," in *ICDE*, 2013.

[41] Z. W. Birnbaum and M. G. Sirken, "Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates," *Vital and Health Statistics*, vol. 2, no. 11, pp. 1–8, 1965.