On the Estimation Accuracy of Degree Distributions from Graph Sampling

Bruno Ribeiro and Don Towsley

Computer Science Department
University of Massachusetts
Amherst, MA, 01003
{ribeiro, towsley}@cs.umass.edu

***Abstract*— Estimating characteristics of large graphs via sampling is vital in the study of complex networks. In this work, we study the Mean Squared Error (MSE) associated with different sampling methods for the degree distribution. These sampling methods include independent random vertex (RV) and random edge (RE) sampling, and crawling methods such as random walks (RWs) and the widely used Metropolis-Hastings algorithm for uniformly sampling vertices (MHRWu). We see that the RW MSE is proportional to the RE MSE and inversely proportional to the spectral gap of the RW transition probability matrix. We also determine conditions under which RW is preferable to RV. Finally, we present an approximation of the MHRWu MSE. We evaluate the accuracy of our approximations and bounds through simulations on large real world graphs.**

I. Introduction

A number of recent studies [2], [5], [6], [11], [17], [12], [18] (to cite a few) are dedicated to the characterization of network graphs. Here a network is represented by an undirected graph with labeled vertices and edges. Network characteristics of interest include the degree distribution, the average number of copies of a file in a peer-to-peer (P2P) network [6], [17], the assortativity coefficient [14], and the global clustering coefficient [14].

Characterizing graphs requires querying vertices and/or edges; each query has an associated resource cost (time, bandwidth, money). Querying the whole graph is often too costly. As a result, researchers have turned their attention to the estimation of graph characteristics based on incomplete (sampled) data.

RV sampling: In networks where each vertex is assigned a unique user-id (e.g., travelers and their passport numbers, Facebook, MySpace, Flickr, and Livejournal) a widespread practice is to perform random vertex (RV) sampling by querying randomly generated user-ids. However, uniform RV sampling may be undesirable when the user-id space is sparsely populated (in MySpace the ratio between the number valid users retrieved and the total number of queries is 10% [13]). Moreover, queries are often subject to resource constraints (e.g., queries are rate-limited in Flickr, Livejournal [11], and Bittorrent [9]). As we see in this work, even when RV sampling is not severely resource-constrained, some characteristics may be better estimated with other sampling methods (e.g., the tail of the degree distribution of a graph).

RE sampling: In independent Random Edge (RE) sampling, a vertex is sampled by first sampling an edge independent and uniformly from the set of edges, and then randomly choosing one of the edge end points. In practice one should use both end points of a sampled edge. However, in order to simplify our analysis, we consider just one sampled vertex for each sampled edge. In real world networks it is more difficult to randomly sampling edges than vertices. Edges are not often associated to unique IDs that can be queried and online social networks such as Facebook, Twitter, MySpace, Livejournal, and Flickr, among others, do not provide an API that allows randomly sampling of edges.

RW sampling: An alternative, and often resource wise cheaper, way to sample a network is by means of a random walk (RW). RW sampling is preferred to other types of graph crawling, such as the breadth-first crawling used in [11], as one can obtain asymptotically unbiased estimates of a number of graph characteristics such as fraction of vertices with a given label [18], the degree distribution [18], and, more recently, assortativity and global clustering coefficients [14]. A RW samples a graph by moving a particle (walker) from a vertex to a neighboring vertex (over an edge). The probability by which the walker selects the next neighboring vertex determines the probability by which vertices and edges are sampled. We denote *standard RW* or just *RW* a random walk that sample neighbors *uniformly*. A Metropolis-Hastings walker, as seen later, selects the next neighboring vertex using a different rule. RWs are popular for sampling networks [12], [18] in order to estimate their characteristics. One of the reasons behind the popularity of RW sampling is that it does not query invalid users as does RV sampling.

MHRW sampling: The Metropolis-Hastings Random Walk (MHRW) is an accept-reject random walk-based sampling process that samples vertices according to a target distribution γ. In this work we are mostly interested in a MHRW that samples vertices uniformly, which we denote MHRWu. MHRWu has been used to uniformly sample peers in peer-to-peer networks [17] and Web pages [7]. Unfortunately, MHRWu is empirically known to have large estimation errors compared to RW estimates [5].

Contributions

This paper presents the following contributions:

1) In Section III we prove that the Mean Squared Error (MSE) obtained by a stationary sequence of n RW

sampled vertices is upper bounded by the MSE of n RE sampled vertices divided by $(1-\alpha)$, where α is the absolute value of the second most dominant eigenvalue of the RW transition probability matrix.

2) RW estimates have been observed to be more accurate than estimates obtained by MHRWu [5], [12]. We study how the Metropolis-Hastings mechanism tends to induce larger estimation errors than RW and RV sampling.

Outline

The outline of this work is as follows. Section II presents definitions used in this paper. Section III presents an upper bound of the MSE of RW sampling as a function of the MSE of RE sampling and α, the absolute value of the second most dominant eigenvalue of the RW transition probability matrix. In Section IV we study how the Metropolis-Hastings mechanism tends to induce larger estimation errors than RW or even RV sampling. Section VI present simulation results that help corroborate our theoretical analysis. And finally Section VIII presents our conclusions.

II. Definitions

Let $G = (V, E)$ be an undirected connected non-bipartite graph and let d_v, $v \in V$, be the degree of vertex v. We denote $\text{vol}(V) \triangleq \sum_{\forall v \in V} d_v$. We want to estimate

$$F = \sum_{\forall v \in V} f(v). \tag{1}$$

from a sequence of vertices sampled from G. Let $(Z_1, \ldots, Z_n)$ be a stationary sequence of n sampled vertices, where $P[Z_t = v] = \beta_v > 0$, $\forall v \in V$, $t = 1, \ldots, n$. Then

$$\hat{F}(Z_1, \ldots, Z_n) \triangleq \frac{1}{n} \sum_{t=1}^{n} \frac{f(Z_t)}{\beta_{Z_t}}, \; Z_i \in V, i = 1, \ldots, n. \tag{2}$$

is an unbiased estimate of F.

Let $\text{var}(X)$ define the variance of X. The Mean Squared Error (MSE) of $\hat{F}(Z_1, \ldots, Z_n)$ is

$$E[(\hat{F}(Z_1, \ldots, Z_n) - F)^2] = \text{var}(\hat{F}(Z_1, \ldots, Z_n)), \tag{3}$$

as $E[\hat{F}(Z_1, \ldots, Z_n)] = F$.

III. A tight upper bound of the RW estimation error

Let $(X_1, \ldots, X_n)$ be a sequence of vertices sampled by a RW in steady state. Let $\pi = (\pi_v : v \in V)$, denote the steady state probability distribution of the RW. A RW on an undirected non-bipartite graph edges are sampled by the RW with equal probability, $1/|E|$, and the probability that a vertex $v \in V$ is the i-th sampled vertex is $P[X_i = v] = \pi_v = d_v/\text{vol}(V)$, $i = 1, \ldots, n$. To use the estimator in (2) we need d_v, which the RW observes when it visits v, and $\text{vol}(V)$, which unfortunately requires knowing the number of edges in G. When the number of edges is unknown, it is common practice to use the Horvitz-Thompson estimator to obtain an asymptotically unbiased estimate of F:

$$\hat{F}'(X_1, \ldots, X_n) = \frac{\sum_{i=1}^{n} f(X_i)/d(X_i)}{\sum_{i=1}^{n} 1/d(X_i)}, \tag{4}$$

where here $d(v)$ denotes the degree of node $v \in V$. While the Horvitz-Thompson is only asymptotically unbiased, in practice we observe that the error from not knowing $\text{vol}(V)$ is negligible, as seen in Section V. Thus, in what we study the MSE of the estimator in (2) assuming $\text{vol}(V)$ is known.

In what follows we present a tight upper bound of the MSE of a stationary RW. Let $(Y_1, \ldots, Y_n)$ be a sequence of RE sampled vertices. We show that the MSE, eq.(3), of $(X_1, \ldots, X_n)$ is upper bounded by a function of the MSE of $(Y_1, \ldots, Y_n)$ and α, where $0 \leq \alpha < 1$ is the absolute value of the second most dominant eigenvalue of the RW transition probability matrix.

We now focus on the second most dominant eigenvalue of the RW. Let $\mathbf{A} = [A_{ij}], i = 1, \ldots, |V|$, be the adjacency matrix of G, $A_{ij} = 1$ iff $(i, j) \in E$, otherwise $A_{ij} = 0$. Let

$$\mathbf{D} = \begin{bmatrix} d_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_{|V|} \end{bmatrix}$$

be a diagonal matrix whose diagonal elements are the degrees of the vertices in G. Let $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$ be the one-step RW transition probability matrix. The probability that a RW reaches vertex j from i in t steps is

$$p_{ij}^{(t)} = (\mathbf{P}^t)_{ij}\,.$$

The stationary distribution of the RW is $\pi = \mathbf{P}\pi$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{|V|}$ be the eigenvalues of $\mathbf{P}$. It follows from the fact that G is an undirected connected non-bipartite graph (and $\mathbf{P}$ is a stochastic matrix) and the Frobenius-Perron Theorem that $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|V|} > -1$ [10]. The absolute value of the second most dominant eigenvalue is defined as

$$\alpha \triangleq \max(\lambda_2, -\lambda_{|V|}). \tag{5}$$

A RW is fast mixing when α is sufficiently small (we choose to use a vague definition of fast mixing as there are numerous different definitions of "fast mixing" in the literature).

In the following theorem (Theorem III.1) we show that the estimation error of a RW can be upper bounded by the estimation error of RE sampling and α.

Theorem III.1. *Let $G = (V, E)$ be an undirected connected non-bipartite graph. Let $(X_1, \ldots, X_n)$ be a sequence of vertices sampled by a stationary RW on G, $n \geq 1$. Let $(Y_1, \ldots, Y_n)$ be a sequence of RE sampled vertices. Let*

$$\hat{F}(v_1, \ldots, v_n) \triangleq \frac{1}{n} \sum_{t=1}^{n} \frac{f(v_t)}{\pi_{v_t}}, \; v_i \in V, i = 1, \ldots, n$$

and let α be the absolute value of the second most dominant eigenvalue of the RW transition probability matrix.

Then

$$\text{var}(\hat{F}(X_1,\ldots,X_n)) \le \frac{\text{var}(\hat{F}(Y_1,\ldots,Y_n))}{(1-\alpha)}. \quad (6)$$

Proof. Generally, the RW estimation error is in part due to the dependence among sampled vertices $(X_1,\ldots,X_n)$. This dependence can be characterized by the covariance function, $\text{cov}(g(X_u),g(X_t))$, $u,t \in \{1,\ldots,n\}$. Upper and lower bounds of the covariance of $g(X_1)$ and $g(X_t)$, $t=2,\ldots,n$ can be obtained as follows.

Let $\pi = \mathbf{P}\pi$ and $\mathbf{S} = \pi^{1/2}\mathbf{P}\pi^{1/2}$. Matrix $\mathbf{S}$ is symmetric with the same eigenvalues of $\mathbf{P}$. The eigenvector of $\mathbf{S}$ corresponding to eigenvalue $\lambda_1 = 1$ is π. The Courant-Fischer theorem [8, Theorem 4.2.11] gives the second largest eigenvalue of $\mathbf{S}$ and its eigenvector w:

$$\lambda_2 = \max_{w\,:\,\langle w,\pi\rangle=0} \frac{\sum_{\forall i\in V}\sum_{\forall j\in V} w_i w_j \pi_i \mathbf{P}_{i,j}}{\sum_{\forall u\in V} w_u^2 \pi_u} \quad (7)$$

and the smallest eigenvalue of $\mathbf{S}$ and its eigenvector w':

$$\lambda_{|V|} = \min_{w'\,:\,\langle w',\pi\rangle=0} \frac{\sum_{\forall i\in V}\sum_{\forall i\in V} w'_i w'_j \pi_i \mathbf{P}_{i,j}}{\sum_{\forall u\in V} w'^2_u \pi_u}. \quad (8)$$

Let function $f: V\to\mathbb{R}$ be as defined in (1). Define $g(v) = f(v)/\pi_v - F$, $v\in V$. Note that $E[g(X_i)]=0$, $i=1,\ldots,n$. We are interested in the variance of

$$\hat{F}(X_1,\ldots,X_n) = \frac{1}{n}\sum_{t=1}^{n}\frac{f(X_t)}{\pi_{X_t}}, i=1,\ldots,n,$$

where $X_1,\ldots,X_n$ are identically distributed. Let $(Y_1,\ldots,Y_n)$ be a sequence of RE sampled vertices. Note that

$$\frac{1}{n}\text{var}(g(X_1)) = \frac{1}{n}\text{var}(\hat{F}(Y_1)) = \text{var}(\hat{F}(Y_1,\ldots,Y_n))$$

and that

$$\text{var}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right) = \text{var}(\hat{F}(X_1,\ldots,X_n)).$$

The variance of the above sum can be broken into the variance and covariances of its summands [15, pp. 265]

$$\begin{aligned}\text{var}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right) =& \frac{1}{n}\text{var}(g(X_1)) + \\ & \frac{2}{n}\sum_{t=2}^{n}\frac{n-t}{n}\text{cov}(g(X_1),g(X_t)).\end{aligned} \quad (9)$$

These variances and covariances can be written as

$$\text{cov}(g(X_1),g(X_t)) \triangleq \sum_{\forall v\in V}\sum_{\forall u\in V} g(v)g(u)\pi_v p_{v,u}^{(t-1)}$$

for $1 < t \le n$; and

$$\text{var}(g(X_i)) \triangleq \sum_{\forall u\in V} g(v)^2\pi_v,\ i=1,\ldots,n.$$

The variances and covariance of $g(X_1)$ and $g(X_2)$ can be easily bounded by λ_2 and $\lambda_{|V|}$ of (7) and (8):

$$\lambda_2 \ge \frac{\text{cov}(g(X_1),g(X_2))}{\text{var}(g(X_1))} \ge \lambda_{|V|}.$$

The covariance of $g(X_1)$ and $g(X_t)$, $t>2$ can be bounded as follows. Let $\alpha = \max(\lambda_2, -\lambda_{|V|})$, as defined in (5). Then

$$\alpha \ge \frac{\text{cov}(g(X_1),g(X_2))}{\text{var}(g(X_1))}.$$

As λ_2^t and $\lambda_{|V|}^t$ are eigenvalues of P^t and $\text{var}(g(X_1))>0$, we have that

$$\alpha^t \text{ var}(g(X_1)) \ge \text{cov}(g(X_1),g(X_t)). \quad (10)$$

Substituting (10) into (9) yields the following upper bound

$$\begin{aligned}\text{var}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right) &\le \text{var}(g(X_1))\left(\frac{1}{n} + \frac{2}{n}\sum_{t=2}^{n}\frac{n-t}{n}\alpha^t\right) \\ &\le \text{var}(g(X_1))\left(\frac{1}{n} + \frac{2\alpha}{n(1-\alpha)}\right) = \\ &= \text{var}(g(X_1))\frac{1+\alpha}{n(1-\alpha)} \\ &\le \frac{\text{var}(g(X_1))}{n(1-\alpha)},\end{aligned}$$

as $0\le\alpha<1$,

$$\sum_{t=2}^{n}\frac{n-t}{n}\alpha^t = \frac{\alpha}{1-\alpha} - \frac{2\alpha^2-\alpha^3-\alpha^{n+1}}{n(1-\alpha)^2},$$

and $2\alpha^2-\alpha^3-\alpha^{n+1}\ge 0$.

□

The upper bound in Theorem III.1 is tight as $\alpha = 0$ yields $\text{cov}(g(X_1),g(X_t)) = 0$, $t>1$, and thus $\text{var}(\hat{F}(X_1,\ldots,X_n)) = \text{var}(\hat{F}(Y_1,\ldots,Y_n))$, which is what we expect if the RW mixes instantly. Hence, for small enough α the MSE of a RW can be approximated by the MSE of RE.

In what follows we compare the standard RW with the Metropolis-Hastings Random Walk (MHRW) that performs uniform sampling of nodes.

A. Degree distribution: RW v.s. RV sampling

The NMSE is the Normalized Mean Square Error of an estimate $\hat{\theta}_d$ of the fraction of nodes with degree d, θ_d, is defined as

$$\text{NMSE}(d) = \frac{\sqrt{MSE(\hat{\theta}_d)}}{\theta_d}.$$

Let $\bar{d}$ denote the average degree. The NMSE of estimating θ_d using n RE is

$$\text{NMSE}_{\text{re}}(d) = \sqrt{(\bar{d}/(d\theta_d)-1)/n}, \quad d>0. \quad (11)$$

Similarly, the NMSE(d) using RV sampling is

$$\text{NMSE}_{\text{rv}}(d) = \sqrt{(1/\theta_d - 1)/n}. \quad (12)$$

Applying Theorem III.1 to (11) yields

$$\text{NMSE}_{\text{rw}}(d) \le \sqrt{\frac{(\bar{d}/(d\theta_d)-1)}{n(1-\alpha)}}, \quad d>0. \quad (13)$$

From equations (13) and (12) we see that a fast mixing RW more accurately estimates degrees larger than the average

($d > \bar{d}$) while RV sampling more accurately estimates degrees smaller than the average ($d < \bar{d}$). The above analysis explains the empirical observation in [12] that the estimates of the degree distribution tail obtained by a RW are more accurate than the equivalent estimates obtained with RV.

IV. MHRWu v.s. RWs

The standard RW described in Section III is the most common type of RW found in the literature [10]. However, other types of random walks exist, refer to [15, Chapter 7] for example. One example is the Metropolis-Hastings RW that can be used to sample vertices uniformly at random (MHRWu). Stutzbach et al. [17] presents an application of the MHRWu to the study of peer-to-peer networks. MHRWu is found to be less accurate than a RW in estimating some graph characteristics such as the degree distribution [5].

A MHRWu is an accept-reject sampling process that samples vertices uniformly. In this section we explore a parallel between MHRWu and a RE resampling algorithm (presented in Section **??**). The MHRWu works as follows, starting at vertex v we:

- select a neighbor u of v uniformly at random;
- the next sampled vertex (step) is u with probability $\min(d_v/d_u, 1)$, otherwise v is the next (step) sampled vertex.

A. MHRWu

In what follows we present MHRWu as a RW process with "self loops". Let π denote the steady state probability of MHRWu. We know that $\pi = (1/|V|, \ldots, 1/|V|)$. Let p_{uv} be the probability that MHRWu goes from u to v, $u, v \in V$, $u \neq v$. MHRWu is time reversible, thus $\pi_v p_{vu} = \pi_u p_{uv}$. Let $\mathbf{A} = [A_{ij}], i = 1, \ldots, |V|$, be the adjacency matrix of G, $A_{ij} = 1$ iff $(i, j) \in E$, otherwise $A_{ij} = 0$. Note that for $i \neq j$, $i, j \in V$, the probability the a MHRWu walker at vertex i goes to vertex j in one step is

$$1/d_i \times (\mathbf{1}\{d_i \geq d_j\} + d_i/d_j \mathbf{1}\{d_i < d_j\}) = 1/\max(d_i, d_j).$$

Thus, the transition probability matrix of the MHRWu Markov chain is a $|V| \times |V|$ transition probability matrix $\mathbf{P} = [p_{ij}]$, where

$$p_{ij} = \begin{cases} 1/\max(d_i, d_j) & \text{if } i \neq j \\ 1 - \sum_{u \in V} \dfrac{A_{iu}}{\max(d_i, d_u)} & \text{if } i = j\,. \end{cases} \quad (14)$$

Figure 1(a) illustrates some transition probabilities of this Markov chain on a graph with two nodes and one edge. The arrows in Figure 1(a) indicate the walker direction and the probability that the direction is taken. Note that the self-loop p_{uu} adds an average of $1/p_{uu}$ "extra" copies of $u \in V$ for each sample of one of its neighbors. These extra copies allow all vertices to be sampled in average at the same frequency.

An interesting characteristic of MHRWu can be seen in Figure 1(b). The self-loop at u has probability $1 - 1/d_v$, which means that on average $d_v - 1$ extra copies of u are made for each sample of v. This resampling can induce large estimation errors as seen in the next section.

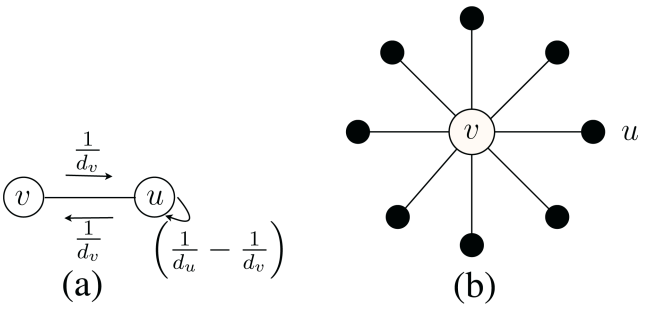

Fig. 1. (a) MHRWu transition probabilities and (b) star graph example.

B. A MHRWu NMSE approximation

In Section III-A we introduced the NMSE of RV and RE, and an upper bound on the RW NMSE. In this section we present an approximation of the MHRWu NMSE. Let $\{H_t\}_1^n$ be a sequence of vertices sampled by MHRWu in steady state. An unbiased estimate of F is

$$\hat{F}^\dagger(H_1, \ldots, H_n) \stackrel{\Delta}{=} \frac{1}{n} \sum_{t=1}^{n} f(H_t). \quad (15)$$

We approximate $\hat{F}^\dagger(H_1, \ldots, H_n)$ as follows. Let be a set of RE sampled vertices and let $K_d(\mathcal{Y})$ be the number of vertices in $\mathcal{Y}$ with degree d. The sampled vertices in process $\{H_t\}_1^n$ can be divided into two categories: (1) vertices that are sampled due to a transition p_{ij}, $i \neq j$; and (2) vertices that were resampled because of the p_{ii}, $i \in V$, probability in (14).

A rough model of $\{H_t\}_1^n$ can be obtained by modeling the sampled nodes of $\{H_t\}_1^n$ as follows: (1) the samples due to a transition p_{ij}, $i \neq j$, are a sequence of RE sampled vertices $\mathcal{Y} = \{Y_t\}_1^n$; and (2) the samples due to p_{ii} are modeled by making some random copies of the vertices in $\mathcal{Y}$. For each RE sampled vertex Y_t, $t \in \{1, \ldots, n\}$, the number of copies is $Z_t - 1$, where Z_t is a Geometric random variable with parameter $d(Y_t)/\text{vol}(V)$, where $d(Y_t)$ is the degree of Y_t. In what follows we show that, just like MHRWu, the above rough model gives an unbiased estimate of F

$$F' = \frac{1}{n} \sum_{t=1}^{n} f(Y_t) Z_t,$$

as shown in the following lemma.

Lemma IV.1. *Let $\mathcal{Y} = \{Y_t\}_1^n$ be a sequence of RE sampled vertices. Let $\{Z_t\}_1^n$ be a sequence of iid Geometric random variables with parameter $d(Y_t)/\text{vol}(V)$, where $d(Y_t)$ is the degree of Y_t. Then*

$$F' = \frac{1}{n} \sum_{t=1}^{n} f(Y_t) Z_t,$$

is an unbiased estimate of F.

Proof. For any $t \geq 1$:

$$\begin{aligned} &E\left[f(Y_t)Z_t\right] = E\left[E[f(Y_t)Z_t|Y_t]\right] \\ &= E\left[f(Y_t)E[Z_t|Y_t]\right] = E\left[f(Y_t)1/\pi_{Y_t}\right] \\ &= \sum_{\forall v \in V} f(v)\pi_v/\pi_v = \sum_{\forall v \in V} f(v). \end{aligned}$$

Thus,

$$E[F'] = \frac{1}{n}\sum_{t=1}^{n} E[f(Y_t)Z_t], = \sum_{\forall v \in V} f(v),$$

which concludes our proof. □

Note that in the above model we have $\sum_{t=1}^{n} Z_t$ samples, where $Z_t \geq 1$, $t = 1, \ldots, n$, while in the original problem there are only n samples. To correct for that we refine the above model.

Let D_d denote the set of vertices with degree d in $\{H_i\}_{i=1}^{n}$. Let $K_d = |D_d|$ be the size of D_d. K_d is a Binomial random variable with parameters n and θ_d; $P[K_d = k] = \binom{n}{k}\theta_d^k(1-\theta_d)^{n-k}$. One difficulty in obtaining closed form expressions using D_d is that vertices in D_d are not independent, as D_d contains the vertices and their copies. To simplify our model we assume independence, i.e., that the number of its copies in $\{H_i\}_{i=1}^{n}$ of vertices with degree d is a sequence $\{Z_t^{(d)}\}_1^{K_d}$ of iid Geometric random variables with parameter $1/d$. Let $\text{NMSE}_{mh}(d)$ denote the NMSE for degree d of this improved model of $\{H_t\}_1^n$.

Theorem IV.2. *Let* $\text{NMSE}_{mh}(d)$ *denote the NMSE for degree* d *of the above model of* $\{H_t\}_1^n$. *Then*

$$\text{NMSE}_{mh}(d) > \sqrt{\frac{2(d-1)^2/\theta_d - 1}{n}}.$$

Proof. Let K_d denote the number of vertices with degree d in $\{H_i\}_{i=1}^{n}$. K_d is a Binomial random variable with parameters n and θ_d; $P[K_d = k] = \binom{n}{k}\theta_d^k(1-\theta_d)^{n-k}$. The total number of replications of vertices with degree d is a sequence $\{Z_t^{(d)}\}_1^{K_d}$ of iid Geometric random variables with parameter $1/d$. We have

$$\text{NMSE}_{mh}(d) = \frac{1}{\theta_d}\sqrt{\frac{1}{n^2}\text{var}\left(\sum_{t=1}^{K_d} Z_t^{(d)}\right)}$$

As $\{Z_t^{(d)}\}_1^{K_d}$ is an iid sequence of random variables that are independent of K_d, the variance of the sum $\sum_{t=1}^{K_d} Z_t^{(d)}$ can be decomposed as [16, pp. 349, Example 4n]

$$\begin{aligned} &\frac{1}{\theta_d}\sqrt{\frac{1}{n^2}\text{var}\left(\sum_{j=1}^{K_d} Z_j'(d)\right)} \\ &= \frac{1}{\theta_d}\sqrt{(1/n^2)\left(E[K_d]\text{var}(Z_1^{(d)}) + E[Z_1^{(d)}]^2\text{var}(K_d)\right)}. \end{aligned}$$

Sampling Method	NMSE error
Node Sampling	$\sqrt{\frac{1/\theta_d - 1}{n}}$
Edge Sampling	$\sqrt{\frac{\bar{d}/(d\theta_d) - 1}{n}}, \quad d > 0$
Random Walk	$\leq \sqrt{\frac{(\hat{d}/(d\theta_d) - 1)}{n(1-\alpha)}}, \quad d > 0$
Metropolis-Hastings RW	$\approx \sqrt{\frac{(d-1)^2/\theta_d - 1}{n}}, \quad d > 0$

TABLE I

SUMMARY OF RESULTS: DEGREE DISTRIBUTION ESTIMATION ERRORS OF VARIOUS SAMPLING METHODS. θ_d IS THE FRACTION OF NODES WITH DEGREE d (QUANTITY THAT IS ESTIMATED), n IS THE NUMBER OF SAMPLED NODES, AND α THE ABSOLUTE VALUE OF THE SECOND MOST DOMINANT EIGENVALUE OF THE RW TRANSITION PROBABILITY MATRIX

The expectation and variance of the above random variables are known thus

$$\begin{aligned} &\frac{1}{\theta_d}\sqrt{(1/n^2)\left(E[K_d]\text{var}(Z_1^{(d)}) + E[Z_1^{(d)}]^2\text{var}(K_d)\right)} \\ &= \sqrt{\frac{n\theta_d(d^2-d) + d^2 n\theta_d(1-\theta_d)}{n^2\theta_d^2}} \\ &= \sqrt{\frac{d^2(2-\theta_d) - d}{n\theta_d}} \\ &> \sqrt{\frac{2(d-1)^2/\theta_d - 1}{n}} \end{aligned}$$

finishing our proof. □

Although there are no guarantees that $\text{NMSE}_{\text{mh}}(d)$ is a good approximation to the true NMSE of MHRWu, our simulations (representative results of our simulations are presented in Section VI) indicate that $\text{NMSE}_{\text{mh}}(d)$ is indeed close to the empirical value of NMSE for large values of d.

It is interesting to take a closer look at the equation

$$\text{NMSE}'_{\text{mh}}(d) > \sqrt{\frac{(d-1)^2/\theta_d - 1}{n}}$$

to note that the MHRWu NMSE grows linearly with d. Contrast the linear growth of $\text{NMSE}'_{\text{mh}}(d)$ as a function of d with the NMSE of a RW:

$$\text{NMSE}_{\text{rw}}(d) \leq \sqrt{(\hat{d}/(d\theta_d) - 1)/(n(1-\alpha))}.$$

which is inversely proportional to d. Also note that

$$\text{NMSE}'_{\text{mh}}(d) > \text{NMSE}_{\text{rv}}(d)\ ,\ \forall d,$$

in fact, the error of MHRWu is almost d times larger than the error of RV sampling. Table I summarizes our analytic results.

V. Current Efforts to Improve RW Accuracy

Sampling a graph using a RW is not without drawbacks. A random walker can get (temporarily) "trapped" inside a subgraph whose characteristics differ from those of the whole graph. Even if the random walker starts in steady state (i.e., is stationary), a "trap" may increase the mean squared error of the estimates. Ideally, the random walker needs to mitigate the effect of these traps on the estimates. Note that in a graph with such "traps", $\alpha \approx 1$ and, as seen in Section III, the RW MSE is upper bounded by $1/(1-\alpha)$. A simple naive solution to the RW "trapping" problem is to sample the graph using multiple independent random walkers [4]. This naive solution, however, can have the opposite effect and exacerbate the problem [14]. The literature, however, provides some promising approaches to cope with this problem if the graph admits a limited (small) amount of RV sampling.

RV sampling has been used to significantly reduce α [1] by allowing the random walks to "jump" to an RV sampled node. The algorithm in [1] differs from the PageRank [3] RW + RV "jumps" in that it obtains unbiased estimates of eq.(1). In [1] it is also shown that, unless the underlying graph is known, the PageRank algorithm must necessarily obtain biased estimates of eq.(1). Another promising approach to improving the RW accuracy is starting m dependent walkers at m RV sampled nodes. This approach, called Frontier Sampling (FS) [14] given in Algorithm 1, introduces a simple dependence among all m walkers in a way that starting the m walkers at m RV sampled nodes is arbitrarily close to starting FS in steady state, provided m is large enough. In our simulations we observe that the FS NMSE for $m = 1000$ is close to the NMSE of a RW with negligible mixing time.

Algorithm 1: Frontier Sampling (FS).

1: $n \leftarrow 0$ {n is the number of steps}
2: Initialize $L = (v_1, \ldots, v_m)$ with m randomly chosen vertices (uniformly)
3: **repeat**
4: Select $u \in L$ with probability $d_u / \sum_{\forall v \in L} d_v$
5: Select an outgoing edge of u, (u, v), uniformly at random
6: Replace u by v in L and add (u, v) to sequence of sampled edges
7: $n \leftarrow n + 1$
8: **until** $n \geq n - mc$

VI. Simulation Results

In what follows we present the results of our simulations of the sampling methods discussed in this paper. The graphs used in our experiments are real-world graphs detailed in Table II. But due to space constraints we restrict our results to the two largest graphs in our datasets: LiveJournal and Flickr. Note that our simulations are performed on disconnected graphs, which can increase the MSE of methods such as RW and MHRWu (FS is designed to mitigate the large MSEs caused by disconnected graphs). Results using the other datasets are similar to LiveJournal and Flickr results. All sampling methods have a budget of n vertices to sample. Each newly sampled vertex deducts one from the budget while resampling a vertex does not count against the budget (i.e., has cost zero). The empirical MSE of our simulations is obtained over $10,000$ runs.

In some of our simulations we use a slightly different MSE metric than the NMSE: the normalized root mean square error of the Complementary Cumulative Distribution Function (CCDF) $\gamma = \{\gamma_d\}_{d \geq 1}$, where $\gamma_d = \sum_{k=d+1}^{\infty} \theta_k$,

$$\text{CNMSE}(d) = \frac{\sqrt{E[(\hat{\gamma}_d - \gamma_d)^2]}}{\gamma_d}, \qquad (16)$$

where $\hat{\gamma}_d$ is the estimate of γ_d. The CNMSE is just the NMSE of γ_d and thus $\text{CNMSE}_{\text{mh}}(d)$, $\text{CNMSE}_{\text{rv}}(d)$, $\text{CNMSE}_{\text{rw}}(d)$, and $\text{CNMSE}_{\text{re}}(d)$ have the same expression as their respective NMSE formulation with θ_d replaced by γ_d (or $\Pi_d = d\theta_d/\bar{d}$ replaced by $d\gamma_d/\bar{d}$).

Note that the graphs in Table II are directed and thus we estimate in-degree and out-degree distributions. These directed graphs, unlike other directed graphs such as the Web graph, are amenable to sampling with RWs as both incoming and outgoing edges are visible to the crawler (random walker). Thus, any degree distribution estimator can be used to also estimate in-degree and out-degree distributions. There are a number of directed graphs that can be crawled as if they were undirected, Twitter and Livejournal are two good examples. For more details on sampling directed graphs with random walks please refer to [14].

Goodness of theoretical approximations

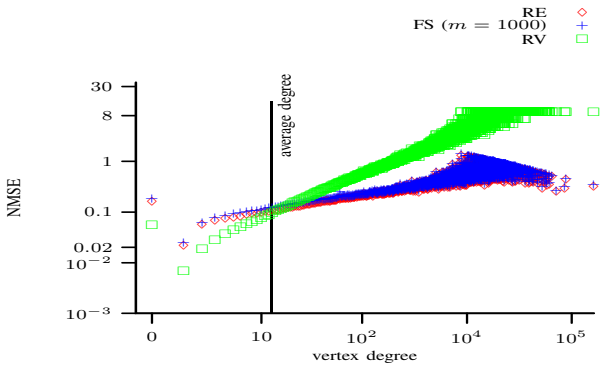

Fig. 2. (Flickr) The log-log plot shows the NMSE of the degree distribution estimation with budget $n = |V|/100 = 18612$ (NMSE over $10,000$ runs).

1) FS v.s. RE: This first set of simulations differ from the remaining simulations in this paper in that we make all directed graph undirected and resampling a vertex reduces the sampling budget by one. In our results we compare the NMSE of FS and RE. Figure 2 shows the log-log plot of the degree NMSE of FS against RE sampling. We observe that the FS NMSE is close to the RE NMSE for all degrees $d > 0$ (note that from Theorem III.1 the RE NMSE is equivalent

Graph	Flickr	LiveJournal	YouTube	Internet RLT
Description	Social Net.	Social Net.	Social Net.	Internet tracert.
Type of graph	Directed	Directed	Directed	Directed
# of Vertices	1,715,255	5,204,176	1,138,499	192,244
Size of LCC	1,624,992	5,189,809	1,134,890	190,914
# of Edges	22,613,981	77,402,652	9,890,764	609,066
Average Degree	12.2	14.6	8.7	3.2
$w_{\max}$	2232	1029	3305	335
% of Original Graph	26.9%	95.4%	NA	NA

TABLE II

SUMMARY OF THE GRAPH DATASETS USED IN OUR SIMULATIONS. "SIZE OF LCC" REFERS TO THE SIZE OF THE LARGEST CONNECTED COMPONENT AND w_{MAX} IS THE VALUE OF THE LARGEST VERTEX DEGREE DIVIDED BY THE AVERAGE DEGREE.

to the MSE of a RW with negligible mixing time ($\alpha \ll 1$)). The same is true in all tested datasets.

2) RE v.s. RV sampling: Figure 2 shows the log-log plot of the in-degree NMSE of RE and RV sampling. As theoretically predicted by our analysis in Section III-A, the NMSE of RV is smaller than the NMSE of RE when d is smaller than the average degree; conversely, the NMSE of RV is larger than the NMSE of RE when d is larger than the average degree as also predicted by our analysis.

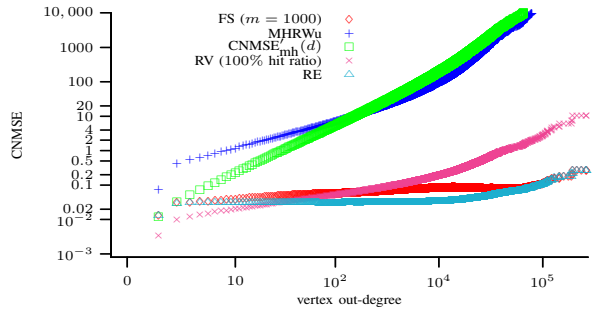

Fig. 3. (Flickr) The log-log plot of the CNMSE of the out-degree distribution estimates with budget $n = |V|/100$. RV with hit ratio 100%.

3) FS v.s. MHRWu : In this simulation on Flickr we compare FS and MHRWu. We simulate FS and MHRWu on the Flickr graph with $n = |V|/100$ samples. Figure 3 plots the the out-degree CNMSE of FS and MHRWu. The estimates of MHRWu are clearly much less accurate than the estimates of FS (for large degrees the MHRWu error is more than 1,000 times that of FS). RW also clearly outperforms MHRWu but less so than FS. Also note that CNMSE$_{\text{re}}(d)$ approximates well the CNMSE of FS, i.e., FS behaves like a RW with spectral gap ≈ 1.

4) MHRWu v.s. NMSE$_{mh}(d)$*:* Figure 3 also assesses the goodness of the theoretical approximation of NMSE$'_{\text{mh}}(d)$ derived in Section IV-B. Note that the approximation of CNMSE$'_{\text{mh}}(d)$ is accurate when $d \geq 10^2$. Note that the CNMSE of MHRWu for large degrees is larger than the scale of the plot, hitting values as large as $10,000$. Clearly MHRWu is unable to accurately sample the tail of the degree distribution.

5) MHRWu v.s. RV sampling: In Figure 3 we also observe that the CNMSE of MHRWu is much larger than the CNMSE of RV and therefore, as expected, RV is preferable to MHRWu for all degrees.

6) Remaining results: Figure 4 plots the in-degree CNMSE of RW, FS, MHRWu, and RV (with 10% hit ratio) on the LiveJournal graph for budget of $n = |V|/1000$. These results confirms the previous findings on Flickr. "RV with (with 10% hit ratio)" represents random vertex sampling when only 1 in 10 queries are valid, i.e., in average only $n/10$ samples are used in the estimator. RV is still the most accurate sampling method for degree $d = 1$ despite the hefty penalty imposed of discarding 9 in 10 samples. For degrees $d > 1$ FS is the most accurate method. FS is slightly more accurate than RW for degrees between 10 and 5×10^3. MHRWu is again the least accurate method (where CNMSE$(d) > 1000$ when $d > 2 \times 10^4$). Figure 5 also shows the error of estimating the in-degree distribution of Flickr. These results are similar to the out-degree distribution results reported in Figure 3.

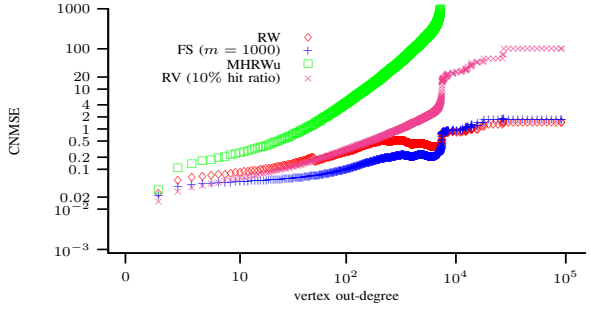

Fig. 4. (LiveJournal) The log-log plot of the CNMSE of the in-degree distribution estimates with budget $n = |V|/1000$.

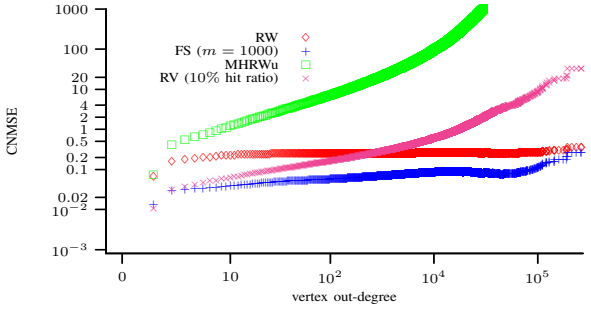

Fig. 5. (Flickr) The log-log plot of the CNMSE of the in-degree distribution estimates with budget $n = |V|/100$.

VII. Related work

A number of works use and contrast standard RW, MHRWu, and RV sampling[2], [5], [6], [11], [17], [12], [18], though their strength is in their empirical findings. In this work we present a principled theoretic comparison of RV, standard RW, RE, and MHRWu sampling methods. We present the first analysis that show that RW sampling MSE is inversely proportional to the spectral gap of the RW transition probability matrix. The first rough approximation to MHRWu sampling using RE sampling is shown in the appendix of Gjoka et al. [5]. The work of Gjoka et al. [5] inspired the refined approximation presented in Section IV-B, where we compute the exact probability of a self-loop, obtaining a better estimate of the average number of resamples of the same node. We use this improved approximation to provide an expression for the MHRWu MSE.

VIII. Conclusions

This paper provides an upper bound for the MSE of a stationary RW as a function of the MSE of RE and the absolute value of the second most dominant eigenvalue of the RW transition probability matrix. We observed that RW and RV sampling are optimal in respect to different weighted MSE optimizations and analyzed when RW is preferable to RV sampling. We also presented an approximation to the MHRWu MSE. Finally, we introduce a novel RW sampling algorithm, Frontier Sampling (FS). Our simulation experiments on large real world graphs showed that FS achieves the MSE of a RW with negligible mixing time.

IX. Acknowledgments

We would like to thank Maciej Kurant, Ananthram Swami for many helpful discussions and Alan Mislove for kindly making available some of the data sets used in this paper. This research was sponsored by the NSF under CNS-1065133, ARO under MURI W911NF-08-1-0233, and the U.S. Army Research Laboratory under Cooperative Agreement W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied of the NSF, ARO, ARL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

[1] Konstantin Avrachenkov, Bruno Ribeiro, and Don Towsley. Improving random walk estimation accuracy with uniform restarts. In *Proc. of the 7th Workshop on Algorithms and Models for the Web Graph*, 2010.

[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proc. of the WWW*, 1998.

[4] Charles J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992.

[5] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Practical recommendations on sampling osn users by crawling the social graph. *JSAC special issue on Measurement of Internet Topologies*, 2011.

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

[7] Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. On near-uniform url sampling. In *Proceedings of the WWW*, pages 295–308, 2000.

[8] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[9] Marlom A. Konrath, Marinho P. Barcellos, and Rodrigo B. Mansilha. Attacking a swarm with a band of liars: evaluating the impact of attacks on bittorrent. In *P2P '07: Proceedings of the Seventh IEEE International Conference on Peer-to-Peer Computing*, pages 37–44, Washington, DC, USA, 2007. IEEE Computer Society.

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

[11] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proc. of the IMC*, October 2007.

[12] Amir H. Rasti, Mojtaba Torkjazi, Reza Rejaie, Nick Duffield, Walter Willinger, and Daniel Stutzbach. Respondent-driven sampling for characterizing unstructured overlays. In *Proc. of the IEEE Infocom*, pages 2701–2705, April 2009.

[13] Bruno Ribeiro, William Gauvin, Benyuan Liu, and Don Towsley. On MySpace account spans and double Pareto-like distribution of friends. In *IEEE Infocom 2010 Network Science Workshop*, Mar 2010.

[14] Bruno Ribeiro and Don Towsley. Estimating and sampling graphs with multidimensional random walks. In *Proc. of the ACM SIGCOMM IMC*, Oct. 2010.

[15] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, 2nd edition, 2005.

[16] Keith Ross. *A First Course in Probability*. Prentice Hall, 5 edition, 1997.

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

[18] Erik Volz and Douglas D. Heckathorn. Probability based estimation theory for Respondent-Driven Sampling. *Journal of Official Statistics*, 2008.