On the Estimation Accuracy of Graph Sampling

Bruno Ribeiro, Don Towsley, *Fellow, IEEE*

***Abstract*—Estimating characteristics of large graphs via sampling is a vital part of the study of complex networks. In this work we present an in-depth study of the Mean Squared Error (MSE) of sampling methods such as independent random vertex (RV) and random edge (RE) sampling and crawling methods such as random walks (RWs), a.k.a. RDS, and the a Metropolis-Hastings algorithm whose target distribution is to uniformly sample vertices (MHRWu). This paper provides an upper bound for the MSE of a stationary RW as a function of the MSE of RE and the module of the second most dominant eigenvalue of the RW transition probability matrix. We see that RW and RV sampling are optimal in respect to different weighted MSE optimizations and show when RW is preferable to RV sampling. We also present an approximation to the MHRWu MSE. Finally, we introduce a novel RW sampling algorithm, Frontier Sampling (FS). Our simulations on large real world graphs show that FS achieves the MSE of a RW with negligible mixing time.**

I. Introduction

A number of recent studies [1], [2], [4], [5], [9], [10], [13], [**?**], [19], [14], [15], [20] (to cite a few) are dedicated to the characterization of network graphs. This paper represents a network as 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 [5], [19], the assortativity coefficient [16], or the global clustering coefficient [16].

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% [15]). Moreover, queries are often subject to resource constraints (e.g., queries are rate-limited in Flickr, Livejournal [13], and Bittorrent [8]). 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, randomly sampling edges can be harder than randomly sampling 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 would allow us to randomly sample edges.

RW sampling: An alternative, and often 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 [13], as one can obtain asymptotically unbiased estimates of a number of graph characteristics such as fraction of vertices with a given label [20], the degree distribution [20], and, more recently discovered, assortativity and global clustering coefficients [16]. 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 [4], [14], [20] in oder to estimate their characteristics. One of the reasons behind the popularity of RWs are that they do not have the drawbacks of RV and RE methods, i.e., in a RW, different than in RV sampling, all queried users are valid.

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 have been used to uniformly sample peers in peer-to-peer networks [19] and Web pages [6]. Unfortunately, MHRWu has large estimation errors compared to RW estimates [4].

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 module of the second most dominant eigenvalue of the RW transition probability matrix.
2) We present the graph sampling problem as the minimization of a weighted MSE sum. We illustrate our approach

The authors are with the Computer Science Department, University of Massachusetts Amherst, Amherst, MA, 01003 USA e-mail: {ribeiro,towsley}@cs.umass.edu

using the degree distribution as an example. For the degree distribution we see that RW sampling minimizes the MSE whose weights are the vertex degree squared (i.e., the weights give more importance to large degree vertices); while RV sampling minimizes the MSE with equal weights.

3) We present a method called Frontier Sampling (FS) that, unlike a standard RW, can start close to steady state even in loosely connected graphs, if a small amount of RV sampling is allowed (in our simulations we use 1000 RV sampled vertices). FS mitigates well known (see [11]) estimation errors and biases of RWs that arise when they get "temporarily trapped" and spend most of their sampling budget exploring a local neighborhood of the graph. In our simulation we also observe that the MSE of FS is close to that of RE, i.e., the MSE of FS is close to the MSE of a fast mixing RW (α small).
4) RW estimates have been observed to be more accurate than estimates obtained by MHRWu [4], [14]. We study how the Metropolis-Hastings mechanism tends to induce larger estimation errors than RW or even 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 module of the second most dominant eigenvalue of the RW transition probability matrix. In Section IV we present the graph sampling problem as minimizing the weighted MSE sum. In Section V we study how the Metropolis-Hastings mechanism tends to induce larger estimation errors than RW or even RV sampling. Section VII 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_a, $a \in V$, be the degree of vertex a. 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 $(Z_1, \dots, Z_n)$ be a stationary sequence of n sampled vertices, where $P[Z_t = v] = \beta_v > 0$, $\forall v \in V$, $t = 1, \dots, n$. Then

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

is an unbiased estimate of F (eq.(1). The estimator $\hat{F}$ in eq.(2) is widely used to estimate F, see [16], [20] and the references therein.

The Mean Squared Error (MSE) of $\hat{F}(Z_1, \dots, Z_n)$ is

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

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

III. A tight upper bound of the RW estimation error

Random walk (RW) sampling and random edge (RE) sampling are closely related. Let $\pi = (\pi_a : a \in V)$, denote the steady state probability distribution of the RW. Because the RW is time reversible, $\pi_v/d_v = \pi_u/d_u$ [12]. A consequence of this is that edge (u, v) is sampled with probability $1/|E|$ (uniformly at random), which is true for all edges in the graph. Thus, RW and RE differ only that edges sampled by a RW samples are correlated.

In what follows we present a tight upper bound of the MSE of a stationary RW. More precisely, let $(X_1, \dots, X_n)$ be a sequence of n vertices sampled by a stationary RW. A RW is stationary iff $X_1 \sim \pi$. Let $(Y_1, \dots, Y_n)$ be a sequence of RE sampled vertices. We show that the MSE, eq.(3), of $(X_1, \dots, X_n)$ is upper bounded by a function of the MSE of $(Y_1, \dots, Y_n)$ and α, where $0 \le \alpha < 1$ is the module of the second most dominant eigenvalue of the RW transition probability matrix.

In what follows we define the RW the module of the second most dominant eigenvalue. Let $\mathbf{A} = [a_{ij}], i = 1, \dots, |V|$, be the adjacency matrix of G, $a_{ij} = 1$ iff $(v_i, v_j) \in E$, otherwise $a_{ij} = 0$. Let

$$\mathbf{D} = \begin{bmatrix} d_{v_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_{v_{|V|}} \end{bmatrix}$$

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 v from u in t steps is

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

The stationary distribution of the RW is $\pi = \mathbf{P}\pi$. Let $\lambda_1 \ge \lambda_2 \ge \dots \ge \lambda_{|V|}$ 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 $1 = \lambda_1 > \lambda_2 \ge \dots \ge \lambda_{|V|} > -1$ [12]. The module of the second most dominant eigenvalue is defined as

$$\alpha \triangleq \max(\lambda_2, -\lambda_{|V|}). \tag{4}$$

A RW is fast mixing when α is sufficiently small (we choose to use a vague definition of fast mixing as there are many contradicting 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 $(X_1, \dots, X_n)$ be a sequence of vertices sampled by a stationary RW on G, $n \ge 1$. Let $(Y_1, \dots, Y_n)$ be a sequence of RE sampled vertices. Let*

$$\hat{F}(Z_1, \dots, Z_n) \triangleq \frac{1}{n} \sum_{t=1}^{n} \frac{f(Z_t)}{\pi_{Z_t}} \ , \ Z_i \in V, i = 1, \dots, n$$

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

Then

$$\text{var}(\hat{F}(X_1,\ldots,X_n)) \le \frac{\text{var}(\hat{F}(Y_1,\ldots,Y_n))}{(1-\alpha)}\,. \tag{5}$$

Proof: Let $\mathbf{S} = \pi^{1/2}\mathbf{P}\pi^{1/2}$. It is easy to verify that $\mathbf{S}$ is a $|V| \times |V|$ symmetric matrix whose eigenvalues are also the eigenvalues of $\mathbf{P}$. The eigenvector of $\mathbf{S}$ corresponding to eigenvalue $\lambda_1 = 1$ is $\pi^{1/2}$. Let $g(v) = f(v)/\pi_v - F$, which yields $E[g(X_i)] = 0$, $i = 1,\ldots,n$. Let $\langle g,\pi\rangle \triangleq \sum_{\forall v\in V} g(v)\pi_v$ be the inner product between g and π. The Courant-Fischer theorem [7, Theorem 4.2.11, pp. 179] gives the module of the second most dominant eigenvalue of $\mathbf{S}$

$$\lambda_2 = \max_{w\,:\,\langle w,\pi\rangle=0} \frac{\sum_{\forall v\in V}\sum_{\forall u\in V} w(v)w(u)\pi_v p_{v,u}}{\sum_{\forall u\in V} w(v)^2\pi_v} \tag{6}$$

and the smallest eigenvalue of $\mathbf{S}$

$$\lambda_{|V|} = \min_{w\,:\,\langle w,\pi\rangle=0} \frac{\sum_{\forall v\in V}\sum_{\forall u\in V} w(v)w(u)\pi_v p_{v,u}}{\sum_{\forall u\in V} w(v)^2\pi_v}\,, \tag{7}$$

as $\langle w,\pi\rangle = \langle r,\pi^{1/2}\rangle$, $r = (w(v_i)\sqrt{\pi_{v_i}} : i = 1,\ldots,|V|)$.

We use eqs.(6) and (7) to define an upper and lower bound of the covariance of $g(X_1)$ and $g(X_t)$, $t = 2,\ldots,n$. We use the following definitions of covariance and variance: For $1 < t \le n$

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

and

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

The bounds are found by replacing the above definitions into eqs.(6) and (7)

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

Let $\alpha = \max(\lambda_2, -\lambda_{|V|})$, as defined in eq.(4). 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

$$\alpha^t\,\text{var}(g(X_1)) \ge \text{cov}(g(X_1), g(X_t))\,. \tag{8}$$

A known property of the variance is [17, pp. 265]

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

Applying eq.(8) into eq.(9) yields

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

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 proof is concluded by noting 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 above proof is valid for any value of α. The upper bound in Theorem III.1 is tight as

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

and $\alpha = \lambda_2 = \lambda_{|V|} = 0$ yields $\text{cov}(g(X_1), g(X_t)) = 0$, $t > 1$. Thus, eq.(9) yields $\text{var}(\hat{F}(X_1,\ldots,X_n)) = \text{var}(\hat{F}(Y_1,\ldots,Y_n))$. Hence, the MSE of a fast mixing RW can be approximated by the MSE of RE, but only for small enough values of α. Otherwise, we need the upper bound in Theorem III.1.

IV. MSE MINIMIZATION

In Section III we provided an upper bound for the RW MSE. So far we considered a RW that samples $v \in V$ proportional to d_v, i.e., vertices are sampled from distribution π. In what follows we denote this type of random walk "*standard RW*".

There are many different ways to sample a graph (e.g., RE, RV, standard RW). In this section we are interested in sampling the graph as to minimizes a given weighted sum of the MSE. The motivation behind this section comes from the several types of stationary RWs that sample vertices, $(X_i)_{i=1}^n$, with (an arbitrary) distribution $X_i \sim \gamma$, $i = 1,\ldots,n$ (e.g., Metropolis-Hastings algorithm and Gibbs sampler are two of such RW types [17]). In Theorem III.1 we proved that the MSE of a standard RW is upper bounded by the MSE of RE sampling times a constant that depends on the graph. Unfortunately, Theorem III.1 cannot be easily extended to $\gamma \ne \pi$. Thus, we make the simplifying assumption of independence in $(X_i)_{i=1}^n$. For MHRWu the independence assumption means that vertices are RV sampled. For RW the independence assumption means that vertices are RE sampled. In our simulations in Section VII we see that the independence assumption gives a good approximation for RWs (specially if vertices are sampled by our proposed RW, frontier sampling, described in Section VI) and a bad approximation for MHRWu.

To illustrate the optimization, consider estimating the degree distribution, θ_d, $d = 0, 1, \ldots$,

$$\theta_d \triangleq \frac{1}{n} \sum_{\forall v \in V} f_d(v) \,,$$

where $f_d(v) = \mathbf{1}(d_v = d)/|V|$ and $\mathbf{1}(d_v = d) = 1$ if $d_v = d$ and zero otherwise, from a sequence of i.i.d. sampled vertices, $(Y_1, \ldots, Y_n)$, where $Y_i \sim \gamma$, $i = 1, \ldots, n$. Note that

$$\hat{F}_d(Y_1, \ldots, Y_n) = \frac{1}{n} \sum_{t=1}^{n} \frac{f_d(Y_t)}{\gamma_{Y_t}} \,.$$

is an unbiased estimate of θ_d. To simplify our analysis assume $\gamma_v = \Gamma_{d_v} > 0$, $v \in V$. From the independence of Y_i, $i = 1, \ldots, n$,

$$\begin{aligned} E[(\hat{F}_d(Y_1, \ldots, Y_n) - F_d)^2] =& \frac{1}{n} \left(\sum_{\forall v \in V} \left(\frac{f_d(v)}{\gamma_v} \right)^2 \gamma_v - \theta_d^2 \right) \\ =& \frac{\theta_d}{\Gamma_d} - \theta_d^2. \end{aligned}$$

Let $\gamma^\star$ be the distribution that minimizes the weighted MSE

$$\gamma^\star = \arg_\gamma \min \sum_{\forall d} \left(\frac{\theta_d}{\Gamma_d} - \theta_d^2 \right) w_d \,,$$

$w_v > 0$, $v \in V$.

Lemma IV.1. *The distribution $\Gamma^\star$ that minimizes the weighted MSE*

$$\Gamma^\star = \arg_\Gamma \min \sum_{\forall d} \left(\frac{\theta_d}{\Gamma_d} - \theta_d^2 \right) w_d \,,$$

with weights $\{w_j\}$ satisfies the following relation

$$\frac{w_i}{w_j} = \left(\frac{\Gamma_i}{\Gamma_j} \right)^2$$

Proof: As Γ_d, $d = 1, 2, \ldots$, is a distribution we add the restriction $\sum_{\forall d} \Gamma_d = 1$ as a Lagrange multiplier in the optimization, which results in the set of equations:

$$h(d) = \sum_{\forall d} \left(\frac{\theta_d}{\Gamma_d} - \theta_d^2 \right) w_d - \lambda \left(\sum_{\forall d} \Gamma_d - 1 \right), \forall u \in V$$

Taking the derivative of $h(d)$ with respect to Γ_d and equating to zero yields

$$\frac{\partial h(d)}{\partial \Gamma_d} = -\frac{1}{\Gamma_d^2} - \lambda = 0, \forall d,$$

thus

$$\frac{w_i}{w_j} = \left(\frac{\Gamma_i}{\Gamma_j} \right)^2$$

All is left is to prove that $\frac{w_i}{w_j} = \left(\frac{\Gamma_i}{\Gamma_j} \right)^2$ is not only a saddle point in $h(d)$. This is easy as $\partial^2 h(d)/\partial^2 \Gamma_d = 2/\Gamma_d^3 > 0$. ■

In particular, $\Gamma_d \propto d$ gives

$$w_i = w_j (i/j)^2 \;\;, i, j = 1, 2, \ldots.$$

Note that $w_i > w_j$ when $i > j$ and $w_i < w_j$ when $d_i < d_j$. Thus, a standard RW and RE sampling optimize a weighted MSE that places larger weights at the tail of the degree distribution. Another particular case occurs when $\Gamma_i = \Gamma_j$, $\forall i, j$:

$$w_v = w_u \,,\; \forall u, v \in V \,,$$

i.e., RV sampling optimizes a weighted MSE with equal weights.

Degree distribution: RW v.s. RV sampling

The NMSE is the Normalized Mean Square Error, defined as $\sqrt{MSE}/F$, where F is the true value. Let $\bar{d}$ be the average degree. Let $\Pi_d = d\theta_d/\bar{d}$ be the probability that a RE (or a RW in steady state) samples a vertex with degree d. From the exposition in Section IV it is straightforward to show that the NMSE estimating θ_d using n RE samples is

$$\text{NMSE}_{\text{re}}(d) = \sqrt{(1/\Pi_d - 1)/n}\,, \quad d > 0. \tag{10}$$

Similarly, the NMSE(d) using RV sampling is

$$\text{NMSE}_{\text{rv}}(d) = \sqrt{(1/\theta_d - 1)/n}\,. \tag{11}$$

Applying Theorem III.1 to eq.(10) yields

$$\text{NMSE}_{\text{rw}}(d) \leq \sqrt{\frac{(1/\Pi_d - 1)}{n(1-\alpha)}}\,, \quad d > 0. \tag{12}$$

Note that $\Pi_d > \theta_d$ if $d > \bar{d}$ and $\Pi_d < \theta_d$ if $d < \bar{d}$. From equations (12) and (11) 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 what has been previously observed in [14].

V. MHRWu v.s. RWs

The RW described in Section III is the most common type of RW found in the literature [12] but there are other types of random walks. For more details refer to [17, Chapter 7]. In [19] a Metropolis-Hastings RW that samples vertices uniformly at random is described, which we denote MHRWu in this work. MHRWu is found to be less accurate than a RW in estimating some graph characteristics such as the degree distribution [4].

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 V-A). 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. This is equivalent to say that we add a copy of vertex v to the sample set with probability $\max(0, 1 - d_v/d_u)$, otherwise u is the next (step) sampled vertex.

A. RW alternative unbiased estimator by resampling

In what follows we present another way to obtain an unbiased estimate of F. This estimator will be later used to provide an approximation to MHRWu MSE. As before, $(Y_i)_{i=1}^n$ is a sequence of vertices sampled by RE. Let K_d be the number of vertices with degree d in $(Y_i)_{i=1}^n$. K_d is a Binomial random variable with parameters n and $p_d = d/\text{vol}(V)$ ($P[K_d = k] = \binom{n}{k} p_d^k (1-p_d)^{n-k}$). Let $Z_i^{(d)} \in \{1, 2, \ldots\}$ be a sequence of i.i.d. Geometric random variable with parameter p_d, $i = 1, \ldots, n$.

Lemma V.1.

$$F' = 1/n \sum_{i=1}^{n} f(Y_i) Z_i^{(d_{Y_i})},$$

where d_{Y_i} is the degree of Y_i, is an unbiased estimate of F (eq.(1)).

Proof: First

$$E\left[f(Y_i)Z_i^{(d_{Y_i})}\right] = E\left[E[f(Y_i)Z_i^{(d_{Y_i})}|Y_i]\right] = E\left[E[f(Y_i)|Y_i]E[Z_i^{(d_{Y_i})}|Y_i]\right] = E\left[f(Y_i)E[Z_i^{(d_{Y_i})}|Y_i]\right] = E\left[f(Y_i)1/\pi_{Y_i}\right] = \sum_{\forall v \in V} f(v)\pi_v/\pi_v = \sum_{\forall v \in V} f(v).$$

As $(Y_i)_{i=1}^n$ is an i.i.d. sequence of random variables $E[F'] = (1/n)nE[f(Y_i)Z_i^{(d_{Y_i})}]$, which concludes our proof. ■

The next lemma provides the NMSE of estimating θ_j using Lemma V.1.

Lemma V.2.

$$\text{NMSE}_{F'}(d) = \sqrt{\frac{2(1-p_d)}{np_d\theta_d^2}}.$$

Proof: The total number of replications of vertices with degree d is $Z_d = \sum_{j=1}^{K_d} Z_j^{(d)}$. Note that $F' = Z_d/n$, which yields [18, pp. 349, Example 4n]

$$\text{var}(F') = \frac{1}{n^2}\text{var}\left(\sum_{j=1}^{K_d} Z_j^{(d)}\right) = (1/n^2)\left(E[K_d]\text{var}(Z_j^{(d)}) + E[Z_j^{(d)}]^2\text{var}(K_d)\right) = (1/n^2)\left((np_d)(1-p_d)/p_d^2 + (1/p_d^2)(np_d(1-p_d))\right) = (1/n^2)\left(n(1-p_d)/p_d + n(1-p_d)/p_d\right) = \frac{2(1-p_d)}{np_d}.$$

Lemma V.1 gives $E[F'] = \theta_k$, which yields

$$\text{NMSE}_{F'}(d) = \sqrt{\text{var}(F')}/\theta_d = \sqrt{\frac{2(1-p_d)}{np_d\theta_d^2}}.$$

■

B. Metropolis-Hasting RW: Uniform vertex sampling (MHRWu)

We now turn our attention to MHRWu of [19]. A different way to present the MHRWu algorithm is as an edge sampling process that samples vertices uniformly. The MHRWu is time reversible, which means that the same number of walkers going from v to u must go from u to v. Let p_{ab} be the probability that the walker goes from a to b, $a, b \in V$. Vertices are sampled uniformly and the Markov chain is time reversible, which yields $p_{vu} = p_{uv}$. To simplify our exposition we assume, without loss of generality, that $d_u < d_v$. It is easy to see that $p_{vu} = p_{uv} = 1/d_v$ satisfies the above conditions and can be implemented by selecting neighbors of v and u uniformly at random. However, because the walker chooses v with probability $1/d_u > 1/d_v$ we are required to add a self-loop at u that has probability $1/d_u - 1/d_v$, as illustrated in Figure 1(a). In Figure 1(a) the arrows indicate the walker direction and the probability that the direction is taken. Thus, the probability that an edge $(v, u) \in E$ is sampled is $1/(d_v|V|)$. The self-loop adds an average of $(d_v - d_u)/(d_v(d_u - 1) + d_u)$ "extra" copies of u for each sample of v, due to the existence of the edge (u, v).

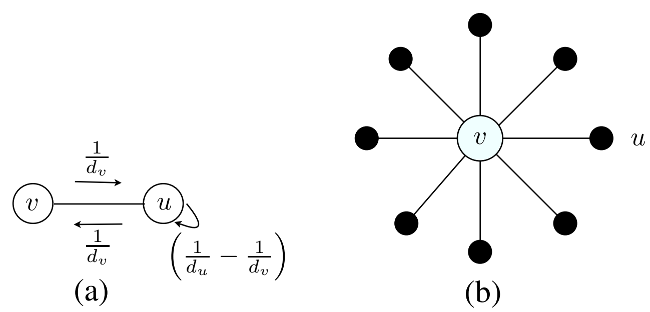

Fig. 1. (a) MHRWu transition probabilities and (b) star graph example.

To illustrate the problem with MHRWu consider the graph 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. Note that the resampling of vertex u can be described by the resampling algorithm in Section V-A. However, as seen next, in general a MHRWu resamples vertices significantly less often than the resampling algorithm described in Section V-A.

C. A MHRWu MSE approximation

Consider an edge-sampling process that samples edge (u, v) with probability $1/(\max(d_v, d_u)|V|)$, $\forall (u, v) \in E$. Edges are sampled independently. The resampling probability is $1/d_v - 1/d_u$ if $d_v < d_u$ and $1/d_u - 1/d_v$ otherwise. While in a MHRWu only the vertex with a self-loop is allowed to *resample* (i.e., make multiple copies of itself), we simplify our analysis by assuming that edge (u, v) incur self-loops on both u and v with probabilities $1/d_u$ and $1/d_v$, respectively. We make the simplifying assumption that vertex v is sampled with probability $1/d_v$ but u is sampled with probability $1/d_u$. Let $(Y_i)_{i=1}^n$ be a sequence of vertices sampled by RE and let

K_d be the number of vertices with degree d in $(Y_i)_{i=1}^n$. Setting $p_d = 1/d$ in Lemma V.2 we get

$$\text{NMSE}'_{\text{mh}}(d) \approx \sqrt{\frac{E[K_d]\text{var}(Y_j^{(d)}) + E[Y_j^{(d)}]^2\text{var}(K_d)}{n^2\theta_d^2}} = \sqrt{\frac{n\theta_d(d^2-d) + d^2 n\theta_d(1-\theta_d)}{n^2\theta_d^2}} = \sqrt{\frac{d^2(2-\theta_d)-d}{n\theta_d}}$$

$$\sqrt{\frac{d^2(2-\theta_d)-d}{n\theta_d}} > \sqrt{\frac{d^2-d}{n\theta_d}}.$$

It is important to emphasize that there are no guarantees that $\text{NMSE}'_{\text{mh}}(d)$ is a good approximation to the true NMSE of MHRWu. The best we can do is to show in Section VII some simulations on real world graphs where $\text{NMSE}'_{\text{mh}}(d)$ is close to the empirical NMSE .

Nevertheless, it is interesting to compare $\text{NMSE}'_{\text{mh}}(d)$ with $\text{NMSE}_{\text{rw}}(d)$ (eq.(12)) and $\text{NMSE}_{\text{rv}}(d)$ (eq.(11)). Clearly $\text{NMSE}'_{\text{mh}}(d) > \text{NMSE}_{\text{rv}}(d)$ as

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

Comparing RW with MHRWu requires some conditions: $\alpha < 1 - 1/d$ and $d > \bar{d} \geq 2$. It is easy to show that $\text{NMSE}_{\text{rw}}(d) < \text{NMSE}'_{\text{mh}}(d)$. Note that

$$\sqrt{\frac{d^2-d}{n\theta_d}} < \sqrt{\frac{d^2}{n\theta_d}}.$$

Replacing $\Pi_d = d\theta_d/\bar{d}$ and $\alpha < 1 - 1/d$ into eq.(12) yields

$$\sqrt{\frac{(\bar{d}/(\theta_d)-1)}{n}} < \sqrt{\frac{2d}{n\theta_d}} < \sqrt{\frac{d^2}{n\theta_d}}.$$

Note that our $\text{NMSE}'_{\text{mh}}(d)$ expression does not depend on α, which shows that our approximation can still be improved. As part of future work we intend to derive the MSE expression of MHRWu from RW samples not RE. Another interesting fact is that $\text{NMSE}'_{\text{mh}}(d)$ is approximately d times larger than $\text{NMSE}_{\text{rv}}(d)$.

VI. Frontier Sampling

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. Note that in such graphs $\alpha \approx 1$. Even if the random walker starts in steady state (i.e., is stationary), this scenario may increase the mean squared error of the estimates. If the random walker does not start in steady state, this scenario may cause an increase in the estimation bias as well as the mean squared error. Ideally, the random walker needs to mitigate the effect of these traps on the estimates. A simple naive solution to the RW "trapping" problem (adopted in [4] to sample Facebook), is to sample the graph using multiple independent random walkers [3]. This naive solution, however, can have the opposite effect and exacerbate the problem [16]. In what follows we propose a method to mitigate the random walk "trapping" problem using m *dependent* random walkers.

In this section we present a new and promising approach to an m-dimensional random walk that benefits from starting its walkers at uniformly sampled vertices. *Frontier Sampling* (FS) performs m *dependent* random walks in the graph. We refer to m as the dimension of the FS random walk. Let c be the cost of randomly sampling a vertex. The FS algorithm, given in Algorithm 1 is a centrally coordinated sampling algorithm that maintains a list of m vertices representing m random walkers. This way FS is less likely to get stuck in loosely connected components than a single random walker.

Algorithm 1: Frontier Sampling (FS).

```
1: n ← 0 {n is the number of steps}
2: Initialize L = (v_1, ..., v_m) with m randomly chosen
   vertices (uniformly)
3: repeat
4:    Select u ∈ L with probability d_u / Σ_{∀v∈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 ← n + 1
8: until n ≥ n − mc
```

FS shares many of the same statistical properties of a single random walker on G.

Theorem VI.1. *In steady state FS has the following properties:*

(I) *edges are sampled uniformly at random and form a stationary sequence and*
(II) *the sequence of sampled edges satisfies the Strong Law of Large Numbers [16, Theorem 4.1].*

Proof: The proof is found in [16]. ■

However, the most striking property of FS is its steady state distribution when $m \gg 1$. When $m \gg 1$, the number of FS walkers in any vertex $v \in V$ is distributed as if the m walkers where distributed uniformly from V [16, Theorem 5.4]. Because of this property and because in Algorithm 1 we initialize FS with vertices sampled uniformly from V, FS starts close to its steady state. The FS algorithm can also be made fully distributed [16]. In Section VII we see that the FS MSE is equivalent to the MSE of a RW with negligible mixing time.

VII. Simulation Results

In what follows we present our the results of our sampling simulations on real world datasets. The graphs used in our experiments are detailed in Table I. 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). The results using the other datasets are similar to LiveJournal and Flickr results, no surprises worth reporting. 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 change 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 $\gamma_d = \sum_{k=d+1}^{\infty} \theta_k$,

$$\text{CNMSE}(d) = \frac{\sqrt{E[(\hat{\gamma}_d - \gamma_d)^2]}}{\gamma_d}, \tag{13}$$

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 $\text{CNMSE}_{\text{re}}(d)$ are have the same equation 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 I are directed. Obtaining directed graph characteristics such as the in-degree distribution from graphs that can be crawled like undirected graphs (e.g., Twitter and Livejournal) is a trivial task, for more details refer to [16].

Goodness of theoretical approximations

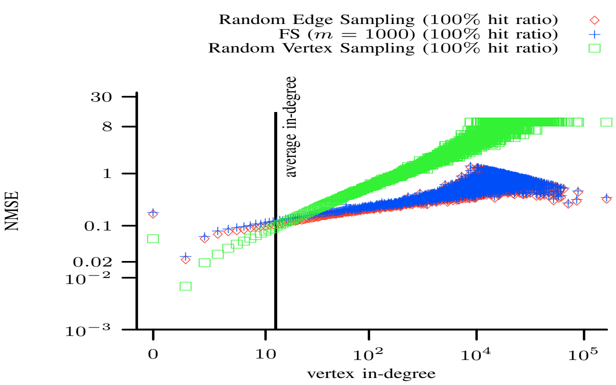

Fig. 2. (Flickr) The log-log plot shows the NMSE of the in-degree distribution estimation with budget $n = |V|/100 = 18612$ (NMSE over $10,000$ runs).

FS v.s. RW and RV sampling: This first set of simulations differ from the remaining simulations in this paper in that resampling a vertex reduces the sampling budget by one. In our results we compare the MSE of FS and RE. In our first result, Figure 2 shows the log-log plot of the in-degree NMSE of FS, RW, and RV sampling. In Figure 2 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 to the MSE of a RW with negligible mixing time ($\alpha \ll 1$)). The same is true in all other datasets. Moreover, as theoretically predicted by the analysis performed in Section IV, the NMSE of RV is smaller than the NMSE of RE when d is smaller than the average degree and the NMSE of RV is larger than the NMSE of RE when d is larger than the average degree.

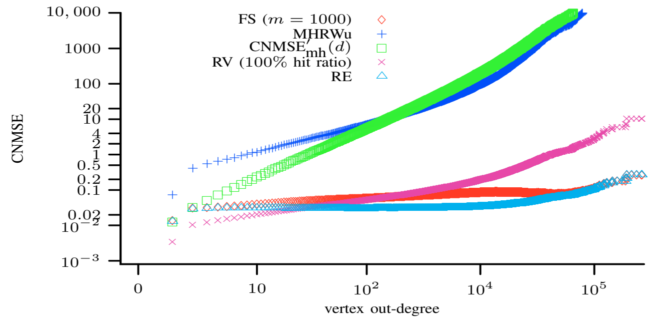

Fig. 3. (Flickr) The log-log plot of the CNMSE of the in-degree distribution estimates with budget $n = |V|/100$. RV with hit ratio 100%.

FS v.s. $\text{CNMSE}_{rw}(d)$ *bound & MHRWu v.s.* $\text{NMSE}'_{mh}(d)$ *and RV sampling:* In this simulation on Flickr we seek to assess the goodness of the theoretical approximation of $\text{NMSE}'_{\text{mh}}(d)$ derived in Section V-C. We also seek to test if assuming that the CNMSE of FS is equivalent to the CNMSE of a RW that mixes fast (i.e., α is small). We simulate FS, MHRWu, and RV on the LiveJournal graph with $n = |V|/100$ samples each. Figure 3 plots the the in-degree CNMSE of FS, MHRWu, RV and also plots $\text{CNMSE}_{\text{re}}(d)$ and $\text{CNMSE}'_{\text{mh}}(d)$. Note that the approximation of $\text{CNMSE}'_{\text{mh}}(d)$ is accurate when $d \geq 10^2$. Remark that the CNMSE of MHRWu is so large that for vertices with degree greater than 8×10^4 it is greater than $10,000$. The estimates of MHRWu are clearly much less accurate than the estimates of FS. As we do not have α, we consider $\alpha = 0$, i.e., $\text{CNMSE}_{\text{rw}}(d) \approx \text{CNMSE}_{\text{re}}(d)$. Note that $\text{CNMSE}_{\text{re}}(d)$ approximates well the CNMSE of FS. From Theorem III.1 we know that this is equivalent to the CNMSE of a RW with α small. Also observe that the CNMSE of MHRWu is much larger than the CNMSE of RV and therefore, as expected, the RV CNMSE is not a good approximation to the MHRWu CNMSE .

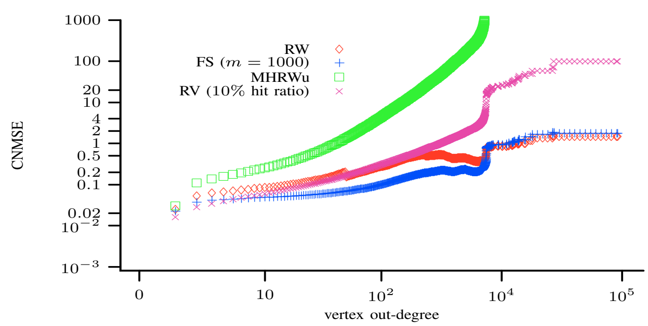

Fig. 4. (LiveJournal) The log-log plot of the CNMSE of the in-degree distribution estimates with budget $n = |V|/1000$.

Accuracy of Graph Sampling Methods

In this simulation we compare RW, FS, MHRWu, and RV (with 10% hit ratio). Figure 4 plots the in-degree CNMSE of RW, FS, MHRWu, and RV (with 10% hit ratio) on the LiveJournal graph for budget of $n = |V|/1000$. "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. We observe that RV (with

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 I
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.

10% hit ratio) is less accurate than RW and FS. FS is slightly more accurate than RW for degrees between 10 and 5×10^3.

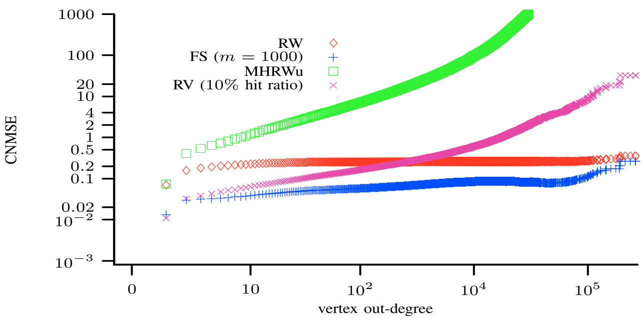

Fig. 5. (Flickr) The log-log plot of the CNMSE of the in-degree distribution estimates with budget $n = |V|/100$.

A similar simulation with $n = |V|/100$ on the Flickr graph reveals a similar picture (results shown in Figure 5). RV (with 10% hit ratio) is the most accurate sampling method for degree $d = 1$ (with FS in a close second place). For degrees $d > 1$ FS is the most accurate method. RW, however, performs poorly when compared to FS (the CNMSE is up to one order of magnitude larger). MHRWu is again the least accurate method (where CNMSE$(d) > 1000$ when $d > 2 \times 10^4$).

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 module 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 for presenting the first problem formulation of the Metropolis-Hastings MSE from RE sampling. We also thank Ananthram Swami for helpful discussions regarding the MSE optimization and Alan Mislove for kindly making available some of the data used in this paper.

References

[1] 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.

[2] Nathan Eagle, Alex S. Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *PNAS*, 106(36):15274–15278, August 2009.

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

[4] Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. A walk in Facebook: Uniform sampling of users in online social networks. In *Proc. of the IEEE Infocom*, March 2010.

[5] 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.

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

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

[8] 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.

[9] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proc. of the KDD*, pages 631–636, 2006.

[10] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proc. of the WWW*, pages 695–704, 2008.

[11] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. AMS, 2009.

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

[13] 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.

[14] 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.

[15] 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.

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

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

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

[19] 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.

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