Estimating and Sampling Graphs with Multidimensional Random Walks

UMass Amherst Technical Report UM-CS-2010-011

Bruno Ribeiro
Computer Science Department
University of Massachusetts at Amherst
Amherst, MA, 01002
ribeiro@cs.umass.edu

Don Towsley
Computer Science Department
University of Massachusetts at Amherst
Amherst, MA, 01002
towsley@cs.umass.edu

ABSTRACT

Estimating characteristics of large graphs via sampling is a vital part of the study of complex networks. Current sampling methods such as (independent) random vertex and random walks are useful but have drawbacks. Random vertex sampling may require too many resources (time, bandwidth, or money). Random walks, which normally require fewer resources *per sample*, can suffer from large estimation errors in the presence of disconnected or loosely connected subgraphs. In this work we propose a new m-dimensional random walk that uses m *dependent* random walkers. We prove that the proposed sampling method, which we call *Frontier sampling*, has all of the nice sampling properties of a regular random walk. At the same time, our simulations over large real world graphs show that, in the presence of disconnected or loosely connected subgraphs, *Frontier sampling* exhibits lower estimation errors than regular random walks. We also argue that *Frontier sampling* is more suitable to sample power-law graphs than random vertex sampling.

Keywords

Multidimensional Random Walks, Graph sampling, Estimation, Degree distribution, Global clustering coefficient, Frontier Sampling

1. INTRODUCTION

A number of recent studies [6, 10, 15, 19, 20, 27, 30, 28, 29, 34] (to cite a few) are dedicated to the characterization of complex networks. A complex network is a network with non-trivial topological features (features that do not occur in simple networks such as lattices or random networks). Examples of such networks include the Internet, the World Wide Web, social, business, and biological networks [6, 28]. This work represents a complex network as a directed graph with labeled vertices and edges. A label can be, for instance, the degree of a vertex or if an individual is HIV positive. Examples of network characteristics include the degree distribution, the fraction of HIV positive individuals in a population [25], the average number of copies of a file in a peer-to-peer network [16], or the percentage of inaccessible pages in a corpus of documents indexed by a search engine [3].

Characterizing the labels of a graph requires querying vertices and/or edges; each query has an associated cost in resources (time, bandwidth, money). Characterizing a large graph by 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. In this work we present a new tool to characterize complex networks. In what follows *random vertex (edge) sampling* refers to sampling vertices (edges) independently and uniformly at random (with replacement).

Distinct sampling strategies have different resource requirements depending on the network being sampled. For instance, in a network where each vertex is assigned a unique user-id (e.g., travelers and their passport numbers) it is a widespread practice to perform random vertex sampling by querying randomly generated user-ids. This approach can be resource intensive if the user-id space is sparsely populated (e.g., less than 10% of all MySpace user-ids between the highest and lowest valid user-ids are currently occupied [30]). Another way to sample a network is by querying edges instead of vertices. Sampling edges randomly is, often, much more difficult as edges do not usually have ids. We summarize the drawbacks of random vertex and edge sampling:

- Random vertex (edge) sampling can be impractical (e.g. Web graph).
- Among networks in which random sampling is possible, its cost can be too high (e.g. MySpace, peer-to-peer network).
- Even when random vertex sampling is reasonably cheap, some characteristics may be better estimated with random edge sampling (e.g. the degree distribution tail of a power-law graph).

An alternative, and often cheaper, way to sample a network is with a random walk (RW). A RW samples a graph by moving a particle (walker) from a vertex to a neighboring vertex (through an edge). By this process edges and vertices are sampled. The probability by which the random walker selects the next neighboring vertex determines the probability by which vertices and edges are sampled. In this work we are interested in random walks that sample *edges* uniformly. These samples can be used to obtain unbiased estimates of a variety of graph characteristics (we present two examples in Section 4).

In order to estimate network characteristics, we assume that a random walker has the ability to query incoming and outgoing edges of a vertex (Section 4 provides the reason behind this assumption). This is possible over graphs such as Twitter, LiveJournal [27], YouTube [27], Facebook [15], MySpace [30], P2P networks [29], and the arXiv citations network. We revisit the theory behind random walks in Section 4.

Sampling graphs with random walks is not without drawbacks. The accuracy of the estimates depends not only on the graph structure but also on the characteristic being estimated. The graph structure can create distortions in the estimates by "trapping" the random walker inside a subgraph. An extreme case happens when the graph consists of two or more disconnected components. For instance, wireless mobile social networks exhibit connection graphs with multiple disconnected subgraphs [10]. But even connected graphs can suffer from the same problem. A random walker can get "temporarily trapped" and spend most of its sampling budget exploring the local neighborhood near where it got "trapped". In such scenario, estimates may be inaccurate if the characteristics of the local neighborhood differ from the overall characteristic of the graph. This problem is well documented in the literature (see [21, 31]) and our goal is to mitigate it.

Contributions

This work proposes a new multidimensional random walk sampling method (*Frontier sampling*) that preserves all of the important statistical properties of a regular random walk, while mitigating the large estimation errors caused by disconnected or loosely connected subgraphs that can "trap" a random walker and distort the estimated graph characteristic. In Section 6 we see that estimates from Frontier sampling have smaller Mean Squared Errors (MSEs) than estimates obtained from regular random walkers (single and multiple independent walkers (reviewed in Section 4.4)) in a variety of scenarios.

We make two additional contributions: (1) we compare random walk-based estimates to random vertex and random edge sampling. We show analytically that the tail of power law graphs is better estimated using random walks (or random edge sampling) than using random vertex sampling. These results help explain recent empirical results [29]; (2) another contribution of our work comes in the form of estimators of graph characteristics. While the literature focuses on vertex-centric estimators for random walks (estimators that use sampled vertices), e.g. Respondent-Driven Sampling (RDS) [34], casting these estimators as edge-centric simplifies the design of edge-centric characteristic estimators such as the global clustering coefficient.

Outline

The outline of this work is as follows. Section 2 presents the notation used in this paper. Section 3 contrasts random vertex with random edge sampling. Section 4 revisits single and multiple independent random walk sampling and estimation. Section 5 introduces *Frontier Sampling* (FS), a sampling process that uses m dependent random walkers in order to mitigate the high estimation errors caused by disconnected or loosely connected subgraphs. Section 5 also shows that FS can be seen as an m-dimensional random walk over the m-th Cartesian power of the graph (formally defined in Section 5). In Section 6 we see that FS outperforms both single and multiple independent random walkers in a variety of scenarios. We also compare independent sampling of vertices and edges with FS sampling. Section 7 reviews the relevant literature. Finally, Section 8 presents our conclusions and future work.

2. NOTATION

We present a formal definition of the sampling problem. Let $G_d = (V, E_d)$ be a labeled directed graph representing the (original) network graph. We assume that each vertex in G_d has at least one incoming or outgoing edge. An edge in G_d is an ordered pair of nodes (u, v) representing a connection from u to v. The in-degree of a vertex u in G_d is the number of distinct edges $(v_1, u), \ldots, (v_k, u)$ into u, and its out-degree is the number of distinct edges $(u, v_1), \ldots, (u, v_k)$ out of u. If a random walker has the ability to retrieve incoming and outgoing edges from a queried vertex (and vertices are distinguishable), then we can represent G_d as an undirected graph. Let $G = (V, E)$ be the undirected counterpart of G_d, i.e., $E = \{(u, v) | (u, v) \in E_d \vee (v, u) \in E_d\}$. Note that G is not necessarily connected. Let $\deg(v)$ denote the degree of vertex v in G. Let $\mathcal{L}$ be a finite set of vertex labels (it is trivial to extend our results to include edge labels). Each vertex in $v \in V$ is associated to a set of labels $\mathcal{L}(v) \subseteq \mathcal{L}$. For instance, a vertex label in G can be its in-degree in the original graph G_d.

Let $\hat{\theta}_l$ be the estimated fraction of vertices with label l. The error metric used in most of our examples is the normalized root mean square error of $\hat{\theta}_l$, which is a normalized measure of the dispersion of the estimates, defined as

$$\mathrm{NMSE}(l) = \frac{\sqrt{E[(\hat{\theta}_l - \theta_l)^2]}}{\theta_l}\,. \tag{1}$$

For the sake of simplicity, in the reminder of this paper we assume that all queries of edges and vertices have unitary cost and that we have a fixed sampling budget B (generalizing the unitary cost assumption is quite straightforward).

3. VERTEX V.S. EDGE SAMPLING

Here we consider a simple estimation problem. We use this to illustrate the tradeoff between random edge and random vertex sampling. Consider the problem of estimating the out-degree distribution of G_d. Let θ_i be the fraction of vertices with out-degree $i > 0$ and $E[D]$ be the average out-degree. We assume that $E[D]$ is known and that a sampled edge (u, v) only provides the out-degree of u. It is easy to see that the probability that random edge sampling samples a vertex with out-degree i is $\pi_i = i\,\theta_i / E[D]$. Random vertex sampling samples a vertex with out-degree i with probability θ_i. A simple calculation shows that the NMSE (equation (1)) of B randomly sampled *edges* with out-degree i is

$$\mathrm{NMSE}(i) = \sqrt{(1/\pi_i - 1)/B}\,, \quad i > 0. \tag{2}$$

Similarly, the NMSE of randomly sampled *vertices* with out-degree i is

$$\mathrm{NMSE}(i) = \sqrt{(1/\theta_i - 1)/B}\,. \tag{3}$$

Now note that $\pi_i/\theta_i = i/E[D]$, which means that $\pi_i > \theta_i$ if $i > E[D]$ and $\pi_i < \theta_i$ if $i < E[D]$. From equations (2) and (3) we see that random edge sampling more accurately estimates large out-degrees ($i > E[D]$) while random vertex sampling more accurately estimates small out-degrees ($i < E[D]$). This means random edge sampling exhibits smaller NMSE when estimating the tail of the out-degree distribution. This characteristic of random edge sampling is also known as importance sampling estimation [31].

The example above is just one of many instances where random edge sampling is preferred over random vertex sampling. Another example: random edge sampling simplifies the estimation of edge-centric graph characteristics such as the global clustering coefficient. Unfortunately, as discussed in Section 1, random edge sampling is rarely practical. In what follows we see that, if G is connected, random walks exhibit similar statistical properties to random edge sampling, without the (costly) need of independence.

4. RANDOM WALK SAMPLING

In this section we review random walk (RW) sampling and estimation over G. In what follows we assume that G is connected and non-bipartite. Sampling G with a RW is a simple task. A random walker with budget B starts at vertex $v_0 \in V$. For the sake of simplicity, in the reminder of this work we assume that all queries of edges and vertices have unitary cost and that we have a fixed sampling budget B (generalizing the unitary cost assumption is quite straightforward). Let $V' = \{v_i\}_{i=1}^{B}$ be a sequence of sampled vertices and $E' = \{(u_i, v_i)\}_{i=1}^{B}$ be the corresponding sequence of sampled edges in a RW. We define V' and E' as sequences because the same vertices (edges) may be sampled multiple times. We refer to $v_i \in V'$ and $(u_i, v_i) \in E'$ as the i-th sampled vertex and edge, respectively. At the n-th step the random walker in vertex v chooses an outgoing edge (v, u) uniformly at random. The walker adds v to V' and (v, u) to E'. At step $n + 1$ the random walker starts at vertex u and the sampling continues until step B. The RW described here is the most common type of RW found in the literature [23]. Other types of random walks differ in the way outgoing edges are sampled (e.g., random walks that mimic random vertex sampling); please refer to [31] for more details.

An important property of a RW is its ability to reach a unique stationary regime. A necessary condition for stationarity is that G must be connected and non-bipartite (the non-bipartite assumption can relaxed in a lazy random walk [23]). In a stationary RW, E' is a stationary sequence. A sequence $X_1, X_2, \ldots$ of random variables is said to be stationary if for any positive integers n and k, the joint distribution of $(X_n, \ldots, X_{n+k})$ is independent of n. Stationarity is a natural generalization of random sampling where the assumption of independence is dropped. Once it reaches steady state, the above RW shares two important properties with random edge sampling. Both sample edges uniformly at random (as shown in Appendix A) and both obey the strong law of large numbers, as we see next.

4.1 Strong Law of Large Numbers

The strong law of large numbers is a powerful tool that states that the sample average of of any function over the samples converges almost surely to its expected value. This property is very useful in building accurate estimators. In this section we see that the average of any function f over the sampled edges (vertices) of a stationary RW converges almost surely to its expected value, under certain constraints [26]. Here we provide details of this known property, which is included here for the sake of completeness. Let X_n be the n-th edge sampled by a stationary RW (a similar result can be obtained for the n-th sampled vertex).

Theorem 4.1 (SLLN). *A stationary RW satisfies the strong law of large numbers, namely that for any function f, where $\sum_{(u,v)\in E} |f(u,v)| < \infty$,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(X_n) \overset{a.s.}{=} \frac{1}{|E|}\sum_{\forall(u,v)\in E} f(u,v)\,,$$

where "a.s." denotes "almost surely" converge, i.e., the event happens with probability one.

Proof. The Markov chain associated to a random walker over G is ergodic, as G is undirected and non-bipartite [23]. Thus we can apply the Strong Law of Large Numbers for ergodic Markov chains [26, Chapter 17] with the fact that any edge $(u, v) \in E$ is sampled with probability $1/|E|$ to prove the result. □

Theorem 4.1 allows us to construct estimators of graph characteristics that converge to their true values as the number of random walk steps goes to infinity ($n \to \infty$). In what follows we apply Theorem 4.1 to estimate graph characteristics; we also present two examples.

4.2 Estimators

An *estimator* is a function that takes the observations (sampled data) as input and outputs an estimate of a unknown population parameter (graph characteristic). In this section we see how to estimate graph characteristics using E' (the sampled edges of a RW). Estimators that take V' as input are commonly used to estimate vertex-oriented metrics (such as the degree distribution) and can be found in the literature [31, 34].

Here we present estimators of two graph characteristics: the vertex (edge) label density (the fraction of vertices (edges) with a given label in the graph) and the global clustering coefficient. The design of the estimator is simple: (1) First we find a function f that computes the characteristic of G assuming $V' = V$ and $E' = E$; (2) later we replace the assumption that $V' = V$ and $E' = E$ with the assumption that V' and E' are sequences drawn from a stationary RW.

4.2.1 Label Density

We illustrate how to build an estimator using a simple example. Recall that we can record the in- and out-degrees of G_d as vertex labels in G. Each vertex in $v \in V$ is associated with a label $\mathcal{L}(v) \subseteq \mathcal{L}$, where $\mathcal{L}$ is the set of labels defined in Section 2. A label can be, for instance, the in-degree of v in the original graph G_d. We seek to calculate, θ_l, the fraction of vertices with label l in G. The following estimator is a simple edge-based version of the vertex-based RDS estimator [34]. Because (for now) we assume that V and E are known, we have

$$\theta_l \equiv \sum_{\forall(u,v)\in E} f(u,v) \equiv \sum_{\forall(u,v)\in E} \left(h_l(v) + h_l(u)\right), \qquad (4)$$

where

$$h_l(v) = \begin{cases} \dfrac{1}{\deg(v)|V|} & \text{if } l \in \mathcal{L}(v) \\ 0 & \text{otherwise.} \end{cases}$$

It is trivial to verify that θ_l is the fraction of vertices with label l. Now we replace the assumption that $V' = V$ and $E' = E$ with the assumption that $E' = \{(u_i, v_i)\}_{i=1}^{B}$ is a

sequence of B edges sampled by a stationary RW. To eliminate the dependence of h_l on any unknown values (e..g $|V|$ and $|E|$) we need to redefine h_l:

$$h'_l(v) = \begin{cases} 1/\deg(v) & \text{if } l \in \mathcal{L}(v) \\ 0 & \text{otherwise.} \end{cases}$$

Using the linearity of the expectation operator we have

$$\hat{\theta}_l \equiv \frac{1}{S\,B} \sum_{\forall (u,v) \in E'} h'_l(v) + h'_l(u)\,. \tag{5}$$

where $S = \sum_{\forall (u,v) \in E'} 1/\deg(v) + 1/\deg(u)$. From Theorem 4.1 we know that this estimator, $\hat{\theta}_l$, is asymptotically unbiased, i.e., $\lim_{B\to\infty} E[\hat{\theta}_l] = \theta_l$.

4.2.2 Global Clustering Coefficient

In the literature the term *clustering coefficient* often refers to the local clustering coefficient [35]. In our example we estimate a different metric: the *global* clustering coefficient. In a social network the global clustering coefficient, C, is the probability that the friend of John's friend is also John's friend [28]. More formally, the global clustering coefficient can be defined as [28]

$$C = \frac{6 \times \text{number of triangles in the graph}}{\text{number of directed paths of length two}},$$

where a triangle is a clique with 3 vertices and a directed path of length two refers to any directed path that connects two vertices in the graph. If $E' = E$ we can calculate the number of triangles

$$\Delta(E) \equiv \sum_{\forall (u,v) \in E} f_\Delta(u,v)/3, \tag{6}$$

where $f_\Delta(u,v)$ is a function that returns the number of common neighbors between u and v. We can also calculate the number of directed paths of length two

$$l(E) \equiv \sum_{\forall (u,v) \in E} f_l(u,v) \equiv \sum_{\forall (u,v) \in E} ((\deg(u)-1)+(\deg(v)-1)), \tag{7}$$

as an edge (u,v) belongs to $2(\deg(u)+\deg(v)-2)$ directed paths of length two and each path is counted twice in the summation. Note that C is well defined only if $l(E) > 0$. As with the previous estimator example, we replace the assumption that $E' = E$ with the assumption that E' is sampled by a stationary RW. Applying Theorem 4.1 we have that $\lim_{B\to\infty} l(E')/B = l(E)|E|$ and that $\lim_{B\to\infty} \Delta(E')/B = \Delta(E)|E|$. From the above we obtain the following lemma.

Lemma 4.2. *Let* $l(E) > 0$ *and*

$$\hat{C} = \frac{6\,\Delta(E')}{l(E')}\,.$$

Then $\hat{C}$ *is an asymptotically unbiased estimator of* C*, i.e.,* $\lim_{B\to\infty} E[\hat{C}] = C$.

PROOF. The proof can be found in Appendix B. □

4.3 Estimator Accuracy & Graph Structure

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. 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 over the estimates.

The above two types of estimation error are well documented in the literature and various solutions are available [13, 31]. For instance, if the random walker does not start in a stationary regime (transient), it is common practice to discard the first w samples [13]. The value of w is called the *burn-in period*. There are two problems with this solution: (1) it only reduces the error related to the non-stationarity of the samples; (2) it is difficult to determine a good value for w when the size and structure of G are unknown.

A simple naive solution to the RW "trapping" problem (adopted in [15] to sample Facebook), is to sample the graph using multiple independent random walkers [13]. In what follows we see that such naive approach can lead to increased estimation errors. In Section 5 we see how to mitigate the random walk "trapping" problem with K *dependent* random walkers.

4.4 Multiple Independent Random Walkers

Here we sample G using K (parallel) independent random walkers (*MultipleRW*). In order to distinguish MultipleRW and sampling using a single random walker, we denote the former *SingleRW*. To simplify our exposition we assume that if B is the sampling budget, each walker takes B/K steps. Let $(V_1, \ldots, V_K)$ be the state of K independent random walkers in steady state. It is easy to verify that, as in the SingleRW case, edges are sampled with probability $1/|E|$.

A drawback of MultipleRW can be explained using a simple example. Consider two random walkers ($K = 2$) walking over a graph that has two loosely connected large (and dense) subgraphs G_A and G_B. Let $\text{vol}(G_A)$ and $\text{vol}(G_B)$ denote the total number of edges in G_A and G_B, respectively. A random walker stuck in G_A samples edges of G_A with probability close to $1/\text{vol}(G_A)$. Another random walker stuck in G_B samples edges of G_B with probability close to $1/\text{vol}(G_B)$. If $1/\text{vol}(G_A) > 1/|E|$ then edges of G_A are oversampled and, consequently, the edges of G_B are undersampled (as $1/\text{vol}(G_B) < 1/|E|$). On the other hand, when $1/\text{vol}(G_A) < 1/|E|$ the edges of G_A are undersampled and the edges of G_B are oversampled. Increasing the sampling budget minimizes the problem only if G is connected.

Moreover, the reduction by a factor of K in the budget of each random walker can exacerbate their non-stationarity. This issue is well documented in the literature and there seems to be no consensus whether MultipleRW estimates are more accurate than SingleRW ones (refer to [13, 31] for a discussion). MultipleRW can also be used to detect the convergence of the estimates to their true value by, say, comparing the estimates obtained by each RW with the estimates combining all K RW together (e.g. Gelman-Rubin convergence diagnostics [12]). Here too there is no consensus. Some authors argue that convergence is better diagnosed by dividing a longer SingleRW into K non-overlapping segments [13].

We say that a graph is *homogeneously explored* by a set of random walkers when the edge sampling probabilities of each sampled edge are similar. In Section 6 we see practical

examples of non-homogeneous exploration by MultipleRW; we also see that this implies large estimation errors. Thus, Sections 4.3 and 4.4 leave us with the following question:

> **Q:** *Is it possible to "homogeneously explore" a graph using multiple random walkers?*

5. FRONTIER SAMPLING

In this section we present a new promising approach to address the above question. *Frontier Sampling* (FS) performs m *dependent* random walks in the graph. We refer to m as the dimension of the FS random walk. The FS algorithm is simple:

(1) $n \leftarrow 0$ *//n is the number of steps//*

(2) Initialize with a collection of m starting vertices $L = (v_1, \ldots, v_m)$

(3) Select $u \in L$ with probability $\deg(u) / \sum_{\forall v \in L} \deg(v)$

(4) Select an outgoing edge of u, (u, v), uniformly at random

(5) Replace u by v in L, add u to V', and add (u, v) to E'

(6) $n \leftarrow n + 1$

(7) While $n < B - K$ goto line (3)

Frontier Sampling (FS) 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 subgraphs than a single random walker. However, unlike m independent random walkers, all m Frontier samplers (random walkers) share the same sampling process and budget. In all of our simulations, presented in Section 6, FS estimates are more accurate than both single and m independent random walkers. Section 8 describes how the FS algorithm can be made fully distributed.

Frontier Sampling: An m-dimensional Random Walk

Now we see that FS shares many of the same statistical properties of a single random walker. The key insight behind Theorem 5.2 below is that the FS stochastic process is equivalent to the stochastic process of a single random walker over the m-th Cartesian power of G, $G^m = (V^m, E_m)$, where

$$V^m = \{(v_1, \ldots, v_m) \,|\, v_1 \in V \wedge \cdots \wedge v_m \in V\}$$

is the m-th Cartesian power of V and $\forall \mathbf{v}, \mathbf{u} \in V^m$, $(\mathbf{v}, \mathbf{u}) \in E_m$ if exists an index i such that $(v_i, u_i) \in E$ and $u_j = v_j$ for $j \neq i$.

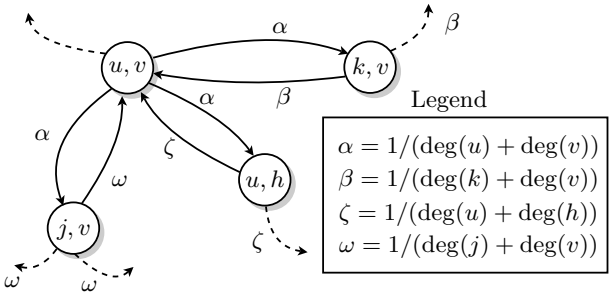

Figure 1: Illustration of the Markov chain associated to the Frontier sampler with dimension $m = 2$.

Lemma 5.1. *The Frontier sampling process is equivalent to the sampling process of a single random walker over G^m.*

PROOF. Consider the $(n-1)$-st step of FS. The reader may find Figure 1 helpful in following the proof. Let $L_n = (v_1, \ldots, v_m)$ be the state of FS before the n-th step. Clearly $L_n \in V^m$. Let $e(L_n)$ denote the collection of all edges associated to the vertices in L_n. We refer to $e(L_n)$ as the edge frontier at the n-th step. We describe the transition from state L_n to state L_{n+1} as follows (lines (3) and (4) of the FS algorithm): Select a vertex $v \in L_n$ with probability proportional to $\deg(v)$ and then replace vertex v in L_n with one of its neighbors (selected uniformly at random). This is equivalent to randomly sampling an edge from $e(L_n)$ with probability

$$p = \frac{1}{|e(L_n)|} = \frac{1}{\sum_{\forall v \in L_n} \deg(v)}.$$

Therefore, L_n is able to transition to state L_{n+1} iff $(L_n, L_{n+1}) \in E_m$ and the transition probability from L_n to L_{n+1} is $1/|e(L_n)|$. Thus, the Markov chain that describes FS is equivalent to the Markov chain of a single random walker over G^m. □

Theorem 5.2. *If G is connected and non-bipartite, then FS is asymptotically stationary and has a unique stable distribution where: (1) edges are sampled with probability $1/|E|$, (2) sampled edges form a stationary sequence, and (3) the sequence satisfies the Strong Law of Large Numbers (Theorem 4.1).*

PROOF. The proof is found in Appendix C. □

6. RESULTS

In this section we compare FS with SingleRW and MultipleRW. We also contrast FS with random vertex and edge sampling. The experiments consist of executing these sampling methods on a variety of real world graphs. The datasets used in the simulations are summarized in Table 1: "Flickr", "Livejournal", and "YouTube" are popular photosharing, blog (weblog), and video sharing websites, respectively. Users are represent as vertices of a graph. In these websites a user can subscribe to other user updates; an edge (u, v) exists between users u and v if user u subscribes to user v. At "Livejournal" and "YouTube" it is possible to query the incoming and outgoing edges of a given user. Further details of these three datasets can be found in [27]. "Hep-th Citations" is a graph of citation references in the ArXiv high energy physics publications archive [36]. "Internet RLT" is a router-level Internet graph collected from traceroute measurements of 23 monitors distributed over the world [11]. Note that some of these graphs contain disconnected components (subgraphs).

In the following simulations the starting vertex of each random walker is chosen uniformly at random from the set of all vertices. Our results show that FS estimates are consistently more accurate than their SingleRW and MultipleRW counterparts. Moreover, when restricted to the largest connected component, FS reaches steady state faster than SingleRW and MultipleRW in the simulations presented in Appendix D.

6.1 In- and Out-degree Distribution Estimates

Here we treat the graphs in Table 1 as undirected graphs. In-degrees and out-degrees are represented as vertex labels. Consider the in-degree distribution. Let $\theta = \{\theta_i\}_{\forall i}$ denote the in-degree distribution, where θ_i is the fraction of vertices

Graph	Flickr	LiveJournal	YouTube	Hep-th Citations	Internet RLT
Description	Social Net.	Social Net.	Social Net.	ArXiv pubs.	Internet tracert.
Type of graph	Directed	Directed	Directed	Directed	Directed
# of Vertices	1,715,255	5,204,176	1,138,499	27,770	192,244
Size of LCC	1,624,992	5,189,809	1,134,890	27,400	609,066
# of Edges	22,613,981	77,402,652	9,890,764	352,807	609,066
Average Degree	12.2	14.6	8.7	12.7	3.2
% of Original Graph	26.9%	95.4%	$\leq$ 79.1%	NA	NA

Table 1: Summary of the graph datasets used in our simulations. "Size of LLC" refers to the size of the largest connected component.

with in-degree i. In our simulations we estimate θ using equation (5). Each simulation consists of 10,000 runs (sample paths) to compute the empirical NMSE (equation (1)), which is then used to compare the accuracy of the estimates obtained from FS (dimension $m \in \{10, 100, 1000\}$), SingleRW, and MultipleRW ($K \in \{10, 100, 1000\}$ walkers). For the sake of conciseness, we restrict our presentation to a handful of representative results.

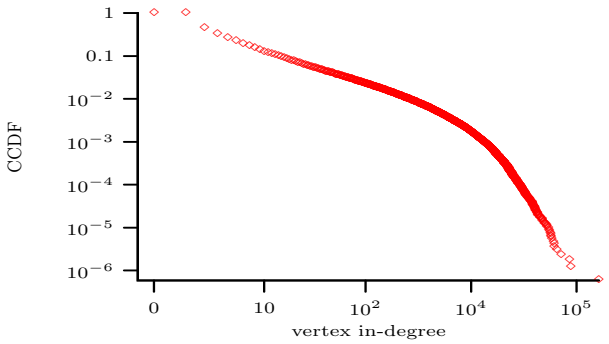

Figure 2: Flickr: Log-log plot of the in-degree CCDF.

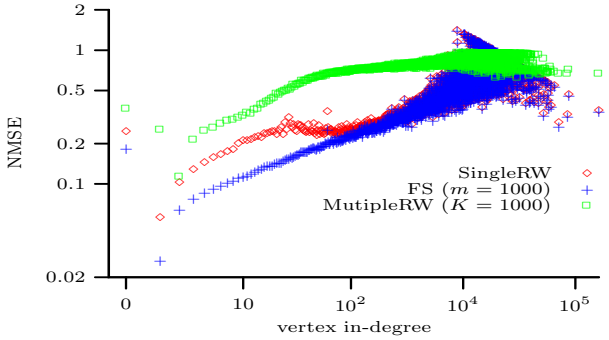

Figure 3: (LCC of Flickr) The log-log plot of the NMSE of the in-degree distribution estimates with budget $B = |V|/100$.

Consider first two representative results from the Flickr graph, whose in-degree CCDF (complementary cumulative distribution function) log-log plot is shown in Figure 2. The sampling budget is $B = 18,123 = |V|/100$, which amounts to sampling 1% of the vertices. In the first simulation, the sampling is restricted to the *Largest Connected Component* (LCC) (which contains 94% of the vertices). The objective is to test if FS can outperform SingleRW and MultipleRW even when there are no disconnected subgraphs. Figure 3 shows a log-log plot of the NMSE of FS ($m = 1000$), SingleRW, and MultipleRW ($K = 1000$). First, we note that the shape of the NMSE for high in-degrees is a consequence of the fact that vertices with high degrees in G tend to have unique high

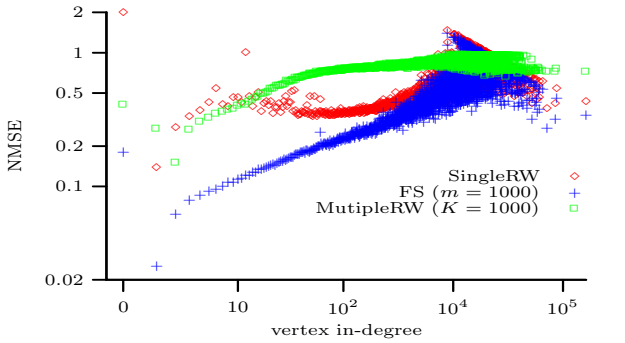

Figure 4: (Flickr) The log-log plot of the NMSE of the in-degree distribution estimates with budget $B = |V|/100$.

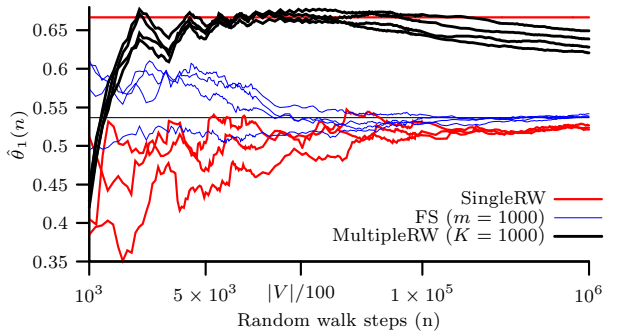

Figure 5: (LCC of Flickr) Four sample paths of $\hat{\theta}_1$ ($\theta_1 = 0.53$) as a function of the number of steps n (horizontal axis in log scale).

in-degree labels and that, similar to random edge sampling, the NMSE decreases with the degree. Figure 3 shows that FS outperforms both SingleRW and MultipleRW (particularly at estimating small in-degrees). It is interesting to note that, for most degrees, estimates obtained by SingleRW are more accurate than the estimates obtained by MultipleRW. Now consider the complete Flickr graph. Figure 4 shows a log-log plot of the NMSE of the in-degree distribution estimates. Contrasting the plots in Figures 3 and 4 we see that the gap between FS and both SingleRW and MultipleRW has significantly increased, favoring FS.

To better understand the differences between these sampling methods, Figure 5 focuses on four runs (sample paths) of the simulation over the complete Flickr graph. Figure 5 plots the evolution of $\hat{\theta}_1$ (the estimate of θ_1) as a function of n (the number of steps in the random walk). At each run of the simulator both FS and MultipleRW start at the same vertices (initially chosen using random vertex sampling). Figure 5 shows that all four FS sample paths (runs) quickly converge to the value of θ_1. For SingleRW, three out of the four runs start inside the LCC. These runs

do not converge to the value of θ_1 as some vertices with in-degree one lie outside the LCC. In one of the runs, SingleRW starts in a small disconnected subgraph and, thus, grossly overestimates the value of θ_1. For a similar reason, i.e., walkers starting at small disconnected subgraphs, MultipleRW grossly overestimates the value of θ_1. The MultipleRW jump around $n = 10^3$ steps needs further investigation. It may be due to the transient of the random walk (discussed in Section 4.4). Even when $n \gg 1$ (not shown in Figure 5) the MultipleRW estimate is unable to converge to θ_1. Modifying both SingleRW and MultipleRW methods to cope with disconnected components is an interesting open problem.

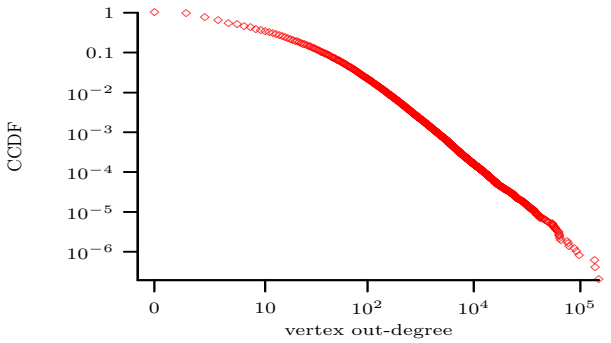

Figure 6: (Livejournal) Log-log plot of the out-degree CCDF.

For the sake of conciseness, we omit the results of the simulations over the remaining graphs (Table 1) as they are similar to the results observed over the Flickr graph. However, consider the *out-degree* distribution estimates of Livejournal. Figure 6 shows a log-log plot of the CCDF of the original out-degrees. The log-log plot of the NMSE is shown in Figure 7 for FS ($m = 100$), SingleRW, and MultipleRW ($K = 100$) with sampling budget $B = |V|/10$. From Figure 7 we see that estimates of vertices with small out-degrees in FS are up to one order of magnitude more accurate than those obtained from both SingleRW and MultipleRW.

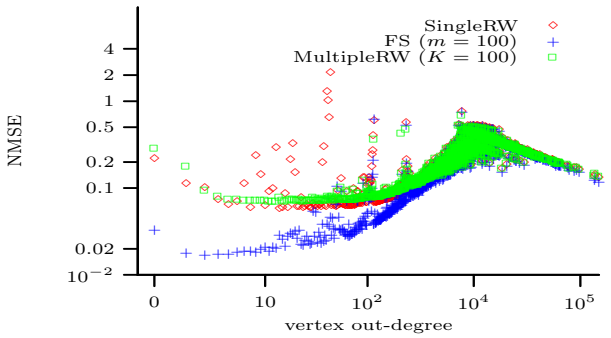

Figure 7: (Livejournal) The log-log plot of the NMSE of the out-degree distribution estimation with sampling budget $B = |V|/10$ (MSE over 10,000 runs).

The next experiment focuses on studying the impact of loosely connected subgraphs over the degree estimates. Consider a graph that consists of two instances of a random undirected Barabási-Albert [4] graph, G_A and G_B, with 5×10^5 vertices each and average degrees 2 and 10, respectively, joined by a single edge connecting the two smallest degree vertices in G_A and G_B (ties are resolved arbitrarily). Henceforth, this graph is referred to as G_{AB}.

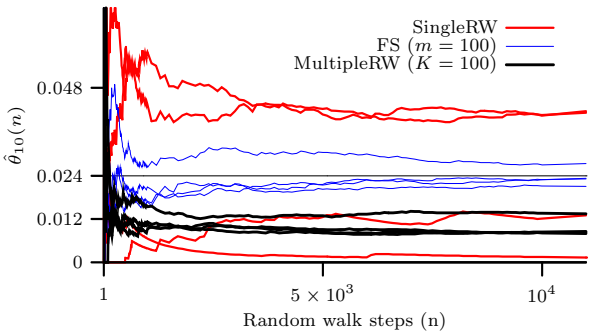

Figure 8: (G_{AB} graph) Four paths of $\hat{\theta}_{10}$ as a function of the number of steps n ($\theta_{10} = 0.024$).

The experiment consists of estimating the degree distribution of G_{AB} using FS ($m = 100$), SingleRW, and MultipleRW ($K = 100$). Again, both FS and MultipleRW start start at the same vertices in each execution of the simulation, which are initially chosen uniformly at random. In this experiment the hypothesis is that, for small sampling budgets, each random walker will see the degree distribution of either G_A or G_B but not the degree distribution of G_{AB}. Moreover, as the starting vertex of each random walker is chosen uniformly at random, G_A, which has the same number of vertices as G_B but 1/5 of the edges, receives more random walkers than its *per edge* "share". Consequently, MultipleRW oversamples G_A.

Figure 8 shows the results of four simulation runs and plots the evolution of the estimates of θ_{10} ($\hat{\theta}_{10}$) as a function of the number of steps. In this simulation note that: (1) FS quickly converges to a value that is close to the correct value; (2) two out of the four SingleRW runs overestimate θ_{10} and the remaining two underestimate it; (3) three out of the four MultipleRW runs converge to the same, incorrect, fraction (underestimating the true value of θ_{10}). FS is designed to be robust to disconnected or loosely connected subgraphs. All of the FS runs quickly converge to a good estimates of θ_{10}. Figure 9 also shows that the NMSE for FS, SingleRW, and MultipleRW, that of FS is consistently lower.

6.2 Frontier v.s. Random Sampling

In Section 3 we showed analytically that random edge sampling is more accurate than random vertex sampling when it comes to estimating the tail of the degree distribution. In this section we observe this to be true even when we replace random edge sampling with FS sampling. Here we estimate the in-degree distribution using the estimator $\hat{\theta}_i$, equation (5) (for the sampled edges, an estimator for sampled vertices is trivial). Let $B = |V|/100$ be the sampling budget; edges and vertices have unitary sampling costs, except when random edge sampling is performed. In the case of random edge sampling the cost of an edge is 2 (as each sampled edge also obtains two sampled vertices).

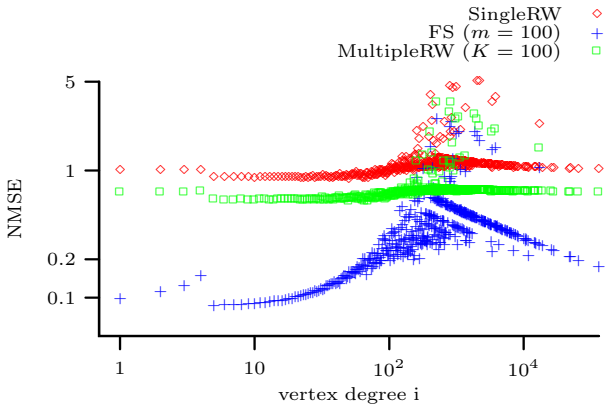

Figure 9: (G_{AB} graph) The log-log plot of the NMSE of the degree distribution estimation with sampling budget $B = |V|/10$ (MSE over 10,000 runs).

Now consider the (complete) Flickr graph. In random vertex sampling, as some complex networks exhibit sparse user-ids, a fraction of the sampling budget B can be spent querying invalid users. Thus, we evalute random vertex sampling over two scenarios: (1) where all queried user-ids are valid (thus, the number of sampled vertices is B), and (2) where only 10% of the queried user-ids are valid (thus, the number of sampled vertices is $B/10$). Recent experiments report that the fraction of invalid user-ids is 10% for the MySpace network [30]. Figure 10 shows a log-log plot of the NMSE. Observe from Figure 10 that FS ($m = 1000$) preforms as well as random edge sampling. As expected, FS and random edge sampling are more accurate at estimating large degrees than random vertex sampling, while random vertex sampling is more accurate at estimating small degrees than FS and random edge sampling.

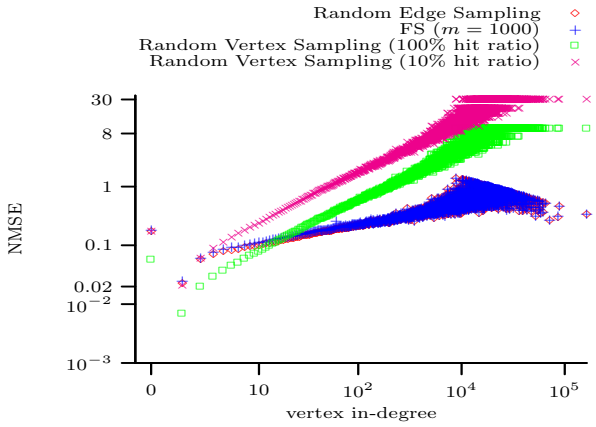

Figure 10: (Flickr) The log-log plot shows the NMSE of the in-degree distribution estimation with budget $B = |V|/100$ (MSE over 10,000 runs).

6.3 Density of Special Interest Groups

In a variety of complex networks, e.g. on-line social networks, each vertex (user) is associated with multiple labels that represent group affiliations, e.g. user interests, user geolocation, among others. For example, in the Flickr graph 21% of the users belong to one or more special interest groups [27]. Let $\mathcal{L}$ denote the set of groups in the Flickr graph and θ_l denote the fraction of vertices that belong to group $l \in \mathcal{L}$. In the simulations we estimate θ_l using FS ($m = 100$), SingleRW, and MultipleRW ($K = 100$) with budget $B = |V|/100$. Figure 11 shows the NMSE (from 10,000 runs) of the most popular 200 groups ordered in decreasing popularity. FS is clearly superior to both SingleRW and MultipleRW. Even when restricting the random walks to the largest connected component, FS still noticeably outperforms MultipleRW ($K = 100$) and SingleRW.

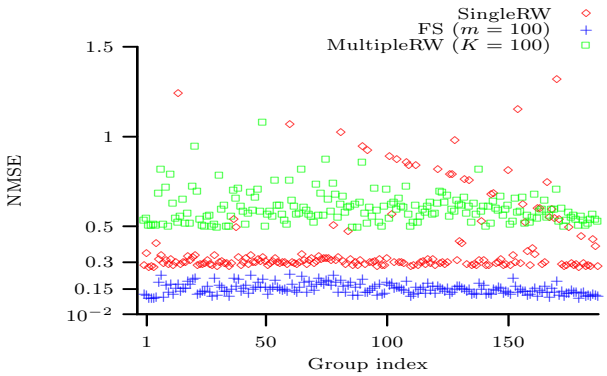

Figure 11: (Flickr) The NMSE of the density estimates of the most popular groups in the Flickr graph.

Graph	Budget(B)	C	$\hat{C} \pm \sqrt{MSE}$
Joint Barabási-Albert	$\|V\|/10$	10^{-4}	$10^{-4} \pm 10^{-5}$
Flickr	$\|V\|/20$	0.05	0.05 ± 10^{-3}
LiveJournal	$\|V\|/50$	0.06	0.06 ± 0.01

Table 2: Frontier sampling: global clustering coefficient estimates. C is the true value of the global clustering coefficient and $\hat{C}$ is its estimated value.

6.4 Global Clustering Coefficient Estimates

In our last set of experiments we evaluate the accuracy of estimating the global clustering coefficient using FS, SingleRW, and MultipleRW. Our simulations show little difference between FS ($m = 100$), SingleRW, MultipleRW ($K = 100$). Table 2 presents the average and the root mean squared error ($\sqrt{MSE}$) over 100,000 runs of the estimates obtained using FS over three graphs. From the results of Table 2 we see that FS accurately estimates the global clustering coefficient.

7. RELATED WORK

This section is devoted to review the related literature. FS can be classified as a Markov Chain Monte Carlo (MCMC) method. Other MCMC-based methods have been applied to characterize complex networks. Applications include, but are not limited to estimating: characteristics of a population [34] (e.g. estimation of HIV seroprevalence among drug users [25]), content density in peer-to-peer networks [16, 29, 33, 24], degree distributions of the Facebook on-line social graph [15], uniformly sampling Web pages from the Internet [17, 32], and uniformly sampling Web pages from a search engine's index [3]. However, they are all variants of a

single random walk and exhibit the same problems described in Section 4.3, namely samples from a single random walker can be collected from a local neighborhood in the graph, which in turn may lead to large estimation errors.

A number of real complex networks are known to have disconnected or loosely connected subgraphs. A large body of MCMC literature is dedicated to overcome the locality problem described in Section 4.3. However, they all either assume that the graph is very structured, e.g., a 2 dimensional lattice, or that the graph is completely known. These assumptions make the solutions inapplicable to our problem. A comprehensive list of MCMC methods and their characteristics can be found in [13, 14, 21, 31].

In a recent work [15], multiple independent random walkers have been used to obtain and test the convergence of Facebook degree distribution estimates. Unfortunately, as seen in Section 4.4, multiple random walks are not suited to sample loosely connected graphs. Moreover, due to the lack of ground truth, [15] compares their results with the results obtained from random vertex sampling and, thus, it is unclear whether their method is accurate.

Projecting a RW onto a higher dimensional space has been used in [8] to turn the Markov chain associated to the random walker nonreversable, which can speed up the mixing of the original RW. This technique, known as "lifting", has been used to accelerate distributed consensus in large networks of autonomous agents [22]. Unfortunately, as "lifting" requires profound changes to the RW, it is unclear if this method can be successfully used to estimate characteristics of complex networks. For instance, the RW in [22] has no known stationary sampling probability.

Multiple random walks also find other applications besides the one presented in this work. They are used to collect Web data [9], search P2P networks [5, 37], and decrease the time to discover "new wireless nodes" [1]. Dependent multiple random walks are also used in percolation theory [2].

8. DISCUSSION AND FUTURE WORK

In this work we presented a new promising random walk-based method (*Frontier sampling*) that mitigates the estimation errors caused by subgraphs that "trap" a random walker. Frontier sampling (FS) uses multiple (m) mutually *dependent* random walker. The dependence between walkers is designed to "better balance" their samples. These samples are shown to be the projection (onto the original graph) of a special type of m-dimensional (single) random walker. Simulations over real world graphs in Section 6 show that Frontier sampling (FS) is more robust than single and multiple independent random walkers to estimate in-degree distributions and the fraction of users that belong to a social group. We also present evidence, using an analytical argument (also substantiated by simulations), that random walks (in particular, FS) are better suited to estimate the tail of power law graphs than random vertex sampling.

Moreover, FS sampling is well suited to be used in large scale (parallel, asynchronous) experiments. This is because FS sampling can be achieved by multiple independent random walkers where the cost of sampling a vertex v is an exponentially distributed random variable with parameter $\deg(v)$. Using the the Uniformization principle of Markov chains [7, Chapter 7.5] and the Poisson decomposition property, and it can be shown that this MultipleRW with random exponential costs is equivalent to the FS sampling process described in Section 5 (Appendix E).

The ideas behind FS can have far reaching implications, from estimating characteristics of dynamic networks to the design of new MCMC-based approximation algorithms.

9. ACKNOLEDGMENTS

We thank Weibo Gong for many helpful discussions and Alan Mislove for kindly making available some of the data used in this paper. This research was sponsored by the ARO under MURI W911NF-08-1-0233, and the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

10. REFERENCES

[1] Chen Avin and Bhaskar Krishnamachari. The power of choice in random walks: An empirical study. *Comput. Netw.*, 52(1):44–60, 2008.

[2] P. Balister, B. Bollobás, and A. Stacey. Dependent percolation in two dimensions. *Prob. Theory and Related Fields*, 117(4):495–513, 2000.

[3] Ziv Bar-Yossef and Maxim Gurevich. Random sampling from a search engine's index. *J. ACM*, 55(5):1–74, 2008.

[4] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[5] Nabhendra Bisnik and Alhussein A. Abouzeid. Optimizing random walk search algorithms in p2p networks. *Computer Networks*, 51(6):1499–1514, 2007.

[6] 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.

[7] Christos G. Cassandras and Stephane Lafortune. *Introduction to Discrete Event Systems*. Springer-Verlag, Inc., 2006.

[8] Fang Chen, László Lovász, and Igor Pak. Lifting Markov chains to speed up mixing. In *Proc. of STOC*, pages 275–281, 1999.

[9] Junghoo Cho and Hector Garcia-Molina. Parallel crawlers. In *Proc. of the WWW*, pages 124–135, 2002.

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

[11] Coperative Association for Internet Data Analysis. CAIDA's Internet topology data kit #0304, 2003.

[12] A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7:457–511, 1992.

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

[14] Walter R. Gilks and Gareth O. Roberts. Strategies for improving MCMC. In *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics Series. 1st edition, December 1995.

[15] 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*, Jun 2010 (*to appear*).

[16] 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.

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

[18] Mark Kac. Random walk and the theory of Brownian motion. *The American Mathematical Monthly*, 54(7):369–391, 1947.

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

[20] 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.

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

[22] Wenjun Li and Huaiyu Dai. Accelerating distributed consensus via lifting Markov chains. In *Proc. of the IEEE ISIT*, pages 2881–2885, June 2007.

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

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

[25] Courtney McKnight, Don Des Jarlais, Heidi Bramson, Lisa Tower, Abu S. Abdul-Quader, Chris Nemeth, and Douglas Heckathorn. Respondent-driven sampling in a study of drug users in New York City: Notes from the field. *Journal of Urban Health*, 83(6):154–159, 2006.

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

[27] 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.

[28] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[29] Amir H. Rasti, Mojtaba Torkjazi, Reza Rejaie, Nick Duffield, Walter Willinger, and Daniel Stutzbach. Graph sampling in unstructured Peer-to-Peer and undirected online social networks. In *IEEE INFOCOM Mini-Conference*, 2009.

[30] Bruno Ribeiro, William Gauvin, Benyuan Liu, and Don Towsley. On MySpace account spans and double Pareto-like distribution of friends. Technical Report UM-CS-2010-001, UMass Amherst, Jan 2010.

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

[32] Paat Rusmevichientong, David M. Pennock, Steve Lawrence, and Lee C. Giles. Methods for sampling pages uniformly from the world wide web. In *AAAI Fall Symposium on Using Uncertainty Within Computation*, pages 121–128, 2001.

[33] 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.

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

[35] D.J. Watts and S.H. Strogatz. Collective dynamics of "small world" networks. *Nature*, 393:440–442, June 1998.

[36] KDD cup 2003 dataset (`http://www.cs.cornell.edu/projects/kddcup/datasets.html`).

[37] Ming Zhong and Kai Shen. Random walk based node sampling in self-organizing networks. *SIGOPS Oper. Syst. Rev.*, 40(3):49–55, 2006.

APPENDIX

A. STATIONARITY OF RANDOM WALKS

Let $\mathbf{P}$ be the transition probability matrix of the Markov chain associated to a random walker over G. Under the condition that G is connected and non-bipartite, the spectral analysis of $\mathbf{P}$ (described in Appendix A.1) shows that the random walker has an unique stationary (aka stable) distribution $\pi\mathbf{P} = \pi$ such that $\sum_{\forall v \in V} \pi_v = 1$. The spectral analysis also shows that the random walk is stationary if one of the two conditions hold: (1) v_0 is chosen from V with probability $\deg(v_0)/2|E|$; (2) starting at some (arbitrary) initial vertex v_0 is asymptotically stationary ($B \to \infty$) [23]. It can also be shown that starting at some (arbitrary) initial vertex v_0, and walking sufficiently many (say n) steps over the graph, then the $n+1$-st step is almost stationary (according to an appropriate metric), provided that n is large enough [23]. In the n-th step the steady state probability that a random walker is at vertices u and v is $\pi_u = \deg(u)/2|E|$ and $\pi_v = \deg(v)/2|E|$, respectively. Thus, the probability that a random walk samples an edge $(u, v) \in E$ at the $(n+1)$-st step is $\pi_u/\deg(u) + \pi_v/\deg(v) = 1/|E|$.

A.1 Spectral decomposition of a random walk

We start with known facts about random walks [23]. Let $\mathbf{A}$ be the adjacency matrix of G and let

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

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 transition probability matrix of the random walk. Let G be a connected, symmetric, and non-bipartite directed graph. A directed graph G is called symmetric if, for every edge that belongs to G, the corresponding inverted edge also belongs to G. The vertex and edge sampling probabilities of the random walk can be obtained as a function of the starting state using the spectral decomposition of $\mathbf{P}$. A transition probability matrix $\mathbf{P}$ of a random walk can be decomposed into its left and right eigenvectors and eigenvalues [18]

$$\varphi_k \mathbf{P} = \lambda_k \varphi_k \quad \text{and} \quad \mathbf{P}\psi_k = \lambda_k \psi_k,\ k = 1, \ldots, |V|,$$

where $\forall k$, $\langle \varphi_k, \varphi_k \rangle = \langle \psi_k, \psi_k \rangle = 1$, and the indexes k are ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{|V|}$. As G is connected, symmetric, and non-bipartite (and because $\mathbf{P}$ is a stochastic matrix), it follows from the Frobenius-Perron Theorem that $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|V|} > -1$ [23]. The probability that a random walk reaches vertex v in n steps, given that it starts from vertex u, is [18, 23]

$$p_{uv}^{(n)} = \frac{\deg(v)}{2|E|} + \sqrt{\frac{\deg(v)}{\deg(u)}} \sum_{k=2}^{|V|} \lambda_k^n \psi_k(u) \varphi_k(v) \,. \qquad (8)$$

As $|\lambda_k| < 1$ for all $k \neq 1$, it is straightforward to see that

$$\lim_{n\to\infty} p_{uv}^{(n)} = \deg(v)/2|E|,$$

which is the same result as solving $\pi\mathbf{P} = \pi$. It is also worth noting that [23]

$$\sum_{\forall u \in V} \pi_u p_{uv}^{(n)} = \deg(v)/(2|E|) \,, \; n = 0, 1, \ldots \,.$$

and if $\pi_v^{(0)}$ denotes the probability that the random walk starts at vertex v, the probability of sampling edge (t, s) is

$$\sum_{\forall u \in V} \pi_u^{(0)} \left(\frac{p_{ut}^{(n)}}{\deg(t)} + \frac{p_{us}^{(n)}}{\deg(s)} \right).$$

Equation (8) clearly shows the dependence between the n-th step sampling probability of v at and the graph structure (represented by $\mathbf{P}$'s spectral decomposition).

B. PROOF OF THE ASYMPTOTIC UNBIASEDNESS OF $\hat{C}$

Lemma. *Let $l(E) > 0$ and*

$$\hat{C} = \frac{6\,\Delta(E')}{l(E')} \,.$$

Then $\hat{C}$ is an asymptotically unbiased estimator of C, i.e., $\lim_{B\to\infty} E[\hat{C}] = C$.

PROOF. Let

$$l^\star = \lim_{B\to\infty} l(E')(|E|/B)$$

and

$$\Delta^\star = \lim_{B\to\infty} \Delta(E')(|E|/B).$$

From Theorem 4.1 we know that

$$l^\star \overset{\text{a.s.}}{=} \Delta(E)$$

and

$$\Delta^\star \overset{\text{a.s.}}{=} \Delta(E) \,.$$

Let $\Gamma_l = |l^\star - l(E)|$ and $\Gamma_\Delta = |\Delta^\star - \Delta(E)|$, note that $\Gamma_l \geq 0$ and $\Gamma_\Delta \geq 0$ are random variables. Almost sure convergence implies convergence in probability, i.e.,

$$P[\Gamma_l \geq \epsilon] = 0, \, \forall \epsilon > 0 \,,$$

i.e., for any $\epsilon > 0$, the probability that $l^\star$ is outside the ball of radius ϵ centered at $l(E)$ is zero. The same is valid for $\Delta^\star$. Now let

$$\hat{C}^- = \frac{6\,\Delta(E) - 6\,\Gamma_\Delta}{l(E) + \Gamma_l}$$

and

$$\hat{C}^+ = \frac{6\,\Delta(E) + 6\,\Gamma_\Delta}{l(E) - \Gamma_l}.$$

It can be shown that $\hat{C}^- \leq C \leq \hat{C}^+$ and that $\hat{C}^- \leq \lim_{B\to\infty} \hat{C} \leq \hat{C}^+$. Now

$$E[\hat{C}^-] = E\left[\frac{6\,\Delta(E)}{l(E) - \Gamma_l}\right] + E\left[\frac{6\,\Gamma_\Delta}{l(E) - \Gamma_l}\right].$$

and expanding the expectation (with a slight abuse of notation)

$$E\left[\frac{6\,\Delta(E)}{l(E) - \Gamma_l}\right] = \int_0^\infty \frac{6\,\Delta(E)}{l(E) - x} P[\Gamma_l = x] dx = \frac{6\,\Delta(E)}{l(E)},$$

as $l(E) > 0$. Similarly,

$$E\left[\frac{6\,\Gamma_\Delta}{l(E) - \Gamma_l}\right] = \int_0^\infty \int_0^\infty \frac{6\,y}{l(E) - x} P[\Gamma_l = x | \Gamma_\Delta = y] P[\Gamma_\Delta = y] = 0,$$

as $l(E) > 0$. The same can be shown for $\hat{C}^+$. Thus, as $E[\hat{C}^-] \leq \lim_{B\to\infty} E[\hat{C}] \leq E[\hat{C}^+]$ and $E[\hat{C}^-] \leq C \leq E[\hat{C}^+]$, we have $\lim_{B\to\infty} E[\hat{C}] = C$. □

C. PROOF OF THE FRONTIER SAMPLING THEOREM

In what follows we restate Theorem 5.2 and present a proof.

Theorem. *If G is connected and non-bipartite, then FS is asymptotically stationary and has a unique stable distribution where: (1) edges are sampled with probability $1/|E|$, (2) sampled edges form a stationary sequence, and (3) the sequence satisfies the Strong Law of Large Numbers (Theorem 4.1).*

PROOF. Consider the $(n-1)$-st step of Frontier sampling. The reader may find Figure 1 helpful in following the proof. Let $L_n = (v_1, \ldots, v_m)$ be the state of Frontier sampling before the n-th step. Clearly $L_n \in V^m$. In what follows let $e(L_n)$ denote the collection of all edges associated to the vertices in L_n. We refer to $e(L_n)$ as the edge frontier at the n-th step. We describe the transition from state L_n to state L_{n+1} as follows (lines 3 and 4 of the frontier sampling algorithm): Select a vertex $v \in L_n$ with probability proportional to $\deg(v)$ and then replace element v in L_n with one of its neighbors (selected uniformly at random). This is equivalent to randomly sampling an edge from $e(L_n)$ with probability

$$p = \frac{1}{|e(L_n)|} = \frac{1}{\sum_{\forall v \in L_n} \deg(v)}.$$

Thus, L_n is able to transition to state L_{n+1} iff $(L_n, L_{n+1}) \in E_m$ and the transition probability from L_n to L_{n+1} is $1/|e(L_n)|$. Thus, we conclude that Frontier sampling is a single random walker over the m-th Cartesian power of G, $G^m = (V^m, E_m)$, where

$$V^m = \{(v_1, \ldots, v_m) \,|\, v_1 \in V \wedge \cdots \wedge v_m \in V\}$$

is the m-ary Cartesian product of V and $\forall \mathbf{v}, \mathbf{u} \in V^m$, $(\mathbf{v}, \mathbf{u}) \in E_m$ if exists an index i such that $(v_i, u_i) \in E$ and $u_j = v_j$ for $j \neq i$.

Now we need to prove that the distribution of L_∞ is stable and unique. For this we only need to show that the random walk over G^m is ergodic. A random walk (Markov chain) is ergodic when it is aperiodic and recurrent non-null. Recall that the random walk over G is ergodic. The probability that Frontier sampling transitions from $L_n \in V^m$ to $L_{n+1} \in V^m$ such that L_n and L_{n+1} only differ in their i-th element is always greater than zero, otherwise there is an infinite increasing degree sequence in the vertices of G. But this is not possible as the random walk over G is recurrent non-null (an infinite increasing degree sequence would be a sink in the random walk over G). Thus, any finite sequence of transitions $\{L_{n+w}\}_{w=1}^{\Delta}$ that only updates its i-th element has probability greater than zero. Thus, as the sequence $\{L_{n+w}\}_{w=1}^{\Delta}$ is also a single random walk over G, it is aperiodic for any chosen $i = 1, \ldots, m$. Thus, a random walker over G^m must also be aperiodic. We can use the same argument to show that the random walk over G^m is recurrent non-null. As random walk over G^m is ergodic, we have that $L^\star$ is distributed according to the steady state distribution of a random walk over G^m

$$P[L_\infty = (v_1, \ldots, v_m)] = \frac{\sum_{i=1}^m \deg(v_i)}{|V|^m \sum_{\forall u \in V} \deg(u)},$$

where $L_\infty = \lim_{n\to\infty} L_n$, which is unique and stable (similar to a single random walker as seen in Section 4).

The rest of the proof is straightforward. Each edge in G^m is actually an edge in G. As each edge in G is copied m times into G^m, we have that edges in G are also sampled uniformly at random in a random walk over G^m. As Frontier sampling is a random walk over G^m, its samples form a stationary sequence and follow the Strong Law of Large Numbers seen in Theorem 4.1. The same is true for the sequence of sampled vertices. □

D. CONVERGENCE TO STATIONARITY

The first question we seek to answer with our simulations is how fast these random walk methods converge to their stationary edge sampling probabilities. In this simulation we set $K \in \{1, 10\}$ (number of independent random walkers), $m = 10$ (Frontier sampling dimension) and restrict our analysis to the largest connected component of the three graphs in our datasets with the smallest number of vertices (in order to speed the computation): "Internet RLT", "YouTube", and "Hep-th". Let $p_{u,v}^{(B)}$ denote the probability that a random walker, whose initial vertex is chosen uniformly at random, samples edge (u, v) at its the end of its sampling budget B. To measure the convergence to the stationary edge sampling probability, we use the largest relative difference between the stationary sampling probability $1/|E|$ and $p_{u,v}^{(B)}$:

$$\max_{(u,v)\in E} 1 - \frac{p_{u,v}^{(B)}}{1/|E|}.$$

Table 3 presents a Monte Carlo estimate of this relative difference. The 95% confidence interval of the Monte Carlo simulation is ±1%. Our estimates show that the difference between the transient and the stationary edge sampling probabilities of independent random walkers are between 5 and 42 times larger than the difference of Frontier sampling. This means that Frontier sampling converges faster to stationarity edge sampling probability.

		Sampling prob. error		
Graph	B (sampling budget)	FS	MRW	SRW
Internet RLT	100	17%	257%	156%
YouTube	20	43%	236%	216%
Hep-Th	20	36%	1510%	781%

Table 3: Relative worst-case difference between the steady state and the transient edge sampling probabilities after $B - K$ steps. Frontier edge sampling probabilities are closer to steady state in all graphs. Legend: (FS) = Frontier sampling ($K = 10$), (SRW) = Single ($K = 1$) Random Walker, and (MRW) = Multiple ($K = 10$) Random Walkers.

E. DISTRIBUTED FRONTIER SAMPLING

In what follows we present an informal proof that FS sampling can be achieved by multiple independent random walkers (MultipleRW) where the cost of sampling a vertex v is an exponentially distributed random variable with parameter $\deg(v)$.

Let $\mathbf{P}$ be the transition probability matrix of the Markov chain associated to a random walker over $G^m = (V^m, E_m)$, the m-th Cartesian power of G. Following Appendix A.1 we have

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{A},$$

where $\mathbf{A}$ is the adjacency matrix of G^m and $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{i,i} = \sum_{\forall j} \mathbf{A}_{i,j}$. According to Lemma 5.1 $\mathbf{P}$ is also the transition probability matrix of FS in G. Let $\mathcal{M} = \{L_n \in V^m : n = 0, \ldots\}$ denote the FS Markov chain (discrete-time), i.e., the transition probability matrix of $\mathcal{M}$ is $\mathbf{P}$. Now let $\chi = \{X(t) \in V^m : t \geq 0\}$ be a continuous-time Markov chain with transition rate matrix

$$\mathbf{Q} = \mathbf{A} - \mathbf{D},$$

observed during the (time) interval $[0, B]$. It is easy to see that the transition probability matrix of the embedded (discrete-time) Markov chain of χ, denoted by χ', is

$$\mathbf{P}' = I - \mathbf{D}^{-1}\mathbf{Q} = \mathbf{P}.$$

In the literature $\mathbf{P}'$ is known as the *Uniformized* counterpart of $\mathbf{Q}$ (with unitary uniformization rate) [7, Chapter 7.5]. Because $\mathbf{P}' = \mathbf{P}$, the stochastic processes χ' and $\mathcal{M}$ are equivalent.

Let $L'_n = (v_1, \ldots, v_m)$ denote the state of χ before the n-th step. Now note that because all off-diagonal non-zero transition rates in $\mathbf{Q}$ are equal to one, the probability that the k-th random walker transitions out of vertex v_k is independent of the state of all the other random walkers in L_n. Thus, we can decompose the Poisson process describing a departure from the state $L'_n = (v_1, \ldots, v_m)$ into m independent stochastic processes, where the i-th process is a Poisson process with rate $\lambda_i = \deg(v_i), i = 1, \ldots, m$. The above is equivalent to the stochastic process of a discrete-time MultipleRW with m random walkers and budget B, where the cost of sampling a vertex v is an exponentially distributed random variable with rate $\mu_v = \deg(v)$.