A General Framework of Hybrid Graph Sampling for Complex Network Analysis

Xin Xu Chul-Ho Lee Do Young Eun

Abstract—**Being able to capture the properties of massive real graphs and also greatly reduce data scale and processing complexity, graph sampling techniques provide an efficient tool for complex network analysis. Random walk-based sampling has become popular to obtain asymptotically uniform samples in the recent literature. However, it produces highly correlated samples and often leads to poor estimation accuracy in sampling large networks. Another widely-used approach is to launch random jump by querying randomly generated user/node ID, but also has the drawback of unexpected cost when the ID space is sparsely populated. In this paper, we develop a hybrid graph sampling framework that inherits the benefit of returning immediate samples from random walk-based crawling, while incorporating the advantage of reducing the correlation in the obtained samples from random jump. We aim to strike the right balance between random jump and crawling by analyzing the resulting asymptotic variance of an estimator of any graph nodal property, in order to give guidelines on the design of better graph sampling methods. We also provide simulation results on real network (graph) to confirm our theoretical findings.**

I. Introduction

Recently, complex networks such as online social networks, the World Wide Web, the Internet, and biological networks have drawn much attention from research community and inspired a number of measurement and theoretical studies, with their increasing popularity and extensive applications. The estimation of their topological properties, among others, has become important, but is often considered a non-trivial task. The huge size of such complex networks makes the complete dataset hard to obtain [1]. Also, commercial companies are typically unwilling to publish their data for some business reasons such as security concern and market competition [2]. As a result, researchers have turned their attention to 'sampled data' that can serve as a model of the whole network, representing its characteristics in a compact manner. Sampling techniques have also become essential for the practical estimation of network properties [1], [2], [3], [4], [5], [6], [7].

In terms of methodology, several graph sampling algorithms have been proposed to obtain a sampled dataset from the networks. Random walk-based crawling is an efficient distributed sampling method that can be used to obtain asymptotically uniform (unbiased) samples. The basic idea is to launch a random walk over a graph, which moves from a node to one of the neighbors in its vicinity, to obtain a set of samples. Examples in the literature for random walk-based crawling include two popular methods: simple random walk (SRW) with re-weighting [2], [6], [5] and Metropolis-Hastings random walk (MHRW) [4], [8], [9], [5], both of which can yield unbiased samples but have their own advantages. It has been shown through numerical simulations that SRW with re-weighting often has better estimation accuracy than MHRW [10], [2], [7]. However, SRW with re-weighting requires an additional overhead related to the re-weighting process.

On the other hand, a notable advantage of using MHRW is its 'ready-to-use' property. Specifically, once uniform node samples are collected by the MHRW for a given graph, they are immediately reusable for the estimation of other nodal properties of the graph. Nonetheless, any type of random walk-based algorithms generally suffers from slow diffusion over the whole network, which makes the obtained samples highly correlated and contributes to poor estimation accuracy.

Another uniform vertex sampling algorithm is 'random jump'-based sampling method [2]. In complex (social) networks such as Livejournal, MySpace and Flickr, each user occupies a unique ID within the value range, which enables us to perform random vertex sampling by querying randomly generated user IDs. This method will produce independent samples – far better than correlated samples in the case of crawling based ones as above, if the random query returns valid IDs. However, if the ID space is sparsely populated, which is often the case in reality,[1] most of the queried IDs are invalid, resulting in a waste of sampling attempts, i.e., high sampling cost.

Few recent studies have proposed to use a combination of crawling and random-jump sampling, taking their own advantages, which in turn improves the sampling performance [12], [13]. Crawling returns immediate samples for sure, but they are highly correlated with previous samples. While jumping, on the other hand, will give "high-quality" sample, independent of all others, but typically followed by many wasted querying attempts. The hybrid approaches aim at achieving the best of both worlds. In particular, they try to avoid the situation of a random walker getting trapped inside a subgraph (leading to the highly correlated samples), while reducing the cost by merely performing the random jump.

Specifically, the authors in [12] propose a version of a

The authors are with Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695-7911, Email: {xxu5, clee4, dyeun}@ncsu.edu

This work was supported in part by National Science Foundation under Grant No. CNS-1217341.

[1] For instance, about 10% of MySpace IDs are valid, and one valid user-ID exists among 77 IDs on average for Flickr [11].

978-1-4799-3360-0/14/$31.00 © 2014 IEEE

hybrid sampling model which incorporates into SRW auxiliary transitions proportional to vertex degree, aiming at accelerating the rate of convergence to the stationary distribution. They prove that the second largest eigenvalue, which is related to the mixing time, can decrease by introducing random restart. However, they do not provide an explicit guide to the choice of the restart rate (especially when taking into account the restart penalty – the cost of the random restart), but only heuristic trials on different simulation settings. Albatross sampling [13], which is an unbiased sampling algorithm, is proposed to improve the convergence time and estimation error for estimating degree distribution under a given sampling cost. However, the effectiveness of the algorithm is only validated over a limited set of numerical simulations without any theoretical support. In addition, sampling cost for the random jump in both of these algorithms is modeled as the *average* number of wasted ID queries to obtain one valid ID. This assumption often simplifies the analysis but is unable to assess the true impact of the sampling cost associated with random jumps, and thus prevents us from maximizing the benefits of the hybrid sampling.

In this paper, we develop an analytical framework for hybrid sampling approaches, which guarantees unbiased graph sampling and dynamically captures the cost of each random jump. Our goal here is to strike the right balance between the crawling-based techniques and random jump, and to analyze the resulting asymptotic variance of any given estimator in terms of the spectral properties of a given graph and the jump probability α. We show that the asymptotic variance is always a convex function of α for an arbitrary choice of nodal property of any graph, which implies that it's always possible to find the optimal α (thus the optimal hybrid sampler) by employing a suite of standard algorithms for minimizing a convex function [14]. We also show that the properties of asymptotic variance readily carry over to the mean squared error of *finite* samples of size t, in which case the error is no larger than $O(1/t)$. We conduct an extensive set of simulations over real graphs, using hybrid sampling algorithms with varying α for estimating degree distribution and clustering coefficients as our test metrics. We explore the relationship among the optimal jump probability, graph spectral properties, density of valid ID space, and the choice of nodal properties to be estimated. We demonstrate that there exists great potential in employing hybrid sampling algorithm with suitably chosen α to reduce the number of required samples to achieve a given level of estimation error, i.e. the sampling cost in a practical setting. To the best of our knowledge, this is the first work to propose a theoretical framework for hybrid sampling algorithms with detailed analysis on their properties, relationship with graph spectral properties, and guideline on how to tune the knob (jump probability) under various network scenarios.

II. Preliminaries

We provide the mathematical background for graph sampling, including a performance metric of interest – the asymptotic variance of an estimator, which will be used throughout the paper. We also review a crawling-based sampling method using the Metropolis-Hastings random walk that is widely used for unbiased graph sampling in the literature, and will also be a basis for our hybrid sampling model.

A. Mathematical Background

Let $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ be a connected, undirected, non-bipartite graph with a set of finite nodes $\mathcal{N} = \{1, 2, \ldots, n\}$ and a set of edges $\mathcal{E}$. We assume that the graph $\mathcal{G}$ has no self-loops and no multiple edges connecting i and j. Let $d(i) \triangleq |\{j \in \mathcal{N} : (i,j) \in \mathcal{E}\}|$ be the degree of node i. Then unbiased graph sampling is to obtain uniform nodal samples to consistently and unbiasedly estimate nodal or topological properties of the graph $\mathcal{G}$. To be precise, for any given desired function $f : \mathcal{N} \to \mathbb{R}$, we want to estimate $\mathbb{E}_{\mathbf{u}}(f) \triangleq \sum_{i\in\mathcal{N}} f(i)/n$, where $\mathbf{u} \triangleq [u(1), u(2), \ldots, u(n)] = [1/n, 1/n, \ldots, 1/n]$ is the uniform distribution over $\mathcal{N}$.

We next briefly explain a basic Markov chain theory that is a mathematical basis for crawling-based (or random walk-based) sampling methods and also for our hybrid sampling model. For graph sampling, we consider a discrete-time Markov chain $\{X_t \in \mathcal{N}, t = 0, 1, 2, \ldots\}$ on $\mathcal{G}$ with transition matrix $\mathbf{P} \triangleq \{P(i,j)\}_{i,j\in\mathcal{N}}$, in which $P(i,j)$ denotes the probability of transition from node i to j. We assume that $\{X_t\}$ is irreducible and aperiodic with its unique stationary distribution $\boldsymbol{\pi} \triangleq [\pi(1), \pi(2), \ldots, \pi(n)]$. Then, consider an estimator based on the Markov chain $\{X_t\}$ as

$$\hat{\mu}_t(f) \triangleq \frac{1}{t}\sum_{s=1}^{t} f(X_s). \quad (1)$$

Since the chain is ergodic (i.e., irreducible and aperiodic), it follows that for any given function f with $\mathbb{E}_{\boldsymbol{\pi}}(|f|) < \infty$,

$$\lim_{t\to\infty} \hat{\mu}_t(f) = \mathbb{E}_{\boldsymbol{\pi}}(f) \triangleq \sum_{i\in\mathcal{N}} f(i)\pi(i) \quad \text{with probability 1} \quad (2)$$

for any initial distribution of the chain [15], [16]. Thus, the estimator $\hat{\mu}_t(f)$ provides asymptotically unbiased estimate of $\mathbb{E}_{\mathbf{u}}(f)$ (ensuring unbiased graph sampling), if $\boldsymbol{\pi} = \mathbf{u}$. In addition, we can measure how good such an estimator is in terms of its asymptotic variance which is given by

$$\begin{aligned} v(f, \mathbf{P}, \boldsymbol{\pi}) &\triangleq \lim_{t\to\infty} t \cdot \mathrm{Var}(\hat{\mu}_t(f)) \\ &= \lim_{t\to\infty} \frac{1}{t}\mathbb{E}\left\{\left[\sum_{s=1}^{t}(f(X_s) - \mathbb{E}_{\boldsymbol{\pi}}(f))\right]^2\right\} \end{aligned} \quad (3)$$

for any function f with $\mathbb{E}_{\boldsymbol{\pi}}(f^2) < \infty$. In particular, it is well known that $\sqrt{t}[\hat{\mu}_t(f) - \mathbb{E}_{\boldsymbol{\pi}}(f)]$ converges in distribution to a Gaussian random variable with zero mean and variance $v(f, \mathbf{P}, \boldsymbol{\pi})$ [17], [15], [16]. Thus, the asymptotic variance $v(f, \mathbf{P}, \boldsymbol{\pi})$ indicates approximately how many samples should be collected in order to achieve a given desired level of accuracy when using the estimator $\hat{\mu}_t(f)$. Throughout this paper, we consider the asymptotic variance $v(f, \mathbf{P}, \boldsymbol{\pi})$ of the estimator $\hat{\mu}_t(f)$ as our primary performance metric.

B. Metropolis-Hastings Random Walk

Random walk-based sampling methods have been widely used for unbiased graph sampling, and generally work as follows. A sampling agent, currently staying at one node, first explores the current node to obtain its nodal sample, and then decides to move (crawl) to one of its neighboring nodes or to stay in the current node (for the next sample), all based on local information. This is complementary to random jump sampling – independent uniform node sampling – if possible, and is sometimes the only viable method for networks where no global topological information is available for the sampling agent. One of the popular random walk-based sampling methods, among others, is to employ the Metropolis-Hasting random walk (MHRW) [18], [8], [4], [5], [6], [9], leading to a Markov chain $\{X_t\}$ whose stationary distribution is uniform over $\mathcal{N}$. The resulting estimator $\hat{\mu}_t(f)$ is thus consistent or asymptotically unbiased in estimating $\mathbb{E}_{\mathbf{u}}(f)$ (in the sense of (2) with $\boldsymbol{\pi}$ replaced by $\mathbf{u}$).

The MHRW works as follows. A random walk agent, currently residing at node i, chooses one of its neighbors (say, node j) uniformly at random with probability $1/d(i)$. This 'proposed move' is then accepted with probability

$$a(i,j) = \min\left\{1, \frac{\pi(j)/d(j)}{\pi(i)/d(i)}\right\}, \tag{4}$$

and rejected with probability $1 - a(i,j)$, in which case the agent stays in the current node i. Thus, for unbiased sampling with uniform target distribution, i.e., $\pi(i) = 1/n$ for all $i \in \mathcal{N}$, the transition matrix $\mathbf{P}$ of this random walk agent is given by

$$P(i,j) = \frac{1}{d(i)}a(i,j) = \min\left\{\frac{1}{d(i)}, \frac{1}{d(j)}\right\} \tag{5}$$

for $(i,j) \in \mathcal{E}$, and $P(i,j) = 0$ if $(i,j) \notin \mathcal{E}$, where $P(i,i) = 1 - \sum_{j \neq i} P(i,j)$. Note that the resulting $\mathbf{P}$ is reversible with respect to $\boldsymbol{\pi} = \mathbf{u}$ since $P(i,j) = P(j,i)$. It is also easy to see that the Markov chain with $\mathbf{P}$ is irreducible and aperiodic (because $\mathcal{G}$ is connected, undirected, and non-bipartite).

III. Hybrid Sampling Model

In this section, we propose a general framework for hybrid graph sampling that allows for the introduction of random jump sampling into the crawling-based sampling with MHRW, and then discuss its several properties.

A. Model Description

Large complex (social) networks such as Facebook, QQ, Myspace, Twitter, assign each user/node a unique user/node ID [2], [11], which provides API to collect uniform samples by generating uniformly random user/node IDs. At each random 'guess' in the ID pool, if the ID is valid, then the random jump succeeds; otherwise, the request is rejected and the current position remains unchanged. This method produces independent and uniform samples by eliminating the correlations between samples regardless of the distribution of valid IDs in user ID pool [6]. However, although this sampling method overcomes the problem of correlations and is easy to implement, crawling is still an indispensable method for sampling because the ID space is often sparsely allocated, resulting in high cost of obtaining each valid sample under random jump sampling.

For fair comparison and optimal exploitation toward hybrid sampling, we set out to properly capture and take into account costs associated with each of the two sampling approaches. One is to crawl the graph to obtain guaranteed but correlated samples, while the other way is to take the risk of being rejected and staying in the original node for random jump, which may lead to 'high-quality' samples if successful, or the same sample as before (100% correlated) if not. With this in mind, our goal is to formally establish a theoretical framework for hybrid samplers, in order to provide a guideline to strike the right balance between the two for maximal sampling efficiency (smallest asymptotic variance) under the same sampling cost.

Let Ω be the set of all possible IDs (entire ID space), including all valid and invalid IDs, and $\mathcal{N} \subset \Omega$ be the set of valid IDs only. At each step, with probability $\alpha \in [0,1]$, a sampling agent attempts to draw an *i.i.d.* sample from Ω, and with probability $1 - \alpha$, the sampling agent continues to crawl the graph according to a Markov chain $\mathbf{P}$. When the agent decides to jump in each step, if the random guess is a valid node index in $\mathcal{N}$ which occurs with probability $\beta = |\mathcal{N}|/|\Omega| \in (0,1]$, the sampling agent teleports to the valid node in one step. If it is not valid (with probability $1 - \beta$), the agent resides in the current position. See Figure 1 for illustration. In this setup, the hybrid sampler follows a Markov chain with its transition matrix given by

$$\mathbf{P}_\alpha = (1-\alpha)\mathbf{P} + \alpha(1-\beta)\mathbf{I} + \frac{\alpha\beta}{n}\mathbf{1}\mathbf{1}^T, \tag{6}$$

where $\mathbf{1} = [1,1,\ldots,1]^T$ is the n-dimensional column vector, and $\mathbf{I}$ is $n \times n$ identity matrix. In this model, β is a system parameter governed by the ID space and the total number of assigned IDs, and can be quickly estimated by measuring the ratio of the number of successful trials to the number of total guesses. On the other hand, the suitable choice of parameter α is our key design issue. By carefully tuning this control knob, the hybrid sampler in (6) can range from a pure crawling based one with $\mathbf{P}_0 = \mathbf{P}$ up to a pure random jump sampling with $\mathbf{P}_1 = (1-\beta)\mathbf{I} + \beta\mathbf{1}\mathbf{1}^T/n$. Throughout the paper, we assume that our base chain $\mathbf{P}_0 = \mathbf{P}$ is MHRW with uniform distribution $\mathbf{u}$. In this case, $\mathbf{P}$ is symmetric, thus $\mathbf{P}_\alpha$ is also symmetric, implying that the stationary distribution of the hybrid sampler $\mathbf{P}_\alpha$ remains untouched and uniform (unbiased) under any choice of $\alpha \in [0,1]$. In fact, any symmetric (or doubly stochastic) matrix $\mathbf{P}$ would serve the purpose.

The well-known PageRank [19], which is an algorithm used by the Google web search engine to rank websites in their search engine results, can also be considered as a combination of the two kernels (crawling and random jump). However, it has intrinsic differences from our model. PageRank or its variation (eg. Personalized PageRank) [19] targets at computing the resulting stationary distribution of the combination of the two kernels, which originally have different stationary

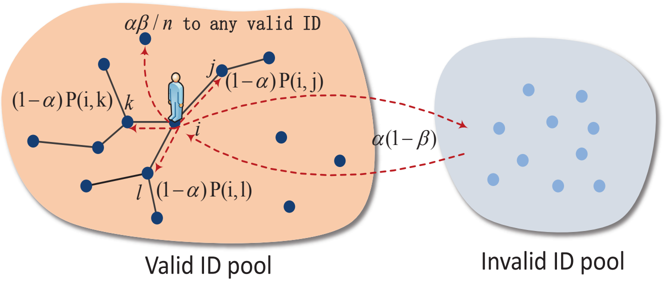

Fig. 1. Dark blue points represent valid IDs forming a graph $\mathcal{G}$, while light blues indicate invalid IDs. A sampling agent moves from node i to one of its neighbors, say j, with probability $(1-\alpha)P(i,j)$. In addition, an attempt of jumping to any node over $\mathcal{G}$ by randomly accessing to one of valid/invalid IDs, which happens with α, will succeed with probability β or fail otherwise in which case the agent resides in its current position.

distributions. While in our case, both the two kernels give uniform distribution, and thus the resulting one is also uniform for the purpose of unbiased estimation. Then we instead focus on the second-order behavior of the chain, such as the asymptotic variance of the estimator as in (1) for any nodal property.

B. Spectral Properties of Hybrid Sampler

While the stationary distribution of the transition matrix $\mathbf{P}_\alpha$ is invariant with respect to α, its corresponding asymptotic variance $\upsilon(f,\mathbf{P}_\alpha,\mathbf{u})$ depends highly on α. To understand this, we here investigate how the entire spectrum of $\mathbf{P}_\alpha$ changes when α varies.

We assume $\mathbf{P}$ is the transition matrix of an irreducible, aperiodic Markov chain by the MHRW, from Perron-Frobenius Theorem [17], its n eigenvalues are given by $1 = \lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n > -1$. For functions $f, g : \mathcal{N} \to \mathbb{R}$, define their scalar product with respect to $\boldsymbol{\pi} = [\pi(1), \ldots, \pi(n)]$ as $\langle f, g\rangle_\pi \triangleq \sum_{i=1}^n f(i)g(i)\pi(i)$. Since $\mathbf{P}$ is symmetric, all its eigenvalues are real and all its n eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ are mutually orthogonal, where $\mathbf{v}_i \in \mathbb{R}^n$ is the eigenvector associated with the eigenvalue λ_i [20]. After a proper normalization, all these vectors can be made orthonormal in the sense of $\langle \mathbf{v}_i, \mathbf{v}_j\rangle_\pi = \delta_{ij}$ for $i,j = 1,2,\ldots,n$, where $\delta_{ij} = 1$ if $i = j$ and zero otherwise. Particularly, $\mathbf{v}_1 = \mathbf{1}$. We then have the following.

Theorem 1: For each given $\alpha \in [0,1]$, let $\{\lambda_i(\alpha), i = 1,2,\ldots,n\}$ and $\{\mathbf{v}_i(\alpha), i = 1,2,\ldots,n\}$ be the set of n eigenvalues and the corresponding eigenvectors of $\mathbf{P}_\alpha$. Then, we have $\lambda_1(\alpha) = 1$ and

$$\lambda_i(\alpha) = (1-\alpha)\lambda_i + \alpha(1-\beta), \quad i = 2,3,\ldots,n, \tag{7}$$

$$\mathbf{v}_i(\alpha) = \mathbf{v}_i, \quad i = 1,2,\ldots,n. \tag{8}$$

Proof: See Appendix A. ■

Theorem 1 tells us that the transition matrix $\mathbf{P}_\alpha$ (for hybrid sampling) has the same set of orthonormal eigenvectors $\{\mathbf{v}_i\}$ as the 'baseline' transition matrix $\mathbf{P}$ (for pure crawling-based sampling with MHRW), while each of its non-principal eigenvalues is now a weighted average of λ_i and $1-\beta$. This relationship will greatly simplify our analysis on the efficiency of hybrid sampling.

IV. Asymptotic Variance of Estimator

Given the hybrid sampling model, in this section, we first show that the asymptotic variance for any nodal property estimator of interest under $\mathbf{P}_\alpha$ is convex in α. We then discuss two extreme cases under which either pure crawling or pure random jump method becomes optimal. We also assess the precision of our estimator using asymptotic variance when a finite number of samples are used in any practical setting.

A. Properties of Asymptotic Variance of Estimator

From the theory of ergodic and reversible Markov chains, e.g., [17, pp.232–235], the asymptotic variance $\upsilon(f,\mathbf{P}_\alpha,\mathbf{u})$ can be expressed in terms of the spectrum of $\mathbf{P}$ as

$$\begin{aligned}\upsilon(f,\mathbf{P}_\alpha,\mathbf{u}) &= \sum_{i=2}^{n} \frac{1+\lambda_i(\alpha)}{1-\lambda_i(\alpha)} |\langle f, \mathbf{v}_i(\alpha)\rangle_{\mathbf{u}}|^2 \\ &= 2\sum_{i=2}^{n} \frac{|\langle f,\mathbf{v}_i\rangle_{\mathbf{u}}|^2}{1-(1-\alpha)\lambda_i-\alpha(1-\beta)} - \sum_{i=2}^{n} |\langle f,\mathbf{v}_i\rangle_{\mathbf{u}}|^2, \end{aligned} \tag{9}$$

where the second equality is from Theorem 1. We define a random variable Λ as

$$P\{\Lambda = \lambda_i\} = |\langle f, \mathbf{v}_i\rangle_{\mathbf{u}}^2|/Z, \quad i = 2,3,\cdots,n, \tag{10}$$

where

$$Z = \sum_{i=2}^{n} |\langle f, \mathbf{v}_i\rangle_{\mathbf{u}}|^2 > 0 \tag{11}$$

is a normalizing constant. The asymptotic variance $\upsilon(f,\mathbf{P}_\alpha,\mathbf{u})$ then becomes

$$\upsilon(f,\mathbf{P}_\alpha,\mathbf{u}) = Z \cdot [2\mathbb{E}_\Lambda\{h(\Lambda,\alpha,\beta)\} - 1], \tag{12}$$

in which $h(\lambda_i,\alpha,\beta) \triangleq [(1-\lambda_i)(1-\alpha)+\alpha\beta]^{-1}$. Now, we have the following.

Theorem 2: $\upsilon(f,\mathbf{P}_\alpha,\mathbf{u})$ is convex in $\alpha \in [0,1]$.

Proof: For notational convenience, we write the asymptotic variance $\upsilon(f,\mathbf{P}_\alpha,\mathbf{u})$ as $\upsilon(\alpha)$ for any given f. Differentiating $\upsilon(\alpha)$ in (12) twice with respect to α gives

$$\upsilon''(\alpha) = 2Z \cdot \mathbb{E}_\Lambda\{h''(\Lambda,\alpha,\beta)\}.$$

Direct calculation yields

$$h''(\lambda_i,\alpha,\beta) = \frac{2(\lambda_i - 1 + \beta)^2}{[(1-\lambda_i)(1-\alpha)+\alpha\beta]^3} \geq 0,$$

because the denominator $(1-\lambda_i)(1-\alpha)+\alpha\beta$ is always positive for $\alpha \in [0,1]$, which is from $\lambda_i \leq \lambda_2 < 1$ and $\beta > 0$. This completes the proof. ■

This convex property of $\upsilon(f,\mathbf{P}_\alpha,\mathbf{u})$ provides us a convenient guideline to search for the following optimal α^*

$$\alpha^* = \alpha^*(f) \triangleq \operatorname*{arg\,min}_{\alpha\in[0,1]} \; \upsilon(f,\mathbf{P}_\alpha,\mathbf{u}) = \operatorname*{arg\,min}_{\alpha\in[0,1]} \; \upsilon(\alpha)$$

by applying any standard convex optimization technique for any given β. Here, the notation $\alpha^*(f)$ is to clearly indicate the dependency of the optimal α on the property f to be estimated. We will simply write α^* instead, whenever no confusion arises.

B. Discussion on Two Extreme Cases

For a large complex network, it is mostly impossible to obtain the set of all its eigenvalues and eigenvectors of the transition matrix $\mathbf{P}_\alpha$. We are thus faced with an optimization problem of minimizing the objective function as given in (9) that, albeit looks simple, defies any closed form expression. Estimating all its eigenvalues and eigenvectors is certainly beyond anyone's reach. Nonetheless, here we are able to find out simple conditions for two special cases under which either the pure-crawling ($\alpha = 0$) or always random jumping ($\alpha = 1$) turns out to be the optimal strategy. The first special case is given next.

Proposition 1: If $1 - \lambda_2 \geq \beta$, then $\alpha^*(f) = 0$ for any f.

Proof: First, fix f. By differentiating $\upsilon(\cdot)$ in (12) with respect to α, we have

$$\begin{aligned}\upsilon'(\alpha) &= 2Z \cdot \mathbb{E}_\Lambda\{h'(\Lambda, \alpha, \beta)\} \\ &= 2Z \cdot \mathbb{E}_\Lambda \left\{ \frac{1 - \beta - \Lambda}{[(1-\Lambda)(1-\alpha) + \alpha\beta]^2} \right\} \end{aligned} \tag{13}$$

Since $1 - \lambda_i \geq 1 - \lambda_2$ for all $i = 2, 3, \ldots, n$, under the stated assumption, we have $1 - \beta - \lambda_i \geq 0$ for all i. Thus, $\upsilon'(0) = 2Z \cdot \mathbb{E}_\Lambda\{\frac{1-\beta-\Lambda}{(1-\Lambda)^2}\} \geq 0$. Since $\upsilon(\alpha)$ is convex in $\alpha \in [0,1]$ from Theorem 2, $\upsilon'(\alpha)$ is increasing (non-decreasing), thus $\upsilon'(\alpha) \geq \upsilon'(0) \geq 0$ for all $\alpha \in [0,1]$, implying that $\upsilon(\alpha)$ is increasing in α. Therefore, the optimal α^* that minimizes $\upsilon(\alpha)$ is $\alpha^* = 0$. This completes the proof. ■

Proposition 1 says that when $1 - \lambda_2 \geq \beta$, always crawling the graph without random jump is optimal for any arbitrary nodal property f to be estimated. Note that this is a sufficient condition. The condition $1 - \lambda_2 \geq \beta$ typically holds when β is very small or λ_2 is away from 1 in that the gap $1 - \lambda_2$ is larger than the density of valid IDs. Equivalently, the condition can be written as $\tau_2 = 1/(1-\lambda_2) \leq 1/\beta$, where τ_2 is called the relaxation time [21], and $1/\beta$ is the expected number of steps until to obtain independent samples under pure jump strategy. In this case, always crawling the graph is the optimal strategy, since there's no benefit of attempting to jump with the hope of getting independent samples. In other words, any attempt to jump with non-zero probability α produces 'worse' samples with higher correlations overall, than the samples obtained by pure-crawling all the time.

We however here point out that this condition is very stringent, ensuring optimality of the crawling method for *any* f. Many large complex networks are known to be 'slow-mixing' [22], implying that λ_2 would be very close to one.[2] Suppose that the opposite condition holds, i.e., $1 - \lambda_2 < \beta$, which we believe to be the case in most interesting scenarios. Note that, even in this case, it is still possible that $\alpha^*(f) = 0$ for certain choice of f. This means $(1-\beta)/\lambda_2 < 1$. Thus we can always find positive number $\alpha \in (0,1)$ such that $(1-\beta)/\lambda_2 < \alpha < 1$, yielding $\lambda_2(\alpha) = (1-\alpha)\lambda_2 + \alpha(1-\beta) < \lambda_2$. That is to say, we can also increase the spectral gap (or decrease the relaxation time) by choosing such α for faster convergence of the resulting chain $\mathbf{P}_\alpha$ than the original $\mathbf{P}$.

We now discuss the other extreme case in which always attempting to jump is the optimal strategy.

Proposition 2: For a given f, $\alpha^*(f) = 1$ if and only if $\mathbb{E}\{\Lambda\} \geq 1 - \beta$, where Λ is defined in (10) and (11).

Proof: Again, using (13), under the stated assumption, we have $\upsilon'(1) = 2Z \cdot \mathbb{E}_\Lambda\{\frac{1-\beta-\Lambda}{\beta^2}\} \leq 0$. Then, following the similar steps in the proof of Proposition 1 and from Theorem 2, we know that $\upsilon(\alpha)$ is decreasing in $\alpha \in [0,1]$. Thus, the optimal α^* that minimizes $\upsilon(\alpha)$ is $\alpha^*(f) = 1$. Conversely, if $\alpha^*(f) = 1$, the function $\upsilon(\alpha)$ must be decreasing since it is convex, and this necessitates $\upsilon'(1) \leq 0$, which then becomes the stated assumption. ■

Proposition 2 tells us that if $\mathbb{E}\{\Lambda\}$ is larger than the density of invalid IDs, then attempting to random jump all the time is the optimal strategy. To better understand the term $\mathbb{E}\{\Lambda\}$, consider a general reversible Markov chain $\mathbf{P}$ with uniform stationary distribution $\boldsymbol{\pi} = \mathbf{u}$. For a given f, let Λ be the random variable as before, and define the autocorrelation coefficient of a stationary sequence $f(X_k)$, $k = 0, 1, \ldots$, as

$$\gamma(k) \triangleq \frac{\mathbb{E}_\mathbf{u}\{(f(X_0) - \mathbb{E}_\mathbf{u}(f))(f(X_k) - \mathbb{E}_\mathbf{u}(f))\}}{\mathrm{Var}_\mathbf{u}(f)}. \tag{14}$$

Then, we have the following.

Lemma 1: For $k = 1, 2, \ldots$, we have $\mathbb{E}\{\Lambda^k\} = \gamma(k)$.

Proof: See Appendix B. ■

From Lemma 1, we know that $\mathbb{E}\{\Lambda\}$ is simply the lag-1 auto-correlation coefficient of the sequence $f(X_n)$, where X_n is the Markov chain with transition matrix $\mathbf{P}$. Then Proposition 2 means if the correlation coefficient of the Markov chain is strong enough to be larger than $1 - \beta$, then the optimal strategy is to always attempt to jump with $\alpha = 1$. In this case, crawling with non-zero probability would render the obtained samples too heavily correlated over time, thus becomes inferior than always attempting to jump (which will be successful only with probability β). In practice, this condition is very easy to check. Let $(X_1, X_2, \cdots, X_t)$ be a stationary sequence of t sampled vertices, then $\hat{r}(k) = \frac{1}{t-k}\sum_{i=0}^{t-k-1} f(X_i)f(X_{i+k})$ is an asymptotically unbiased estimator [24] of the autocorrelation function $r(k) = \mathbb{E}_\mathbf{u}\{f(X_i)f(X_{i+k})\}$ for $k = 0, 1, \cdots$, thus one would simply need to find $\hat{r}(1)/\hat{r}(0)$ for a consistent estimator of $\gamma(1) = \mathbb{E}\{\Lambda\}$.

C. Estimation with Finite Samples

Although the asymptotic variance defined in (3) is of important analytical use, researchers would have to use the mean squared error based on *finite* number of samples to

[2] The condition says the spectral gap of the MHRW on the graph is $O(1)$, independent of the size n of the graph, suggesting that the graph is an expander graph [23].

approximate the asymptotic variance. Here, we will discuss in detail the impact of such sample space constraint.

Let t be the sample size. Assuming that the initial distribution of the Markov chain $\mathbf{P}$ is drawn from its uniform stationary distribution, the re-scaled mean squared error (or variance) of the estimator $\hat{\mu}_t(f)$ in (1) under the chain $\mathbf{P}$ becomes

$$t\cdot \mathrm{Var}(\hat{\mu}_t(f)) = \mathrm{Var}_{\mathbf{u}}(f)\cdot\left[1+2\sum_{k=1}^{t-1}\left(1-\frac{k}{t}\right)\gamma(k)\right], \tag{15}$$

where $\gamma(k)$ is defined in (14). On the other hand, the asymptotic variance defined in (3) can be written as

$$\upsilon(f,\mathbf{P},\mathbf{u}) = \mathrm{Var}_{\mathbf{u}}(f)\cdot\left[1+2\sum_{k=1}^{\infty}\gamma(k)\right]. \tag{16}$$

Our next result shows that the error between the asymptotic variance and the (re-scaled) mean squared error based on t samples is bounded by $O(t^{-1})$.

Proposition 3: For any f and any reversible chain $\mathbf{P}$ with uniform stationary distribution $\mathbf{u}$,

$$|\upsilon(f,\mathbf{P},\mathbf{u}) - t\cdot\mathrm{Var}(\hat{\mu}_t(f))| < \frac{4\mathrm{Var}_{\mathbf{u}}(f)}{t(1-\lambda_2)^2}. \tag{17}$$

Proof: Using (15) and (16), observe that, from Lemma 1

$$\begin{aligned}\frac{\upsilon(f,\mathbf{P},\mathbf{u}) - t\cdot\mathrm{Var}(\hat{\mu}_t(f))}{2\mathrm{Var}_{\mathbf{u}}(f)} &= \sum_{k=1}^{t-1}\frac{k}{t}\gamma(k) + \sum_{k=t}^{\infty}\gamma(k)\\ = \sum_{k=1}^{t-1}\frac{k}{t}\mathbb{E}\{\Lambda^k\} + \sum_{k=t}^{\infty}\mathbb{E}\{\Lambda^k\} &= \mathbb{E}_\Lambda\left\{\frac{\Lambda(1-\Lambda^t)}{t(1-\Lambda)^2}\right\}.\end{aligned}$$

By taking absolute values on both sides, we get

$$\begin{aligned}\frac{|\upsilon(f,\mathbf{P},\mathbf{u}) - t\cdot\mathrm{Var}(\hat{\mu}_t(f))|}{2\mathrm{Var}_{\mathbf{u}}(f)} &\le \mathbb{E}_\Lambda\left\{\left|\frac{\Lambda(1-\Lambda^t)}{t(1-\Lambda)^2}\right|\right\}\\ \le \frac{\mathbb{E}_\Lambda\left\{|\Lambda| + |\Lambda^{t+1}|\right\}}{t(1-\lambda_2)^2} &< \frac{2}{t(1-\lambda_2)^2},\end{aligned}$$

where the second inequality follows from $1-\Lambda \ge 1-\lambda_2$, and the last inequality is from $|\Lambda| < 1$. This completes the proof. ■

Proposition 3 says the error between the asymptotic variance of an estimator and its finite-sample counterpart becomes negligible as the number of samples t increases. We thus expect that the convexity of the asymptotic variance and its relevant spectral properties carry over into the (re-scaled) mean square error $t\cdot\mathrm{Var}(\hat{\mu}_t(f))$ and also the original mean squared error $\mathrm{Var}(\hat{\mu}_t(f))$ for a fixed sample size t, when t is large which is typically the case in real large graphs.

V. Numerical Results

In this section, we provide numerical results to support our theoretical findings and compare the performance of the sampling algorithm under different parameter settings.

A. Data Set

We performed our experiments on real world dataset from Google web graph. [25]. Google web graph is composed of web pages with directed hyperlinks between them. The 875713 nodes in the graph represent the sampled web pages and the 5105039 edges represent the hyperlinks. The largest connected component (LCC) contains 855802 nodes. In the following simulations, we use the undirected version of the graph and consider only its LCC for the fair comparison between hybrid sampling, MHRW and random jump [7], [12], [13]. Each point in the following figures is the average of 10^4 independent tests.

B. Performance Metrics and Estimation Error

We concentrate on the degree distribution $\mathbf{P}\{D_{\mathcal{G}} = d\}$ estimates and global clustering coefficient given by $C = \frac{1}{|\mathcal{N}|}\sum_{i=1}^{|\mathcal{N}|} c_i$ where $c_i = \triangle(i)/\binom{d_i}{2}$ for $d_i \ge 2$, otherwise, $c_i = 0$. Here, $\triangle(i) = |(j,k) \in \mathcal{E} : (i,j) \in \mathcal{E} \text{ and } (i,k) \in \mathcal{E}|$ is the number of triangles that contain vertex i, and $\binom{d_i}{2} = d_i(d_i-1)/2$ is the number of possible triangles composed of vertex i and its neighbors. For estimation of $\mathbf{P}\{D_{\mathcal{G}} = d\}$, we choose $f(i) = \mathbf{1}_{\{d_i = d\}}$ as the corresponding estimator, while we set $f(i) = c_i$ as the estimator for the clustering coefficient.

To measure the estimation accuracy, we use the normalized root mean squared error (NRMSE) [26], [12] defined by

$$\mathrm{NRMSE}(\hat{x}_j(t)) = \sqrt{\mathbb{E}\left[(\hat{x}_j(t) - x_j)^2\right]}/x_j,\ j = 1, 2, \cdots,$$

which measures the relative error of the estimator $\hat{x}_j$ with respect to its 'ground truth' value x_j when the number of samples is t. In our unbiased sampling case, $x_j = \lim_{t\to\infty}\hat{x}_j(t)$. In the literature, this metric is generally preferred over the mean squared error, as the NRMSE enables us to compare the errors for different functions on a common scale.

C. Simulation Results

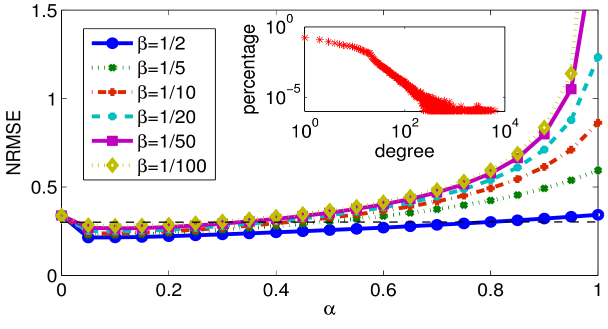

Fig. 2. NRMSE vs. α in estimating degree distribution (inset: degree distribution).

In Figure 2, for different values of β, we present the NRMSE curves that vary as a function of jump probability α for the estimation of degree distribution. The degree distribution of the graph is also given in the inset figure. Here each point is the average NRMSE taken over all degrees with 8.5×10^6 samples. It shows that the curves (linear combination of convex functions) are convex in α as predicted

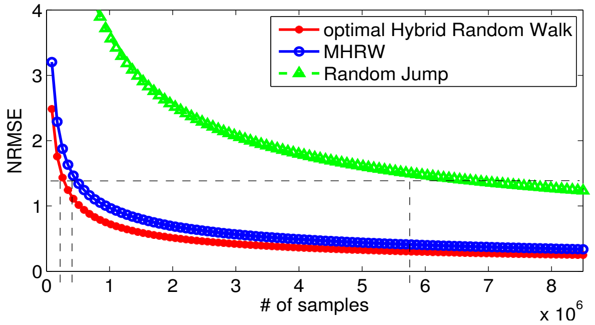

Fig. 3. NRMSE vs. t under $\beta = 1/20$ in estimating degree distribution.

in Theorem 2. the minimum NRMSE values are achieved at around $\alpha = 0.05$ for the six curves, and the corresponding improvements over MHRW and random jump is obvious. In Figure 3, the relation between the average NRMSE and the sample size t is investigated under $\beta = 1/20$. We can see that the optimal hybrid sampling reduces the sample size to 50% that of MHRW and 5% that of random jump to achieve the same overall NRMSE, thus the cost saving is clearly evident.

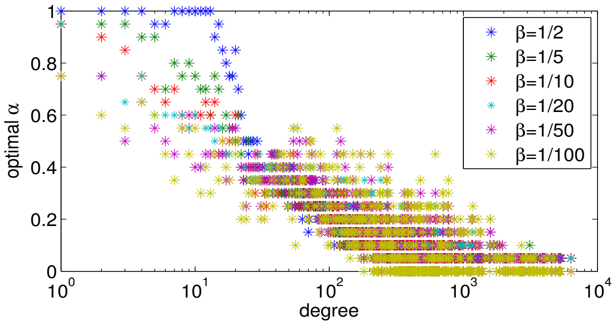

Fig. 4. Optimal α vs. degree in estimating degree distribution.

Figure 4 shows the optimal α in estimating $P\{D_{\mathcal{G}} = d\}$ for each degree d. For relatively small degrees (high percentage over the whole graph), the optimal jump probability $\alpha^* = 1$, namely, employing random jump will be most beneficial, while for relatively large degree (small percentage), MHRW will serve as the optimal sampling strategy. In between is the transition phase. Here we give an explanation for the phenomenon using Figure 5. For the set of nodes which occupy a small portion over the graph, the frequency that a random walker visits the set is relatively small ($f(X_t) = 1$ and represented by red circles if the nodes in this set are hit as in Figure 5(a)), then the samples do not have much interplay under MHRW. While for the relatively larger set (Figure 5(b)), because of the repetition of the same samples from MHRW as well as the assortativity between pairs of linked nodes, the correlation of consecutive steps is large. This leads to poor performance employing just pure random walk. Thus the optimal α is inclined to be 1.

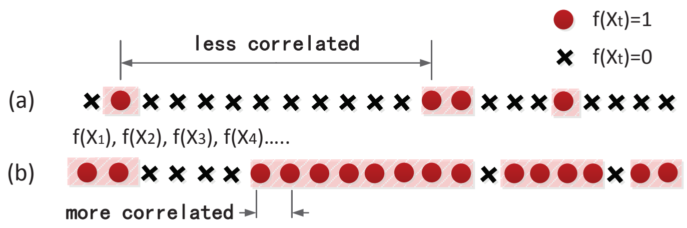

Fig. 5. Illustration for a Markov chain $\{X_t\}$ on a graph with the red circle as $f(X_t) = 1$ and the black cross as $f(X_t) = 0$: (1) When the percentage of the set of nodes is relatively small, (2) When the percentage of the set of nodes is relatively large.

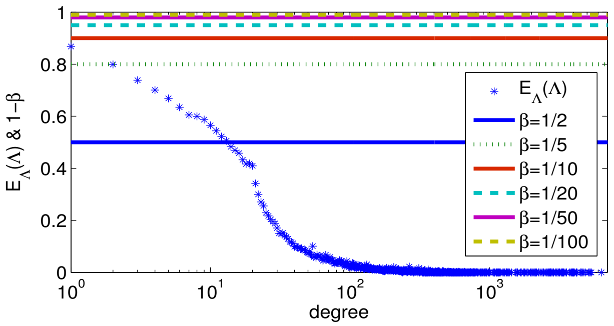

Fig. 6. $E_\Lambda\{\Lambda\}$ and $1 - \beta$ vs. degree.

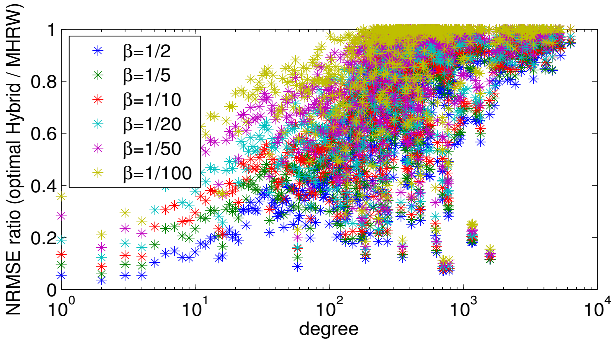

Fig. 7. NRMSE ratio of optimal hybrid sampling to MHRW.

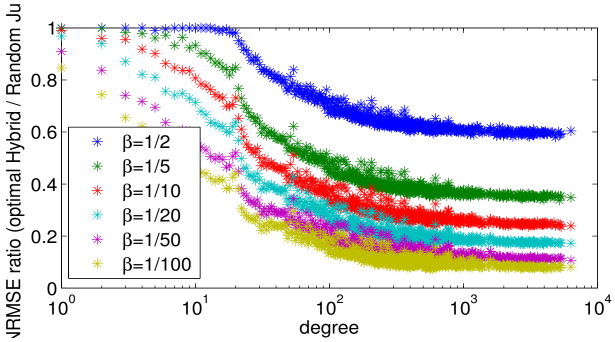

Fig. 8. NRMSE ratio of optimal hybrid sampling to random jump.

In order to compare the simulation results and our presented condition in Proposition 2, we estimate $E_\Lambda\{\Lambda\}$ using the estimator for $\gamma(1)$ in (14) and compare it with $1 - \beta$ (represented as the lines) in Figure 6. The result is in accordance with what is observed in Figure 4. For instance, for the degrees smaller than 15, $E_\Lambda\{\Lambda\} > 1 - \beta$ for $\beta = 1/2$, in which case the best sampling strategy is supposed to be random jump. Moreover, the corresponding NRMSE ratios of optimal hybrid sampling to MHRW and to random jump are depicted in Figure 7 and 8 respectively.

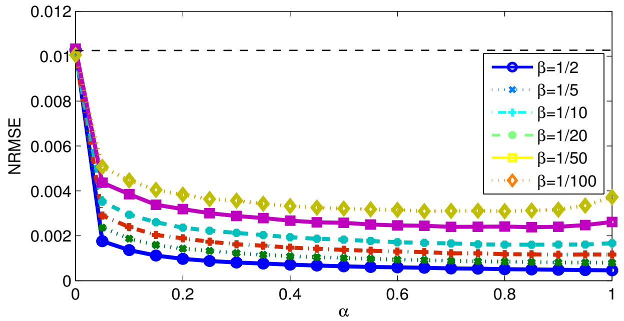

Fig. 9. NRMSE vs. α in estimating clustering coefficient.

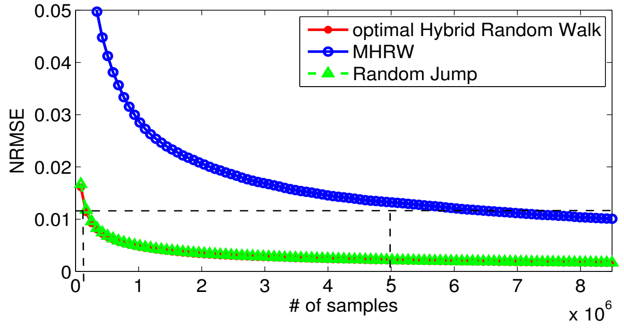

Fig. 10. NRMSE vs. t under $\beta = 1/20$ in estimating clustering coefficient.

We then set our target function as the global clustering coefficient in Figure 9 and 10. From Figure 9, we see a sharp reduce in NRMSE when a small probability random jump is introduced to MHRW, and then its change with α is relatively mild. This trend says the introduction of random jump can bring out decent benefit and when α changes over a wide range, the benefit doesn't actually have much fluctuation. Thus for this case, the estimation of the optimal α is not necessarily in strict accuracy. Again, the results are in good agreement with our theoretical finding that the NRMSE is a convex function of α. We also estimate $E_{\Lambda}\{\Lambda\} = 0.838$ in this case, and expect that for $\beta = 1/2$ and $1/5$, the NRMSE curves are monotonically decreasing with the minimum value taken at $\alpha = 1$ as predicted in Proposition 2. This guess is justified in Figure 9. In Figure 10, we also plot the NRMSE varying with the increase of t for $\beta = 1/20$ in estimating the clustering coefficient. The optimal sampling strategy (random jump) reads about 96% cost saving compared to MHRW to achieve 0.013 in NRMSE. Thus, we can conclude that hybrid sampling exhibits great potential in improving the estimation error of both degree distribution and clustering coefficient.

We also repeat the same set of simulations over Road-PA graph and Wiki-Vote graph, and we observe similar trends. Due to the space limit, we refer to our technical report [27] for more details.

VI. Conclusion

We have proposed a general framework for hybrid graph sampling methods, which enables us to correctly evaluate the potential benefits of crawling-based sampling with random jump. In particular, we analyzed the entire spectrum of a generic hybrid sampling model, and found out the convex property of the asymptotic variance of its resulting estimator. We also obtained the conditions under which pure crawling or random jump serves as the best sampling strategy (to minimize the asymptotic variance), so that before collecting samples, a preprocessing procedure can lead to a quick decision for optimal sampling performance under the extreme cases. While our theoretical analysis was mostly done for the asymptotic variance, we also demonstrated a small error between the asymptotic variance and the (re-scaled) variance of finite samples, which shifts our attention from the theoretical analysis into its practical use. In addition, we provided simulation results over real-world network dataset to reveal the great potential of hybrid graph sampling in improving the estimation accuracy, and also support our analytical findings. We expect that our results would serve as guidelines for the design of more efficient hybrid sampling methods.

References

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

[2] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in facebook: a case study of unbiased sampling of osns," in *IEEE INFOCOM*, Mar. 2010.

[3] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *ACM SIGKDD*, Aug. 2006.

[4] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger, "On unbiased sampling for unstructured peer-to-peer networks," *IEEE/ACM Trans. on Networking*, vol. 17, no. 2, pp. 377–390, 2009.

[5] A. H. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach, "Respondent-driven sampling for characterizing unstructured overlays," in *IEEE INFOCOM*, Apr. 2009.

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

[7] C.-H. Lee, X. Xu, and D. Y. Eun, "Beyond random walk and metropolis-hastings samplers: Why you should not backtrack for unbiased graph sampling," in *ACM SIGMETRICS/Performance*, June 2012.

[8] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[9] M. A. Hasan and M. J. Zaki, "Output space sampling for graph patterns," in *VLDB*, Aug. 2009.

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

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

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

[13] L. Jin et al., "Albatross sampling: Robust and effective hybrid vetex sampling for social graphs," in *ACM HotPlanet*, June 2011.

[14] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2010.

[15] G. L. Jones, "On the Markov chain central limit theorem," *Probability Surveys*, vol. 1, pp. 299–320, 2004.

[16] G. O. Roberts and J. S. Rosenthal, "General state space Markov chains and MCMC algorithms," *Probability Surveys*, vol. 1, pp. 20–71, 2004.

[17] P. Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer-Verlag, 1999.

[18] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, “Equation of state calculations by fast computing machines,” *Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: Bringing order to the web,” 1999.

[20] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[21] D. Aldous and J. Fill, *Reversible Markov Chains and Random Walks on Graphs*. monograph in preparation.

[22] A. Mohaisen, A. Yun, and Y. Kim, “Measuring the mixing time of social graphs,” in *IMC*, Nov. 2010.

[23] S. Hoory, N. Linial, and A. Wigderson, “Expander graphs and their applications,” *Bull. Amer. Math. Soc.*, vol. 43, no. 4, pp. 439–561, 2006.

[24] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 2nd ed. John Wiley & Sons, 2000.

[25] “Stanford Large Network Dataset Collection,” http://snap.stanford.edu/data/.

[26] B. Ribeiro and D. Towsley, “Estimating and sampling graphs with multidimensional random walks,” in *IMC*, Nov. 2010.

[27] X. Xu, C.-H. Lee, and D. Y. Eun, “A general framework of hybrid graph sampling for complex network analysis,” Technical Report, http://www4.ncsu.edu/~dyeun/pub/techrep-hybrid-sampling13.pdf.

[28] R. Bru, R. Canto, R. L. Soto, and A. M. Urbano, “A brauer’s theorem and related results,” *Central European Journal of Mathematics*, vol. 10, no. 1, pp. 312–321, 2012.

[29] Y. Saad, *Numerical methods for large eigenvalue problem*, 2nd ed. SIAM, 2011.

Appendix

A. Proof of Theorem 1:

We need the following result regarding rank-one update of a matrix and its resulting eigenvalues and eigenvectors.

Proposition 4: [28], [29] Let $\mathbf{A}$ be an $n \times n$ arbitrary matrix with eigenvalues $\theta_1, \theta_2, \ldots, \theta_n$ and their corresponding eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Pick one of the eigenvectors, $\mathbf{x}_k$, and set $\mu_k = \theta_k + \mathbf{x}_k^T \mathbf{q}$ for any n-dimensional vector $\mathbf{q}$, where $\mu_k \neq \theta_i$, $i = 1, 2, \ldots, n$.

Then, the matrix $\mathbf{A}' = \mathbf{A} + \mathbf{x}_k \mathbf{q}^T$ (rank-one update) has eigenvalues of $\{\theta_1, \theta_2, \ldots, \theta_{k-1}, \mu_k, \theta_{k+1}, \ldots, \theta_n\}$, and the corresponding eigenvectors of $\{\mathbf{w}_1, \ldots, \mathbf{w}_{k-1}, \mathbf{x}_k, \mathbf{w}_{k+1}, \ldots, \mathbf{w}_n\}$, where

$$\mathbf{w}_i = \mathbf{x}_i - \frac{\mathbf{q}^T \mathbf{x}_i}{\mu_k - \theta_i} \mathbf{x}_k, \quad i \neq k.$$

Here we define $\mathbf{A} = (1-\alpha)\mathbf{P} + \alpha(1-\beta)\mathbf{I}$. Clearly, the eigenvalues of $\mathbf{A}$ are

$$\theta_i = (1-\alpha)\lambda_i + \alpha(1-\beta), \quad i = 1, 2, \ldots, n, \tag{18}$$

associated with the same eigenvectors $\mathbf{x}_i = \mathbf{v}_i$ of $\mathbf{P}$. We now apply Proposition 4 to the matrix $\mathbf{A}$ and set $k = 1$, $\mathbf{q} = \frac{\alpha\beta}{n}\mathbf{1}$. Note that $\mathbf{x}_1 = \mathbf{1}$ with $\lambda_1 = 1$ here. Then we have

$$\mathbf{P}_\alpha = \mathbf{A} + \frac{\alpha\beta}{n}\mathbf{1}\mathbf{1}^T = \mathbf{A} + \mathbf{x}_1 \mathbf{q}^T.$$

Observe

$$\mu_1 = \theta_1 + \mathbf{x}_1^T \mathbf{q} = (1-\alpha) + \alpha(1-\beta) + \frac{\alpha\beta}{n}\mathbf{1}^T\mathbf{1} = 1.$$

Clearly, $\mu_1 \neq \theta_i$ for all $i = 1, 2, \ldots, n$. Thus, Proposition 4 asserts that $\mathbf{v}_1(\alpha) = \mathbf{1}$ is an eigenvector of the matrix $\mathbf{P}_\alpha$ associated with the eigenvalue $\lambda_1(\alpha) = \mu_1 = 1$, and the eigenvectors of $\mathbf{P}_\alpha$ associated with $\lambda_i(\alpha) = \theta_i$, $i \neq 1$, are:

$$\mathbf{v}_i(\alpha) = \mathbf{v}_i - \frac{\alpha\beta \mathbf{1}^T \mathbf{v}_i}{n(1-\lambda_i(\alpha))}\mathbf{1}, \; i = 2, 3, \ldots, n. \tag{19}$$

Since $\mathbf{P}$ is symmetric, all the eigenvectors are orthogonal, thus $\mathbf{1}^T \mathbf{v}_i = 0$ for $i = 2, \ldots, n$. This means $\mathbf{v}_i(\alpha) = \mathbf{v}_i$ for $i = 1, \ldots, n$, and the corresponding eigenvalues of $\mathbf{P}_\alpha$ are $\lambda_i(\alpha)$. From (18) and our definition of $\lambda_i(\alpha)$ in (7), we are done.

B. Proof of Lemma 1:

We prove the result here for a slightly more general case with arbitrary π, not necessarily with the uniform $\boldsymbol{\pi} = \mathbf{u}$. For any function $f : \mathcal{N} \to \mathbb{R}$, we interpret $\mathbf{P}$ as an operator from $\mathcal{N}$ to $\mathcal{N}$ defined as

$$(\mathbf{P}^k f)(i) \triangleq \sum_{j \in \mathcal{N}} P^{(k)}(i,j) f(j) = \mathbb{E}\{f(X_k) | X_0 = i\}, \tag{20}$$

where $P^{(k)}(i,j)$ is the k-step transition probability from i to j. Then, from (20), we have

$$\begin{aligned} \langle f, \mathbf{P}^k f \rangle_\pi &\triangleq \sum_{i,j \in \mathcal{N}} \pi(i) f(i) P^{(k)}(i,j) f(j) \\ &= \mathbb{E}_{\boldsymbol{\pi}}\{f(X_0) f(X_k)\}. \end{aligned} \tag{21}$$

On the other hand, since the set of eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\}$ forms an orthonormal basis of $\mathbb{R}^n$, any vector $f \in \mathbb{R}^n$ (or a function $f : \mathcal{N} \to \mathbb{R}$) can be expressed as $f = \sum_{i=1}^n \alpha_i \mathbf{v}_i$, where $\alpha_i = \langle f, \mathbf{v}_i \rangle_\pi$. It thus follows that

$$\mathbb{E}_\pi(f^2) = \langle f, f \rangle_\pi = \sum_{i=1}^n |\langle f, \mathbf{v}_i \rangle_\pi|^2. \tag{22}$$

Similarly, note that

$$|\langle f, \mathbf{v}_1 \rangle_\pi|^2 = |\langle f, \mathbf{1} \rangle_\pi|^2 = \mathbb{E}_\pi^2(f). \tag{23}$$

Thus, the normalizing constant Z becomes

$$Z = \sum_{j=2}^n |\langle f, \mathbf{v}_j \rangle_\pi|^2 = \mathbb{E}_\pi(f^2) - \mathbb{E}_\pi^2(f) = \mathrm{Var}_\pi(f). \tag{24}$$

Now, observe

$$\begin{aligned} \langle f, \mathbf{P}^k f \rangle_\pi &= \left\langle \sum_{i=1}^n \langle f, \mathbf{v}_i \rangle_\pi \mathbf{v}_i, \; \mathbf{P}^k \left(\sum_{i=1}^n \langle f, \mathbf{v}_i \rangle_\pi \mathbf{v}_i \right) \right\rangle_\pi \\ &= \left\langle \sum_{i=1}^n \langle f, \mathbf{v}_i \rangle_\pi \mathbf{v}_i, \; \sum_{i=1}^n \lambda_i^k \langle f, \mathbf{v}_i \rangle_\pi \mathbf{v}_i \right\rangle_\pi \\ &= \sum_{i=1}^n \lambda_i^k |\langle f, \mathbf{v}_i \rangle_\pi|^2 = |\langle f, \mathbf{v}_1 \rangle_\pi|^2 + Z \cdot \mathbb{E}_\Lambda\{\Lambda^k\} \\ &= \mathbb{E}_{\boldsymbol{\pi}}^2(f) + \mathrm{Var}_{\boldsymbol{\pi}}(f) \mathbb{E}_\Lambda\{\Lambda^k\}, \end{aligned} \tag{25}$$

where the second last equality is from (10)–(11), and the last equality is from (23)–(24). Thus, (14) follows by combining (21) and (25), and by noting that

$$\begin{aligned} &\mathbb{E}_\pi\{(f(X_0) - \mathbb{E}_\pi(f))(f(X_k) - \mathbb{E}_\pi(f))\} \\ &\quad = \mathbb{E}_{\boldsymbol{\pi}}\{f(X_0) f(X_k)\} - \mathbb{E}_{\boldsymbol{\pi}}^2(f). \end{aligned}$$

In our case, $\boldsymbol{\pi} = \mathbf{u}$, This completes the proof.