# Challenging the Limits: Sampling Online Social Networks with Cost Constraints

Xin Xu Chul-Ho Lee Do Young Eun

***Abstract*—Graph sampling techniques via random walk crawling have been popular for analyzing statistical characteristics of large online social networks due to simple implementation and provable guarantees on unbiased estimates. Despite the growing popularity, the 'cost' of sampling and its true impact on the accuracy of estimates still have not been carefully studied. In addition, the random walk-based methods inherently suffer from the sluggish nature of random walks and the 'slow-mixing' structure of social graphs, thereby leading to high correlation in the samples obtained. With these in mind, in this paper, we develop a mathematical framework such that the cost of sampling is properly taken into account, which in turn re-defines a widely used asymptotic variance into a cost-based asymptotic variance. Our new metric enables us to compare a class of sampling policies under the same cost constraint, integrating "random skipping" (bypassing nodes without sampling) into the random walk-based sampling. We obtain an optimal policy striking the right balance between sampling quality (less correlation) and sampling quantity (higher cost per sample), which greatly improves over the usual skip-free crawling-based samplers. We further extend our framework, enabling one to design more sophisticated sampling strategies with an array of control knobs, which all produce unbiased estimates under the same cost constraint.**

## I. Introduction

Recently, online social networks (OSNs) such as Facebook, Twitter, and Digg, have triggered a tremendous amount of attention in various disciplines because of their extensive applications and massive useful data. A large number of research studies have been conducted into this area, aiming at exploring the underlying social structure and network characteristics or improving on information retrieval tasks of social data. However, the sheer size of such complex networks often refrains researchers and developers from obtaining the complete database to thoroughly study their properties [1], forcing them to resort to graph sampling techniques in order to obtain "sampled data" for the purpose of estimating the characteristics of the networks in a compact manner [2], [3].

On the other hand, most of today's OSNs usually provide restrictive, local-neighborhood-only access interfaces, albeit public, to researchers and developers [4], [5]. Even worse, the size of such a network is changing dynamically and is unknown to the public. Thus it is often infeasible to perform an *ideal* user ID-based uniform sampling or its variants, sampling users uniformly at random from the network, which requires prior knowledge about the username configuration of the whole system [4]. Under these circumstances, "crawling" the network has become the most realistic solution to collect samples, which only requires exploring the network's neighborhood structure (e.g., [6], [4], [7], [8], [9] and references therein). In particular, random walk-based crawling methods have been more widely used than any other crawling ones, as they can be easily implemented in a distributed manner, while ensuring unbiased samples of graph properties with proper post-processing if necessary. The basic idea here is to launch a random walker (or multiple parallel walkers) moving from a node to one of its direct neighbors to obtain a set of samples. Metropolis-Hastings random walk (MHRW) and simple random walk (SRW) with reweighting are the most popular of this kind [10], [11].

### A. Motivation

Although the graph sampling techniques greatly reduce the workload of data analysis tasks to uncover network characteristics, we still need to take into account cost or resource consumption associated with the 'sampling' operation, which constrains the total sample size and consequently the accuracy of the estimates. The cost/resource restrictions can be in many forms. One prime example for the cost is the time and/or memory consumption associated with public API requests or HTML screen-scraping for crawling [6]. When resorting to public API requests, API rate limiting would also be an important consideration [5], [12].* In addition, retrieving and processing the information for each sample usually requires a different amount of resources determined by factors such as the sampling technique itself, crawler's location and target quantity to estimate. Other examples include the budgetary cost in purchasing a dataset of users, and the number of servers used as web-crawlers, to name a few.

Consider a typical scenario of crawling an OSN via Web scraping. In order to transit a user by looking into his/her friend list (or move to one of its neighbors), one needs to download a Web page whose URL address is, for example, "http://www.facebook.com/user-id/friends" to get all the friends of the current '*user-id*', regardless of whether any *sampling* operation for that user is performed. However, if we are interested in more than just a list of friends, e.g., the user's education background, affiliation, social activities, etc., we also need to download additional Web pages, e.g., the Web page of "http://www.facebook.com/user-id/about". This is usually the case in sampling the current user, i.e., retrieving the user's information, and then moving to the next user to

Xin Xu and Do Young Eun are with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, and Chul-Ho Lee is with the Department of Electrical and Computer Engineering, Florida Institute of Technology, Melbourne, FL. Email: {xxu5, dyeun}@ncsu.edu; clee@fit.edu. This work was supported in part by National Science Foundation under Grant Nos. CNS-1217341 and CNS-1423151.

*For instance, Twitter allows only 15 API requests to retrieve IDs of a user's followers every 15 minutes [13].

continue crawling. Thus, sampling a user (e.g., downloading both 'friends' and 'about' pages) clearly costs more than simple traversal (e.g., downloading the 'friends' page only), and the cost of sampling can also be different depending on the kind of information extracted from the user's Web pages (and even from his/her friends' Web pages).

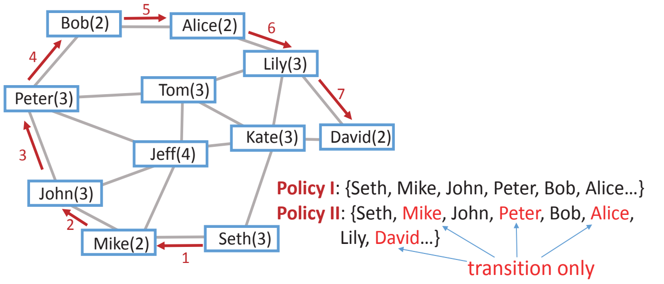

Bob(2)
Alice(2)
Lily(3)
Peter(3)
Tom(3)
Kate(3)
David(2)
Jeff(4)
John(3)
Mike(2)
Seth(3)
1
2
3
4
5
6
7
Policy I: {Seth, Mike, John, Peter, Bob, Alice...}
Policy II: {Seth, Mike, John, Peter, Bob, Alice, Lily, David...}
transition only

Fig. 1. Illustrating two different sampling policies when crawling a social graph.

Figure 1 shows an illustrative example on how such a cost difference could affect the sampling performance. Here, a crawler (or random walker) collects a sequence of samples along a path whose direction is depicted by red arrows, yet under two different sampling policies – Policies I and II. For illustration purposes, we only consider time as the sole factor for the 'cost'. The crawler at a user may opt out the sampling operation; instead, it directly moves to one of its neighbors *without retrieving* any information at that user. We assume that such a transition-only operation just costs 1 unit of time as there is no information retrieval required. In contrast, *sampling* a user and then moving to one of its neighbors costs *more* than 1 unit of time. The exact cost for both sampling and moving appears inside the parenthesis right next to the name of each user in Figure 1.

Policy I here means that the crawler always takes a sample at each and every visited user, while Policy II refers to a case where the crawler can proceed to the next user without sampling (transition only) from time to time. As can be seen from Figure 1, Policy I collects 6 samples from Seth, Mike, John, Peter, Bob and Alice, with the total cost budget of 15 units of time. In contrast, Policy II leads to a collection of only 4 samples from Seth, John, Bob and Lily under the same total budget, since their corresponding costs are 3, 3, 2 and 3 units of time respectively, along with four intermediate 'transition-only' operations wasting 4 units of time. Observe that Policy II ends up with fewer samples than Policy I, but these four samples are *less correlated* than those six in Policy I. In other words, Policies I and II exhibit a tradeoff between 'superior' yet fewer samples and 'inferior' yet more samples, or, more generally, a tradeoff between sampling "quality" and sampling "quantity", which should be carefully dealt with when resources are concerned.

Currently, almost all the existing sampling algorithms [4], [14], [15], [8], however, assume that only some constant cost is deducted for each sample. This kind of model apparently does not fit all the situations of sampling problems. Under a fixed cost constraint, although more sophisticated algorithms usually result in a set of "superior" samples, the number of total samples obtained could be less. It is thus not guaranteed that these algorithms can serve as better sampling strategies. In addition, most previous studies [4], [14], [15], [16], [17], [9] aimed at comparing different algorithms assuming the same number of samples acquired (or their asymptotic performance in the limit) become *inapplicable* to the cases when cost constraints come into play. This observation motivates us to develop a general framework from which one can design more effective sampling strategies, judiciously exploiting the tradeoff between sample quality and sample quantity so as to improve the sampling performance under a cost constraint.

### B. Related Work

There have been few studies that attempt to take into account the sampling cost when designing graph sampling algorithms. For example, the authors in [18] have proposed an algorithm that combines independent uniform node sampling into random walk-based crawling, in order to improve the rate of convergence to the stationary distribution. They assumed unit cost for crawling and ran a sequence of simulations with different restart cost settings, but this practical concern about cost is considered only via numerical results without any theoretical support. A hybrid sampling method in [14] incorporates random jump into crawling for the purpose of reducing the asymptotic variance of any given estimator. Differently from [18], the cost of a failed random jump is to repeat the previously reported sample, but this setting does not extend to any more general case. Among other current literature, [4], [15], [16], [17], [9], [8] assume that only a unit cost is deducted for each sample. Some other works targeting at comparing the performance of different graph sampling algorithms, although not mentioning the terminology of "cost", assume that the walk length (also called sample size or iterations) is fixed for a fair comparison [19], [20], [21], [22], [23], [24]. This assumption basically admits the existence of cost restrictions to collect desired number of samples. However, they oversimplified the problem and assumed that collecting the same number of samples consumes the same amount of resources over different sampling strategies under consideration.

### C. Our Contributions

We first demonstrate that the so-called asymptotic variance, widely used to evaluate and rank the performance of samplers in the MCMC and graph sampling literature, is deficient when the cost of sampling is taken into account. As an alternative and corrective metric, we rigorously define a "cost-based" asymptotic variance on which different sampling strategies can be fairly compared under the same cost constraint. With this metric in hand, we integrate "random skipping" (intentionally bypassing nodes without sampling), having the benefit of less correlations in the samples obtained yet with higher costs per sample, into the off-the-shelf random walk-based sampling, and then find an optimal policy striking the right balance between sample quality and quantity. Our mathematical framework is general enough, thus being applicable to any random walk methods even requiring a reweighting procedure for unbiased estimation. We extend this framework to further challenge the cost limits by allowing a sampling decision tailored to each location of the crawler, which in turn enables one to design a wide collection of sophisticated, location-dependent strategies, all producing unbiased estimates. The simulation results based on several graph datasets justify our theoretical findings, and show that by judiciously taking advantage of random skipping, we can reap a significant amount of efficiency improvements under stringent cost constraints. To the best of our knowledge, this is the first work to develop a general, theoretical framework, enabling one to compare different graph sampling strategies under the same sampling budget and also providing diverse design choices for a new sampling strategy.

## II. Preliminaries

In this section, we first provide a theoretical background for random walk-based graph sampling to unbiasedly estimate the statistics of a graph via random walk crawling. We then present two representative sampling methods of its kind in the literature.

### A. Theoretical Background

We define an undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ to model the structure of an OSN, where $\mathcal{N} = \{1, 2, \ldots, n\}$ is the set of nodes (users) and $\mathcal{E}$ is the set of edges (users' social relationships). For node $i$, let $N(i) = \{j \in \mathcal{N} : \exists (i,j) \in \mathcal{E}\}$ be the set of its neighbors, and $d(i) = |N(i)|$ its degree. The goal of unbiased graph sampling here is to unbiasedly estimate the statistics of the graph $G$, without obtaining complete graph data. To be precise, for any given function $f : \mathcal{N} \rightarrow \mathbb{R}$, we are interested in estimating $\mathbb{E}_{\mathbf{u}}(f) \triangleq \sum_{i \in \mathcal{N}} f(i)/n$, where $\mathbf{u} \triangleq [u(1), u(2), \cdots, u(n)] = [1/n, 1/n, \cdots, 1/n]$ is a uniform distribution over $\mathcal{N}$. To this end, there have been several random walk-based sampling methods proposed in the literature, which are almost all described by the theory of Markov chains [25], [26].

Consider a discrete-time Markov chain (or a discrete-time random walk) $\{X_t \in \mathcal{N}, t = 0, 1, 2 \ldots\}$ on $\mathcal{G}$ with transition probability matrix $\mathbf{P} \triangleq \{P(i,j)\}_{i,j \in \mathcal{N}}$ where $P(i,j) = \mathbb{P}\{X_{t+1} = j | X_t = i\}$ is a transition probability from node $i$ to $j$. We assume that the Markov chain $\{X_t\}$ is irreducible and aperiodic (and hence ergodic), having a unique stationary distribution $\boldsymbol{\pi} \triangleq [\pi(1), \pi(2), \cdots, \pi(n)]$. Then one can construct an estimator based on $\{X_t\}$, which is given by

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

Since the chain is ergodic, the above estimator converges to its statistical average, i.e.,

$$\lim_{t \rightarrow \infty} \hat{\mu}_t(f) = \mathbb{E}_{\boldsymbol{\pi}}(f) \triangleq \sum_{i \in \mathcal{N}} f(i)\pi(i) \quad \text{a.s.} \tag{2}$$

for any function $f$ with $\mathbb{E}_{\boldsymbol{\pi}}(|f|) < \infty$ and any initial distribution. It follows that the estimator $\hat{\mu}_t(f)$ provides an asymptotically unbiased approximation for $\mathbb{E}_{\mathbf{u}}(f)$ if $\boldsymbol{\pi} = \mathbf{u}$.

While $\hat{\mu}_t(f)$ has been widely used to infer the statistics of $\mathcal{G}$, its estimate may fluctuate around its true value *even* in the long run. It is always desirable to have smaller fluctuation or higher accuracy of the estimate. In assessing the accuracy of the estimator $\hat{\mu}_t(f)$, the so-called asymptotic variance has been one of the most important and popular criterions in the MCMC and graph sampling literature (e.g., [25], [27], [26], [28], [14], [8]), and is defined by, for $\mathbb{E}_{\boldsymbol{\pi}}(f^2) < \infty$,

$$\begin{aligned} v(f, \mathbf{P}, \boldsymbol{\pi}) &\triangleq \lim_{t \rightarrow \infty} t \cdot \mathrm{Var}(\hat{\mu}_t(f)) \\ &= \mathrm{Var}_{\boldsymbol{\pi}}(f) + 2 \sum_{k=1}^{\infty} \mathrm{Cov}_{\boldsymbol{\pi}} \left[ f(X_0), f(X_k) \right], \end{aligned} \tag{3}$$

where $\mathrm{Var}_{\boldsymbol{\pi}}(f) = \mathbb{E}_{\boldsymbol{\pi}}(f^2) - \mathbb{E}_{\boldsymbol{\pi}}(f)^2$ is the marginal variance of a function $f$ with respect to $\boldsymbol{\pi}$, and $\mathrm{Cov}_{\boldsymbol{\pi}} [f(X_0), f(X_k)] = \mathbb{E}_{\boldsymbol{\pi}} [(f(X_0) - \mathbb{E}_{\boldsymbol{\pi}}(f))(f(X_k) - \mathbb{E}_{\boldsymbol{\pi}}(f))]$ is the lag $k$ autocovariance of the stationary sequence $\{f(X_t)\}$. From the Central Limit Theorem for a Markov chain [27], [26], $\sqrt{t}[\hat{\mu}_t(f) - \mathbb{E}_{\boldsymbol{\pi}}(f)]$ converges in distribution to a Gaussian random variable with zero mean and variance $v(f, \mathbf{P}, \boldsymbol{\pi})$.

### B. Random Walk-based Sampling Methods

A most famous sampling method is the Metropolis-Hastings random walk (MHRW) [25], [8], which is based on the celebrated Metropolis-Hastings algorithm enabling one to sample from a desired probability distribution. It is a two-step procedure: a proposal for next candidate move and its follow-up decision to accept or reject the proposal, leading to the following transition probabilities:

$$P(i,j) = \begin{cases} \min\left\{\frac{1}{d(i)}, \frac{1}{d(j)}\right\} & \text{if } j \in N(i), \\ 1 - \sum_{k \in N(i)} P(i,k) & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

This chain is reversible with respect to $\boldsymbol{\pi} = \mathbf{u}$.[†] Thus the samples obtained by the MHRW can be directly used 'as is', and the resulting estimator in (1) is simply unbiased.

The other famous sampling method is to use the simple random walk (SRW) with a post-hoc reweighting procedure. In the SRW, the next node is chosen uniformly at random among the neighbors of the current node. Assuming the current node $i$, the probability of moving from $i$ to $j$ is

$$P(i,j) = \begin{cases} 1/d(i) & \text{if } j \in N(i), \\ 0 & \text{otherwise.} \end{cases}$$

This chain is also reversible but with stationary probabilities $\pi(i) = d(i)/(2|\mathcal{E}|)$. The stationary distribution here is proportional to the node degree, indicating that the estimator $\hat{\mu}_t(f)$ in (1) is *biased* toward high-degree nodes. Thus a reweighting procedure becomes necessary to correct the bias, and it is essentially using the following estimator instead of the one in (1):

$$\frac{\hat{\mu}_t(f_w)}{\hat{\mu}_t(w)} = \frac{\sum_{s=1}^{t} w(X_s) f(X_s)}{\sum_{s=1}^{t} w(X_s)} = \frac{\sum_{s=1}^{t} f(X_s)/d(X_s)}{\sum_{s=1}^{t} 1/d(X_s)}, \tag{5}$$

where $f_w \triangleq w \circ f$ is the Hadamard product of $w$ and $f$, with $f_w(i) = w(i)f(i)$ and $w(i) = 1/d(i)$ for $i \in \mathcal{N}$. It is known

[†] A Markov chain is *reversible* if $\pi(i)P(i,j) = \pi(j)P(j,i)$ for all $i, j$.

that $\hat{\mu}_t(f_w)/\hat{\mu}_t(w) \to \mathbb{E}_{\mathbf{u}}(f)$ almost surely as $t \to \infty$, i.e., the estimator in (5) is asymptotically unbiased [19], [11], [8].

## III. Skipping vs. Sampling

We provide a mathematical model to incorporate "random skipping" (or passing by the currently visited node without sampling) into the random walk-based sampling, having the benefit of less correlation in the samples obtained but paying higher costs per sample. We then develop a unified framework enabling us to compare such sampling policies with the same cost constraint, from which we obtain an optimal sampling policy. We also explain how these results can be carried over to the case of requiring a post-hoc reweighting procedure as is the one for SRW.

### A. Mathematical Model

Consider a sampling agent that moves over $\mathcal{G}$ according to the transition matrix $\mathbf{P}_0$ of a reversible Markov chain $\{X_s, s=0,1,\ldots\}$, where $X_s$ is the location of the agent at time $s$. We consider a class of sampling policies $\mathcal{P} = \mathcal{P}(p) \in \mathscr{P}$ that allow for skipping the currently visited node *without* sampling or retrieving any information at the node, indexed by a decision probability $0 < p \leq 1$. While traversing over $G$ according to $\mathbf{P}_0$, the agent samples the current node $X_s = i$ with probability $p$, and skips node $i$ with probability $1-p$. Note that we here focus on the 'state-independent' strategies (with the same $p$) for ease of exposition. We will relax this later in Section IV and allow the decision probability to be state-dependent (location-dependent).

Now, we introduce the notion of 'cost' associated with the currently visited node, which is different depending on whether to skip or take a sample there. Let $\{c_s, s=1,2\ldots\}$ be a sequence of incurred costs, corresponding to $\{X_s\}$. At each node $i$, we assume that the cost of 'sampling' is $a(i)$, while the cost of 'skipping' is $b(i)$. We also assume that $0 < b(i) \leq a(i)$ for all $i$, and every cost is bounded, i.e., $a(i) \leq \bar{c}$ for some finite constant $\bar{c}$. To summarize, when the sampling agent enters node $i$ at time $s$ ($X_s = i$), it performs sampling there with probability $p$, incurring cost $c_s = a(i)$. Otherwise, it simply passes by node $i$, with $c_s = b(i)$.

Let $\{G_t, t=1,2,\ldots\}$ be an *i.i.d.* sequence of random variables indicating the length of intervals between two consecutive samples obtained. Clearly, $\mathbb{P}\{G_t = l\} = (1-p)^{l-1}p$. We then define a sequence of *sampled nodes* $\{Z_t, t=0,1,\ldots\}$, where $Z_t = X_{L_t}$ indicates the location at which the 'sampling' operation is performed, with $L_t = G_1+G_2+\ldots+G_t$. Assume $Z_0 = X_0$. Note that the sampling accuracy under a given policy $\mathcal{P}(p)$ is decided by its corresponding $\{Z_t\}$. Letting $C_t$ be the cost to obtain the $t$-th sample given by

$$C_t = \sum_{s=L_{t-1}+1}^{L_t} c_s,$$

$\{C_t, t=1,2,\ldots\}$ then becomes the sequence of costs associated with $\{Z_t\}$. For instance, consider Policy II in Figure 1. We see that $\{X_1, X_2, X_3, \ldots\} = \{$Seth, Mike, John, $\ldots\}$, with associated costs $\{c_1, c_2, c_3, \ldots\} = \{3,1,3,\ldots\}$. Contrastingly, $\{Z_1, Z_2, Z_3, \ldots\} = \{$Seth, John, Bob, $\ldots\}$, and their associated costs become $\{C_1, C_2, C_3, \ldots\} = \{3,4,3,\ldots\}$. For instance, $C_2 = c_2 + c_3 = 4$ is the sum of costs for passing by Mike (1 unit of time) and sampling John (3 unites of time).

Let $1 = \lambda_{1,0} > \lambda_{2,0} \geq \cdots \geq \lambda_{n,0} > -1$ be the $n$ eigenvalues of the reversible chain $\{X_s\}$ with transition matrix $\mathbf{P}_0$, and $\mathbf{v}_{i,0}$ ($i = 1,2,\ldots$) be the corresponding eigenvectors. Let $\boldsymbol{\pi}$ be the stationary distribution of $\{X_s\}$. Unless otherwise stated we consider $\boldsymbol{\pi} = \mathbf{u}$ as is the case for MHRW (but this is not the only Markov chain leading to the uniform stationary distribution $\mathbf{u}$). We then have the following. Due to space constraint, we refer to our technical report [29] for all the proofs in this paper.

*Lemma 1:* $\{Z_t\}$ is a reversible Markov chain with transition matrix $\mathbf{P}(p)$ given by

$$\mathbf{P}(p) = \sum_{l=1}^{\infty} \mathbf{P}_0^l p(1-p)^{l-1}, \tag{6}$$

and the same stationary distribution $\boldsymbol{\pi}(p) = \boldsymbol{\pi}$. The corresponding eigenvalues and eigenvectors are

$$\lambda_i(p) = \frac{p\lambda_{i,0}}{1-(1-p)\lambda_{i,0}}, \text{ and } \mathbf{v}_i(p) = \mathbf{v}_{i,0}, \tag{7}$$

respectively. □

Note that as a special case, if $p = 1$, then $Z_t = X_t$ for all $t \geq 0$, with $\mathbf{P}(1) = \mathbf{P}_0$.

### B. Asymptotic Variance Under Cost Constraint

Let $M$ be the total cost budget allowed for the sampling purpose. Define

$$T = T(M) = \max\{u : C_1 + C_2 + \cdots + C_u \leq M\} \tag{8}$$

to be the stopping time to spend all $M$, i.e., *the number of samples collected* when the sampling process is terminated. We require that under any feasible policy of consideration, $T \to \infty$ almost surely as $M \to \infty$. This rules out trivial policies in which, for instance, a sampling agent always makes a transition to next node without sampling at all, which still consumes costs, until it reaches the total cost budget. We next develop a new cost-oriented performance metric, which is a refined version of the asymptotic variance in (3), for a fair comparison of sampling policies.

Since Lemma 1 guarantees that all possible policies $\mathcal{P}(p) \in \mathscr{P}$ under consideration possess the same stationary distribution $\boldsymbol{\pi}$ regardless of $p > 0$, the new estimator

$$\hat{\mu}_T(f) = \frac{1}{T}\sum_{t=1}^{T} f(Z_t) \to \mathbb{E}_{\boldsymbol{\pi}}(f) \qquad \text{a.s.}, \tag{9}$$

also converges to the same target quantity as that of the estimator based on $\{X_s\}$. (See (1) and (2).) Thus, it would be tempting to consider the asymptotic variance of this new estimator as was used in other existing studies (e.g., [28], [14], [8]), to evaluate and compare the performance of different sampling policies and also to find the optimal policy. When sampling cost constraints come into play, however, the tradeoff between sample "quality" and "quantity" makes this problem more subtle.

Intuitively, when the 'skip' rate approaches 1 (i.e., $p \to 0$), the samples obtained behave more like those from the *ideal* (uniform) random, independent samples directly drawn from $\mathcal{N}$ [4], thereby leading to smaller asymptotic variance than that from any random walk-based sampling method producing correlated samples. Under a given cost constraint, however, such a policy would spend almost all the budget $M$ on making transitions only, with very few samples actually collected in the end. This clearly demonstrates that one cannot simply use the asymptotic variance 'as is' to rank the sampling policies under the same ground. Instead, one would have to consider the variance of the estimator with a different number of samples obtained under a given cost budget. Therefore, we renormalize the asymptotic variance in a way that any two competing sampling policies are to be compared under the same but large amount of 'cost', not the number of samples obtained as would be the case in the literature.

We notice that from (3), we have $\text{Var}(\hat{\mu}_T(f)) \to 0$, as $M \to \infty$. Thus, in order to compare the resulting estimators under different policies but under the same total cost constraint, we consider the variance of $\sqrt{M}(\hat{\mu}_T(f) - \mathbb{E}_{\boldsymbol{\pi}}(f))$ instead and obtain the following:

*Theorem 1:* Under the aforementioned setting with policy $\mathcal{P}(p)$, suppose $M/T(M)$ converges, as $M \to \infty$, in probability, to a constant $c(p)$ – the long-term average cost to obtain one sample. Then $\sqrt{M}\,(\hat{\mu}_T(f) - \mathbb{E}_{\boldsymbol{\pi}}(f))$ converges in distribution to a Gaussian random variable with mean 0 and variance equal to $c(p) \cdot \upsilon(f, \mathbf{P}(p), \boldsymbol{\pi})$. □

An immediate observation is that if $C_t = 1$ for all $t$, then $T = M$ (constant), and thus Theorem 1 reduces to the usual Central Limit Theorem for a Markov chain as mentioned at the end of Section II-A. For the rest of the paper, we call the new asymptotic variance as "cost-based asymptotic variance" given by

$$\Psi(p) \triangleq c(p) \cdot \upsilon(f, \mathbf{P}(p), \boldsymbol{\pi}). \tag{10}$$

### C. Variance Analysis and Optimal Policy

Since the cost-based asymptotic variance is now the product of $c(p)$ and $\upsilon(f, \mathbf{P}(p), \boldsymbol{\pi})$, one may expect that it captures the tradeoff between variance and cost per sample. In other words, when intentionally skipping nodes more often (or decreasing the value of $p$), it is expected to have less correlation (or equivalently, smaller asymptotic variance as seen from (3)) yet with higher cost per sample. In what follows, we show that this is indeed the case by obtaining its closed-form expression. We also find the optimal solution $p^*$ (or the optimal policy $\mathcal{P}(p^*)$) to minimize $\Psi(p)$ in (10).

Let $A \triangleq \mathbb{E}_{\boldsymbol{\pi}}(a) = \sum_{i \in \mathcal{N}} a(i)\pi(i)$ be the average cost of the sampling operation with respect to the stationary distribution $\boldsymbol{\pi}$, and similarly $B \triangleq \mathbb{E}_{\boldsymbol{\pi}}(b)$ for the case of the transition-only operation. Clearly, $A \geq B$ since $a(i) \geq b(i)$ from our assumption. We then have the following:

*Theorem 2:* Under a policy $\mathcal{P}(p)$, the cost-based asymptotic variance $\Psi(p) = c(p) \cdot \upsilon(f, \mathbf{P}(p), \boldsymbol{\pi})$ is given by

$$c(p) = B(1-p)/p + A, \text{ and} \tag{11}$$

$$\upsilon(f, \mathbf{P}(p), \boldsymbol{\pi}) = \gamma + \alpha p, \tag{12}$$

where $\gamma = \text{Var}_{\boldsymbol{\pi}}(f)$ and $\alpha = \upsilon(f, \mathbf{P}_0, \boldsymbol{\pi}) - \text{Var}_{\boldsymbol{\pi}}(f)$. □

The explicit expressions in Theorem 2 allow us to better explain why the original asymptotic variance might not work under a cost constraint. Note that $\alpha$ and $\gamma$ depend only on the sampling function $f$ and the graph structure, but are not functions of $p$. Thus $\upsilon(f, \mathbf{P}(p), \boldsymbol{\pi})$ in (12) is increasing in $p \in (0, 1]$. When $p \to 0$, namely, almost always performing transitions over $\mathcal{G}$ *without sampling* will result in the smallest asymptotic variance. This would make sense for the usual asymptotic variance as in (3), since the 'transition-only' operations will break the correlation between consecutive samples, leading to the ideal (uniform) random sampling in the limit, assuming there are still a large number of samples obtained. However, under a fixed cost constraint, this will incur a very high cost to obtain one sample on average as $c(p)$ in (11) grows indefinitely, thus offsetting smaller $\upsilon(f, \mathbf{P}(p), \boldsymbol{\pi})$.

*Proposition 1:* $\Psi(p)$ is convex in $p \in (0, 1]$. When $A = B$, its optimal solution becomes $p^* = 1$. If $A > B$, we have

$$p^* = \begin{cases} 1 & \text{if } \alpha/\gamma \leq \beta, \\ \sqrt{\beta\gamma/\alpha} & \text{if } \alpha/\gamma > \beta, \end{cases}$$

where $\beta = \frac{B}{A-B}$. □

Proposition 1 presents the optimal solution $p^*$ in a closed form. First, if obtaining $f(i)$ at node $i$ does not require any additional cost (e.g., $f(i) = d(i)$, the degree of node $i$), we have $A = B$ for which the optimal strategy is to sample all the time ($p^* = 1$). This is consistent with our intuition in that there is no point of skipping any node if the costs of sampling and skipping are the same. When the cost of sampling is higher than that of passing by, as illustrated with Figure 1, the optimal sampling strategy will depend on the subtle interplay between the amount of additional cost for the sampling operation encoded into the parameter $\beta$, and the graph topological properties and the sampling function $f$ captured by the parameter $\alpha/\gamma$. From the proof of Theorem 2, we observe that $\alpha$ and $\gamma$ involves the entire spectrum of the chain $\mathbf{P}_0$. The following result provides a sufficient condition if the second largest eigenvalue $\lambda_{2,0}$ of $\mathbf{P}_0$ is available.

*Corollary 1:* Let $A = (1+\delta)B$ for some $\delta > 0$. If $\lambda_{2,0} \leq 1/(1+2\delta)$, then $p^* = 1$. □

For a given chain $\mathbf{P}_0$ (e.g., the MHRW on $\mathcal{G}$), the second largest eigenvalue $\lambda_{2,0}$ is highly related to the so-called mixing time of the chain $X_s$ (or the speed of convergence of the distribution of $X_s$ to its stationary one $\boldsymbol{\pi}$), and there are a number of techniques available to estimate or bound this quantity. Corollary 1 suggests that if the additional cost for sampling over skipping is not too large (i.e., $\delta$ is small) and the chain is not so slow mixing (i.e., $\lambda_{2,0}$ is bounded away from 1 by more than $2\delta/(1+2\delta)$, then sampling all the time with $p^* = 1$ is optimal. This implies that in this case there is no benefit of intentionally skipping nodes (with cost $b(i)$) in order to reduce the correlation among samples, as the chain $\mathbf{P}_0$ is not so slow anyway ($\lambda_{2,0}$ is not too close to one) and the additional cost for sampling is not too much. Otherwise, again, it reduces down to a subtle interplay between the property of the crawler chain $\alpha/\gamma$ and the quantity $\beta$ depending on the average costs, which would be the typical case as social graphs generally do not possess the fast mixing property [30].

### *D. Non-Uniform Stationary Distribution*

All the discussions so far directly apply for *unbiased* graph sampling, with $\boldsymbol{\pi} = \mathbf{u}$. For example, the estimator $\hat{\mu}_T(f)$ in (9) is asymptotically unbiased according to Theorem 1. However, when the crawler chain $\{X_s\}$ has a non-uniform stationary distribution (e.g., SRW), a reweighting procedure is required to correct a bias in the samples obtained, as outlined in Section II-B. We below show all the preceding arguments still remain intact even with the reweighing procedure. To this end, we assume that a sampling agent traverses over $\mathcal{G}$ in a SRW fashion. That is, $\mathbf{P}_0$ is the transition matrix of the SRW, with stationary probabilities $\pi(i) = d(i)/(2|\mathcal{E}|)$.

We first define a new estimator as

$$\frac{\hat{\mu}_T\left(f_w\right)}{\hat{\mu}_T\left(w\right)} = \frac{\frac{1}{T}\sum_{t=1}^{T} f_w(Z_t)}{\frac{1}{T}\sum_{t=1}^{T} w(Z_t)}, \tag{13}$$

which is in the same form as (5) except that here $T$ is a random variable (stopping time) as in (8), indicating the number of samples obtained until all the budget $M$ is spent. Recall that $f_w = w \circ f$, with $w(i) = 1/d(i)$ as defined in Section II-B. Let us briefly explain how we get the estimator. Once the sampling agent obtains a new sample $f(i)$ of node $i$, we simply reweight (multiply) this value with $1/d(i)$. At the same time, we add the term $1/d(i)$ to the denominator in (13). We can then show that this new estimator under a policy $\mathcal{P}(p)$ is still an unbiased estimator, i.e.,

$$\hat{\mu}_T\left(f_w\right)/\hat{\mu}_T\left(w\right) \to \mathbb{E}_{\mathbf{u}}(f) \text{ almost surely, as } M \to \infty, \tag{14}$$

for any $p \in (0, 1]$. This result holds by following the similar lines in Section II-B, noting that $T \to \infty$ almost surely as $M \to \infty$ and the stationary distribution of $\{Z_t\}$ is the same as that of $\{X_s\}$ for any $p \in (0, 1]$ from Theorem 1.

As before, for all the possible estimators in (13) indexed (implicitly) by $p \in (0, 1]$ being *unbiased*, we can again turn to their 'cost-based' asymptotic variances to compare their sampling accuracy.

*Theorem 3:* As the total cost $M \to \infty$, under a policy $\mathcal{P}(p)$, $\sqrt{M}\left[\hat{\mu}_T\left(f_w\right)/\hat{\mu}_T\left(w\right) - \mathbb{E}_{\mathbf{u}}(f)\right]$ converges in distribution to a Gaussian random variable with zero mean and variance $c(p) \cdot v(\mathcal{H}, \mathbf{P}(p), \boldsymbol{\pi})$, where the function $\mathcal{H} : \mathcal{N} \to \mathbb{R}$ is given by

$$\mathcal{H}(i) = \bar{d}\left[f_w(i) - \mathbb{E}_{\mathbf{u}}(f) w(i)\right], \ i \in \mathcal{N}, \tag{15}$$

and $\bar{d} = 2|\mathcal{E}|/n$ is the average degree. □

*Corollary 2:* The cost-based asymptotic variance (or the objective function to minimize) becomes

$$\tilde{\Psi}(p) = \left(B\frac{(1-p)}{p} + A\right)\left(\tilde{\gamma} + \tilde{\alpha} p\right),$$

where $\tilde{\gamma} = \mathrm{Var}_{\boldsymbol{\pi}}(\mathcal{H})$ and $\tilde{\alpha} = v(\mathcal{H}, \mathbf{P}_0, \boldsymbol{\pi}) - \mathrm{Var}_{\boldsymbol{\pi}}(\mathcal{H})$. □

Theorem 3 and Corollary 2 tell us that after the reweighting procedure, our cost-based asymptotic variance $\tilde{\Psi}(p)$ is still given by the product of the expected cost per sample and the original asymptotic variance, but now with the sampling function $f$ replaced by $\mathcal{H}$ in (15). Thus, all the properties associated with $\Psi(p)$, including the convexity and condition for the optimal policy $p^*$, carry over in the same way to $\tilde{\Psi}(p)$.

## IV. State-Dependent Sampling

Up till now, we have focused on the state-independent sampling policies (the probability of sampling node $i$ is $p$, regardless of the current location $i$ of a sampling crawler). Recall that by construction, the sampling cost $a(i)$ is a function of node $i$, capturing the situation that some node incurs a higher cost for sampling. For example, if the sampling function $f$ requires an exploration of all the neighbors of node $i$, the cost of sampling node $i$ would be proportional to its degree $d(i)$. In this case, one may want to deliberately use different probabilities of sampling node $i$ such that nodes with *high* sampling costs are sampled with small probabilities. Such a state-dependent strategy can be brought in to further challenge the cost limits, to squeeze more samples out of a given budget and reduce the estimation error, as long as we can remove any resulting bias by a proper reweighting procedure.

As before, assume that a crawler moves over $\mathcal{G}$ according to the transition matrix $\mathbf{P}_0$ of a reversible chain $\{X_s\}$ with stationary distribution $\boldsymbol{\pi} = [\pi(1), \ldots, \pi(n)]$. Note that the stationary distribution $\boldsymbol{\pi}$ does not need to be just a unform $\mathbf{u}$ but can be arbitrary. When residing in node $i$, the crawler performs 'sampling' with probability $p(i) \in (0, 1]$, which is now *state-dependent (location-dependent)*. Again, the sampling operation with probability $p(i)$ will cost $a(i)$, while simply passing by node $i$ with probability $1 - p(i)$ will cost $b(i)$. The crawler continues this process until all the budget $M$ is spent. We consider a set of state-dependent sampling policies $\mathcal{P}(\boldsymbol{p})$, parameterized by $\boldsymbol{p} = [p(1), p(2), \cdots p(n)]$. We then have the following:

*Theorem 4:* The sequence of sampled nodes $\{Z_t\}$ under a policy $\mathcal{P}(\boldsymbol{p})$ is a reversible Markov chain with stationary probabilities $\pi'(i) \propto \pi(i)p(i), \ i \in \mathcal{N}$. □

Theorem 4 asserts that under the state-dependent policy $\mathcal{P}(\boldsymbol{p})$, the stationary probabilities of the resulting chain $\{Z_t\}$ are given by

$$\pi'(i) = \frac{\pi(i)p(i)}{\sum_{j \in \mathcal{N}} \pi(j)p(j)} = \frac{\pi(i)p(i)}{\mathbb{E}_{\boldsymbol{\pi}}(p)},$$

which are not uniform unless $p(i) \propto 1/\pi(i), \ i \in \mathcal{N}$. Due to a bias originating from the non-uniform stationary distribution, we cannot simply use the samples obtained 'as is', so we employ the usual reweighting procedure to remove the bias, along with its mathematical properties, as was used for the SRW. (See Section II-B and also Section III-D.)

Specifically, we set the weight function $w(i) \propto 1/(\pi(i)p(i))$ so that $\hat{\mu}_T\left(f_w\right)/\hat{\mu}_T\left(w\right)$ converges to $\mathbb{E}_{\mathbf{u}}(f)$ almost surely. For example, if the crawler chain $\{X_s\}$ is the SRW, we set

$w(i) = 1/(d(i)p(i))$. For the MHRW with $\boldsymbol{\pi} = \mathbf{u}$, it becomes $w(i) = 1/p(i)$. In addition, by following the similar lines as in the proof of Theorem 3, we can show that $\hat{\mu}_T(f_w)/\hat{\mu}_T(w)$ converges in distribution to a Gaussian random variable with zero mean and variance given by

$$\Psi'(\boldsymbol{p}) = c(\boldsymbol{p}) \cdot v(\mathcal{H}', \mathbf{P}(\boldsymbol{p}), \boldsymbol{\pi}'),$$

where $c(\boldsymbol{p})$ is the long-term average cost to obtain one sample, and $v(\mathcal{H}', \mathbf{P}(\boldsymbol{p}), \boldsymbol{\pi}')$ is the usual asymptotic variance of $\{Z_t\}$ with $\mathcal{H}'(i) = [f_w(i) - \mathbb{E}_{\mathbf{u}}(f)w(i)]/\mathbb{E}_{\boldsymbol{\pi}'}(w),\ i \in \mathcal{N}$. Note that the complicated form of the transition probabilities of $\{Z_t\}$ (see the proof of Theorem 4) prohibits any usable closed-form expression of the cost-based asymptotic variance $\Psi'(\boldsymbol{p})$. Nonetheless, the importance of our extension here is to enable one to design more sophisticated *state-dependent* sampling strategies, equipped with a full array of $n$ tunable knobs $p(i)$, all producing *unbiased estimates* with proper reweighting as outlined above.

## V. Numerical Results

### A. Simulation Setting

***Datasets:*** We conduct experiments over four real world dataset from Youtube, Slashdot, Wikipedia Talk [31] and Digg[‡] [32]. We summarize the characteristics of each dataset in Table I. To ensure the connectivity of each graph, we use its largest connected component.

TABLE I
Statistics of the dataset

| | Youtube | Slashdot | Wiki Talk | Digg |
|---|---|---|---|---|
| # of users | 1,134,890 | 77,360 | 2,394,385 | 270,535 |
| # of edges | 2,987,625 | 905,468 | 5,021,410 | 1,731,658 |

***Performance Metrics and Estimation Error:*** We validate our framework over various network settings under the following two sampling functions (network properties to estimate):

- Membership probability $\mathbb{P}\{i \in D\}$ for a set of nodes $D$ satisfying a certain pre-defined condition, e.g., membership. In words, we want to estimate the probability that a randomly chosen node belongs to the set $D$.
- Average clustering coefficient $\Phi \triangleq \sum_{i \in \mathcal{N}} \omega_i/|\mathcal{N}|$, where $\omega_i = \triangle(i)/\binom{d(i)}{2}$ for node $i$ with degree $d(i) \geq 2$, and $\omega_i = 0$ if otherwise. Here, $\triangle(i) = |\{(j,k) \in \mathcal{E} : (i,j) \in \mathcal{E} \text{ and } (i,k) \in \mathcal{E}\}|$ is the number of total connections among the neighbors of $i$, and $\binom{d(i)}{2} = d(i)(d(i)-1)/2$ is the maximum possible connections among all neighbors of $i$.

For the case of estimating the membership probability $\mathbb{P}\{i \in D\}$, we randomly choose a subset $D$ with size $|D| = 0.3|\mathcal{N}|$ a priori. We use a set of different constant values for the cost of sampling, namely, $a(i) = 1, 2, 10, 50, 100$ for all $i$. The sampling function $f$ here becomes $f(i) = \mathbf{1}_{\{i \in D\}}, i \in \mathcal{N}$. On the other hand, for the case of estimating the average clustering coefficient $\Phi$, we set $f(i) = \omega_i$ as in [4], [23], and set the 'sampling' cost to be $a(i) = d(i), i \in \mathcal{N}$, which reflects that the cost of sampling is equivalent to that of exploring all the neighbors of each node, proportional to its degree (location-dependent sampling cost). For both cases, the cost of skipping a node without sampling (transition only) is assumed to be 1, i.e., $b(i) = 1$ for all $i$.

In order to see if the cost-based asymptotic variance can help finding the optimal sampling probability $p^*$, we compare it with the mean squared error (MSE) of an estimator, given by $\mathsf{MSE}(\hat{x}) = \mathbb{E}[(\hat{x} - x)^2]$, where $\hat{x}$ represents the estimated value and $x$ is its ground truth. Here we use MSE instead of the normalized root mean square error (NRMSE) [22], [8], [28] to measure the estimation accuracy, to be consistent with our choice of the cost-based asymptotic variance $\Psi(p)$. Observe that NRMSE, which is defined as $\mathsf{NRMSE}(\hat{x}) = \sqrt{\mathbb{E}[(\hat{x} - x)^2]}/x$, is nothing but the square root of MSE normalized by a constant term, and thus MSE can capture a relative change of the estimation error computed by NRMSE. For the random walk-based crawlers, we use both MHRW and SRW with reweighting in our simulations. In each case, the initial position of each random walker is drawn from its stationary distribution. Each data point reported here is obtained by averaging over $10^4$ independent simulations.

***Estimating The Cost-based Asymptotic Variance:*** To obtain the cost-based asymptotic variance $\Psi(p)$ in terms of the sampling probability $p$, we first need to estimate all the involving terms such as $A$, $B$, $\alpha$, $\gamma$ as shown in Theorem 2 (or $\tilde{\alpha}$, $\tilde{\gamma}$ in Theorem 3). Estimating $A$ and $B$ (the expected costs with respect to $\boldsymbol{\pi}$) is relatively a trivial task if we run the chain over $\mathcal{G}$ and collect a sequence of sample values for the cost. Their sample mean would serve the purpose. $\gamma$ is nothing but the variance of function $f$ with respect to the stationary distribution $\boldsymbol{\pi}$, and thus can be approximated by observed sample variance. In addition, note that $v(f, \mathbf{P}_0, \boldsymbol{\pi})$ in $\gamma$ is the asymptotic variance of the estimator when sampling operation is always performed ($p = 1$), and thus $\alpha$ can also be well estimated.

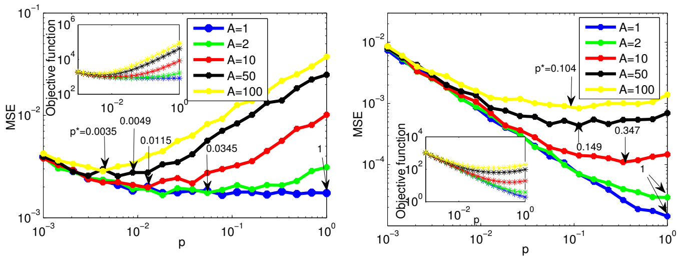
Estimating $\tilde{\gamma}$ and $\tilde{\alpha}$ would be more subtle, but the fundamental mechanism remains the same. Again, the training period would be based on a crawler under policy $\mathcal{P}(1)$. A sequence of training samples can serve the purpose to estimate $\mathbb{E}_{\mathbf{u}}(f)$ by appropriately re-weighting each sample and taking their sample mean, as illustrated in Section III-D. Then, estimating the sample variance with respect to the function $\mathcal{H}$ and the asymptotic variance $v(\mathcal{H}, \mathbf{P}_0, \boldsymbol{\pi})$ would help finding $\tilde{\gamma}$ and $\tilde{\alpha}$ in a similar way.

### B. Simulation Results

Due to space constraint, we here mainly present our simulation results for Youtube graph. We observe similar trends for other graphs, which are reported in our technical report [29].

***State-Independent Sampling:*** In Figure 2(a), we compare the MSE of the estimator $\hat{\mu}_T(f)$ in (9) with our cost-based asymptotic variance $\Psi(p)$ in Theorem 2 when the underlying chains are MHRW and the function of interest is $f(i) = \mathbf{1}_{\{i \in D\}}$. The total budget is $M = 10^6$. We first run crawlers over the graph in order to estimate $v(f, \mathbf{P}_0, \mathbf{u})$, $\mathrm{Var}_{\mathbf{u}}(f)$ and $A$,

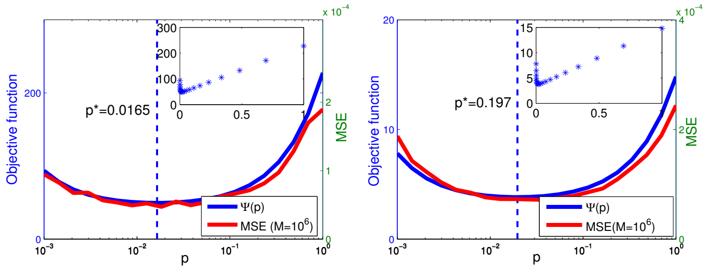
[‡] In this work, we use an undirected version of this graph.



(a) MHRW (log-log)

(b) SRW (log-log)

Fig. 2. $\Psi(p)$ vs. MSE when estimating $\mathbb{P}\{i \in D\}$.



(a) MHRW (log-linear)

(b) SRW (log-linear)

Fig. 3. $\Psi(p)$ vs. MSE when estimating $\Phi$.

and hence $\Psi(p)$ as discussed above. We set the sampling cost $a(i) = 1, 2, 10, 50, 100$ for all $i$ (hence $A = 1, 2, 10, 50, 100$), and apparently $B = 1$ since the cost for each transition is 1. We observe that our cost-based objective function (inset figures) closely captures the trend of MSE for the estimators, and the optimal solution $p^*$ that minimizes $\Psi(p)$ matches well with the theory given in Proposition 1. When $A = 1$, MSE is always monotone decreasing, since the crawler will spend 1 unit of cost anyway to transit the nodes, thus skipping a node without collecting any sample will not bring any benefit. In order to evaluate the improvement of the optimal strategy $\mathcal{P}(p^*)$, we mainly use the percentage of decrease in the MSE of the estimator compared with that of $\mathcal{P}(1)$, i.e., $[\text{MSE}(\mathcal{P}(1)) - \text{MSE}(\mathcal{P}(p^*))]/\text{MSE}(\mathcal{P}(1))$. In case of the optimal policy with $p^* = 0.034, 0.011, 0.0049, 0.0035$ for $A$=2, 10, 50, 100, respectively, we observe the MSE of the estimator in (9) decreases 46.6%, 87.9%, 97% and 98.2% compared with MHRW ($p = 1$), respectively, which clearly shows significant improvement.

In Figure 2(b), we also compare the modified version of the cost-based asymptotic variance $\Psi(p)$ under SRW with reweighting (defined in Corollary 2) with the MSE of estimator $\hat{\mu}_T(f_w)/\hat{\mu}_T(w)$ as defined in (13), when estimating $f(i) = \mathbf{1}_{\{i \in D\}}$ with $M = 10^6$. The estimation error decreases by 18.4%, 33.7% and 37.7% compared with MHRW with the corresponding optimal $p^* = 0.35$, 0.15 and 0.1049 for $A$ = 10, 50, 100. For $A$ = 1, 2, the optimal solution is $p^* = 1$, as expected from Proposition 1. Another observation is that with the increase of $A$, the optimal solution $p^*$ is inclined to shift toward zero. One intuitive explanation is that when average cost is large, it would be more beneficial to spend some budget on breaking the ties between samples, because the cost spent on transitions, compared with collecting sample, is relatively minor.

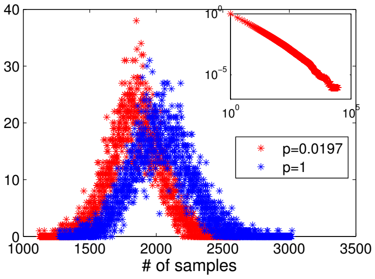


Fig. 4. # of samples with $M = 10^6$

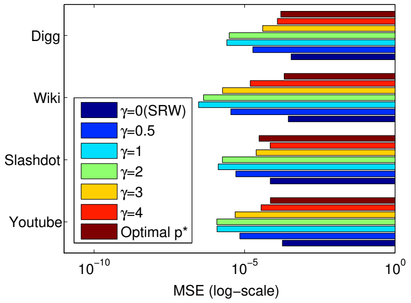


Fig. 5. State-dependent sampling ($p(i) = d^{-r}$).

In Figure 3, we repeat the same simulations for state-independent sampling with both MHRW and SRW with reweighting implemented, but change the sampling function of interest to estimate the average clustering coefficient $\Phi$. Again, we observe that, if properly re-scaled, $\Psi(p)$ is almost identical to the MSE of the estimator. We also calculate the optimal probability $p^*$ that minimizes $\Psi(p)$ according to Proposition 1. For MHRW (Figure 3(a)) and SRW with reweighting (Figure 3(b)), the optimal solutions $p^*$ are 0.0165 and 0.0197 and the optimal policies with these sampling rates result in approximately 70% and 68% reductions in the MSE compared with the case that sampling is always performed ($p = 1$), respectively. Note here the change is quite sharp when $p$ is close to 0, so we plot the figure with $x$-axis in log-scale. In order to check the convexity property that we find in Proposition 1, we also plot $\Psi(p)$ with $x$-axis in linear scale in the inset figure of Figure 3, which clearly justifies our theoretical findings.

For a better understanding on how the number of samples collected after all $M$ budget is spent (i.e., $T = T(M)$) varies with probability $p$, we present its empirical distribution over $10^4$ runs in Figure 4. When $p = 0.0197$, the mean value for the size of sample is approximately 1900, while it increases to about 2100 when $p = 1$. This again corroborates our motivation that the comparison between different sampling algorithms assuming same number of samples collected is usually unfair.

We repeat the same simulation over other graphs under both MHRW and SRW with reweighting. We observe similar trends and also good match between our cost-based asymptotic variance and the MSE of the estimators under a given budget. We refer to our technical report [29] for more details.

***State-Dependent Sampling:*** In Figure 5, we investigate how the state-dependent sampling can help in further improving the performance of the sampler when the quantity to estimate is the average clustering coefficient $\Phi$. Here we use SRW as the underlying chain $\mathbf{P}_0$ and heuristically set the probability for sampling each node $i$ to be $p(i) = d(i)^{-r}$, where $r = 0, 0.5, 1, 2, 3, 4$. Correspondingly, the weight function $w(i) = 1/(d(i)p(i)) = d(i)^{r-1}$, as mentioned in Section IV. Note that when $r = 0$, the sampling strategy degenerates to the usual SRW with reweighting. We observe that the sampler achieves the minimum estimation error at about $r = 1$ for these four graphs, with $99.3\%$, $98.1\%$, $99.8\%$ and $99.2\%$ reductions in MSE compared with SRW with reweighting for Youtube, Slashdot, Wiki talk and Digg graphs, respectively. We observe that the improvement is significant when estimating $\Phi$. This is reasonable, since SRW inclines to sample nodes with large degree more often, while at the same time sampling the large-degree nodes costs more. Recall that the sampling cost here is $a(i) = d(i)$. Thus, intuitively, the stack of these two factors greatly magnifies the average cost per sample. In other words, intentionally skipping those large-degree nodes with higher probability of $1 - p(i) = 1 - 1/d(i)$ while at the same time removing the bias by a reweighting procedure will bring significant benefit. We also compare the MSE of the estimators when $r = 1$ with the optimal solution for the state-independent sampling, and observe $98.3\%, 93.2\%, 99.7\%$ and $98.2\%$ decreases in MSE for the four graphs, respectively. We here note that, although sampling large-degree nodes less often may bring significant improvement in the estimation accuracy for *unbiased sampling*, the crawler might miss some opportunity of sampling 'important' nodes with large degrees. While their target goals are originally different, we expect that our general framework for cost-effective unbiased sampling strategies can also be integrated into such application scenarios where different nodes have different levels of importance.

## VI. Conclusion

We have provided a general, mathematical framework with a new "cost-based asymptotic variance" to compare different graph sampling strategies under the same cost constraint. After integrating "random skipping", which generates superior yet more expensive samples, into the popular random walk-based sampling, we were able to find an optimal sampling policy striking the right balance between sample quality and quantity, which in turn greatly improves over the original skip-free random walk sampling. We have further demonstrated that our framework is applicable to any random walk methods having a reweighting procedure for unbiased estimation, and also extended to state-dependent sampling policies, which still have room for improvement. We expect that our work provides a first step toward the correct understanding of graph sampling under practical cost-oriented scenarios, and also shed light on the design of more effective sampling strategies under cost constraints.

## References

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

[2] C. Hubler, H.-P. Kriegel, K. Borgwardt, and Z. Ghahramani, "Metropolis algorithms for representative subgraph sampling," in *IEEE ICDM*, 2008.

[3] M. A. Bhuiyan, M. Rahman, M. Rahman, and M. A. Hasan, "Guise: Uniform sampling of graphlets for large graph analysis," in *IEEE ICDM*, Dec. 2012.

[4] 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.

[5] A. Nazi, Z. Zhou, S. Thirumuruganathan, N. Zhang, and G. Das, "Walk, not wait: Faster sampling over online soical networks," *PVLDB*, vol. 8, pp. 678–689, Feb. 2015.

[6] A. Mislov, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *ACM IMC*, Oct. 2007.

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

[8] 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*, Jun. 2012.

[9] M. Papagelis, G. Das, and N. Koudas, "Sampling online social networks," *IEEE Trans. on Knowledge and Data Engineering*, vol. 25, no. 3, pp. 662–676, Mar. 2013.

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

[11] 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, Oct. 2011.

[12] C. Reuter and S. Scholl, "Technical limitations for designing applications for social media," in *Mensch & Computer 2014 - Workshopband*, 2014.

[13] "Twitter API," http://dev.twitter.com/rest/public/rate-limits.

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

[15] B. Ribeiro and D. Towsley, "On the estimation accuracy of degree distributions from graph sampling," in *IEEE CDC*, Dec. 2012.

[16] D. Figueiredo, P. Nain, B. Ribeiro, E. de Souza e Silva, and D. Towsley, "Characterizing continuous time random walks on time varying graphs," in *ACM SIGMETRICS*, Jun. 2012.

[17] A. H. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach, "Evaluating sampling techniques for large dynamic graphs," CIS-TR-08-01, University of Oregon, Tech. Rep., Sep. 2008.

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

[19] 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.

[20] 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.

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

[22] M. Kurant, M. Gjoka, C. T. Butts, and A. Markopoulou, "Walking on a graph with a magnifying glass: Stratified sampling via weighted random walks," in *ACM SIGMETRICS*, Jun. 2011.

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

[24] P. Rusmevichientong, D. M. Pennock, S. Lawrence, and L. C. Giles, "Methods for sampling pages uniformly from the world wide web," in *AAAI Fall Symposium*, Nov. 2001.

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

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

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

[28] X. Wang, R. T. B. Ma, Y. Xu, and Z. Li, "Sampling online social networks via heterogeneous statistics," in *IEEE INFOCOM*, Apr. 2015.

[29] X. Xu, C.-H. Lee, and D. Y. Eun, "Challenging the limits: Sampling online social networks with cost constraints," *Tech. Rep.*, Jul. 2016, https://www.dropbox.com/sh/rm9br6x4qc0f5d9/AAC441e4JK4iKiEkrv0tukwFa/techreport.pdf?raw=1.

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

[31] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," http://snap.stanford.edu/data, Jun. 2014.

[32] T. Hogg and K. Lerman, "Social dynamics of Digg," *EPJ Data Science*, vol. 1, no. 5, pp. 1–26, 2012.