Inference in OSNs via Lightweight Partial Crawls

Konstantin Avrachenkov, Bruno Ribeiro, Jithin Sreedharan

Inference in OSNs via Lightweight Partial Crawls

Konstantin Avrachenkov
INRIA
Sophia Antipolis,France
k.avrachenkov@inria.fr

Bruno Ribeiro
Dept. of Computer Science
Purdue University
West Lafayette, IN, USA
ribeiro@cs.purdue.edu

Jithin K. Sreedharan
INRIA
Sophia Antipolis,France
jithin.sreedharan@inria.fr

ABSTRACT

Are Online Social Network (OSN) A users more likely to form friendships with those with similar attributes? Do users at an OSN B score content more favorably than OSN C users? Such questions frequently arise in the context of Social Network Analysis (SNA) but often crawling an OSN network via its Application Programming Interface (API) is the only way to gather data from a third party. To date, these partial API crawls are the majority of public datasets and the synonym of lack of statistical guarantees in incomplete-data comparisons, severely limiting SNA research progress. Using regenerative properties of the random walks, we propose estimation techniques based on short crawls that have proven statistical guarantees. Moreover, our short crawls can be implemented in massively distributed algorithms. We also provide an adaptive crawler that makes our method parameter-free, significantly improving our statistical guarantees. We then derive the Bayesian approximation of the posterior of the estimates, and in addition, obtain an estimator for the expected value of node and edge statistics in an equivalent configuration model or Chung-Lu random graph model of the given network (where nodes are connected randomly) and use it as a basis for testing null hypotheses. The theoretical results are supported with simulations on a variety of real-world networks.

Keywords

Bayesian inference, Graph sampling, Random walk on graphs, Social network analysis

1. INTRODUCTION

What is the fraction of male-female connections against that of female-female connections in a given Online Social Network (OSN)? Is the OSN assortative or disassortative? Edge, triangle, and node statistics of OSNs find applications in computational social science (see e.g. [30]), epidemiology [31], and computer science [5, 15]. Computing these statistics is a key capability in large-scale social network analysis and machine learning applications. But because data collection in the wild is often limited to partial OSN crawls through Application Programming Interface (API) requests, observational studies of OSNs – for research purposes or market analysis – depend in great part on our ability to compute network statistics with incomplete data. Case in point, most datasets available to researchers in widely popular public repositories are partial OSN crawls[1]. Unfortunately, these incomplete datasets have unknown biases and no statistical guarantees regarding the accuracy of their statistics. To date, the best methods for crawling networks ([4, 14, 32]) show good real-world performance but only provide statistical guarantees asymptotically (i.e., when the entire OSN network is collected).

This work addresses the fundamental problem of obtaining unbiased and reliable node, edge, and triangle statistics of OSNs via partial crawling. *To the best of our knowledge our method is the first to provide a practical solution to the problem of computing OSN statistics with strong theoretical guarantees from a partial network crawl.* More specifically, we (a) provide a provable finite-sample unbiased estimate of network statistics (and their spectral-gap derived variance) and (b) provide the approximate posterior of our estimates that performs remarkably well in all tested real-world scenarios.

More precisely, let $G = (V, E)$ be an undirected labeled network – not necessarily connected – where V is the set of vertices and $E \subseteq V \times V$ is the set of edges. Unlike the usual definition of E where each edge is only present once, to simplify our notation we consider that if $(u, v) \in E$ then $(v, u) \in E$. Both edges and nodes can have labels. Network G is unknown to us except for $n > 0$ *arbitrary* initial seed nodes in $I_n \subseteq V$. Nodes in I_n must span all the different connected components of G. From the seed nodes we crawl the network starting from I_n and obtain a set of crawled edges $\mathcal{D}_m(I_n)$, where $m > 0$ is a parameter that regulates the number of website API requests. With the crawled edges $\mathcal{D}_m(I_n)$ we seek an unbiased estimate of

$$\mu(G) = \sum_{(u,v)\in E} g(u, v)\,, \tag{1}$$

for any function $g(u, v)$ over the node pair (u, v). Note that functions of the form eq. (1) are general enough to compute node statistics

$$\mu_{\text{node}}(G) = \sum_{(u,v)\in E} h(v)/d_v\,,$$

where d_u is the degree of node $u \in V$, $h(v)$ is any function of the node v, and statistics of triangles such as the local

[1]Public repositories such as SNAP [26] and KONECT [24] contain a majority of partial website crawls, not complete datasets or uniform samples.

clustering coefficient of G first provided by [32]

$$\mu_{\triangle}(G) = \frac{1}{|V|}\sum_{(u,v)\in E}\Big\{\frac{\mathbf{1}(d_v>2)}{d_v} \frac{\sum_{a\in N_v}\sum_{b\in N_v, b\neq a}\mathbf{1}((v,a)\in E\cap(v,b)\in E\cap(a,b)\in E)}{\binom{d_v}{2}}\Big\},$$

where the expression inside the sum is zero when $d_v < 2$ and N_v are the neighbors of $v \in V$ in G. Our task is to find estimates of general functions of the form $\mu(G)$ in eq. (1).

Contributions

In our work we provide a partial crawling strategy using short dynamically adjustable random walk tours starting at a "virtual" super-node without invoking the notion of lumpability [21]. A random walk tour is a random walk sample path that starts and ends at the same node on the graph. We use these tours to compute a frequentist unbiased estimator of $\mu(G)$ (including its variance) regardless of the number of nodes, $n > 0$, in the seed set and regardless of the value of $m > 0$, unlike previous asymptotically unbiased methods [4, 14, 25, 32, 33]. We also provide a Bayesian approximation of the posterior of $\mu(G)$ given the observed tours $P[\mu(G)|\mathcal{D}_m(I_n)]$, which is shown to be consistent. In our experiments we note that the posterior is remarkably accurate using a variety of networks large and small. Furthermore, when the network is formed by randomly wiring connections while preserving degrees and attributes of the nodes in the observed network, we devise an estimation technique for the expected true value with partial knowledge of the original graph.

Related Work

The works of Massoulié et al. [28] and Cooper et al. [10] are the ones closest to ours. Massoulié et al. [28] estimates the size of a network based on the return times of random walk tours. Cooper et al. [10] estimates number of triangles, network size, and subgraph counts from weighted random walk tours using results of Aldous and Fill [1, Chapter 2 and 3]. The previous works on finite-sample inference of network statistics from incomplete network crawls [16, 22, 23, 18, 19, 27, 34] need to fit the partial observed data to a probabilistic graph model such as ERGMs (exponential family of random graphs models). Our work advances the state-of-the-art in estimating network statistics from partial crawls because: (a) we estimate statistics of arbitrary edge functions without assumptions about the graph model or the underlying graph; (b) we do not need to bias the random walk with weights as in Cooper et al.; this is particularly useful when estimating multiple statistics reusing the same observations; (c) we derive upper and lower bounds on the variance of estimator, which both show the connection with the spectral gap; and, finally, (d) we compute a posterior over our estimates to give practitioners a way to access the confidence in the estimates without relying on unobtainable quantities like the spectral gap and without assuming a probabilistic graph model.

The remainder of the paper is organized as follows. In Section 2 we introduce key concepts and defines the notation used throughout this manuscript. Section 3 presents the algorithms to build the super-node and proves the equivalence between them. The frequentist estimators and their properties are explained in Section 4. Section 5 contains the main result of the posterior distribution in Bayesian framework. Section 6 consists of experimental results over real-world networks. Finally, in Section 7 we present our conclusions.

2. SUPER-NODE RATIONALE

In this section we present definitions and concepts using throughout the reminder of the paper. Then we substantiate an intuitive reasoning that our random walk tours are shorter than the "regular random walk tours" because the "node" that they start from is an amalgamation of a multitude of nodes in the graph.

Preliminaries

Let $G = (V, E)$ be an unknown undirected graph with N number of nodes and $M = |E|/2$, number of edges. Our goal is to find an unbiased estimate of $\mu(G)$ in eq. (1) and its posterior by crawling a small fraction of G. We are given a set of $n > 0$ initial *arbitrary* nodes denoted $I_n \subset V$. If G has disconnected components I_n must span all the different connected components of G.

Unless stated otherwise our network crawler is a classical random walk (RW) over the following augmented multigraph $G' = (V', E')$ with N' nodes and M' edges. A multigraph is a graph that can have multiple edges between two nodes. In $G'(I_n)$ we aggregate all nodes of I_n into a single node, denoted hereafter S_n, the *super-node*. Thus, $V'(I_n) = \{V\backslash I_n\} \cup \{S_n\}$. The edges of $G'(I_n)$ are $E'(I_n) = E\backslash\{E \cap \{I_n \times V\}\} \cup \{(S_n, v) : \forall(u, v) \in E$, s.t. $u \in I_n$ and $v \in V\backslash I_n\}$, i.e., $E'(I_n)$ contains all the edges in E including the edges from the nodes in I_n to other nodes, and I_n is merged into the super-node S_n. Note that $G'(I_n)$ is necessarily connected as I_n spans all the connected components of G. For compactness of notation we sometimes refer to $G'(I_n)$ as G' when I_n is clear from the context. *We also use S_n and I_n interchangeably to denote both the super-node at $G'(I_n)$ and each individual nodes of I_n at G.*

A random walk on $G'(I_n)$ has transition probability from node u to an adjacent node v, $p_{uv} := \alpha_{u,v}/d_u$, where d_u is the degree of u and $\alpha_{u,v}$ is the number of edges between $u \in V'$ and $v \in V'$. Let $\mathbf{P} = \{p_{uv}\}$. We note that the theory presented in the paper can be extended to more sophisticated random walks as well. Let π_i be the stationary distribution at node i in the random walk on $G'(I_n)$.

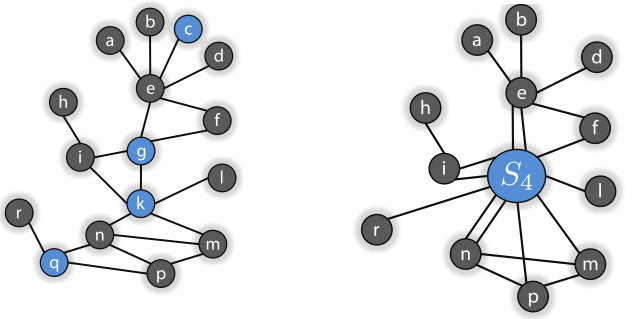
A random walk *tour* is defined as the sequence of nodes $X_1^{(k)}, \ldots, X_{\xi_k}^{(k)}$ visited by the random walk during successive k-th and $k+1$-st visits to the super-node S_n. Here $\{\xi_k\}_{k\geq 1}$ denote the successive return times to S_n. Tours have a key property: from the renewal theorem tours are independent since the returning times act as renewal epochs. Moreover, let $Y_1, Y_2, \ldots, Y_n$ be a random walk on $G'(I_n)$ in steady state.

Note that the random walk on $G'(I_n)$ is equivalent to a random walk on G where all the nodes in I_n are treated as *one single node*. Figure 1 shows an example of the formation of $G'(I_n)$.

Why a Super-node

The introduction of super-node is primary motivated by the following closely-related reasons:

- *Tackling disconnected or low-conductance graphs:* When the graph is not well connected or has many connected



(a) The original graph G with $I_4 = \{c, g, j, q\}$.

(b) The modified graph $G'(I_n)$ with super-node S_4.

Figure 1: Graph modification with the super-node

components, forming a super-node with representatives from each of the components make the modified graph connected and suitable for applying random walk theory. Even when the graph is connected, it might not be well-knit, i.e., it has low conductance. Since the conductance is closely related to mixing time of Markov chains, such graph will prolong the mixing of random walks. But with proper choice of super-node, we can reduce the mixing time and, as we show, improve the estimation accuracy.

If we consider self loops from S_n to itself while forming the modified graph G', i.e. all the connections between I_n ($E \cap \{I_n \times I_n\}$), then G' becomes a contracted graph [9]. Then [9, Lemma 1.15] says that if S_n is formed from $n = 2$ vertices, the spectral gaps of the two graphs $\delta_{G'}$ and δ_G are related as follows: $\delta_{G'} \geq \delta_G$. The spectral gap $\delta_{G'} = 1 - \lambda_2$, where λ_2 is the second largest eigenvalue of $\mathbf{P}$, and δ_G can be defined accordingly on the random walk on the original graph. The above argument with spectral gaps can be extended to $n > 2$ by induction and hence $\delta_{G'} \geq \delta_G$ follows. The improvement in the spectral gap proves that the modified graph will become well-knit (low conductance). Note that G' in our case does not involve self loops around S_n, but this is for the ease of computation as the function values over the self loops are known (from the knowledge of I_n and further API queries with them), and hence allowing self loops will only slower down the estimation of $\mu(G)$ outside the super-node.

- *No lumpability for random walks:* The theory of lumpability [21, Section 6.3] provides ways to exactly partition a Markov chain. Unfortunately, lumping states in a classical random walk will not give an accurate representation of the random walk Markov chain and, thus, we consider a super-node S_n where all the nodes inside the super-node are merged into one node rather than partitioning the states. The edges from S_n are the collection of all the edges from the nodes inside the super-node which are connected to nodes outside super-node, and Markov chain property still holds on this formation. The graph modification with S_n is illustrated with an example in Figure 1.

- *Faster estimate with shorter tours:* The expected value of the k-th tour length $\mathbb{E}[\xi_k] = 1/\pi_{S_n}$ is inversely proportional to the degree of the super-node d_{S_n}. Hence, by forming a massive-degree super-node we can significantly shorten the average tour length. This property is of great practical importance as it reduces the number of API queries required per tour.

3. STATIC AND DYNAMIC SUPER-NODES

In what follows we describe the algorithms to build super-nodes. The static super-node technique selects the nodes I_n before starting the experiment, while the dynamic super-node recruits the nodes on the fly.

3.1 Static Super-node Algorithm

The static super-node is selected by n nodes from G without replacement to form I_n. If the graph is disconnected, I_n must contain at least one node of each component of interest. To construct I_n we can crawl each component of the graph G. For instance, one can make I_n be the n largest degree nodes seen in a set of random walks with a total of $k > n$ steps (as in Avrachenkov et al. [3]). Because random walks are biased towards large-degree nodes the resulting super-node S_n tends to have large degrees.

Once I_n is chosen we form the virtual graph $G'(I_n)$ and start m random walk tours from the virtual super-node S_n. We stop each tour when the walk comes back to S_n. One practical issue in building I_n is knowing how many nodes we need to recruit to keep the random walk tours short. To ameliorate the situation in what follows we consider a dynamic algorithm to select the super-node.

3.2 Dynamic Super-node Algorithm

In a dynamic super-node, nodes are added into the super-node *on-demand* using a different random walk called the super-node recruiting walk. The super-node S_j starts with $j \geq 1$ nodes. S_j must span nodes in all graph components. The algorithm is as follows:

1. Run a super-node recruiting walk independent of all previous tours starting from S_n, $n \geq j$. Once a node of interest i, or set of nodes, are reached, stop the super-node recruiting walk.

2. Add a newly recruited node i to the super-node S_n, $n \geq j$, $S_{n+1} = S_n \cup \{i\}$. If node i appears in any of the previous tours, break these tour into multiple tours where i either ends or starts a new tour.

3. Generate a random number k_{redo} from the negative binomial distribution with number of successes as the number of previous tours (not counting the broken tours) and probability of success $d_{S_n}/d_{S_{n+1}}$, where $d_{S_{n+1}}$ is the degree of the new super-node that includes i and d_{S_n} is the degree of the super-node without i.

4. Perform $k_{\text{redo}} - k_{\text{broken}} > 0$ tours, where k_{broken} is the number of broken tours that start with node i and have length greater than two. These tours start at node i in $G'(S_i)$ with a first step into nodes in $N(i) \backslash S_{i+1}$, where $N(i)$ are the neighbors of i in $G'(S_i)$. The tour ends at either S_n or i. Note that in $G'(S_{i+1})$ these tours start at S_{i+1} and end at S_{i+1} with length greater than two. If $k_{\text{redo}} - k_{\text{broken}} < 0$ then randomly remove tours starting at i until only k_{redo} tours remain.

5. Redo steps 2–4 until all recruited nodes are added to the super-node.

6. We can now proceed normally with new super-node tours (or recruit more nodes if necessary by redoing steps 1–4).

The step 4 calculates the number of tours that might have happened in the past when the new node i was part of S_n. This retrospective event can be recreated by sampling from a negative binomial distribution with appropriate parameters.

3.3 Equivalence Between Dynamic and Static Super-node Sample Paths

In what follows we show that the tours of a dynamic super-node S_n^{dyn} and the tours of the same super-node as a static super-node have the same probability distribution.

Theorem 1. *Let $\mathcal{D}_m^{(dyn)}$ denote the set of m tours according to the super-node dynamic algorithm over $n \geq 1$ steps, resulting in super-node S_n and $\mathcal{D}_m^{(st)}$ denote the set of m tours according to the static super-node algorithm using super-node S_n. The dynamic super-node sample paths and the static super-node sample paths are equivalent in distribution, that is, $P[\mathcal{D}_m^{(dyn)}(S_n) = Q] = P[\mathcal{D}_m^{(st)}(S_n) = Q]$, $n \geq 1, \forall S_n \subset V, \forall Q$, where $m > 1$ is the number of tours.*

Proof. We prove by induction. Let $\sigma(\omega, S, \mathcal{E})$ be a deterministic function that is given an infinite random vector ω, where $\omega(1), \omega(2), \ldots \sim \text{Uniform}(0,1)$ are i.i.d. random variables, and a vector of starting nodes S and terminal nodes $\mathcal{E}$ as inputs and outputs a sample path of a random walk on the original graph G that starts at a node $u \in S$ with probability proportional to d_u and ends when it reaches any node in $\mathcal{E}$.

In what follows I_i denotes a set of i nodes as well as a vector of i nodes, $i \geq 1$. We add an arbitrary node outside I_i, $v \in V \backslash I_i$, into the first position $I_{i+1} = (v, I_i)$ and consider the deterministic sample path function:

$$\sigma^{(\text{dyn})}(\omega, I_i, v) = \begin{cases} (v, \sigma') & \text{, if } \omega(1) \leq d_v/\text{vol}(I_{i+1}) \\ \sigma(\omega', I_i, I_{i+1}) & \text{, otherwise,} \end{cases}$$

where $\text{vol}(S) = \sum_{t \in S} d_t$, $\sigma' = \sigma((\omega(2), \ldots), \{v\}, I_{i+1})$, $\omega' = ((\omega(1) - p_v)/(1 - p_v), \omega(2), \ldots)$, with $p_v = d_v/\text{vol}(I_{i+1})$. Note that by construction

$$\mathcal{D}_m^{(\text{st})}(I_i) = \{\sigma(\omega_k, I_i, I_i) : k = 1, \ldots, m, |\sigma(\omega_k, I_i, I_i)| > 2\}$$

and if we aggregate the nodes I_i into a single super-node S_i these are independent sample paths of a random walk on the super-node graph $G'(I_i)$ starting from super-node S_i. Similarly, if the choice of nodes in I_i are independent of the random vectors $\{\omega_k\}_{k=1}^m$ then

$$\mathcal{D}_m^{(\text{dyn})}(I_i) = \{r : r := \sigma^{\text{dyn}}(\omega_k, I_{i-1}, u), k = 1, \ldots, m, |r| > 2\},$$

where $u \in I_i \backslash I_{i-1}$, and $\mathcal{D}_m^{(\text{dyn})}(I_i)$ are the sample paths of the random walk described by the dynamic super-node algorithm with node addition sequence I_i.

Our proof is by induction on i. For $S_1 = \{v\}$, $v \in V$ it is clear that $\sigma^{\text{dyn}}(\omega, \emptyset, v) = \sigma(\omega, I_1, I_1)$, $\forall \omega$. By induction assume that $\sigma^{\text{dyn}}(\omega, I_{i-1}, u) = \sigma(\omega, I_i, I_i)$, $\forall \omega$, $i \geq 2$ and $u \in I_i \backslash I_{i-1}$. By construction the only possible difference between the sample paths of σ and σ^{dyn} is how they select the sample paths starting with u, the first node in the vector I_i. But by our induction assumption these two deterministic functions are equivalent and u is selected with probability $d_u/\text{vol}(I_i)$. Thus, using the same deterministic rule σ selects v, the first element of vector I_{i+1}, with probability $d_v/\text{vol}(I_{i+1})$ making $\sigma^{\text{dyn}}(\omega, I_i, v) = \sigma(\omega, I_{i+1}, I_{i+1})$ and yielding

$$P[\mathcal{D}_m^{(\text{dyn})}(I_n) = Q] = P[\mathcal{D}_m^{(\text{st})}(I_n) = Q], \quad \forall n, I_n, Q.$$

To finish the proof note that for $v \in I_{i+1} \backslash I_i$, the the deterministic rule σ guarantees that we are selecting tours starting from v with probability $d_v/\text{vol}(I_{i+1})$. This rule is equivalent to the dynamic super-node algorithm that starts k_{redo} tours from v once v is added to the super-node, where k_{redo} is a negative binomial random variable with success probability $\text{vol}(I_i)/\text{vol}(I_{i+1})$. The success probability in the algorithm is $d_{I_i}/d_{I_{i+1}}$ because by definition the algorithm, like our definition of $\mathcal{D}_m^{(\text{dyn})}(I_{i+1})$, disregards tours of size two which only contain nodes in I_{i+1}. □

4. FREQUENTIST APPROACH

In what follows we present our main results for the estimators.

4.1 Estimator of $\mu(G)$

Theorem 2 proposes an unbiased estimator of $\mu(G)$ in equation (1) via random walk tours. Later in Section 5 we present the approximate posterior distribution of this unbiased estimator.

To compute an estimate of $\mu(G)$ using super-node tours we define a function f and a set H over the modified graph $G'(I_n)$ as follows.

(a) If $\forall (u,v) \in E$, s.t. $u \in I_n$, $v \in V \backslash I_n$ the function $g(u,v)$ can be obtained with little overhead (using extra API queries to find all the neighbors of $u \in I_n$ and further querying them to find the function $g(u,v)$), then we define f as

$$f(u,v) = \begin{cases} g(u,v) & \text{, if } u \neq S_n, v \neq S_n \\ 0 & \text{, if } u \text{ or } v = S_n\,. \end{cases} \tag{2}$$

Define $H = \{(u,v) : (u,v) \in E \text{ s.t. } u \in I_n \textbf{ or } v \in I_n\}$.

(b) Otherwise we define f as

$$f(u,v) = \begin{cases} g(u,v) & \text{if } u \neq S_n, v \neq S_n \\ \frac{1}{k_{xS}} \sum_{w \in I_n} g(u,w) & \text{if } u \text{ or } v = S_n, \end{cases} \tag{3}$$

where $x = u$ if $u \neq S_n, x = v$ if $v \neq S_n$ and k_{yS} is the number of neighbors of node $y \in V \backslash I_n$ that are in I_n. Define $H = \{(u,v) : (u,v) \in E \text{ s.t. } u, v \in I_n\}$.

Let $\mu(G')$ be contribution from G' to the true value $\mu(G)$ and $\mu(G') = \sum_{(u,v) \in E'} f(u,v)$.

Theorem 2. *Let G be an unknown undirected graph where $n > 0$ initial* arbitrary *set of nodes is known $I_n \subseteq V$ which span all the different connected components of G. Consider a random walk on the augmented multigraph G' described in Section 2 starting at super-node S_n. Let $(X_t^{(k)})_{t=1}^{\xi_k}$ be the k-th random walk tour until the walk first returns to S_n and let $\mathcal{D}_m(S_n)$ denote the collection of all nodes in $m \geq 1$ such*

tours, $\mathcal{D}_m(S_n) = \left((X_t^{(k)})_{t=1}^{\xi_k}\right)_{k=1}^{m}$. *Then,*

$$\hat{\mu}(\mathcal{D}_m(S_n)) = \overbrace{\frac{d_{S_n}}{2m}\sum_{k=1}^{m}\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})}^{\text{Estimate from crawls}} + \overbrace{\sum_{(u,v)\in H} g(u,v)}^{\substack{\text{Given knowledge}\\ \text{from nodes in } I_n}}, \tag{4}$$

is an unbiased estimate of $\mu(G)$*, i.e.,* $\mathbb{E}[\hat{\mu}(\mathcal{D}_m(S_n))] = \mu(G)$*. Moreover the estimator is strongly consistent, i.e.,* $\hat{\mu}(\mathcal{D}_m(S_n)) \to \mu(G)$ *a.s. for* $m \to \infty$*.*

The proof of Theorem 2 is in Appendix A. Theorem 2 provides an unbiased estimate of network statistics from random walk tours. The length of tour k is short if it starts at a massive super-node as the expected tour length is inversely proportional to the degree of the super-node, $\mathbb{E}[\xi_k] \propto 1/d_{S_n}$. This provides a practical way to compute unbiased estimates of node, edge, and triangle statistics using $\hat{\mu}(\mathcal{D}_m(S_n))$ (eq. (4)) while observing only a small fraction of the original graph. Because random walk tours can have arbitrary lengths, we show in Lemma 2, Section 4.4, that there are upper and lower bounds on the variance of $\hat{\mu}(\mathcal{D}_m(S_n))$. For a bounded function f, the upper bounds are shown to be always finite.

4.2 Confidence interval of the estimator

In what follows we give confidence intervals for the estimator presented in Theorem 2. Let

$$\bar{f}_m = m^{-1}\sum_{k=1}^{m}\left(\frac{d_{S_n}}{2}\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})\right),$$

$$\hat{\sigma}_m^2 = m^{-1}\sum_{k=1}^{m}\left(\frac{d_{S_n}}{2}\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)}) - \bar{f}_m\right)^2.$$

If $\Phi(x)$ is the CDF of the standard Gaussian distribution, then for a known constant $c > 0$ [6]:

$$\sup_x \left| P\left\{\sqrt{m}\left(\frac{\bar{f}_m - \mu(G')}{\hat{\sigma}_m}\right) < x\right\} - \Phi(x)\right| \le \frac{c\beta}{\sigma^3\sqrt{m}}, \tag{5}$$

where

$$\beta = \mathbb{E}\left[\left|\frac{d_{S_n}}{2}\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)}) - \mu(G')\right|^3\right],$$

$$\sigma^2 = \mathrm{Var}\left[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})\right].$$

Moreover, $\sigma^2 < \infty$ for a bounded function f as we will prove in Lemma 2(i) in Section 4.4 and $\beta < \infty$ through the C_r inequality [17, Chapter 3, Theorem 2.2].

Therefore, with $\epsilon > 0$ and large m (number of tours) within the confidence interval $[\hat{\mu}(\mathcal{D}_m(S_n)) - \epsilon, \hat{\mu}(\mathcal{D}_m(S_n)) + \epsilon]$ yields

$$P\left(|\mu(G) - \hat{\mu}(\mathcal{D}_m(S_n))| \le \epsilon\right) \approx 1 - 2\Phi\left(\frac{\epsilon\sqrt{m}}{\hat{\sigma}_m}\right).$$

with the rate of convergence given by equation (5).

4.3 Estimation and hypothesis testing in random graph models

Here we study $\mu(G)$ when the connections in graph G are made randomly while keeping the node attributes and node degrees the same as the observed graph. Two types of random graph generation are considered here: configuration model and Chung-Lu model. These models can be regarded as null hypotheses in graph hypothesis testing problem. First we estimate the expected value $\mathbb{E}[\mu(G)]$ in these random graph models. Then we seek, *with how much certainty* the value $\mu(G)$ of the observed graph could possibly belong to a random network with the same node attributes and degrees as that of the observed graph, all this with partial knowledge of the original graph.

4.3.1 Estimators for Configuration model and Chung-Lu random graphs

Configuration model is an important class of random graph model. For a given degree sequence over the nodes, the configuration model creates random edge connections by uniformly selecting pairs of half edges. We assume that the number of nodes N and number of edges M are known (The estimation of N and M can be done explicitly, for instance using the techniques in [28] and [10]). The probability that the nodes u and v are connected in the configuration model is $d_u d_v/(2M-1)$ if $u \neq v$ and the probability of a self-edge from node u to itself is $\binom{d_u}{2}/(2M-1)$.

Another important model, Chung-Lu random graph [8] is a generalized version of Erdős-Renyi graphs and is closely related to configuration model. Chung-Lu model takes the positive weights $w_1, \ldots, w_N$ corresponding to nodes $1, \ldots, N$ as input and generates a graph with average degrees as these weights. The edges are created between any two vertices u and v independently of all others with probability $w_u w_v / \sum_{k=1}^{N} w_k$, when $u \neq v$, and for the self loops at node u, with a probability $\binom{w_u}{2} / \sum_{k=1}^{n} w_k$.

In the case of Chung-Lu random graphs, from [2], it is known that the weights $w_1, \ldots, w_N$ in fact becomes the actual degrees $d_1, \ldots, d_N$ asymptotically and the following concentration bound exists: for $c > 0$,

$$\mathbb{P}\left(\max_{1\le i\le N}\left|\frac{d_i}{w_i} - 1\right| \ge \beta\right) \le \frac{2}{N^{c/4} - 1}, \text{ if } \beta \ge \frac{c\log N}{w_{\min}} = o(1).$$

Thus we take the sequence $\{w_k\}$ as $\{d_k\}$ of the observed graph. One main advantage in using Chung-Lu model compared to configuration model is that the edges are independent to each other.

For brevity we will use G_{obs} for the observed graph, G_{conf} for an instance of the configuration model with the same degree sequence $\{d_k\}$ as that of G and $G_{\text{C-L}}$ for the Chung-Lu graph sample with weights as $\{d_k\}$. Note that $\mu(G_{\text{conf}})$ and $\mu(G_{\text{C-L}})$ are random variables. Thus we look for $\mathbb{E}[\mu(G_{\text{conf}})]$ and $\mathbb{E}[\mu(G_{\text{C-L}})]$, where the expectation is with respect to the probability distribution of the configuration model and Chung-Lu model respectively. The values of $\mathbb{E}[\mu(G_{\text{C-L}})]$ and $\mathbb{E}[\mu(G_{\text{conf}})]$ are nearly the same, but for higher moments the values are different since configuration model introduces correlation between edges.

Now the expected value in the Chung-Lu model is

$$\mathbb{E}[\mu(G_{\text{C-L}})] = \sum_{\substack{(u,v)\in E\cup E^c\\ u\neq v}} g(u,v)\,\frac{d_u d_v}{2M} + \sum_{\substack{(u,v)\in E\cup E^c\\ u=v}} g(u,v)\,\frac{\binom{d_u}{2}}{2M}. \tag{6}$$

In order to calculate $\mathbb{E}[\mu(G_{\text{C-L}})]$, we need to know the missing edge set E^c. The set E^c is revealed once the entire graph is crawled, which is not possible in the context of this paper. The idea is to estimate $\mathbb{E}[\mu(G_{\text{C-L}})]$ from the tours of a random walk. Since the classical random walk which we have used so far, could sample only from E, we resort to a new random walk that could make samples from E^c as well.

We use random walk with uniform restarts (RWuR) [4] in which if the crawler is at a node i, with a probability $d_i/(d_i+\alpha)$ the crawler chooses one of the neighbors uniformly (RW strategy) and with a probability $\alpha/(d_i+\alpha)$, the crawler chooses the next node by uniformly sampling all the nodes. The parameter $\alpha > 0$ controls the rate of uniform sampling (which has higher cost in many OSNs).

Define a new function f' whose value depends on the crawling strategy as follows: let u, v be the nodes chosen by the crawling technique (RWuR or RW) in order,

$$f'(u,v) = \begin{cases} g(u,v)\frac{d_u d_v}{2M-1} & \text{if } u,v \neq S_n \\ \sum_{w\in I_n} g(u,w)\frac{d_u d_w}{2M-1} & \text{if } u \neq S_n, v = S_n \\ & (u,v) \text{ selected} \\ & \text{by unif. sampling} \\ \frac{1}{k_{uS}}\sum_{w\in I_n} g(u,w)\frac{d_u d_w}{2M-1} & \text{if } u \neq S_n, v = S_n \\ & (u,v) \text{ selected by RW} \end{cases} \tag{7}$$

where k_{uS} is defined in (3). In the new graph G' there will be k_{uS} multiple edges between u and S_n and k_{uS} is introduced in the last term is to take into account this. In case of classical random walk, the second criteria does not exist.

We denote $W'_k = \sum_{t=2}^{\xi_k} f'(X_{t-1}^{(k)}, X_t^{(k)})$ when RWuR is employed as the random walk technique for crawling the graph and $W''_k = \sum_{t=2}^{\xi_k} f'(X_{t-1}^{(k)}, X_t^{(k)})$, when the classical random walk is used for crawling. Let

$$R = \frac{1}{2}\sum_{\substack{(u,v)\in I_n\times I_n \\ u\neq v}} g(u,v)\frac{d_u d_v}{(2M-1)} + \frac{1}{2}\sum_{\substack{(u,v)\in I_n\times I_n \\ u= v}} g(u,v)\frac{\binom{d_u}{2}}{(2M-1)}.$$

The value R can be calculated a priori from the knowledge of I_n. In the lemma below we propose an estimator for $\mathbb{E}[\mu(G_{\text{C-L}})]$ and proves that it is unbiased.

Lemma 1. *The estimator*

$$\hat{\mu}_C(\mathcal{D}_m(S_n)) = \frac{1}{m}\sum_{k=1}^{m}\Big[\frac{N'(d_{S_n}+\alpha)}{2\alpha}W'_k - \frac{N' d_{S_n}}{2\alpha}W''_k + R\Big],$$

is an unbiased estimator of $\mathbb{E}[\mu(G_{C\text{-}L})]$ of the Chung-Lu model.

Proof. See Appendix A □

4.3.2 A hypothesis testing problem for the Chung-Lu model

The Chung-Lu model or configuration model can be regarded as a null hypothesis model and comparing $\mu(G_{\text{obs}})$ to $\mathbb{E}[\mu(G_{\text{C-L}})]$ or $\mathbb{E}[\mu(G_{\text{conf}})]$ answers many questions like whether the connections are formed based on degrees alone with no other influence or whether the edges are formed purely at random?

Let $G_{\text{C-L}}(V_{\text{C-L}}, E_{\text{C-L}})$ be a sample of the Chung-Lu model with weights $\{d_k\}$. Like the estimator of $\mathbb{E}[\mu(G_{\text{C-L}})]$, the estimator of $\text{Var}(G_{\text{C-L}})$, $\widehat{\text{Var}}_{\text{C-L}}(\mathcal{D}_m(S_n))$ can be constructed as follows: modify $g(u,v)\frac{d_u d_v}{2M}$ to $g^2(u,v)\frac{d_u d_v}{2M}\left(1-\frac{d_u d_v}{2M}\right)$ in the function f' in (7).

By invoking Lindeberg central limit theorem [17, Chapter 7, Section 2], for independent non-identically distributed Bernoulli random variables[2], we get

$$\sum_{(u,v)\in E_{\text{C-L}}} f(u,v) \sim \text{Normal}\left(\mathbb{E}[\mu(G_{\text{C-L}})], \text{Var}(G_{\text{C-L}})\right).$$

Hence a natural check is whether $\mu(G_{\text{obs}})$ is a sample from the above distribution. Since we have only one sample $\mu(G_{\text{obs}})$, a simple test is to check whether

$$|\hat{\mu}(\mathcal{D}(S_n)) - \hat{\mu}_C(\mathcal{D}(S_n))| \leq a\sqrt{\widehat{\text{Var}}_{\text{C-L}}(\mathcal{D}_m(S_n))},$$

holds for any $a = 1, 2, 3$ and for a large value of m . This condition is satisfied by a Gaussian sample with probabilities 0.6827, 0.9545, or 0.9973 respectively. On the other hand, the lower a is, the more certain that the sample belongs to this particular Gaussian distribution.

4.4 Impact of spectral gap on variance

In this section we derive results on higher moments of the estimator $\hat{\mu}(\mathcal{D}(S_n))$. Lemma 2, which follows, introduces upper and lower bounds on the variance of the i.i.d. the tour sum $\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})$, and also shows that all the moments exist. Moreover, the results in the lemma establish a connection between the estimator variance and the spectral gap.

Let $\mathbf{S} = \mathbf{D}^{1/2}\mathbf{P}\mathbf{D}^{-1/2}$, where $\mathbf{P} = \{p_{uv}\}$ is the random walk transition probability matrix as defined in Section 2 and $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_{|V'|})$ is a diagonal matrix with the node degrees of G'. The eigenvalues $\{\lambda_i\}$ of $\mathbf{P}$ and $\mathbf{S}$ are same and $1 = \lambda_1 > \lambda_2 \geq \ldots \geq \lambda_{|V'|} \geq -1$. Let jth eigenvector of $\mathbf{S}$ be $(w_{ji}), 1 \leq i \leq |V|$. Let δ be the spectral gap, $\delta := 1 - \lambda_2$. Let the left and right eigenvectors of $\mathbf{P}$ be v_j and u_j respectively. $d_{tot} := \sum_{v\in V'} d_v$. Define $\langle r, s\rangle_{\hat{\pi}} = \sum_{(u,v)\in E'} \hat{\pi}_{uv} r(u,v)s(u,v)$, with $\hat{\pi}_{uv} = \pi_u p_{uv}$, and matrix $\mathbf{P}^*$ with (j,i)th element as $p^*_{ji} = p_{ji}f(j,i)$. Also let $\hat{f}$ be the vector with $\hat{f}(j) = \sum_{i\in V'} p^*_{ji}$.

Lemma 2. *The following holds*

(i). Assuming the function f is bounded, $\max_{(i,j)\in E'} f(i,j) \leq B < \infty$, $B > 0$ and for tour $k \geq 1$,

$$\begin{aligned} &\text{Var}\left[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})\right] \\ &\leq \frac{1}{d_{S_n}^2}\left(2d_{tot}^2 B^2 \sum_{i\geq 2}\frac{w_{S_n i}^2}{(1-\lambda_i)} - 4\mu^2(G_{S_n})\right) \\ &\quad - \frac{1}{d_{S_n}}B^2 d_{tot} + B^2 \\ &< B^2\left(\frac{2d_{tot}^2}{d_{S_n}^2\delta} + 1\right). \end{aligned}$$

Moreover,

$$\mathbb{E}\left[\left(\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})\right)^l\right] < \infty \quad \forall l \geq 0.$$

[2]The necessary condition to hold this central limit theorem, the Lindeberg condition is satisfied by the sequence of independent Bernoulli random variables with different success probabilities $\{p_k\}$, if $0 < p_k < 1$. This is always true in our case when we assume $d_k > 0$ for all k.

(ii).

$$\begin{aligned}
&\mathrm{Var}\left[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)}))\right] \\
&\geq 2\frac{d_{tot}}{d_{S_n}}\sum_{i=2}^{r}\frac{\lambda_i}{1-\lambda_i}\langle f, v_i\rangle_{\hat{\pi}}\,(u_i^{\intercal}\hat{f}) + \frac{1}{d_{S_n}}\sum_{(u,v)\in E'} f(u,v)^2 \\
&+ \frac{1}{d_{tot}d_{S_n}}\left\{\sum_{(u,v)\in E'} f(u,v)^2\right\}^2 \\
&+ \frac{1}{d_{tot}d_{S_n}}\sum_{u\in V'} d_u\Big\{\sum_{u\sim v} f(u,v)\Big\}^2 \\
&- \frac{4}{d_{S_n}^2}\left\{\sum_{(u,v)\in E'} f(u,v)\right\}^2 \\
&- \frac{8}{d_{tot}}\left\{\sum_{(u,v)\in E'} f(u,v)\right\}^2\sum_{i\geq 2}\frac{w_{S_n i}^2}{(1-\lambda_i)} \\
&- \frac{4}{d_{tot}d_{S_n}}\left\{\sum_{(u,v)\in E'} f(u,v)\right\}^2. \qquad (8)
\end{aligned}$$

Proof. See Appendix B. □

5. BAYESIAN APPROACH

In this section we consider Bayesian formulation of our problem and derive the posterior of $\mu(G)$ given the tours and provide a consistent maximum a posteriori estimator (MAP).

Approximate posterior

For the same scenario of Theorem 2 for $m \geq 2$ tours let

$$\hat{F}_h = \frac{d_{S_n}}{2\lfloor\sqrt{m}\rfloor}\sum_{k=((h-1)\lfloor\sqrt{m}\rfloor+1)}^{h\lfloor\sqrt{m}\rfloor}\sum_{t=2}^{\xi_h} f(X_{t-1}^{(k)}, X_t^{(k)}) + \sum_{(u,v)\in H} g(u,v)\,,$$

which is similar to equation (4) but first sums a range of $\lfloor\sqrt{m}\rfloor$ tours rather than all m tours. Let σ_F^2 be the variance of $\hat{F}_h$. Assuming priors

$$\begin{aligned}
\mu(G)|\sigma_F^2 &\sim \text{Normal}(\mu_0, \sigma_F^2/m_0) \\
\sigma_F^2 &\sim \text{Inverse-gamma}(\nu_0/2, \nu_0\sigma_0^2/2),
\end{aligned}$$

then for large values of m, the marginal posterior density of $\mu(G)$ can be approximated by a *non-standardized t - distribution*

$$\phi(x|\nu,\widetilde{\mu},\widetilde{\sigma}) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\widetilde{\sigma}\sqrt{\pi\nu}}\left(1+\frac{(x-\widetilde{\mu})^2}{\widetilde{\sigma}^2\nu}\right)^{-\frac{\nu+1}{2}}, \qquad (9)$$

with degrees of freedom parameter

$$\nu = \nu_0 + \lfloor\sqrt{m}\rfloor,$$

location parameter

$$\widetilde{\mu} = \frac{m_0\mu_0 + \lfloor\sqrt{m}\rfloor\hat{\mu}(\mathcal{D}_m(S_n))}{m_0+\lfloor\sqrt{m}\rfloor},$$

and scale parameter

$$\widetilde{\sigma}^2 = \frac{\nu_0\sigma_0^2 + \sum_{k=1}^{\lfloor\sqrt{m}\rfloor}(\hat{F}_k - \hat{\mu}(\mathcal{D}_m(S_n)))^2 + \frac{m_0\lfloor\sqrt{m}\rfloor(\hat{\mu}(\mathcal{D}_m(S_n))-\mu_0)^2}{m_0+\lfloor\sqrt{m}\rfloor}}{(\nu_0+\lfloor\sqrt{m}\rfloor)(m_0+\lfloor\sqrt{m}\rfloor)}.$$

The derivation is detailed in Section 5.1 later.

Remark 1. Note that the approximation (9) is Bayesian and Theorem 2 is its frequentist counterpart. In fact, the motivation of our Bayesian approach comes from the frequentist estimator. From the approximate posterior in (9), the Bayesian MAP estimator is

$$\hat{\mu}_{\text{MAP}} = \arg\max_x \phi(x|v,\widetilde{\mu},\widetilde{\sigma}) = \widetilde{\mu}.$$

Thus for large values of m, the Bayesian estimator $\hat{\mu}_{\text{MAP}}$ is essentially the frequentist estimator $\hat{\mu}(\mathcal{D}_m(S_n))$, which is unbiased, and hence the MAP estimator is consistent.

The above remark shows that the approximate posterior in (9) provides a way to access the confidence in the estimate $\hat{\mu}(\mathcal{D}_m(S_n))$. The Normal prior for the average gives the largest variance given a given mean. The inverse-gamma is a non-informative conjugate prior if the variance of the estimator is not too small [13], which is generally the case in our application. Other choices of prior, such as uniform, are also possible yielding different posteriors without closed-form solutions [13].

Remark 2. Another asymptotic result in Bayesian analysis, the classic Bernstein-von Mosses Theorem [36, Chapter 10] is not useful in our scenario. The Bernstein-von Mosses Theorem states that irrespective of the prior distribution, when μ is the random parameter of likelihood, then posterior distribution of $\sqrt{m}(\mu - \hat{\mu}_m^{\text{MLE}})$ converges to Normal$(0, I(\mu_0)^{-1})$, where $\hat{\mu}_m^{\text{MLE}}$ is the maximum likelihood estimator (MLE) and $I(\mu_0)$ is the Fisher information at the true value μ_0. But note that in cases like ours, where the distribution of $W_k = \sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})$ is unknown, $k \geq 1$, the Fisher information is also unknown. In contrast, our approximate posterior of $\mu(G)$ uses only the available information and does not need to guess the distribution of W_k.

5.1 Derivation of approximate posterior

In this section we derive the approximation (9) of the posterior. The approximation relies first on showing that $\hat{\mu}(\mathcal{D}_m(S_n))$ has finite second moment. By Lemma 2 the variance of $\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})$, $k \geq 1$, is also finite.

We are now ready to give the approximation in equation (9). Let $m' = \lfloor\sqrt{m}\rfloor$,

$$\hat{F}_h = \frac{d_{S_n}}{2m'}\sum_{k=((h-1)m'+1)}^{hm'}\sum_{t=2}^{\xi_h} f(X_{t-1}^{(k)}, X_t^{(k)}) + \sum_{(u,v)\in H} g(u,v)\,.$$

and $\{\hat{F}_h\}_{h=1}^{m'}$ and because the tours are i.i.d. the marginal posterior density of μ is

$$P[\mu|\{\hat{F}_h\}_{h=1}^{m'}] = \int_0^\infty P[\mu|\sigma_F^2, \{\hat{F}_h\}_{h=1}^{m'}]P[\sigma_F^2|\{\hat{F}_h\}_{h=1}^{m'}]d\sigma_F^2\,.$$

For now assume that $\{\hat{F}_h\}_{h=1}^{m'}$ are i.i.d. normally distributed random variables, and let

$$\hat{\sigma}_{m'} = \sum_{h=1}^{m'}(\hat{F}_h - \hat{\mu}(\mathcal{D}_m(S_n)))^2,$$

then [20, Proposition C.4]

$$\mu|\sigma_F^2, \{\hat{F}_h\}_{h=1}^{m'} \sim \text{Normal}\left(\frac{m_0\mu_0 + \sum_{h=1}^{m'}\hat{F}_h}{m_0+m'}, \frac{\sigma_F^2}{m_0+m'}\right),$$

$$\sigma_F^2|\{\hat{F}_h\}_{h=1}^{m'} \sim \text{Inverse-Gamma}\Bigg(\frac{\nu_0+m'}{2}, \frac{\nu_0\sigma_0^2+\hat{\sigma}_{m'}+\frac{m_0m'}{m_0+m'}(\mu_0-\hat{\mu}(\mathcal{D}_m(S_n)))^2}{2}\Bigg),$$

are the posteriors of parameters μ and σ_F^2, respectively. The non-standardized t-distribution can be seen as a mixture of normal distributions with equal mean and random variance inverse-gamma distributed [20, Proposition C.6]. Thus, if $\{\hat{F}_h\}_{h=1}^{m'}$ are i.i.d. normally distributed then the posterior of $\mu(G)$ given $\mathcal{D}_{m'}(S_n)$ is a non-standardized t-distributed with parameters

$$\widetilde{\mu} = \frac{m_0\mu_0+\sum_{h=1}^{m'}\hat{F}_h}{m0+m'},$$
$$\widetilde{\sigma}^2 = \frac{\nu_0\sigma_0^2+\sum_{k=1}^{m'}(\hat{F}_k-\hat{\mu}(\mathcal{D}_m(S_n)))^2 + \frac{m_0m'(\hat{\mu}(\mathcal{D}_m(S_n))-\mu_0)^2}{m_0+m'}}{(\nu_0+m')(m_0+m')}, \quad \nu=\nu_0+m'$$

where $\widetilde{\mu}, \widetilde{\sigma}^2$ and ν are location, scale and degree of freedom parameters of the student-t distribution. Left to show is that $\{\hat{F}_h\}_{h=1}^{m'}$ converge *in distribution* to i.i.d. normal random variables as $m \to \infty$. As the spectral gap of $G'(I_n)$ is greater than zero, $|\lambda_1 - \lambda_2| > 0$, Lemma 2 shows that for $W_k = \sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})$, $\sigma_W^2 = \text{Var}(W_k) < \infty$, $\forall k$. From the renewal theorem we know that $\{W_k\}_{k=1}^{m}$ are i.i.d. random variables and thus any subset of these variables is also i.i.d.. By construction $\hat{F}_1, \ldots, \hat{F}_{m'}$ are also i.i.d. with mean $\mu(G)$ and finite variance. Applying the Lindeberg-Lévy central limit theorem [11, Section 17.4] yields

$$\sqrt{m'}(\hat{F}_h - \mu(G)) \xrightarrow{d} \text{Normal}(0, \sigma_W^2), \quad \forall h\,.$$

Thus, for large values of m (recall that $m' = \lfloor\sqrt{m}\rfloor$), $\{\hat{F}_h\}_{h=1}^{m'}$ are approximately i.i.d. normally distributed

$$\hat{F}_h \sim \text{Normal}(\mu(G), \sigma_W^2/m')\,, \quad \forall h\,.$$

This completes the derivation of the approximation (9). In what follows we present our results over real-world networks.

6. EXPERIMENTS ON REAL-WORLD NETWORKS

In this section we demonstrate the effectiveness of the theory developed above with the experiments on real data sets of various social networks. We have chosen to work with the datasets where the value $\mu(G)$ is available. This way it is possible to check the correctness of the results obtained via experiments. We assume the contribution from super-node to the true value is known a priori and hence we look for $\mu(G')$ in the experiments. In the case that the edges of the super-node are unknown, the estimation problem is easier and can be taken care separately. One option is to start multiple random walks in the graph and form connected subgraphs. Later, in order to estimate the bias created by this subgraph, do some random walk tours from the largest degree node in each of these sub graph and use the idea in Theorem 3.

In the figures we display both the approximate posterior generated from $\hat{F}_h$ with one only run of the experiment and empirical posterior created from multiple runs. For the approximate posterior, we have used the following parameters $m_0 =, \nu_0 = 0, \mu_0 = 0.1, \sigma_0 = 1$ (see (9)). The green line in the plots shows the actual value $\mu(G')$.

We have used the dynamic super-node algorithm explained in Section 3.2. From the experiments, it is observed that the static super-node and dynamic super-node produces similar results which is in corroboration with Theorem 1. In the following experiments, we opt a simple strategy to decide when to run super-node recruiting walk: run the super-node recruiting walk when the number of original tours reaches multiples of a fixed integer and it stops when a node of degree exceeding a specific threshold is reached.

6.1 Friendster

First we study a network of moderate size, a connected subgraph of Friendster network with 64,600 nodes and 1,246,479 edges (data publicly available at the SNAP repository [26]). Friendster is an online social networking website where nodes are individuals and edges indicate friendship. Here, we consider two types of functions:

1. $f_1 = d_{X_t}.d_{X_{t+1}}$ 2. $f_2 = \begin{cases} 1 \text{ if } d_{X_t} + d_{X_{t+1}} > 50 \\ 0 \text{ otherwise} \end{cases}$

These functions reflect assortative nature of the network. The final super-node size is 10,000. Figures 2 and 3 display the results for functions f_1 and f_2, respectively. A good match between the approximate and empirical posteriors can be observed from the figures. Moreover the true value $\mu(G')$ is also fitting well with the plots. The percentage of graph crawled is 7.43% in terms of edges and is 24.44% in case of static super-node with uniform sampling.

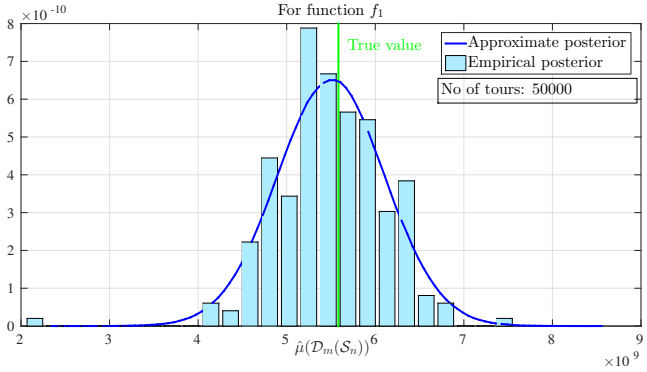


Figure 2: Friendster subgraph, function f_1

6.2 Dogster network

The aim of this example is to check whether there is any affinity for making connections between the owners of same breed dogs [12]. The network data is based on the social networking website Dogster. Each user (node) indicates the dog breed; the friendships between dogs' owners form the edges. Number of nodes is 415,431 and number of edges is 8,265,511.

In Figure 4, two cases are plotted. Function f_1 counts the number of connections with different breeds as pals and function f_2 counts connections between same breeds. The

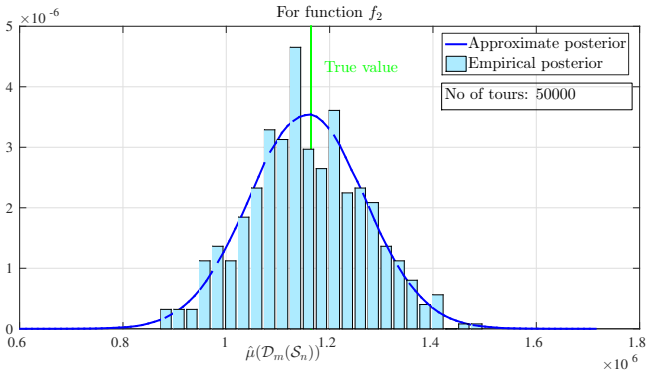


Figure 3: Friendster subgraph, function f_2

final super-node size is 10,000. The percentage of the graph crawled in terms of edges is 2.72% and in terms of nodes is 14.86%. While using the static super-node technique with uniform sampling, the graph crawled is 5.02% 2.72% (in terms of edges) and 37.17% (in terms of nodes) with the same super-node size. These values can be reduced much further if we allow a bit less precision in the match between approximate distribution and histogram.

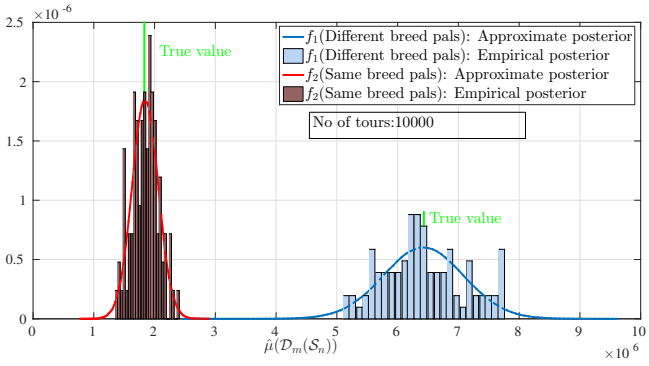


Figure 4: Dog pals network

In order to better understand the correlation in forming edges, we now consider the *configuration model*. We use the estimator $\hat{\mu}_C(\mathcal{D}(S_n))$ proposed in Section 4.3.1 (note that the estimator is same for Chung-Lu model and configuration model). It is important to recollect that this estimator does not require the knowledge of the complete network (in fact the Figures 5 and 6 are based on estimates from RWuR crawls which covered 8.9% of the graph). This is shown in blue line in Figure 5 and red line in Figure 6, and the true value is the net expected value given by (6). Moreover we run our original estimator $\hat{\mu}(\mathcal{D}(S_n))$ and calculated the approximate posterior on one random instance of the configuration model with same degree sequence of the original graph. Figure 5 compare function f_2 for the configuration model and original graph. The figure shows that in the correlated case (original graph), the affinity to form connection between same breed owners is around 7.5 times more than that in the uncorrelated case. Figure 6 shows similar figure in case of f_1.

6.3 ADD Health data

Though our main result in the approximation (9) holds

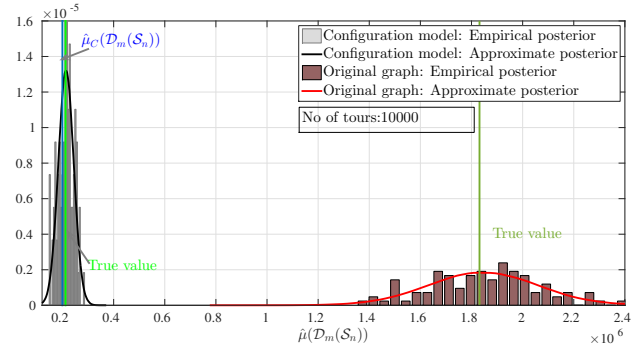


Figure 5: Dog pals network: Comparison between configuration model and original graph for f_2, number of connection between same breeds

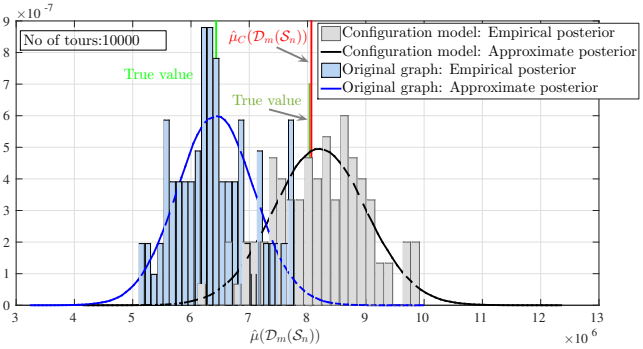


Figure 6: Dog pals network: Comparison between configuration model and original graph for f_1, number of connection between different breeds

when the number of tours is large, in this section we check with a small dataset. We consider ADD network project[3], a friendship network among high school students in US. The graph has 1545 nodes and 4003 edges.

We take two types of functions. Figure 7 shows the affinity in same gender or different gender friendships and Figure 8 displays the inclination towards same race or different race in friendships. The random walk tours covered around 10% of the graph. We find that the theory works reasonably well for this network data, for instance, the true values in both the cases in Figure 7 are nearly the same, and this is evident from the approximate posterior calculated from only one run. We have not added the empirical posterior in the figures since for such small sample sizes, the empirical distribution does not lead to a meaningful histogram.

6.4 Check for Chung-Lu random graph model in Dogester

We use the same dataset and functions as in Section 6.2. Consider the function f_1, which is one when the owners of different breed dogs form connection, zero otherwise. For the Chung-Lu model, $\hat{\mu}_C(\mathcal{D}(S_n))$, the estimator of $\mathbb{E}[\mu(G_{\text{C-L}})]$ is 8.066×10^6 and $\widehat{\text{Var}}_C(\mathcal{D}(S_n))$, the estimator of $\text{Var}_C[\mu(G_{\text{C-L}})]$ is 6.3938×10^{11}. For the original graph, the estimated value

[3]http://www.cpc.unc.edu/projects/addhealth

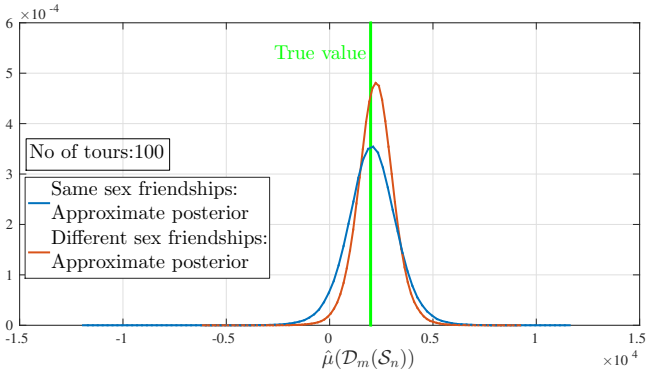


Figure 7: ADD network: effect of gender in relationships

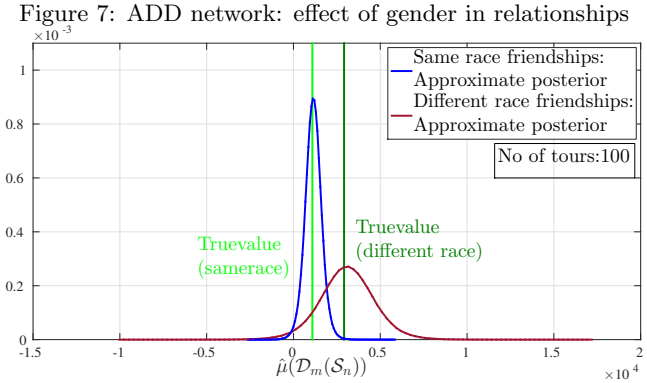


Figure 8: ADD network: effect of race in friendships

$\mu(\mathcal{D}(S_n)) = 6.432 \times 10^6$. Now

$$|\hat{\mu}_C(\mathcal{D}(S_n)) - \mu(\mathcal{D}(S_n))| \leq a\sqrt{\widehat{\mathrm{Var}}_C(\mathcal{D}(S_n))},$$

is satisfied for $a = 3$, but not for $a = 1, 2$. This implies there is a slight probability (0.0428) that $\mu(G)$ belongs to the values from different configurations of Chung-Lu model.

For function f_2, which is one when the owners of same breed dogs form connection, zero otherwise, $\hat{\mu}_C(\mathcal{D}(S_n)) = 1.995 \times 10^5$, $\widehat{\mathrm{Var}}_C(\mathcal{D}(S_n)) = 2.9919 \times 10^4$ and for the original graph $\mu(\mathcal{D}(S_n)) = 1.831 \times 10^6$. We find that

$$|\hat{\mu}_C(\mathcal{D}(S_n)) - \mu(\mathcal{D}(S_n))| \not\leq a\sqrt{\widehat{\mathrm{Var}}_C(\mathcal{D}(S_n))} \text{ for } a = 1, 2, 3.$$

Hence the probability that $\mu(G)$ belongs to the values generated by the random network made from Chung-Lu model is less than 0.0027, which is negligible. These two inferences can also be observed in Figures 6 and 5.

7. CONCLUSIONS

In this work we have introduced a method that by crawling a fraction of a large network can produce, to the best of our knowledge, the first non-asymptotic unbiased estimates of network node and edge characteristics. Our method is based on random walk tours and a dynamic super-node algorithm. We derive variance lower and upper bounds of this estimator and show its connection to the spectral gap of a random walk on the graph. One of our contributions is introducing an approximate Bayesian posterior of the network metric of interest using crawled data (random walk tours). We also derived a technique to study how a network looks "random" to a metric by estimating the same metric if the network was drawn from a Chung-Lu network or a configuration model with the same node labels and node degrees, all using random walk crawls without ever knowing the full original network. Our simulations over real-world networks show great accuracy of our estimators and approximations. In particular, the simulations clearly show that the derived posterior distribution fits very well with the data even when as few as 2.7% of the edges and less than 15% of the nodes in the network are observed by the crawl.

8. ACKNOWLEDGMENTS

This work was supported in part by NSF grant CNS-1065133 and ARL Cooperative Agreement W911NF-09-2-0053 and ADR "Network Science" of the Alcatel-Lucent Inria joint Lab. Research was partially conducted within the context of the THANES Associate Team, jointly supported by INRIA (France) and FAPERJ (Brazil). The authors would like to thank Arun Kadavankandy for the useful discussions.

9. REFERENCES

[1] D. Aldous and J. A. Fill. Reversible markov chains and random walks on graphs, 2002. Unfinished monograph, recompiled 2014, available at http://www.stat.berkeley.edu/~aldous/RWG/book.html.

[2] K. Avrachenkov, A. Kadavankandy, L. O. Prokhorenkova, and A. Raigorodskii. Pagerank in undirected random graphs. In *WAW*, 2015.

[3] K. Avrachenkov, N. Litvak, M. Sokol, and D. Towsley. Quick detection of nodes with large degrees. *Internet Mathematics*, 10(1-2):1–19, 2014.

[4] K. Avrachenkov, B. Ribeiro, and D. Towsley. Improving random walk estimation accuracy with uniform restarts. In *Algorithms and Models for the Web-Graph*, pages 98–109. Springer, 2010.

[5] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *ACM SIGCOMM IMC*, pages 49–62. ACM, 2009.

[6] V. Bentkus and F. Gotze. The berry-esseen bound for student's statistic. *The Annals of Probability*, pages 491–503, 1996.

[7] P. Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queue*. Springer, 2013.

[8] F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, 2002.

[9] F. R. Chung. *Spectral graph theory*, volume 92. American Mathematical Society, 1997.

[10] C. Cooper, T. Radzik, and Y. Siantos. Fast low-cost estimation of network properties using random walks. In *Algorithms and Models for the Web Graph*, pages 130–143. Springer, 2013.

[11] H. Cramér. *Mathematical methods of statistics*. Princeton university press, 1999.

[12] Dogster and C. friendships network dataset KONECT. http://konect.uni-koblenz.de/networks/petster-carnivore, May 2015.

[13] A. Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article

by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

[14] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Practical recommendations on crawling online social networks. *IEEE Journal on Selected Areas in Communications*, 29(9):1872–1892, 2011.

[15] M. Gjoka, M. Kurant, and A. Markopoulou. 2.5 k-graphs: from sampling to generation. In *IEEE INFOCOM*, pages 1968–1976, 2013.

[16] S. Goel and M. J. Salganik. Respondent-driven sampling as Markov chain Monte Carlo. *Statistics in medicine*, 28(17):2202–2229, 2009.

[17] A. Gut. *Probability: A Graduate Course*. Springer Science & Business Media, 2013.

[18] M. S. Handcock and K. J. Gile. Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5–25, mar 2010.

[19] D. D. Heckathorn. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2):174–199, 05 1997.

[20] S. Jackman. *Bayesian analysis for the social sciences*. John Wiley & Sons, 2009.

[21] J. G. Kemeny and J. L. Snell. *Finite Markov Chains*. Springer, 1983.

[22] J. H. Koskinen, G. L. Robins, and P. E. Pattison. Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. *Statistical Methodology*, 7(3):366–384, may 2010.

[23] J. H. Koskinen, G. L. Robins, P. Wang, and P. E. Pattison. Bayesian analysis for partially observed network data, missing ties, attributes and actors. *Social Networks*, 35(4):514–527, oct 2013.

[24] J. Kunegis. Konect: the Koblenz network collection. In *WWW*, pages 1343–1350, 2013.

[25] 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 Evaluation Review*, volume 40, pages 319–330, 2012.

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

[27] J. G. Ligo, G. K. Atia, and V. V. Veeravalli. A controlled sensing approach to graph classification. *IEEE Transactions on Signal Processing*, 62(24):6468–6480, 2014.

[28] L. Massoulié, E. Le Merrer, A.-M. Kermarrec, and A. Ganesh. Peer counting and sampling in overlay networks: random walk methods. In *ACM PODS*, pages 123–132, 2006.

[29] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[30] R. Ottoni, J. P. Pesce, D. B. Las Casas, G. Franciscani Jr, W. Meira Jr, P. Kumaraguru, and V. Almeida. Ladies first: Analyzing gender roles and behaviors in pinterest. In *ICWSM*, 2013.

[31] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[32] B. Ribeiro and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In *ACM SIGCOMM IMC 2010*, Nov 2010.

[33] B. Ribeiro, P. Wang, F. Murai, and D. Towsley. Sampling directed graphs with random walks. In *IEEE INFOCOM*, pages 1692–1700, 2012.

[34] S. K. Thompson. Targeted random walk designs. *Survey Methodology*, 32(1):11, 2006.

[35] H. C. Tijms. *A first course in stochastic models*. John Wiley and Sons, 2003.

[36] A. W. Van der Vaart. *Asymptotic statistics*. Cambridge university press, 2000.

APPENDIX

A. PROOF OF THEOREM 2

First, in Lemma 3 we show that the estimate of $\mu(G')$ from each tour is unbiased.

Lemma 3. *Let $X_1^{(k)}, \ldots, X_{\xi_k}^{(k)}$ be the nodes traversed by the k-th random walk tour on G', $k \geq 1$ starting at super-node S_n. Then the following holds, $\forall k$,*

$$\mathbb{E}\Big[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})\Big] = \frac{2}{d_{S_n}}\mu(G'). \tag{10}$$

Proof. The random walk starts from the super-node S_n, thus

$$\mathbb{E}\Big[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})\Big] = \sum_{(u,v)\in E'} \mathbb{E}\Big[\Big(\text{No. of times Markov chain crosses } (u,v) \text{ in the tour}\Big) f(u,v)\Big]. \tag{11}$$

Consider a renewal reward process with inter-renewal time distributed as $\xi_k, k \geq 1$ and reward as the number of times Markov chain crosses (u,v). From renewal reward theorem,

$$\{\text{Asymptotic frequency of transitions from } u \text{ to } v\} = \mathbb{E}[\xi_k]^{-1}\mathbb{E}\Big[\Big(\text{No. of times Markov chain crosses } (u,v) \text{ in the tour}\Big) f(u,v)\Big].$$

The left-hand side is essentially $2\pi_u p_{uv}$. Now (11) becomes

$$\mathbb{E}\Big[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})\Big] = \sum_{(u,v)\in E'} f(u,v)\, 2\pi_u\, p_{uv}\, \mathbb{E}[\xi_k] = \frac{2}{d_{S_n}} \sum_{(u,v)\in E'} f(u,v),$$

which concludes our proof. □

In what follows we prove Theorem 2 using Lemma 3.

Proof Theorem 2. By Lemma 3 the estimator $W_k = \sum_{t=2}^{\xi_k - 1} f(X_{t-1}^{(k)}, X_t^{(k)})$ is an unbiased estimate of $(2/d_{S_n})\mu(G')$. By the linearity of expectation the average estimator $\bar{W}(m) = m^{-1}\sum_{k=1}^{m} W_k$ is also unbiased.

We now consider two cases depending on f is defined as (2) or (3). When f is as in (2), it is trivial. For the function described in (3), $\mathbb{E}[W_k]$ can be rewritten as,

$$\frac{2}{d_{S_n}}\mathbb{E}[W_k] = \sum_{\substack{(u,v)\in E'\\ u\neq S_n, v\neq S_n}} g(u,v) + \sum_{\substack{(u,v)\in E'\\ u\neq S_n, v= S_n}} \frac{1}{k_{u\mathcal{S}}}\sum_{w\in I_n} g(u,w).$$

Note that in the graph G' there are $k_{u\mathcal{S}}$ multiple edges between u and S_n, when u and S_n are connected, and each contributes $\sum_{w\in I_n} g(u,w)$ to the net expectation. Moreover the multiplying factor of two in the left-hand side of the above expression takes into account edges in both the directions since the random walk is reversible and graph is undirected. Hence

$$\frac{2}{d_{S_n}}\mathbb{E}[W_k] = \sum_{\substack{(u,v)\in E\\ u\notin I_n, v\notin I_n}} g(u,v) + \sum_{\substack{(u,v)\in E\\ u\notin I_n, v\in I_n}} g(u,v).$$

Finally for the estimator

$$\hat{\mu}(\mathcal{D}_m(S_n)) = \frac{d_{S_n}}{2m}\bar{W}(m) + \sum_{\substack{(u,v)\in E\\ \text{s.t. } u,v\in I_n}} f(u,v).$$

has average

$$\mathbb{E}[\hat{\mu}(\mathcal{D}_m(S_n))] = \sum_{\substack{(u,v)\in E\\ \text{s.t. } u\notin I_n \text{ or } v\notin I_n}} g(u,v) + \sum_{\substack{(u,v)\in E\\ \text{s.t. } u,v\in I_n}} g(u,v) = \mu(G).$$

Furthermore, by strong law of large numbers with $\mathbb{E}[W_k] < \infty$, $\hat{\mu}(\mathcal{D}_m(S_n)) \to \mu(G)$ a.s. as $m\to\infty$. This completes our proof. □

sectionProof of Lemma 1

Proof. For RWuR, stationary distribution of node u, $\hat{\pi}_u = \frac{d_i+\alpha}{2M'+N'\alpha}$ and transition probability from node u to v, $\hat{p}_{uv} = \frac{\alpha/N'+1}{d_u+\alpha}$ if u and v are connected, $\frac{\alpha/N'}{d_u+\alpha}$ otherwise [4].

Let $f''(u,v) = g(u,v)\frac{d_u d_w}{2m-1}$ and f' as defined in (7). Let $V'' = V - I_n$. We have

$$\begin{aligned} V\times V &= \{V''\cup I_n\}\times\{V''\cup I_n\}\\ &= \{V''\times V''\}\cup\{V''\times I_n\}\cup\{I_n\times V''\}\cup\{I_n\times I_n\}. \end{aligned}$$

Now the value in the set $\{V''\times I_n\}$ can be expresses in terms of $V'\times V'$ as,

$$\begin{aligned} \sum_{(u,v)\in V''\times I_n} f''(u,v) &= \sum_{u\in V''}\sum_{w\in I_n} f''(u,w)\\ &= \sum_{\substack{u\neq S_n, v=S_n\\ (u,v)\in V'\times V'}}\sum_{w\in I_n} f''(u,w). \end{aligned}$$

$$\begin{aligned} &\mathbb{E}[W'_k]\\ &= \sum_{(u,v)\in E'} f'(u,v)2\hat{\pi}_u\hat{p}_{uv}\mathbb{E}[\xi_k] + \sum_{(u,v)\in (E')^c} f'(u,v)2\hat{\pi}_u\hat{p}_{uv}\mathbb{E}[\xi_k]\\ &= \frac{2\alpha/N'}{d_{S_n}+\alpha}\sum_{(u,v)\in E'\cup(E')^c} f'(u,v) + \frac{2}{d_{S_n}+\alpha}\sum_{(u,v)\in E'} f'(u,v)\\ &= 2\sum_{\substack{(u,v)\in E'\\ u\neq S_n, v\neq S_n}} f''(u,v)\frac{\alpha/N'+1}{d_{S_n}+\alpha} + 2\sum_{\substack{(u,v)\in E'\\ u\neq S_n, v= S_n}} \frac{\alpha/N'+k_{u\mathcal{S}}}{d_{S_n}+\alpha}\sum_{w\in I_n} f''(u,w)\\ &\quad + 2\sum_{\substack{(u,v)\in (E')^c\\ u\neq S_n, v\neq S_n}} \frac{\alpha/N'}{d_{S_n}+\alpha} f''(u,v) + 2\sum_{\substack{(u,v)\in (E')^c\\ u\neq S_n, v= S_n}} \frac{\alpha/N'}{d_{S_n}+\alpha}\sum_{w\in I_n} f''(u,w)\\ &= \frac{\alpha/N'}{d_{S_n}+\alpha}\left[\sum_{(u,v)\in V''\times V''} f''(u,v) + \sum_{(u,v)\in V''\times I_n} f''(u,v)\right]\\ &\quad + \frac{1}{d_{S_n}+\alpha}\left[\sum_{\substack{(u,v)\in E'\\ u\neq S_n, v= S_n}} k_{u\mathcal{S}}\left(\frac{1}{k_{u\mathcal{S}}}\sum_{w\in I_n} f''(u,w)\right)\right]. \end{aligned}$$

A multiplying factor 2 will be added to the first term in the above expression since RWuR is reversible and the graph under consideration is undirected. The last term can be removed by using classical random walk tours $\{W''_k\}$ with appropriate bias. The unbiasedness of the estimator then follows from the linearity of expectation. □

B. PROOF OF LEMMA 2

(i). The variance of the estimator at tour $k\geq 1$ starting from node S_n is

$$\begin{aligned} &\mathrm{Var}_{S_n}\left[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)}))\right]\\ &\quad\leq B^2\mathbb{E}[(\xi_k-1)^2] - \left(\mathbb{E}\left[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)}))\right]\right)^2. \end{aligned} \quad (12)$$

It is known from [1, Chapter 2 and 3] that

$$\mathbb{E}[\xi_k^2] = \frac{2\sum_{i\geq 2} w_{S_n i}^2(1-\lambda_i)^{-1}+1}{\pi_{S_n}^2}.$$

Using Theorem 3 eq. (12) can be written as

$$\begin{aligned} &\mathrm{Var}\left[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)}))\right]\\ &\leq \frac{1}{d_{S_n}^2}\left(2d_{\mathrm{tot}}^2B^2(\sum_{i\geq 2} w_{S_n m}^2(1-\lambda_i)^{-1}) - 4\mu^2(G')\right)\\ &\qquad -\frac{1}{d_{S_n}}B^2 d_{\mathrm{tot}} + B^2. \end{aligned}$$

The latter can be upper-bounded by $B^2(2d_{\mathrm{tot}}^2/(d_i^2\delta)+1)$.

For the second part, we have

$$\begin{aligned} \mathbb{E}\left[\left(\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})\right)^l\right] &\leq B^l\mathbb{E}[(\xi_k-1)^l)]\\ &\leq C(\mathbb{E}[(\xi_k)^l]+1), \end{aligned}$$

for a constant $C>0$ using C_r inequality [17, Chapter 3, Theorem 2.2]. From [29], it is known that there exists an $a>0$, such that $\mathbb{E}[\exp(a\,\xi_k)]<\infty$, and this implies that $\mathbb{E}[(\xi_k)^l]<\infty$ for all $l\geq 0$. This proves the theorem.

(ii). We denote $\mathbb{E}_\pi f$ for $\mathbb{E}_\pi[f(Y_1,Y_2)]$ and Normal(a,b) indicates Gaussian distribution with mean a and variance b. With the trivial extension of the central limit theorem of Markov chains [29] of node functions to edge functions, we have for the ergodic estimator $\bar{f}_n = n^{-1}\sum_{t=2}^n f(Y_{t-1},Y_t)$,

$$\sqrt{n}(\bar{f}_n - \mathbb{E}_\pi f) \xrightarrow{d} \text{Normal}(0,\sigma_a^2), \quad (13)$$

where

$$\sigma_a^2 = \mathrm{Var}(f(Y_1, Y_2)) + 2\sum_{l=2}^{n-1} \frac{(n-1)-l}{n} \mathrm{Cov}(f(Y_0, Y_1), f(Y_{l-1}, Y_l)) < \infty.$$

We derive σ_a^2 in Lemma 4. Note that σ_a^2 is also the asymptotic variance of the ergodic estimator of edge functions.

Consider a renewal reward process at its k-th renewal, $k \geq 1$, with inter-renewal time ξ_k and reward as $W_k = \sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})$. Let $\bar{W}(n)$ be the average cumulative reward gained up to m-th renewal, i.e., $\bar{W}(m) = m^{-1}\sum_{k=1}^{m} W_k$. From the central limit theorem for the renewal reward process [35, Theorem 2.2.5], with $l_n = \mathrm{argmax}_k \sum_{j=1}^{k} \mathbf{1}(\xi_j \leq n)$, after n total number of steps yields

$$\sqrt{n}(\bar{W}(l_n) - \mathbb{E}_\pi f) \xrightarrow{d} \mathrm{Normal}(0, \sigma_b^2), \tag{14}$$

with $\sigma_b^2 = \dfrac{\nu^2}{\mathbb{E}[\xi_k]}$ and

$$\begin{aligned}\nu^2 &= \mathbb{E}[(W_k - \xi_k \mathbb{E}_\pi f)^2] = \mathbb{E}_i\left[\left(W_k - \xi_k \frac{\mathbb{E}[W_k]}{\mathbb{E}[\xi_k]}\right)^2\right] \\ &= \mathrm{Var}_{S_n}(W_k) + (\mathbb{E}[W_k])^2 + \left(\frac{\mathbb{E}[W_k]}{\mathbb{E}[\xi_k]}\right)^2 \mathbb{E}[(\xi_k)^2] \\ &\quad -2\frac{\mathbb{E}[W_k]}{\mathbb{E}[\xi_k]}\mathbb{E}[W_k \xi_k].\end{aligned}$$

In fact it can be shown that (see [29, Proof of Theorem 17.2.2])

$$|\sqrt{n}(\bar{f}_n - \mathbb{E}_\pi f) - \sqrt{n}(\bar{W}(l_n) - \mathbb{E}_\pi f)| \to 0 \quad \text{a.s.}.$$

Therefore $\sigma_a^2 = \sigma_b^2$. Combining this result with Lemma 4 shown below we get (8). □

Lemma 4.

$$\begin{aligned}&\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}_\pi\left(\sum_{k=2}^{n} f(Y_{k-1}, Y_k)\right) \\ &= 2\sum_{i=2}^{r} \frac{\lambda_i}{1-\lambda_i}\langle f, v_i\rangle_{\hat{\pi}} (u_i^\intercal \hat{f}) + \frac{1}{d_{tot}} \sum_{(i,j)\in E} f(i,j)^2 \\ &\quad + \frac{1}{d_{tot}^2}\Big(\sum_{(i.j)\in E} f(i,j)^2\Big)^2 + \frac{1}{d_{tot}^2}\sum_{i\in V} d_i\Big(\sum_{i\sim j} f(i,j)\Big)^2\end{aligned}$$

Proof. We extend the arguments in the proof of [7, Theorem 6.5] to the edge functions. We have,

$$\begin{aligned}&\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}_\pi\left(\sum_{k=2}^{n} f(Y_{k-1}, Y_k)\right) \\ &= \frac{1}{n}\Big(\sum_{k=2}^{n} \mathrm{Var}_\pi(f(Y_{k-1}, Y_k)) \\ &\quad + 2\sum_{\substack{k,j=2 \\ k<j}} \mathrm{Cov}_\pi(f(Y_{k-1}, Y_k), f(Y_{j-1}, Y_j))\Big) \\ &= \mathrm{Var}_\pi(f(Y_{k-1}, Y_k)) \\ &\quad + 2\sum_{l=2}^{n-1} \frac{(n-1)-l}{n} \mathrm{Cov}_\pi(f(Y_0, Y_1), f(Y_{l-1}, Y_l)).\end{aligned} \tag{15}$$

Now the first term in (15) is

$$\mathrm{Var}_\pi(f(Y_{k-1}, Y_k)) = \langle f, f\rangle_{\hat{\pi}} - \langle f, \Pi\hat{f}\rangle_{\hat{\pi}}, \tag{16}$$

where $\Pi = \mathbf{1}\pi^\intercal$.

For the second term in (15),

$$\begin{aligned}&\mathrm{Cov}_\pi(f(Y_0, Y_1), f(Y_{l-1}, Y_l)) \\ &\quad = \mathbb{E}_\pi(f(Y_0, Y_1), f(Y_{l-1}, Y_l)) - (\mathbb{E}_\pi[f(Y_0, Y_1)])^2.\end{aligned}$$

$$\begin{aligned}&\mathbb{E}_\pi(f(Y_0, Y_1), f(Y_{l-1}, Y_l)) \\ &= \sum_i \sum_j \sum_k \sum_m \pi_i\, p_{ij}\, p_{jk}^{(l-2)}\, p_{km} f(i,j) f(k,m) \\ &= \langle f, P^{(l-2)}\hat{f}\rangle_{\hat{\pi}}.\end{aligned} \tag{17}$$

Therefore,

$$\mathrm{Cov}_\pi(f(Y_0, Y_1), f(Y_{l-1}, Y_l)) = \langle f, (\mathbf{P}^{(l-2)} - \Pi)\hat{f}\rangle_{\hat{\pi}}.$$

Taking limits, we get

$$\begin{aligned}&\lim_{n\to\infty} \sum_{l=2}^{n-1} \frac{n-l-1}{n} (\mathbf{P}^{(l-2)} - \Pi) \\ &\overset{(a)}{=} \lim_{n\to\infty} \sum_{k=1}^{n-1} \frac{n-k}{n} (\mathbf{P}^k - \Pi) + (\mathbf{I} - \Pi) - \lim_{n\to\infty} \frac{3}{n} \sum_{k=1}^{n-3} (\mathbf{P}^k - \Pi) \\ &= (\mathbf{Z} - \mathbf{I}) + (\mathbf{I} - \Pi) = \mathbf{Z} - \Pi,\end{aligned} \tag{18}$$

where the first term in (a) follows from the proof of [7, Theorem 6.5] and since $\lim_{n\to\infty}(\mathbf{P}^n - \Pi) = 0$, the last term is zero using Cesaro's lemma [7, Theorem 1.5 of Appendix].

We have,

$$\mathbf{Z} = \mathbf{I} + \sum_{i=2}^{r} \frac{\lambda_i}{1-\lambda_i} v_i u_i^\intercal,$$

Thus

$$\begin{aligned}&\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}_\pi\left(\sum_{k=2}^{n} f(Y_{k-1}, Y_k)\right) \\ &= \langle f, f\rangle_{\hat{\pi}} - \langle f, \Pi\hat{f}\rangle_{\hat{\pi}} + 2\langle f, \left(\mathbf{I} + \sum_{i=2}^{r} \frac{\lambda_i}{1-\lambda_i} v_i u_i^\intercal - \Pi\right)\hat{f}\rangle_{\hat{\pi}} \\ &= \frac{1}{d_{tot}} \sum_{(i,j)\in E} f(i,j)^2 + \frac{1}{d_{tot}^2}\Big(\sum_{(i.j)\in E} f(i,j)^2\Big)^2 \\ &+ \frac{1}{d_{tot}^2}\sum_{i\in V} d_i\Big(\sum_{i\sim j} f(i,j)\Big)^2 + 2\sum_{i=2}^{r} \frac{\lambda_i}{1-\lambda_i}\langle f, v_i\rangle_{\hat{\pi}} (u_i^\intercal \hat{f})\end{aligned}$$

□