



Bayesian Inference of Online Social Network Statistics via Lightweight Random Walk Crawls

Konstantin Avrachenkov, Bruno Ribeiro, Jithin K. Sreedharan

► To cite this version:

Konstantin Avrachenkov, Bruno Ribeiro, Jithin K. Sreedharan. Bayesian Inference of Online Social Network Statistics via Lightweight Random Walk Crawls. [Research Report] RR-8793, Inria Sophia Antipolis; Purdue University. 2015. hal-01216285v1

HAL Id: hal-01216285

<https://inria.hal.science/hal-01216285v1>

Submitted on 16 Oct 2015 (v1), last revised 17 Dec 2015 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Bayesian Inference of Online Social Network Statistics via Lightweight Random Walk Crawls

Konstantin Avrachenkov, Bruno Ribeiro,
Jithin K. Sreedharan

**RESEARCH
REPORT**

N° 8793

October 2015

Project-Team Maestro



Bayesian Inference of Online Social Network Statistics via Lightweight Random Walk Crawls

Konstantin Avrachenkov*, Bruno Ribeiro†,
Jithin K. Sreedharan‡

Project-Team Maestro

Research Report n° 8793 — October 2015 — 20 pages

Abstract: Online social networks (OSN) contain extensive amount of information about the underlying society that is yet to be explored. One of the most feasible technique to fetch information from OSN, crawling through Application Programming Interface (API) requests, poses serious concerns over the the guarantees of the estimates. In this work, we focus on making reliable statistical inference with limited API crawls. Based on regenerative properties of the random walks, we propose an unbiased estimator for the aggregated sum of functions over edges and proved the connection between variance of the estimator and spectral gap. In order to facilitate Bayesian inference on the true value of the estimator, we derive the asymptotic posterior distribution of the estimate. Later the proposed ideas are validated with numerical experiments on inference problems in real-world networks.

Key-words: Bayesian inference, Random walk on graphs, Social network analysis, Graph sampling

The authors acknowledge the partial support by ADR "Network Science" of Inria - Bell Labs and by the joint Brazilian-French research team Thanos

* Inria Sophia Antipolis, France. Email: k.avrachenkov@sophia.inria.fr

† Purdue University, IN, USA Email: ribeiro@cs.purdue.edu

‡ Inria Sophia Antipolis, France. Email: jithin.sreedharan@inria.fr

**RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE**

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Inférence bayésienne de statistiques des réseaux sociaux par les marches aléatoires avec complexité légère

Résumé : Les réseaux sociaux en ligne contiennent une grande quantité d'informations à propos de la société sous-jacente, qui demande encore à être explorée. Une des techniques, parmi les plus faisables pour récupérer des informations, est d'explorer le réseau à l'aide des Application Programming Interface (API). Dans ce travail, nous nous concentrons sur l'inférence statistique fiable avec un nombre limité d'appels à l'API. Sur la base des propriétés régénératrices des marches aléatoires, nous proposons un estimateur sans biais pour la somme agrégée de fonctions sur les arêtes du graphe. Nous montrons la connexion entre la variance de l'estimateur et le trou spectral. Afin de faciliter l'inférence bayésienne sur la vraie valeur de l'estimateur, nous calculons la distribution postérieure asymptotique de l'estimation. Finalement, les idées proposées sont validées par des expériences numériques sur les problèmes d'inférence dans les réseaux sociaux réels.

Mots-clés : Inférence bayésienne, marche aléatoire sur les graphes, analyse de réseau social.

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. [25]), epidemiology [26,], and computer science [4, 11, 27,]. 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 crawls¹. 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 ([3, 10, 28]) 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 asymptotic posterior of our estimates that performs remarkably well 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. 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

$$\mu(G) = \sum_{(u,v) \in E} f(u, v). \quad (1)$$

Note that functions of the form eq. (1) are general enough to compute node statistics

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

where d_u is the degree of node $u \in V$, and statistics of triangles such as the local clustering coefficient of G first provided by [28]

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

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).

¹The majority of the datasets in the public repositories SNAP [21,] and KONECT [19,] are partial website crawls, not complete datasets or uniform samples.

Contributions

In our work we provide a partial crawling strategy using random walk tours whose posterior

$$P[\mu(G)|\mathcal{D}_m(I_n)]$$

is shown to have an unbiased maximum a posteriori estimate (MAP) $\hat{\mu}_{\text{MAP}}(\mathcal{D}_m(I_n))$ regardless of the number of nodes in the seed set $n > 0$ and regardless of the value of $m > 0$, i.e., $E[\hat{\mu}_{\text{MAP}}(\mathcal{D}_m(I_n))] = \mu(G)$, $\forall n, m > 0$. Note that we guarantee that our MAP estimate is unbiased in the finite-sample regime unlike previous asymptotic methods [3, 10, 20, 28, 29,]. Moreover, we provide the posterior $P[\mu(G)|\mathcal{D}_m(I_n)]$ for the large m regime and prove its convergence in distribution showing its convergence rate. In our experiments we note that the posterior is remarkably accurate using a variety of networks large and small. We also provide upper and lower bounds for $P[\mu(G)|\mathcal{D}_m(I_n)]$.

Related Work

The works of [23] and [6] are the ones closest to ours. [23] estimates the size of a network based on the return times of random walk tours. [6] estimates number of triangles, network size, and subgraph counts from weighted random walk tours using results of [1]. The previous works on non-asymptotic inference of network statistics from incomplete network crawls [12, 17, 18, 13, 14, 22, 30,] 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; 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 our main theorems and supporting lemmas and proofs. In Section 3 we introduce artificial illustrative examples to aid understanding our method. In Section 4 we introduce our results using simulations over real-world networks. Finally, in Section 5 we present our conclusions.

2 Network Estimation from Partial Crawls

In this section we present our main results. The outline of this section is as follows. Section 2.1 introduces key concepts and defines the notation used throughout this manuscript. Section 2.2 introduces our main results in the form of two theorems: Theorem 1 presents an unbiased estimator of any function over edges of an undirected graph using random walk tours. 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. Theorem 2 introduces the asymptotic posterior of the estimator in Theorem 1. Section 2.3 proves Theorem 1, introducing important upper and lower bounds of the estimator variance in Section 2.4.1 and showing the effect of the spectral gap. Finally, Section 2.4 proves Theorem 2 also using the bounds obtained in Section 2.4.1.

2.1 Preliminaries

Let $G = (V, E)$ be an unknown undirected graph. 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 .

Our network crawler is a classical random walk over the following augmented multigraph $G' = (V', E')$. A multigraph is a graph that can have multiple edges between two nodes. In G' we aggregate all nodes of I_n into a single node, denoted hereafter \mathcal{S}_n , the *super-node*. Thus, $V' = \{V \setminus I_n\} \cup \{\mathcal{S}_n\}$. The edges of G' are $E' = E \setminus \{E \cap \{I_n \times V\}\} \cup \{(\mathcal{S}_n, v) : \forall (u, v) \in E, \text{ s.t. } u \in I_n \text{ and } v \in V \setminus I_n\}$, i.e., E' 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 \mathcal{S}_n . Note that G' is necessarily connected as I_n spans all the connected components of G .

A random walk on G' has transition probability from node u to an adjacent node v , $p_{uv} := \mathbf{P}_{u,v}$ with $\alpha_{u,v}/d_u$, where d_u is the degree of u and $\alpha_{u,v}$ is the number of edges between $u \in V'$ and $v \in V'$. 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' . Let the nodes crawled the random walk be $X_1, X_2 \dots$.

A random walk *tour* is defined as the sequence of nodes $X_1^{(k)}, \dots, X_{\xi_k}^{(k)}$ visited by the random walk during successive k -th and $k+1$ -st visits to the super-node \mathcal{S}_n . Here $\{\xi_k\}_{k \geq 1}$ denote the successive return times to \mathcal{S}_n . Tours have a key property: from the renewal theorem tours are independent since the returning times act as renewal epochs.

Note that the random walk on G' is equivalent to a random walk on G where all the nodes in I_n are treated as **one single node**.

Super-node Motivation

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

- *Tackling disconnected or low-conductance graphs*: When the graph is not strongly connected or has many connected 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. This idea is illustrated with a Dumbell graph example in Section 3.
- *Faster Estimate with Shorter Tours*: The expected value of the k -th tour length $E[\xi_k] = 1/\pi_{\mathcal{S}_n}$ is inversely proportional to the degree of the super-node $d_{\mathcal{S}_n}$. Hence, by forming a massive-degree super-node we can significantly shorten the average tour length.

2.2 Main Results

In what follows we present our main results in two theorems. Theorem 1 proposes an unbiased estimator of edge characteristics $\mu(G)$ via random walk tours. Theorem 2 presents the asymptotic posterior distribution of the unbiased estimator presented in Theorem 1.

Theorem 1. *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 2.1 starting at super-node \mathcal{S}_n . Let*

$(X_t^{(k)})_{t=1}^{\xi_k}$ be the k -th random walk tour until the walk first returns to \mathcal{S}_n and let $\mathcal{D}_m(\mathcal{S}_n)$ denote the collection of all nodes in $m \geq 1$ such tours, $\mathcal{D}_m(\mathcal{S}_n) = ((X_t^{(k)})_{t=1}^{\xi_k})_{k=1}^m$. Then,

$$\hat{\mu}(\mathcal{D}_m(\mathcal{S}_n)) = \overbrace{\frac{d_{\mathcal{S}_n}}{2m} \sum_{k=1}^m \sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})}^{\text{Estimate from crawls}} + \overbrace{\sum_{(u,v) \in E \text{ s.t. } u,v \in I_n} f(u,v)}^{\text{Edges between initial nodes in original } G} \quad (2)$$

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

Theorem 1 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, $E[\xi_k] \propto 1/d_{\mathcal{S}_n}$. This provides a practical way to compute unbiased estimates of node, edge, and triangle statistics using $\hat{\mu}(\mathcal{D}_m(\mathcal{S}_n))$ (eq. (2)) while observing only a small fraction of the original graph. Because random walk tours can have arbitrary lengths, we show in Lemma 2, Section 2.4, that there are upper and lower bounds on the variance of $\hat{\mu}(\mathcal{D}_m(\mathcal{S}_n))$. For a bounded function f , the upper bounds are shown to be always finite.

In what follows Theorem 2 shows the asymptotic posterior of the estimator in Theorem 1. In Section 4 we shall see that the asymptotic posterior matches very well the empirical posterior using simulations over real-world networks while crawling $< 10\%$ of the nodes in the network.

Theorem 2. *In the scenario of Theorem 1 for $m \geq 2$ tours and assuming priors $\hat{\mu}(\mathcal{D}_m(\mathcal{S}_n))|\sigma^2 \sim \text{Normal}(\mu_0, \sigma^2/m_0)$, $\sigma^2 \sim \text{Inverse-gamma}(\nu_0, \sigma_0^2)$, and parameter $m_0 > 0$, then the marginal posterior density of $\hat{\mu}(\mathcal{D}_{\lfloor \sqrt{m} \rfloor}(\mathcal{S}_n))$ as $m \rightarrow \infty$ converges in distribution to a non-standardized t -distribution*

$$\phi(x|\nu, \hat{\mu}, \hat{\sigma}) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2}) \hat{\sigma} \sqrt{\pi\nu}} \left(1 + \frac{(x - \hat{\mu})^2}{\hat{\sigma}^2 \nu} \right)^{-\frac{\nu+1}{2}} \quad (3)$$

with degrees of freedom parameter

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

, location parameter

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

and scale parameter

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

where

$$\hat{S}_h = \frac{d_{\mathcal{S}_n}}{2 \lfloor \sqrt{m} \rfloor} \sum_{k=((h-1)\lfloor \sqrt{m} \rfloor + 1)}^{h \lfloor \sqrt{m} \rfloor} \sum_{t=2}^{\xi_h} f(X_{t-1}^{(k)}, X_t^{(k)}).$$

The asymptotic posterior in Theorem 2 provides a way to access the confidence in the estimate $\hat{\mu}(\mathcal{D}_m(\mathcal{S}_n))$. 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 [9,], 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 [9,]. The posterior is conservative as $\hat{\mu}(\mathcal{D}_m(\mathcal{S}_n))$ is calculated from m tours while the posterior considers only \sqrt{m} tours. Being conservative, however, is advised as the posterior is asymptotic and the conservative estimate better protects us from finite-sample anomalies and perform very well in practice as we see in Section 4.

In what follows we provide the proofs of our main results.

2.3 Proof of Theorem 1

The outline of this proof is as follows. In Lemma 1 we show that the estimate of $\mu(G)$ from each tour is unbiased.

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

$$E\left[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})\right] = \frac{2}{d_{\mathcal{S}_n}} \mu(G'), \quad (4)$$

where $\mu(G')$ is the true value $\mu(G)$ outside the subgraph formed by the nodes that were merged into the super-node, and $\mu(G') = \sum_{(u,v) \in E'} f(u, v)$.

Proof. The random walk starts from the super-node \mathcal{S}_n , thus

$$E\left[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})\right] = \sum_{(u,v) \in E'} E\left[\left(\text{No. of times Markov chain crosses } (u, v) \text{ in the tour}\right) f(u, v)\right] \quad (5)$$

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,

$$\begin{aligned} & \{\text{Asymptotic frequency of transitions from } u \text{ to } v\} \\ &= \frac{E\left[\left(\text{No. of times Markov chain crosses } (u, v) \text{ in the tour}\right) f(u, v)\right]}{E[\xi_k]} \end{aligned}$$

Here the left-hand side is essentially $2\pi_u p_{uv}$. Now (5) becomes

$$\begin{aligned} E\left[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)})\right] &= \sum_{(u,v) \in E'} f(u, v) 2\pi_u p_{uv} E[\xi_k] \\ &= 2 \sum_{(u,v) \in E'} f(u, v) \frac{d_u}{\sum_j d_j} \frac{1}{d_u} \frac{\sum_j d_j}{d_{\mathcal{S}_n}} = \frac{2}{d_{\mathcal{S}_n}} \sum_{(u,v) \in E'} f(u, v), \end{aligned}$$

which concludes our proof. \square

In what follows we prove Theorem 1 using Lemma 1.

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

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

has average

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

Furthermore, by strong law of large numbers with $E[W_k] < \infty$, $\hat{\mu}(\mathcal{D}_m(\mathcal{S}_n)) \rightarrow \mu(G)$ a.s. for $m \rightarrow \infty$. This completes our proof. \square

2.4 Proof of Theorem 2

The proof of Theorem 2 relies first on showing that $\hat{\mu}(\mathcal{D}_m(\mathcal{S}_n))$ has finite first and second moments. We go further and in Lemma 2 we introduce upper and lower bounds on the variance of $\hat{\mu}(\mathcal{D}_m(\mathcal{S}_n))$. By Theorem 1 the first moment of $\hat{\mu}(\mathcal{D}_m(\mathcal{S}_n))$ is finite as $E[\hat{\mu}(\mathcal{D}_m(\mathcal{S}_n))] = \mu(G)$.

To show that the second moment is finite we prove that the estimate $W_k = \sum_{t=2}^{\xi_k-1} f(X_{t-1}^{(k)}, X_t^{(k)})$, $k \geq 1$, whose variance is $m \cdot \text{Var}(\hat{\mu}(\mathcal{D}_m(\mathcal{S}_n)))$, has finite second moment. The results in Lemma 2 are of interest on their own because they establish a connection between the estimator variance and the spectral gap.

2.4.1 Impact of spectral gap on variance

Let $\mathbf{S} = \mathbf{D}^{1/2} \mathbf{P} \mathbf{D}^{-1/2}$, where \mathbf{P} is the random walk transition probability matrix as defined in Section 2.1 and $\mathbf{D} = \text{diag}(d_1, d_2, \dots, d_{|V'|})$ is a diagonal matrix with the node degrees of G' . The eigenvalues $\{\lambda_i\}$ of \mathbf{P} and \mathbf{S} are same and $1 = \lambda_1 > \lambda_2 \geq \dots \geq \lambda_{|V'|} \geq -1$. Let j th 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 $\langle f, g \rangle_{\hat{\pi}} = \sum_{(u,v) \in E'} \hat{\pi}_{uv} f(u,v) g(u,v)$, 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, $\max_{(i,j) \in E'} f(i,j) \leq B < \infty$, and for tour $k \geq 1$,*

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

(ii).

$$\begin{aligned}
 & \text{var}_{\mathcal{S}_n} \left[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)}) \right] \\
 & \geq 2 \frac{d_{\text{tot}}}{d_{\mathcal{S}_n}} \sum_{i=2}^r \frac{\lambda_i}{1 - \lambda_i} \langle f, v_i \rangle_{\hat{\pi}} (u_i^\top \hat{f}) + \frac{1}{d_{\mathcal{S}_n}} \sum_{(u,v) \in E'} f(u,v)^2 + \frac{1}{d_{\text{tot}} d_{\mathcal{S}_n}} \left(\sum_{(u,v) \in E'} f(u,v)^2 \right)^2 \\
 & \quad + \frac{1}{d_{\text{tot}} d_{\mathcal{S}_n}} \sum_{u \in V'} d_u \left(\sum_{u \sim v} f(u,v) \right)^2 - \frac{4}{d_{\mathcal{S}_n}^2} \left(\sum_{(u,v) \in E'} f(u,v) \right)^2 \\
 & \quad - \frac{8}{d_{\text{tot}}} \left(\sum_{(u,v) \in E'} f(u,v) \right)^2 \sum_{i \geq 2} \frac{w_{\mathcal{S}_n i}^2}{(1 - \lambda_i)} - \frac{4}{d_{\text{tot}} d_{\mathcal{S}_n}} \left(\sum_{(u,v) \in E'} f(u,v) \right)^2. \tag{6}
 \end{aligned}$$

Proof. (i). The variance of the estimator at tour $k \geq 1$ starting from node \mathcal{S}_n is

$$\text{var}_{\mathcal{S}_n} \left[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)}) \right] \leq B^2 E[(\xi_k - 1)^2] - \left(E \left[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)}) \right] \right)^2. \tag{7}$$

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

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

By Theorem 1 eq. (7) can be written as

$$\begin{aligned}
 & \text{Var} \left[\sum_{t=2}^{\xi_k} f(X_{t-1}^{(k)}, X_t^{(k)}) \right] \\
 & \leq B^2 \frac{Z_{\mathcal{S}_n, \mathcal{S}_n}}{\pi_{\mathcal{S}_n}^2} - \frac{B^2}{\pi} + B^2 - \frac{4}{d_{\mathcal{S}_n}^2} \mu^2(G') \\
 & = \frac{1}{d_{\mathcal{S}_n}^2} \left(2d_{\text{tot}}^2 B^2 \left(\sum_{i \geq 2} u_{\mathcal{S}_n i}^2 (1 - \lambda_i)^{-1} \right) - 4\mu^2(G') \right) - \frac{1}{d_{\mathcal{S}_n}} B^2 d_{\text{tot}} + B^2,
 \end{aligned}$$

where $Z_{uu} = \sum_{i \geq 2} w_{ui}^2 (1 - \lambda_i)^{-1}$. The latter can be upper-bounded by $B^2(2d_{\text{tot}}^2/(d_i^2 \delta) + 1)$.

(ii). Let Y_1, Y_2, \dots, Y_n be a random walk on G' in steady state. We denote $E_\pi f$ for $E_\pi[f(Y_1, Y_2)]$ and $\text{Normal}(a, b)$ indicates Gaussian distribution with mean a and variance b . With the trivial extension of the central limit theorem of Markov chains [16,] of node functions to edge functions, we have for the ergodic estimator $\bar{f}_n = n^{-1} \sum_{t=2}^n f(Y_{t-1}, Y_t)$,

$$\sqrt{n}(\bar{f}_n - E_\pi f) \xrightarrow{d} \text{Normal}(0, \sigma_a^2), \tag{8}$$

where

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

We derive σ_a^2 in Lemma 3. 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 $W_k = \sum_{t=2}^{\xi_k-1} f(X_{t-1}, X_t)$. 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 [31, Theorem 2.2.5], we have

$$\sqrt{m}(\bar{W}(m) - E_\pi f) \xrightarrow{d} \text{Normal}(0, \sigma_b^2), \quad (9)$$

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

$$\begin{aligned} \nu^2 &= E[(W_k - \xi_k E_\pi f)^2] = E_i \left[\left(W_k - \xi_k \frac{E[W_k]}{E[\xi_k]} \right)^2 \right] \\ &= \text{var}_{\mathcal{S}_n}(W_k) + (E[W_k])^2 + \left(\frac{E[W_k]}{E[\xi_k]} \right)^2 E[(\xi_k)^2] - 2 \frac{E[W_k]}{E[\xi_k]} E[W_k \xi_k]. \end{aligned}$$

In fact it can be shown that (see [24, Proof of Theorem 17.2.2])

$$|\sqrt{n}(\bar{f}_n - E_\pi f) - \sqrt{l_n}(\bar{W}(l_n) - E_\pi f)| \rightarrow 0 \quad \text{a.s.},$$

where $l_n = \arg\max_k \sum_{j=1}^k \mathbf{1}(\xi_j \leq n)$. Therefore $\sigma_a^2 = \sigma_b^2$. Combing this result with Lemma 3 shown in the appendix we get (6). \square

We are now ready to prove Theorem 2.

Theorem 2. Let $m' = \lfloor \sqrt{m} \rfloor$. Given

$$\hat{F}_h = \frac{d\mathcal{S}_n}{2\lfloor \sqrt{m} \rfloor} \sum_{k=((h-1)\lfloor \sqrt{m} \rfloor + 1)}^{h\lfloor \sqrt{m} \rfloor} \sum_{t=2}^{\xi_h} f(X_{t-1}^{(k)}, X_t^{(k)}).$$

and $\{\hat{F}_h\}_{h=1}^{m'}$ and because the tours are i.i.d. $\hat{\mu}(\mathcal{D}_{\lfloor \sqrt{m} \rfloor}(S_n))$ the marginal posterior density of μ is

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

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 [15, Proposition C.4]

$$\begin{aligned} \mu|\sigma^2, \{\hat{F}_h\}_{h=1}^{m'} &\sim \text{Normal} \left(\frac{m_0\mu_0 + \sum_{h=1}^{m'} \hat{F}_h}{m_0 + m'}, \frac{\sigma^2}{m_0 + m'} \right), \\ \sigma^2|\{\hat{F}_h\}_{h=1}^{m'} &\sim \text{Inverse-Gamma} \left(\frac{\nu_0 + m'}{2}, \frac{\nu_0\sigma_0^2 + \hat{\sigma}_{m'} + \frac{m_0 m'}{m_0 + m'} (\mu_0 - \hat{\mu}(\mathcal{D}_m(S_n)))^2}{2} \right) \end{aligned}$$

are the posteriors of parameters μ and σ^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 [15, Proposition C.6]. Thus, if $\{\hat{F}_h\}_{h=1}^{m'}$ are i.i.d. normally distributed then

the posterior of their average μ and variance σ^2 is non-standardized t -distributed with parameters

$$t\left(\mu = \frac{m_0\mu_0 + \sum_{h=1}^{m'} \hat{F}_h}{m_0 + m'}, \sigma^2 = \frac{\nu_0\sigma_0^2 + \hat{\sigma}_{m'} + \frac{m_0m'}{m_0+m'}(\mu_0 - \hat{\mu}(\mathcal{D}_m(S_n)))^2}{2}, \nu = \nu_0 + \lfloor \sqrt{m} \rfloor\right). \quad (10)$$

Left to show is that $\{\hat{F}_h\}_{h=1}^{m'}$ are converge *in distribution* to i.i.d. normal random variables as $m \rightarrow \infty$. As the spectral gap of G_{S_n} is greater than zero, $|\lambda_1 - \lambda_2| > 0$, Lemma 2 shows that

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

for any tour k of length ξ_k . From the renewal theorem we know that $\{R_k\}_{k=1}^m$ are i.i.d. random variables and thus any subset of these variables is also i.i.d.. By construction $\hat{F}_1, \dots, \hat{F}_{m'}$ are also i.i.d. with mean $\mu(G_{S_n})$ and finite variance $0 < \sigma_{m'}^2 < \infty$. Applying the Lindeberg-Lévy central limit theorem [7, Section 17.4] yields

$$\hat{F}_h - \mu(G_{S_n}) \xrightarrow{d} \text{Normal}(0, \sigma_{m'}^2 / \sqrt{m'}), \quad \forall k,$$

where $\text{var}(\hat{F}_h) = \sigma_{m'}^2$. Thus, in the limit as $m \rightarrow \infty$ and $m' \rightarrow \infty$ (recall that $m' = \lfloor \sqrt{m} \rfloor$), the variables $\{\hat{F}_h\}_{h=1}^{m'}$ are i.i.d. normally distributed with

$$\hat{F}_h \sim \text{Normal}(\mu(G_{S_n}), \sigma_{m'}^2 / \sqrt{m'}), \quad \forall k,$$

and

$$\sum_{(u', v') \in E \text{ s.t. } u', v' \in I_n} f(u', v')$$

is constant and known, which concludes our proof. \square

3 Illustrative example

Here we consider the classical example of low-conductance graph: the *dumbbell* graph to show how the super-node solves the *variance* problem for random walk tours on graphs. A dumbbell graph consists of two complete graphs K_n on n vertices connected by a single edge. The spectral gap δ ($(1 - \lambda_2)$, λ_2 being the second largest absolute eigenvalue of \mathbf{P}) of this graph is roughly $O(1/n^2)$.

It is known that the variance of the return time ξ_k of tour $k > 0$ is related to δ as [1,]

$$\text{var}_{S_n} \xi_k \leq \frac{2}{\delta \pi_{S_n}^2} + \frac{1}{\pi_{S_n}}.$$

With the super-node, the new graph G_{S_n} will be more connected and hence δ improves, thus the variance.

Another way to view this from forming super-node is that the covering time with a normal random walk on dumbbell graph is $\Theta(n^2)$. If $k = O(\log n)$ random walks run in parallel with some conditions on distributing them, the covering time can be reduced to $O(n)$ [2,]. Forming a super-node with the initial seeds of these random walks can be viewed as starting multiple random walks from these nodes, and thus covering the graph in short runs.

4 Experiments on Real-world Networks

In this section we demonstrate the effectiveness of the theory developed above with real data sets of various social networks. We assume the contribution from super-node to the true value is known a priori and hence we look for $\mu(G_{\mathcal{S}_n})$ in the experiments. Even in the case that the edges of the super-node is known, 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 sub graphs. Later, in order to estimate the bias created by this sub graph, do some RW tours from the largest degree node in each of these sub graph and use the idea in Theorem 1.

In the figures we display both asymptotic posterior and empirical posterior generated from \hat{F}_h . For the asymptotic posterior, we have used the following parameters $n_0 = 0, \nu_0 = 0, \mu_0 = 0, \sigma_0 = 1$ (see Theorem 2). In the figures, the green line shows the actual value $\mu(G_{\mathcal{S}_n})$.

4.1 Friendster

First we study a network of moderate size, a connected subgraph of Friendster network with 64,600 nodes and 1,246,479 edges. Two types of functions are considered:

1. $f_1 = d_{X_t} \cdot d_{X_{t+1}}$
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 super-node is formed from 10,000 uniformly sampled nodes. Figures 1 and 2 display the results for functions f_1 and f_2 , respectively. A good match between the asymptotic and empirical posteriors can be observed from the figures. Moreover the true value $\mu(G_{\mathcal{S}_n})$ is also fitting well with the plots. The percentage of graph crawled is 24.44% in terms of edges and this drops to 7.43% if we use random walk based super-node formation.

4.2 Network of dogs

The aim of this example is to check whether there is any affinity for making connections between the owners of same breed dogs [8,]. Number of nodes is 415,431 and number of edges is 8,265,511.

In Figure 3, 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 super-node is formed by 10,000 nodes which are uniformly selected at random. The percentage of the graph crawled in terms of edges is 5.02% and in terms of nodes is 37.17%. While using the random walk based super-node formation, the graph crawled drops to 2.72% (in terms of edges) and 14.86% (in terms of nodes) with the same super-node size. But these values can be reduced much further if we allow a bit less precision in the matching between theoretical distribution and histogram.

In order to better understand the correlation in forming edges, we now consider the *configuration model*. The configuration model is formed as follows: all the edges in the graph is cut and what is left is half edges for each node, and the number of half edges is the degree of the node. Now these half edges are paired uniformly. Such a configuration will create a graph whose edges are formed without any correlation between the end-nodes. We run our estimator on the configuration model and plot the histogram and distribution as we did for the original graph. Figure 4 compare function f_2 for the configuration model and original graph. The figure shows

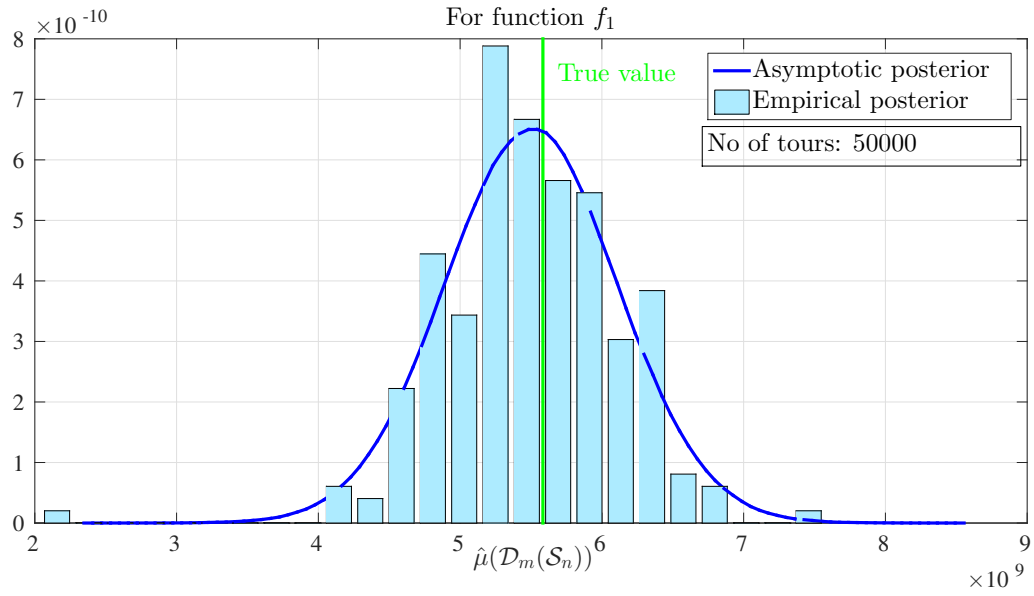


Figure 1: Friendster subgraph, function f_1

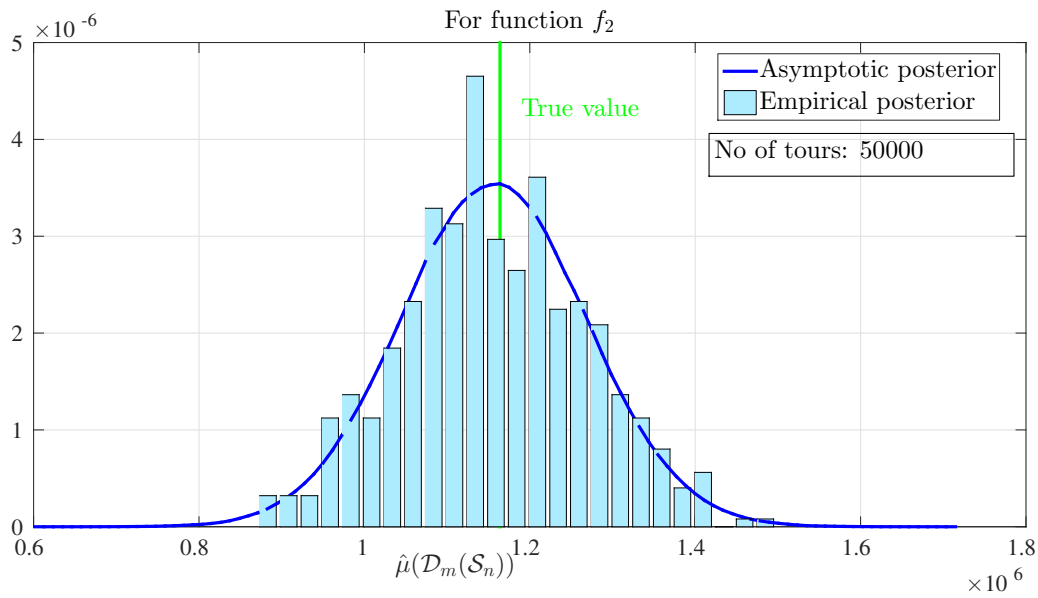


Figure 2: Friendster subgraph, function f_2

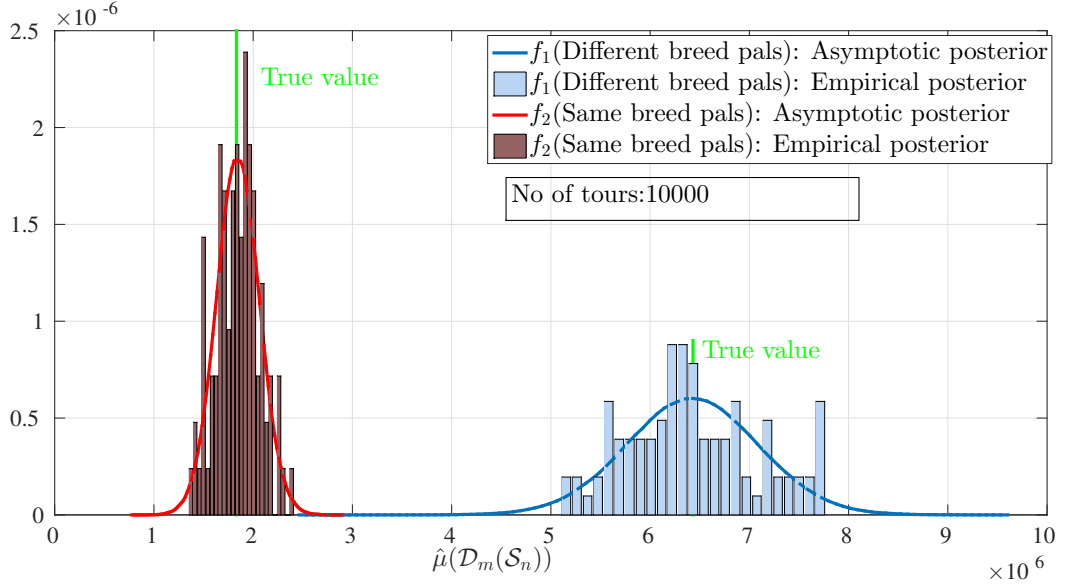


Figure 3: Dog pals network

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 5 shows similar figure in case of f_1 . It is important to note that one can create a configuration network model from the crawl and the knowledge of the complete network is not necessary.

4.3 ADD Health data

Though our main result in Theorem 2 is asymptotic, in this section we check with a small dataset. We consider ADD network project (<http://www.cpc.unc.edu/projects/addhealth>), a friendship network among high school students in US. The graph has 1545 nodes and 4003 edges.

We take two types of functions. Figure 6 shows the affinity in same gender or different gender friendships and Figure 7 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.

5 Conclusions

In this work we have studied an efficient way to make the statistical inference on networks. We introduced a technique to make a quicker inference by crawling a very small percentage of a network which may be disconnected or having low conductance. The proposed method can also be performed in parallel. In the paper, first we presented a non-asymptotic unbiased estimator and derived the bounds on its variance showing the connection with the spectral gap. Later we proved that the asymptotic posterior of the estimator given the crawled data converges to a non-centralized student's t distribution. The numerical experiments on real-world networks of different sizes, large and small, demonstrate the correctness of the estimator. In particular, the

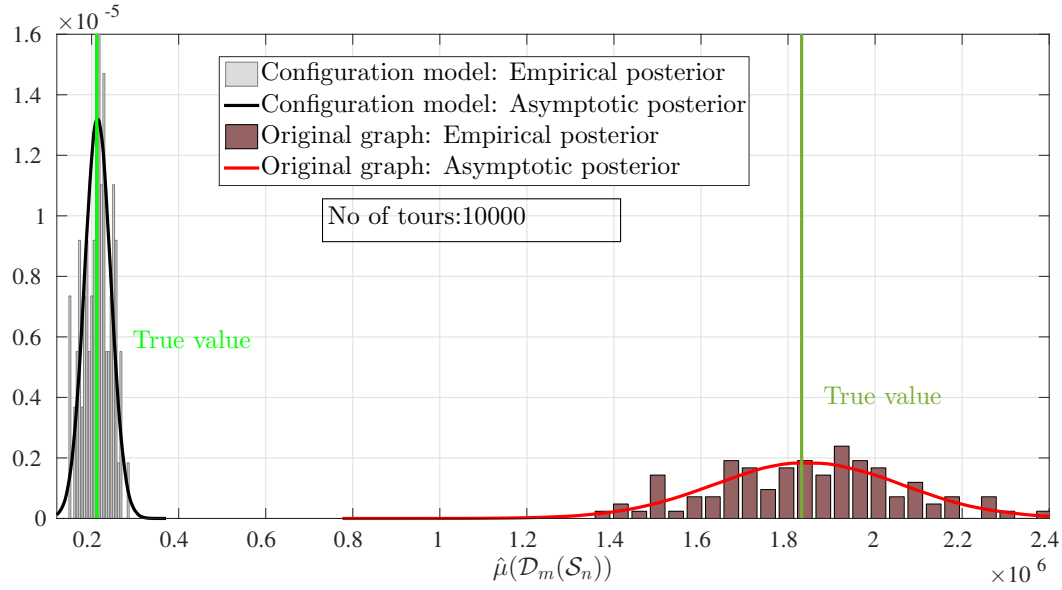


Figure 4: Dog pals network: Comparison between configuration model and original graph for f_2 , number of connection between same breeds

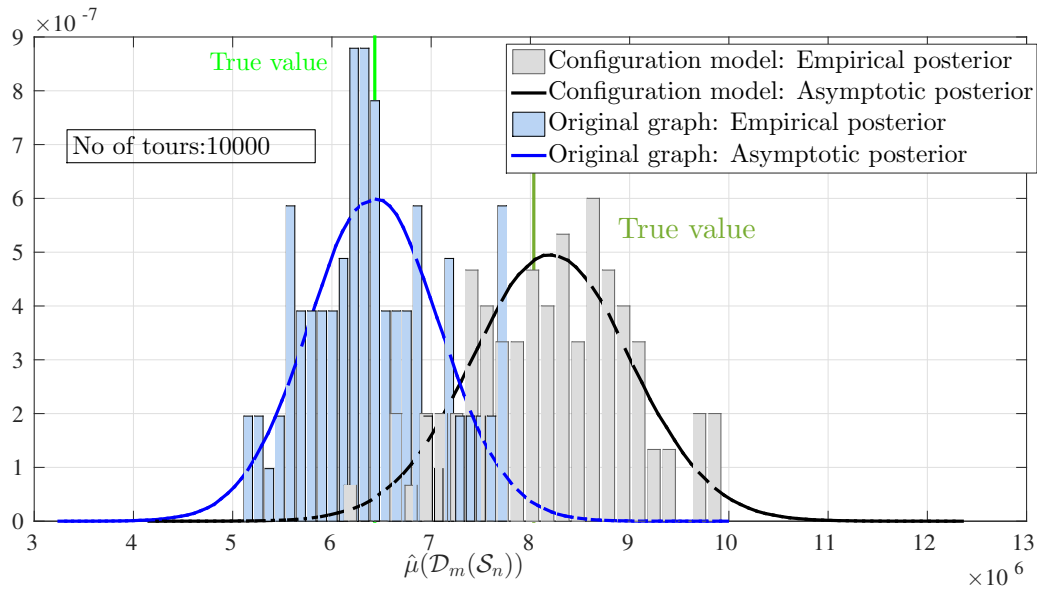


Figure 5: Dog pals network: Comparison between configuration model and original graph for f_1 , number of connection between different breeds

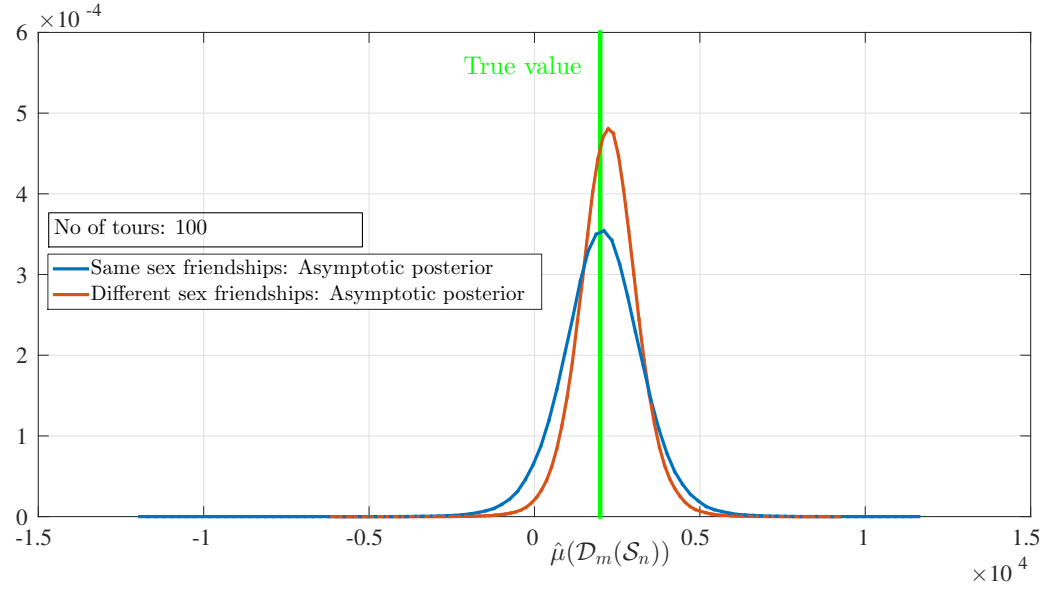


Figure 6: ADD network: effect of gender in relationships

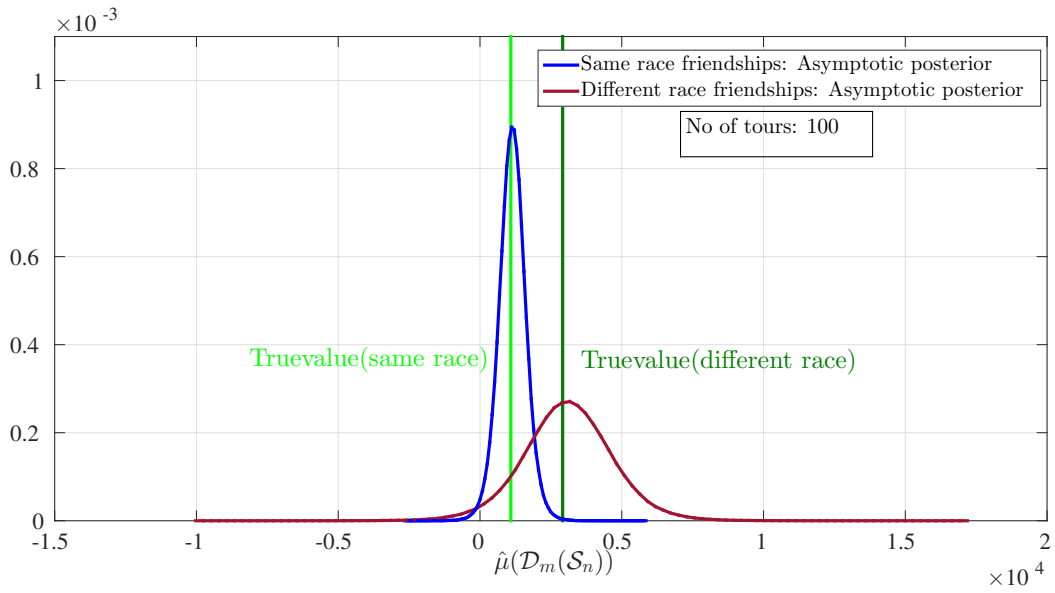


Figure 7: ADD network: effect of race in friendships

simulations clearly show that the derived posterior distribution fits very well with the data even while crawling a small percentage of the graph.

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] N. Alon, C. Avin, M. Koucký, G. Kozma, Z. Lotker, and M. R. Tuttle. Many random walks are faster than one. *Combinatorics, Probability and Computing*, 20(04):481–502, 2011.
- [3] 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.
- [4] 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.
- [5] P. Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer Science & Business Media, 2013.
- [6] 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.
- [7] H. Cramér. *Mathematical methods of statistics*. Princeton university press, 1999.
- [8] Dogster and C. friendships network dataset KONECT. <http://konect.uni-koblenz.de/networks/petster-carnivore>, May 2015.
- [9] A. Gelman. Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper), 2006.
- [10] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Practical recommendations on crawling online social networks. *IEEE JSAC*, 29(9):1872–1892, 2011.
- [11] M. Gjoka, M. Kurant, and A. Markopoulou. 2.5 k-graphs: from sampling to generation. In *Proceedings IEEE INFOCOM*, pages 1968–1976. IEEE, 2013.
- [12] S. Goel and M. J. Salganik. Respondent-driven sampling as Markov chain Monte Carlo. *Stat. Med.*, 28(17):2202–2229, 2009.
- [13] M. S. Handcock and K. J. Gile. Modeling social networks from sampled data. *Ann. Appl. Stat.*, 4(1):5–25, mar 2010.
- [14] D. D. Heckathorn. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2):174–199, 05 1997.
- [15] S. Jackman. *Bayesian analysis for the social sciences*, volume 846. John Wiley & Sons, 2009.
- [16] C. Kipnis and S. S. Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, 104(1):1–19, 1986.

- [17] J. H. Koskinen, G. L. Robins, and P. E. Pattison. Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. *Stat. Methodol.*, 7(3):366–384, may 2010.
- [18] J. H. Koskinen, G. L. Robins, P. Wang, and P. E. Pattison. Bayesian analysis for partially observed network data, missing ties, attributes and actors. *Soc. Networks*, 35(4):514–527, oct 2013.
- [19] J. Kunegis. Konect: the Koblenz network collection. In *Proceedings of the World Wide Web Companion*, pages 1343–1350, 2013.
- [20] 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. ACM, 2012.
- [21] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. <http://snap.stanford.edu/data>, June 2014.
- [22] J. G. Ligo, G. K. Atia, and V. V. Veeravalli. A controlled sensing approach to graph classification. *Signal Processing, IEEE Transactions on*, 62(24):6468–6480, 2014.
- [23] L. Massoulié, E. Le Merrer, A.-M. Kermarrec, and A. Ganesh. Peer counting and sampling in overlay networks: random walk methods. In *Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing*, pages 123–132. ACM, 2006.
- [24] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.
- [25] 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.
- [26] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.
- [27] J. Putzke, K. Fischbach, D. Schoder, and P. A. Gloor. Cross-cultural gender differences in the adoption and usage of social media platforms—an exploratory study of last. fm. *Computer Networks*, 75:519–530, 2014.
- [28] B. Ribeiro and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In *Proceedings of ACM SIGCOMM Internet Measurement Conference 2010*, pages 390–403, November 2010.
- [29] B. Ribeiro, P. Wang, F. Murai, and D. Towsley. Sampling directed graphs with random walks. In *IEEE INFOCOM*, pages 1692–1700. IEEE, 2012.
- [30] S. K. Thompson. Targeted random walk designs. *Survey Methodology*, 32(1):11, 2006.
- [31] H. C. Tijms. *A first course in stochastic models*. John Wiley and Sons, 2003.

6 Appendix

Lemma 3.

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{n} \text{var}_\pi \left(\sum_{k=2}^n f(X_{k-1}, X_k) \right) \\ &= 2 \sum_{i=2}^r \frac{\lambda_i}{1 - \lambda_i} \langle f, v_i \rangle_{\hat{\pi}} (u_i^\top \hat{f}) + \frac{1}{d_{\text{tot}}} \sum_{(i,j) \in E} f(i,j)^2 + \frac{1}{d_{\text{tot}}^2} \left(\sum_{(i,j) \in E} f(i,j)^2 \right)^2 \\ & \quad + \frac{1}{d_{\text{tot}}^2} \sum_{i \in V} d_i \left(\sum_{i \sim j} f(i,j) \right)^2 \end{aligned}$$

Proof. We extend the arguments in the proof of [5, Theorem 6.5] to the edge functions. When the initial distribution is π , we have

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{n} \text{var}_\pi \left(\sum_{k=2}^n f(X_{k-1}, X_k) \right) \\ &= \frac{1}{n} \left(\sum_{k=2}^n \text{var}_\pi(f(X_{k-1}, X_k)) + 2 \sum_{\substack{k,j=2 \\ k < j}}^n \text{cov}_\pi(f(X_{k-1}, X_k), f(X_{j-1}, X_j)) \right) \\ &= \text{var}_\pi(f(X_{k-1}, X_k)) + 2 \sum_{l=2}^{n-1} \frac{(n-1)-l}{n} \text{cov}_\pi(f(X_0, X_1), f(X_{l-1}, X_l)). \end{aligned} \quad (11)$$

Now the first term in (11) is

$$\text{var}_\pi(f(X_{k-1}, X_k)) = \langle f, f \rangle_{\hat{\pi}} - \langle f, \Pi \hat{f} \rangle_{\hat{\pi}}, \quad (12)$$

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

For the second term in (11),

$$\text{cov}_\pi(f(X_0, X_1), f(X_{l-1}, X_l)) = E_\pi(f(X_0, X_1), f(X_{l-1}, X_l)) - (E_\pi[f(X_0, X_1)])^2. \quad (13)$$

$$\begin{aligned} & E_\pi(f(X_0, X_1), f(X_{l-1}, X_l)) \\ &= \sum_i \sum_j \sum_k \sum_m \pi_i p_{ij} p_{jk}^{(l-2)} p_{km} f(i, j) f(k, m) \\ &= \sum_i \sum_j \sum_k \pi_i p_{ij} f(i, j) p_{jk}^{(l-2)} \hat{f}(k) \\ &= \sum_{(i,j) \in E} \hat{\pi}_{ij} f(i, j) (P^{(l-2)} \hat{f})(j) \\ &= \langle f, P^{(l-2)} \hat{f} \rangle_{\hat{\pi}}. \end{aligned} \quad (14)$$

Therefore,

$$\text{cov}_\pi(f(X_0, X_1), f(X_{l-1}, X_l)) = \langle f, (\mathbf{P}^{(l-2)} - \Pi) \hat{f} \rangle_{\hat{\pi}}.$$

Taking limits, we get

$$\lim_{n \rightarrow \infty} \sum_{l=2}^{n-1} \frac{n-l-1}{n} (\mathbf{P}^{(l-2)} - \Pi)$$

$$\begin{aligned}
&= \lim_{n \rightarrow \infty} \sum_{k=1}^{n-3} \frac{n-k-3}{n} (\mathbf{P}^k - \Pi) + (\mathbf{I} - \Pi) \\
&\stackrel{(a)}{=} \lim_{n \rightarrow \infty} \sum_{k=1}^{n-1} \frac{n-k}{n} (\mathbf{P}^k - \Pi) + (\mathbf{I} - \Pi) - \lim_{n \rightarrow \infty} \frac{3}{n} \sum_{k=1}^{n-3} (\mathbf{P}^k - \Pi) \\
&= (\mathbf{Z} - \mathbf{I}) + (\mathbf{I} - \Pi) = \mathbf{Z} - \Pi,
\end{aligned} \tag{15}$$

where the first term in (a) follows from the proof of [5, Theorem 6.5] and since $\lim_{n \rightarrow \infty} (\mathbf{P}^n - \Pi) = 0$, the last term is zero using Cesaro's lemma [5, Theorem 1.5 of Appendix].

We have,

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

Thus

$$\begin{aligned}
&\lim_{n \rightarrow \infty} \frac{1}{n} \text{var}_\pi \left(\sum_{k=2}^n f(X_{k-1}, X_k) \right) \\
&= \langle f, f \rangle_{\hat{\pi}} - \langle f, \Pi \hat{f} \rangle_{\hat{\pi}} + 2 \langle f, \left(\mathbf{I} + \sum_{i=2}^r \frac{\lambda_i}{1 - \lambda_i} v_i u_i^\top - \Pi \right) \hat{f} \rangle_{\hat{\pi}} \\
&= \langle f, f \rangle_{\hat{\pi}} + \langle f, \Pi \hat{f} \rangle_{\hat{\pi}} + 2 \langle f, \hat{f} \rangle_{\hat{\pi}} + 2 \sum_{i=2}^r \frac{\lambda_i}{1 - \lambda_i} \langle f, v_i \rangle_{\hat{\pi}} (u_i^\top \hat{f}) \\
&= \frac{1}{d_{tot}} \sum_{(i,j) \in E} f(i,j)^2 + \frac{1}{d_{tot}^2} \left(\sum_{(i,j) \in E} f(i,j)^2 \right)^2 + \frac{1}{d_{tot}^2} \sum_{i \in V} d_i \left(\sum_{i \sim j} f(i,j) \right)^2 \\
&\quad + 2 \sum_{i=2}^r \frac{\lambda_i}{1 - \lambda_i} \langle f, v_i \rangle_{\hat{\pi}} (u_i^\top \hat{f})
\end{aligned} \tag{16}$$

□



**RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE**

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399