Graph sampling by lagged random walk

Li-Chun Zhang[1,2,3]

[1] *Statistisk sentralbyraa, Norway*
[2] *University of Southampton, UK (L.Zhang@soton.ac.uk)*
[3] *Universitetet i Oslo, Norway*

Abstract

We propose a family of lagged random walk sampling methods in simple undirected graphs, where transition to the next state (i.e. node) depends on both the current and previous states—hence, lagged. The existing random walk sampling methods can be incorporated as special cases. We develop a novel approach to estimation based on lagged random walks at equilibrium, where the target parameter can be any function of values associated with finite-order subgraphs, such as edge, triangle, 4-cycle and others.

Key words: random jump, stationary distribution, non-Markovian process, generalised ratio estimator, capture-recapture estimator

1 Introduction

Let $G = (U, A)$ be a simple undirected graph, with known node set U and *order* $N = |U|$, but unknown edge set A or *size* $R = |A|$. Discrete-time random walk (RW) in G can only move from the current state $X_t = i$ to the next $X_{t+1} = j$ if $(ij) \in A$, where $i, j \in U$, such that the walk is confined within the component of G where the initial state X_0 is located, given any $X_0 \in U$. For a *targeted random walk (TRW,* Thompson, 2006), the transition probabilities $\Pr(X_{t+1} = j|X_t = i)$ at each time step t can be subject to other devices, such as random jumps or an acceptance-rejection mechanism for the proposed moves. The TRW stationary distribution has many applications to large often dynamic networks, such as PageRank (Brin and Page, 1998); see Masuda et al. (2017) and Newman (2010) for reviews.

$$1 \text{ —— } 2 \xrightarrow{p} 3 \xrightarrow{1-p} 4 \text{ —— } 5$$

Figure 1: Lagged random walk on a subway line, $(X_{t-1}, X_t) = (2, 3)$

We call a random walk *lagged* if the transition probabilities at time step t are instead given by $\Pr(X_{t+1} = j|X_t = h, X_{t-1} = i)$. The process $\{X_t : t \geq 0\}$ is non-Markovian when the states are generated by a *lagged random walk (LRW)*. Figure 1 provides a simple setting of LRW on a subway line of 5 stations. Fixing $p = 0.5$ for $X_t \in \{2, 3, 4\}$ yields pure RW, whereas normal subway service corresponds to

$$p = \Pr(X_{t+1} = i \mid X_t = h, X_{t-1} = i) = 0$$

for any $h \in \{2, 3, 4\}$. The LRW has a tendency to persist in the same travel direction if $p < 0.5$, to backtrack to the previous station if $0.5 < p < 1$, or stucks between two stations if $p = 1$.

Generally, by an extra parameter that controls the tendency for backtracking in arbitrary graphs, LRW yields a family of sampling methods which includes TRW as a special case.

Let $G(M)$ be a subgraph induced by M, with node set M, $M \subset U$, and edge set $A \cap \{(ij) : i, j \in M\}$. The specific characteristics of $G(M)$ is called *motif* (Zhang and Patone, 2017; Zhang, 2021), denoted by $[M]$, and the order of $[M]$ is $|M|$, i.e. that of $G(M)$. Figure 2 illustrates some low-order motifs. For instance, edge is a 2nd-order motif which can be given by $[\{i, j\} : a_{ij} = 1]$, where $a_{ij} = 1$ if $(ij) \in A$ and 0 otherwise, for $i, j \in U$; any particular edge in A is an *occurrence* of the edge motif in G. Or, triangle is a 3rd-order motif given by $[\{i, j, k\} : a_{ij}a_{jk}a_{ki} = 1]$, where $i, j, k \in U$; any particular triangle in G is an occurrence of the triangle motif.

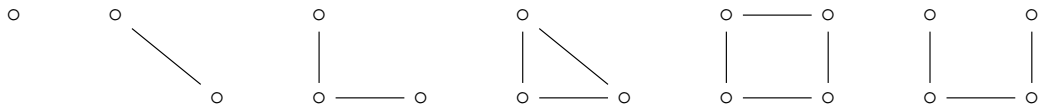

Figure 2: From left to right, motif node, edge, 2-star, triangle, 4-cycle and 3-path

Let Ω contain all the occurrences of the motifs of interest in G. For each $\kappa \in \Omega$, let y_κ be a fixed value associated with κ. The corresponding *graph total*, denoted by θ, is given by

$$\theta = \sum_{\kappa \in \Omega} y_\kappa \tag{1}$$

whereas we refer to any function of $\{y_\kappa : \kappa \in \Omega\}$ as a *graph parameter*, denoted by μ.

For instance, let $\Omega = U$, and let $y_\kappa = 1$ or 0 be a binary indicator of the type of node κ, then θ by (1) is the total number of type-1 nodes in G, and $\mu = \theta/N$ is a 1st-order graph parameter measuring their prevalence. Let $\Omega = A$ and $y_\kappa = 1$ for any $\kappa \in A$, then $\theta = R$ is the size of G. Let Ω contain all the triangles and 2-stars in G, and let θ be the graph total of triangles and θ' that of 2-stars, then $\mu = \theta/(\theta + \theta')$ is a measure of the transitivity of G.

Previously, Thompson (2006) and Avrachenkov et al. (2010) have considered the estimation of 1st-order graph parameters based on TRW sampling, and Avrachenkov et al. (2016) that of the 2nd-order graph totals. Cooper et al. (2016) propose an approach using the times of the first returns to high-degree nodes. Below, in Section 2, we propose a general family of LRW sampling methods. We develop in Section 3 an approach to estimating finite-order graph parameters and totals as defined above. Numerical illustrations are given in Section 4, and some remarks on future research in Section 5.

2 LRW sampling

Let d_i be the degree of node i in U. Let $r > 0$ and $0 \leq w \leq 1$ be two chosen constants regulating random jump and backtracking, respectively. Let the transition probability of LRW be

$$\begin{aligned} &\Pr(X_t = j \mid X_t = h, X_{t-1} = i) \\ &\quad = \begin{cases} \frac{r}{d_h + r}\left(\frac{1}{N}\right) + \frac{a_{hj}}{d_h + r} & \text{if } d_h = 1 \\ \frac{r}{d_h + r}\left(\frac{1}{N}\right) + \mathbb{I}(j = i)\frac{wa_{hj}}{d_h + r} + \mathbb{I}(j \neq i)\frac{d_h - wa_{ih}}{d_h + r}\frac{a_{hj}}{d_h - a_{ih}} & \text{if } d_h > 1 \end{cases} \end{aligned} \tag{2}$$

That is, the walk either jumps randomly to any node j in U–including the current X_t–with probability $r/(d_h + r)$, or it moves to an adjacent node j with probability $d_h/(d_h + r)$. If $d_h = 1$, then there is only one adjacent j; if $d_h > 1$, then it either backtracks to the previous $X_{t-1} = i$ with probability $w/(d_h + r)$, or it moves randomly to one of the other adjacent nodes.

2.1 Stationary distribution

TRW of Avrachenkov et al. (2010) is the special case of LRW by (2) given $w = 1$, where the previous state X_{t-1} (if adjacent) is treated as any other adjacent nodes at step t and $\{X_t : t \geq 0\}$

form a Markov chain. The stationary probability is given by

$$\pi_h \coloneqq \Pr(X_t = h) = \frac{d_h + r}{2R + rN} \tag{3}$$

The process $\{X_t : t \geq 0\}$ is non-Markovian if $w < 1$. However, let $\boldsymbol{x}_t = (X_{t-1}, X_t)$ for $t \geq 1$. Given any initial $\boldsymbol{x}_1 = (X_0, X_1)$, LRW (2) generates a Markov chain $\{\boldsymbol{x}_t : t \geq 1\}$, since

$$\Pr(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, ..., \boldsymbol{x}_1) = \Pr(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) = \Pr(X_{t+1}|\boldsymbol{x}_t)$$

Due to random jumps, this Markov chain is irreducible, such that there exists a unique stationary distribution, denoted by

$$\pi_{\boldsymbol{x}} \coloneqq \Pr(\boldsymbol{x}_t = \boldsymbol{x}) \qquad \text{and} \qquad \sum_{\boldsymbol{x}} \pi_{\boldsymbol{x}} = 1$$

A unique stationary distribution of X_t follows, which satisfies the *mixed* equation

$$\pi_h = \sum_{i \in U} \Pr\big(\boldsymbol{x}_t = (i, h)\big) = \sum_{\substack{\boldsymbol{x}=(i,h) \\ i \in \nu_h}} \pi_{\boldsymbol{x}} + \sum_{i \notin \nu_h} \frac{\pi_i}{d_i + r}\big(\frac{r}{N}\big) \tag{4}$$

where $\nu_h = \{i \in U : a_{ih} = 1\}$ is the neighbourhood of adjacent nodes (to h), and the transition to h from any node outside ν_h can only be accomplished by a random jump.

Lemma 1. *For LRW by* (2) *with any* $0 \leq w \leq 1$ *in undirected simple graphs, the stationary probability* π_h *is given by* (3) *for any* $h \in U$*, and* $\pi_{\boldsymbol{x}}$ *for any* $\boldsymbol{x} = (i, h)$ *is given by*

$$\pi_{\boldsymbol{x}} \propto \begin{cases} 1 + r/N & \text{if } \ i \in \nu_h \\ r/N & \text{if } \ i \notin \nu_h \end{cases}$$

Proof. Under LRW (2), balanced flows between $\boldsymbol{x}_t = (i, h)$ and $\boldsymbol{x}_{t+1} = (h, j)$ are the same as balanced flows over (X_{t-1}, X_t, X_{t+1}) in the corresponding directions. To show the values $\{p_h = d_h + r : h \in U\}$ satisfy the balanced flows through (X_{t-1}, X_t, X_{t+1}) at equilibrium, one needs to consider three situations.

I. $\qquad (i = j) \text{ —— } h \qquad\qquad\qquad i \text{ —— } h \text{ —— } j$

I. $\{i, j\} \in \nu_h$, as illustrated. If $i = j$, then (i, h, j) and (j, h, i) are backtracking in either direction, the probability of which is the same by (2), so that they are always balanced. If $i \neq j$, then neither (i, h, j) nor (j, h, i) is backtracking, with the same transition probability but $\pi_{\boldsymbol{x}_t}$ differ to $\pi_{\boldsymbol{x}_{t+1}}$ generally. However, $\boldsymbol{x} = (i, h)$ for any given $i \in \nu_h$ is involved in one flow in *either* direction, such that *altogether* these flows are balanced.

II. $\qquad i \qquad h \qquad j$

II. $\{i, j\} \notin \nu_h$, as illustrated, including $i = h$ or $j = h$, in which case (i, h, j) and (j, h, i) can only take place by random jumps, which are balanced given $p_h = d_h + r$, $\forall h \in U$, since

$$p_i \frac{r}{d_i + r}\big(\frac{1}{N}\big) \frac{r}{d_h + r}\big(\frac{1}{N}\big) = p_j \frac{r}{d_j + r}\big(\frac{1}{N}\big) \frac{r}{d_h + r}\big(\frac{1}{N}\big)$$

III. $\qquad i \text{ —— } h \qquad j \qquad\qquad\qquad i \text{ —— } (h = j)$

III. One of (i, j), say, i belongs to ν_h but not the other, including when $j = h$, as illustrated.

Setting $p_i = d_i + r$ for π_i in the second term on the right-hand side of (4), we have

$$\big(\pi_h - \sum_{\substack{\boldsymbol{x}=(i,h)\\ i\in\nu_h}} \pi_{\boldsymbol{x}}\big)\big(\frac{d_h}{d_h+r} + \frac{r}{d_h+r}\frac{d_h}{N}\big) = r\frac{N-d_h}{N}\big(\frac{d_h}{d_h+r} + \frac{r}{d_h+r}\frac{d_h}{N}\big)$$

i.e. summing over *all* possible flows (j,h,i) where $j \notin \nu_h$ and $i \in \nu_h$. Whereas, summing over all the possible flows (i,h,j) in the other direction, we have

$$\big(\sum_{\substack{\boldsymbol{x}=(i,h)\\ i\in\nu_h}} \pi_{\boldsymbol{x}}\big)\frac{r}{d_h+r}\big(\frac{N-d_h}{N}\big)$$

Balancing (i.e. equating) the two groups of flows at equilibrium yields

$$\sum_{\substack{\boldsymbol{x}=(i,h)\\ i\in\nu_h}} \pi_{\boldsymbol{x}} = d_h + r\frac{d_h}{N}$$

Replacing this into the right-hand side of (4), we obtain $p_h = d_h + r$. Summarising all the three situations, the values p_h satisfy the balanced flows, so that $\pi_h \propto d_h + r$ as in (3). Finally, $\pi_{\boldsymbol{x}}$ for any $\boldsymbol{x} = (i,h)$ follows from (4) by symmetry. □

2.2 Sample graph by LRW

LRW can be used to select subgraphs of G following the definition of sample graph given by Zhang (2021). By the *observation procedure (OP)* of LRW, we observe all the edges incident to $X_t = i$ at step t, i.e. the entire row and column of the adjacency matrix of G corresponding to node i. Let $s = \{X_0, ..., X_T\}$ be the *seed sample* of LRW, which are defined as the nodes to which the OP is applied. For any given s, the observed part of the adjacency matrix can be specified as $s \times U \cup U \times s$, such that the sampled (or observed) edges are given by

$$A_s = A \cap \big(s \times U \cup U \times s\big)$$

The *sample graph* by LRW, denoted by G_s, is given by

$$G_s = (U_s, A_s) \qquad \text{where} \qquad U_s = s \cup \mathrm{Inc}(A_s) \tag{5}$$

i.e. the node sample is the union of s and all the nodes incident to the edges in A_s.

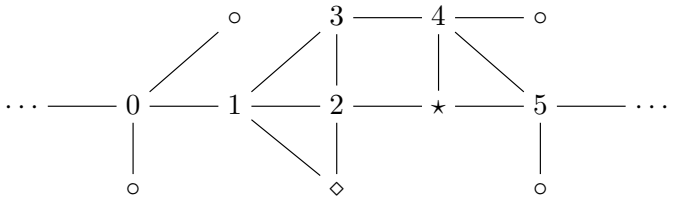

Figure 3: Illustration of sample observation by LRW

Figure 3 illustrates how one can observe various finite-order motifs by LRW sampling. Let $X_{t+q} = q$ for $q = 0, 1, ..., 5$. In addition to the nodes $\{X_t, ..., X_{t+5}\}$ we observe many other motifs, such as edge $[\{0,1\}]$ at $X_t = 0$ and again at $X_{t+1} = 1$, triangle $[\{1,2,3\}]$ and $[\{1,2,\diamond\}]$ given $(X_{t+1}, X_{t+2}) = (1,2)$ and $[\{1,2,3\}]$ again given $(X_{t+2}, X_{t+3}) = (2,3)$, 4-cycle $[\{2,3,4,\star\}]$ given $(X_{t+2}, X_{t+3}, X_{t+4}) = (2,3,4)$. All these motifs are observed in the sample graph G_s by LRW. Denote by Ω_s the elements of Ω in (1) which are observed in G_s.

3 Estimation

We consider first the estimation of R, i.e. the size of G, hence the stationary probability (3) by LRW sampling. Next, we consider other finite-order graph parameters and totals. In applications one can use the mean of a given estimator from any number, say, K independent LRWs as the final estimator, and use the empirical variance of the K estimators for variance estimation.

3.1 Size of graph

Let $\{X_{t_1}, ..., X_{t_{n_x}}\}$ be an extraction of n_x states from LRW $\{X_t : t \geq 0\}$ at equilibrium, where $t_1, ..., t_{n_x}$ do not need to be consecutive. Similarly, let $\{Y_{\tau_1}, ..., Y_{\tau_{n_y}}\}$ be those of a separate independent LRW $\{Y_\tau : \tau \geq 0\}$. We have

$$E(m) \coloneqq E\Big(\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\frac{1}{d_h + r}\mathbb{I}(Y_{\tau_j} = h \mid X_{t_i} = h)\Big) = \frac{n_x n_y}{2R + rN}$$

where m can be referred to as the number of collisions by $\{X_t\}$ and $\{Y_\tau\}$. This yields

$$\hat{R}_1 = (n_x n_y/m - rN)/2 \tag{6}$$

as a *capture-recapture (CR)* estimator. Meanwhile, since

$$E(\bar{d}_w) \coloneqq E\left(\frac{\sum_{i=1}^{n_x}\frac{d_{X_{t_i}}}{d_{X_{t_i}}+r} + \sum_{j=1}^{n_y}\frac{d_{Y_{\tau_j}}}{d_{Y_{\tau_j}}+r}}{\sum_{i=1}^{n_x}\frac{1}{d_{X_{t_i}}+r} + \sum_{j=1}^{n_y}\frac{1}{d_{Y_{\tau_j}}+r}}\right) \approx \frac{2R}{N}$$

we obtain another *generalised ratio (GR)* estimator of R as

$$\hat{R}_2 \ = N\bar{d}_w/2 \tag{7}$$

Finally, combining the two, we obtain a GR-CR estimator, denoted by $\hat{R}_3$, by incorporating an estimator $\hat{N}$ although N is known, i.e.

$$\begin{cases} 2m\hat{R}_3 + rm\hat{N} = n_x n_y \\ -2\hat{R}_3 + \bar{d}_w\hat{N} = 0 \end{cases} \quad \Rightarrow \quad \hat{N} = \frac{n_x n_y}{m(r + \bar{d}_w)} \quad \text{and} \quad \hat{R}_3 \ = \frac{n_x n_y \bar{d}_w}{2m(r + \bar{d}_w)} \tag{8}$$

3.2 Finite-order graph parameters and totals

Below we first clarify generally the basis of estimation under LRW sampling, before we define the estimators of finite-order graph parameters and totals.

3.2.1 Basis of estimation

Let $M = \{X_{t+1}, ..., X_{t+q}\}$ be a set of successive states by LRW. At equilibrium, we have

$$\pi_M = \Pr(X_{t+1}, ..., X_{t+q}) = \pi_{X_{t+1}}\prod_{i=1}^{q-1} p_{X_{t+i}X_{t+i+1}} \tag{9}$$

where $p_{X_{t+i}X_{t+i+1}}$ is the transition probability from X_{t+i} to X_{t+i+1}, which is known regardless the unvisited nodes of G. We shall refer to π_M by (9) as the *stationary successive sampling probability (S3P)*, where the probability $\pi_{X_{t+1}}$ is known exactly up to the proportionality constant $2R+rN$, or it can be estimated via $\hat{R}$ in which case we write $\hat{\pi}_{X_{t+1}}$ and $\hat{\pi}_M$.

Apart from the actual sequences in a LRW, the S3P (9) is also available for any *hypothetical* sequence M, given $M \subseteq s$, because the part of the transition probability matrix corresponding to $s \times s$ is already observed. For instance, given the actual sequence $(X_t, X_{t+1}, X_{t+2}) = (1, 2, 3)$ in Figure 3, we can as well calculate $\pi_3 p_{32} p_{21}$ of a sampling sequence $(X_t, X_{t+1}, X_{t+2}) = (3, 2, 1)$ hypothetically. Given the seed sample s, all the sequences for which the S3P is available belong to the *generating (sequences of) states* of LRW sampling given by

$$\mathcal{C}_s = \{M : M \subseteq s\}$$

The subset containing the parts of the actual walk is given by

$$\mathcal{C}_w = \big\{\{X_t, ..., X_{t+q}\} : 0 \leq t \leq t+q \leq T\big\}$$

For example, suppose LRW with $X_0 = 1$ and $X_T = 4$ in Figure 3, for which we have

$$\mathcal{C}_w = \big\{\{1\}, \{1,2\}, \{1,2,3\}, \{1,2,3,4\}, \{2\}, \{2,3\}, \{2,3,4\}, \{3\}, \{3,4\}, \{4\}\big\}$$

Let the sample motif κ, $\kappa \in \Omega_s$, be observed from the *actual sampling sequence of states (AS3)* $s_\kappa = (X_t, ..., X_{t+q})$, for some time step t and $q = |s_\kappa| - 1$. An *equivalent sampling sequence of states (ES3)*, denoted by $\tilde{s}_\kappa \sim s_\kappa$ including $s_\kappa \sim s_\kappa$, is any possible sequence of states of the same length $|\tilde{s}_\kappa| = |s_\kappa|$, such that the motif κ would be observed if $(X_t, X_{t+1}, ..., X_{t+q}) = \tilde{s}_\kappa$ but not based on any subsequence of $\tilde{s}_\kappa$. In other words, a motif κ observed from AS3 s_κ can be sampled otherwise sequence-by-sequence, in which respect its ES3 set

$$F_\kappa = \{\tilde{s}_\kappa : \tilde{s}_\kappa \sim s_\kappa\}$$

constitute the multiplicity of access to κ under LRW sampling, similarly to the concept of multiplicity under indirect sampling (Birnbaum and Sirken, 1965).

Take the triangle κ with $M = \{1, 2, 3\}$ in Figure 3, where $s_\kappa = (X_t, X_{t+1}) = (1, 2)$ is an AS3, for which we have $F_\kappa = \{(1,2), (2,1), (1,3), (3,1), (2,3), (3,2)\}$, since the same κ would be observed if (X_t, X_{t+1}) is any of these 6 sequences in F_κ.

3.2.2 Estimators

Given LRW sampling from G, choose a seed sample of successive states at equilibrium which is of size n, denoted by $s = \{X_1, ..., X_n\}$. Obtain the corresponding motif sample Ω_s, the generating states $\mathcal{C}_s$ and its subset $\mathcal{C}_w$.

For any $\kappa \in \Omega_s$ observed from AS3 $s_\kappa = (X_t, ..., X_{t+q})$, where $s_\kappa \in \mathcal{C}_w$, let F_κ be its ES3 set. Let M be a possible sequence of states $(X_t, ..., X_{t+q})$. Let the sampling indicator $\delta_M = 1$ if M is realised and 0 otherwise, with the associated S3P π_M, over hypothetically repeated sampling. Let the observation indicator $I_\kappa(M) = 1$ if $M \in F_\kappa$ and 0 otherwise. Let $w_{M\kappa}$ be the *incidence weight* for any $M \in F_\kappa$, where $w_{M\kappa} = 0$ if $M \notin F_\kappa$, such that for any $\kappa \in \Omega$, we have

$$\sum_{M \in F_\kappa} w_{M\kappa} = \sum_M I_\kappa(M) w_{M\kappa} = 1$$

An estimator of the graph total (1) is given by

$$\hat{\theta}(X_t, ..., X_{t+q}) = \sum_{\kappa \in \Omega} \sum_M \frac{\delta_M}{\pi_M} I_\kappa(M) w_{M\kappa} y_\kappa \tag{10}$$

which is unbiased for θ, since

$$E\big(\hat{\theta}(X_t,...,X_{t+q})\big) = \sum_{\kappa\in\Omega} y_\kappa \sum_M I_\kappa(M) w_{M\kappa} = \sum_{\kappa\in\Omega} y_\kappa$$

Although there are infinitely many possibilities in principle, the two basic choices of $w_{M\kappa}$ are the *multiplicity weight* given by

$$w_{M\kappa} = 1/|F_\kappa| \tag{11}$$

for any M in F_κ, and the *proportional-to-probability weight (PPW)* given by

$$w_{M\kappa} = \pi_M/\pi_{(\kappa)} \qquad \text{and} \qquad \pi_{(\kappa)} = \sum_{M\in F_\kappa} \pi_M \tag{12}$$

Notice that the PPW (12) is only possible if $F_\kappa \subseteq \mathcal{C}_s$.

Given the AS3 s_κ is of the order $|s_\kappa| = q+1$ and $|s| = n$, there are at most $n-q$ possible estimators based on $(X_t,...,X_{t+q})$ by (10), denoted by $\hat{\theta}_t$ for $t = 1,...,n-q$. Let $\mathbb{I}_t = 1$ indicate that $(X_t,...,X_{t+q})$ can lead to the observation of at least one motif in Ω, and $\mathbb{I}_t = 0$ otherwise. An estimator of θ combining all the $\hat{\theta}_t$ is given by

$$\hat{\theta} = \Big(\sum_{t=1}^{n-q} \mathbb{I}_t\hat{\theta}_t\Big)/\Big(\sum_{t=1}^{n-q} \mathbb{I}_t\Big)\ . \tag{13}$$

Generally, the estimators (10) and (13) can only be calculating using $\hat{\pi}_M$ instead of π_M, which requires an estimate of R. However, it is also possible to use π_M directly to estimate graph parameters regardless the proportionality constant $2R + rN$. For instance, let $\mu = \theta/N_\Omega$, where $N_\Omega = |\Omega|$. A generalised ratio estimator is given by

$$\hat{\mu} = \hat{\theta}/\hat{N}_\Omega$$

where $\hat{N}_\Omega$ is also given by (10) and (13) but with $y_\kappa \equiv 1$. In the special case of $\Omega = U$, this reduces to the GR estimator of population mean $\sum_{i\in U} y_i/N$ (Thompson, 2006), as a 1st-order graph parameter. Generally one can estimate any μ that is a function of graph totals, as long as the plug-in $\hat{\mu}$ is invariant when each total involved is replaced by its estimator (13).

4 Illustration

Below we explore the convergence rate of LRW and illustrate some associated estimators of 1st, 2nd and 3rd-order graph totals and parameters.

Let $G = (U, A)$ be simple and undirected with $N = |U| = 100$. Let $y = 1$ be the value associated with the first 20 nodes $i = 1,...,20$, to be referred to as the cases, and let $y = 0$ be the value for the rest 80 nodes or the noncases. The edges (ij) are generated randomly, with different probabilities according to $y_i + y_j = 2$, 1 or 0. The resulting graph has 299 edges, $R = |A| = 299$; the cases have an average degree 13.5, and the noncases have an average degree 4.1. This valued graph G with a mild core-periphery structure is fixed for the illustrations below.

4.1 Convergence to equilibrium

Let $p_{t,i} = \Pr(X_t = i)$, for $i \in U$. We have $p_{t,i} \to \pi_i \propto d_i + r$ by LRW, as $t \to \infty$. How quickly a walk reaches equilibrium is affected by the selection of the initial state X_0. In the extreme case the walk is at equilibrium from the beginning if $\Pr(X_0 = i) = \pi_i$ for $i \in U$. To explore the speed of convergence, consider two other choices $p_{0,i} = \Pr(X_0 = i) \equiv 1/N$ or $p_{0,1} = 1$.

To track the convergence empirically, we use B independent simulations of LRW to estimate $E(Y_t) = \sum_{i \in U} p_{t,i} y_i$, which is targeted at the equilibrium expectation $E(Y_\infty) = \sum_{i \in U} \pi_i y_i$.

Table 1: $E(Y_t)$ by t, r and initial X_0, 10^5 simulations of LRW with $w = 1$

	$r = 1,\ E(Y_\infty) = 0.415$				$r = 0.1,\ E(Y_\infty) = 0.447$			
Initiation	$t = 1$	$t = 4$	$t = 8$	$t = 16$	$t = 1$	$t = 4$	$t = 8$	$t = 16$
$p_{0,i} = \pi_i$	0.420	0.419	0.421	0.414	0.449	0.451	0.450	0.448
$p_{0,i} = 1/N$	0.341	0.394	0.408	0.413	0.355	0.412	0.433	0.449
$p_{0,1} = 1$	0.714	0.481	0.433	0.413	0.741	0.529	0.471	0.448

Table 1 shows the results for $t = 1, 4, 8, 16$ and $r = 1, 0.1$, each based on 10^5 simulations of LRW with $w = 1$. The equilibrium expectation $E(Y_\infty)$ varies with r. Of the two choices here, the value $r = 1$ yields the degree+1 walk, whereas the value $r = 0.1$ tunes the walk closer to pure RW. It can be seen that the walk stays at equilibrium if $p_{0,i} = \pi_i$. Under the current set-up, convergence to equilibrium can apparently be expected at $t = 16$, whether the initial X_0 is selected completely randomly from U given $p_{0,i} = 1/N$, or fixed at $i = 1$ given $p_{0,1} = 1$. Neither does the speed of convergence vary much for the values of r here.

4.2 Estimation of case prevalence

Consider the 1st-order parameter $\mu = \sum_{i \in U} y_i/N$. Let $s = \{X_0, ..., X_T\}$, where X_0 is simply drawn with π_i now that convergence to equilibrium is quick and what matters for the estimation is mainly the size of the seed sample size itself. Let ψ be the *traverse* of the walk, given as the ratio between the number of distinct nodes visited by the walk and N, which indicates how extensively the walk has travelled through G.

Given (T, r, w), generate LRW independently B times, each resulting in a replicate of $\hat{\mu}$, the mean of which is an estimate of $E(\hat{\mu})$ by LRW sampling at equilibrium, denoted by Mean($\hat{\mu}$), and the square root of their empirical variance–SD($\hat{\mu}$)–is an estimate of the standard error of $\hat{\mu}$.

Table 2: Estimation of case prevalence $\mu = 0.2$, 10^3 simulations

	$r = 0.1$			$r = 6$		
$T = 50$	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	Mean(ψ)	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	Mean(ψ)
$w = 1$	0.211	0.093	0.321	0.203	0.070	0.383
$w = 0.01$	0.210	0.079	0.367	0.203	0.067	0.395
	$r = 0.1$			$r = 6$		
$T = 100$	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	Mean(ψ)	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	Mean(ψ)
$w = 1$	0.204	*0.066*	0.499	0.200	0.049	0.607
$w = 0.01$	0.203	*0.054*	0.559	0.201	0.048	0.617

Table 2 gives the results for $T = 50$ or 100, $r = 6$ or 0.1, and $w = 1$ or 0.01, each based on 10^3 simulations. Since the case nodes have larger degrees than the noncase nodes in G, the stationary probability π_i depends on y_i and LRW sampling is informative in this sense. Nevertheless, this is not an issue for design-based estimation of graph parameters, and the consistency of $\hat{\mu}$ is already evident at $T = 50$.

Mean(ψ) shows that in the current setting a TRW (i.e. $w = 1$) of length $T = 50$ is expected to visit only about a third of all the nodes in G, and one of length $T = 100$ can reach about half of the nodes. How quickly TRW traverses the graph depends on the chance of visiting isolated nodes by random jumps as well as the chance of backtracking to the previous adjacent node, the probabilities of both are reduced given small r and w. By reducing the chance of

backtracking, LRW with $w = 0.01$ can speed up the traverse. However, increasing the chance of random jumps, e.g. $r = 6$ instead of $r = 0.1$, can speed up the traverse even more. Note that since the average degree $2R/N$ is about 6 in G here, setting $r = 6$ in (2) makes a random jump on average at least as probable as an adjacent move at each time step.

Given that we are estimating a 1st-order parameter here, every distinct node encountered contributes more effectively to the estimation than revisiting a node. As can be seen from $\mathrm{SD}(\hat{\mu})$, increasing r from 0.1 to 6 is more beneficial to the sampling efficiency than reducing w from 1 to 0.01. Nevertheless, reducing w can greatly reduce the sampling variance of TRW at $r = 0.1$, e.g. by about 1/3 given $T = 100$ as highlighted (italic) in Table 2.

4.3 Estimation of graph size

Consider the three estimators (6), (7) and (8) of the 2nd-order graph total R. We use simply $n_x = n_y = n$ for each pair of independent $\{X_t\}$ and $\{Y_t\}$. Table 3 gives the results for $n = 50$ or 100, $r = 6$ or 0.1, and $w = 1$ or 0.01, each based on 10^4 simulations, where an estimate of the expectation of each $\hat{R}$ is given together with its simulation error in the parentheses.

Table 3: Estimation of graph size $R = 299$ (SD in parentheses), 10^4 simulations

	$r = 0.1$			$r = 6$		
$n = 50$	$\hat{R}_1$	$\hat{R}_2$	$\hat{R}_3$	$\hat{R}_1$	$\hat{R}_2$	$\hat{R}_3$
$w = 1$	344 (1.64)	306 (0.41)	344 (*1.61*)	333 (1.61)	300 (0.26)	316 (0.81)
$w = 0.01$	319 (0.93)	304 (0.36)	319 (*0.91*)	329 (1.43)	300 (0.26)	314 (0.72)
	$r = 0.1$			$r = 6$		
$n = 100$	$\hat{R}_1$	$\hat{R}_2$	$\hat{R}_3$	$\hat{R}_1$	$\hat{R}_2$	$\hat{R}_3$
$w = 1$	311 (0.61)	303 (0.31)	311 (0.60)	307 (0.70)	299 (0.19)	303 (0.36)
$w = 0.01$	305 (0.42)	302 (0.27)	305 (0.41)	306 (0.65)	299 (0.18)	302 (0.33)

The traverse for given (T, r, w) is already reported earlier. At $n = 50$, the average number of collisions between $\{X_t\}$ and $\{Y_t\}$ is about 4 given $r = 0.1$, which is about halved given $r = 6$; at $n = 100$, it is about 16 given $r = 0.1$, and about halved given $r = 6$. Reducing $w = 1$ to 0.01 improves the CR estimator $\hat{R}_1$ more than increasing $r = 0.1$ to 6. The improvement is about the same with either device for the GR-CR estimator $\hat{R}_3$. The GR estimator is the best of the three here, because the number of collisions is still quite low whether $n = 50$ or 100.

Regardless the estimator, LRW sampling with $w = 0.01$ is more efficient than TRW sampling (i.e. $w = 1$). The gain is largest for the two CR-related estimators, which is most useful when the TRW sampling variance is largest, e.g. as highlighted in Table 3.

4.4 Estimation of a 3rd-order graph total and a related parameter

Let θ be the total number of triangles in G, and θ_1 that of the case triangles, i.e. $\prod_{i \in M} y_i = 1$ if $[M]$ is triangle. The 3rd-order graph parameter $\mu = \theta_1/\theta$ is a measure of the transitivity among cases compared to the overall transitivity in G. We have $(\theta, \theta_1, \mu) = (170, 140, 0.824)$ here.

Let Ω contain all the triangles in G. Given $(X_t, X_{t+1}) = (i, j)$ of two adjacent nodes under LRW, one observes all the triangles involving i and j. For instance, both the triangles of $\{1, 2, 3\}$ and $\{1, 2, \diamond\}$ in Figure 3 are observed from $(X_t, X_{t+1}) = (1, 2)$, such that for (10) we have $\delta_M = 1$ if $M = (1, 2)$ and $I_\kappa(M) = 1$ if κ is either of these two triangles.

Whereas μ can be estimated using π_M directly, $\hat{\pi}_M = \pi_M(\hat{R})$ is needed for θ or θ_1. Given any $(X_t, X_{t+1}) = (i, j)$ as the AS3 of a triangle κ, the ES3s are the six possible adjacent moves along the triangle. The incidence weight is the same, i.e. $w_{M\kappa} \equiv 1/6$, either by (11) or (12),

because the S3P (9) of LRW (2) is a constant of $M = (i, j)$, i.e.

$$\pi_M = \pi_i p_{ij} \propto 1 + r/N$$

Table 4: Estimation of $\mu = \theta_1/\theta = 0.824$ ($B = 10^3$) and $\theta = 170$ ($B = 10^4$)

	$r = 0.1$			$r = 6$		
$T = 50$	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	$\hat{\theta}$	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	$\hat{\theta}$
$w = 1$	0.796	0.108	185 (0.91)	0.796	0.125	174 (0.84)
$w = 0.01$	0.797	0.101	180 (0.78)	0.799	0.124	175 (0.82)
	$r = 0.1$			$r = 6$		
$T = 100$	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	$\hat{\theta}$	Mean($\hat{\mu}$)	SD($\hat{\mu}$)	$\hat{\theta}$
$w = 1$	0.825	*0.082*	177 (0.63)	0.815	0.101	172 (0.59)
$w = 0.01$	0.826	*0.073*	175 (0.55)	0.811	0.096	173 (0.58)

Table 4 gives the results for $T = 50$ or 100, $r = 6$ or 0.1, and $w = 1$ or 0.01. The number of simulations is $B = 10^3$ for μ and $B = 10^4$ for θ. For $\hat{\theta}$ an estimate of Mean($\hat{\theta}$) is given together with its simulation error in the parentheses. The speed of convergence is noticeably quicker for $\hat{\theta}$ with $r = 6$ instead of 0.1, but does not differ much for $\hat{\mu}$. On the one hand, intuitively a large chance of random jump is unlikely to be efficient as the order of motif increases; on the other hand, the total estimator $\hat{\theta}$ uses $\pi_M(\hat{R})$ instead of π_M, and its convergence rate seems to have benefitted from the correlation with $\hat{R}$. Again, the extra flexibility of LRW can be useful in certain situations, e.g. as highlighted in Table 4.

5 Some remarks on future research

Theoretical results on the convergence rate only seem to be available for pure random walks in connected graphs; see e.g Boyd et al. (2004), Ben-Hamou et al. (2019a). An extreme case of lagged random work, referred to as non-backtracking random walk, has been shown to mix faster in regular graphs (Alon et al., 2007) and certain irregular graphs (Ben-Hamou et al., 2019; Palahan, 2015). The empirical results of LRW sampling here suggest that longer walks are needed for estimating totals using $\hat{\pi}_M$ than estimating parameters using π_M directly, but general theoretical results are lacking at the moment.

What is a good choice of w can conceivably vary for different motifs. Using a small w with nearly no backtracking if possible seems to be reasonable in all the illustrations considered here, especially when the walk has a relatively low probability for random jumps. More systematic investigation is needed in this respect.

The choice of incidence weights $w_{M\kappa}$ over F_κ is nontrivial generally, although the two basic choices (11) or (12) happen to coincide for the triangle motif. For instance, let $M = \{i, j, g, h\}$ form a 4-cycle, which can be observed based on AS3 consisting of two successive adjacent moves, say, $(X_t, X_{t+1}, X_{t+2}) = (i, j, g)$. The corresponding S3P (9) is $\pi_i p_{ij} p_{jg} = \frac{1}{d_j + r}\big(1 + \frac{r}{N}\big)^2$, such that the PPW (12) would differ to the multiplicity weight (11). Since the PPW requires $F_\kappa \subseteq \mathcal{C}_s$ but not the equal weights, the latter can be applied to a larger number of sample motifs. In practice one can calculate several estimators using different incidence weights and choose among them depending on the given situation.

At any moment the traverse of a walk is either known given N or can be estimated together with $\hat{R}_3$. A prediction form of (13) would be helpful, which takes into account the fact that uncertainty only arises due to the part of graph that is yet unobserved.

Acknowledgement I am grateful to two reviewers for helpful comments and suggestions.

References

[1] Alon, N., Benjamini, I., Lubetzky, E., and Sodin, S. (2007). Non-backtracking random walks mix faster. *Communications in Contemporary Mathematics*, 9:585-603.

[2] Avrachenkov, K., Ribeiro, B., and Sreedharan, J. K. (2016). Inference in OSNs via lightweight partial crawls. *ACM SIGMETRICS Performance Evaluation Review*, 44:165-177.

[3] Avrachenkov, K., Ribeiro, B. and Towsley, D. (2010). Improving Random Walk Estimation Accuracy with Uniform Restarts. *Research report*, RR-7394, INRIA. inria-00520350

[4] Ben-Hamou, A., Oliveira, R., and Peres, Y. (2019a). Estimating graph parameters with random walks. *Mathematical Statistics and Learning*, 1:375-399.

[5] Ben-Hamou, A., Lubetzky, E, and Peres, Y. (2019b) Comparing mixing times on sparse random graphs. *Annales de l'Institut Henri Poincaré Probability and Statistics*, 55:1116-1130.

[6] Birnbaum, Z.W. and Sirken, M.G. (1965). *Design of Sample Surveys to Estimate the Prevalence of IRareDiseases: Three Unbiased Estimates*. Vital and Health Statistics, Ser. 2, No.11. Washington:Government Printing Office.

[7] Boyd, S., Diaconis, P. and Xiao, L. (2004). Fastest mixing Markov chain on a Graph. *SIAM Review*, 46:667-689.

[8] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107-117.

[9] Cooper, C., Radzik, T., and Siantos, Y. (2016). Fast low-cost estimation of network properties using random walks. *Internet Mathematics*, 12:221-238.

[10] Masuda, N., Porter, M.A. and Lambiotte, R. (2017) Random walks and diffusion on networks. *Physics Reports*, 716-717: 1-58. `http://dx.doi.org/10.1016/j.physrep.2017.07.007`

[11] Newman, M.E.J. (2010). *Networks: An Introduction*. Oxford University Press.

[12] Palahan, S. (2015). Estimating degree distributions of large networks using non-backtracking random walk with non-uniform jump. *IEEE International Conference on Electro/Information Technology (EIT), 2015*, pp. 667-671.

[13] Thompson, S.K. (2006). Targeted random walk designs. *Survey Methodology*, 32, 11–24.

[14] Zhang, L.-C. (2021b). *Graph Sampling.* Chapman & Hall/CRC, in press.

[15] Zhang, L.-C. and Patone, M. (2017). Graph sampling. *Metron*, 75:277-299.