Estimating Network Parameters using Random Walk Techniques

Aditya Maitra supervised by Peter Taylor

2023/24 Mathematics and Statistics Vacation Scholarship Program, The University of Melbourne

Introduction

How can we estimate the total size of a network when it is impossible to see its entire structure?

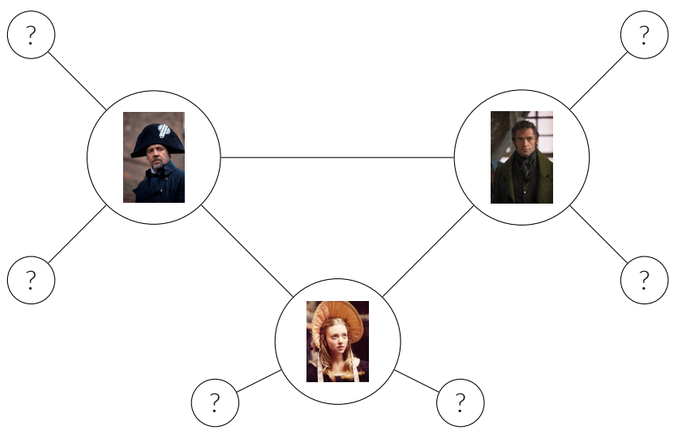
Consider a network consisting of objects and links between them. The network can be represented as a graph $G = (V, E)$.

This poster explores how random walk techniques can be used to estimate global properties of the graph whilst only having access to information about the node we are currently at.

The Sample Dataset

Numerical results in this poster are based on the sample Les Miserables Network. In the Les Mis network, nodes are characters of the novel, and edges are formed if two characters appear in the same chapter of the novel.

The Les Mis Network has $|V| = 77$ and $|E| = 254$.



Random Walks on a Graph

Consider the multi-graph $G = (V, E)$. Let

- $d(i) =$ the degree of node i
- $f(i, j) =$ the number of edges between i and j
- $n = |V| =$ the number of nodes in G
- $m = |E| =$ the number of edges in G

We consider two types of random walks on G.

The Discrete Time Random Walk (DTRW)

Suppose the DTRW is at node i after n steps. At step $n+1$, the walk jumps to a neighbour of i uniformly at random.

The position X_n of the SRW after n steps is described by a Discrete Time Markov Chain with state space V and transition matrix

$$P_{ij} = \begin{cases} f(i,j)/d(i), & \text{if } ij \in E, \\ 0, & \text{if } ij \notin E \end{cases} \tag{1}$$

Properties of the DTRW

The stationary distribution of the DTRW is

$$\pi(i) = \frac{d(i)}{2|E|} = \frac{d(i)}{2m}$$

The expected return time to a node i is

$$E(T_i) = \frac{1}{\pi(i)} = \frac{2m}{d(i)} \tag{2}$$

The Continuous Time Random Walk (CTRW)

Suppose the CTRW arrives in node i at time t. The walk stays at i for an exponentially distributed amount of time with parameter $d(i)$, before jumping to a neighbour of i uniformly at random.

The position X_t of the CTRW at time t is described by a Continuous Time Markov Chain with state space V and generator

$$Q_{ij} = \begin{cases} -d(i), & \text{if } i = j, \\ f(i,j), & \text{if } i \neq j \text{ and } ij \in E, \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Properties of the CTRW

The stationary distribution of the CTRW is

$$\pi(i) = \frac{1}{|V|} = \frac{1}{n}$$

The expected return time to a node i is

$$E(T_i) = \frac{|V|}{d(i)} = \frac{n}{d(i)} \tag{4}$$

References

- Avrachenkov, K., Borkar, V.S., Kadavankandy, A. et al. (2018) Revisiting random walk based sampling in networks: evasion of burn-in period and frequent regenerations. Computational Social Networks, 5(4).
- Hunter, J. J. (1969). On the Moments of Markov Renewal Processes. Advances in Applied Probability, 1(2), 188–210.

Estimating the Number of Edges $|E|$

The form of (2) suggests that the number of edges in G can be estimated using a DTRW as

$$\hat{m}(i) = \frac{\hat{E}(T_i)d(i)}{2} \tag{5}$$

To estimate m in practice we use the following algorithm

1. Start at node i and simulate DTRW until returning to i
2. Record first time of return T_i^1
3. Repeat k times to get k return observations $[T_i^1, ..., T_i^k]$
4. Terminate simulation once the estimate has 95% confidence of being within $\pm\epsilon$ of the true value.

Theoretical Performance of the Estimator

The Markov property implies that each T_i^j for j∈{1,..,k} is independently and identically distributed. Using the Central Limit Theorem with (3) and assuming a large enough k, the distribution of $\hat{m}(i)$ can be approximated by the below

$$\hat{m}(i) \sim N\left(m, \frac{d(i)^2\sigma_D(i)^2}{4k}\right) \sim N\left(m, \frac{m \times d(i)\sigma_D(i)^2}{2s(i,k)}\right) \tag{6}$$

- $\sigma_D(i)^2$ gives variance of first return time to i under the DTRW.
- $s(i,k) = \frac{d(i)\times k}{2m}$ gives the expected number of steps in k returns to node i.

Then if we simulate returns from node i it takes roughly

$$s_{95}(i,k) = \frac{1.96^2 \times md(i)\sigma_D(i)^2}{2\epsilon^2} \tag{7}$$

steps to have 95% confidence that $|\hat{m} - m| \leq \epsilon$.

Estimating the Number of Nodes $|V|$

Likewise, the form of (4) suggests that the number of nodes in G can be estimated using a CTRW as

$$\hat{n}(i) = \hat{E}(T_i)d(i) \tag{8}$$

To estimate n in practice we use the same algorithm as above but using a CTRW instead of the DTRW.

Theoretical Performance of the Estimator

Once again, the Markov property implies that each T_i^j for j∈{1,..,k} is independently and identically distributed. Using the Central Limit Theorem with (8) and assuming a large enough k, the distribution of $\hat{n}(i)$ can be approximated by the below

$$\hat{n}(i) \sim N\left(n, \frac{d(i)^2\sigma_C(i)^2}{K}\right) \sim N\left(n, \frac{2 \times m \times d(i)\sigma_C(i)^2}{s(i,k)}\right) \tag{9}$$

- $\sigma_C(i)^2$ gives variance of first return time to i under the CTRW
- $s(i,k) = \frac{d(i)\times k}{2m}$ gives the expected number of jumps in k returns to node i.

Then, if we simulate returns from node i it takes roughly

$$s_{95}(i,k) = \frac{1.96^2 \times 2md(i)\sigma_C(i)^2}{\epsilon^2} \tag{10}$$

steps to have 95% confidence that $|\hat{n} - n| \leq \epsilon$

Comparing Estimation Speed

In both techniques, the properties of the estimator depend on the node we begin the random walk from.

(6) and (9) demonstrate that both estimators are unbiased. However their speeds of convergence are both given by the term

$$\alpha(i) = d(i)\sigma(i)^2 \tag{11}$$

Nodes with smaller values of $\alpha(i)$ converge faster.

Variance of first return times

The term $\alpha(i)$ depends on the variance of the first return time to i under the given random walk. In both the discrete and continuous case, this variance can be computed, but lacks a simple expression.

Importantly, $\sigma(i)^2$ *mostly* decreases with $d(i)$ faster than linearly. In turn $\alpha(i)$ decreases with $d(i)$, and so in both methods, starting at nodes with higher degrees gives faster convergence!

Numerical Results: Required Steps

We run the estimators above on the Les Mis Network, starting from three different nodes. Tables show required steps for each node with $\epsilon = 5$.

Node	Degree	$\sigma(i)^2$	Steps
A	36	370	259,630
B	10	23,241	4,535,557
C	1	514,248	10,035,726

Table 1. Steps to Estimate m with $\epsilon = 5$

Node	Degree	$\sigma(i)^2$	Steps
A	36	8.4	23,654
B	10	442.9	345,728
C	1	10,781.4	841,613

Table 2. Steps to Estimate n with $\epsilon = 5$

Note, even best case scenario requires large number of steps.

Numerical Results: Levels of Accuracy

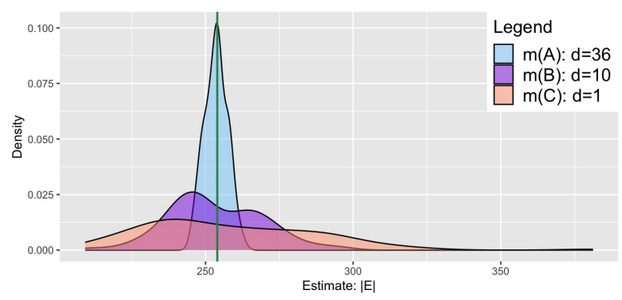


Figure 1. Distribution of Estimators after $100,000$ steps of DTRW. Green bar shows true value $m = 254$.

The SuperNode Technique

Starting from a single node yields slow convergence of estimators. Avrachenkov, et all (2018) propose the SuperNode method for faster convergence.

In the SuperNode method we contract the graph as follows:

- Choose a set of nodes from G: $S_n = \{i, j, k\} \subset V$.
- Contract G so that these nodes are combined into one node S
- Edges out of S are given by the set of edges from nodes in S_n to nodes out of S_n. Any edges between nodes in S_n are removed.
- Call the contracted graph $G' = (V', E')$
- Let $e' = |E| - |E'|$ and $v' = |V| - |V'| = n(S_n) - 1$

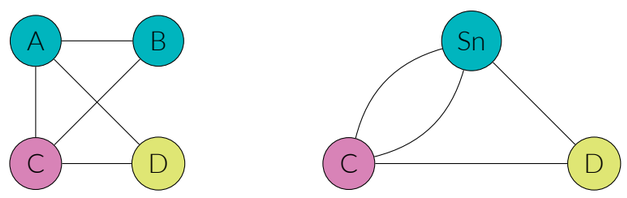
We can estimate the parameters of the contracted graph G' by starting random walks at the SuperNode S. The higher degree of the SuperNode means these estimates will converge faster than before.

Then we can map the parameters of G' to the parameters of G

$$|\hat{E}| = |\hat{E'}| + e' \text{ and } |\hat{V}| = |\hat{V'}| + v'$$

A Toy Example

Consider $G = (V, E)$ with $|V| = 4$ and $|E| = 5$. Contract graph so that $S_n = \{A, B\}$. The contracted graph $G' = (V', E')$ has $|V'| = 3$ and $|E'| = 4$.



Numerical Results with SuperNode

There are a variety of ways we can construct the SuperNode. Here we take S_n to be the three nodes of G with the highest degree.

The tables show the required steps for each node with $\epsilon = 5$, and the figure shows the distribution of the estimators after $50,000$ steps of the walk, in the discrete time case.

Node	Degree	$\sigma(i)^2$	Steps
S	71	55	75,269
A'	36	370	259,630
B'	22	936	402,005
C'	19	1,006	372,932

Table 3. Steps to Estimate m with $\epsilon = 5$

Node	Degree	$\sigma(i)^2$	Steps
S	71	2.6	14,108
A'	36	8.4	23,654
B'	22	30.7	52,794
C'	19	33.35	49,470

Table 4. Steps to Estimate n with $\epsilon = 5$

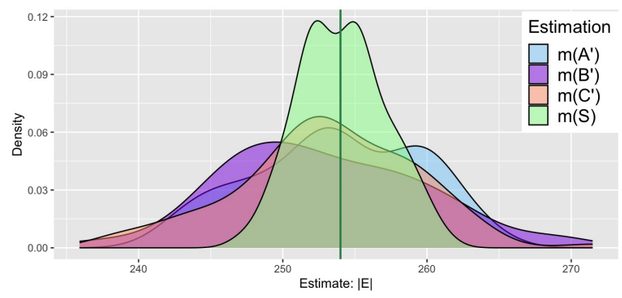


Figure 2. Distribution of Estimators after $50,000$ steps of DTRW. Green bar shows true value $m = 254$.

Further Extensions

- Explore how the convergence speed of estimators scales with graph size.
- Explore optimal procedures for construction of a SuperNode.
- Understand properties of estimators for other graph parameters.
- Investigate how random walk theory can be used to estimate parameters for directed graphs.

amaitra@student.unimelb.edu.au