Estimate Graph Property considering Private Node by Random Walk

2021

Table of Contents

Random Walk on Graph

- Start from any vertex v_{x_1} in graph G.
- Move to next node in neighborhood of current node *uniformly and randomly*.
- Repeat $r-1$ times to get a list of nodes $\{v_{x_1}, \ldots, v_{x_r}\}$.
- By Markov Chain, we can prove that the stationary distribution is $\Pr[x_i = k] = \frac{d_k}{\sum_{i=1}^{n} d_i}$. And the random walk will converge to this distribution.

Random Walk on Graph

Why random walk?

- Control how many nodes we sampled and reduce running time and memory.
- In real world, data of the *whole graph* is usually hard to acquire. Sometimes the reason is that graph is too large, but more commonly, whole data is not provided due to commercial interest and privacy.
- As an alternative, platforms and websites usually provide an API to request information for one node at a time, which is just suitable for random walk algorithm.
- Using random walk, we can give good estimation on many graph properties with small proportion of samples.

Estimate Average Degree

Definition

d_i is degree of vertex i, D is total number of degree of all vertices.

$$d_{avg} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{D}{n}$$

The expectation of the sum of $1/d_{x_i}$ in our sampled $\{x_1, \ldots, x_r\}$ is

$$E[\sum_{i=1}^{r} \frac{1}{d_{x_i}}] = \sum_{i=1}^{r} E[\frac{1}{d_{x_i}}] = \sum_{i=1}^{r} \sum_{j=1}^{n} \Pr[x_i = j] \frac{1}{d_j} = \sum_{i=1}^{r} \sum_{j=1}^{n} \frac{1}{D} = \frac{rn}{D} = \frac{r}{d_{avg}}$$

So $\frac{r}{\sum_{i=1}^{r} \frac{1}{d_{x_i}}}$ can be a good estimation of d_{avg}.

Estimate Clustering Coefficient

Changing values calculated from samples, we can estimate variouss properties. Clustering coefficient is a common and important property of social graph.

Definition related to clustering coefficient

- l_i: pairs (number of edges) between neighbors of v_i.
- Local clustering coefficient: $c_i = \frac{2l_i}{d_i(d_i-1)}$.
- Network average: $\bar{C} = \frac{1}{n}\sum_{i=1}^{n} c_i$.
- Global: $C = \frac{2\sum_{i=1}^{n} l_i}{\sum_{i=1}^{n} d_i(d_i-1)}$.

Estimate Clustering Coefficient

Network Average Estimator

Let $\phi_k = 1$ if $x_{r_{k-1}}$ and $x_{r_{k+1}}$ are connected. 0 o.w.

$$E[\phi_k \frac{1}{d_{x_k} - 1}] = \sum_{i=1}^{n} \frac{d_i}{D} \frac{2l_i}{{d_i}^2} \frac{1}{d_i - 1} = \frac{1}{D} \sum_{i=1}^{n} \frac{2l_i}{d_i(d_i - 1)} = \frac{1}{D} \sum_{i=1}^{n} c_i$$

and we know $E[\frac{1}{d_{x_k}}] = \frac{n}{D}$, so

$$E[\frac{\phi_k \frac{1}{d_{x_k}-1}}{\frac{1}{d_{x_k}}}] \approx \frac{E[\phi_k \frac{1}{d_{x_k}-1}]}{E[\frac{1}{d_{x_k}}]} = \frac{1}{n} \sum_{i=1}^{n} c_i = \bar{C}$$

Estimate Clustering Coefficient

Global Estimator

$$E[\phi_k d_{x_k}] = \sum_{i=1}^{n} \frac{d_i}{D} \frac{2l_i}{{d_i}^2} d_i = \frac{1}{D} \sum_{i=1}^{n} 2l_i$$

$$E[d_{x_k} - 1] = \sum_{i=1}^{n} \frac{d_i}{D}(d_i - 1) = \frac{1}{D} \sum_{i=1}^{n} d_i(d_i - 1)$$

Similar,

$$E[\frac{\phi_k d_{x_k}}{d_{x_k} - 1}] \approx \frac{2\sum_{i=1}^{n} l_i}{\sum_{i=1}^{n} d_i(d_i - 1)} = C$$

Private Node Problem

Private node is common in real world, especially in circumstance of social network. For our setting, the main problem is that private node hides information of its neighborhood:

- If we (unfortunately) sample a private node in random walk, as its neighbors are hidden, how we sample (move to) the next node?
- The neighborhood of one node now includes both public and private nodes, and we don't know which of them are public.
- The degree of private node is unknown.

Random Walk with Private Node

- Simple idea: Ignore and avoid private node, only consider public part of graph.
- In practice, as we don't know which nodes in neighbors are public, we first sample randomly in neighborhood, and repeat until a public node is sampled.
- The sampling trying process can also be used to estimate public degree, by how much ratio we succeed for this node.

Definition

- G^*: largest public nodes component of G.
- d_i^*: public degree of node v_i.
- D^*: sum of public degrees, $D^* = \sum_{v_i \in G^*} d_i^*$

Random Walk with Private Node

Intuitively, if we run same random walk algorithms on G^*, we will get properties of G^*. What's the relationship of proprieties of G^* and G?

Random Walk with Private Node

Assume each node has p possibility becoming private independently, we have

Property relationship of G and G^*

- $d^*_{avg} = \frac{D^*}{n^*}$, $E_{pri}[d^*_{avg}] \approx \frac{E_{pri}[D^*]}{E_{pri}[n^*]} = \frac{(1-p)^2 D}{(1-p)n} = (1-p)d_{avg}$.
- $E_{pri}[|V^*|] = (1-p)|V|$
- $E_{pri}[l^*_i] = (1-p)^2 l_i$
- $E_{pri}[C^*] \approx \frac{E_{pri}[\sum_{i=1}^{n} l^*_i]}{E_{pri}[\sum_{v_i} d^*_i(d^*_i-1)]} = \frac{(1-p)^2\sum_{i=1}^{n} l_i}{(1-p)^2\sum_{i=1}^{n} d_i(d_i-1)} = C$
- $E_{pri}[\bar{C}^*] \approx \frac{\sum_{v_i\in G^*}[1-p^{d_i}-d_i(1-p)^{d_i-1}]c_i}{\sum_{v_i\in G^*} c_i}\bar{C}$ (Nakajima and Shudo, 2021)

Random Walk with Private Node

- If we know the exact value of private possibility p, it's easy getting estimation of G from results of G^* (by relationship in last slide).
- However, we don't know p value in real world!
- Compared to missing node problem, we can still estimate some information of private node from its public neighbors.
- The basic idea of reducing affects of private node is offsetting p in estimator's dividing.

Proposed Estimator of Average Degree

Theorem

In random walk on G^*, $\Pr[x_i = k] = \frac{d_k^*}{D^*}, x_k \in V^*$

Smooth estimator for d_{avg} (Dasgupta, Kumar and Sarlos, 2014)

$$d_{avg}^{smooth} = \frac{r}{\sum_{i=1}^{r} \frac{1}{d_{x_i^*}}}$$

$$E[d_{avg}^{smooth}] \approx d_{avg}^*, E_{pri}[d_{avg}^*] \approx (1-p) d_{avg}$$

Modified smooth estimator for d_{avg} (Nakajima and Shudo, 2020)

$$d_{avg}^{new} = \frac{r}{\sum_{i=1}^{r} \frac{1}{d_{x_i}}}, E_{pri}[E[d_{avg}^{new}]] \approx d_{avg}$$

Proposed Estimator of Average Degree

Modified smooth estimator for d_{avg} (Nakajima and Shudo, 2020)

$$d_{avg}^{new} = \frac{r}{\sum_{k=1}^{r} \frac{1}{d_{x_k}}}, E_{pri}[E[d_{avg}^{new}]] \approx d_{avg}$$

While it seems natural, the proof is not so direct:

Proof.

$$E[\frac{1}{d_{x_k}}] = \sum_{v_i \in V^*} \frac{d_i^*}{D^*} \frac{1}{d_i} = \frac{1}{D^*} \sum_{v_i \in V^*} \frac{d_i^*}{d_i}$$

$$E[d_{avg}^{new}] \approx \frac{r}{\frac{r}{D^*} \sum_{v_i \in V^*} \frac{d_i^*}{d_i}} = \frac{D^*}{\sum_{v_i \in V^*} \frac{d_i^*}{d_i}}$$

$$E_{pri}[\frac{D^*}{\sum_{v_i \in V^*} \frac{d_i^*}{d_i}}] \approx \frac{(1-p)^2 D}{(1-p)n[(1-p) \cdot 1]} = \frac{D}{n} = d_{avg}$$

Proposed Estimator of Clustering Coefficient

Because $E_{pri}[C^*] \approx C$, we mainly focus on $\bar{C}$ here.

Original Estimator for $\bar{C}$ (Hardiman and Katzir, 2013)

$$\bar{C}^{ori} = \frac{\frac{1}{r-2}\sum_{k=2}^{r-1} \phi_k \frac{1}{d_{x_k^*}-1}}{\frac{1}{r}\sum_{k=1}^{r} \frac{1}{d_{x_k^*}}}$$

$$E[\bar{C}^{ori}] \approx \bar{C}^*,\ E_{pri}[\bar{C}^*] \approx \frac{\sum_{v_i \in V^*}[1 - p^{d_i} - d_i(1-p)^{d_i - 1}]c_i}{\sum_{v_i \in V^*} c_i} \bar{C}$$

Proposed Estimator of Clustering Coefficient

Original Estimator for $\bar{C}$ (Hardiman and Katzir, 2013)

$$\bar{C}^{ori} = \frac{\frac{1}{r-2}\sum_{k=2}^{r-1} \phi_k \frac{1}{d_{x_k^*}-1}}{\frac{1}{r}\sum_{k=1}^{r} \frac{1}{d_{x_k^*}}}$$

Proposed Estimator for $\bar{C}$

$$\bar{C}^{new} = \frac{\frac{1}{r-2}\sum_{k=2}^{r-1} \phi_k^{new} \frac{1}{d_{x_k}-1}}{\frac{1}{r}\sum_{k=1}^{r} \frac{1}{d_{x_k}}}$$

Proposed Estimator of Clustering Coefficient

Proposed Estimator for $\bar{C}$

$$\bar{C}^{new} = \frac{\frac{1}{r-2}\sum_{k=2}^{r-1} \phi_k^{new} \frac{1}{d_{x_k}-1}}{\frac{1}{r}\sum_{k=1}^{r} \frac{1}{d_{x_k}}}$$

- In the sample step k, if we sample a private node v', we actually know if $v_{x_{k-1}}$ and v' are connected by seeing neighbors of $v_{x_{k-1}}$, just like we test $v_{x_{k-1}}$ and $v_{x_{k+1}}$. (Assume it's a undirected graph)
- When p is large, there's a high percentage of sampling are private nodes. Taking advantage of this information can greatly improve our estimation.

Definition

ϕ_k^{new} is that, in all sampling tries in neighborhood of $v_{x_{r_k}}$, the ratio that sampled v' is connected to $x_{r_{k-1}}$.

Proposed Estimator of Clustering Coefficient

Let w_i^* be the number of pairs in the neighborhood of x_i that has at least one node is public.

Theorem

$$E_{pri}[E[\bar{C}^{new}]] = \bar{C}$$

Proof.

$$E[\phi_k^{new} \frac{1}{d_{r_k} - 1}] = \sum_{v_i \in V^*} \frac{d_i^*}{D^*} \frac{w_i^*}{d_i d_i^*} \frac{1}{d_i - 1} = \frac{1}{D^*} \sum_{v_i \in V^*} \frac{w_i}{d_i(d_i - 1)}$$

$$E[\bar{C}^{new}] \approx \frac{E[\frac{1}{r-2} \sum_{k=2}^{r-1} \phi_k^{new} \frac{1}{d_{r_k}-1}]}{E[\frac{1}{r} \sum_{k=1}^{r} \frac{1}{d_{x_k}}]} = \frac{\sum_{v_i \in V^*} \frac{w_i^*}{d_i(d_i-1)}}{\sum_{v_i \in V^*} \frac{d_i^*}{d_i}}$$

Proposed Estimator of Clustering Coefficient

Proof cont.

$$E_{pri}[\frac{\sum_{v_i \in V^*} \frac{w_i^*}{d_i}}{\sum_{v_i \in V^*} \frac{d_i^*}{d_i}}] \approx \frac{\sum_{v_i \in V^*} \frac{1}{d_i(d_i-1)} E_{pri}[w_i]}{(1-p)^2 n}$$

$$E_{pri}[w_i^*] = (1-p)^2(2l_i) + 2p(1-p)l_i + p^2 \cdot 0 = 2(1-p)l_i$$

$$E_{pri}[\frac{\sum_{v_i \in V^*} \frac{w_i^*}{d_i(d_i-1)}}{n^*}] \approx \frac{(1-p)\sum_{v_i \in V^*} \frac{2l_i}{d_i(d_i-1)}}{(1-p)^2 n} = \bar{C}$$

Compared to original estimator of $\bar{C}$, our proposed estimator has no shift in theory and works better in reality, especially when p is large.

Experiments

- Estimator: Average degree, average and global clustering coefficient, and size (which is not mentioned in slides).
- Dataset: Youtube, CAIDA and GitHub data from SNAP.
- Setting:
  - Sample size $r = 1\%|V|$.
  - Repeat random walk 100 times independently and calculate mean of results.
  - Ground truth value is computed by exact algorithms.
  - Error is defined as $(\frac{x}{x_{gt}} - 1)^2$.

Estimator of average degree:

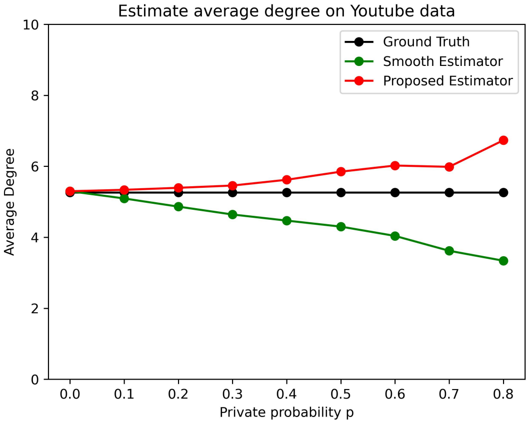

Figure: Average degree estimation on Youtube data

Experiments on Youtube Data

Estimator of average degree:

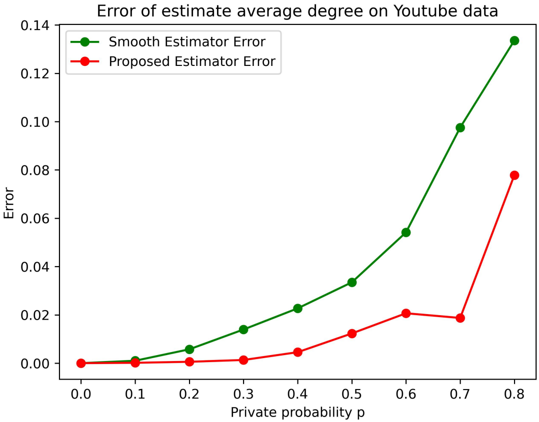

Figure: Error of average degree estimation on Youtube data

Experiments on Youtube Data

Estimator of average cluster coefficient:

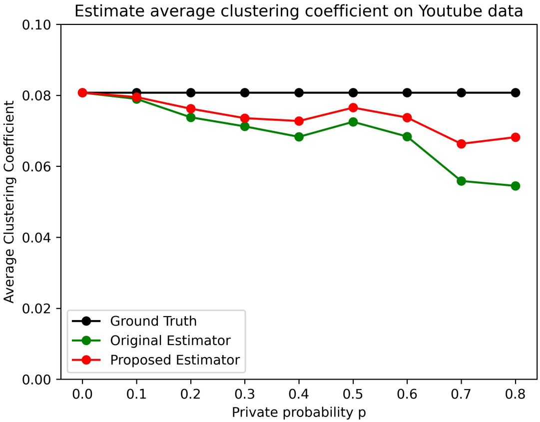

Figure: Average cluster coefficient estimation on Youtube data

Estimator of average cluster coefficient:

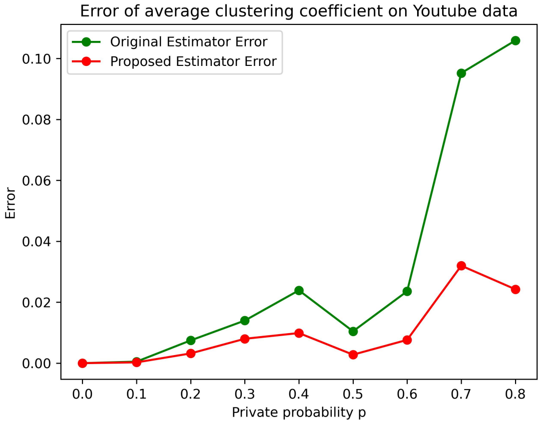

Figure: Error of average cluster coefficient estimation on Youtube data

Experiments on GitHub Data

Estimator of average degree:

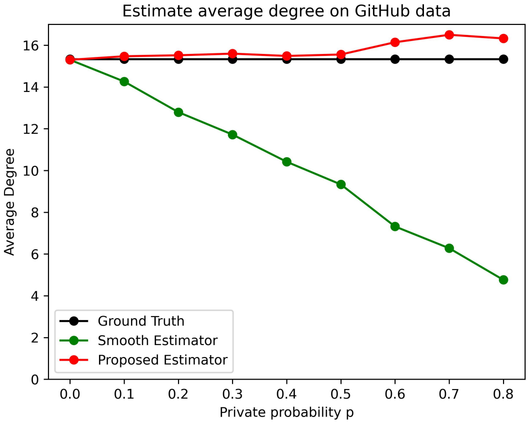

Figure: Average degree estimation on GitHub data

Experiments on GitHub Data

Estimator of average degree:

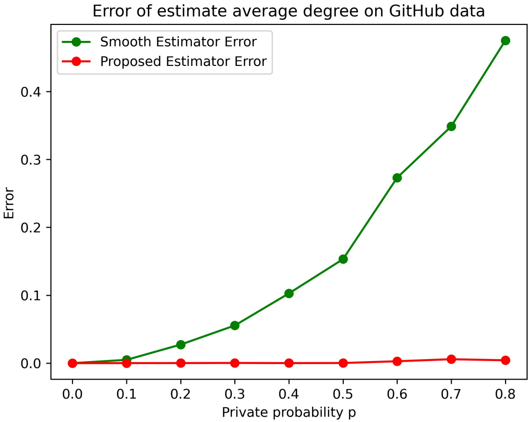

Figure: Error of average degree estimation on GitHub data

Experiments on GitHub Data

Estimator of average cluster coefficient:

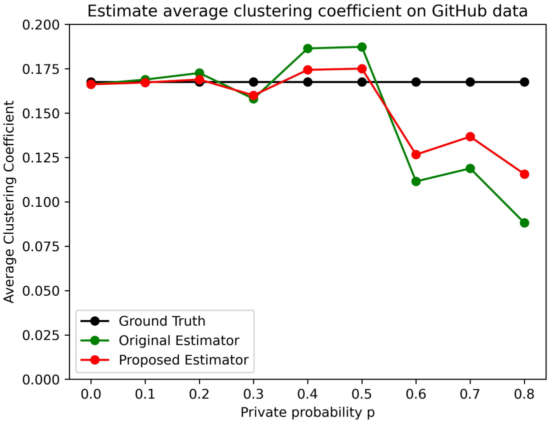

Figure: Average cluster coefficient estimation on Youtube data

Experiments on GitHub Data

Estimator of average cluster coefficient:

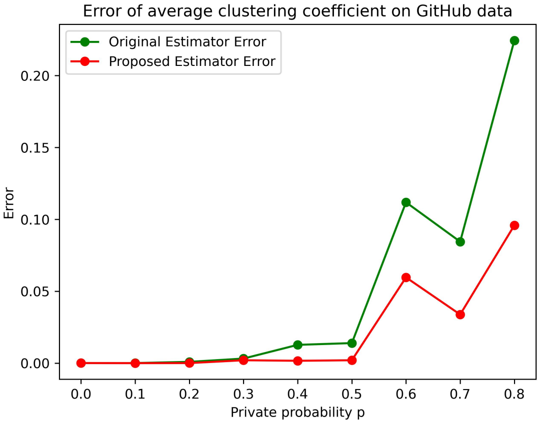

Figure: Error of average cluster coefficient estimation on GitHub data

Reference

- Nakajima, K., Shudo, K. (2020, August). Estimating properties of social networks via random walk considering private nodes. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (pp. 720-730).
- Nakajima, K., Shudo, K. (2021). Measurement error of network clustering coefficients under randomly missing nodes. Scientific Reports, 11(1), 1-14.
- Hardiman, S. J., Katzir, L. (2013, May). Estimating clustering coefficients and size of social networks via random walk. In Proceedings of the 22nd international conference on World Wide Web (pp. 539-550).
- Dasgupta, A., Kumar, R., Sarlos, T. (2014, April). On estimating the average degree. In Proceedings of the 23rd international conference on World wide web (pp. 795-806).