Estimating the Clustering Coefficient of a Social Network by a Non-Backtracking Random Walk

Kenta Iwasaki, and Kazuyuki Shudo
Tokyo Institute of Technology

***Abstract*—The clustering coefficient of social networks can be estimated via an unbiased sampling technique such as random walk sampling. To this end, we propose an algorithm that assumes no prior knowledge of the network and that accesses the network only through a publicly available interface. The clustering coefficient of a network is estimated by counting triangles in a non-backtracking random walk (NBRW). A social network is viewed as an undirected graph and the NBRW is retrieved through a public interface, which theoretically guarantees its retrieval. In a simulation study on real social networks, the proposed method achieved higher efficiency and accuracy than the prior state-of-the-art method based on a simple random walk.**

***Index Terms*—social network, clustering coefficient, random walk**

I. Introduction

Online social networks (OSNs) have become increasingly popular in recent years. The most popular OSN, Facebook, counted more than 1.8 billion members in January 2017 [1]. This popularity has sparked growing interest in the properties of OSNs. There are prior work [6], [10] that determines structural measures of OSNs, including the clustering coefficient. However, estimating network characteristics of OSNs is a difficult task, because such networks are typically too large to measure and their complete datasets are typically unavailable, primarily because of privacy concerns [2]. For example, independent sampling which directly obtains uniform, independent node samples for unbiased estimation, is infeasible from such a large unknown network. In practical scenarios, the social network may be available only through a public interface such as an application programming interface (API), which queries and obtains information about a target node. In most social networks, the public interface provides a list of a user's neighboring nodes. By applying this function iteratively to a random member of the neighbor list, one can effectively perform a random walk on the network. Considering the delay and query limit of APIs, the sample should be small but representative. In the example of Fig. 1, the researchers obtain a neighbor list by repeatedly querying the API.

However, to calculate the clustering coefficient of a node, the researcher must query nodes other than the sampled nodes. Therefore, the clustering coefficient of the network is difficult to obtain. Hardiman et al. [8] estimated the clustering coefficient of social networks from the sampling queries alone. They achieved efficient estimation by counting triangles during random walk sampling.

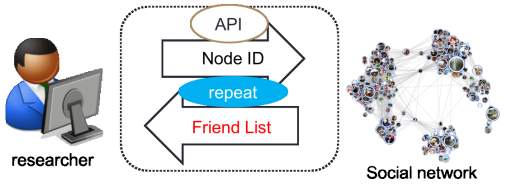

Fig. 1. Sampling through the API of a social network.

Conversely, a simple random walk usually diffuses slowly across the network, leading to poor estimation accuracy. In particular, because the next node is selected in purely random fashion, the walk often returns to the previous node that was just visited. This backtracking produces many duplicate samples over a short to moderate time span, reducing the estimation accuracy. Obviously, such backtracking transitions should be avoided whenever possible, and the walk should be steered toward unvisited nodes without biasing the estimation. Lee et al. [9] proposed a non-backtracking random walk (NBRW) with re-weighting that guaranteed not only unbiased graph sampling but also smaller asymptotic variance of the estimators than a simple random walk.

In this paper, we propose an efficient and accurate method for estimating the clustering coefficients of networks. While sampling, our method counts triangles via a NBRW. Combining this way of counting triangles with NBRW is not trivial. We combined these two ideas and adjusted to be able to estimate unbiased. We theoretically proved that the proposed method can perform the unbiased estimation and showed theoretical and experimental results that it is more efficient and better than the existing method.

This paper is organized as follows. Section II shows related work. Section III provides notations and preliminaries. Section IV describes our method. Section V compares the method with the existing method. Section VI summarizes our contributions.

II. Related Work

This section describes related work, in which the average clustering coefficient is estimated by the crawling-based graph

This work was supported by JSPS KAKENHI Grant Numbers 2570008 and 16K12406. This work was supported by New Energy and Industrial Technology Development Organization (NEDO).

[1] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Proc. 2018 IEEE Int'l Conf. on Big Data and Smart Computing
(IEEE BigComp 2018), pp.114-118, January 2018

sampling method. Ribeiro et al. [10] estimated the network average clustering coefficient using a simple random walk (SRW). Their computation requires augmenting the set of sampled nodes with further exploration of the ego network. To calculate the local clustering coefficient at a special node v, a range of ego networks of v must be known. Gjaka et al. [6] explored an OSN graph using a Metropolis-Hastings random walk that generates uniform samples from the node set. This method also requires further exploration of the ego network. Such methods are inefficient because exploring the ego network requires additional API queries.

Hardiman et al. [8] estimated the network average clustering coefficient by a random walk that did not explore the ego network. Their method outperformed competing approaches [6], [10] on all social networks in their study [8]. However, they used an SRW, which (as mentioned above) diffuses slowly over the space and compromises the estimation accuracy. During a transition, the SRW frequently returns to the just-visited node.

In this paper, we perform a NBRW that avoids the previous node whenever possible. The proposed method is conceptualized in Section 3 and developed in Section 4. In Section 5, we show that the proposed algorithm outperforms Hardiman et al.'s approach [8] on various social networks. The paper briefly concludes with Section 6.

III. Notations and Preliminaries

This section describes the notations, definitions, and the basic idea the of the proposed method (counting triangles during NBRW).

A. Notations

The social graph of an OSN can be modeled as a connected, undirected graph $G = (V, E)$ with a set of nodes $V = \{v_1, v_2, \ldots, v_n\}$ and a set of edges E. We assume that $3 \leq |V| = n < \infty$. We also assume that graph G has no self-loops and no multi-edges. Let $N(v) \stackrel{\text{def}}{=} \{w \in V : (v, w) \in E\}$ be the neighbor set of node $v \in V$, $d_i = k_{v_i} = |N(v_i)|$ be the degree of node v.

A triplet of nodes (v_j, v_i, v_k) is called connected if there is an edge between v_j and v_i, and an edge between v_i and v_k, and $j < k$.

A triangle is a connected triplet (v_j, v_i, v_k) in which v_j and v_k are connected by an edge. Let $\triangle_i$ be the number of edges between neighbors of v_i.

The local clustering coefficient c_i [5] of node v_i defines the ratio of the number of (v_j, v_i, v_k) triangles to the number of (v_j, v_i, v_k) connected triplets. Formally,

$$c_i = \begin{cases} 0 & d_i = 0 \text{ or } d_i = 1 \\ \dfrac{2\triangle_i}{d_i(d_i-1)} & \text{otherwise} \end{cases} \qquad c_i \in [0,1] \qquad (1)$$

The network average clustering coefficient C [5] is defined as

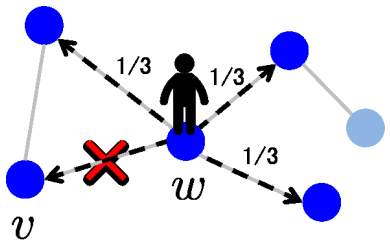

Fig. 2. Transitions of an NBRW over the nodes of G. The walker is currently located at node w (with $k_w = 4$) and has just visited node v. From w, it will move to any neighbor except node v with equal probability.

$$C = \frac{1}{n}\sum_{i=1}^{n} c_i \qquad (2)$$

Our proposed method estimates the network average clustering coefficient.

The first step of the estimation algorithm generates a random walk. A random walk with r steps on G, denoted by $R = (x_1, x_2, \ldots, x_r)$, starts from an arbitrary node v_{x_1} and then moves to a chosen probabilistically selected neighboring node. This process repeats $r - 1$ times.

Herein, we apply two random walks: the SRW and the non-backtracking random walk (NBRW). The SRW on G moves from a node to a neighboring node, which is randomly and uniformly selected with with probability $\frac{1}{d_i}$. The NBRW avoids backtracking to the previous node whenever possible.

Let $\Pr[A]$ denote the occurrence probability of event A. The distribution induced by random walk, is defined as

$$\pi = (\mathbf{P}r[x_r = 1], \mathbf{P}r[x_r = 2], \ldots, \mathbf{P}r[x_r = n]). \qquad (3)$$

After many random walks, the probability $\mathbf{P}r[x_r = i]$ converges to a certain value $\pi(i)$. The vector $\pi = (\pi(1), \pi(2), \ldots, \pi(n))$ is called the stationary distribution of G. As is well known, the stationary distribution of the SRW is $d_i/2|E|$ [3].

B. Non-Backtracking Random Walk

This subsection overviews the graph sampling method with the NBRW proposed in [9] and presents its algorithm. Our proposed algorithm is based on the NBRW, which minimizes backtracking to the previous node while preserving the same stationary distribution as the SRW.

In an NBRW with more than two neighbors ($k_w \geq 2$), the walker at current node w randomly selects the next node to visit among the neighbors of node w, except for the previous node v. If the current node w has only one neighbor ($k_w = 1$), the walk always returns to the previous node v. The non-backtracking nature of the NBRW as it traverses the nodes of G is depicted in Fig. 2. The NBRW moves from its initial node, which can be arbitrarily chosen, to any of its neighboring nodes with equal probability (because no previous nodes have been visited).

C. Counting Triangles

Herein, we apply the "counting triangles" technique to the NBRW. Previously, Hardiman et al. [8] estimated the network average clustering coefficient by this technique in an SRW.

Let (x_{i-1}, x_i, x_{i+1}) be any set of consecutive triplets in a random walk $R = (x_1, x_2, \ldots, x_r)$. By the nature of random walks, $(v_{x_{i-1}}, v_{x_i}, v_{x_{i+1}})$ is a connected triplet. Moreover, if nodes $v_{x_{i-1}}$ and $v_{x_{i+1}}$ are connected by an edge, then $(v_{x_{i-1}}, v_{x_i}, v_{x_{i+1}})$ is a triangular triplet (x_1, x_2, x_3) is a triangular triplet because an edge connects node v_{x_1} to node v_{x_3}. Conversely, the consecutive triplet (x_2, x_3, x_4) is not a triangular triplet because no edge exists between nodes v_{x_2} and v_{x_4}.

In practical situations, when the neighbor list includes the two most recently sampled nodes during transition to the next node, then nodes v_{x_1} and v_{x_3} are connected. For example, consider the walker located at node $v_{x_3} = v_3$ in Fig. 3. The neighbor list of node v_{x_3} is (v_1, v_2, v_3, v_4), which includes the two most recently visited nodes $v_{x_1} = v_1$. Therefore, we can confirm that during the random walk sampling, an edge exists between nodes v_{x_3} and v_{x_1}.

It is possible to ascertain the existence of an edge by checking whether the last two sampled nodes have been included in the neighbor list when making transition to next node. For example, In Fig. 3, when a walker stays at a node $v_{x_3} = v_3$, a neighbor list of the node v_{x_3} is (v_1, v_2, v_3, v_4), and the last two sampled node $v_{x_1} = v_1$ is included in the neighbor list. Therefore we can confirm that an edge exists between the nodes v_{x_3} and the node v_{x_1} while sampling via random walk.

The counting triangles technique estimates the network average clustering coefficient because the local clustering coefficient defines the triangle-membership proportion of a node. However, this method is problematic when transiting to the previous node in an SRW. For example, in Fig. 4, the consecutive triplet (x_1, x_2, x_3) can never be a triangular triplet because the SRW backtracks to the previous node. Generally, backtracking to the previous node prevents a triangular path. Therefore, unless the edges are weighted, the proportion of triangles in the SRW is expected to be low.

In fact, by upward weighting of the counted triangles, the SRW-based approach proposed in [8] enables unbiased estimation of the network average clustering coefficient, lowering the estimation accuracy.

This problem is overcome by the NBRW, as shown the following section.

IV. Proposed Method

We now propose a method that estimates the network average clustering coefficient of social networks via an NBRW. Our idea stems from the counting triangles technique [8]. We prove that our estimate converges to the correct value, and demonstrate the superior effectiveness and accuracy our method over the prior method.

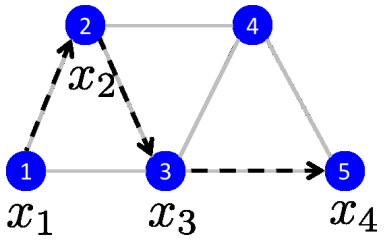

Fig. 3. Example of the counting triangles technique in random walk sampling.

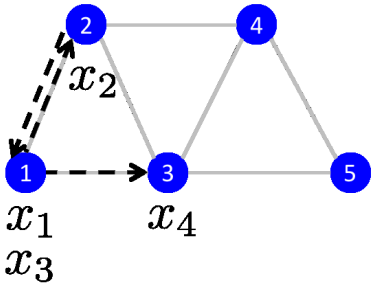

Fig. 4. Example of transition to the previous node in the SRW.

Given a NBRW $R' = (x_1, x_2, \ldots, x_r)$, we define a new variable ϕ'_k as follows: for $2 \leq \forall k \leq r-1$, if nodes $v_{x_{k-1}}$ and $v_{x_{k+1}}$ are connected, ϕ'_k is 1, otherwise it is 0.

When v_{x_k} and v_i are equal, the expected value of ϕ'_k is

$$\mathbb{E}[\phi'_k \mid x_k = i] = \frac{2\triangle_i}{d_i(d_i-1)} = c_i \tag{4}$$

The first equality holds because there are $d_i(d_i-1)$ equally probable combinations of $(x_{k-1}, v_i, x_{k+1})(x_{k-1} \neq x_{k+1})$, of which only $2\triangle_i$ form a triangle (v_j, v_i, v_k) or a reverse triangle (v_k, v_i, v_j). The third equality holds by definition of clustering coefficient of node v_i.

To estimate C, we introduce two variables, i.e., the weighted sum Φ of ϕ_js and the sum Ψ of the reciprocal degrees of the sampled nodes:

$$\Phi = \frac{1}{r-2} \sum_{k=2}^{r-1} \phi'_k \frac{1}{d_{x_k}} \tag{5}$$

$$\Psi = \frac{1}{r} \sum_{k=1}^{n} \frac{1}{d_{x_k}} \tag{6}$$

The expectations of Φ and Ψ are easily determined by linearity of the expectation and Eq (4):

$$\mathbb{E}[\Phi]=\mathbb{E}[\phi'_k \frac{1}{d_{x_k}}]$$
$$=\sum_{i=1}^{n} \pi'(i)\mathbb{E}[\phi'_k \mid x_k = i]\frac{1}{d_i} \quad (7)$$
$$=\sum_{i=1}^{n} \frac{d_i}{2|E|} c_i \frac{1}{d_i} \quad (8)$$
$$=\frac{1}{2|E|}\sum_{i=1}^{n} c_i \quad (9)$$
$$\mathbb{E}[\Psi]=\mathbb{E}[\frac{1}{d_{x_k}}]$$
$$=\sum_{i=1}^{n} \pi'(i)\frac{1}{d_i} \quad (10)$$
$$=\sum_{i=1}^{n} \frac{d_i}{2|E|}\frac{1}{d_i} \quad (11)$$
$$=\frac{n}{2|E|} \quad (12)$$

As proven in [9], the NBRW and SRW can have the same stationary distribution π', namely, $\pi'(i) = d_i/2|E|$.

From the above equations, C is derived as

$$C = \frac{1}{n}\sum_{i=1}^{n} c_i = \frac{\mathbb{E}[\Phi]}{\mathbb{E}[\Psi]} \quad (13)$$

Intuitively, both Φ and Ψ converge to their expected values and the estimator Φ/Ψ converges to C, as shown below. The estimator $\hat{C}$ of C, is defined as follows:

$$\hat{C} \stackrel{\text{def}}{=} \frac{\Phi}{\Psi} \quad (14)$$

To show that it is more effective than the existing method by SRW, we calculate the variance value for the local clustering coefficient in Eq (4).

$$Var[\phi'_k]=\mathbb{E}[(\phi'_k)^2] - \{\mathbb{E}[\phi'_k]\}^2$$
$$=\frac{2\triangle_i}{d_i(d_i-1)} \cdot 1^2 - {c_i}^2 \quad (15)$$

Conversely, the expectation value of the local clustering coefficient by SRW in [9] is as follows:

$$\mathbb{E}[\phi_k \frac{d_i}{d_i-1}] = c_i \quad (16)$$

ϕ_k is defined as follows: given an SRW $R = (x_1, x_2, \ldots, x_r)$ for $2 \leq \forall k \leq r-1$, and , if nodes $v_{x_{k-1}}$ and $v_{x_{k+1}}$ are connected, ϕ'_k is 1, otherwise it is 0.

This variance value is

$$Var[\phi_k \frac{d_i}{d_i-1}]=\mathbb{E}[(\phi_k \frac{d_i}{d_i-1})^2] - \{\mathbb{E}[\phi_k \frac{d_i}{d_i-1}]\}^2$$
$$=\frac{2\triangle_i}{{d_i}^2} \cdot \{\frac{d_i}{d_i-1}\}^2 - {c_i}^2 \quad (17)$$

According to Eq(15) and Eq(17),

$$Var[\phi'_k] < Var[\phi_k \frac{d_i}{d_i-1}] \quad (18)$$

holds. Therefore, the proposed method has smaller variance.

V. Evaluation

Finally, we support our theoretical findings in a simulation study. We use various social network datasets published by by the Stanford Network Analysis Project (SNAP) [1]. The dataset statistics are listed in Table I.

A. Networks on public datasets

The effectiveness of each estimator was evaluated on social networks with known structure published by the Stanford Network Analysis Project (SNAP) [1]. The dataset statistics are listed in Table I .

1) Amazon Network: Based on the *Customers Who Bought This Item Also Bought* feature of the Amazon website, this network was collected by crawling the Amazon website. If a product i is frequently co-purchased with product j, the graph contains an undirected edge from i to j. Each product category provided by Amazon defines a ground-truth community [1].

2) Digital Bibliography and Library Project (DBLP): DBLP provides a comprehensive list of research papers in computer science. In the DBLP co-authorship network, two authors are connected if they have published at least one paper together.

3) Gowalla: Gowalla is a location-based social networking website in which users share their locations by checking in. This undirected friendship network was collected through the public API of the network. However, the service closed in January 2012.

4) LiveJournal: LiveJournal is a free online blogging community in which users declare and share their friendships.

TABLE I
Network statistics

Network	n	D/n	C
Amazon	334,863	5.530	0.3967
DBLP	317,080	6.622	0.6324
Gowalla	196,591	9.668	0.2367
LiveJournal	3,997,962	17.35	0.2843

Fig. 5 plots the the normalized root mean squared error (NRMSE) versus the number of steps in the random walk [4], [10]. The NRMSEs calculated as $\frac{1}{C}\sqrt{\mathbb{E}[(\hat{C}-C)^2]}$, we determined in 1000 independently simulations of each case.

To estimate the NRMSE, each simulation was run independently 1000 times. In all simulations, the initial position of each random walk was drawn from the stationary distribution as described in [4], unless otherwise specified. In practical implementations, one can specify a burn-in period that allows the random walk to reach the steady-state [7].

Fig. 5 compares the NRMSEs of the proposed algorithm and the previous algorithm, which uses the counting triangles

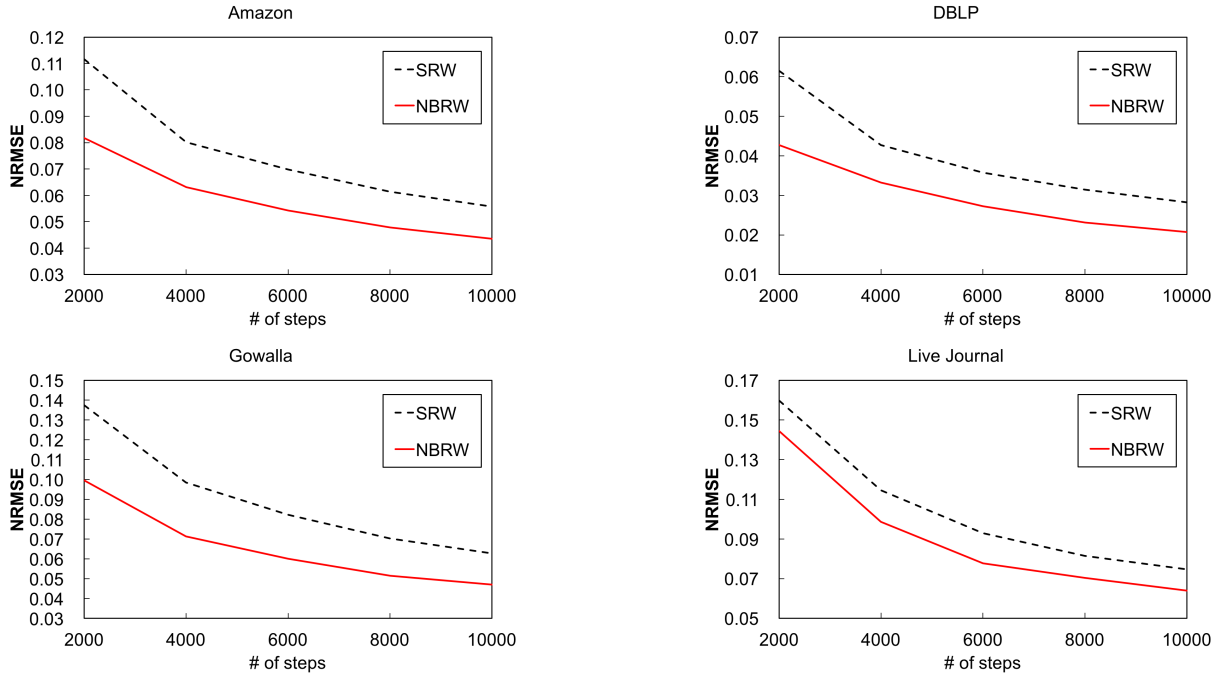

Fig. 5. Normalized root mean squared error (NRMSE) in the network clustering coefficient versus step number of the random walk.

technique in the SRW [8]. The results are presented for both algorithms on all datasets. The proposed NBRW estimator consistently outperformed the SRW estimator.

VI. Conclusions

We proposed a method that estimates the network average clustering coefficient via a NBRW. The proposed algorithm considerably outperformed the prior state-of-the-art method on various social networks.

References

[1] Stanford large network dataset collection. https://snap.stanford.edu/data/.

[2] M. Al Hasan. Methods and applications of network sampling. In *Optimization Challenges in Complex, Networked and Risky Systems*, pages 115–139. INFORMS, 2016.

[3] D. Aldous and J. Fill. Reversible markov chains and random walks on graphs, 2002.

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

[5] F. Costa, L, F. A. Rodrigues, G. Travieso, and P. R. V. Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

[6] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou. Walking in Facebook: A case study of unbiased sampling of OSNs. In *Proceedings IEEE Infocom*, pages 1–9, 2010.

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

[8] S. J. Hardiman and L. Katzir. Estimating clustering coefficients and size of social networks via random walk. In *Proceedings of the 22nd international conference on World Wide Web*, pages 539–550. International World Wide Web Conferences Steering Committee, 2013.

[9] C.-H. Lee, X. Xu, and D. Y. Eun. Beyond random walk and metropolis-hastings samplers: why you should not backtrack for unbiased graph sampling. In *ACM SIGMETRICS Performance Evaluation Review*, volume 40, pages 319–330, 2012.

[10] B. Ribeiro and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In *Internet Measurement Conference*, pages 390–403, 2010.