Average Path Length Estimation of Social Networks by Random Walk

Toshiki Matsumura, Kenta Iwasaki, and Kazuyuki Shudo
Tokyo Institute of Technology

***Abstract*—Average path length (APL) is an index of small-world networks. Calculating APL accurately requires measuring all of the shortest path lengths between two arbitrary nodes in a network. However, obtaining an entire social network is difficult because of security and privacy protection restrictions. Therefore, sampling a portion of a network to estimate features can be effective. In this research, we propose a method for estimating APL using random walk, one of the crawling-based sampling methods.**

***Index Terms*—social network, complex network, graph sampling, average path length**

I. Introduction

The number of users of social networking services (SNS) is increasing every year, and interest in analyzing such huge networks is increasing. For example, the number of Facebook users exceeds 1.8 billion in 2017. There is a method for analyzing the structure of a graph in which an SNS user is a node and the relation between two users is an edge. Networks with huge and complex structures, such as a network of human relations, are called complex networks [1], and most social networks are such networks.

In complex networks, average path length (APL) is an index of the small-world property, representing the "size" of a network [2]. APL is the average of the shortest path lengths (SPLs) between all two nodes in a network. Calculating APL accurately requires obtaining all of the SPLs, but this is unrealistic on a social network because information on all nodes cannot be obtained owing to security and privacy protection restrictions. Therefore, using graph sampling [3] to estimate this feature of a network using only a portion of the network can be effective.

There are two types of graph sampling, random sampling and sampling by crawling. Random sampling is a method of selecting nodes and edges from the network with independent probabilities and sampling them. However, in a social network, it is almost impossible to obtain information on all nodes in the network. Sampling by crawling, a method of selecting and sampling adjacent nodes and edges without needing to know the entire network can thus be applied to social networks.

Breadth-first sampling (BFS) and snowball sampling, one of the typical crawling algorithm, can not accurately estimate features of the original network because the sample is biased around the initial node. On the contrary, random walk (RW) is likely to spread to every corner of the original network, and thus might be able to estimate features of the original network by analyzing the sampling list thoroughly. For these reasons, we propose a method for estimating APL of a social network using RW.

II. Related work

According to [4], there are two scenarios for sampling large networks: *full access scenario* and *restricted access scenario*. In full access scenario, we can access to every node (or edge) of the network, and the main purpose of sampling is visualization and acceleration. However, this scenario is not related to our research, because the network we analyze is SNS, which is mostly limited in access. Thus, our goal is to find a way to obtain APL "accurately" in restricted access scenario.

Qi and colleagues experimented with APL estimation through graph sampling [5]. They estimated APL using four sampling methods including BFS. Except for BFS, these methods assumes full access scenario and are difficult to apply to social networks. Also, BFS is unstable in accuracy in terms of network features estimation.

III. Preparation

In this section, we describe RW and the definition of APL.

A. RW

RW is a method of transitioning from the initial node to one of the adjacent nodes and collecting the information regarding that node. The simplest random walk, in which the transition probabilities are uniformly random, is called simple random walk (SRW). Let p_{v_i,v_j} be the transition probability of SRW from node v_i to node v_j:

$$p_{v_i,v_j} = \begin{cases} \dfrac{1}{d_{v_i}} & v_j \in N(v_i) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where d_{v_i} and $N(v_i)$ are the degrees of v_i and the adjacent node list, respectively.

The greatest advantage of random walk is that the visit probability to each node can be calculated mathematically. For example, the distribution π' of the visit probability to each node after step t by SRW is $\pi' = (Pr[x_t = 1], Pr[x_t = 2], ..., Pr[x_t = n])$. $\pi'^{(i)}$ converges to d_{v_i}/D [6], where $\pi'^{(i)}$ is the ith element of π'. D is the sum of the degrees and is a constant. That is, in SRW, the visit probability to each node is proportional to the degree.

The number of steps T until the visit probability converges to a stationary distribution is called the *mixing time* [7].

Proc. 2018 IEEE Int'l Conf. on Big Data and Smart Computing
(IEEE BigComp 2018), pp.611-614, January 2018

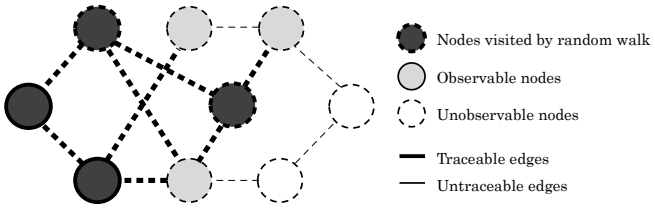

Fig. 1. An example of graph sampling by random walk: A subgraph comprising colored nodes and thick edges is denoted as G'.

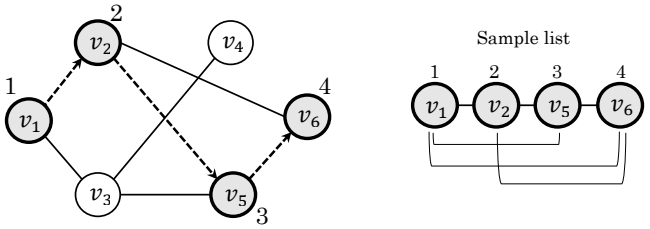

Fig. 2. Example of elements in I, if $m = 2$

The order of mixing time is approximately $O(\log^2 n)$. Even for two nodes in the sample list by random walk, if the index is greater than the mixing time, it is known that they can be regarded as random sampling from the network [8].

The sampling image of RW is shown in Fig. 1.

B. APL

APL is an indicator of the small-world property and is the average of SPLs between all of the nodes in the network. In this study, since the graph is assumed to be a graph without weight, SPL is also called the shortest step number (SSN). Letting S_{v_i,v_j} be the SSN from node v_i to node v_j, APL L is defined as

$$L = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} S_{v_i,v_j} \tag{2}$$

When the total number of nodes in the network n increases, and when L is at most as high as $\log n$, the network satisfies the small-world condition. In this paper, we estimate APL by graph sampling.

IV. Proposed Method

In this section, we propose a method for etimating the APL of social networks by random walk.

A. Process of APL Estimation

The proposed method analyzes the sample list after sampling by SRW.

1) Sampling: We acquire a sample list for r steps from graph G by SRW. For the sampled node, the node ID and the node ID of the adjacent node are held.

2) Selection of Node Pairs: To obtain the SSN, node pairs are selected from the sample list. However, node pairs whose indices are too close in the sample list are strongly correlated. For example, there is a tendency for the SSN between the ith node and the $(i+1)$th node in the sample list to be one necessarily and for the SSN between the node pairs having close indices to be small. Therefore, in this method, we use "some distance" node pairs in the sample list collected by SRW. This condition is necessary to ensure that the nodes in the pairs are chosen randomly from the graph according to the stationary distribution of SRW, i.e., that they are sampled by random sampling. For the sample list $(v_{x_1}, v_{x_2}, ..., v_{x_r})$, we define a set I of node pairs that are more than a particular threshold m above the mixing time as follows:

$$I = \{(v_{x_i}, v_{x_j}) | m \le j - i \wedge 1 \le i < j \le r, x_i \ne x_j\} \tag{3}$$

These are the node pairs used for analysis. For example, when a sample list as shown in the right-hand figure of Fig. 2 is acquired, if $m = 2$, the pair of nodes to be used as samples is $\{(v_1, v_5), (v_1, v_6), (v_2, v_6)\}$, where the indices in the list are separated by two or more.

3) Calculation of the SSN: The average of S_{v_i,v_j} in the node pairs in I is the value to be calculated; however, it is impossible to determine that S_{v_i,v_j} does not obtain the topology of the entire G. Therefore, the SSN is obtained from only the information held. Specifically, we obtain S'_{v_i,v_j} for G' formed from nodes visited by SRW and its adjacent nodes. . S'_{v_i,v_j} satisfies the relation of $S_{v_i,v_j} \le S'_{v_i,v_j}$.

4) Weighting and Calculation of the Estimated Value $\hat{L}$ of L: SRW samples in a manner biased toward high-degree nodes. In other words, node pairs selected by this method cannot successfully select samples of low-degree nodes. Therefore, we consider using a weight ω to eliminate this bias.

Let $\phi(\omega)$ be the average of S'_{v_i,v_j} in node pairs in I multiplied by the weight ω and ψ be the average of $1/d_{v_i}$ in the sample list:

$$\phi(\omega) = \frac{1}{|I|} \sum_{(v_i,v_j) \in I} \omega S'_{v_i,v_j} \tag{4}$$

$$\psi = \frac{1}{r} \sum_{i=1}^{r} \frac{1}{d_{v_i}} \tag{5}$$

These expected values are as follows:

$$E[\phi(\omega)] = E[\omega S'_{v_i,v_j}] \quad (6)$$

$$= \sum_{1\le i<j\le n, i\neq j} \omega S'_{v_i,v_j} \pi^{(i)}\pi^{(j)} \quad (7)$$

$$= \sum_{1\le i<j\le n, i\neq j} \omega S'_{v_i,v_j} \frac{d_{v_i}}{D}\frac{d_{v_j}}{D} \quad (8)$$

$$E[\psi] = E[\frac{1}{d_{v_i}}] \quad (9)$$

$$= \sum_{i=1}^{r} \frac{1}{d_{v_i}} \pi^{(i)} \quad (10)$$

$$= \sum_{i=1}^{r} \frac{1}{d_{v_i}} \frac{d_{v_i}}{D} \quad (11)$$

$$= \frac{n}{D} \quad (12)$$

Define $\hat{L}$ as follows:

$$\hat{L} = \frac{\phi(\frac{1}{d_{v_i}d_{v_j}})}{\psi^2} \quad (13)$$

This expected value as follows with attention to $S_{v_i,v_j} \le S'_{v_i,v_j}$:

$$E[\hat{L}] = \frac{1}{n^2} \sum_{1\le i,j\le n, i\neq j} S'_{v_i,v_j} \quad (14)$$

$$= \frac{n-1}{n}\frac{2}{n(n-1)} \sum_{1\le i<j\le n} S'_{v_i,v_j} \quad (15)$$

$$\ge \frac{n-1}{n}\frac{2}{n(n-1)} \sum_{1\le i<j\le n} S_{v_i,v_j} \quad (16)$$

$$= \frac{n-1}{n} L \quad (17)$$

That is, the estimated value $\hat{L}$ converges to a value greater than $\frac{n-1}{n}L$. Since n is sufficiently large in SNS, it can be regarded as $\frac{n-1}{n}L \simeq L$. In addition, as we increase the sample number by SRW, S'_{v_i,v_j} converges to S_{v_i,v_j}, with the result that $\hat{L}$ approaches L.

B. S_{v_i,v_j} and S'_{v_i,v_j}

The shortest path often includes nodes that are hubs of the network. Considering a complex network, the subgraph G' of G has many hubs, since SRW has the property that it is easy to visit a node with a high degree. In other words, it is considered that there is no extremely large difference between the shortest path in G' and that in G.

C. How to Determine Threshold m

Since the mixing time is a value that varies depending on the random walk method and the network to be used, it is difficult to determine. If m is less than the mixing time, the node pairs that are not independent, i.e., node pairs connected by a relatively short number of steps, are sampled, with the result that the estimated value converges to a smaller value than expected. On the contrary, if m is too large, the elements of I decreases in number and the estimated value must be obtained with only a few samples, with the result that the variance becomes large. Therefore, we must increment m, estimate APL, and observe the change in $\hat{L}$.

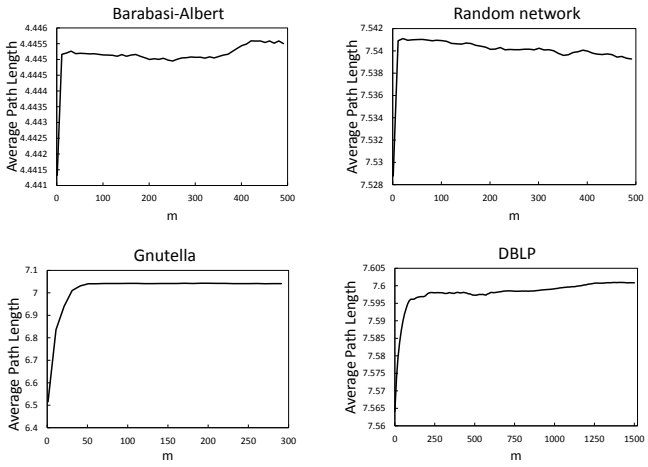

Fig. 3. Transformation of $\hat{L}$ resulting from increasing m from one in each network

V. Experiment

In this section, we examine the method of determining m mentioned in Section IV and we evaluate the estimate by the proposed method using this m. For comparison, we use two APL estimation methods, using BFS and determining APL of a subgraph using SRW.

A. Network Datasets

For the experiment, we use four datasets, a network created according to the Barabasi-Albert (BA) model, one of the generation models of complex networks, a random network, and two datasets of the Stanford Network Analysis Project (SNAP) [9]. The characteristics of each network are summarized in Table II.

TABLE I
Summary of datasets

Network	n	D/n	L
BA model	100,000	9.9995	4.2989
Random Network	100,000	9.9924	5.2631
Gnutella	62,581	4.7275	5.9355
DBLP	317,080	6.6221	6.7815

B. Preliminary Experiment

We experiment with each network on the method of determining m, as described in Section IV. Specifically, we sampled nodes by SRW as much as 5% of the total number of nodes n, and observed changes in value resulting increasing m. The results are shown in Fig. 3. According to this result, increasing m in any network shows that the estimated value begins to converge at some point. For example, looking at

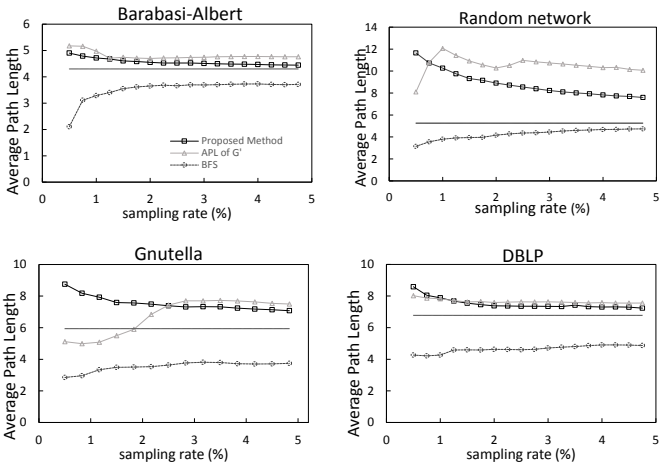

Fig. 4. Transformation of $\hat{L}$ with respect to sample rate in each network

the Gnutella network in Fig. 3, we understand that the value begins to converge with m approximately equal to 50. In other words, the mixing time of the SRW in the Gnutella network is roughly 50. Assuming that m is 100, it can be considered that sufficiently independent node pairs can be selected. In this way, m is determined for each network.

C. Estimation Experiment of Average Distance

We observe the change of $\hat{L}$ mentioned in Section IV resulting from changing the sample rate. Fig. 4 shows the transition of $\hat{L}$ for each network. m is determined as described in Section V.

D. Results and Discussion

As the sample rate increases, the proposed method shows a trend gradually approaching from a value larger than the true value; in the three networks except the random network at the sample rate of 5%, it turns out to be an accurate method. As mentioned in Section IV, in a random network with uniform degree and no large hub, it is considered to be a factor that there is a large difference between S_{v_i,v_j} and S'_{v_i,v_j}. Actually, Fig. 5 shows a perfect match rate between S_{v_i,v_j} and S'_{v_i,v_j}. This shows that while the perfect match rate achieves 80% to 90% in the scale-free network, in the random network , it is extremely low. However, social networks often have extremely many friendships for celebrities and others, and mostly they are scale-free networks. Therefore, the proposed method can be expected to be effective for social networks.

VI. Summary and Future Challenges

In this paper, we proposed a method for estimating APL by graph sampling using random walk for social networks for which random sampling is difficult. Since the proposed method uses a crawling-based graph sampling method, it can be applied to real SNS. Furthermore, the proposed method is an effective method for scale-free networks. The contribution of this research is that more accurate values can be got than any other sampling method In real social networks.

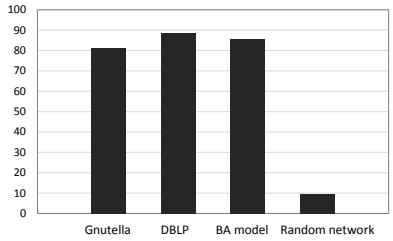

Fig. 5. Near perfect match rate between S_{v_i,v_j} and S'_{v_i,v_j}

The experiment reported in this study was conducted with the sample rate set to 5% uniformly in each network. However, when graph sampling is actually applied to SNS, it is desirable to set the time for random walk to crawl the network as, e.g., one week, and to estimate features using the obtained sample.

Finally, although the proposed method uses only portion of the network, the calculation time takes $O(r^3)$ for the number of samples r using Dijkstra's method, the simplest method. Solving this problem requires considering the efficiency of the algorithm. For example, instead of selecting node pairs separated by more than m steps from the sample list by SRW, if we select node pairs that are just m steps away, accuracy will be slightly worse but the number of calculations of the SSN can be expected to decrease. Evaluation of this trade-off is a subject for a future study.

Acknowledgment

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).

References

[1] Réka Albert and Albert-László Barabási. Statistical Mechanics of Complex Networks. *Reviews of Modern Physics*, Vol. 74, No. 1, p. 47, 2002.

[2] Xiao Fan Wang and Guanrong Chen. Complex Networks: Small-World, Scale-Free and Beyond. *IEEE Circuits and Systems Magazine*, Vol. 3, No. 1, pp. 6–20, 2003.

[3] Pili Hu and Wing Cheong Lau. A Survey and Taxonomy of Graph Sampling. *arXiv preprint arXiv:1308.5865*, 2013.

[4] Mohammad Al Hasan. Methods and Applications of Network Sampling. In *Optimization Challenges in Complex, Networked and Risky Systems*, pp. 115–139. INFORMS, 2016.

[5] Qi Ye, Bin Wu, and Bai Wang. Distance Distribution and Average Shortest Path Length Estimation in Real-World Networks. In *International Conference on Advanced Data Mining and Applications*, pp. 322–333. Springer, 2010.

[6] David Aldous and Jim Fill. Reversible Markov Chains and Random Walks on Graphs, 2002.

[7] Abedelaziz Mohaisen, Aaram Yun, and Yongdae Kim. Measuring the Mixing Time of Social Graphs. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, pp. 383–389. ACM, 2010.

[8] Stephen J Hardiman and Liran Katzir. 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. ACM, 2013.

[9] Stanford Large Network Dataset Collection. https://snap.stanford.edu/data/.