Estimating Top-k Betweenness Centrality Nodes in Online Social Networks

Kazuki Nakajima
Tokyo Institute of Technology
Tokyo, Japan
kazuibasou@gmail.com

Kenta Iwasaki
Tokyo Institute of Technology
Tokyo, Japan
iwaken59371518@gmail.com

Toshiki Matsumura
Tokyo Institute of Technology
Tokyo, Japan
toshiki7507@gmail.com

Kazuyuki Shudo
Tokyo Institute of Technology
Tokyo, Japan
shudo@is.titech.ac.jp

***Abstract*—Betweenness centrality is a widely used network measure in social network analysis. There are many algorithms for calculating or approximating this measure, but most of these algorithms assume that all network information is known. Therefore, since we can obtain little information for online social networks because of security and privacy concerns, we must consider a crawling-based algorithm.**

Herein, we propose a new crawling-based algorithm for estimating the top-k nodes with the highest betweenness centrality in online social networks. The proposed algorithm approximates the ego betweenness centrality of the nodes sampled via a random walk and approximates the top-k nodes with the highest betweenness centrality in a graph as the top-k nodes with the highest approximated ego betweenness centrality in the sampled nodes. This algorithm makes the same number of requests to an application programming interface as existing algorithms because we reuse sampled nodes for approximation. Our experimental results show that the proposed algorithm can estimate the top-k nodes of real social networks more accurately than existing methods when sample size is very small.

***Index Terms*—social network, betweenness centrality, ego betweenness centrality, random walk**

I. Introduction

Online social networks (OSNs) have recently gained considerable attention. For example, Twitter had over 300 million monthly active users in 2018 [1]. There are many studies that have analyzed a particular OSN as a graph with nodes of users and edges of relationships among users. In [2]–[4], the authors investigate structural measures of OSNs such as the clustering coefficient, average shortest path length, and graphlets.

Identifying influential users in an OSN is important in marketing [5], [6]. Centrality indicates the importance of nodes in a graph. In particular, the betweenness centrality proposed by Freeman [7] has been actively used in network analysis [8]–[10]. The betweenness centrality of a node is defined as the sum of the ratio of the shortest paths between any two nodes in a graph that pass through that node. Users with high betweenness centrality in OSNs are instrumental for the spread of information because these nodes are present on many of the shortest paths in a graph.

Our goal is to estimate the top-k nodes with the highest betweenness centrality in OSNs. If we have all the information about a network, we can estimate the top-k betweenness centrality nodes using an exact algorithm [11] or approximation algorithms [12]–[17]. However, these methods cannot be applied to OSNs because privacy and security concerns prevent one from obtaining complete network information. In practical scenarios, we can obtain the information about a target user by making a request to the application programming interface (API). In most OSNs, we can sample the information using a crawling algorithm that traverses a list of neighboring users provided by the API. However, this gives considerably less than total information because the number of requests per unit time is limited for most OSNs.

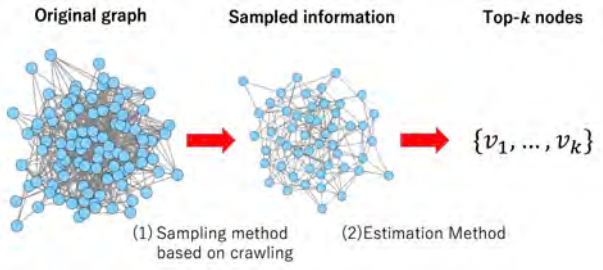


Fig. 1. Overview of crawling-based algorithms that estimate top-k nodes.

Therefore, we need crawling-based algorithms that consider the limitation of requests. As shown in Fig. 1, crawling-based algorithms consist of two methods: (1) the sampling method in which the nodes or edges are sampled from a graph by crawling and (2) the estimation method in which the top-k nodes with the highest betweenness centrality in the graph are estimated from the sampled information. Two indices are used for the evaluation of each method: *the collection ratio* and *the overlap ratio* [18]. The former is the ratio of the top-k nodes that are sampled, whereas the latter is the ratio of the top-k nodes that are accurately estimated. A sampling method needs to be chosen whose collection ratio is overall high in most graphs because the nodes that are not sampled cannot be estimated. In addition, we need an estimation method whose overlap ratio is high.

In this paper, we propose a new crawling-based algorithm to estimate the top-k betweenness centrality nodes in OSNs. The proposed method approximates the ego betweenness centrality [16] of the nodes sampled via a random walk and then approximates the top-k betweenness centrality nodes in a graph as the nodes with the highest approximated top-k ego

Proc. IEEE ISPA-IUCC-BDCloud-SocialCom-SustainCom 2018,
pp.1128-1135, December 2018

betweenness centrality in the sampled nodes. Our experimental results show that the overlap ratio of the proposed method is higher for real social networks when sample size is very small than that of existing methods in [18] and [19].

II. RELATED WORK

A number of studies have focused on estimating the top-k betweenness centrality nodes. In this section, we describe the cases in which all the network information can be obtained and discuss cases in which only a limited amount of network information can be obtained.

A. Cases Wherein All Network Information Can Be Obtained

Currently, the fastest known algorithm for exactly calculating the betweenness centrality of all the nodes is Brandes' algorithm [11]. This algorithm solves the single-source shortest path problem (SSSP) from every node (v) and then traverses backward on these paths to efficiently compute the contribution of the shortest paths from v to the betweenness centrality of other nodes. The algorithm requires at least $O(nm)$ time for the unweighted graphs and $O(nm+n^2 \log n)$ time for the weighted graphs, where n is the number of nodes and m is the number of edges. Therefore, for considerably large graphs, Brandes' algorithm is not practical for performing exact calculations. For this reason, many approximation algorithms have been proposed [12]–[17].

There are two approximation approaches that are used to estimate the top-k betweenness centrality nodes. The first approach is to approximate the betweenness centrality of all nodes via a random sampling [12]–[14]. Brandes and Pich solved the SSSP using a small set of nodes randomly sampled from nodes in a graph [12]. Yoshida randomly sampled node pairs and computed their shortest paths by using the breadth-first search algorithm [13]. Riondato and Kornaropoulos approximated the betweenness centrality of all nodes by randomly sampling the shortest paths between any two nodes [14].

The second approach is to approximate the top-k betweenness centrality nodes as the top-k nodes of another centrality [15]–[17]. The other centrality needs to be calculated easily for each node from little information, and many of the top-k nodes of another centrality should be top-k betweenness centrality nodes. Pfeffer and Carley used the k-betweenness centrality (proposed by Borgatti and Everett [20]), which selected node pairs with the shortest path length of k or less from a graph [15]. Everett and Borgatti used the ego betweenness centrality of each node, which selects the node pairs from their neighbors and is equivalent to 2-betweenness centrality [16]. Kourtellis et al. proposed the k-path centrality [17] and used it for approximation. The k-path centrality considers random paths in a graph whose length is k or less but which need not necessarily be the shortest paths. In addition, Kourtellis et al. proposed an algorithm that approximates the k-path centrality of all nodes via a random sampling.

In this study, we used ego betweenness centrality; however, our method is different from that proposed in previous research [16] in terms of the assumption that considerably limited information is available from OSNs. Everett and Borgatti assumed that all the network information was available when they calculated the ego betweenness centrality of all the nodes. Therefore, it is impractical to apply this method to OSNs because we can obtain the little information. For the same reason, other exact or approximation algorithms [11]–[15], [17] can not be applied.

B. Cases Wherein Limited Network Information Can Be Obtained

There are two crawling-based algorithms which can be applied to OSNs [18], [19]. Each crawling-based algorithm is composed of two methods: (1) the sampling method in which the nodes or edges are sampled from a graph by crawling and (2) the estimation method in which the top-k nodes are estimated from the sampled information. Lim et al. introduced the following indicators for evaluating each method [18].

- **Collection ratio**: The ratio of the number of top-k nodes that are sampled to k.
- **Overlap ratio**: The ratio of the number of top-k nodes that are accurately estimated to k.

The collection ratio depends on the sampling method, whereas the overlap ratio depends on the estimation method.

First, we describe sampling methods. Ideally, we should choose the sampling method whose collection ratio is overall high for many networks with small sample size. There are a number of crawling-based sampling methods, such as breadth-first search sampling and random walk sampling. Previous studies [18], [19] have compared the collection ratios of various crawling-based sampling methods, and the collection ratio of random walk sampling was found to be overall high in most graphs. In addition, when we use a random walk sampling, we have the advantage of being able to perform a theoretical analysis because the stationary distribution of a random walk can be calculated.

Next, we describe estimation methods. Maiya and Berger-Wolf induced a subgraph, G', from a set of sampled nodes, V', in the graph $G = (V, E)$ and approximated the top-k betweenness centrality nodes in G as the top-k betweenness centrality nodes in G' [19]. A subgraph induced by V' in G is defined as $G' = (V', E')$, where $E' = \{(v_i, v_j) \in E : v_i, v_j \in V'\}$.

Lim et al. approximated the top-k nodes with the highest betweenness centrality nodes in G as the top-k nodes with the highest degree centrality among the nodes sampled via a random walk [18]. Nodes with high degree centrality frequently have high betweenness centrality [21], [22]. In addition, the degree centrality of each sampled node can be calculated exactly because all neighboring nodes of the sampled nodes are obtained via a random walk.

Our method approximates the top-k betweenness centrality in a graph as the top-k nodes with the highest approximated ego betweenness centrality in the nodes sampled via a random walk. In Section V, we show that the overlap ratio of our

proposed method is higher than that of existing methods [18], [19] on real social networks.

III. Preliminaries

In this section, we describe the notation, several centrality definitions, and random walk.

A. Notations and Definitions

OSNs can be modeled as an undirected and unweighted graph, $G = (V, E)$, with a set of nodes, $V = \{v_1, v_2, ..., v_n\}$, and a set of edges E. We assume that $|V| = n$ and that the graph G is a simple and connected graph. Let $N(i) = \{v_j \in V : (v_i, v_j) \in E\}$ denote the set of neighbors of the nodes $v_i \in V$ and $d_i = |N(i)|$ denote the degree of node v_i.

The betweenness centrality of v_i is defined as the sum of the ratio of the shortest path between all pairs of nodes in the graph that pass through v_i [7]. Let $\sigma_{j,k}$ denote the number of shortest paths between v_j and v_k. For $v_i \in V$, let $\sigma_{j,k}(i)$ denote the number of shortest paths between v_j and v_k that pass through v_i. For any $v_i \in V$ and $v_j \in N(i)$, we define $\frac{\sigma_{j,j}(i)}{\sigma_{j,j}} = 0$.

Definition 1. *The betweenness centrality of v_i is defined as*

$$BC(i) = \sum_{v_j, v_k \in V \setminus \{v_i\}} \frac{\sigma_{j,k}(i)}{\sigma_{j,k}}.$$

The k-betweenness centrality is defined as the sum of the ratio of the shortest paths between node pairs having the shortest path length as k or less [20]. Let $dist_{i,j}$ denote the shortest path length between v_i and v_j.

Definition 2. *The k-betweenness centrality of v_i is defined as*

$$BC_k(i) = \sum_{v_{i_1}, v_{i_2} \in V \setminus \{v_i\}, dist_{i_1,i_2} \leq k} \frac{\sigma_{i_1,i_2}(i)}{\sigma_{i_1,i_2}}.$$

An ego network is a network comprised of a certain node and its neighbors. The ego betweenness centrality of v_i is the betweenness centrality of v_i in the ego network of v_i [16].

Definition 3. *The ego betweenness centrality of v_i is defined as*

$$eBC(i) = \sum_{v_j, v_k \in N(i)} \frac{\sigma_{j,k}(i)}{\sigma_{j,k}}.$$

We define $N_j(i) = N(i) \setminus (N(j) \cup \{v_j\})$. For a simple graph, the following proposition is obtained from this definition:

Proposition 1.

$$eBC(i) = \sum_{v_j \in N(i)} \sum_{v_k \in N_j(i)} \frac{1}{|N(j) \cap N(k)|}.$$

Proof. When $v_k \in N_j(i)$, the number of shortest paths between v_j and v_k that pass through v_i equals 1 because there are no multiple edges and $(v_j, v_k) \notin E$. In addition, there is a shortest path between v_j and v_k for each common neighbor of v_j and v_k. Therefore, $\sigma_{j,k} = |N(j) \cap N(k)|$ holds. When $v_k \notin N_j(i)$, the number of shortest paths between v_j and v_k that pass through v_i equals 0 because $(v_j, v_k) \in E$. Therefore, for each $v_j \in N(i)$,

$$\frac{\sigma_{j,k}(i)}{\sigma_{j,k}} = \begin{cases} \frac{1}{|N(j) \cap N(k)|} & (v_k \in N_j(i)) \\ 0 & (otherwise) \end{cases}.$$

□

We also obtain following proposition.

Proposition 2. *For any nodes v_i, the following inequality is satisfied:*

$$0 \leq eBC(i) \leq d_i(d_i - 1).$$

Proof. For nodes $v_i \in V$ and $v_j, v_k \in N(i)$, $0 \leq \frac{\sigma_{j,k}(i)}{\sigma_{j,k}} \leq 1$ holds. From Definition 3, the lower bound is obtained if and only if $\frac{\sigma_{j,k}(i)}{\sigma_{j,k}} = 0$ for all $v_j, v_k \in N(i)$. For any $v_j \in N(i)$, we defined $\frac{\sigma_{j,j}(i)}{\sigma_{j,j}} = 0$. Therefore, the upper bound is obtained if and only if $\frac{\sigma_{j,k}(i)}{\sigma_{j,k}} = 1$ for all $v_j \in N(i)$ and $v_k \in N(i) \setminus \{v_j\}$. □

The degree centrality of v_i is defined as

$$DC(i) = d_i.$$

B. Random Walk

In a random walk, a walker uniformly selects a neighbor in a random fashion and traverses the node. This process repeats until the target number of the sampled nodes is reached. We sample the information, such as IDs in OSNs, of nodes which traversed by random walk. Here let $R = (x_1, x_2, ..., x_r)$ be a sequence of indices of nodes sampled via a random walk with r steps. The transition probability of node v_i to node v_j in a random walk is defined as

$$p_{i,j} = \begin{cases} \frac{1}{d_i} & (v_j \in N(i)) \\ 0 & (otherwise) \end{cases}.$$

Let $Pr[A]$ denote the probability that event A occurred. We denote the distribution induced by R as follows:

$$\pi_R = (Pr[x_r = 1], Pr[x_r = 2], ..., Pr[x_r = n]).$$

After many steps, the probability $Pr[x_r = i]$ converges to a certain value $\pi(i)$. The vector $\pi = (\pi(1), \pi(2), ..., \pi(n))$ is called the stationary distribution of G. In particular, the stationary distribution of a random walk is given as follows [23]:

$$\pi(i) = \frac{d_i}{2|E|}.$$

Herein, we assume that the initial node v_{x_1} is drawn from the stationary distribution of G, as described in [23]. In OSNs, it is impossible to do this in practice. Therefore, we randomly select the initial node from V and repeat the random walk process with some steps until it converges to the stationary distribution, after which we start sampling. The number of the steps depends on the mixing time of G and many OSNs are known to have low mixing time [24].

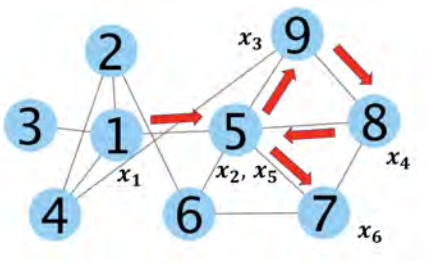


Fig. 2. Example of a random walk.

IV. Proposed Method

Our estimation method approximates the top-k nodes with the highest betweenness centrality in a graph as the top-k nodes with the highest approximated ego betweenness centrality in the nodes sampled via a random walk. Many OSNs have the low average shortest path length, which are called small world networks [25]. For example, in Twitter, it is suggested that the average shortest path length is 4.12, which is a considerably small value [26]. For such small world networks, many nodes with high k-betweenness centrality [20] also have high betweenness centrality, where the value of k is considerably small (e.g., 2, 3, and 4) [15], [27]. This is because high betweenness centrality nodes mediate many node pairs whose shortest path length is small. We use the ego betweenness centrality that is equal to 2-betweenness centrality, because we need to make additional requests in comparison with those required for the existing methods [18], [19] when k was larger than 2. However, even though k is 2, as shown in our experiment, ego betweenness centrality is good approximation for betweenness centrality in real social networks.

A. Approximating Ego Betweenness Centrality

Here, we describe how to approximate the ego betweenness centrality of nodes sampled via a random walk. Let $R = (x_1, x_2, ..., x_r)$ be a sequence of indices of the nodes sampled via a random walk with r steps. We assume that the initial node v_{x_1} is drawn from the stationary distribution of G. We define the set $I(i)$ for each sampled node v_i as follows:

$$I(i) = \{l : x_l = i, 2 \leq l \leq r-1\}.$$

When we visit v_i at l step $(2 \leq l \leq r-1)$, l is an element of $I(i)$. For each $l \in I(i)$, we define a variable $\phi_l(i)$ as follows:

$$\phi_l(i) = \begin{cases} \frac{1}{|N(x_{l-1}) \cap N(x_{l+1})|} & (v_{x_{l+1}} \in N_{x_{l-1}}(i)) \\ 0 & (otherwise) \end{cases}.$$

We define a variable $\Phi(i)$ as follows:

$$\Phi(i) = \frac{1}{|I(i)|} \sum_{l \in I(i)} \phi_l(i)$$

For each sampled node v_i which the condition $I(i) \neq \phi$ holds, let $\widetilde{eBC(i)}$ be the approximation for $eBC(i)$.

Definition 4. *We define* $\widetilde{eBC(i)}$ *as*

$$\widetilde{eBC(i)} = {d_i}^2 \Phi(i).$$

For example, let v_i be i in Fig. 2. Additionally, let $R = (x_1, x_2, ..., x_6) = (1, 5, 9, 8, 5, 7)$ be a sequence of indices of the nodes sampled by a random walk of six steps. In this case, $I(5) = \{2, 5\}, I(8) = \{4\}$, and $I(9) = \{3\}$ holds. For example, we calculate $\widetilde{eBC(5)}$. When $l = 2$, the conditions $i = x_2 = 5$, $v_{x_1} = 1$, $v_{x_3} = 9$, and $v_{x_3} \in N_{x_1}(5)$ hold. $|N(x_1) \cap N(x_3)| = |\{4, 5\}| = 2$; therefore, we conclude that $\phi_2(5) = \frac{1}{2}$. When $l = 5$, the conditions $i = x_5 = 5$, $v_{x_4} = 8$, $v_{x_6} = 7$, and $v_{x_6} \notin N_{x_4}(5)$ hold. Thus, we conclude that $\phi_5(5) = 0$. Consequently, we can obtain $\widetilde{eBC(5)} = \frac{{d_5}^2}{|I(5)|}\Phi(5) = \frac{25}{2}(\phi_2(5) + \phi_5(5)) = \frac{25}{4}$. Similarly, we obtain $\widetilde{eBC(8)} = \widetilde{eBC(9)} = 0$. In addition, $\widetilde{eBC(1)}$, $\widetilde{eBC(2)}$, $\widetilde{eBC(3)}$, $\widetilde{eBC(4)}$, $\widetilde{eBC(6)}$, and $\widetilde{eBC(7)}$ are not defined.

We can obtain the following lemma:

Lemma 1. *For any sampled nodes* $v_i \in V$*, we have that*

$$E[\widetilde{eBC(i)}] = eBC(i).$$

Proof. For each $l \in I(i)$, both $v_{x_{l-1}}$ and $v_{x_{l+1}}$ are neighbors of v_i because a random walk traverses the neighboring node. Therefore, we obtain

$$\begin{aligned} E[\widetilde{eBC(i)}] &= E[{d_i}^2 \Phi(i)] = E[{d_i}^2 \phi_l(i)] \\ &= {d_i}^2 \sum_{v_j, v_k \in N(i)} (Pr[x_{l-1} = j, x_{l+1} = k | x_l = i] \\ &\qquad \times E[\phi_l(i) | x_{l-1} = j, x_{l+1} = k]) \qquad (1) \\ &= {d_i}^2 \sum_{v_j \in N(i)} \sum_{v_k \in N_j(i)} (Pr[x_{l-1} = j, x_{l+1} = k | x_l = i] \\ &\qquad \times \frac{1}{|N(j) \cap N(k)|}). \qquad (2) \end{aligned}$$

Eq. (1) holds because of the law of total expectation. Eq. (2) holds due to the definition of $\phi_l(i)$. The condition $Pr[x_{l-1} = j, x_{l+1} = k | x_l = i]$ holds:

$$\begin{aligned} & Pr[x_{l-1} = j, x_{l+1} = k | x_l = i] \\ &= \frac{Pr[x_{l-1} = j, x_l = i, x_{l+1} = k]}{Pr[x_l = i]} \qquad (3) \\ &= \frac{Pr[x_{l-1} = j] \cdot p_{j,i} \cdot p_{i,k}}{Pr[x_l = i]} \qquad (4) \\ &= \frac{\frac{d_j}{2|E|} \cdot \frac{1}{d_j} \cdot \frac{1}{d_i}}{\frac{d_i}{2|E|}} \qquad (5) \\ &= \frac{1}{{d_i}^2}. \end{aligned}$$

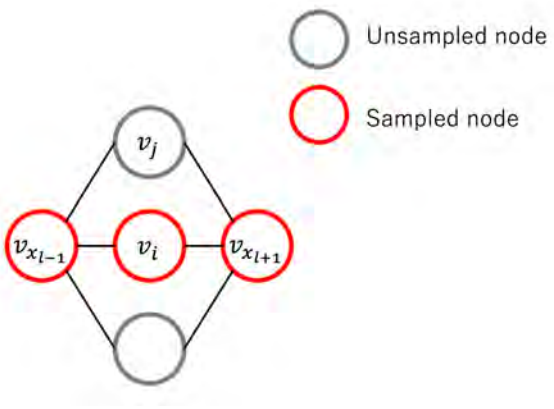


Fig. 3. An example of calculating $\phi_l(i)$.

Eq. (3) holds due to the definition of the conditional probability. Eqs. (4) and (5) holds because of the transition probability and stationary distribution of a random walk. Therefore,

$$\begin{aligned} E[\widetilde{eBC(i)}] &= {d_i}^2 \sum_{v_j \in N(i)} \sum_{v_k \in N_j(i)} \frac{1}{{d_i}^2} \frac{1}{|N(j) \cap N(k)|} \\ &= \sum_{v_j \in N(i)} \sum_{v_k \in N_j(i)} \frac{1}{|N(j) \cap N(k)|} \\ &= eBC(i). \end{aligned}$$

□

The approximation error is relatively large for nodes whose $|I(i)|$ is small. However, small $I(i)$ means that v_i is visited little, and such nodes have often low degree because nodes with low degree is rarely traversed in a random walk. Their exact ego betweenness centrality is also small frequently (See Proposition 2).

B. Number of Requests to API

In OSNs, we can obtain the IDs for given user's neighbors by making a request to the APIs. In a crawling-based algorithm, we need to consider the total number of requests because we are allowed only a limited number of requests per unit time in most OSNs.

To avoid repetitive requests for the nodes that have already been traversed once, we store the IDs of the neighboring nodes for each sampled node. Therefore, the number of requests to the API of sampling via r steps of a random walk is equal to the number of sampled nodes n' ($n' \leq r$): all algorithms need to make requests at the least n' times.

As well as existing crawling-based algorithms [18], [19], our algorithm only makes requests n' times because there are no additional requests in approximating ego betweenness centrality of each sampled nodes. In our algorithm, we calculate $\phi_l(i)$ for each $l \in I(i)$. We consider the case of Fig. 3. We can check $v_{x_{l+1}} \in N_{x_{l-1}}(i)$ and calculate $\phi_l(i) = \frac{1}{3}$ without sampling additional common neighbors such as v_j. This is because we know $N(x_{l-1})$ and $N(x_{l+1})$ since we sampled them.

TABLE I
DATASETS.

Network	$\|V\|$	$\|E\|$
Epinions [28]	75,877	405,739
soc-buzznet [29]	101,163	2,763,066
Gowalla [28]	196,591	950,327
soc-academia [29]	200,167	1,022,440
Dogster [28]	426,485	8,543,321
soc-flickr [29]	513,969	3,190,452

V. Experimental Results

We used real OSNs to evaluate our method. We focus on undirected and simple graphs by removing the directions of edges if the graphs are directed, treating multiple edges as a single edge, and deleting the loops. Additionally, for an unconnected graph, we deleted the nodes that are not contained in the largest connected component of the graph. Table I lists number of nodes and number of edges.

To compare our method with existing methods [18], [19], we use the collection ratio and the overlap ratio. Each crawling-based algorithm consists of a sampling method and an estimation method. Both our and existing algorithms use the same sampling method, namely random walk: these algorithms have same collection ratio. Therefore, in order to compare the overlap ratio of each estimation method, we calculate the overlap ratio of each estimation methods from the same sample nodes in each simulation. Each simulation is performed as follows:

1) We sample n' nodes from an original graph by a random walk. we select the initial node of a random walk from the stationary distribution of the graph.
2) We estimate top-k nodes from the same n' sampled nodes by each estimation method (namely the proposed method, Lim et al.'s method, and Maiya et al.'s method).
3) We calculate overlap ratios of each estimation method.

We ran the simulation independently 1000 times and compared the average of overlap ratios of each estimation method. We discuss the average of each method when n' is changed and when k is changed, respectively.

First, we discuss the experimental result when n' is changed. Fig. 4 shows the average of the collection ratio and overlap ratios of the proposed method and those of Lim et al. and Maiya et al. when we vary n' from 1000 to 5000 in increments of 1000 (Indicated by a black dotted line, a black solid line, a red solid line, a black dashed line, a black dash-dot line). As mentioned above, each algorithm has the same collection ratio. Note that the collection ratio is the maximum value of the overlap ratio, and it is advantageous to have these two values close to each other.

Firstly, we discuss the collection ratio of a random walk. In all graphs, we could collect many top-10 nodes with the highest betweenness centrality in spite of small sample sizes. We often traverse nodes with high degree in a random walk: therefore, nodes with high betweenness centrality often have high degree or are close to nodes with high degree.

TABLE II
THE AVERAGE OVERLAP RATIO OF TOP-k NODES IN 1000 SIMULATIONS (n' = 5000).

Network	Ours k=10	Lim et al. k=10	Maiya et al. k=10	Ours k=20	Lim et al. k=20	Maiya et al. k=20	Ours k=50	Lim et al. k=50	Maiya et al. k=50
Epinions	**0.881**	0.800	0.849	**0.817**	0.750	0.793	0.764	0.740	**0.781**
soc-buzznet	**0.903**	0.900	0.800	**0.921**	0.900	0.784	**0.792**	0.740	0.719
Gowalla	**0.856**	0.800	0.801	0.764	**0.796**	0.727	**0.740**	0.719	0.716
soc-academia	**0.851**	0.801	0.810	0.847	**0.854**	0.853	0.739	**0.740**	0.656
Dogster	**0.915**	0.900	0.878	0.851	0.753	**0.859**	**0.881**	0.844	0.804
soc-flickr	**0.763**	0.682	0.697	**0.609**	0.509	0.594	0.559	0.335	**0.565**

TABLE III
THE OVERLAP RATIO BETWEEN EXACT TOP 10 BETWEENNESS CENTRALITY NODES AND EXACT TOP 10 EGO BETWEENNESS CENTRALITY (DEGREE CENTRALITY) NODES.

Network	Ego betweenness	Degree
Epinions	**1.0**	0.8
soc-buzznet	**0.9**	**0.9**
Gowalla	**0.9**	0.8
soc-academia	**0.8**	**0.8**
Dogster	**1.0**	0.9
soc-flickr	**0.9**	0.7

Then, we compare the overlap ratios of the proposed method and existing methods. We obtained three results. First, the overlap ratios of proposed method are higher than those of Lim et al.'s method in all graphs when n' = 5000. The overlap ratio of the proposed method converges to the overlap ratio between the exact top 10 betweenness centrality nodes and the exact top 10 ego betweenness centrality nodes in each graph. Similarly, the overlap ratio of Lim et al.'s method converges to the overlap ratio between the exact top 10 betweenness centrality nodes and the exact top 10 degree centrality nodes. Table III shows the exact overlap ratio of each method and the higher overlap ratio is shown in bold. The proposed method achieved higher overlap ratios than Lim et al.'s method, because the exact overlap ratios of ego betweenness centrality are higher overall than those of degree centrality.

Second, the overlap ratios of the proposed method are lower than those of Lim et al.'s method in all graphs when n' = 1000. This is caused by rank error among the top nodes with approximated ego betweenness centrality because of approximation error. In the future, we would like to improve the approximation error of the proposed method.

Third, the overlap ratios of the proposed method are higher overall than those of Maiya et al.'s method in all graphs. The overlap ratio of Maiya et al.'s method depends on the induced subgraph, and the betweenness centrality is defined globally in a graph. Therefore, the overlap ratio of their method is unstable when the number of sampled nodes is small.

Finally, we discuss the experimental result when k is changed. Table II shows the average of overlap ratios of the proposed method and those of Lim et al.'s method, and Maiya et al.'s method when we set k as 10, 20 and 50, respectively. The highest average overlap ratio among these methods iss shown in bold. The overlap ratios of the proposed method are higher than those of other methods in all graphs when k is equal to 10. Also, the proposed method has overall higher ratios than those of the other methods when k is equal to 20 and 50. For some graphs, the overlap ratio of the proposed method is lower than that of another method; however, this is caused by the approximation error of ego betweenness centrality. Future work consists of improving the approximation error of ego betweenness centrality of sampled nodes when the number of sampled nodes is small.

VI. CONCLUSION

We proposed a new method to estimate the top-k betweenness centrality nodes via a random walk in OSNs. We approximated the ego betweenness centrality of the nodes sampled via a random walk and the top-k betweenness centrality nodes as the top-k nodes of the approximated ego betweenness centrality in the sampled nodes. In our approximation method, we reused the sampled nodes to avoid making additional requests. For OSNs wherein only limited information about these networks can be obtained, our experimental results show that the proposed method can estimate top-k nodes more accurately than existing methods. In future work, we would like to improve the approximation error of ego betweenness centrality of sampled nodes and propose estimation algorithms for directed graphs.

VII. ACKNOWLEDGMENT

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

REFERENCES

[1] "Q2 2018 letter to shareholders," http://files.shareholder.com/downloads/AMDA-2F526X/6357113850x0x982955/2C0E130D-5400-4185-9C99-7C00EF47461B/Q2_2018_Shareholder_Letter.pdf, 2018.

[2] K. Iwasaki and K. Shudo, "Estimating the clustering coefficient of a social network by a non-backtracking random walk," in *Big Data and Smart Computing (BigComp), 2018 IEEE International Conference on*. IEEE, 2018, pp. 114–118.

[3] T. Matsumura, K. Iwasaki, and K. Shudo, "Average path length estimation of social networks by random walk," in *Big Data and Smart Computing (BigComp), 2018 IEEE International Conference on*. IEEE, 2018, pp. 611–614.

[4] X. Chen, Y. Li, P. Wang, and J. Lui, "A general framework for estimating graphlet statistics via random walk," *Proceedings of the VLDB Endowment*, vol. 10, no. 3, pp. 253–264, 2016.

[5] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[6] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 261–270.

[7] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[8] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang, "High-betweenness proteins in the yeast protein interaction network," *BioMed Research International*, vol. 2005, no. 2, pp. 96–103, 2005.

[9] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, pp. 026–113, 2004.

[10] R. Guimera and L. A. N. Amaral, "Modeling the world-wide airport network," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 381–385, 2004.

[11] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of mathematical sociology*, vol. 25, no. 2, pp. 163–177, 2001.

[12] U. Brandes and C. Pich, "Centrality estimation in large networks," *International Journal of Bifurcation and Chaos*, vol. 17, no. 07, pp. 2303–2318, 2007.

[13] Y. Yoshida, "Almost linear-time algorithms for adaptive betweenness centrality using hypergraph sketches," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1416–1425.

[14] M. Riondato and E. M. Kornaropoulos, "Fast approximation of betweenness centrality through sampling," *Data Mining and Knowledge Discovery*, vol. 30, no. 2, pp. 438–475, 2016.

[15] J. Pfeffer and K. M. Carley, "k-centralities: Local approximations of global measures based on shortest paths," in *Proceedings of the 21st International Conference on World Wide Web*. ACM, 2012, pp. 1043–1050.

[16] M. Everett and S. P. Borgatti, "Ego network betweenness," *Social networks*, vol. 27, no. 1, pp. 31–38, 2005.

[17] N. Kourtellis, T. Alahakoon, R. Simha, A. Iamnitchi, and R. Tripathi, "Identifying high betweenness centrality nodes in large social networks," *Social Network Analysis and Mining*, vol. 3, no. 4, pp. 899–914, 2013.

[18] Y. S. Lim, D. S. Menasché, B. Ribeiro, D. Towsley, and P. Basu, "Online estimating the k central nodes of a network," in *Network Science Workshop (NSW), 2011 IEEE*. IEEE, 2011, pp. 118–122.

[19] A. S. Maiya and T. Y. Berger-Wolf, "Online sampling of high centrality individuals in social networks," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2010, pp. 91–98.

[20] S. P. Borgatti and M. G. Everett, "A graph-theoretic perspective on centrality," *Social networks*, vol. 28, no. 4, pp. 466–484, 2006.

[21] K.-I. Goh, E. Oh, B. Kahng, and D. Kim, "Betweenness centrality correlation in social networks," *Physical Review E*, vol. 67, no. 1, p. 017101, 2003.

[22] M. Barthelemy, "Betweenness centrality in large complex networks," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 163–168, 2004.

[23] 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*. ACM, 2013, pp. 539–550.

[24] A. Mohaisen, A. Yun, and Y. Kim, "Measuring the mixing time of social graphs," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 2010, pp. 383–389.

[25] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 29–42.

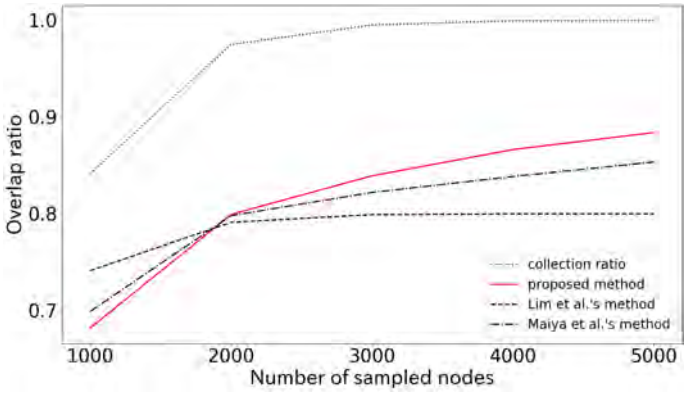
[26] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 591–600.

[27] M. Ercsey-Ravasz and Z. Toroczkai, "Centrality scaling in large networks," *Physical review letters*, vol. 105, no. 3, p. 038701, 2010.

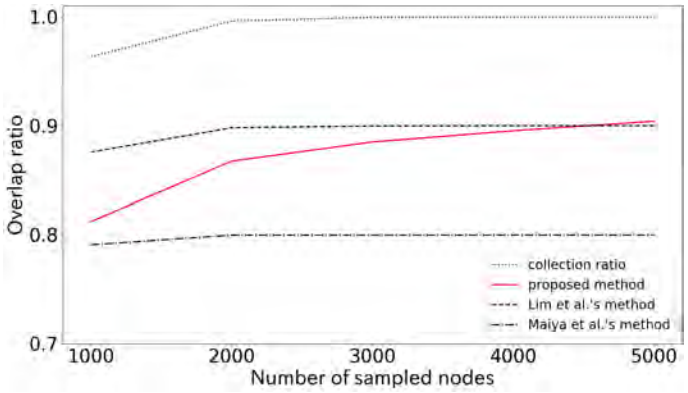
[28] "Konect datasets: The koblenz network collection," http://konect.uni-koblenz.de/.

[29] "The network data repository with interactive graph analytics and visualization," http://networkrepository.com.

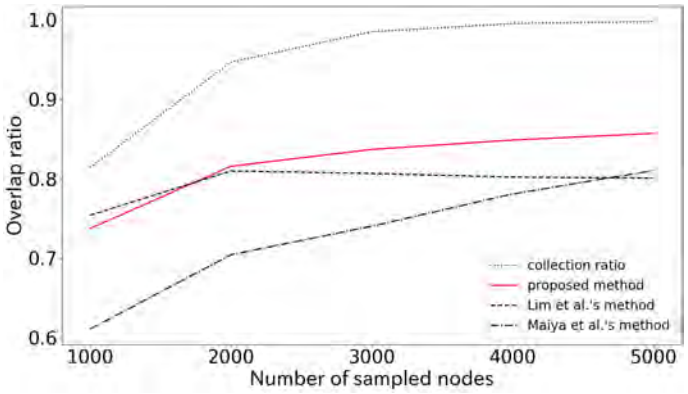
Proc. IEEE ISPA-IUCC-BDCloud-SocialCom-SustainCom 2018,
pp.1128-1135, December 2018



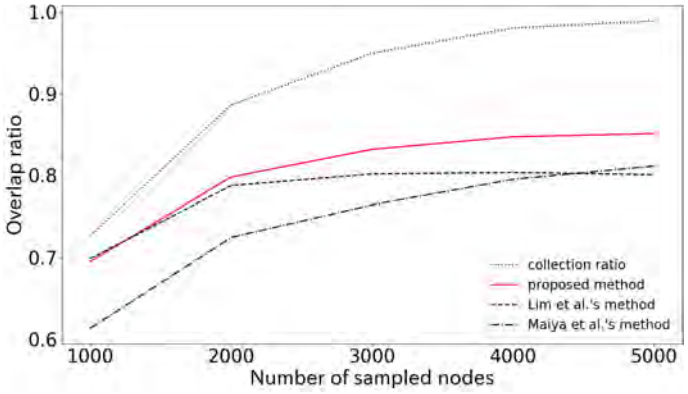
(a) Epinions



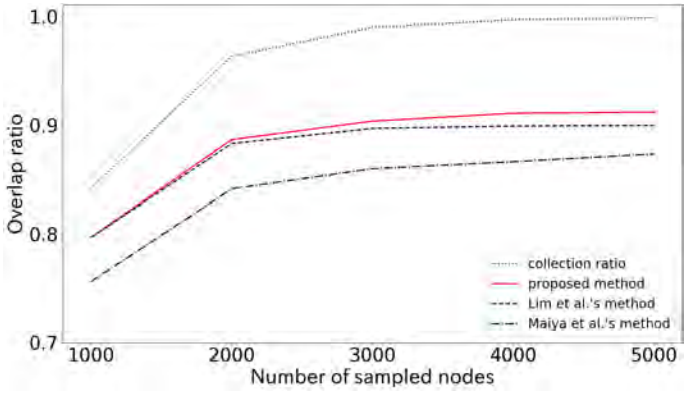
(b) soc-buzznet



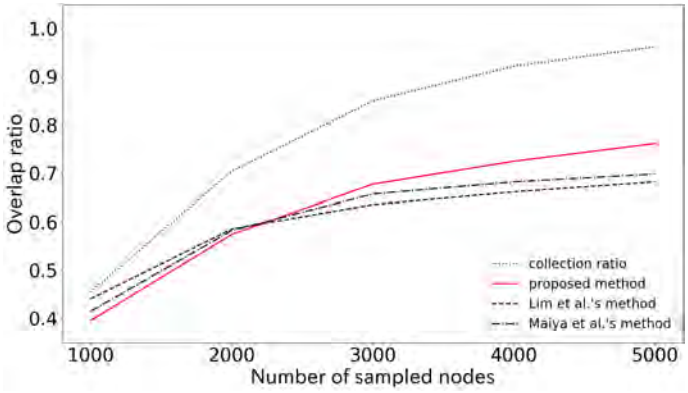
(c) Gowalla



(d) soc-academia



(e) Dogster



(f) soc-flickr

Fig. 4. The average collection ratio and average overlap ratio value of each methods in 1000 simulations (k=10).