Electronic Preprint for Journal of Information Processing Vol.28

Regular Paper

Estimating High Betweenness Centrality Nodes via Random walk in Social Networks

Kazuki Nakajima[1,a)] Kazuyuki Shudo[1,b)]

Received: December 7, 2019, Accepted: March 27, 2020

Abstract: The betweenness centrality is a widely used property to identify important nodes in social networks. Several algorithms have been studied to efficiently compute the top-k nodes with the highest betweenness centrality on a graph where all the data is available. However, all the graph data of real social networks are not typically available to third parties such as researchers or marketers, and hence, an estimation algorithm based on sampling the graph data is required. Accurately estimating the top-k nodes with the highest betweenness centrality from a small sample of a graph is a challenging task. First, the top-k nodes need to be included in the small sample. Second, nodes with the high betweenness centrality that is defined on the whole graph need to be accurately identified from the small sample. We propose a random walk-based algorithm to estimate the top-k nodes with the highest betweenness centrality by utilizing the ego betweenness centrality that has a high correlation with the betweenness centrality in social networks. The proposed algorithm firstly obtains a small sample that includes many of top-k nodes with the highest betweenness centrality via a random walk on a social network. Then, we obtain unbiased estimates of the ego betweenness centrality of sampled nodes and approximate the top-k nodes with the highest betweenness centrality as the top-k nodes with the highest estimated ego betweenness centrality. The proposed estimator efficiently estimates the ego betweenness centrality of each sample without additionally sampling the graph data by utilizing the neighbor data of the previous and the next samples. The experiments using real social network datasets show that the proposed algorithm estimates more accurately the top-k nodes with the highest betweenness centrality than existing algorithms when the sample size is small.

Keywords: social networks, betweenness centrality, ego betweenness centrality, sampling, estimation, random walk

1. Introduction

Online social networks (OSNs) have been primarily studied to understand the nature of the social structure such as human connections and behaviors [1], [17], [25], [28]. The very large-scale OSNs, such as Facebook with over 2 billion active monthly users as of December 2019 [15], significantly improve the research extent and accuracy of social network analysis [7]. A basic and effective approach to analyze the structure is to calculate the properties of the graph that consists of nodes as users and edges as users' connections. This study focuses on the *betweenness centrality* that is a property to measure the importance of nodes on a graph.

We aim to identify the top-k nodes with the highest betweenness centrality in OSNs. The betweenness centrality of a node is defined as the sum of the ratio of the shortest paths that pass through that node between any two nodes in a graph [16]. This property has been used to analyze various networks, e.g., protein interaction networks [20] and airport networks [19], and social networks for clustering and community detection [31]. Users with high betweenness centrality in OSNs have a considerable advantage in terms of the spread of information or influence because nodes with high betweenness centrality exist on the shortest paths between many node pairs.

However, the accurate identification of the top-k nodes with the highest betweenness centrality in social networks is a challenging task because of the access limitations to the graph data. Several efficient algorithms [3], [8], [10], [14], [23], [32], [34] have been studied to efficiently compute the top-k nodes with the highest betweenness centrality on a graph where all the data is available; however, all the graph data of social networks is not typically available to third parties such as researchers or marketers due to the privacy or security concerns. In practical scenarios, we sample a part of graph data through the application programming interfaces (APIs) [17] and then estimate top-k nodes from the sample.

It is non-trivial to accurately estimate the top-k nodes with the highest betweenness centrality from a small sample of the graph data. First, we need to obtain a sample that includes many of the top-k nodes because we cannot estimate top-k nodes that are not in the sample. Then, we need to accurately identify the top-k nodes that are included in the sample. The betweenness centrality of nodes in the small sample may differ greatly from the betweenness centrality on the whole graph.

There are two existing algorithms to estimate top-k nodes with the highest betweenness centrality via random walk-based sampling [26], [27]. A random walk is an effective sampling approach to obtain a sample that includes many of the top-k nodes with the highest betweenness centrality on social networks where

[1] Tokyo Institute of Technology, Meguro, Tokyo 152–8550, Japan
[a)] nakajima.k.an@m.titech.ac.jp
[b)] shudo@is.titech.ac.jp

the neighbor data of users can be obtained through querying the APIs [26]. Maiya et al. [27] proposed an algorithm that approximates the top-k nodes with the highest betweenness centrality on the original graph as the top-k nodes with the highest betweenness centrality on a subgraph induced from the sample. Lim et al. [26] proposed an algorithm that approximates the top-k nodes with the highest betweenness centrality on the original graph as the top-k nodes with the highest degree in sampled nodes.

The existing algorithms have room to improve estimation accuracy. Maiya et al.'s algorithm has a low estimation accuracy with the small sample size because the betweenness centrality of nodes in the induced subgraph has large errors between the original betweenness centrality due to missing the large part of the graph data [5], [12]. Lim et al.'s algorithm has achieved the improvement of the estimation accuracy by utilizing the top-k nodes with the highest degree that are accurately estimated with the small sample size and are largely overlapping with the top-k nodes with the highest betweenness centrality. We aim to further improve the estimation accuracy by utilizing the top-k nodes with the highest another centrality that are more largely overlapping with the top-k nodes with the highest betweenness centrality than those with the highest degree.

We propose a random walk-based algorithm that approximates the top-k nodes with the highest betweenness centrality as the top-k nodes with the highest estimated *ego betweenness centrality*. The ego betweenness centrality of a node is defined as the sum of the ratio of the shortest paths that pass through that node between only the neighbor pairs [14]. We have observed that more of the top-k nodes with the highest ego betweenness centrality are the top-k nodes with the highest betweenness centrality than those with the highest degree in real social networks. Further, the proposed algorithm obtains unbiased estimators of the ego betweenness centrality of sampled nodes by a random walk. The proposed estimator efficiently estimates the ego betweenness centrality of each sample without additionally sampling the graph data by utilizing the neighbor data of previous and next samples. The experimental results show that the proposed algorithm improves the estimation accuracy of top-k nodes with the highest betweenness centrality in real social network datasets when the sample size is small.

This paper is an extended version of our previous study [29]. The differences are as follows. First, we provide a time complexity analysis of the proposed algorithm in Section 4.2. Second, we empirically show that more of the top-k nodes with the ego betweenness centrality in real social networks are the top-k nodes with the betweenness centrality than those with the highest degree in Section 5.2. Finally, we evaluate the computation time of the existing and proposed algorithms in Section 5.3.

2. Related Work

Brandes' algorithm is currently the fastest known algorithm to exactly calculate the betweenness centrality of all the nodes in a graph [8]. 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 $O(nm + n^2 logn)$ time for the weighted graphs, where n is the number of nodes and m is the number of edges. The exact identification of the top-k nodes with the highest betweenness centrality on a large graph takes considerably computation time.

Several algorithms have been studied to reduce the computation time by approximating the top-k nodes with the highest betweenness centrality [3], [10], [14], [23], [32], [34]. Previous algorithms are classified into two main approaches. The first approach is to approximate the betweenness centrality of all nodes via a random sampling [3], [10], [34]. Brandes and Pich solved the SSSP using a small set of nodes randomly sampled from nodes in a graph [10]. Bader et al. proposed an adaptive sampling algorithm that computes the approximation for nodes with high betweenness centrality by keeping track of the partial contribution of each sampled node [3]. Riondato and Kornaropoulos approximated the betweenness centrality of all nodes by randomly sampling the shortest paths between any two nodes [34].

The second approach is to approximate the top-k nodes with the highest betweenness centrality as the top-k nodes with the highest another centrality that can be fast calculated and are largely overlapping the top-k nodes with the highest betweenness centrality [14], [23], [32]. Everett and Borgatti proposed and utilized the ego betweenness centrality that is defined as the sum of the proportion of the shortest paths that pass through a node between the neighbor pairs [14]. Pfeffer and Carley utilize the k-betweenness centrality, proposed by Borgatti and Everett [6], that is defined as the sum of the proportion of the shortest paths that pass through a node between node pairs whose shortest path length is not more than k [32]. Kourtellis et al. proposed and utilized the k-path centrality which is defined on random paths in a graph whose length is k or less but which need not necessarily be the shortest paths [23].

The proposed algorithm that utilizes the ego betweenness centrality corresponds to the second approach described above; however, our study is different from a previous study [14] in terms of the assumption that only a small part of the graph data is available. While the previous study [14] exactly computes the ego betweenness centrality of nodes, the proposed algorithm obtains unbiased estimators of the ego betweenness centrality of sampled nodes via a random walk. Similarly, our study is different from the previous studies mentioned above because of the assumption that the available graph data is limited.

Another main related work is the previous studies that aimed to obtain unbiased estimators of the properties of social networks via random walk-based sampling [2], [11], [13], [17], [21], [22], [26], [27], [33], [37]. Gjoka et al. designed a practical framework to obtain unbiased estimators of properties of social networks via the re-weighted random walk scheme where each sample obtained by a random walk is re-weighted to remove the sampling bias [17]. The algorithms based on re-weighted random walk have been studied for several graph properties [11], [13], [21], [22], [33], [37]. The proposed algorithm is based on the re-weighted random walk and is inspired by the algorithm to estimate the clustering coefficient via a random walk [21]; however, the unbiased estimator of the ego between-

ness centrality has not been studied. There are several studies [2], [26], [27] that focus on estimating the centrality of nodes via a random walk. Avrachenkov et al. studied an algorithm to accurately estimate high degree nodes via a random walk with the small sample size [2]; however, they do not focus on the betweenness centrality that is our interest.

3. Preliminaries

3.1 Notations and Definitions

We represent a social network as a connected, simple, and undirected graph $G = (V, E)$, where $V = \{v_1, v_2, ..., v_n\}$ denotes the set of n nodes (users), and E denotes the set of edges (friendship). We assume that G is static. Let $N(i) = \{v_j \in V : (v_i, v_j) \in E\}$ denote the set of neighbors of a node v_i and $d_i = |N(i)|$ denote the degree of a node v_i. We define the sum of degrees as $D = \sum_{v_i \in V} d_i$. Let d_{max} denote the maximum degree of a node in the graph. Let $N_j(i) = N(i) \backslash (N(j) \cup \{v_j\})$ denote the set of neighbors of v_i that are not v_j and not neighbors of v_j. Let $\sigma_{j,k}$ denote the number of shortest paths between v_j and v_k, and $\sigma_{j,k}(i)$ denote the number of shortest paths between v_j and v_k that pass through v_i. If $j = k$, let $\sigma_{j,k} = 1$ by convention [8], [9].

The betweenness centrality [16] of v_i is defined as the sum of the proportion of the shortest path between all pairs of nodes in the graph that pass through v_i as follows:

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

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

The ego betweenness centrality [14] of v_i is defined as the sum of the proportion of the shortest path between the neighbor pairs of v_i as follows:

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

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

3.2 Random Walk Sampling and Our Goal

We sample indices and their neighbors of nodes by performing a random walk on a graph G. In a random walk, a walker repeatedly moves to a randomly selected neighbor. The transition probability of node v_i to node v_j in a random walk is defined as

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

Let $R = \{x_s\}_{s=1}^r$ be a sequence of indices of r sampled nodes via random walk, where x_s denotes an index of the s-th sampled node. Let $Pr[A]$ denote the probability that event A occurred. We denote the distribution induced by a sample sequence, R, as follows:

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

After many steps of a random walk, the probability $Pr[x_r = i]$ converges to a certain value, $\frac{d_i}{D}$, for each node v_i [21]. The following vector $\boldsymbol{\pi}$ is called the stationary distribution of G:

$$\boldsymbol{\pi} = \left(\frac{d_1}{D}, \frac{d_2}{D}, ..., \frac{d_n}{D}\right).$$

Our goal is to accurately estimate the top-k nodes with the highest betweenness centrality on G from a sequence of indices and neighbors of r sampled nodes by a random walk, denoted by $\{(x_s, N(x_s))\}_{s=1}^r$.

3.3 Existing Algorithms

There are two existing algorithms to estimate top-k nodes with the highest betweenness centrality via random walk-based sampling in social networks [26], [27].

Maiya and Berger-Wolf proposed an algorithm that approximates the top-k nodes with the highest betweenness centrality on the original graph as the top-k nodes with the highest betweenness centrality on a subgraph induced from sampled nodes [27]. The running time of computing the betweenness centrality of all nodes in the induced subgraph is $O(n'm')$ by using Brande's algorithm [8], where n' and m' are the number of nodes and edges in the induced subgraph, respectively.

Lim et al. proposed an algorithm that approximates the top-k nodes with the highest betweenness centrality on the original graph as the top-k nodes with the highest degree in sampled nodes [26]. The running time of computing the degree centrality of r sampled nodes is $O(r)$. This algorithm achieves the improvement of the estimation accuracy by utilizing the top-k nodes with the highest degree that can be accurately estimated with the small sample size and are largely overlapping with the top-k nodes with the highest betweenness centrality in real social networks [4], [18], [30].

4. Proposed Algorithm

We propose an algorithm that approximates the top-k nodes with the highest betweenness centrality as the top-k nodes with the highest estimated ego betweenness centrality in sampled nodes. The proposed algorithm obtains unbiased estimators of the ego betweenness centrality of sampled nodes via random walk. The proposed estimator efficiently estimates the ego betweenness centrality of each sample, v_{x_s}, without additional sampling the graph data by calculating the ratio of the shortest paths that pass through v_{x_s} between the previous and next samples, $v_{x_{s-1}}$ and $v_{x_{s+1}}$.

4.1 Unbiased Estimation of Ego Betweenness Centrality

We propose an unbiased estimator of the ego betweenness centrality of each sampled node via a random walk.

First, we show the following lemma regarding the ego betweenness centrality of a node v_i on a simple graph:

Lemma 1. *If G is a simple graph, i.e., G has no loops and multiple edges, the ego betweenness centrality of each node v_i is as follows:*

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

Proof. If $v_k = v_j$, the number of shortest paths between v_j and v_k that pass through v_i equals 0 because there is a shortest path from v_j to v_k because of the definitions and the path does not pass through v_i. If $v_k \in N(j)$, the number of shortest paths between v_j and v_k that pass through v_i equals 0 because $(v_j, v_k) \in E$. If

$v_k \in N_j(i)$, 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 $(v_j, v_k) \notin E$. Additionally, $\sigma_{j,k} = |N(j) \cap N(k)|$ holds because there is a shortest path between v_j and v_k for each common neighbor of v_j and v_k. Therefore, for each $v_j \in N(i)$, it holds

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

□

Then, we define the set of the ordinal numbers of a sample sequence between 2 and $r-1$ where a node v_i is sampled as:

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

If $s \in I(i)$, it means that a node v_i is sampled at s-th step of a random walk, where s is between 2 and $r-1$.

For each $s \in I(i)$, we define the proportion of the shortest paths that pass through v_{x_s} between the previous and next samples, $v_{x_{s-1}}$ and $v_{x_{s+1}}$, as a variable $\phi_s(i)$ as follows:

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

We define the estimate of the ego betweenness centrality, $\widetilde{eBC(i)}$, of a node v_i as the average of $\phi_s(i)$ that is weighted with d_i^2 to remove the sampling bias due to a random walk:

Definition 3. *We define the estimate of the ego betweenness centrality as*

$$\widetilde{eBC(i)} = \begin{cases} \frac{1}{|I(i)|} \sum_{s \in I(i)} {d_i}^2 \phi_s(i) & (|I(i)| > 0) \\ 0 & (\text{otherwise}) \end{cases}.$$

We obtain the following lemma regarding the estimate of the ego betweenness centrality, $\widetilde{eBC(i)}$:

Lemma 2. *$\widetilde{eBC(i)}$ is an unbiased estimator of the ego betweenness centrality of v_i, $eBC(i)$.*

Proof. We show $E[\widetilde{eBC(i)}] = eBC(i)$. We obtain

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

The first equation holds because of the linearity of the expectation. The second equation holds because of the law of total expectation and both $v_{x_{s-1}}$ and $v_{x_{s+1}}$ are neighbors of v_{x_s} for each $s \in I(i)$. The third equation holds because of the definition of $\phi_s(i)$. $Pr[x_{l-1} = j, x_{l+1} = k | x_l = i]$ is derived as follows:

$$\begin{aligned} &Pr[x_{s-1} = j, x_{s+1} = k | x_s = i] = \frac{Pr[x_{s-1} = j, x_s = i, x_{s+1} = k]}{Pr[x_s = i]} \\ &= \frac{Pr[x_{s-1} = j] Pr[x_s = i | x_{s-1} = j] Pr[x_{s+1} = k | x_{s-1} = j]}{Pr[x_s = i]} \\ &= \frac{\frac{d_j}{D} \cdot \frac{1}{d_j} \cdot \frac{1}{d_i}}{\frac{d_i}{D}} = \frac{1}{{d_i}^2}. \end{aligned}$$

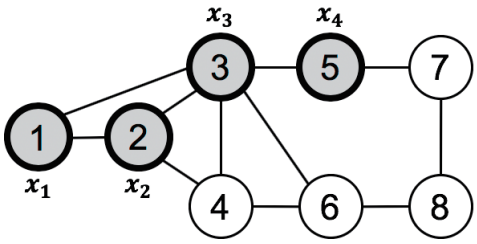


Fig. 1 Example of a random walk. x_s denotes the index of s-th sample.

The first equation holds because of the definition of the conditional probability. The second and third equations hold because of the transition probability and stationary distribution of a random walk. Therefore, we obtain the following equation by using the above equations and Lemma 1:

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

□

It is remarkable that we can obtain an unbiased estimator of the ego betweenness centrality, that needs the neighbor data in the calculation, without additionally sampling the neighbor data. The proposed estimator avoids additionally sampling the neighbor data by calculating the proportion of the shortest paths that pass through each sample between the previous and next samples, $\phi_s(i)$.

Algorithm 1 describes the algorithm to obtain an unbiased estimator of the ego betweenness centrality of each sampled node via a random walk.

Example: Let $v_i = i$ $(1 \leq i \leq 8)$ as shown in the graph in **Fig. 1**. Let $R = (x_1, x_2, x_3, x_4) = (1, 2, 3, 5)$ be a sequence of indices of the nodes sampled by a random walk with four steps. It holds $I(2) = \{2\}, I(3) = \{3\}$ and $I(1)$ and $I(5)$ are empty sets. First, we calculate $\widetilde{eBC(2)}$. It holds $\phi_2(2) = 0$ because $v_{x_3} = 3$ is not in $N_{x_1}(2) = N_1(2) = \{4\}$. Thus, we conclude that $\widetilde{eBC(2)} = 0$. Then, we calculate $\widetilde{eBC(3)}$. It holds $\phi_3(3) = 1$, because $v_{x_4} = 5$ is in $N_{x_2}(3) = N_2(3) = \{5, 6\}$ and $N(2) \cap N(5) = \{3\}$. Thus, we conclude that $\widetilde{eBC(3)} = d_3^2 \phi_3(3) = 25$. Finally, $\widetilde{eBC(1)} = \widetilde{eBC(5)} = 0$ because $I(1)$ and $I(5)$ are empty sets.

4.2 Number of Queries and Time Complexity

The proposed algorithm does not perform additional queries on obtaining unbiased estimators of the ego betweenness centrality of sampled nodes because we utilize the neighbor data of the previous and next sampled nodes, $v_{x_{s-1}}$ and $v_{x_{s+1}}$, to calculate a variable, $\phi_s(i)$, of each sampled node, v_{x_s} (see the definition of $\phi_s(i)$). Therefore, the proposed algorithm performs the same number of queries as existing algorithms [26], [27].

The running time of the proposed algorithm with r samples is $O(rd_{max})$, because the running time of computing $\phi_s(x_s)$ for each sample v_{x_s}, at lines 7 and 8 in Algorithm 1, is $O(d_{max})$. The proposed algorithm computes considerably faster than Maiya et al.'s algorithm which requires $O(n'm')$ computation time, where n' and m' are the number of nodes and edges in the induced subgraph, respectively. The proposed algorithm is slower than Lim et al.'s algorithm which requires $O(r)$ computation time; however,

Algorithm 1 Unbiased estimation of the ego betweenness centrality of sampled nodes via a random walk.

Input: A sequence of the sets of an index and neighbors sampled by a random walk with r steps, denoted by $\{(x_s, N(x_s))\}_{s=1}^{r}$.

Output: A set of unbiased estimators of the ego betweenness centrality of sampled nodes

// initialization

for $s = 1$ to r **do**
  $r(x_s) \leftarrow 0$
  $\widehat{eBC(x_s)} \leftarrow 0$
end for

// unbiased estimation of ego betweenness centrality

for $s = 2$ to $r - 1$ **do**
  $r(x_s) \leftarrow r(x_s) + 1$
  if $v_{x_{s+1}} \in N_{x_{s-1}}(x_s)$ **then**
    $\widehat{eBC(x_s)} \leftarrow \widehat{eBC(x_s)} + \frac{|N(x_s)|^2}{|N(x_{s-1}) \cap N(x_{s+1})|}$
  end if
end for
for $s = 1$ to r **do**
  if $r(x_s) > 0$ **then**
    $\widehat{eBC(x_s)} \leftarrow \frac{\widehat{eBC(x_s)}}{r(x_s)}$
  else
    $\widehat{eBC(x_s)} \leftarrow 0$
  end if
end for
return $\{(v_{x_s}, \widehat{eBC(x_s)})$, for each sampled node $v_{x_s}\}$

the difference is subtle with the small sample size.

5. Experiments

We evaluate the proposed algorithm using real social network datasets from the following two viewpoints:

(1) Estimation accuracy: We show that the proposed algorithm improves the estimation accuracy of the top-k nodes with the highest betweenness centrality with the small sample size.

(2) Computation time: We show that the proposed algorithm performs considerably fast with the small sample size.

5.1 Experimental Setup

We use publicly available datasets[*1] of Epinions, Buzznet, Gowalla, Academia, Dogster, and Flickr. For these six datasets, we focus on undirected, simple, and connected graphs by performing the following preprocessing: (1) we remove the directions of edges if the graphs are directed; (2) we treat multiple edges as a single edge and delete the loops; and then (3) we delete the nodes that are not contained in the largest connected component of the original graphs. **Table 1** lists the number of nodes and edges of six datasets that were used in our experiments. We conducted experiments on a Linux server with an Intel Xeon E5-2698 (2.20 GHz) processor and 503 GB of main memory. All algorithms were implemented in C++.

We performed independently the following simulations 100 times for various values of sample size, n', and k:

(1) We sample n' nodes, not including duplicates, from a target graph by a random walk. We select randomly a seed of a random walk from nodes on a graph.

(2) We estimate top-k nodes from the same samples by using each algorithm, Maiya et al.'s [27], Lim et al.'s [26], and the proposed algorithm.

We use the overlap coefficient [36] to evaluate the accuracy of the estimated top nodes obtained by the algorithms. The overlap coefficient between two sets is defined as the size of the intersection divided by the size of the smaller set and measures the similarity between two finite sets. The overlap coefficient is from 0.0 to 1.0 and a higher value means that two sets are more similar.

Table 1 Datasets.

Network	$\|V\|$	$\|E\|$
Epinions [24]	75,877	405,739
Buzznet [35]	101,163	2,763,066
Gowalla [24]	196,591	950,327
Academia [35]	200,167	1,022,440
Dogster [24]	426,485	8,543,321
Flickr [35]	513,969	3,190,452

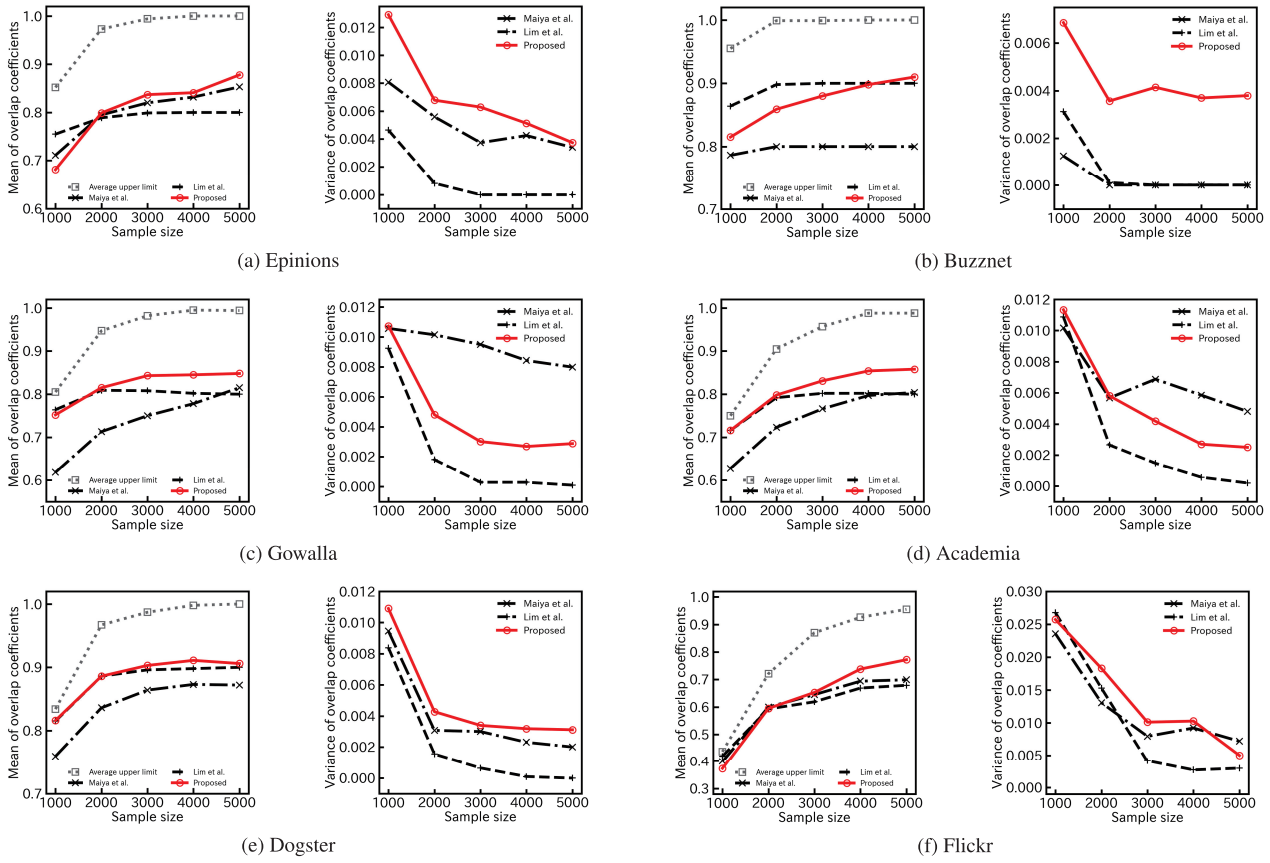
5.2 Estimation Accuracy

Figure 2 shows the mean and variance of overlap coefficients of each algorithm between two sets of the exact and estimated top-10 nodes with the highest betweenness centrality when the sample size, n', is changed from 1,000 to 5,000 in increments of 1,000. The upper limit of the mean is calculated by the overlap coefficient between a set of exact top-10 nodes with the highest betweenness centrality and a set of all the sampled nodes. We note that the upper limits of the three methods are equal because the top-k nodes by each method are estimated from the same sample. We show the average upper limit of the mean over 100 runs on each dataset. **Table 2** shows the mean and variance of overlap coefficients of each algorithm between two sets of exact and estimated top-k nodes with the highest betweenness centrality for various values of k when the sample size is 1,000 and 5,000, respectively.

First, we observe that many of the top-k nodes with the highest betweenness centrality are collected via random walk-based sampling with the small sample size (see Fig. 2). For example, Fig. 2 (e) shows that 80% of top-10 nodes on average are contained in only 1,000 sampled on Dogster. We see that almost all the top-10 nodes with the highest betweenness centrality are collected with 5,000 samples on all the datasets. The previous study [26] similarly observed that a random walk can collect many of the top-k nodes with the highest betweenness centrality with the small sample size.

The proposed algorithm has a higher mean of overlap coefficients than that of Lim et al.'s algorithm with 5,000 samples for various values of k (see Fig. 2 and Table 2). The improvement of the estimation accuracy results from the following two facts. The first fact is that more of the top-k nodes with the highest ego betweenness centrality are the top-k nodes with the highest betweenness centrality than the top-k nodes with the highest degree in real social networks. **Table 3** shows the overlap coefficients between a set of exact top-k nodes with the highest betweenness

[*1] The Epinions, Buzznet, Gowalla, Academia, Dogster, and Flickr are publicly available at
http://konect.uni-koblenz.de/networks/soc-Epinions1,
http://networkrepository.com/soc-buzznet.php,
http://konect.uni-koblenz.de/networks/loc-gowalla_edges,
http://networkrepository.com/soc-academia.php,
http://konect.uni-koblenz.de/networks/petster-friendships-dog,
http://networkrepository.com/soc-flickr.php.



(a) Epinions (b) Buzznet

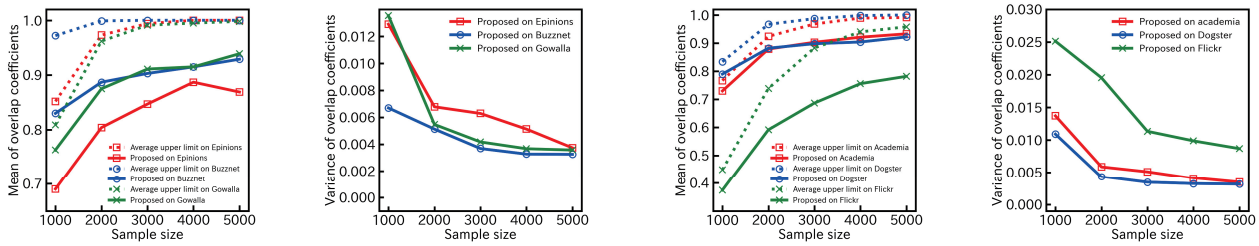
(c) Gowalla (d) Academia

(e) Dogster (f) Flickr

Fig. 2 The mean and variance of overlap coefficients of each algorithm between two sets of exact and estimated top-10 nodes with the highest betweenness centrality for various sample sizes.

Table 2 The mean (variance) of overlap coefficients of each algorithm between two sets of the exact and estimated top-k nodes with the highest betweenness centrality for various values of k when the sample size is 1,000 and 5,000, respectively. The highest mean is shown in bold.

		1,000 sample			5,000 sample		
Dataset	k	Maiya et al. [27]	Lim et al. [26]	Proposed	Maiya et al. [27]	Lim et al. [26]	Proposed
Epinions	$k = 10$	0.706 (0.009)	**0.754** (0.004)	0.662 (0.023)	0.840 (0.003)	0.800 (0.000)	**0.879** (0.004)
	$k = 20$	0.626 (0.006)	**0.677** (0.003)	0.627 (0.007)	0.795 (0.001)	0.750 (0.000)	**0.825** (0.002)
	$k = 30$	0.590 (0.004)	**0.653** (0.005)	0.583 (0.006)	0.768 (0.001)	0.767 (0.000)	**0.814** (0.002)
	$k = 40$	0.582 (0.003)	**0.621** (0.003)	0.556 (0.004)	0.791 (0.001)	0.775 (0.000)	**0.792** (0.001)
	$k = 50$	0.551 (0.003)	**0.594** (0.003)	0.527 (0.003)	**0.784** (0.001)	0.740 (0.000)	0.763 (0.001)
Buzznet	$k = 10$	0.789 (0.001)	**0.867** (0.003)	0.801 (0.007)	0.800 (0.000)	0.900 (0.000)	**0.905** (0.003)
	$k = 20$	0.740 (0.002)	**0.880** (0.001)	0.806 (0.003)	0.788 (0.001)	0.900 (0.000)	**0.923** (0.001)
	$k = 30$	0.750 (0.001)	**0.800** (0.001)	0.765 (0.002)	0.737 (0.000)	0.800 (0.000)	**0.853** (0.001)
	$k = 40$	0.641 (0.001)	0.689 (0.001)	**0.708** (0.002)	0.655 (0.000)	0.725 (0.000)	**0.784** (0.001)
	$k = 50$	0.654 (0.001)	**0.722** (0.001)	0.700 (0.001)	0.719 (0.000)	0.740 (0.000)	**0.795** (0.001)
Gowalla	$k = 10$	0.582 (0.008)	**0.731** (0.009)	0.710 (0.009)	0.809 (0.008)	0.801 (0.000)	**0.850** (0.003)
	$k = 20$	0.476 (0.008)	**0.615** (0.008)	0.578 (0.007)	0.723 (0.004)	**0.795** (0.000)	0.773 (0.001)
	$k = 30$	0.452 (0.008)	**0.563** (0.007)	0.540 (0.006)	0.757 (0.002)	**0.794** (0.000)	0.777 (0.001)
	$k = 40$	0.414 (0.007)	**0.516** (0.005)	0.494 (0.005)	0.721 (0.002)	0.736 (0.000)	**0.754** (0.001)
	$k = 50$	0.392 (0.004)	**0.492** (0.006)	0.470 (0.005)	0.717 (0.002)	0.721 (0.000)	**0.740** (0.001)
Academia	$k = 10$	0.607 (0.010)	**0.700** (0.014)	0.692 (0.015)	0.795 (0.005)	0.801 (0.000)	**0.847** (0.002)
	$k = 20$	0.495 (0.009)	**0.575** (0.011)	0.573 (0.011)	0.842 (0.003)	**0.856** (0.000)	0.851 (0.002)
	$k = 30$	0.411 (0.006)	**0.484** (0.008)	**0.484** (0.008)	0.748 (0.002)	0.814 (0.001)	**0.830** (0.001)
	$k = 40$	0.342 (0.004)	**0.414** (0.005)	0.413 (0.005)	0.647 (0.002)	**0.775** (0.001)	0.772 (0.001)
	$k = 50$	0.320 (0.003)	**0.385** (0.005)	0.383 (0.005)	0.653 (0.002)	0.737 (0.001)	**0.740** (0.001)
Dogster	$k = 10$	0.762 (0.011)	0.799 (0.010)	**0.800** (0.013)	0.869 (0.002)	0.900 (0.000)	**0.907** (0.004)
	$k = 20$	0.668 (0.008)	**0.689** (0.007)	0.675 (0.007)	**0.858** (0.002)	0.752 (0.000)	0.848 (0.002)
	$k = 30$	0.591 (0.005)	**0.631** (0.005)	0.607 (0.004)	**0.850** (0.001)	0.793 (0.000)	0.814 (0.001)
	$k = 40$	0.544 (0.004)	0.575 (0.005)	**0.576** (0.007)	0.807 (0.000)	0.781 (0.000)	**0.814** (0.001)
	$k = 50$	0.547 (0.004)	**0.563** (0.003)	0.558 (0.004)	0.806 (0.002)	0.844 (0.000)	**0.877** (0.001)
Flickr	$k = 10$	0.442 (0.024)	**0.451** (0.023)	0.408 (0.020)	0.692 (0.005)	0.679 (0.002)	**0.751** (0.008)
	$k = 20$	**0.323** (0.010)	0.297 (0.011)	0.316 (0.009)	0.583 (0.003)	0.503 (0.001)	**0.598** (0.003)
	$k = 30$	**0.288** (0.007)	0.268 (0.007)	0.285 (0.007)	0.591 (0.002)	0.414 (0.001)	**0.602** (0.003)
	$k = 40$	**0.271** (0.005)	0.257 (0.005)	0.269 (0.005)	**0.588** (0.002)	0.351 (0.001)	0.561 (0.003)
	$k = 50$	0.262 (0.004)	0.253 (0.004)	**0.265** (0.004)	0.558 (0.002)	0.334 (0.000)	**0.564** (0.002)



(a) Epinions, Buzznet, and Gowalla (b) Academia, Dogster, and Flickr

Fig. 3 The mean and variance of overlap coefficients of the proposed algorithm between two sets of the exact and estimated top-10 nodes with the highest ego betweenness centrality for various sample sizes.

Table 3 The overlap coefficients between a set of the exact top-k nodes with the betweenness centrality and each set of the exact top-k nodes with the degree (Degree) and ego betweenness centrality (EBC) for various values of k. The highest value is shown in bold.

Dataset	k	Degree	EBC
Epinions	$k = 10$	0.800	**1.000**
	$k = 20$	0.750	**0.900**
	$k = 30$	0.767	**0.900**
	$k = 40$	0.775	**0.850**
	$k = 50$	0.740	**0.840**
Buzznet	$k = 10$	**0.900**	**0.900**
	$k = 20$	0.900	**0.950**
	$k = 30$	0.800	**0.867**
	$k = 40$	0.725	**0.800**
	$k = 50$	0.740	**0.820**
Gowalla	$k = 10$	0.800	**0.900**
	$k = 20$	**0.800**	**0.800**
	$k = 30$	**0.800**	**0.800**
	$k = 40$	0.725	**0.775**
	$k = 50$	0.740	**0.780**
Academia	$k = 10$	**0.800**	**0.800**
	$k = 20$	**0.850**	**0.850**
	$k = 30$	0.800	**0.833**
	$k = 40$	0.825	**0.850**
	$k = 50$	0.780	**0.820**
Dogster	$k = 10$	0.900	**1.000**
	$k = 20$	0.750	**0.900**
	$k = 30$	0.800	**0.900**
	$k = 40$	0.775	**0.850**
	$k = 50$	0.860	**0.920**
Flickr	$k = 10$	0.700	**0.900**
	$k = 20$	0.500	**0.700**
	$k = 30$	0.433	**0.767**
	$k = 40$	0.375	**0.775**
	$k = 50$	0.320	**0.740**

Table 4 The mean (variance) of the overlap coefficients of the proposed algorithm between two sets of the exact and estimated top-k nodes with the highest ego betweenness centrality for various values of k when the sample size is 1,000 and 5,000, respectively.

		1,000 sample	5,000 sample
Dataset	k	Mean (Variance)	Mean (Variance)
Epinions	$k = 10$	0.662 (0.018)	0.879 (0.004)
	$k = 20$	0.649 (0.007)	0.881 (0.002)
	$k = 30$	0.600 (0.006)	0.872 (0.002)
	$k = 40$	0.579 (0.005)	0.860 (0.001)
	$k = 50$	0.563 (0.004)	0.858 (0.001)
Buzznet	$k = 10$	0.807 (0.007)	0.937 (0.003)
	$k = 20$	0.833 (0.003)	0.959 (0.001)
	$k = 30$	0.804 (0.002)	0.915 (0.001)
	$k = 40$	0.783 (0.002)	0.899 (0.001)
	$k = 50$	0.783 (0.001)	0.912 (0.001)
Gowalla	$k = 10$	0.747 (0.012)	0.930 (0.003)
	$k = 20$	0.642 (0.009)	0.922 (0.002)
	$k = 30$	0.580 (0.008)	0.879 (0.002)
	$k = 40$	0.532 (0.006)	0.861 (0.001)
	$k = 50$	0.497 (0.006)	0.838 (0.001)
Academia	$k = 10$	0.713 (0.014)	0.934 (0.003)
	$k = 20$	0.595 (0.011)	0.928 (0.002)
	$k = 30$	0.509 (0.007)	0.919 (0.001)
	$k = 40$	0.444 (0.007)	0.889 (0.002)
	$k = 50$	0.399 (0.005)	0.840 (0.002)
Dogster	$k = 10$	0.800 (0.013)	0.907 (0.004)
	$k = 20$	0.680 (0.009)	0.883 (0.002)
	$k = 30$	0.612 (0.005)	0.849 (0.001)
	$k = 40$	0.593 (0.004)	0.851 (0.001)
	$k = 50$	0.558 (0.004)	0.888 (0.001)
Flickr	$k = 10$	0.410 (0.022)	0.771 (0.009)
	$k = 20$	0.354 (0.012)	0.744 (0.005)
	$k = 30$	0.312 (0.007)	0.684 (0.004)
	$k = 40$	0.289 (0.006)	0.648 (0.004)
	$k = 50$	0.278 (0.004)	0.620 (0.003)

centrality and each set of exact top-k nodes with the highest degree and ego betweenness centrality. We see that more of the top-k nodes with the highest ego betweenness centrality are included in a set of the top-k nodes with the highest betweenness centrality than the top-k nodes with the highest degree for various values of k on all the datasets. The second fact is that the proposed algorithm accurately estimates the top-k nodes with the highest ego betweenness centrality with the small sample size. **Figure 3** and **Table 4** show that the mean and variance of overlap coefficients of the proposed algorithm between two sets of exact and estimated top-k nodes with the highest ego betweenness centrality when the sample sizes and the values of k are changed, respectively. The proposed algorithm achieves the high mean of overlap coefficients with the small sample size and various values of k.

The proposed algorithm has a lower mean of overlap coefficients with 1,000 sample size and a higher variance with most sample sizes than those of Lim et al.'s algorithm (see Fig. 2 and Table 2). The main reason for this result is that Lim et al.'s algorithm obtains the exact degree of each sample while the proposed algorithm obtains the estimates of the ego betweenness centrality. The proposed algorithm causes errors of the rank order the estimated top nodes because of the relatively large errors of the estimates of the ego betweenness centrality of each sample when the sample size is considerably small.

Finally, we observe that Maiya et al.'s algorithm has a lower mean of overlap coefficients than that of other algorithms in all datasets (see Fig. 2 and Table 2). This is because the betweenness centrality of nodes in the subgraph induced from small samples has typically large errors between the betweenness centrality in the original graph [5], [12].

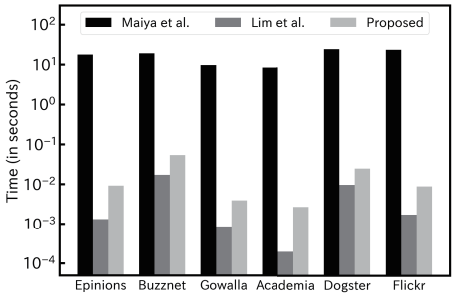
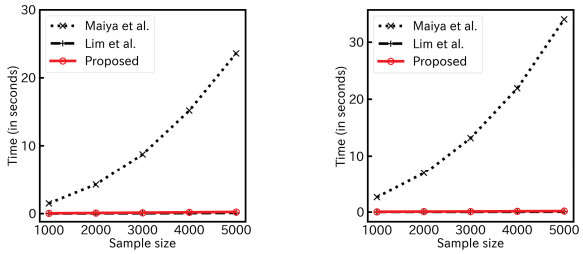


Fig. 4 Computation time of each algorithm when the sample size is 5,000.



(a) Dogster (b) Flickr

Fig. 5 Computation time of each algorithm for various sample sizes.

5.3 Computation Time

Figure 4 shows the running time to estimate the top-k nodes from samples in each algorithm when the sample size is 5,000. **Figure 5** shows the computation time of each algorithm for various sample sizes on (a) Dogster and (b) Flickr. The proposed algorithm and Lim et al.'s algorithm are faster than Maiya et al.'s algorithm for all the sample sizes. Although the computation time of Maiya et al.'s algorithm can be reduced by utilizing some algorithms [3], [10], [34] for approximating the betweenness centrality, the estimation accuracy of the top-k nodes with the highest betweenness centrality should fall because of the approximation. The proposed algorithm is slower than the Lim et al.'s algorithm; however, the difference is subtle with small sample size.

6. Conclusion

We have proposed a random walk-based algorithm to estimate the top-k nodes with the highest betweenness centrality in social networks. The proposed algorithm obtains unbiased estimators of the ego betweenness centrality of sampled nodes via a random walk and approximates the top-k nodes with the highest betweenness centrality as the top-k nodes with the highest estimated ego betweenness centrality. The experimental results show that the proposed algorithm improves the estimation accuracy of top-k nodes with the highest betweenness centrality in real social network datasets.

Acknowledgments This work was supported by New Energy and Industrial Technology Development Organization (NEDO).

References

[1] Ahn, Y.-Y., Han, S., Kwak, H., Moon, S. and Jeong, H.: Analysis of topological characteristics of huge online social networking services, *WWW*, pp.835–844 (2007).
[2] Avrachenkov, K., Litvak, N., Sokol, M. and Towsley, D.: Quick detection of nodes with large degrees, *Internet Mathematics*, Vol.10, No.1-2, pp.1–19 (2014).
[3] Bader, D.A., Kintali, S., Madduri, K. and Mihail, M.: Approximating betweenness centrality, *International Workshop on Algorithms and Models for the Web-Graph*, pp.124–137 (2007).
[4] Barthelemy, M.: Betweenness centrality in large complex networks, *The European Physical Journal B-Condensed Matter and Complex Systems*, Vol.38, No.2, pp.163–168 (2004).
[5] Borgatti, S.P., Carley, K.M. and Krackhardt, D.: On the robustness of centrality measures under conditions of imperfect data, *Social Networks*, Vol.28, No.2, pp.124–136 (2006).
[6] Borgatti, S.P. and Everett, M.G.: A graph-theoretic perspective on centrality, *Social Networks*, Vol.28, No.4, pp.466–484 (2006).
[7] Boyd, D.M. and Ellison, N.B.: Social network sites: Definition, history, and scholarship, *Journal of Computer-mediated Communication*, Vol.13, No.1, pp.210–230 (2007).
[8] Brandes, U.: A faster algorithm for betweenness centrality, *Journal of Mathematical Sociology*, Vol.25, No.2, pp.163–177 (2001).
[9] Brandes, U.: On variants of shortest-path betweenness centrality and their generic computation, *Social Networks*, Vol.30, No.2, pp.136–145 (2008).
[10] Brandes, U. and Pich, C.: Centrality estimation in large networks, *International Journal of Bifurcation and Chaos*, Vol.17, No.7, pp.2303–2318 (2007).
[11] Chen, X., Li, Y., Wang, P. and Lui, J.: A general framework for estimating graphlet statistics via random walk, *PVLDB*, Vol.10, No.3, pp.253–264 (2016).
[12] Costenbader, E. and Valente, T.W.: The stability of centrality measures when networks are sampled, *Social Networks*, Vol.25, No.4, pp.283–307 (2003).
[13] Dasgupta, A., Kumar, R. and Sarlos, T.: On estimating the average degree, *WWW*, pp.795–806 (2014).
[14] Everett, M. and Borgatti, S.P.: Ego network betweenness, *Social Networks*, Vol.27, No.1, pp.31–38 (2005).
[15] Facebook: Facebook Reports Fourth Quarter and Full Year 2019 Results (2019), available from ⟨https://s21.q4cdn.com/399680738/files/doc_financials/2019/q4/FB-12.31.2019-Exhibit-99.1-r61_final.pdf⟩.
[16] Freeman, L.C.: A set of measures of centrality based on betweenness, *Sociometry*, Vol.40, No.1, pp.35–41 (1977).
[17] Gjoka, M., Kurant, M., Butts, C.T. and Markopoulou, A.: Practical recommendations on crawling online social networks, *IEEE Journal on Selected Areas in Communications*, Vol.29, No.9, pp.1872–1892 (2011).
[18] Goh, K.-I., Oh, E., Kahng, B. and Kim, D.: Betweenness centrality correlation in social networks, *Physical Review E*, Vol.67, No.1, p.017101 (2003).
[19] Guimera, R. and Amaral, L.A.N.: Modeling the world-wide airport network, *The European Physical Journal B-Condensed Matter and Complex Systems*, Vol.38, No.2, pp.381–385 (2004).
[20] Joy, M.P., Brock, A., Ingber, D.E. and Huang, S.: High-betweenness proteins in the yeast protein interaction network, *BioMed Research International*, Vol.2005, No.2, pp.96–103 (2005).
[21] Katzir, L. and Hardiman, S.J.: Estimating clustering coefficients and size of social networks via random walk, *ACM Trans. Web*, Vol.9, No.4, p.19 (2015).
[22] Katzir, L., Liberty, E., Somekh, O. and Cosma, I.A.: Estimating sizes of social networks via biased sampling, *Internet Mathematics*, Vol.10, No.3-4, pp.335–359 (2014).
[23] Kourtellis, N., Alahakoon, T., Simha, R., Iamnitchi, A. and Tripathi, R.: Identifying high betweenness centrality nodes in large social networks, *Social Network Analysis and Mining*, Vol.3, No.4, pp.899–914 (2013).
[24] Kunegis, J.: KONECT — The Koblenz Network Collection, *WWW*, pp.1343–1350 (2013).
[25] Kwak, H., Lee, C., Park, H. and Moon, S.: What is Twitter, a social network or a news media?, *WWW*, pp.591–600 (2010).
[26] Lim, Y.S., Menasché, D.S., Ribeiro, B., Towsley, D. and Basu, P.: Online estimating the k central nodes of a network, *Network Science Workshop*, pp.118–122 (2011).
[27] Maiya, A.S. and Berger-Wolf, T.Y.: Online sampling of high centrality individuals in social networks, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp.91–98 (2010).
[28] Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P. and Bhattacharjee, B.: Measurement and analysis of online social networks, *IMC*, pp.29–42 (2007).
[29] Nakajima, K., Iwasaki, K., Matsumura, T. and Shudo, K.: Estimating top-k betweenness centrality nodes in online social networks, *SocialCom*, pp.1128–1135 (2018).
[30] Newman, M.E.: A measure of betweenness centrality based on random walks, *Social Networks*, Vol.27, No.1, pp.39–54 (2005).
[31] Newman, M.E. and Girvan, M.: Finding and evaluating community structure in networks, *Physical Review E*, Vol.69, No.2, pp.26–113 (2004).

[32] Pfeffer, J. and Carley, K.M.: k-centralities: Local approximations of global measures based on shortest paths, *WWW*, pp.1043–1050 (2012).
[33] Ribeiro, B. and Towsley, D.: Estimating and sampling graphs with multidimensional random walks, *IMC*, pp.390–403, ACM (2010).
[34] Riondato, M. and Kornaropoulos, E.M.: Fast approximation of betweenness centrality through sampling, *Data Mining and Knowledge Discovery*, Vol.30, No.2, pp.438–475 (2016).
[35] Rossi, R.A. and Ahmed, N.K.: The Network Data Repository with Interactive Graph Analytics and Visualization, *AAAI* (2015).
[36] Simpson, G.G.: Mammals and the nature of continents, *American Journal of Science*, Vol.241, No.1, pp.1–31 (1943).
[37] Wang, P., Lui, J., Ribeiro, B., Towsley, D., Zhao, J. and Guan, X.: Efficiently estimating motif statistics of large networks, *ACM Trans. Knowledge Discovery from Data*, Vol.9, No.2, p.8 (2014).

Kazuki Nakajima is currently a Ph.D. student at Tokyo Institute of Technology. His research interest is graph sampling algorithms for social networks.

Kazuyuki Shudo received B.E. in 1996, M.E. in 1998, and a Ph.D. degrees in 2001 all in computer science from Waseda University. He worked as a Research Associate at the same university from 1998 to 2001. He later served as a Research Scientist at National Institute of Advanced Industrial Science and Technology. In 2006, he joined Utagoe Inc. as a Director, Chief Technology Officer. Since December 2008, he currently serves as an Associate Professor at Tokyo Institute of Technology. His research interests include distributed computing, programming language systems and information security. He has received the best paper award at SACSIS 2006, Information Processing Society Japan (IPSJ) best paper award in 2006, the Super Creator certification by Japanese Ministry of Economy Trade and Industry (METI) and Information Technology Promotion Agency (IPA) in 2007, IPSJ Yamashita SIG Research Award in 2008, Funai Prize for Science in 2010, The Young Scientists' Prize, The Commendation for Science and Technology by the Minister of Education, Culture, Sports, and Technology in 2012, and IPSJ Nagao Special Researcher Award in 2013. He is a member of IEEE, IEEE Computer Society, IEEE Communications Society and ACM.

(Editor in Charge: *Yasunori Ishihara*)