Resampling-based Framework for Estimating Node Centrality of Large Social Network

Kouzou Ohara[1], Kazumi Saito[2], Masahiro Kimura[3], and Hiroshi Motoda[4]

[1] Department of Integrated Information Technology, Aoyama Gakuin University
ohara@it.aoyama.ac.jp
[2] School of Administration and Informatics, University of Shizuoka
k-saito@u-shizuoka-ken.ac.jp
[3] Department of Electronics and Informatics, Ryukoku University
kimura@rins.ryukoku.ac.jp
[4] Institute of Scientific and Industrial Research, Osaka University
School of Computing and Information Systems, University of Tasmania
motoda@ar.sanken.osaka-u.ac.jp

Abstract. We address a problem of efficiently estimating value of a centrality measure for a node in a large social network only using a partial network generated by sampling nodes from the entire network. To this end, we propose a resampling-based framework to estimate the approximation error defined as the difference between the true and the estimated values of the centrality. We experimentally evaluate the fundamental performance of the proposed framework using the closeness and betweenness centralities on three real world networks, and show that it allows us to estimate the approximation error more tightly and more precisely with the confidence level of 95% even for a small partial network compared with the standard error traditionally used, and that we could potentially identify top nodes and possibly rank them in a given centrality measure with high confidence level only from a small partial network.

Keywords: Error estimation, resampling, node centrality, social network analysis

1 Introduction

Recently, Social Media such as Facebook, Digg, Twitter, Weblog, Wiki, etc. becomes increasingly popular on a worldwide scale, and allows us to construct large-scale social networks in cyberspace. An article that is posted on social media can rapidly and widely spread through such networks and can be shared by a large number of people. Since such information can substantially affect our thought and decision making, a large number of studies have been made by researchers in many different disciplines such as sociology, psychology, economy, and computer science [8, 4] to analyze various aspects of social networks and information diffusion on them.

In the domain of social network analysis, several measures called centrality have been proposed so far [7, 5, 1, 3, 13]. They characterize nodes in a network based on its structure, and give an insight into network performance. For example, a centrality provides us with the information of how important each node is through node ranking derived directly from the centrality. It also provides us with topological features of a network. For example, scale free property is derived from the degree distribution. As a social network in World Wide Web easily grows in size, it is becoming pressingly important that we are able to efficiently compute values of a centrality to analyze such a large social network. However, if a centrality measure is based not only on local structure around a target node, e.g. its neighboring nodes, but also on global structure of a network, e.g. paths between arbitrary node pairs, its computation becomes harder as the size of the network increases. Thus, it is crucial to reduce the computational cost of such centralities for large social networks. Typical examples are the closeness and the betweenness centralities which we consider in this paper (explained later).

It is worth noting that such a centrality is usually defined as a summarized value of more primitive ones that are derived from node pairs in a network. For example, the closeness centrality is defined as the average of the shortest path lengths from a target node to each of the remaining nodes in a network. Considering this fact, it is inevitable to employ a sampling-based approach as a possible solution of this kind of problem on scalability. It is obvious that using only a limited number of nodes randomly sampled from a large social network can reduce the computational cost. However, the resulting value is an approximation of its true value, and thus it becomes important to accurately estimate the approximation error. It is well known from the statistical view point that the margin of error (difference between sample mean and population mean) is $\pm 2 \times \sigma / \sqrt{N}$ with the confidence level of 95%, where σ and N are the standard deviation of a population and the number of samples, respectively. However, this traditional boundary does not necessarily give us a tight approximation error.

In this paper, we propose a framework that provides us with a tighter error estimate of how close the approximation is to the true value. The basic idea is that we consider all possible partial networks of a fixed size that are generated by resampling nodes according to a given coverage ratio, and then estimate the approximation error, referred to as *resampling error*, using centrality values derived from those partial networks. We test our framework using two well-known centrality measures, the closeness and the betweenness centralities, both of which require to use the global structure of a network for computing the value of each node. Extensive experiments were performed on three real world social networks varying the sample coverage for each centrality measure. We empirically confirmed that the proposed framework is more promising than the traditional error bound in that it enables us to give a tighter approximation error with a higher confidence level than the traditional one under a given sampling ratio. The framework we proposed is not specific to computation of node centralities for social network analysis. It is very generic and is applicable to any other estimation problems that require aggregation of many (but a finite number of) primitive computations.

The paper is organized as follows. Section 2 gives the formal definitions of both the resampling-based framework that we propose and the traditional bound of approximation error. Section 3 explains the closeness and the betweenness centralities we used to evaluate our framework and presents how to estimate their approximation error. Section 4 reports experimental results for these centralities on three real world networks. Section 5 concludes this paper and addresses the future work.

2 Related work

As mentioned above, it is crucial to employ a sampling-based approach when analyzing a large social network. Many kinds of sampling methods have been investigated and proposed so far [6, 11, 10]. Non-uniform sampling techniques give higher probabilities to be selected to specific nodes such as high-degree ones. Similarly, results by traversal/walk-based sampling are biased towards high-degree nodes. In our problem setting the goal is to accurately estimate centralities of an original network and thus uniform sampling that selects nodes of a given network uniformly at random is essential because biased samples might skew centrality values derived from a resulting network.

This motivates us to propose the framework that ensures the accuracy of the approximations of centrality values under uniform sampling. Although we use a simple random sampling here, our framework can adopt a more sophisticated technique such as MH-sampling [6] in so far as it falls under uniform sampling. In this sense, our framework can be regarded as a meta-level method that is applicable to any uniform sampling technique.

3 Resampling-based estimation framework

For a given set of objects S whose number of elements is $L = |S|$, and a function f which calculates some associated value of each object, we first consider a general problem of estimating the average value μ of the set of entire values $\{f(s) \,|\, s \in S\}$ only from its arbitrary subset of partial values $\{f(t) \,|\, t \in T \subset S\}$. For a subset T whose number of elements is $N = |T|$, we denote its partial average value by $\mu(T) = (1/N)\sum_{t\in T} f(t)$. Below, we formally derive an expected estimation error $RE(N)$ which is the difference between the average value μ and the partial average value $\mu(T)$, with respect to the number of elements N. Hereafter, the estimated error based on $RE(N)$ is referred to as resampling error.

Now, let $\mathcal{T} \subset 2^S$ be a family of subsets of S whose number of elements is N, that is, $|T| = N$ for $T \in \mathcal{T}$. Then, we obtain the following estimation formula for the expected error:

$$
\begin{aligned}
RE(N) &= \sqrt{\langle(\mu - \mu(T))^2\rangle_{T\in\mathcal{T}}} \\
&= \sqrt{\binom{L}{N}^{-1} \sum_{T\in\mathcal{T}} \left(\mu - \frac{1}{N}\sum_{t\in T} f(t)\right)^2} = \sqrt{\binom{L}{N}^{-1} \frac{1}{N^2} \sum_{T\in\mathcal{T}} \left(\sum_{t\in T}(f(t) - \mu)\right)^2} \\
&= \sqrt{\binom{L}{N}^{-1} \frac{1}{N^2} \left(\binom{L-1}{N-1}\sum_{s\in S}(f(s)-\mu)^2 + \binom{L-2}{N-2}\sum_{s\in S}\sum_{t\in S, t\neq s}(f(s)-\mu)(f(t)-\mu)\right)} \\
&= \sqrt{\binom{L}{N}^{-1} \frac{1}{N^2} \left(\left(\binom{L-1}{N-1} - \binom{L-2}{N-2}\right)\sum_{s\in S}(f(s)-\mu)^2 + \binom{L-2}{N-2}\left(\sum_{s\in S}(f(s)-\mu)\right)^2\right)} \\
&= C(N)\sigma. \qquad (1)
\end{aligned}
$$

Here the factor $C(N)$ and the standard deviation σ are given as follows:

$$C(N) = \sqrt{\frac{L-N}{(L-1)N}}, \quad \sigma = \sqrt{\frac{1}{L}\sum_{s\in S}(f(s)-\mu)^2}.$$

In this paper we consider a huge social network consisting of millions of nodes as a collection of a large number of objects, and propose a framework in which we use the partial average value as an approximate solution with an adequate confidence level using the above estimation formula, Equation (1). More specifically, we claim that for a given subset T whose number of elements is N, and its partial average value $\mu(T)$, the probability that $|\mu(T) - \mu|$ is larger than $2 \times RE(N)$, is less than 5%. This is because the estimation error of Equation (1) is regarded as the standard deviation with respect to the number of elements N. Hereafter this framework is referred to as the resampling estimation framework.

In order to confirm the effectiveness of the proposed resampling estimation framework, we also consider a standard approach based on the i.i.d. (independently identical distribution) assumption for comparison purpose. More specifically, for a given subset T whose number of elements is N, we assume that each element $t \in T$ is independently selected according to some distribution $p(t)$ such as an empirical distribution $p(t) = 1/L$. Then, by expressing elements of T as $T = \{t_1, \cdots, t_N\}$, we obtain the following estimation formula for the expected error:

$$\begin{aligned}
SE(N) &= \sqrt{\langle(\mu-\mu(T))^2\rangle} \\
&= \sqrt{\sum_{t_1\in S}\cdots\sum_{t_N\in S}\left(\mu-\frac{1}{N}\sum_{n=1}^{N}f(t_n)\right)^2\prod_{n=1}^{N}p(t_n)} = \sqrt{\frac{1}{N^2}\sum_{t_1\in S}\cdots\sum_{t_N\in S}\left(\sum_{n=1}^{N}(f(t_n)-\mu)\right)^2\prod_{n=1}^{N}p(t_n)} \\
&= \sqrt{\frac{1}{N^2}\sum_{t_1\in S}\cdots\sum_{t_N\in S}\left(\sum_{n=1}^{N}(f(t_n)-\mu)^2+\sum_{n=1}^{N}\sum_{m=1,m\neq n}^{N}(f(t_n)-\mu)(f(t_m)-\mu)\right)\prod_{n=1}^{N}p(t_n)} \\
&= D(N)\sigma,
\end{aligned} \qquad (2)$$

where $D(N) = 1/\sqrt{N}$ and σ is the standard deviation. Hereafter, the estimated error based on $SE(N)$ is referred to as standard error. The difference between Equations (1) and (2) is only their coefficients, $C(N)$ and $D(N)$. We can easily see that $C(N) \leq D(N)$, $C(L) = 0$ and $D(L) \neq 0$. For more details, we empirically compare these resampling error $RE(N)$ and standard error $SE(N)$ through experiments on node centrality calculation of social networks as described below. Note that the standard deviation σ is needed in both Equations (1) and (2). We are assuming that $|S|$ is too large to compute σ. Otherwise, sampling is not needed. We can use, instead of σ, the standard deviation σ' that is derived from a subset S' $(\subset S)$ such that $|S'| = L'$ is small enough to compute σ' within a reasonable time.

4 Application to node centrality estimation

We investigate our proposed resampling framework on node centrality estimation of a social network represented by a directed graph $G = (V, E)$, where V and E $(\subset V \times V)$

are the sets of all the nodes and the links in the network, respectively. When there is a link (u, v) from node u to node v, u is called a *parent node* of v and v is called a *child node* of u. For any node $v \in V$, let $A(u)$ and $B(v)$ respectively denote the set of all child nodes of u and the set of all parent nodes of v in G, *i.e.*, $A(u) = \{v \in V; (u, v) \in E\}$ and $B(v) = \{u \in V; (u, v) \in E\}$.

4.1 Closeness centrality estimation

The closeness $cls_G(u)$ of a node u on a graph G is defined as

$$cls_G(u) = \frac{1}{(|V| - 1)} \sum_{v \in V, v \neq u} \frac{1}{spl_G(u, v)}, \tag{3}$$

where $spl_G(u, v)$ stands for the shortest path length from u to v in G. Namely, the closeness of a node u becomes high when a large number of nodes are reachable from u within relatively short path lengths. Here note that we set $spl_G(u, v) = \infty$ when node v is not reachable from node u on G. Thus, in order to naturally cope with this infinite path length, we employ the inverse of the harmonic average as shown in Equation (3).

The burning algorithm [12] is a standard technique for computing $cls_G(u)$ of each node $u \in V$. More specifically, after initializing a node subset X_0 to $X_0 \leftarrow \{u\}$, and path length d to $d \leftarrow 0$, this algorithm repeatedly calculates a set X_{d+1} of newly reachable nodes from X_d and set $d \leftarrow d + 1$ unless X_d is empty. Here, newly reachable nodes from X_{d-1} is defined by $X_d = (\bigcup_{v \in X_{d-1}} A(v)) \setminus (\bigcup_{c<d} X_c)$. Then the shortest path length of node $v \in X_d$ from u is obtained as $spl_G(u, v) = d$. Here recall that $spl_G(u, v) = \infty$ if v is not reachable from u. Since the computational complexity of computing $cls_G(u)$ for each node $u \in V$ become $O(|E|)$, it takes a large amount of computation time for a huge social networks consisting of millions of nodes.

Now, we present a method for computing $cls_G(u)$ of each node $u \in V$ under our resampling estimation framework. The method first constructs the reverse network of $G = (V, E)$ by reversing the direction of each link from (u, v) to (v, u). Namely, the reverse network is defined by $H = (V, F)$ and $F = \{(v, u) \,|\, (u, v) \in E\}$. Then, by using the burning algorithm starting from node v over the reverse network, we can calculate each shortest path length from v to u as $spl_H(v, u)$. Clearly, $spl_H(v, u)$ is the shortest path length from node u to v, *i.e.*, $spl_G(u, v)$. Namely, for each node $u \in V$, by setting $S_u = V \setminus \{u\}$ and $f_u(v) = spl_H(v, u)$, we can calculate partial average value from an arbitrary subset $T \subset S_u \cup \{u\}$. Here note that, due to the nature of the burning algorithm, we can obtain such partial average value simultaneously for all nodes $u \in V$.

4.2 Betweenness centrality estimation

The betweenness $btw_G(u)$ of a node u on a graph G is defined as

$$btw_G(u) = \frac{1}{(|V| - 1)(|V| - 2)} \sum_{v \in V, v \neq u} \left(\sum_{w \in V, w \neq u, w \neq v} \frac{nsp_G(v, w; u)}{nsp_G(v, w)} \right), \tag{4}$$

where $nsp_G(v, w)$ is the number of the shortest paths from v to w in G and $nsp_G(v, w; u)$ is the number of the shortest paths from v to w in G that passes through node u. Namely, the betweenness of a node u becomes high when a large number of shortest paths between two nodes pass through node u. Here note that although $cls_G(u)$ and $cls_H(u)$ is not generally equal, since any node pair (v, w) is examined in Equation (4) we can easily see that $btw_G(u) = btw_H(u)$.

The Brandes algorithm [2] is a standard technique for computing $btw_G(u)$ of each node $u \in V$. The algorithm utilizes a series of node subsets $(X_0, \cdots, X_D)$ produced by the burning algorithm described in Section 4.1 starting from node $v \in V$, where D stands for the maximum burning step. Then, after setting $nsp_G(v, w) \leftarrow 1$ for $w \in X_1$, the algorithm in turn computes $nsp_G(v, w) \leftarrow \sum_{x \in B(w) \cap X_{d-1}} nsp_G(v, x)$ for $w \in X_d$ from $d = 2$ to D. Next, we define the following betweenness $btw_G(u; v)$ of node u, which restricts its starting node to v,

$$btw_G(u; v) = \sum_{w \in V, w \neq u, w \neq v} \frac{nsp_G(v, w; u)}{nsp_G(v, w)}. \tag{5}$$

Then, after setting $btw_G(u; v) \leftarrow 0$ for $u \in X_D$, the algorithm in turn computes $btw_G(u; v) \leftarrow \sum_{x \in A(u) \cap X_{d+1}} (nsp_G(v, u)/nsp_G(v, x))(1 + btw_G(x; v))$ for $u \in X_d$ from $d = D - 1$ to 2. Finally, by computing and summing $btw_G(u; v)$ by changing the starting node v, we can obtain the betweenness $btw_G(u)$ of each node $u \in V$. Again, the computational complexity of computing $btw_G(u)$ for each node $u \in V$ become $O(|E|)$.

Now, we present a method based on the Brandes algorithm for computing $btw_G(u)$ of each node $u \in V$ under our resampling estimation framework. Namely, for each node $u \in V$, by setting $S_u = V \setminus \{u\}$ and $f_u(v) = btw_G(u; v)/(|V| - 2)$, we can calculate partial average value from an arbitrary subset $T \subset S_u \cup \{u\}$. Again note that, due to the nature of the Brandes algorithm, we can obtain such partial average value simultaneously for all nodes $u \in V$.

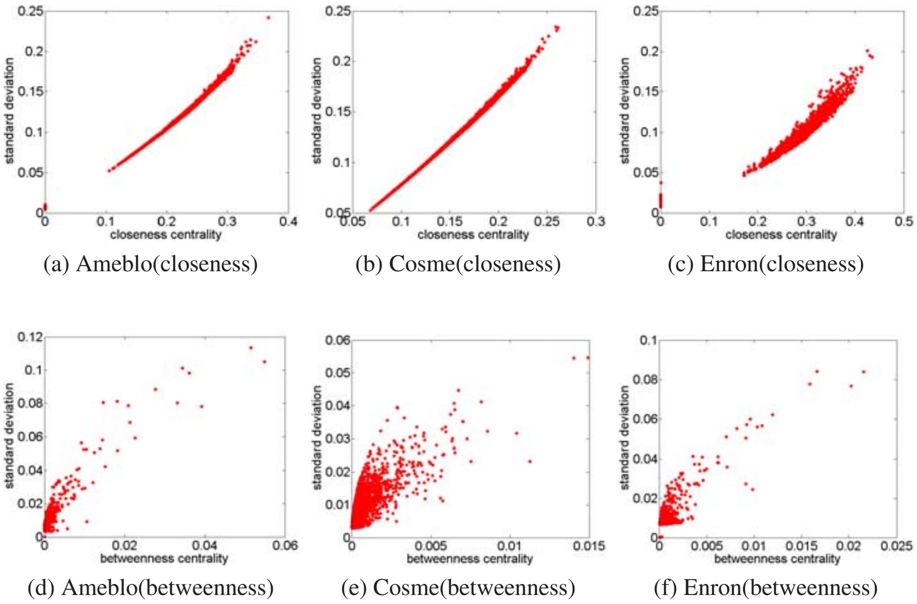
5 Experiments

5.1 Datasets

To experimentally evaluate the methods proposed in the previous sections, we employed three datasets of real networks, where all networks are represented as directed graphs. The first one is a reader network extracted from a Japanese blog service site "Ameba"[5], in which each blog can have a list of reader links. A reader link is directional and a link is constructed from blog u to blog v if blog v registers blog u as her favorite one. We crawled the lists of 117,374 blogs of "Ameba" in June 2006, and extracted a large connected network that has 56,604 nodes and 734,737 directed links. We refer to this network as the Ameblo network. The second one is a network extracted from "@cosme"[6], a Japanese word-of-mouth communication site for cosmetics, in which each user page can have *fan links*. A fan link (u, v) means that user v registers user u as her favorite user. We traced up to ten steps in the fan-link network from a randomly chosen user in December 2009, and extracted a large connected network consisting

[5] http://www.ameba.jp/

[6] http://www.cosme.net/



(a) Ameblo(closeness) (b) Cosme(closeness) (c) Enron(closeness)

(d) Ameblo(betweenness) (e) Cosme(betweenness) (f) Enron(betweenness)

Fig. 1. Results for "centrality value vs. standard deviation"

of 45, 024 nodes and 351, 299 directed links. We refer to this directed network as the Cosme network. The last one is a network derived from the Enron Email Dataset [9], in which an email address that appears in the dataset as either a sender or a recipient is regarded as a node and two email addresses u and v are linked by a directional link (u, v) if u sent an email to v. We refer to this directed network as the Enron network, which has 19, 603 nodes and 210, 950 links. These three networks are not very huge, *i.e.*, networks with millions of nodes. We dare chose them to investigate the basic performance from various angles.

5.2 Statistical Analysis

For each of the three real networks, $G = (V, E)$, we first computed the value of the closeness centrality $cls_G(u)$ and betweenness centrality $btw_G(u)$ of each node $u \in V$ by means of the algorithms presented in Sections 4.1 and 4.2, respectively. In addition, we investigated their standard deviations given by

$$\sigma_{cls}(u) = \sqrt{\frac{1}{|V|-1} \sum_{v \in V, v \neq u} \left(\frac{1}{spl_G(u, v)} - cls_G(u) \right)^2}$$

for the closeness centrality, and

$$\sigma_{btw}(u) = \sqrt{\frac{1}{|V|-1} \sum_{v \in V, v \neq u} \left(\frac{btw_G(u; v)}{|V|-2} - btw_G(u) \right)^2}$$

for the betweenness centrality. Figures 1(a) to 1(c) plot the pair $(cls_G(u), \sigma_{cls}(u))$ for the Ameblo, Cosme, and Enron networks, and Figs. 1(d) to 1(f) plot the pair $(btw_G(u), \sigma_{btw}(u))$ for the same three networks. In each figure, the horizontal and vertical axes indicate the values of corresponding centrality, $cls_G(u)$ or $btw_G(u)$, and its standard deviation, $\sigma_{cls}(u)$ or $\sigma_{btw}(u)$, respectively.

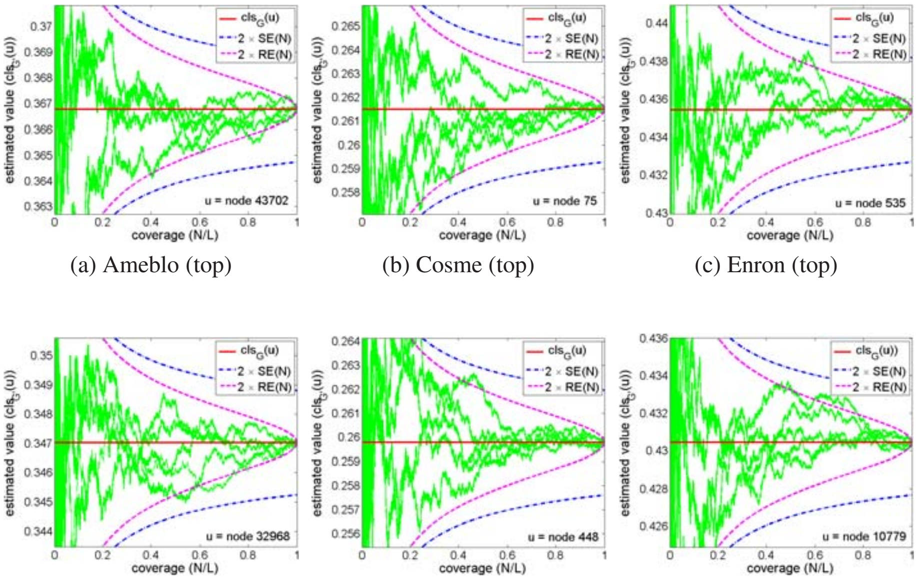
We can observe that there exists positive correlation between the centrality value of each node and its standard deviation. This tendency can be found more clearly in the results for the closeness centrality compared to the results for the betweenness centrality in which nodes are scattered over a larger area. It is noted that, for every network, higher-ranked nodes in each centrality measure are distinguishable from each other because of their distinctive values of the centrality, while it looks hard to do the same for lower-ranked nodes. This implies that there is a possibility that we can detect a cluster of such high ranked nodes or estimate their ranking with a high confidence level only using a smaller partial network if we can secure a tight approximation error.

5.3 Results

In this section, we evaluated the fundamental performance of the resampling error $RE(N)$, *i.e.*, how tightly and accurately it estimates the approximation error, using the closeness and betweenness centralities on the three networks. To this end, we considered a problem of estimating $\mu_G(u)$, the true value of a centrality measure for node u in network $G(V, E)$ using its partial network G' generated by sampling N nodes from V, and empirically investigated whether or not the estimation $\mu_{G'}(u)$, the partial average derived from G', falls within the range of $\mu_G(u) \pm 2 \times RE(N)$. Here, $\mu_G(u)$ stands for either $cls_G(u)$ or $btw_G(u)$. In addition, we considered the range of $\mu_G(u) \pm 2 \times SE(N)$ for comparison.

Figures 2 and 3 show the results for the closeness and betweenness centralities, respectively. In this experiment, we considered the top and second nodes in each network that respectively have the largest and second largest true values of the corresponding centrality in Fig. 1. In each figure, the horizontal axis “coverage” means the ratio of the number of sampled nodes N to the total number of nodes L, *i.e.*, N/L, in each network, while the vertical axis means the value of the centrality, and how the estimated value fluctuates as a function of the coverage is depicted. We conducted five independent trials for each of these two nodes in each network, and plotted estimated values $\mu_{G'}(u)$ for a given coverage N/L with green jagged lines. The red horizontal center line in each figure presents the true value of the centrality $\mu_G(u)$ for node u, while the red broken and blue chain lines show the ranges of $\mu_G(u) \pm 2 \times RE(N)$ and $\mu_G(u) \pm 2 \times SE(N)$, respectively.

From these results, we can confirm that the boundary determined by $RE(N)$ estimates the approximation error more tightly and converges to 0.0 as the coverage approaches 1.0, while the boundary by $SE(N)$ is looser and does not converge to 0.0 even if the coverage becomes 1.0. Furthermore, in most cases, the estimated value falls within the range of $\mu_G(u) \pm 2 \times RE(N)$ for every network regardless of the centrality used. From these results, we can say that the resampling error $RE(N)$ provides us with a better error bound with the confidence level of 95% compared to the standard error $SE(N)$. Besides, it is found that in Fig. 2(d) the value of the upper-bound of the range



(a) Ameblo (top) (b) Cosme (top) (c) Enron (top)

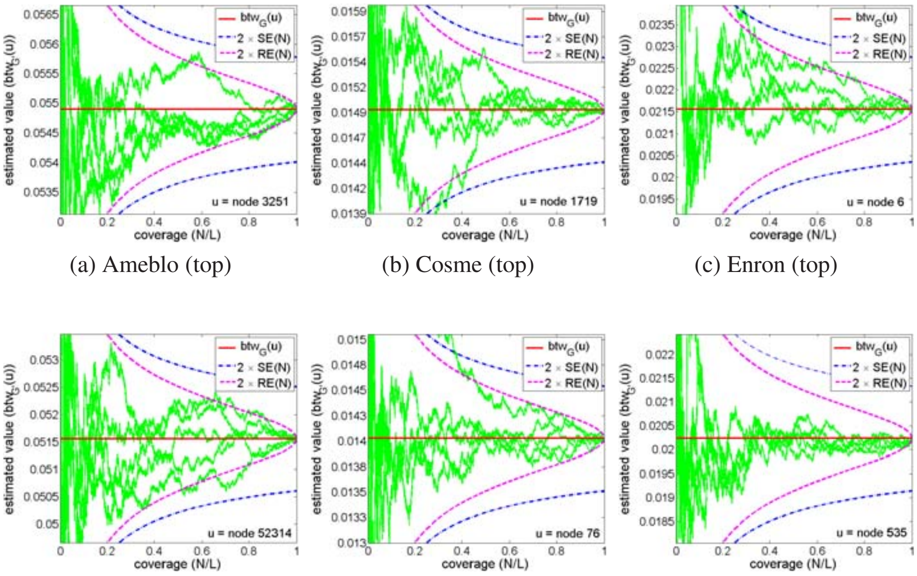
(d) Ameblo (second) (e) Cosme (second) (f) Enron (second)

Fig. 2. Fluctuation of the estimated value of the closeness centrality as a function of the coverage for the top and second nodes that respectively have the largest and second largest true values of the centrality in the Ameblo, Cosme, and Enron networks.

given by $RE(N)$ is approximately 0.3505 when the coverage is 0.2, and it is smaller than the corresponding value of the lower-bound of the range given by $RE(N)$ in Fig. 2(a), which is approximately 0.3628. These observations enable us to decide that in the Ameblo network the value of the closeness centrality of node 43702 (the top node) is higher than the value of node 32968 (the second node) with the confidence level of 95% only from the results obtained under the coverage of 0.2 because their error bounds derived from $RE(N)$ for the confidence level of 95% do not overlap each other. The same holds for the results of the Ameblo network in Fig. 3 although the coverage must be slightly larger in this case. This may not necessarily generalize to other networks, but it suggests that we could potentially detect top-K nodes and possibly their ranking in a given centrality measure with such a high confidence level even under a small coverage.

Next, we quantitatively confirmed the accuracy of the proposed resampling error in Fig. 4, in which it is shown how the difference $\delta(N)$ between the true approximation error and the estimated error fluctuates as a function of coverage in the same fashion as in Figs. 2 and 3. Here, we computed RMSE (Root Mean Squared Error) by conducting $R = 1,000$ independent trials for each value of N, which is defined as follows:

$$\mathcal{E}_{RMSE}(N) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\mu_{G',r}(u) - \mu_G(u))^2},$$



(a) Ameblo (top) (b) Cosme (top) (c) Enron (top)

(d) Ameblo (second) (e) Cosme (second) (f) Enron (second)

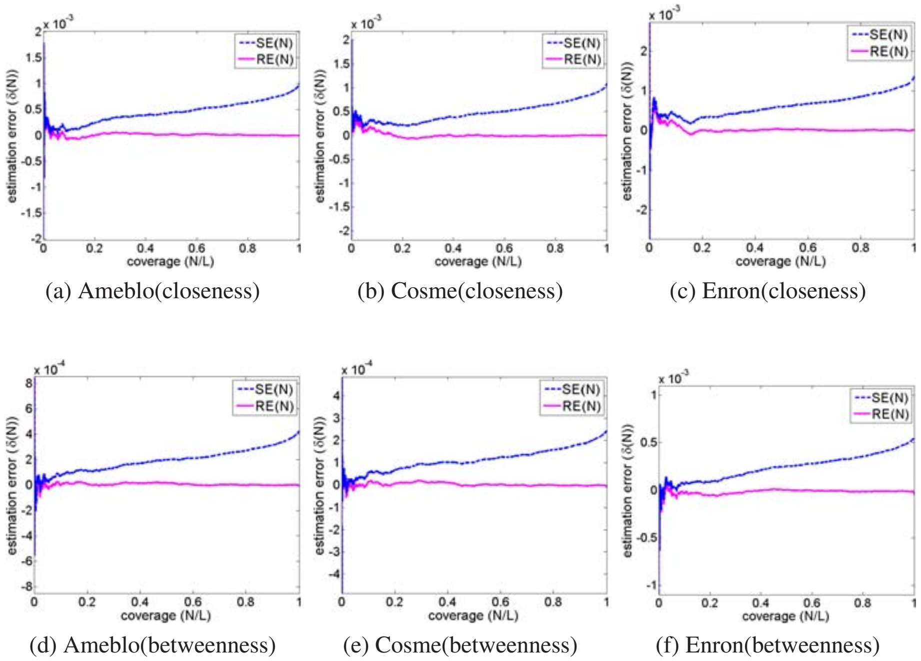
Fig. 3. Fluctuation of the estimated value of the betweenness centrality as a function of the coverage for the top and second nodes that respectively have the largest and second largest true values of the centrality in the Ameblo, Cosme, and Enron networks.

where $\mu_{G',r}(u)$ denotes the estimated value of the centrality of node u for partial graph G' in the r-th trial. Then, we used $\mathcal{E}_{RMSE}(N)$ as the true approximation error, and $RE(N)$ and $SE(N)$ as the estimated error.

Namely, in Fig. 4, the difference $\delta(N)$ is defined as $RE(N) - \mathcal{E}_{RMSE}(N)$ for the resampling error (the red curves), and $SE(N) - \mathcal{E}_{RMSE}(N)$ for the standard error (the blue broken curves). Here, we only show the results for the top node of each network in both centralities and omit the results for the second node because the tendency observed for the second node was quite similar to the one for the top nodes.

From these results, we can observe that the difference fluctuates when the value of coverage is less than 0.2 in both cases of $RE(N)$ and $SE(N)$, but for a larger coverage it becomes remarkably stable and almost equal to 0.0 in the case of $RE(N)$, while it increases as the value of coverage becomes larger in the case of $SE(N)$. This tendency is common to every network regardless of the centrality used. These results show that the proposed resampling error can precisely estimate the approximation error from the true values of a centrality measure if the coverage is larger than a certain threshold, say 0.2, while the standard error tends to overestimate the true approximation error.

Consequently, we can say that the resampling error we proposed is more promising than the standard error in this kind of estimation problem, and can give a tighter and more precise estimate of the approximation error with high confidence level than the standard error does.



(a) Ameblo(closeness) (b) Cosme(closeness) (c) Enron(closeness)

(d) Ameblo(betweenness) (e) Cosme(betweenness) (f) Enron(betweenness)

Fig. 4. Fluctuation of the difference between the true and the estimated approximation errors as a function of the coverage for the top node that has the largest true value of the centrality in the Ameblo, Cosme, and Enron networks.

6 Conclusion

In this paper, we addressed a problem of estimating the value of a centrality measure for a node in a social network. Centrality measure plays an important role in social network analysis since it characterizes nodes in a network and its values indicate the importance of nodes in some respects. Thus, it is crucial to efficiently calculate the value of a centrality measure for each node, but its computation could be intractable for those centrality measures that require use of a global network structure for their computation when the network becomes very large. It is inevitable to take a sampling-based approach to deal with the scalability problem, in which we approximate the true value of a centrality only from a partial network that can be generated by sampling nodes from the whole network. What is important is that we ensure the accuracy of the approximations without knowing the truth. To this end, we proposed a resampling-based framework to estimate the approximation error of the estimated values of a centrality measure for each node. We have conducted extensive experiments on three real world networks varying the coverage ratio of nodes to be sampled, and evaluated the proposed framework by comparing it with the standard error known in statistics using two typical centrality measures, the closeness and betweenness centralities. We empirically confirmed that the proposed framework enables us to estimate the approximation error more tightly and more precisely with the confidence level of 95% even for a partial network whose coverage is small, say 0.2, than using the standard error estimate. Furthermore, the experimental results suggest that we could potentially estimate top-K nodes for a small K, say 10, and possibly their ranking in a given centrality measure with high confidence level only from a small partial network. It is noted that the framework we proposed is not specific to computation of centrality measures. Indeed, it is very generic and applicable to any other estimation problems that require aggregation of many (but a finite number of) primitive computations. We believe that the conclusion obtained in this paper can generalize but we have yet to test out the proposed framework in a broader setting and also in different domains, too.

Acknowledgments

This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research under Grant No. AOARD-13-4042, and JSPS Grant-in-Aid for Scientific Research (C) (No. 26330261).

References

1. Bonacichi, P.: Power and centrality: A family of measures. Amer. J. Sociol. 92, 1170–1182 (1987)
2. Brandes, U.: A faster algorithm for betweenness centrality. Journal of Mathematical Sociology 25, 163–177 (2001)
3. Brin, S., L.Page: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems 30, 107–117 (1998)
4. Chen, W., Lakshmanan, L., Castillo, C.: Information and influence propagation in social networks. Synthesis Lectures on Data Management 5(4), 1–177 (2013)
5. Freeman, L.: Centrality in social networks: Conceptual clarification. Social Networks 1, 215–239 (1979)
6. Henzinger, M.R., Heydon, A., Mitzenmacher, M., Najork, M.: On near-uniform url sampling. The International Journal of Computer and Telecommunications Networking 33(1-6), 295–308 (2000)
7. Katz, L.: A new status index derived from sociometric analysis. Sociometry 18, 39–43 (1953)
8. Kleinberg, J.: The convergence of social and technological networks. Communications of ACM 51(11), 66–72 (2008)
9. Klimt, B., Yang, Y.: The enron corpus: A new dataset for email classification research. In: Proceedings of the 2004 European Conference on Machine Learning (ECML'04). pp. 217–226 (2004)
10. Kurant, M., Markopoulou, A., Thiran, P.: Towards unbiased bfs sampling. IEEE Journal on Selected Areas in Communications 29(9), 1799–1809 (2011)
11. Leskovec, J., Faloutsos, C.: Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06). pp. 631–636 (2006)
12. Newman, M.E.J.: Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. Physical Review E 64, 016132 (2001)
13. Zhuge, H., Zhang, J.: Topological centrality and its e-science applications. Journal of the American Society of Information Science and Technology 61, 1824–1841 (2010)