Resampling-Based Gap Analysis for Detecting Nodes with High Centrality on Large Social Network

Kouzou Ohara[1(✉)], Kazumi Saito[2], Masahiro Kimura[3], and Hiroshi Motoda[4,5]

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

Abstract. We address a problem of identifying nodes having a high centrality value in a large social network based on its approximation derived only from nodes sampled from the network. More specifically, we detect gaps between nodes with a given confidence level, assuming that we can say a gap exists between two adjacent nodes ordered in descending order of approximations of true centrality values if it can divide the ordered list of nodes into two groups so that any node in one group has a higher centrality value than any one in another group with a given confidence level. To this end, we incorporate confidence intervals of true centrality values, and apply the resampling-based framework to estimate the intervals as accurately as possible. Furthermore, we devise an algorithm that can efficiently detect gaps by making only two passes through the nodes, and empirically show, using three real world social networks, that the proposed method can successfully detect more gaps, compared to the one adopting a standard error estimation framework, using the same node coverage ratio, and that the resulting gaps enable us to correctly identify a set of nodes having a high centrality value.

Keywords: Gap analysis · Error estimation · Resampling · Node centrality

1 Introduction

Recently, social media such as Facebook, Digg, Twitter, etc. becomes an extremely popular communication tool on a global scale, and generates large-scale social networks on the web. Such networks allow us to share a wide variety of topics

T. Cao et al. (Eds.): PAKDD 2015, Part I, LNAI 9077, pp. 135–147, 2015.
DOI: 10.1007/978-3-319-18038-0_11

that have been posted on social media because those topics can rapidly and widely spread through the networks. Thus, in recent years, social media plays an important role as information infrastructure, and social networks constructed on it have been extensively investigated from various angles [4,8].

In such social network analysis, we can get an insight into some features of a given network by using the node centrality [1,3,5,7,14], which characterizes nodes in the network based on its topology. Typical ones include the degree, closeness, and betweenness centralities. Some of them such as the degree centrality are based only on the information of neighboring nodes of a target node, but some others are also on global structure of a network. For example, to compute the betweenness centrality, we have to enumerate paths between arbitrary node pairs, which is computationally very expensive. Since a social network on the web can easily grow in size, it is crucial to efficiently compute values of such a centrality to analyze a large network.

To this kind of problem on scalability, sampling-based approaches have been proposed so far [6,10,11], which investigate sampling methods that can obtain better approximations of true centrality values. Those methods are roughly categorized into uniform sampling, non-uniform sampling, and traversal/walk-based sampling. In contrast to them, we proposed a framework that ensures the accuracy of the approximations under uniform sampling [13], in which we estimated the approximation error referred to as resampling error by considering all possible partial networks of a fixed size that are generated by resampling nodes according to a given coverage ratio and approximated centrality values derived from them. It is empirically shown that the resampling-based framework provides a tighter approximation error with a higher confidence level than the traditional standard error in statistics under a given sampling ratio.

Unlike these existing approaches, in this paper, we consider detecting a set of nodes having a high centrality value only from approximations derived from sampled nodes with an adequate confidence level, instead of trying to accurately estimate the centrality value itself. We are interested in such nodes because they tend to play an important role for information diffusion on the network. To this end, we consider a list of nodes in descending order of the approximate centrality value, and devise an algorithm to efficiently detect gaps that exist between two adjacent nodes in the list. Here, we say a gap, or a boundary exists between two adjacent nodes in the list if it can divide the ordered list of nodes into two groups so that any node belonging to one group has a higher centrality value than any node in another group with a given confidence level. We incorporate confidence intervals of true centrality values for each node to detect such gaps, and adopt the above resampling-based estimation framework to estimate the confidence intervals as accurately as possible. The results of extensive experiments on three real world social networks demonstrate that using the resampling error for detecting gaps outperforms using the standard error in terms of the number of gaps detected, and that the resulting gaps allow us to correctly identify nodes having a high centrality value.

2 Resampling-Based Estimation Framework

In this section, according to the work [13], we revisit the resampling-based framework for estimating an approximation error with a given confidence level and its application to computing the node centrality.

2.1 General Framework

Let S be a set of objects such that $|S| = L$, and f a function that assigns a value to each object $s \in S$. Then, the problem we address is estimating the average μ over the set of entire values $\{f(s) \,|\, s \in S\}$ only from its arbitrary subset of partial values $\{f(t) \,|\, t \in T \subset S\}$. Let $\mu(T)$ be the partial average over a subset T whose number of elements is N, *i.e.*, $\mu(T) = (1/N) \sum_{t \in T} f(t)$. Then, we consider using this partial average $\mu(T)$ as an approximate solution of the true average μ and estimating an expected approximation error $RE(N)$, referred to as resampling error, which is the difference between μ and $\mu(T)$, with respect to the number of elements N, if L is too large to compute μ. Given $\mathcal{T} \subset 2^S$ that is a family of subsets of S such that $|T| = N$ for $T \in \mathcal{T}$, the resampling error $RE(N)$ is defined as follows:

$$RE(N) = \sqrt{\langle (\mu - \mu(T))^2 \rangle_{T \in \mathcal{T}}} = \sqrt{\binom{L}{N}^{-1} \sum_{T \in \mathcal{T}} \left(\mu - \frac{1}{N} \sum_{t \in T} f(t) \right)^2} = C(N)\sigma, \tag{1}$$

where the factor $C(N) = \sqrt{(L-N)/((L-1)N)}$ and $\sigma = \sqrt{L^{-1} \sum_{s \in S} (f(s) - \mu)^2}$ is the standard deviation. Note that since the estimation error of Equation (1) is regarded as the standard deviation with respect to the number of elements N, we can claim from a statistical viewpoint that for a given subset T such that $|T| = N$, and its partial average value $\mu(T)$, the probability that $|\mu(T) - \mu|$ is larger than $1.96 \times RE(N)$, is less than 5%. In other words, the range of $\mu(T) \pm 1.96 \times RE(N)$ is regarded as the 95% confidence interval of μ.

On the other hand, we can consider a standard approach to this problem that is based on the i.i.d. (independently identical distribution) assumption. More specifically, for a given subset T that has N elements, that is, $T = \{t_1, \cdots, t_N\}$, it is assumed 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, the standard error $SE(N)$ based on this assumption is defined as follows:

$$SE(N) = \sqrt{\langle (\mu - \mu(T))^2 \rangle} = \sqrt{\sum_{t_1 \in S} \cdots \sum_{t_N \in S} \left(\mu - \frac{1}{N} \sum_{n=1}^{N} f(t_n) \right)^2 \prod_{n=1}^{N} p(t_n)} = D(N)\sigma, \tag{2}$$

where $D(N) = 1/\sqrt{N}$ and σ is the standard deviation.

It is noted that the difference between Equations (1) and (2) is only their coefficient terms, $C(N)$ and $D(N)$, and that $C(N) \leq D(N)$, $C(L) = 0$ and $D(L) \neq 0$. Namely, $RE(N) \leq SE(N)$ for any N, and $RE(N)$ becomes 0 when

$N = L$, but not $SE(N)$. Note that the true standard deviation σ is needed in both Equations (1) and (2), but in practice, 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 if $|S|$ is too large to compute σ, which is just the case where sampling is needed.

2.2 Application to Node Centrality Estimation

Next, we present the way to apply the above estimation framework to node centrality estimation of a social network that is represented as 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. Here, we consider two node centrality measures, the closeness centrality and the betweenness centrality as in [13].

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, and we set $spl_G(u, v) = \infty$ when node v is unreachable from node u on G. Intuitively, a node u has a high value for this closeness centrality if a large number of nodes are reachable from u within relatively short path lengths. A standard technique for computing $cls_G(u)$ of each node $u \in V$ is the burning algorithm [12] whose computational complexity is $O(|E|)$. Thus, it takes a large amount of computation time for a huge social network consisting of millions of nodes. To apply the above estimation framework to the computation of an approximation of the closeness centrality $cls_G(u)$ of each node $u \in V$, we instantiate the set of objects S and the function f to this problem. In fact, we consider $S_u = V \setminus \{u\}$ as the set S and $f_u(v) = 1/spl_G(u, v)$ as the function f, and thereby can calculate a partial average value $cls_G(u; T)$ from an arbitrary subset $T \subset S_u \cup \{u\}$ and its approximation error, $RE(u; |T|)$ and $SE(u; |T|)$, according to the above framework.

Next, 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 that pass through node u. Thus, the betweenness of a node u becomes high if a large number of shortest paths between two nodes pass through node u. The Brandes algorithm [2] is a standard technique for computing $btw_G(u)$ of each node $u \in V$ and its computational complexity is $O(|E|)$. Thus, it requires a large amount of computation time for a large social network, too. Again, we consider instantiating S and f of the above estimation framework for computing an approximation

of the betweenness centrality $btw_G(u)$. More specifically, we regard the expression inside the large parentheses in Equation (4) as a function $btw_G(u;v)$, the betweenness of node u that restricts its starting node to v. Then, by considering $S_u = V \setminus \{u\}$ and $f_u(v) = btw_G(u;v)/(|V|-2)$, we can calculate a partial average value $btw_G(u;T)$ from an arbitrary subset $T \subset S_u \cup \{u\}$ and its estimation error, $RE(u;|T|)$ and $SE(u;|T|)$, based on the above estimation framework.

3 Gap Detection Method

In this section, we consider the way to detect a set of nodes having a high centrality value with a given confidence level based only on centrality values estimated from a subset of nodes in a network. First of all, we formally define the problem we address here. For a network $G(V,E)$, let $\mu_G(v)$ be the true value of a certain centrality measure for node $v \in V$, $\mu_G(v;T)$ be its estimation derived only from a subset of nodes $T \subseteq V$, and $\sigma(v;|T|)$ be its approximation error such as $RE(v;|T|)$ and $SE(v;|T|)$. In addition, given a node v, let $V_H(v;T) = \{u \in V; \mu_G(u;T) \geq \mu_G(v;T)\}$ and $V_L(v;T) = \{w \in V; \mu_G(w;T) < \mu_G(v;T)\}$ be disjoint partitions of V with respect to $\mu_G(v;T)$. Then, incorporating the confidence interval estimation in statistics, the problem can be defined as finding out all nodes $v \in V$ that satisfy the following inequality for $\forall u \in V_H(v;T)$ and $\forall w \in V_L(v;T)$:

$$\mu_G(u;T) - z(\alpha) \cdot \sigma(u;|T|) > \mu_G(w;T) + z(\alpha) \cdot \sigma(w;|T|) \tag{5}$$

where $0 < \alpha < 1$ and $z(\alpha)$ is the upper $\alpha/2$ critical value of the standard normal distribution. In other words, $\mu_G(u) > \mu_G(w)$ holds for $\forall u \in V_H(v;T)$ and $\forall w \in V_L(v;T)$ with the confidence level $C = 100(1-\alpha)\%$. Here, the upper half set $V_H(v;T)$ is a set that we want to identify, and we say that a gap exists between v and $v' \in \arg\max_{w \in V_L(v;T)} \mu_G(w;T)$. It is obvious that a straightforward approach to this problem requires the computational complexity of $O(|V|^3)$ because it has to check $|V_H(v;T)||V_L(v;T)|$ pairs of nodes for each v, which is not acceptable when a given social network is very large.

To cope with this, we first consider a lower error bound of $V_H(v;T)$ and an upper error bound of $V_L(v;T)$, respectively defined as $LB(V_H(v;T);\alpha) = \min_{u \in V_H(v)}(\mu_G(u;T) - z(\alpha)\sigma(u;|T|))$ and $UB(V_L(v;T);\alpha) = \max_{w \in V_L(v)}(\mu_G(w;T) + z(\alpha)\sigma(w;|T|))$. Hereafter, for simplicity, $LB(V_H(v;T);\alpha)$ and $UB(V_L(v;T);\alpha)$ are denoted by $LB(V_H(v);T,\alpha)$ and $UB(V_L(v);T,\alpha)$, respectively. Then, we focus on the fact that the above problem is reduced to finding all nodes $v \in V$ that satisfy the relation $LB(V_H(v);T,\alpha) > UB(V_L(v);T,\alpha)$ for given α. Since both $LB(V_H(v);T,\alpha)$ and $UB(V_L(v);T,\alpha)$ can be simultaneously computed for arbitrary $v \in V$ by making only one pass through V, the total computational complexity becomes $O(|V|^2)$, which is smaller than $O(|V|^3)$, but it is still hard to find all of such nodes when the size of a network gets larger.

Thus, we further consider an ordered list $(v_1, v_2, \cdots, v_{|V|})$ of nodes in V resulted from sorting them in descending order of the value of $\mu_G(v;T)$, *i.e.*, $\mu_G(v_i;T) \geq \mu_G(v_{i+1};T)$ for $i \in \{1, \cdots, |V|-1\}$. Then, $LB(V_H(v_k);T,\alpha)$ is

recursively defined as $LB(V_H(v_k);T,\alpha) = \min(LB(V_H(v_{k-1});T,\alpha),\ \mu_G(v_k;T) - z(\alpha)\sigma(v_k;|T|))$. As well, $UB(V_L(v_k);T,\alpha)$ is defined as $UB(V_L(v_k);T,\alpha) = \max(UB(V_L(v_{k+1});T,\alpha),\ \mu_G(v_{k+1};T) + z(\alpha)\sigma(v_{k+1};|T|))$. Considering these definitions, we can compute $LB(V_H(v);T,\alpha)$ and $UB(V_L(v);T,\alpha)$ for every node $v \in V$ by making only one pass, each, through the list $(v_1, v_2, \cdots, v_{|V|})$, respectively, which implies that we can detect all gaps by making two passes through the ordered list. More specifically, in the first pass, referred to as the forward step, we compute $LB(V_H(v_k);T,\alpha)$ varying k from 1 to $|V|-1$, and then, in the second pass called the backward step, we compute $UB(V_L(v_k);T,\alpha)$ and detect a gap if $LB(V_H(v_k);T,\alpha) > UB(V_L(v_k);T,\alpha)$ holds varying k from $|V|$ to 2. The computational complexity of this method is governed by that of its sorting process, and thus becomes $O(|V|\log|V|)$, which enables the practical gap analysis even for a large social network. The procedure is summarized as follows:

1. $A \leftarrow \emptyset$, $LB(V_H(v_1);T,\alpha) = \mu_G(v_1;T) - z(\alpha)\sigma(v_1;|T|))$, and $UB(V_L(v_{|V|});T,\alpha) = 0$;
2. (Forward step) For $k = 2$ to $|V|-1$,
   $LB(V_H(v_k);T,\alpha) = \min(LB(V_H(v_{k-1});T,\alpha),\ \mu_G(v_k;T) - z(\alpha)\sigma(v_k;|T|))$;
3. (Backward step) For $k = |V|-1$ to 2,
   (a) $UB(V_L(v_k);T,\alpha) = \max(UB(V_L(v_{k+1});T,\alpha),\ \mu_G(v_{k+1};T) + z(\alpha)\sigma(v_{k+1};|T|))$;
   (b) $A \leftarrow A \cup \{v_k\}$ if $LB(V_H(v_k);T,\alpha) > UB(V_L(v_k);T,\alpha)$;
4. Output A, and terminate.

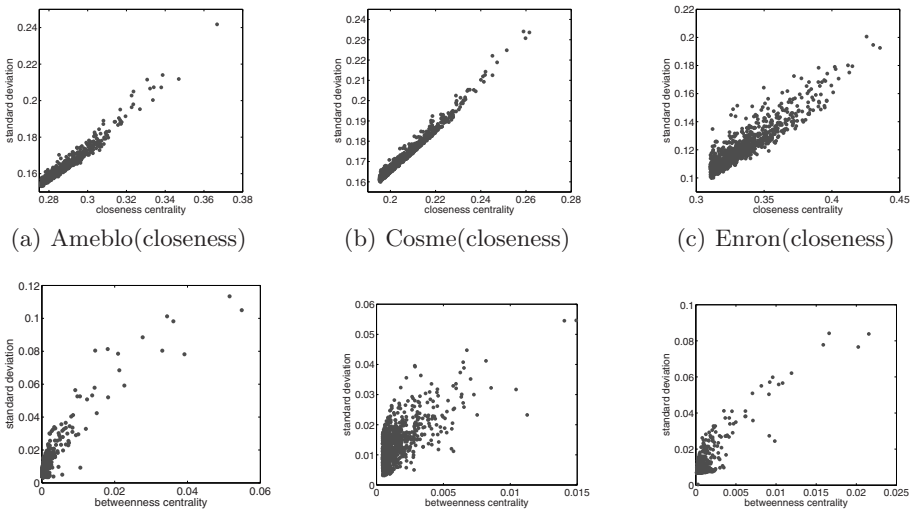
We consider three kinds of methods by adopting different definitions of the estimated error $\sigma(v;|T|)$, which are $\sigma(v;|T|) = 0$, $\sigma(v;|T|) = SE(v;|T|)$, and $\sigma(v;|T|) = RE(v;|T|)$. We refer to these methods as the naive, SE, and RE method, respectively. Note that the naive method assumes $\mu_G(v;T) = \mu_G(v)$. Thus, it determines that there exists a gap between nodes v_k and v_{k+1} for every k such that $\mu_G(v_k;T) \neq \mu_G(v_{k+1};T)$. On the other hand, since $SE(v;|T|)$ overestimates the approximation error of $\mu_G(v;T)$ compared to $RE(v;|T|)$, the number of gaps detected by the SE method becomes less than that by the RE method. For more details, we empirically compare these methods through experiments on real world social networks as described below.

4 Experiments

4.1 Datasets

We empirically evaluated the three gap detection methods described in the previous section on three datasets of real world networks that are represented as directed graphs. The first dataset is a network extracted from a Japanese blog service site "Ameba"[1], which has 56,604 nodes representing blogs in "Ameba" and 734,737 directed links among them. Each directed link is constructed from

[1] `http://www.ameba.jp/`



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

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

Fig. 1. Centrality values and their standard deviations of the top 1,000 nodes in descending order of the true value of each centrality in the Ameblo, Cosme, and Enron networks

blog u to blog v if blog u is registered as a favorite one in blog v. We refer to this network as the Ameblo network. The second one is a network extracted from a Japanese word-of-mouth communication site for cosmetics, "@cosme"[2], consisting of 45,024 nodes representing its users and 351,299 directed links, in which a link (u, v) means that user v registers user u as her favorite one. We refer to this directed network as the Cosme network. The last one is a network derived from the Enron Email Dataset [9], which has 19,603 nodes and 210,950 links. In this network, a node is an email address that appears in the dataset as either a sender or a recipient, while a directional link (u, v) between two email addresses u and v means that u sent an email to v. We refer to this directed network as the Enron network. These three networks are not very huge, but large enough to investigate the basic performance of the three methods from various angles. We thus simply use the standard deviation σ derived from S to compute the resampling and standard errors.

Figures 1(a) to 1(c) show the top 1,000 nodes in descending order of true value of the closeness centrality in the Ameblo, Cosme, and Enron networks, respectively, while Figures 1(d) to 1(f) show the top 1,000 nodes in descending order of true value of the betweenness centrality for the same three networks. We only plotted the top 1,000 nodes because we are interested in nodes having high centrality values. In each figure, the horizontal axis indicates the values of corresponding centrality, and the vertical axis shows its standard

[2] `http://www.cosme.net/`

deviation defined as $\sigma_{\mu_G}(u) = \sqrt{(|V|-1)^{-1}\sum_{v\in V, v\neq u}\left(f_u(v)-\mu_G(u)\right)^2}$, where $\mu_G(u)$ sands for either $cls_G(u)$ or $btw_G(u)$, and $f_u(v)$ is $1/spl_G(u,v)$ for $cls_G(u)$ and $btw_G(u;v)/(|V|-2)$ for $btw_G(u)$. From these figures, we can observe that higher-ranked nodes in each centrality measure are distinguishable from each other in every network because of their distinctive values of the centrality, while it looks hard to do the same for lower-ranked nodes. This tendency can be found more clearly in the plots for the betweenness centrality in which nodes are scattered over a larger area. From these observations, we can expect that it is harder to detect gaps that exist between lower-ranked nodes compared to the ones between higher-ranked nodes and that more gaps can be detected for the betweenness centrality than for the closeness centrality.

4.2 Results

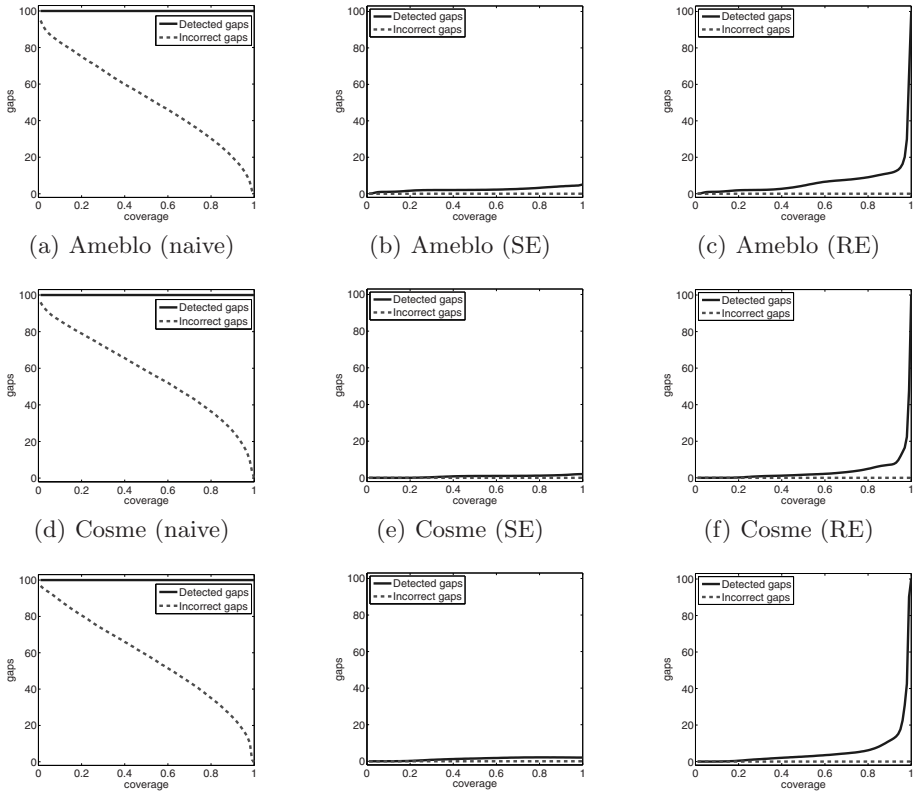
We applied the naive, SE, and RE methods to the three networks mentioned above for the closeness and betweenness centralities, and examined the number of gaps they detected and how many gaps among them were correct. A correct gap is the one that the resulting upper half set $V_H(v_k;T)$ corresponds exactly to the true upper half set that is a set of the top k nodes in the descending order of the true centrality value. In this experiment, we adopted the confidence level of 95% ($\alpha = 0.05$) as a typical one and fixed it, while we varied the coverage $|T|/|V|$ from 0.01 to 1.00 by 0.01 points to see how the number of gaps detected changes according to the coverage. More precisely, we randomly sampled nodes from V without replacement, added it to the subset T one by one, and counted the number of gaps detected and the number of gaps correctly detected each time the coverage increases by 0.01. Since we are interested in nodes having a high centrality value, we considered only the top K nodes in descending order of the estimated value of the corresponding centrality at each coverage. We repeated this process $R = 1,000$ times and computed the average over them.

Figure 2 shows the results for the closeness centrality in the case of $K = 100$. The horizontal axis means the coverage, and the vertical axis means the number of gaps. The blue solid line and the red broken line represent the number of gaps detected and the number of gaps incorrectly detected by the corresponding method, respectively, which are defined as follows:

$$(\#\text{ of gaps detected}) \quad \frac{1}{R}\sum_{r=1}^{R}\frac{|A(c,r)|}{|A_{nv}(c,r)|}\times K \tag{6}$$

$$(\#\text{ of gaps incorrectly detected}) \quad \frac{1}{R}\sum_{r=1}^{R}\frac{|A(c,r)\setminus A^*(c,r)|}{|A_{nv}(c,r)|}\times K, \tag{7}$$

where $A(c,r)$ is the set of nodes corresponding to gaps, *i.e.,* A in the algorithm in Section 3 detected by the respective method at coverage c in the r-th iteration, while $A^*(c,r)$ is the set of nodes correctly detected among them. It is noted that since some of the top K nodes may have the same estimation, these numbers



(a) Ameblo (naive) (b) Ameblo (SE) (c) Ameblo (RE)

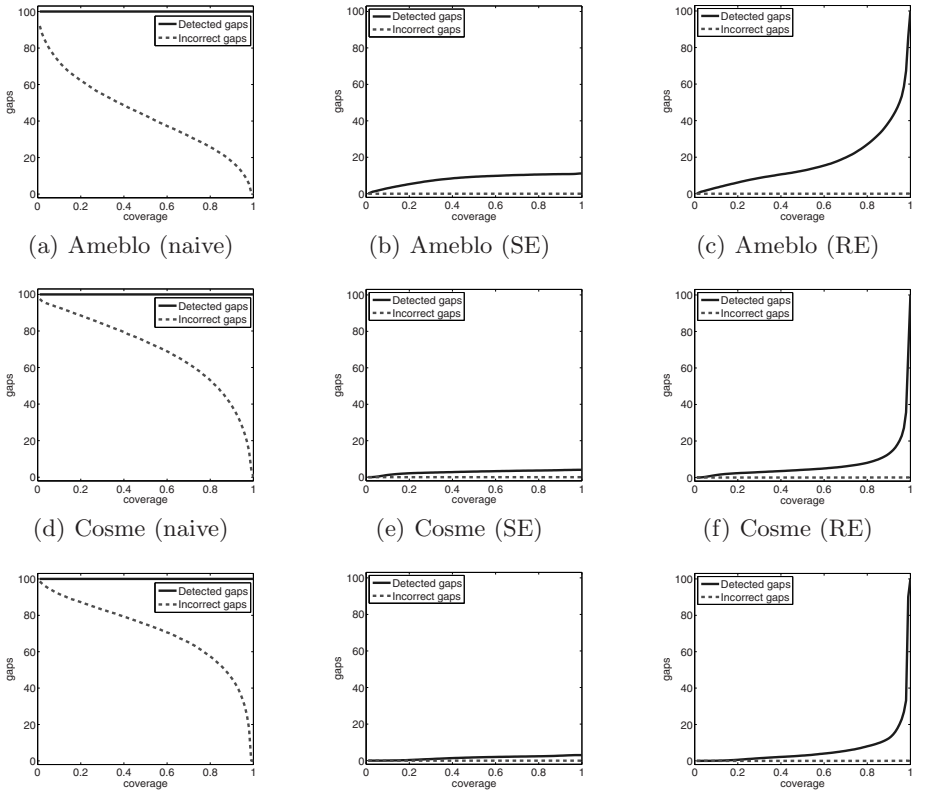
(d) Cosme (naive) (e) Cosme (SE) (f) Cosme (RE)

(g) Enron (naive) (h) Enron (SE) (i) Enron (RE)

Fig. 2. Fluctuation of the number of gaps detected by the naive, SE, and RE methods as a function of the coverage for the top $K = 100$ nodes in descending order of the estimated value of the *closeness* centrality in the Ameblo, Cosme, and Enron networks

are normalized by the number of gaps detected by the naive method $|A_{nv}(c, r)|$ that corresponds to the number of node pairs v_i and v_{i+1} having different estimations. Thus, the blue solid line for the naive method always exhibits the best performance ($=K$).

From these results, it is found that although the number of gaps incorrectly detected by the naive method decreases as the coverage becomes larger, it is much larger than the ones by the other two methods that are almost exactly 0. Whereas, the number of gaps detected either by the SE or RE method is very small compared to the one by the naive method. Especially, the number of gaps detected by the SE method increases only a very little even if the coverage becomes closer to 1.0. On the other hand, the number of gaps detected by th RE method is slightly larger than the one by the SE method while the coverage is small, but it rapidly increases at around $c = 0.9$ and finally becomes 100 while the number of gaps incorrectly detected remains almost 0. This difference



(a) Ameblo (naive) (b) Ameblo (SE) (c) Ameblo (RE)

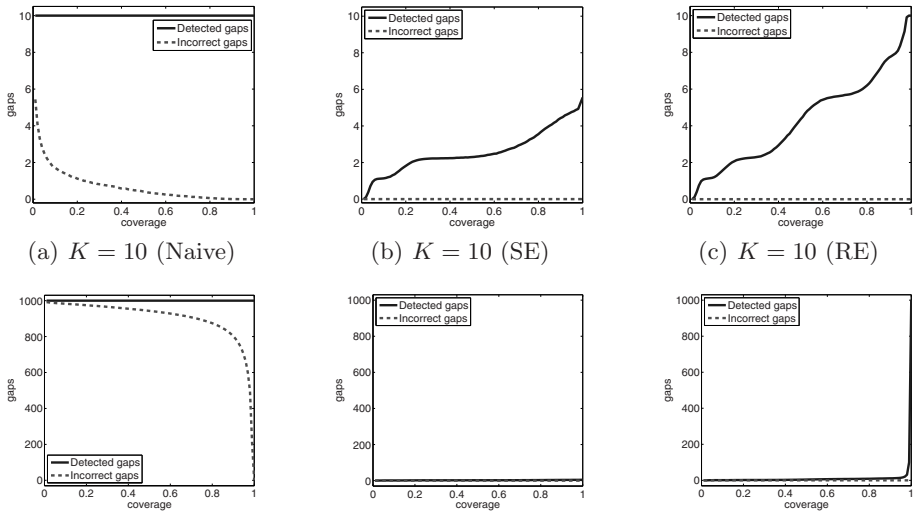
(d) Cosme (naive) (e) Cosme (SE) (f) Cosme (RE)

(g) Enron (naive) (h) Enron (SE) (i) Enron (RE)

Fig. 3. Fluctuation of the number of gaps detected by the naive, SE, and RE methods as a function of the coverage for the top $K = 100$ nodes in descending order of the estimated value of the *betweenness* centrality in the Ameblo, Cosme, and Enron networks

comes from their nature that the resampling error $RE(v; |T|)$ converges to 0 as $|T|$ approaches to $|V|$, while the standard error $SE(v; |T|)$ does not. These tendencies are also observed in the results for the betweenness centrality shown in Fig. 3.

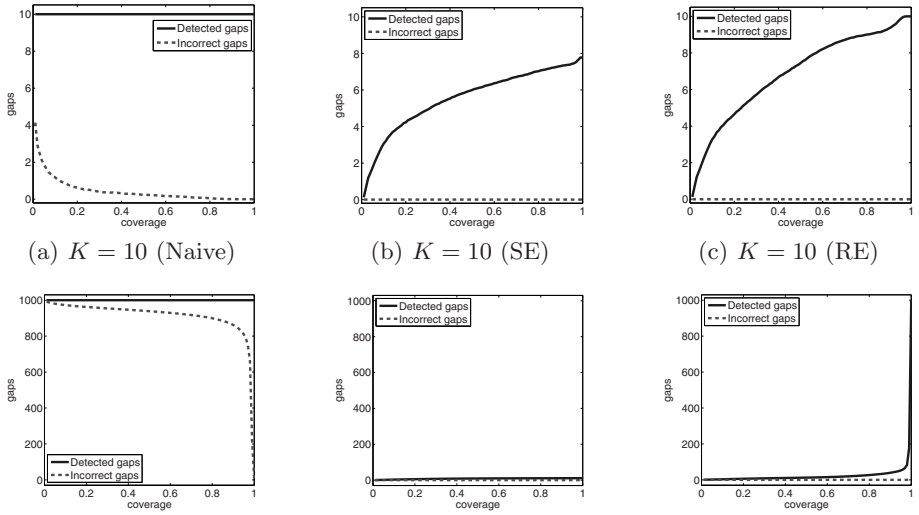
Next, we examined in the cases of $K = 10$ and $1,000$. Due to the page limitation, we will show only the results for the Ameblo network here, but we observed the same tendencies for the others. Figures 4 and 5 show the results for the closeness centrality and for the betweenness centrality, respectively. From Figs. 4(a) and 5(a), the number of gaps incorrectly detected by the naive method is relatively small compared to the results for $K = 100$ although it is still larger than the ones by the other methods that are almost 0 in this case, too. This is because the higher-ranked nodes in the true centrality value are distinguishable as shown in Fig. 1. Due to the same reason, the number of gaps detected either by the SE or RE method is relatively large compared to the case of $K = 100$.



(a) $K = 10$ (Naive) (b) $K = 10$ (SE) (c) $K = 10$ (RE)

(d) $K = 1,000$ (Naive) (e) $K = 1,000$ (SE) (f) $K = 1,000$ (RE)

Fig. 4. Fluctuation of the number of gaps detected by the naive, *SE*, and *RE* methods as a function of the coverage for the top $K = 10$ and $K = 1,000$ nodes in descending order of the estimated value of the *closeness* centrality in the Ameblo network



(a) $K = 10$ (Naive) (b) $K = 10$ (SE) (c) $K = 10$ (RE)

(d) $K = 1,000$ (Naive) (e) $K = 1,000$ (SE) (f) $K = 1,000$ (RE)

Fig. 5. Fluctuation of the number of gaps detected by the naive, *SE*, and *RE* methods as a function of the coverage for the top $K = 10$ and $K = 1,000$ nodes in descending order of the estimated value of the *betweenness* centrality in the Ameblo network

It is more clearly found that the RE method can correctly detect more gaps than the SE method does at the same coverage by comparing Figs. 4(b) and 4(c) for the closeness centrality, and by comparing Figs. 5(b) and 5(c) for the betweenness centrality. Furthermore, as expected above, by comparing Figs. 4(b) and 5(b), we can observe that the number of gaps detected by the SE method for the betweenness centrality is larger than that for the closeness centrality. The similar tendency can be observed for the RE method from Figs. 4(c) and 5(c). On the other hand, we can observe from the results for $K = 1,000$ that the number of gaps incorrectly detected by the naive method is relatively large, and the number of gaps detected by the other methods is relatively small, compared to the other results. This result demonstrates our expectation that it is harder to correctly detect gaps that exist between lower-ranked nodes.

To summarize the above results, the naive method is not reliable for a large K. It can detect many gaps correctly for a small K, say 10, but it detects incorrect gaps if the coverage is low. This is not desirable as a means to reduce the computational cost for detecting nodes having a high centrality value. On the other hand, the SE and RE methods satisfactorily detect gaps correctly regardless of the value of coverage. The SE method is more conservative by overestimating the error margin and less useful than the RE method in terms of the number of gaps detected at the same coverage. Note that although the number of gaps detected by the RE method is limited for a low coverage, the resulting gaps are more likely to appear between nodes having a high centrality value, which is desirable for us to detect important nodes in a network.

5 Conclusion

In this paper, we addressed a problem of identifying nodes having a high centrality value in a social network based only on its approximation derived from a limited number of sampled nodes. To this end, we focused on confidence intervals of true centrality value for each node, and considered detecting gaps that divide a set of nodes into two groups so that any node in one group has a higher centrality value than any one in another does with a given confidence level. To estimate confidence intervals as accurately as possible, we employed the resampling-based framework for estimation of the approximation error, and devised an algorithm that can efficiently detect gaps whose computational complexity is $O(|V|log|V|)$ for the number of nodes in a network, $|V|$, which is much less than $O(|V|^3)$ of the straightforward approach. Through extensive experiments on three real world social networks for the closeness and betweenness centralities, we empirically confirmed that the proposed method can correctly detect gaps that exist between high-ranked nodes with the confidence level of 95% even for a partial network whose coverage is small, say 0.2, and can detect more gaps compared to the one that uses the standard error to estimate confidence intervals at the same coverage ratio. Especially, the ratio of gaps incorrectly detected to the total number of detected gaps is almost 0 for both the methods. It is noted that the method we proposed is not only specific to identification of nodes having a high

centrality value, but also applicable to any other estimation problems to which the resampling-based estimation framework is applicable. We believe that the conclusions obtained in this paper can generalize but we have yet to test out the proposed method in a broader setting and 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., Page, L.: 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: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 217–226. Springer, Heidelberg (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 2006), 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. Ohara, K., Saito, K., Kimura, M., Motoda, H.: Resampling-based framework for estimating node centrality of large social network. In: Džeroski, S., Panov, P., Kocev, D., Todorovski, L. (eds.) DS 2014. LNCS, vol. 8777, pp. 228–239. Springer, Heidelberg (2014)
14. 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)