人工知能学会研究会資料
SIG-KBS-B402-06

リサンプリング法に基づくギャップ分析を用いた高中心性ノードの同定

Identifying High Centrality Nodes using Resampling-based Gap Analysis

大原剛三[1*]　斉藤和巳[2]　木村昌弘[3]　元田浩[4]
Kouzou Ohara[1], Kazumi Saito[2], Masahiro Kimura[3], and Hiroshi Motoda[4]

[1] 青山学院大学理工学部
[1] College of Science and Engineering, Aoyama Gakuin University
[2] 静岡県立大学経営情報学部
[2] School of Management and Information, University of Shizuoka
[3] 龍谷大学理工学部
[3] Faculty of Science and Technology, Ryukoku University
[4] 大阪大学産業科学研究所
[4] The Institute of Scientific and Industrial Research, Osaka University

Abstract: We address a problem of identifying high centrality nodes in a large social network based on approximated centrality values derived from a small portion of nodes sampled uniformly at random from the whole set. To this end, we apply our resampling-based framework to estimate the approximation error, and detect gaps between nodes with a given confidence level. Here, a gap means a clear difference between two nodes in terms of a centrality measure, and gap detection means, given two nodes, determining which node has a greater centrality value than the other with a given confidence level. On two real world social networks, we empirically show that the proposed method can successfully detect more gaps only from several tens of percent of the node, compared to the one adopting a standard error estimation framework, and that the resulting gaps enable us to correctly identify a set of nodes having a high centrality value.

1　はじめに

今日，Facebook や Twitter などのソーシャルメディアの普及により，インターネット上に巨大な社会ネットワークが形成されつつある．ソーシャルメディアに一旦投稿された情報は，その上に展開している社会ネットワークを通して急速，かつ広範囲に拡散され，多数の人々に共有され得る．そのような情報は我々の日常における意思決定にも多大な影響を与えるため，近年，社会学のみならず計算機科学も含めた多様な分野において社会ネットワークの分析が進められている [7, 4].

それらの社会ネットワーク分析においては，様々な中心性と呼ばれる指標が提案され，利用されている [6, 5, 1, 3, 11]．中心性指標はネットワーク構造に基づきノードを特徴づけるものであり，たとえば，その大きさから各ノードがどのような意味で，どの程度重要かについての情報を我々にもたらしてくれる．また，ネットワークに対するスケールフリー性が次数分布から導かれるように，ネットワーク全体の構造的特徴を知る手がかりともなる．その一方で，近接中心性や媒介中心性などのように，その値を求めるために対象ノードの近傍情報（隣接ノードの情報）のみならず，任意のノード間の最短経路などのようなネットワーク全体にわたる情報を必要とするものがあり，それらに関しては，ネットワークが大きくなるとその計算が困難になる．

実際には，そのような計算コストの高い中心性指標は，ノードペアなどから導かれる値を基礎に，その平均値として定義されることが多い．このことから，その計算コスト軽減に対する 1 つのアプローチとしては，サンプリングによるノード数の削減が考えられる．ノード数を制限することにより中心性指標の計算は容易になるが，得られるのは近似値となるため，真の指標値との近似誤差を精度良く推定することが重要となる．この問題に対して，我々は近似誤差を精度よく推定するリサンプリング法に基づいた枠組みを提案し，それにより得られる近似誤差（以下，リサンプリング誤差）が独立同分布の下でのサンプリングを仮定した標準的な近似誤差（以下，標準誤差）よりも正確な誤差範囲を

*連絡先：青山学院大学理工学部情報テクノロジー学科
〒 252-5258 相模原市中央区淵野辺 5-10-1
E-mail: ohara@it.aoyama.ac.jp

与えることを実験的に示している [10].

一方，社会ネットワーク分析では，全ノードの中心性の値を知ることよりも，高い中心性をもつノードが興味の対象となることが多い．そのため，本稿では，高い中心性をもつノード集合をサンプリングにより得られた中心性指標の近似値から精度よく同定することを考える．具体的には，中心性指標値に大小関係がある 2 つのノード間には**ギャップがある**とし，そのようなギャップとそのギャップをもつノードの大小関係を中心性指標の近似値から一定の精度で検出することを試みる．統計的な観点からは，これは，与えられた信頼度の下で各ノードの中心性指標値の信頼区間を求め，その重複関係を調べることに相当する．そこで本研究では，その信頼区間の導出に前述のリサンプリング誤差を導入し，実際の社会ネットワークを用いた評価実験を通して，標準誤差を利用するよりも多くのギャップを検出し，かつ検出したギャップが高い中心性指標値をもつノード集合の同定に有用であることを示す．

2 リサンプリング法に基づいた近似誤差推定

本節では，文献 [10] に従い，リサンプリング法に基づいた近似誤差推定の一般的な枠組み，およびその近接中心性と媒介中心性への適用について述べる．

2.1 一般的枠組み

いま，ある集合 S（$|S| = L$）に対して，f を S 中の各要素に何らかの値を対応付ける関数とする．このとき，S に対する f の平均値 $\mu = (1/L)\sum_{s\in S} f(s)$ を，S の任意の部分集合 T（$|T| = N$）に対する f の値 $\{f(t)|t \in T, T \subset S\}$ のみから推定することを考える．実際には，T に対する f の値から μ を直接推定することはできないため，μ と $\mu(T) = (1/N)\sum_{t\in T} f(t)$ 間の近似誤差を，μ を仮定せずに推定する．そのために，任意の $T \in \mathcal{T}$ に対して，$T \subset S$，かつ $|T| = N$ であるような S の部分集合族 $\mathcal{T} \subset 2^S$ を考える．このとき，μ と $\mu(T)$ の近似誤差 $RE(N)$ を以下のように定義する．

$$\begin{aligned} RE(N) &= \sqrt{\langle(\mu - \mu(T))^2\rangle_{T\in\mathcal{T}}} \\ &= \sqrt{\binom{L}{N}^{-1} \sum_{T\in\mathcal{T}} \left(\mu - \frac{1}{N}\sum_{t\in T} f(t)\right)^2} \end{aligned} \quad (1)$$

この式は，S から N 個の要素をリサンプリングすることで得られる $\mathcal{T}$ に対して，$T \in \mathcal{T}$ に対する部分平均 $\mu(T)$ と真の平均 μ との 2 乗平均平方根誤差（RMSE）を計算していると解釈できる．実際には，この式は以下のように変形できる．

$$RE(N) = \sqrt{\frac{L-N}{(L-1)N}} \times \sqrt{\frac{1}{L}\sum_{s\in S}(f(s)-\mu)^2} \quad (2)$$

ここで，右辺のうち N に依存するのは第 1 項のみであり，第 2 項は N に依存しないこと，および，この第 2 項が全体集合 S に対する関数 f の値の標準偏差となっていることに注意されたい．以下，第 2 項を定数項 σ，第 1 項をその係数項 $C(N)$ とし，$RE(N) = C(N)\sigma$ とする．このことから，実際には部分集合 T をリサンプリングせず，定数 L，σ，およびサンプリング数 N が与えられた時点で $RE(N)$ の値を計算可能なことがわかる．以下では，この $RE(N)$ をリサンプリング誤差と呼ぶ．

一方，より一般には，独立同分布の下でのサンプリングを前提に，μ と $\mu(T)$ の近似誤差の期待値を計算する．具体的には，$t \in T$ がある確率分布 $p(t)$ に従って独立に S から選択されたと仮定する．$p(t)$ としては，$p(t) = 1/L$ のような経験的な一様分布などが考えられる．このとき，T の要素を $T = \{t_1, \cdots, t_N\}$ とすると，μ と $\mu(T)$ の近似誤差の期待値は次式のように定義できる．

$$\begin{aligned} SE(N) &= \sqrt{\langle(\mu - \mu(T))^2\rangle} \\ &= \sqrt{\sum_{t_1\in S}\cdots\sum_{t_N\in S}\left(\mu - \frac{1}{N}\sum_{n=1}^{N} f(t_n)\right)^2 \prod_{n=1}^{N} p(t_n)} \\ &= \sqrt{\frac{1}{N}} \times \sqrt{\frac{1}{L}\sum_{s\in S}(f(s)-\mu)^2} \end{aligned} \quad (3)$$

この式も式 (2) 同様，右辺の第 1 項のみが N に依存し，第 2 項は関数 f の値の標準偏差となっていることから，実際には T をサンプリングすることなく，その値を求めることができる [1]．以下，$SE(N)$ を標準誤差と呼び，式 (2) 同様，右辺の第 2 項を定数項 σ，第 1 項をその係数項 $D(N)$ とし，$SE(N) = D(N)\sigma$ とする．ここで，$C(N) \le D(N)$ であり，$C(L) = 0$ であるのに対し $D(L) \neq 0$ であることに注意されたい．すなわち，ある N に対して $RE(N) \le SE(N)$ であり，$N = L$ のとき $RE(N)$ は 0 となるが，$SE(N)$ は 0 とはならない．

2.2 中心性指標への適用

次に，上記の近似誤差推定の枠組みを社会ネットワークにおけるノード中心性の推定問題に適用する．以下では，社会ネットワークを有向グラフ $G = (V, E)$ により表現する．ここで，V，および $E \subseteq V \times V$ はそれぞれネットワーク中のノード集合と有向リンク集合である．

[1] $RE(N)$, $SE(N)$ いずれの計算においても σ が必要となるが，$|S| = L$ が大きい場合はそもそも σ の計算が困難であるため，実際にはその近似値として，$|S'| = L'$ が十分小さい部分集合 $S' \subset S$ から現実的な計算時間で得られる標準偏差 σ' を近似値として用いる．

2.2.1 近接中心性

まず，G 中のノード $u \in V$ に対して次式で定義される近接中心性を考える．

$$cls_G(u) = \frac{1}{(|V|-1)} \sum_{v \in V, v \neq u} \frac{1}{spl_G(u,v)} \tag{4}$$

ここで，$spl_G(u,v)$ はグラフ G におけるノード u からノード v までの最短経路長を表し，v が u から到達可能でなければ $spl_G(u,v) = \infty$ とする．直観的には，ネットワーク中の他のどのノードにも比較的短い経路長で到達可能なノードほど近接中心性は大きな値となる．この近接中心性を計算する一般的な方法としては，基点ノードから 1 つのリンクを辿ることで新たに到達可能となるノード集合を漸進的に求める burning アルゴリズム [9] が知られているが，各ノード u に対する近接中心性 $cls_G(u)$ を求める計算量は $O(|E|)$ であり，巨大な社会ネットワークに対しては膨大な計算時間を要する．

この近接中心性に対して，前述のリサンプリングに基づいた近似誤差推定の枠組みを適用することを考える．ここでは，対象ノード u を除く V からサンプリングしたノード集合 T（$|T| = N$）のみから求められる u の近接中心性の近似値 $cls_G(u;T)$ と真の値 $cls_G(u)$ の近似誤差を考えることになる．そのために，前節における全サンプル集合 S，評価関数 f を近接中心性の計算に合わせて具体化する．まず，近接中心性はノード集合全体に対する値ではなく，各ノードに対する値であるため，S に関しては，対象ノードを u としたとき，$S_u = V \setminus \{u\}$ とする．ここで，$\setminus$ は集合差を意味する．一方，$cls_G(u)$ はその定義より，ノード u 以外のノード v に対して求められる $1/spl_G(u,v)$ の平均値であるため，評価関数 f に関しては，$f_u(v) = 1/spl_G(u,v)$ とする．これにより，$cls_G(u;T)$ を $(1/N)\sum_{v \in T} f_u(v)$ として求めることができ，式 (2)，および (3) に従い，$RE(N)$，$SE(N)$ をそれぞれ計算することが可能となる．

2.2.2 媒介中心性

次に，次式で定義されるノード u の媒介中心性について考える．

$$btw_G(u) = \frac{1}{(|V|-1)(|V|-2)} \sum_{v \in V, v \neq u} \left(\sum_{\substack{w \in V, w \neq u \\ w \neq v}} \frac{nsp_G(v,w;u)}{nsp_G(v,w)} \right) \tag{5}$$

ここで，$nsp_G(v,w)$ はグラフ G におけるノード v から w までの最短経路数，$nsp_G(v,w;u)$ はそのうちノード u を経由する最短経路の数を表す．直観的には，ノード u を経由する 2 ノード間の最短経路数が多いほど，u の媒介中心性 $btw_G(u)$ の値は大きくなる．この媒介中心性を求める標準的な方法としては，Brandes のアルゴリズム [2] が知られており，各ノード u に対して $btw_G(u)$ を求める計算量は近接中心性同様 $O(|E|)$ である．

いま，ノード u の真の媒介中心性の値 $btw_G(u)$ と，u を除く V の部分集合 T（$|T| = N$）から求められるその近似値 $btw_G(u;T)$ の近似誤差を 2.1 節の枠組みに基づいて推定することを考える．全サンプル集合 S に関しては，媒介中心性も全ノード集合ではなく個々のノードに対して定まる値であるため，近接中心性と同様に対象ノード u に対して $S_u = V \setminus \{u\}$ とする．一方，式 (5) は，以下のように書き換えることができる．

$$btw_G(u) = \frac{1}{|V|-1} \sum_{v \in V, v \neq u} \left(\frac{1}{|V|-2} btw_G(u;v) \right) \tag{6}$$

$$btw_G(u;v) = \sum_{\substack{w \in V, w \neq u \\ w \neq v}} \frac{nsp_G(v,w;u)}{nsp_G(v,w)}. \tag{7}$$

これより，$btw_G(u)$ は，ノード u 以外のノード v に対して求められる $btw_G(u;v)/(|V|-2)$ の平均と考えられる．したがって，$f_u(v) = btw_G(u;v)/(|V|-2)$ と定義することで，任意の部分集合 T に対するノード u の媒介中心性 $btw_G(u;T)$ を $(1/N)\sum_{v \in T} f_u(v)$ として求めることができ，式 (2)，(3) に従い $RE(N)$，$SE(N)$ をそれぞれ計算することができる．

3 信頼度つきギャップ検出法

本節では，ネットワーク $G = (V,E)$ が与えられたとき，ノード集合 V の部分集合 T（$|T| = N$）を用いて求めた各ノードの中心性指標の近似値に基づき，真に高い中心性指標値をもつノード集合を高精度で同定することを考える．そのためには，真の中心性の値とその近似値の近似誤差のみならず，近似値間の大小関係も精度よく推定することが必要となる．なぜなら，2 つのノードに対する中心性指標の近似値の大小関係は，真の値の下では逆転し得るからである．以下では，真の中心性の値に大小関係がある 2 つのノード間にはギャップがあるとし，ギャップとそのギャップをもつ 2 つのノードの大小関係を与えられた確信度の下で精度よく検出することを考える．

実際には，中心性の近似値の大小関係を精度よく推定するためには，その差に加え，各値の近似誤差を考慮する必要がある．その誤差を過大評価した場合，真の中心性の値の下でその大小関係が逆転する可能性が減少するため，ギャップ検出の精度は高くなる一方，再現率，すなわち検出できるギャップ数は減少することが予想される．逆に，近似誤差を過小に見積もった場合，再現率は向上するが精度が下がることが予想される．そのため，適切な近似誤差の推定が重要となる．

一方，統計的な観点からは，この問題は，与えられた確信度の下で各ノードの真の中性指標値の信頼区間を求め，その重複関係を調べることに相当する．いま，G 中のノード v に対するある中心性の真の値を $\mu_G(v)$，V の部分集合 T（$|T| = N$）から求められるその近似値を $\mu_G(v;T)$，両者の近似誤差の推定値を $\hat{\sigma}_v(N)$，α を $0 < \alpha < 1.0$ としたとき，信頼度 $100(1-\alpha)\%$ の下での $\mu_G(v)$ の信頼区間は以下の式で与えられる．

$$\mu_G(v;T) - z(\alpha)\cdot\hat{\sigma}_v(N) \le \mu_G(v) \le \mu_G(v;T) + z(\alpha)\cdot\hat{\sigma}_v(N) \quad (8)$$

ここで，$z(\alpha)$ は確率値 α に対して標準正規分布表から得られる値である．したがって，$\mu_G(v_i;T) > \mu_G(v_j;T)$ である 2 つの異なるノード v_i と v_j の中心性の真の値の信頼区間が重ならないことは，以下の関係が成り立つことを意味する．

$$\mu_G(v_i;T) - z(\alpha)\cdot\hat{\sigma}_{v_i}(N) > \mu_G(v_j;T) + z(\alpha)\cdot\hat{\sigma}_{v_j}(N) \quad (9)$$

この式を近似値の差に着目して変形すると次式が導ける．

$$\mu_G(v_i;T) - \mu_G(v_j;T) > z(\alpha)(\hat{\sigma}_{v_i}(N) + \hat{\sigma}_{v_j}(N)) \quad (10)$$

したがって，ノード v_i と v_j に対してこの関係が成り立つとき，信頼度 $100(1-\alpha)\%$ でこのノード間にはギャップが存在すると言える．

以上の考えに基づき，式 (10) の右辺を閾値 δ に一般化したギャップ検出手順を以下にまとめる．

1. V から N 個のノードをサンプリングし T を生成．
2. すべての $v \in V$ に対して $\mu_G(v;T)$ を計算．
3. $v \in V$ を $\mu_G(v;T)$ の降順にソートし，その結果を $v_1, v_2, \cdots, v_{|V|}$ とする（$\mu_G(v_1:T) \ge \mu_G(v_{|V|}:T)$）．
4. ギャップ集合 P を $P = \emptyset$ として初期化．
5. すべてのノードペア v_i，v_{i+1}（$i = 1, \cdots, |V|-1$）に対して，次の規則に従い P を更新する（ギャップ検出）．
   if $\mu_G(v_i:T) - \mu_G(v_{i+1};T) > \delta$ **then** $P = P \cup \{v_i\}$
6. P を出力して終了．

以下，閾値 δ を $\delta = z(\alpha)(RE_v(N) + RE_{v_{i+1}}(N))$ とするものを RE 法，$\delta = z(\alpha)(SE_v(N) + SE_{v_{i+1}}(N))$ とするものを SE 法と呼ぶ．また，比較のために，$\delta = 0$ とするナイーブ法も評価実験では用いる．ナイーブ法は，$\mu_G(v_i;T) \neq \mu_G(v_{i+1};T)$ のときには，必ずギャップを検出するため，すべてのギャップを検出できる一方，明らかに誤検出が多くなる．ナイーブ法，SE 法，RE 法の出力結果をそれぞれ P_{nv}，P_{se}，P_{re} とすると，その閾値 δ の値から，$P_{se} \subseteq P_{re} \subseteq P_{nv}$ という関係が成り立つ．次節では，これらの手法を実世界の社会ネットワークを用いて実験的に評価する．

4 評価実験

4.1 実験設定

前節で提案したギャップ検出法を，2 つの実ネットワークを用いて実験的に評価した．1 つは，化粧品に関する口コミサイトである“@cosme”[2] から抽出したお気に入りネットワークである．本実験では，2009 年 12 月の時点でランダムに選択したユーザを起点にお気に入りリンクを 10 ステップまで辿ることで生成したノード数 45,024，リンク数 351,299 の有向ネットワークを用いた．以下，コスメネットワークと呼ぶ．もう 1 つは，エンロン E メールデータセット [8] に対して，メールの送信者，もしくは受信者をノードとし，各メールの送信者と受信者間にリンクを生成したノード数 19,603，リンク数 210,950 の有向ネットワークである．以下，エンロンネットワークと呼ぶ．

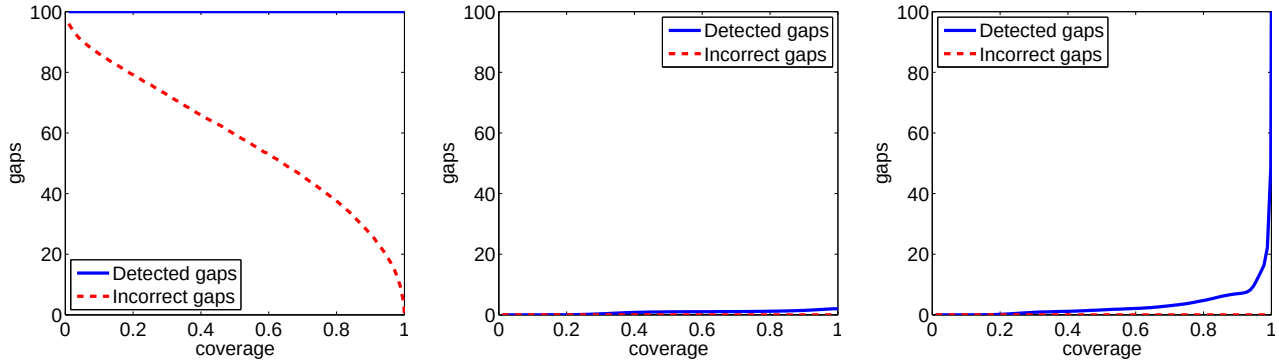
本実験では，これらのネットワーク中のノードのうち，すべてのノードから求めた真の中心性の値において上位 100 ノードを対象にナイーブ法，SE 法，RE 法の各手法を評価した．実験手順としては，ノード被覆率（coverage）を 0.01 から 1.0 まで 0.01 単位で変化させ，各被覆率において各手法が検出したギャップ数（検出数），およびその中で不正解であったギャップ数（不正解数）を調べた．ここでは，各手法の出力であるギャップ集合 P の要素 $v_k \in P$ に対して，その被覆率で計算した中心性の近似値上位 k 個のノード集合と，真の中心性の上位 k 個のノード集合が一致した場合にそのギャップを“正解”，そうでない場合を“不正解”とした．これは，ギャップはノード列の局所的な順序関係のみを表し，そのような局所的な順序関係の精度のみでは，高中心性ノード集合の同定という本来の目的に対する評価としては不十分なためである．実験では，これを $R = 1,000$ 回試行し，信頼度 95%（$\alpha = 0.05$）の下での各被覆率ごとのギャップ検出数，不正解数の平均を求めた．

4.2 実験結果

図 1 にコスメネットワークに対する近接中心性の結果を示す．グラフの横軸は被覆率であり，縦軸は各試行でのナイーブ法の検出数が 100 となるように各手法の検出数，不正解数を正規化した値の 1,000 回試行における平均値である．したがって，被覆率 c，r 回目の試行におけるナイーブ法の検出数を $D_{nv}(c,r)$，各手法の検出数を $D(c,r)$，不正解数を $I(c,r)$ としたとき，グラフ中の実線は次式の値を示し，

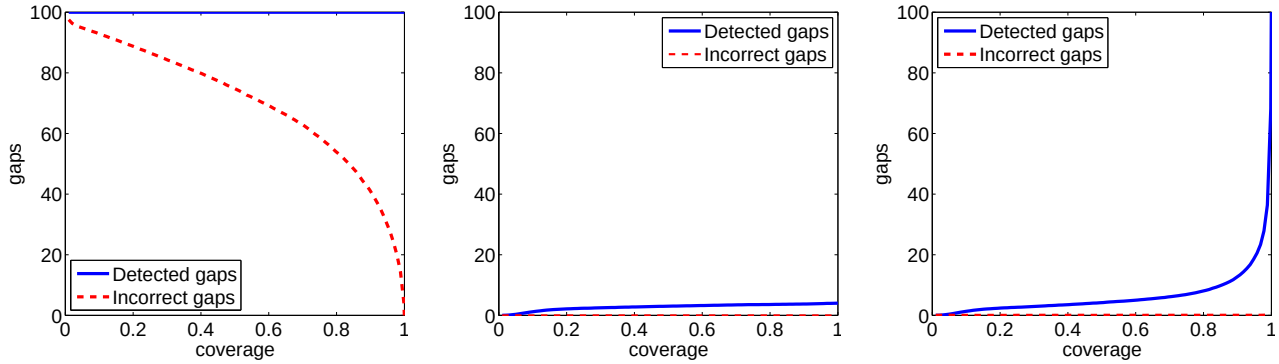
$$D(c) = \frac{1}{R}\sum_{r=1}^{R}\frac{D(c,r)}{D_{nv}(c,r)} \times 100 \quad (11)$$

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



(a) ナイーブ法　(b) SE 法　(c) RE 法

図 1: コスメネットワークに対する近接中心性の実験結果



(a) ナイーブ法　(b) SE 法　(c) RE 法

図 2: コスメネットワークに対する媒介中心性の実験結果

破線は次式の値を示していることになる．

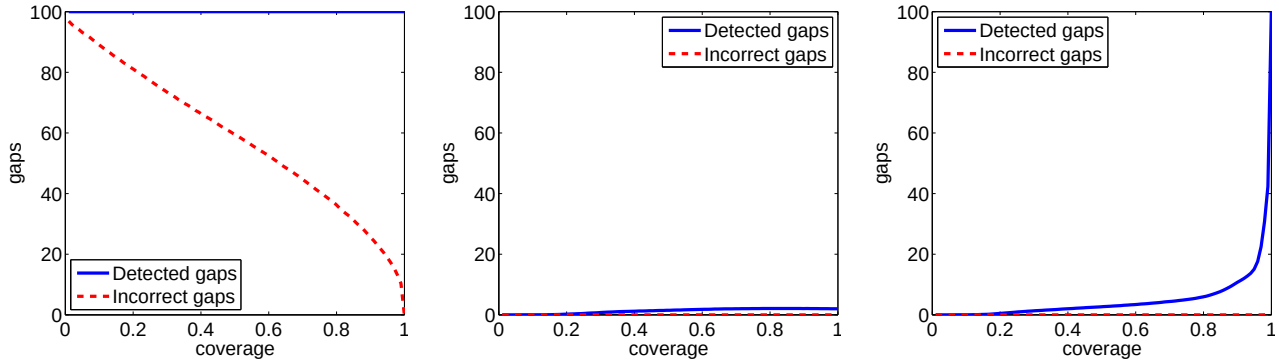
$$I(c) = \frac{1}{R}\sum_{r=1}^{R}\frac{I(c,r)}{D_{nv}(c,r)} \times 100 \qquad (12)$$

ここで，試行回数は $R = 1,000$ である．

各手法を比較すると，ナイーブ法は予想通り検出数はいずれの被覆率でも高いものの，不正解数も多い．被覆率が高くなるにつれて不正解数は減少するが，SE 法，RE 法と比較してその値は非常に大きい．一方，SE 法と RE 法に関しては，被覆率が 0.2 あたりまではほぼ同様の傾向を示し，両者ともに検出数はそれほど多くない．しかし，SE 法が被覆率が高くなっても検出数があまり多くならないのに比べて，RE 法は被覆率が高くなるにつれて検出数は増加し，0.9 を超えるあたりからその数は急増し，最終的には 100 となっている．これは，RE 法が用いるリサンプリング誤差が誤差範囲をより厳密に評価するのに対し，SE 法が用いる標準誤差は誤差範囲を過大評価する傾向にあるためである．そのため，被覆率が高くなり，中心性の近似値が真の値に近くなっても，SE 法における信頼区間は多くの場合重複してしまい，ギャップの検出が困難となっている．これに対して RE 法が用いるリサンプリング誤差は誤差範囲を厳密に評価するものの過小評価はしないため，被覆率が高くなるにつれて検出数が多くなる一方，不正解数はほとんんど増えず，ほぼ 0 という結果になっている．より安全な誤差範囲を想定する SE 法でも，不正解数に関しては同様の傾向となっている．不正解数が少ないことは，検出したギャップにより高中心性ノード集合が精度よく同定できていることを意味する．なお，同じ被覆率では，RE 法の検出数は SE 法の検出数以上となっている．これらの傾向は，図 2 に示す媒介中心性に対する結果にも共通しており，また，エンロンネットワークに対する結果を示す図 3，4 からも同様の傾向が読み取れる．

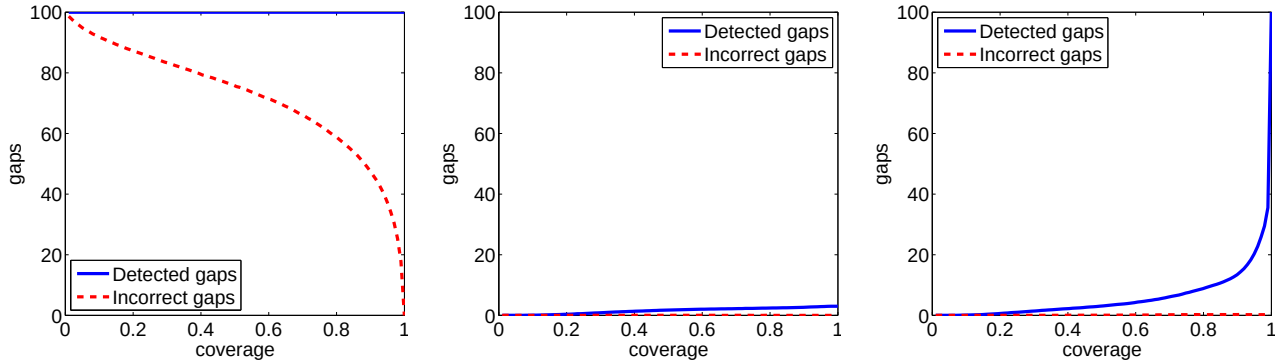
5　まとめ

本論文では，社会ネットワークにおけるノード中心性に関して，サンプリングした一部のノード集合から求められるその近似値のみを用いて，真の中心性の値が高いノード集合を高精度で同定する手法を提案した．提案法では，真の中心性の値とその近似値の近似誤差として我々が提案したリサンプリング誤差を用い，与えられた確信度の下での真の中心性の値の信頼区間を計算することで，高い精度でノード間のギャップを検出



(a) ナイーブ法　(b) SE 法　(c) RE 法

図 3: エンロンネットワークに対する近接中心性の実験結果



(a) ナイーブ法　(b) SE 法　(c) RE 法

図 4: エンロンネットワークに対する媒介中心性の実験結果

することを可能としている．実世界の社会ネットワークを用いた評価実験では，信頼度 95% の下，提案法が標準誤差を用いる手法より高い再現率でギャップを検出し，かつ検出したギャップにより高中心性ノード集合を精度よく同定できることを示した．

今後の課題としては，より大規模なネットワークにおける提案法の実験的評価，ならびに他のサンプリングに基づくアプローチとの比較が挙げられる．

謝辞

本研究は，科学研究費補助金基盤研究 (C)(No. 26330261) の補助を受けた．

参考文献

[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] Katz, L.: A new status index derived from sociometric analysis. Sociometry 18, 39–43 (1953)

[7] Kleinberg, J.: The convergence of social and technological networks. Communications of ACM 51(11), 66–72 (2008)

[8] 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)

[9] Newman, M.E.J.: Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. Physical Review E 64, 016132 (2001)

[10] Ohara, k., Saito, K., Kimura, M., and Motoda, H.: Resampling-based Framework for Estimating Node Centrality of Large Social Network, *Proc. of DS2014*, pp.228–239(2014).

[11] 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)