

ランダムウォークを用いた媒介中心性上位ノードの推定

中嶋 一貴[†] 岩崎 謙汰^{††} 松村 俊樹^{††} 首藤 一幸^{††}

[†] 東京工業大学 理学部 情報科学科

^{††} 東京工業大学 情報理工学院 数理・計算科学系

〒 152-8552 東京都目黒区大岡山 2-12-1-W8-43

あらまし 媒介中心性はネットワーク上におけるノードの重要度を表す指標の一つで、ソーシャルネットワークなどによく適用される。媒介中心性の上位 K ノードを推定する手法として厳密手法やランダムサンプリングを用いて推定する手法は既に存在する。しかし、ソーシャルネットワークはセキュリティやプライバシー保護の観点からネットワークの全体にランダムにアクセスすることはできないため、既存手法を適用することはできない。そのため、隣接ノードを順次辿っていく推定手法を考えなければならない。現時点で最も精度の優れた手法はランダムウォークで訪れたノードの次数から推定する手法である。本論文では媒介中心性に対して次数より相関が強いエゴ媒介中心性という統計量をランダムウォークしながら近似し、媒介中心性の上位ノードを推定する手法を提案する。

キーワード ソーシャルネットワーク, 媒介中心性, ランダムウォーク

1. はじめに

ソーシャルネットワーキングサービス (SNS) の利用者は年々増加しており、例えば Twitter^(注1) の利用者は 2017 年時点で 3 億人を超えている。このような巨大なソーシャルネットワークをグラフとして表して解析する研究が盛んに行われている。

ソーシャルネットワークにおいて、影響力の高いユーザを特定することはマーケティングなどをする上で非常に重要である [1, 2]。グラフにおけるノードの重要度を測るものが中心性であり、特に Freeman によって提案された媒介中心性 [3] はネットワークの分析において盛んに使用されている [4–6]。あるノードの媒介中心性は、グラフ内の他の 2 ノード間の最短経路上に含まれる割合によって定義され、グラフの広い範囲の情報を反映している。

本研究ではソーシャルネットワーク上の影響力の高いユーザを媒介中心性上位 K ノードとし、これらを推定することを目標とする。ネットワーク全体の情報が得られる状況では、各ノードの媒介中心性を厳密に計算する手法 [7]、ノードをランダムにサンプリングして近似する手法 [8–10] を適用することで上位ノードを求めることができる。しかし、ソーシャルネットワークではプライバシーといった理由から企業などが保持するネットワークの全ての情報を得られないことがほとんどであり、これらの手法を適用することはできない。この場合、隣接ノードを辿っていくクロールリングによってグラフから一部をサンプリングし、サンプリングした情報から上位ノードを推定する手法を考えなければならない。

グラフの全体の情報が得られない状況では、元のグラフの媒介中心性と相関が強い統計量をサンプリングしたノードやエッジから計算し、それを用いて推定することが有効である。Arun らはサンプリングしたノード集合から部分グラフを生成し、その

部分グラフ内の媒介中心性上位 K ノードを推定ノードとする手法を提案した [11]。Lim らはランダムウォークで訪れたノードを次数を記録し、サンプリングしたノードの次数上位 K ノードを推定ノードとする手法を提案した [12]。媒介中心性はグラフの広い範囲の情報をを用いるため、サンプリング率が低い状況では元のグラフと部分グラフのそれぞれの媒介中心性の相関が低くなる [13]。一方で、次数は最も局所的で媒介中心性と相関が強い統計量であるため、高い精度で媒介中心性上位ノードを推定することができる [12]。

本論文では、次数より媒介中心性と相関が強いエゴ媒介中心性 [14] を用いて推定する手法を提案する。エゴ媒介中心性はエゴネットワークにおけるエゴの媒介中心性である。提案手法ではランダムウォークで遷移したノードのエゴ媒介中心性の近似を行う。この近似により近似手法と同じコストで推定を行うことができ、また実験から既存の推定手法より優位な推定精度を実現することを示した。

本論文の構成は以下の通りである。2 章では関連研究について述べる。3 章では媒介中心性などの定義や他の統計量の相関、ランダムウォークによるサンプリングについて述べる。4 章では提案手法について述べる。5 章では提案手法の実験による評価と考察を行う。6 章では本研究のまとめと今後の課題を述べる。

2. 関連研究

クロールリングによる媒介中心性上位ノード推定の既存手法について述べる。上位ノードの推定には以下のような手順を踏む。

(1) クロールリングによりノードやエッジをサンプリングする。

(2) サンプリングしたノードやエッジの集合から媒介中心性上位ノードを推定する。

媒介中心性上位ノードを推定する手法を評価する上で、Lim らは以下の 2 つの指標を考えている [12]。

(注1) : <https://twitter.com/>

- 回収率: クローリングによるサンプリングで媒介中心性上位 K ノードを回収できた割合

- 重複率: 媒介中心性上位 K ノード集合と推定した上位 K ノード集合の重複している割合

回収率はクローリングによるサンプリング手法に、重複率は推定手法に依存する。このため、回収率と重複率を総合的に評価して手法を考える必要がある。

また、実社会のソーシャルネットワークでは、SNS の API へアクセスすることでクローリングで遷移したノードの ID とその隣接ノード集合の ID を取得できるが、多くの SNS では単位時間ごとにアクセスできる回数に制限がかかっている。このため本論文では、API へのアクセスの合計回数も考慮して手法を評価する。

2.1 サンプリング手法

クローリングによるサンプリング手法は数多く存在する。Arun ら、Lim らの研究において、ランダムウォークサンプリング、Random-walk sampling with uniform restarts (RWU)、Random-walk sampling with jumps (RWJ)、Frontier sampling (FS)、expansion sampling (XS)、Backlink Count (BLC)、OPIC algorithm (OPIC)、幅優先探索サンプリング (BFS)、深さ優先探索サンプリング (DFS)、Adjusted Expansion Sampling (AXS)、Randomized Expansion Sampling (RXS) が用いられており、総じてランダムウォークサンプリングの回収率が優れているとされている。よって、本研究ではランダムウォークサンプリングを用いてグラフの一部をサンプリングする。ランダムウォークサンプリングの詳細は 3 章で述べる。

2.2 媒介中心性上位ノードの推定手法

Arun ら、Lim らがそれぞれ提案した推定手法について述べる。Arun らはクローリングによるサンプリングで生成される部分グラフから推定する手法を提案した [11]。グラフ $G = (V, E)$ に対してクローリングでサンプリングしたノード集合を $V' \subset V$ 、エッジ集合を $E' = \{(v_i, v_j) \in E : v_i \in V', v_j \in V'\}$ とし、クローリングから生成されるグラフ G の部分グラフ G' を $G' = (V', E')$ と定義する。Arun らは、 G' の媒介中心性の上位 K ノードを推定ノードとしている。しかし、サンプリング率が 5% 未満といった状況では、元のグラフと部分グラフのそれぞれの媒介中心性の相関は低くなり [13]、推定精度は低い。

Lim らはクローリングの遷移時に次数を記録し、サンプリング終了時の次数の上位 K ノードを推定ノードとする手法を提案した [12]。以下、次数手法と呼ぶ。多くのネットワークでは次数と媒介中心性の相関が強いこと [15, 16]、クローリングの遷移時にノードの次数がわかることを利用しており、Arun らの手法より優れた精度で媒介中心性上位ノードを推定することができる。

我々は次数より媒介中心性と相関が強いエゴ媒介中心性を用いて媒介中心性上位ノードを推定する。5 章ではランダムウォークサンプリングで得られた同一のサンプルノード集合に対して、Lim らの次数手法と提案手法の重複率を比較し提案手法が優れていることを示す。

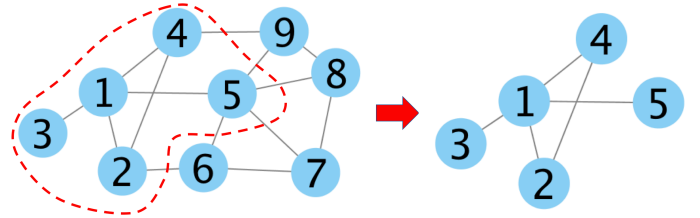


図 1: あるネットワークの 1 のエゴネットワーク

3. 準備

本章では、対象とするネットワークの前提、媒介中心性とエゴ媒介中心性の定義、各統計量と媒介中心性の相関、そしてランダムウォークサンプリングについて述べる。記法は表 1 にまとめた。

3.1 前提

本論文では静的なネットワークを無向グラフ $G = (V, E)$ で表す。グラフ G は連結、セルフループが存在しない、エッジは重みを持たない、多重辺を持たないとする。 $V = \{v_1, v_2, \dots, v_n\}$ はノード集合で E はエッジ集合であり、ノード数を n 、エッジ数を m とする。ノード v_i の隣接ノード集合を、 $N(i) = \{v_j \in V : (v_i, v_j) \in E\}$ とする。

また実社会のソーシャルネットワークでは、SNS の API へアクセスすることでクローリングで遷移したノードの ID や隣接ノード集合の ID を取得できる。このため、対象とするネットワークでは、クローリング以外で情報を取得することはできないものとする。

3.2 媒介中心性

あるノード v_i に対して v_i と異なる 2 個のノード v_j, v_k を取り、 v_i が v_j, v_k 間の最短経路上にあるとき v_i を通る最短経路数の割合に応じて加点される。例えば v_j, v_k 間に最短経路が 10 通りあり、このうち v_i を通る最短経路が 5 本あるとき $\frac{5}{10}$ を v_i に加点する。この加点を v_i と異なるすべてのノードの組 (v_j, v_k) について行う。 $\sigma_{j,k}$ は v_j, v_k 間の最短経路数を表し、 $\sigma_{j,k}(i)$ は v_j, v_k 間の最短経路のうち、ノード v_i を通る最短経路の個数とする。ただし、 $v_j = v_k$ のとき $\sigma_{j,k} = 0$ 、また $\frac{0}{0} = 0$ とする。このとき、グラフ G 上のノード v_i の媒介中心性 $C_B(i)$ は

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

で表される。

3.3 エゴ媒介中心性

エゴネットワークとはあるノードとその隣接ノードから成るネットワークである。エゴとはエゴネットワーク内で中心となっているノードをいう。例えば、図 1 の左のネットワークにおいて、1 をエゴとするエゴネットワークとはノード 1, 2, 3, 4, 5 からなる図 1 の右のようなネットワークでノード集合 $V = \{1, 2, 3, 4, 5\}$ 、 $E = \{(1, 2), (1, 3), (1, 4), (1, 5), (2, 4)\}$ とするグラフ $G = (V, E)$ で表せる。

v_i をエゴとするエゴネットワークにおける v_i の媒介中心性を

エゴ媒介中心性と呼ぶ。 v_i の媒介中心性を計算するときは 2 個のノード v_j, v_k を v_i 以外のすべてのノードを選んだが、エゴ媒介中心性を計算するときは v_j, v_k を v_i の隣接ノード集合のみに限定する。このとき、 G のノード v_i のエゴ媒介中心性は

$$eBC(i) = \sum_{v_j \in N(i)} \sum_{v_k \in N(i)} \frac{\sigma_{j,k}(i)}{\sigma_{j,k}} \quad (1)$$

と表せる。

ノード v_i, v_j に対して、 $N_i(j) = N(i) \setminus (N(j) \cup \{v_j\})$ とすると以下の式が成り立つ。

$$eBC(i) = \sum_{v_j \in N(i)} \sum_{v_k \in N_i(j)} \frac{1}{|N(j) \cap N(k)|} \quad (2)$$

Proof. 式 (1) の $\sigma_{j,k}(i)$ について考える。 $v_k \in N_i(j)$ のとき、 $(v_j, v_k) \notin E$ であり、多重辺が無いことから v_i を通る最短経路が 1 通り存在する。 $v_k \notin N_i(j)$ のとき、 $(v_j, v_k) \in E$ であり、 v_i を通る最短経路は存在しない。

よって、

$$\sigma_{j,k}(i) = \begin{cases} 1 & (v_k \in N_i(j)) \\ 0 & (otherwise) \end{cases}$$

となる。

次に $\sigma_{j,k}$ について考える。 $v_k \notin N_i(j)$ のとき v_j, v_k 間にエッジがあり、多重辺がないことから $\sigma_{j,k} = 1$ である。 $v_k \in N_i(j)$ のとき、 v_j, v_k の共通隣接ノード 1 つにつき最短経路が 1 通りあるので v_j, v_k の共通隣接ノード数を調べれば良い。 よって、

$$\sigma_{j,k} = \begin{cases} |N(j) \cap N(k)| & (v_k \in N_i(j)) \\ 1 & (otherwise) \end{cases}$$

よって、

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

これより、 (2) の式を得る。 $\square$

3.4 ランダムウォークサンプリング

ランダムウォークサンプリングは、あるノードの隣接ノードからランダムに 1 つノードを選択して遷移することを繰り返し遷移したノードをサンプリングする手法である。 $R = (x_1, x_2, \dots, x_r)$ を r ステップのランダムウォークで訪れたノードのインデックスの列とする。ランダムウォークのノード v_i からノード v_j への遷移確率 $p_{i,j}$ は、ノード v_i の次数を d_i と表すとすると

$$p_{i,j} = \begin{cases} \frac{1}{d_i} & (v_j \in N(i)) \\ 0 & (otherwise) \end{cases} \quad (3)$$

と表せる。事象 A が起こる確率を $Pr[A]$ と表すとすると、 r ステップ後のランダムウォークの分布は次のように表すことができる。

$$\pi' = (Pr[x_r = 1], Pr[x_r = 2], \dots, Pr[x_r = n])$$

ステップを十分多く重ねると、 $Pr[x_r = i]$ はある値 $\pi(i)$ に近づく。ベクトル $\pi = (\pi(1), \pi(2), \dots, \pi(n))$ は G の定常分布と呼ばれ、

$$\pi(i) = \frac{d_i}{2|E|} \quad (4)$$

である [17]。

本論文では、初期ノード v_{x_1} は G の定常分布から選ぶとする。実際に SNS 上でサンプリングを行う場合は初期ノードを定常分布から選ぶことは不可能であるため、この場合は初期ノードをランダムに取り、定常分布に収束するまで遷移を繰り返したのちにサンプリングを開始する。定常分布に収束するまでのステップ数 T を *mixing time* と呼び、パラメータ ϵ を用いて次のように定義される。

$$T(\epsilon) = \max_i \min_t \{t : |\pi - \pi(i)P^t|\}$$

ここで π はランダムウォークの定常分布、 P^t は t ステップ後の遷移確率行列である。実際のソーシャルネットワークの *mixing time* は $\Theta(\log^2 n)$ であることが知られている。 [18]

3.5 各統計量と媒介中心性の相関

多くの関連研究ではある統計量と媒介中心性の相関をピアソンの相関係数で論じている [12, 14, 19]。しかし、本研究では媒介中心性上位 K ノードに関心があり、下位ノードの相関まで反映するピアソンの相関係数は適切とは言えない。

Lim らの次数手法では、グラフ全体における各統計量の上位ノードと媒介中心性上位ノードの要素がどれだけ重複しているかが精度に影響する。例えば次数手法では、グラフ全体における次数上位 K ノードと媒介中心性上位 K ノードが全て一致していれば高い精度で推定が可能である。なぜならランダムウォークの定常分布 (式 (4)) から次数が高いノードはサンプリングされやすく、サンプルされたノードの次数が正しい値で算出できるからである。このためグラフ全体における次数上位 K ノードと媒介中心性上位 K ノードの重複率は次数と媒介中心性の相関を表す良い指標となり得る。

しかし、上位 K ノードをすべて確実にサンプリングできるわけではない。例えばあるグラフ全体における媒介中心性上位 5 ノードが順位順に $\{1, 2, 3, 4, 5\}$ 、次数上位 10 ノードが $\{1, 2, 3, 6, 7, 4, 5, 8, 9, 10\}$ とするとそれぞれの上位 5 ノードの重複率は $\frac{3}{5}$ である。しかし、サンプリング時にノード 6, 7 がサンプリングされずノード 1, 2, 3, 4, 5 がサンプリングされた場合、サンプルノード集合内の次数上位 5 ノードは 1, 2, 3, 4, 5 となり重複率は 1 となる。このため K より少し幅を広げて、媒介中心性上位 K ノード集合と次数上位 m ノード集合 ($m \geq K$) の重複率を考慮する必要がある。エゴ媒介中心性と媒介中心性との相関についても同様である。

これらを考慮して新しい指標を定義する。グラフ全体における媒介中心性上位 K ノードの集合を $top_{bc}(K)$ 、次数上位 K_1 ノードの集合を $top_d(K_1)$ 、エゴ媒介中心性上位 K_2 ノードの集合を $top_{eBC}(K_2)$ とする。ある定数 m を用いて次数とエゴ媒

表 1: 記法

$G(V, E)$	重みなしの無向グラフ
V, E	ノード集合, エッジ集合
n, m	ノード数, エッジ数
K	推定上位ノード数 ($K \ll n$)
v_i	G のノード
d_i	ノード v_i の次数
$N(i)$	ノード v_i の隣接ノード集合
$N_i(j)$	$N(i) \setminus (N(j) \cup \{v_j\})$
$\sigma_{j,k}$	v_j と v_k 間の最短経路数
$\sigma_{j,k}(i)$	$\sigma_{j,k}$ のうち, v を通る経路数
$C_B(i)$	v_i の媒介中心性
$eBC(i)$	v_i の エゴ媒介中心性
V'	ランダムウォークでサンプリングしたノード集合
E'	$\{(v_i, v_j) \in E : v_i \in V', v_j \in V'\}$
G'	V', E' から生成される部分グラフ
R	ランダムウォーク
$I(i)$	$\{l : x_l = i, 2 \leq l \leq r-1\}$
$\widetilde{eBC}(i)$	$eBC(i)$ の近似値

介中心性それぞれと媒介中心性の相関を表す指標 $correlation_d$, $correlation_{eBC}$ を以下のように定義する.

$$correlation_d = \frac{\sum_{i=K}^m \frac{|top_{bc}(K) \cap top_d(i)|}{K} * \frac{1}{i}}{\sum_{i=K}^m \frac{1}{i}}$$

$$correlation_{eBC} = \frac{\sum_{i=K}^m \frac{|top_{bc}(K) \cap top_{eBC}(i)|}{K} * \frac{1}{i}}{\sum_{i=K}^m \frac{1}{i}}$$

$\frac{|top_{bc}(K) \cap top_d(i)|}{K}$ は媒介中心性上位 K ノード集合と次数上位 i ノード集合の重複率である. $\frac{1}{i}$ を掛けているのは同じ重複率であっても i が大きくなるほど値が小さくなるようにするためである. 分母は規格化定数である. 本論文ではこれらを用いて次数とエゴ媒介中心性のそれぞれの媒介中心性との相関を論じる.

4. 提案手法

本章では提案手法の概要について述べる. 次数手法では, ランダムウォークで訪れたノードの次数から媒介中心性の上位ノードを推定した. しかしあるノードの次数がグラフ内で上位であっても, そのノードのエゴネットワークが完全グラフとなるとき媒介中心性は 0 となる [9]. このようなノードのエゴ媒介中心性は 0 となるため, エゴ媒介中心性を用いることで次数は高いが媒介中心性が低いノードを推定する可能性を減らすことができる. また, エゴネットワーク内で加点を行うため次数が高く媒介中心性が高いノードはエゴ媒介中心性も高くなりやすい. このためエゴ媒介中心性は次数より媒介中心性と相関が強くなると考えられる.

Algorithm 1 提案手法

Require: グラフ $G = (V, E)$, 目標サンプルノード数 n' , 上位ノード数 K

Ensure: 推定上位 K ノード集合

$v \leftarrow G$ の定常分布から最初のノードを選ぶ

$i \leftarrow 1$

$V' \leftarrow$ 空集合

$topk \leftarrow$ 空集合

$n1 \leftarrow$ 空集合

$n2 \leftarrow$ 空集合

$n3 \leftarrow$ 空集合

while $|V'| < n'$ **do**

V' に v を追加

$w_i \leftarrow v$

$n1 \leftarrow n2$

$n2 \leftarrow n3$

$n3 \leftarrow N(v)$

$w \leftarrow v$ の隣接ノードから一様を選ぶ

$v \leftarrow w$

if $2 < i$ **then**

if w_{i-1} に初めて訪れた **then**

$s_{w_{i-1}} \leftarrow 0$

$t_{w_{i-1}} \leftarrow 0$

end if

$s_{w_{i-1}} \leftarrow s_{w_{i-1}} + 1$

if $w_{i-2} \neq w_i$ and $w_{i-2} \notin n3$ **then**

$\sigma \leftarrow |n1 \cap n3|$

$d \leftarrow |n2|$

$t_{w_{i-1}} \leftarrow t_{w_{i-1}} + \frac{d^2}{\sigma}$

end if

$i \leftarrow i + 1$

end if

end while

for all $v \in V'$ **do**

$\widetilde{eBC}(v) \leftarrow \frac{t_v}{s_v}$

end for

for $i = 1$ to K **do**

$v \leftarrow \arg \max_{v \in V'} \widetilde{eBC}(v)$

$topk$ に v を追加

$V' \leftarrow V' \setminus \{v\}$

end for

return $topk$

4.1 ランダムウォークを用いたエゴ媒介中心性の近似計算

ランダムウォークを用いたエゴ媒介中心性の近似計算について述べる. $R = (x_1, x_2, \dots, x_r)$ を r ステップのランダムウォークで訪れたノードのインデックスの列とする. ただし, 初期ノード v_{x_1} はグラフ G の定常分布 $\pi = (\pi(1), \pi(2), \dots, \pi(n))$ から選ぶとする. 各ノード v_i について重複を許す集合 $I(i)$ を以下のように定義する.

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

ノード v_i, v_j に対して, $N_i(j) = N(i) \setminus (N(j) \cup \{v_j\})$ とし, $l \in I(i)$ に対して変数 $\phi_l(i)$ を

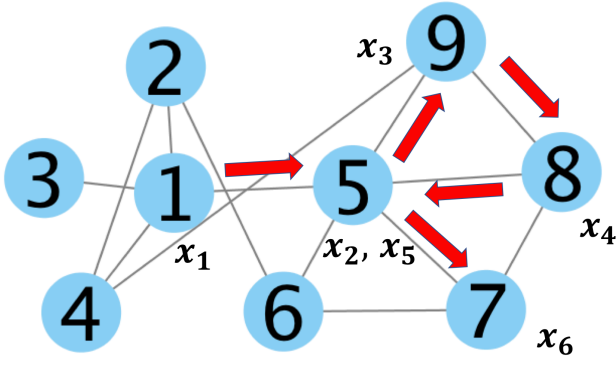


図 2: ランダムウォークの例

$$\phi_l(i) = \begin{cases} \frac{d_i^2}{|N(x_{l-1}) \cap N(x_{l+1})|} & (v_{x_{l+1}} \in N_i(x_{l-1})) \\ 0 & (\text{otherwise}) \end{cases}$$

とする。ノード v_i の $eBC(i)$ の近似値 $e\widetilde{BC}(i)$ を以下のように定義する。

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

ただし $I(i) = \phi$ であるノードは $\phi_l(i), e\widetilde{BC}(i)$ を定義しない。例えば、図 2 のネットワークで $v_i = i$ ($1 \leq i \leq 9$) とする。6 ステップのランダムウォークを行い、インデックスの列 $R = (1, 5, 9, 8, 5, 7)$ が得られたとする。このとき、 $I(5) = \{2, 5\}, I(8) = \{4\}, I(9) = \{3\}, I(1) = I(2) = I(3) = I(4) = I(6) = I(7) = \phi$ である。 $e\widetilde{BC}(5)$ について考える。 $I(5) = \{2, 5\}$ である。 $l = 2$ のとき、 $i = x_2 = 5, v_{x_1} = 1, v_{x_3} = 9$ であり、 $N_5(x_1) = \{6, 7, 8, 9\}$ より $v_{x_3} \in N_5(x_1)$ である。 $N(x_1) \cap N(x_3) = \{4, 5\}, d_5 = 5$ から、 $\phi_2(5) = \frac{d_5^2}{|N(x_1) \cap N(x_3)|} = \frac{25}{2}$ である。また、 $l = 5$ のとき、 $i = x_5 = 5, v_{x_4} = 8, v_{x_6} = 7$ であり、 $v_{x_6} \notin N_i(x_4)$ であるから、 $\phi_5(5) = 0$ である。よって、 $e\widetilde{BC}(5) = \frac{1}{2}(\phi_2(5) + \phi_5(5)) = \frac{25}{4}$ となる。

$\phi_l(i)$ の定義から各 l ステップ目 ($2 \leq l \leq r-1$) に対して $l-1$ ステップ目と $l+1$ ステップ目の隣接ノードの情報が必要である。提案手法では各ステップで 1 ステップ前と 1 ステップ後の隣接ノードを記憶しておくことで API への追加のアクセスを無くし、推定に必要なアクセス回数を Lim らの次数手法と等しくなるようにしている。

ステップ数 r を増やしていくと $e\widetilde{BC}(i)$ は $eBC(i)$ に収束することを示す。大数の法則より $e\widetilde{BC}(i)$ の期待値が式 (3) に等しいことを示せばよい。すなわち

$$E[e\widetilde{BC}(i)] = \sum_{v_j \in N(i)} \sum_{v_k \in N_i(j)} \frac{1}{|N(j) \cap N(k)|}$$

を示す。

Proof. 事象 B における事象 A の条件付き確率を $Pr[A|B]$ と表すことにする。ランダムウォークは隣接ノードを辿っていくことと $I(i)$ の定義より v_j, v_k は v_i の隣接ノードであることから

$$\begin{aligned} E[e\widetilde{BC}(i)] &= \sum_{v_j \in N(i)} \sum_{v_k \in N_i(j)} Pr[x_{l-1} = j, x_{l+1} = k | x_l = i] E[\phi_l(i) | x_{l-1} = j, x_{l+1} = k] \\ &= \sum_{v_j \in N(i)} \sum_{v_k \in N_i(j)} Pr[x_{l-1} = j, x_{l+1} = k | x_l = i] \frac{d_i^2}{|N(j) \cap N(k)|} \quad (5) \end{aligned}$$

$Pr[x_{l-1} = j, x_{l+1} = k | x_l = i] = \frac{1}{d_i^2}$ を示す。条件付き確率の定義より

$$Pr[x_{l-1} = j, x_{l+1} = k | x_l = i] = \frac{Pr[x_{l-1} = j, x_l = i, x_{l+1} = k]}{Pr[x_l = i]} \quad (6)$$

ランダムウォークはマルコフ連鎖であるからその定義より

$$Pr[x_{l-1} = j, x_l = i, x_{l+1} = k] = p_{i,k} p_{j,i} Pr[x_{l-1} = j]$$

ランダムウォークの遷移確率と定常分布 (式 (3), (4)) より

$$\begin{aligned} p_{i,k} p_{j,i} Pr[x_{l-1} = j] &= \frac{1}{d_i} \frac{1}{d_j} \frac{d_j}{2|E|} \\ &= \frac{1}{2d_i|E|} \end{aligned}$$

また定常分布 (式 (4)) から $Pr[x_l = i] = \frac{d_i}{2|E|}$ より、式 (6) から $Pr[x_{l-1} = j, x_{l+1} = k | x_l = i] = \frac{1}{d_i^2}$ が成り立つ。よって式 (5) は

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

ステップ数 r を増やすことで $|I(i)|$ も増えていくため、近似値 $e\widetilde{BC}(i)$ は $eBC(i)$ に収束する。□

また、実装のアルゴリズムを Algorithm 1 に示した。

5. 実験

複数のネットワークに対して提案手法の精度を評価した。

5.1 データセット

Stanford Network Analysis Project (SNAP) のデータセット [20] で公開されている複数のネットワーク、複雑ネットワークの生成モデルの 1 つである Barabasi-Albert モデル (BA モデル) [21] を用いて実験を行なった。有向グラフは、各辺に対して反対側を向く辺を追加し無向グラフとして扱った。非連結なグラフは最大連結成分をグラフとして扱った。また、ループ、多重辺は削除し、辺の重みは全て 1 とした。表 2 に各ネットワークの特徴をまとめた。

5.2 実験

与えられたグラフの媒介中心性上位 10 ノードを推定する。各グラフであるノード数をランダムウォークサンプリングし、その同一のサンプルノード集合に対して、回収率、Lim らの次数手法 [12] と提案手法の重複率をそれぞれ算出することを 1000 回行った。同一のサンプルノード集合に対して各手法で推定するため回収率に差は生じない。サンプルノード数は n の 1%, 1.5%, 2%, 2.5%, 3% とした。ランダムウォークの初期ノード

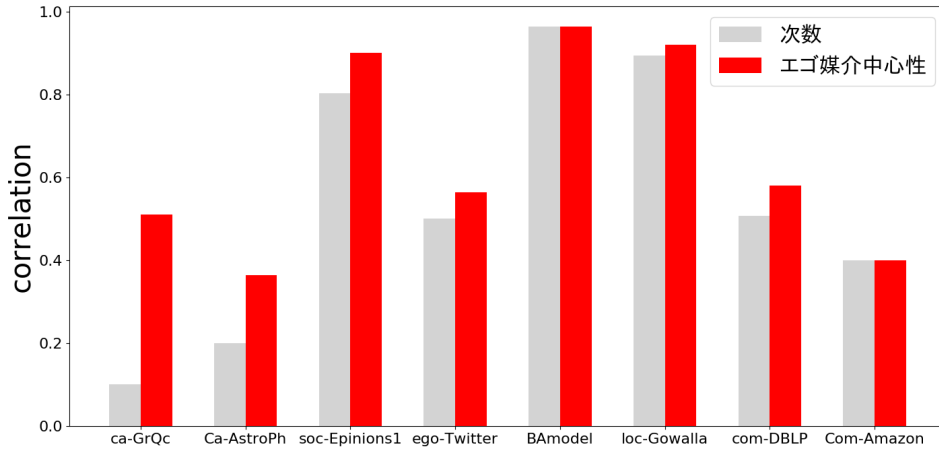


図 3: 各グラフの $correlation_d$, $correlation_{eBC}$

表 2: データセットの概要

ネットワーク	種類	ノード数	エッジ数
ca-GrQc	共著ネットワーク	4158	13422
ca-AstroPh	共著ネットワーク	17903	196972
soc-Epinions1	ソーシャルネットワーク	75877	405739
ego-Twitter	ソーシャルネットワーク	81306	1342296
BAmode1	ネットワーク生成モデル	100000	199996
loc-Gowalla	ソーシャルネットワーク	196591	950327
com-DBLP	共著ネットワーク	317080	1049866
com-Amazon	商品ネットワーク	334863	925872

はグラフの定常分布から選んだ。回収率、重複率の定義は以下の通りである。

- **回収率**: ランダムウォークサンプリングで媒介中心性上位 10 ノードを回収できた割合
- **重複率**: 媒介中心性上位 10 ノード集合と推定した上位 10 ノード集合の重複している割合

まず、図 3 に各グラフの $correlation_d$, $correlation_{eBC}$ (それぞれ図 3 の灰色、赤色) を示した。図 3 よりエゴ媒介中心性の相関が次数より高いことがわかる。本論文では次数とエゴ媒介中心性のそれぞれの媒介中心性との相関の指標として $correlation_d$, $correlation_{eBC}$ を以下のように定義している。

$$correlation_d = \frac{\sum_{i=10}^{20} \frac{|top_{bc}(10) \cap top_d(i)|}{10} * \frac{1}{i}}{\sum_{i=10}^{20} \frac{1}{i}}$$

$$correlation_{eBC} = \frac{\sum_{i=10}^{20} \frac{|top_{bc}(10) \cap top_{eBC}(i)|}{10} * \frac{1}{i}}{\sum_{i=10}^{20} \frac{1}{i}}$$

図 4 に各グラフにおける回収率 (図 2 の青い線) と、次数手法 (図 2 の赤い線) と提案手法 (図 2 の黒い点線) の重複率の 1000

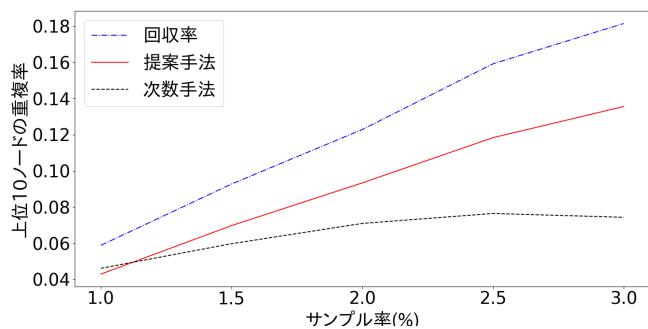
回の平均値を示した。重複率の最大値は回収率でありこの値に近い手法ほど推定精度が優れている。提案手法は、ca-GrQc, ca-AstroPh, soc-Epinions1, ego-Twitter, loc-Gowalla, com-DBLP において優位な重複率を示し、BAmode1, com-Amazon において同程度の重複率を示した。またグラフごとの重複率の優劣が図 3 の correlation の優劣と一致しており $correlation_d$, $correlation_{eBC}$ が推定精度を論じるに適切な指標であることが言える。またランダムウォークの定常分布 (式 (4)) より次数が高いノードに訪れやすいため、ca-GrQc や ca-AstroPh といった $correlation_d$ が低いグラフでは回収率が低いこともわかる。

6. まとめと今後の課題

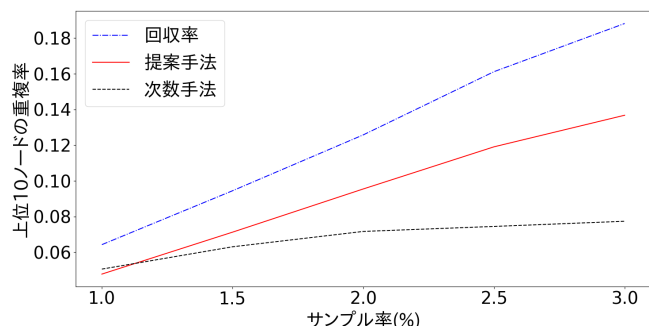
本論文ではクロールベースの新たな媒介中心性上位ノード推定手法を提案した。次数より媒介中心性と相関が強いエゴ媒介中心性をランダムウォークしながら近似計算することで、近似手法と同じコストかつ優位な推定精度を達成することができた。また、次数、エゴ媒介中心性それぞれと媒介中心性の相関の新しい指標を提案した。この指標は媒介中心性上位ノードに焦点を当てた指標で、回収率の高低、次数手法と提案手法の重複率の差を論じることができた。API へのアクセスが制限されサンプルノード数が限られるオンラインソーシャルネットワークでは、エゴ媒介中心性といった局所的で媒介中心性と相関が強い統計量から媒介中心性上位ノードを推定することは 1 つの有効な手段と言える。今後は有向グラフや多重辺、ループがある場合にも対応できる手法を考えることを課題とする。

謝 辞

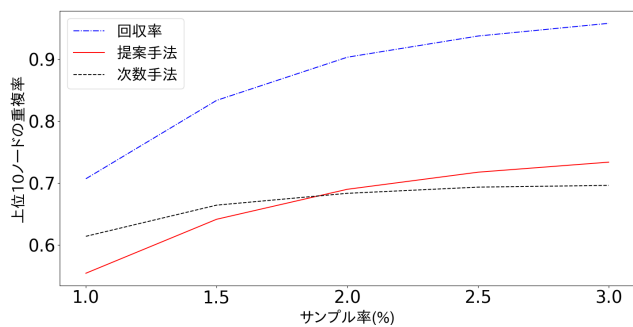
本研究の一部は、国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務として行われました。本研究は JSPS 科研費 25700008 および 16K12406 の助成を受けたものです。



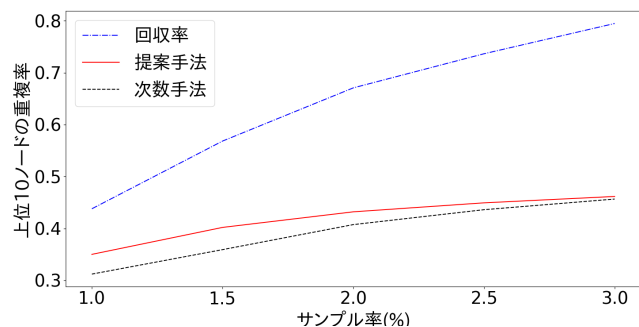
(a)ca-GrQc



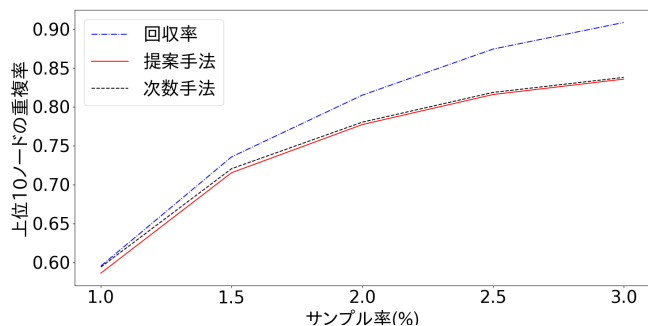
(b)ca-AstroPh



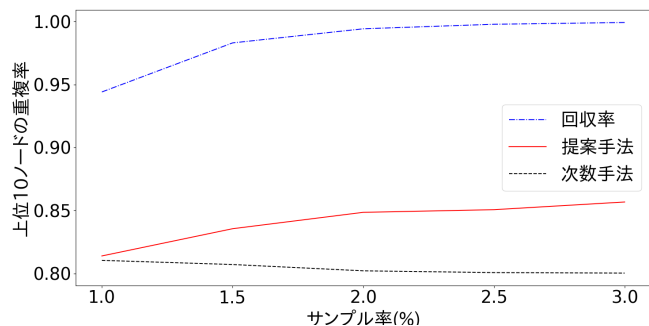
(c)soc-Epinions1



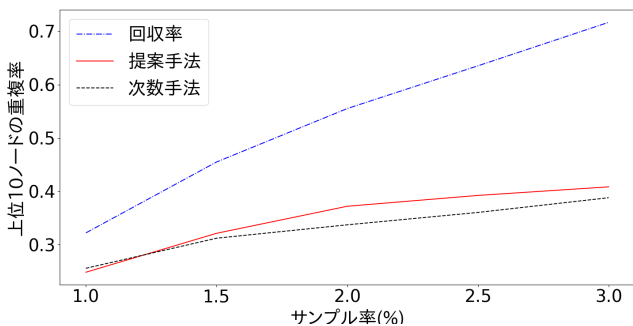
(d)ego-Twitter



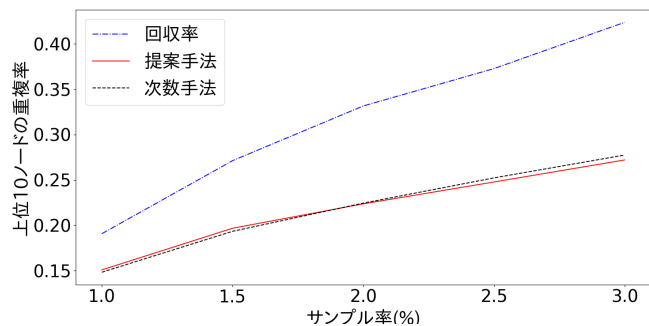
(e)BAmodel



(f)loc-Gowalla



(g)com-DBLP



(h)com-Amazon

図 4: 各グラフにおける次数手法, 提案手法の重複率

- [1] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146. ACM, 2003.
- [2] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twit-terrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pp. 261–270. ACM, 2010.
- [3] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pp. 35–41, 1977.
- [4] Maliackal Poulo Joy, Amy Brock, Donald E Ingber, and Sui Huang. High-betweenness proteins in the yeast protein interaction network. *BioMed Research International*, Vol. 2005, No. 2, pp. 96–103, 2005.
- [5] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, Vol. 69, No. 2, p. 026113, 2004.
- [6] Roger Guimera and Luis A Nunes Amaral. Modeling the world-wide airport network. *The European Physical Journal B-Condensed Matter and Complex Systems*, Vol. 38, No. 2, pp. 381–385, 2004.
- [7] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, Vol. 25, No. 2, pp. 163–177, 2001.
- [8] Ulrik Brandes and Christian Pich. Centrality estimation in large networks. *International Journal of Bifurcation and Chaos*, Vol. 17, No. 07, pp. 2303–2318, 2007.
- [9] David A Bader, Shiva Kintali, Kamesh Madduri, and Milena Mihail. Approximating betweenness centrality. In *International Workshop on Algorithms and Models for the Web-Graph*, pp. 124–137. Springer, 2007.
- [10] Yuichi Yoshida. Almost linear-time algorithms for adaptive betweenness centrality using hypergraph sketches. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1416–1425. ACM, 2014.
- [11] Arun S Maiya and Tanya Y Berger-Wolf. Online sampling of high centrality individuals in social networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 91–98. Springer, 2010.
- [12] Yeon-sup Lim, Daniel S Menasché, Bruno Ribeiro, Don Towsley, and Prithwish Basu. Online estimating the k central nodes of a network. In *Network Science Workshop (NSW), 2011 IEEE*, pp. 118–122. IEEE, 2011.
- [13] Stephen P Borgatti, Kathleen M Carley, and David Krackhardt. On the robustness of centrality measures under conditions of imperfect data. *Social networks*, Vol. 28, No. 2, pp. 124–136, 2006.
- [14] Martin Everett and Stephen P Borgatti. Ego network betweenness. *Social networks*, Vol. 27, No. 1, pp. 31–38, 2005.
- [15] K-I Goh, Eulsik Oh, Byungnam Kahng, and Doochul Kim. Betweenness centrality correlation in social networks. *Physical Review E*, Vol. 67, No. 1, p. 017101, 2003.
- [16] Marc Barthelemy. Betweenness centrality in large complex networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, Vol. 38, No. 2, pp. 163–168, 2004.
- [17] David Aldous and Jim Fill. Reversible markov chains and random walks on graphs.
- [18] Stephen J Hardiman and Liran Katzir. Estimating clustering coefficients and size of social networks via random walk. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 539–550. ACM, 2013.
- [19] Jürgen Pfeffer and Kathleen M Carley. k-centralities: Local approximations of global measures based on shortest paths. In *Proceedings of the 21st International Conference on World Wide Web*, pp. 1043–1050. ACM, 2012.
- [20] Stanford large network dataset collection. <https://snap.stanford.edu/data/>.
- [21] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, Vol. 286, No. 5439, pp. 509–512, 1999.