ソーシャルグラフ向けクラスタ係数推定手法の効率化

岩崎　謙汰†　　華井　雅俊††　　首藤　一幸††

† 東京工業大学 理学部 情報科学科
†† 東京工業大学 大学院情報理工学研究科 数理・計算科学専攻
〒 152-8552 東京都目黒区大岡山 2-12-1-W8-43
E-mail: †iwasaki.k.ah@m.titech.ac.jp

あらまし　ソーシャルネットワーク全体の特徴はクラスタ係数といった指標で表すことができる．現実のソーシャルネットワークサービスでは，グラフ全体を入手することは現実的ではないため，オンラインでサンプリングした一部分から指標を算出する他ない．オンラインでのサンプリングでは，頂点や辺の収集はグラフをクロールして行うこととなる．特にクラスタ係数の推定ではランダムウォークベースの手法が有効である．現在最も精度が高いクラスタ係数推定手法は，Hardiman らが提案した Simple random walk (SRW) を利用した手法である．しかし SRW では一つ前のノード戻る遷移が精度を下げる原因となる．本研究では既存手法の問題点に着目し，一つ前のノードに戻らないランダムウォーク Non-backtracking random walk (NBRW) を利用したクラスタ係数推定手法を提案した．そして実験により提案手法が既存手法より効率的かつ高精度であることを示した．

キーワード　ソーシャルネットワーク，グラフサンプリング，クラスタ係数

1. はじめに

ソーシャルネットワークの利用者は増加を続け，例えば Facebook の利用者は 10 億人に達した (注1)．利用者が増えるに従い，ソーシャルネットワークをグラフ構造として分析することの関心が高まっている．例えばユーザをノード，ユーザの友達関係をエッジとするグラフが考えられる．

ソーシャルネットワークの分析において，複雑ネットワークの特徴の一つであるグラフのクラスタ係数は重要な指標の一つである．クラスタ係数は特定の一つのノードに対して定義できる値であり，隣接ノード間に辺が張られている割合によって表される指標である．グラフの全ノードのクラスタ係数の平均が高いほどそのネットワークは密であると言える．クラスタ係数はネットワークの分析に盛んに使用されている [1–3].

しかし大規模なソーシャルネットワークに対して，計算量の問題から平均クラスタ係数を厳密に計算することは困難である．具体的には頂点数 n に対して $O(n^3)$ の計算量 [4] がかかる．そのため，グラフの全体から一部分のノードとエッジを取得し特徴を抽出するグラフサンプリングによってクラスタ係数を推定する手法が盛んに研究されてきた．

ソーシャルネットワークに対するグラフサンプリング手法を考える場合には，グラフデータを保有する企業がプライバシーに関する規律や法律を定めているため，ネットワークのトポロジ情報が事前に与えられないことがあることを考慮しなければならない．同様の理由でネットワークの全ての情報にアクセスすることもできない．例えばネットワークの全体のノード数すら事前に知ることができないソーシャルネットワークがほとんどである [5]．それゆえ [6] で紹介されている，ノードを一様独立に選択しサンプリングする手法は現実的ではない [7]．したがって隣接ノードの情報を元にノードの隣接関係を辿っていくクローリングによるグラフサンプリング手法を考える必要がある [6,8]．現実のソーシャルネットワークでも，幅優先サンプリング (BFS) やランダムウォークなどのクローリングによるグラフサンプリング手法が実際に使われている [1–3, 9–11].

我々が知る限り，グラフサンプリングによるクラスタ係数推定手法で最も効率的かつ高精度な手法は Hardiman ら [12] が提案した Simple random walk (SRW) を行いながら三角形を数える手法である [13]．しかし，SRW は一つ前に訪問したノードに遷移する可能性があるため，三角形を数える計算に無駄が生じ推定精度を下げる原因となる．したがって一つ前に訪問したノードに戻らないランダムウォークである Non-backtracking random walk (NBRW) によって三角形を数える手法が有効であると考えられる．また NBRW はその性質から SRW よりも一定のノード数をサンプルするのに必要なステップ数が少なくなることが期待できる．

本研究は NBRW をしながら三角形を数えるクラスタ係数推定手法を提案した．そして実験結果から既存手法より効率的かつ高精度なクラスタ係数推定手法であることを示した．

本論文の構成は以下の通りである．2 章では関連研究を紹介する．3 章では本論文で使用する表記や数式の定義，サンプリング手法の紹介を行う．4 章では我々の提案手法を述べる．5 章では提案手法の評価実験を行う．6 章では本研究のまとめと今後の課題について述べる．

2. 背　　景

本章では，既存のクローリングによるサンプリング手法とクラスタ係数推定手法について述べる．

(注1)：http://www.adweek.com/socialtimes/3q-2014-earnings/438899

2.1 クローリングによるサンプリング手法

クローリングによるグラフサンプリング手法は大きく分けてに 2 種類に分類できる．走査ベースによるサンプリング手法 [6, 8, 14] とランダムウォークベースによるサンプリング手法 [15, 16] である．

2.1.1 走査ベースのグラフサンプリング

走査ベースのサンプリング手法は，幅優先サンプリング (BFS)，深さ優先サンプリング (DFS)，などが挙げられる．特に BFS は基礎的であり最も広く使われている手法である [1–3]．一つの理由として，BFS はグラフの特定の部分を全て収集してくれる点にある．しかし，BFS は次数が高いノードに収集が偏りやすいことがわかっている [17]．さらに，どのように収集するノードの偏るのか分析できないため，クラスタ係数や次数分布などのグラフの構造的特徴を求める場合には不向きである [5, 18]．

2.1.2 ランダムウォークベースのグラフサンプリング

ランダムウォークとは，まず最初に一つもしくは複数のノードを恣意的に選択し，現在いるノードから隣接しているノードに対して確率的に一つのノードを選択し次のノードとして移動する動作を繰り返すことによってノードを収集する手法のことである．次のノードを選ぶ時に一様ランダムに選択するランダムウォークを Simple random walk (SRW) と呼ぶ．SRW は最も基本的なランダムウォークである．SRW は BFS 同様次数が高いノードに収集が偏りやすいサンプリング手法である．しかしマルコフ連鎖による解析から，ノードの訪問確率はそのノードの次数に比例することが解明されている [19]．それに対し Metropolis-Hasting (MH) [20] アルゴリズムを使用し，定常分布が一様分布になるようにノード選択の確率を調整したランダムウォークが Metropolis-Hasting random walk (MHRW) である．MHRW は一様にサンプリングするので理想的と思えるが，ノード遷移を行う時に次数が高いノードを選択した場合は遷移を拒否する可能性があり，SRW より多くのグラフ情報を取得しなければならないという欠点がある [21] [13]．

Non-backtracking random walk (NBRW) は次のノードに遷移する時に，一つ前に訪問したノード以外の隣接ノードから一様ランダムにノードを選択するランダムウォークである．NBRW は一つ前のステップのノードに遷移確率が依存するため，ノードの状態空間ではマルコフ連鎖ではない．しかしエッジの状態空間とするとマルコフ連鎖を作ることができるため，NBRW による推定手法が可能となる [21]．その定常分布は SRW と同じである．また NBRW による推定は SRW による推定より分散が小さくなる [21]．

2.2 クラスタ係数推定手法

クローリングによるクラスタ係数推定手法は [2, 22] で提案されている．Ribeiro ら [22] は SRW によるノード収集を行い，Gjoka ら [2] は MHRW によるノード収集を行う手法を提案した．これらの手法はノード収集を行った後にクラスタ係数を推定するためにさらにエゴネットワークにあたるノードとエッジも収集する必要がある．これは一つのノードに対するクラスタ係数を求めるために，隣接ノードに加えて隣接ノードが持つエッジの情報が必要なためである [2]．したがってランダムウォークのステップ数以上の情報が必要となり，取得情報量が制限されているソーシャルネットワークの条件下では推定精度が低下する．例えば図 1 を見ると，ランダムウォークによってのみ見えている情報はノードが 2 個エッジが 5 本だが，エゴネットワークの範囲も含めると見えている情報はノードが 4 個エッジが 7 本と増えているのが分かる．

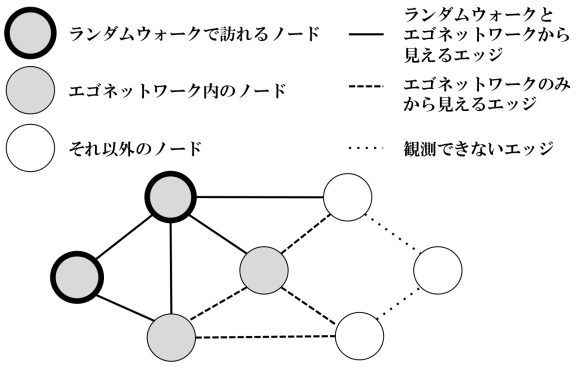

図 1　エゴネットワーク

これに対し Hardiman ら [12] は SRW によるノードの収集を行った後，エゴネットワークの情報を使用せずにクラスタ係数を推定する手法を提案した．この手法はランダムウォークをしながら 1 ステップ前と 1 ステップ後のノードが繋がってる場合の数をカウントする．つまり三角形を数えることでクラスタ係数を推定する手法である．しかし，SRW では次のノードに移動する時に一つ前に訪問したノードに遷移する可能性があり，その場合三角形を数えることにおいて無駄なステップを踏むことになる．また同様の理由で訪問するノードの範囲が狭まり，偏ったサンプリング結果を引き起こす可能性がある．

第 6 章では，我々の NBRW による提案手法と Hardiman ら [12] のクラスタ係数推定手法を比較する．Ribeiro ら [22] の手法と Gjoka ら [2] の手法との比較は [12] の論文により，Hardiman ら [12] の手法の方が精度が高いことが判明しているため，本論文では比較しない．

3. 準　　備

本章では，数式やグラフの表現の定義と本論文で使用するサンプリング手法について述べる．この章に記した定義は表 1 にまとめた．

3.1 表　　記

本論文ではネットワークを無向グラフ $G(V, E)$ で表す．グラフ G は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仮定する．$V = \{v_1, v_2, ..., v_n\}$ はノード（頂点）の集合であり，全ノード数を $n = |V|$ とする．E はエッジの集合である．頂点 $v \in V$ と隣接しているノードの集合を $N(v) \stackrel{\text{def}}{=} \{w \in V : (v, w) \in E\}$ とする．頂点 v_i の次数を d_i または k_{v_i} と表すとする．$d_i = k_{v_i} = |N(v_i)|$ である．全てのノードの次数の総和を $D \stackrel{\text{def}}{=} \sum_{i=1}^{n} d_i = 2|E|$ とする．グラフ G の $n \times n$ 隣接行列を A とする．すなわちノード v_i と v_k の間にエッジが存在すれば $A_{i,k} = A_{k,i} = 1$ であり，エッジが存在

しなければ $A_{i,k} = A_{k,i} = 0$ である．

［定義 3.1］ 3 つのノードの組 (v_j, v_i, v_k) に対して，$j < k$ のとき v_j と v_i，v_i と v_k 間のエッジが存在することを「(v_j, v_i, v_k) は連結」と定義する．隣接行列で表すと $j < k$ の時 $A_{j,i} = 1, A_{i,k} = 1$ である．

［定義 3.2］ 3 つのノードの組 (v_j, v_i, v_k) に対して (v_j, v_i, v_k) が連結していて，v_j と v_k の間にエッジが存在する時，「(v_j, v_i, v_k) は三角形」と定義する．隣接行列で表すと $A_{j,k} = 1$ である．

3 つノードの組 (v_j, v_i, v_k) が連結であるとは $A_{j,i}A_{i,k} = 1 (j < k)$ と同義である．3 つノードの組 (v_j, v_i, v_k) が三角形であるとは $A_{i,j}A_{i,k}A_{j,k} = 1\,(j < k)$ と同義である．特定のノード v_i に対して連結している (v_j, v_i, v_k) の数は $\sum_{j<k} A_{j,i}A_{i,k}$ と表すことができ，$\sum_{j<k} A_{j,i}A_{i,k} = d_i(d_i - 1)/2$ である．特定のノード v_i を含む (v_j, v_i, v_k) の三角形の数を $l_i \stackrel{\text{def}}{=} \sum_{j<k} A_{i,j}A_{i,k}A_{j,k}$ と定義する．これは v_i の隣接ノード間のエッジの数に等しい．

3.2 クラスタ係数

［定義 3.3］ v_i に対するクラスタ係数 [23] を c_i と表し

$$c_i = \begin{cases} 0 & d_i = 0 \text{ または } d_i = 1 \\ \dfrac{2l_i}{d_i(d_i-1)} & \text{otherwise} \end{cases} \qquad c_i \in [0,1]$$

とする．これは連結している (v_j, v_i, v_k) の数に対する (v_j, v_i, v_k) の三角形の数の割合を意味する．

［定義 3.4］ 平均クラスタ係数を C と表し

$$C = \frac{1}{n}\sum_{i=1}^{n} c_i$$

と定義する．本研究で推定するのはこのグラフの平均クラスタ係数である．

3.3 グラフサンプリング手法

本論文の実験のために実装したランダムウォークベースのグラフサンプリングアルゴリズムを二つ紹介する．

3.3.1 Simple random walk

$R = (x_1, x_2, ..., x_r)$ を SRW によって訪れたノードのインデックスの列とする．r は合計ステップ回数である．最初にランダムに選ばれたノード v_{x_1} からスタートし，隣接ノードから一様ランダムに一つノードを選択し移動する．これを繰り返す．SRW は既約なマルコフ連鎖である．遷移確率行列を $\mathbf{P} = \{P(v,w)\}_{v,w \in V}$ とすると $P(v,w)$ は

$$P(v,w) = \begin{cases} \frac{1}{k_v} & w \in N(v) \\ 0 & \text{otherwise} \end{cases}$$

と表される．事象 A が起きる確率を $\Pr[A]$ と表すとすると，r 回ステップ後の SRW による分布は次のように表すことができる．

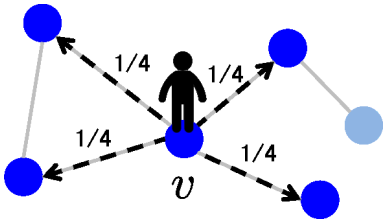

図 2 SRW の遷移確率

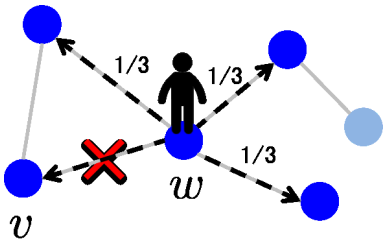

図 3 ノード v からノード w へ遷移した直後の NBRW の遷移確率

$$\boldsymbol{\pi} = (\Pr[x_r = 1], \Pr[x_r = 2], ..., \Pr[x_r = n])$$

十分多くランダムウォークのステップ数を重ねた後の確率 $Pr[x_r = i]$ を $\pi(i)$ と定義し，SRW の場合 $\pi(i) = d_i/D$ に収束する [19]．ベクトル $\boldsymbol{\pi} = (\pi(1), \pi(2), ..., \pi(n))$ はグラフ G の定常分布と呼ばれる．SRW では次数が高いノードに訪れやすい．例えば他のノードに比べて次数が 2 倍のノードは訪れる確率も 2 倍になる．

3.3.2 Non-backtracking random walk(NBRW)

$R' = (x_1, x_2, ..., x_r)$ を NBRW によって訪れたノードのインデックスの列とする．NBRW は次のノードを選択する時に一様ランダムに一つ選択する点は SRW と同じだが，一つ前に訪れたノードを選択しないのが特徴である．すなわち $x_{t+1} \neq x_{t-1}\,(t \geqq 1)$ を満たす．ただし現在いるのノード v に対して次数 k_v=1 の場合は一つ前のノードしか遷移先がないため，例外となる．また最初のノードから次のノードを選択する時は，一つ前のノードが存在しないため $1/d_{x_1}$ の確率で遷移する．SRW と NBRW の遷移確率の違いを図 2 図 3 に示した．

3.4 Hardiman らのクラスタ係数推定手法

Hardiman ら [12] は SRW をしながら三角形の数えるクラスタ係数推定手法を提案した．その概要について述べる．

SRW によるランダムウォーク $R = (x_1, x_2, ..., x_r)$ に対して新しい変数 ϕ_k を定義する．

$$\phi_k \stackrel{\text{def}}{=} A_{x_{k-1}}A_{x_{k+1}} (2 \leqq \forall k \leqq r-1). \qquad (1)$$

ランダムウォークの $k-1$ ステップ目のノードと $k+1$ ステップ目のノード間にエッジが存在する場合 $\phi_k = 1$ となり，存在しない場合は 0 となる．例えば図 4 のように x_1, x_2, x_3, x_4 の順番にランダムウォークした場合，x_2 の前後である x_1 と x_3 は隣接しているので $\phi_2 = 1$．x_3 の前後である x_2 と x_4 は隣接していないので $\phi_3 = 0$ となる．これはランダムウォークしながら三角形を数えることになる．そしてクラスタ係数の推定値を次のように定義する．ステップ数を r として

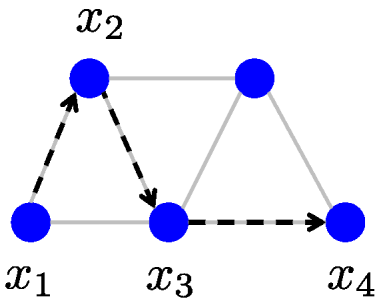

図 4　ϕ_k の計算例

$$\hat{C} = \frac{\frac{1}{r-2}\sum_{k=2}^{r-1}\phi_k/(d_{x_k}-1)}{\frac{1}{r}\sum_{k=1}^{r}1/d_{x_k}} \tag{2}$$

とした．

表 1　定義のまとめ

G	無向グラフ
n	グラフのノード数
A	G の隣接行列
v_i	G のノード
$N(v)$	ノード v の隣接ノードの集合
d_i	ノード v_iの次数
D	全てのノードの次数の総和 $\sum_{i=1}^{n} d_i$
r	ランダムウォークの全ステップ数
x_k	ランダムウォークの k 番目ノードのインデックス
R	SRW によるランダムウォーク
R'	NBRW によるランダムウォーク
$\boldsymbol{\pi}$	定常分布
l_i	v_i の隣接ノード間のエッジの数
c_i	ノード v_i に対するクラスタ係数
C	ネットワークのクラスタ係数
$\hat{C}$	C の推定量

4. 提案手法

既存手法の Hardiman ら [12] では SRW を使用してクラスタ係数を推定する．しかし SRW ではひとつ前のノードに戻る場合無駄が生じる．提案手法では三角形の数えるアイデアと NBRW を組み合わせた新しい推定手法を提案し，クラスタ係数推定の精度の向上を目指した．本章では，提案手法の概要について述べる．また，実装のアルゴリズムを Algorithm1 に示した．

4.1 クラスタ係数推定

NBRW によるランダムウォーク $R'(x_1, x_2, ..., x_r)$ に対して新しい変数 ϕ'_k を定義する．

$$\phi'_k \stackrel{\text{def}}{=} A_{x_{k-1}x_{k+1}} (2 \leqq \forall k \leqq r-1).$$

これは式 (1) と同じ定義である．違いは SRW か NBRW によるものかである．次に ϕ_kとϕ'_kの $v_{x_k} = v_i$の時の期待値を求める．

$$E[\phi_k \mid x_k = i] = \frac{2l_i}{d_i^2} \tag{3}$$

$$E[\phi'_k \mid x_k = i] = \frac{2l_i}{d_i(d_i-1)} = c_i \tag{4}$$

(3)(4) の式の分子の $2l_i$は (v_j, v_i, v_k) の三角形の数とそれを反転させた (v_k, v_i, v_j) の数を足した数である．それぞれの分母は連結している (v_j, v_i, v_k) の数のうち (x_{k-1}, v_i, x_{k+1}) が選ばれる場合の数である．NBRW は一つ前のノードに戻らない分だけ出て行く場合の数が $d_i - 1$ となっている．また ϕ'_kの期待値は v_{x_k}のクラスタ係数の c_i と一致するため，SRW より精度が良くなると予想できる．

続いてネットワークのクラスタ係数 C を推定するために次の二つの変数を導入する．

$$\Phi = \frac{1}{r-2}\sum_{k=2}^{r-1}\phi'_k\frac{1}{d_{x_k}}$$

$$\Psi = \frac{1}{r}\sum_{k=1}^{n}\frac{1}{d_{x_k}}.$$

そして推定値 $\hat{C}$ を次のように定義する．

$$\hat{C} \stackrel{\text{def}}{=} \frac{\Phi}{\Psi}$$

推定値がステップ数を増やしていくと真値のクラスタ係数 C に収束することを示す．大数の法則より期待値と真値が一致することを示せば良い．

Φ の期待値は

$$\begin{aligned} E[\Phi] &= E[\phi'_k\frac{1}{d_{x_k}}] \\ &= \sum_{i=1}^{n}\pi'(i)E[\phi_k \mid x_k = i]\frac{1}{d_i} \\ &= \sum_{i=1}^{n}\frac{d_i}{D}c_i\frac{1}{d_i}. \\ &= \frac{1}{D}\sum_{i=1}^{n}c_i \end{aligned}$$

となり，Ψ の期待値は

$$\begin{aligned} E[\Psi] &= E[\frac{1}{d_{x_k}}] \\ &= \sum_{i=1}^{n}\pi'(i)\frac{1}{d_i} \\ &= \sum_{i=1}^{n}\frac{d_i}{D}\frac{1}{d_i}. \\ &= \frac{n}{D} \end{aligned}$$

従って，

$$C = \frac{1}{n}\sum_{i=1}^{n}c_i = \frac{E[\Phi]}{E[\Psi]} \tag{5}$$

ここで π' は NBRW の定常分布であるが [21] より，$\pi'(i) = \pi(i) = d_i/D$ であることがわかっている．この証明は付録 1. に示した．(5) より推定値 $\hat{C}$ が真値 C に収束することが示された．

5. 実験・評価

提案手法のクラスタ係数推定の精度を評価するために，複数

Algorithm 1 提案手法

Require: 無向グラフ $G(V, E)$, ステップ数 r
Ensure: グラフ G のクラスタ係数の近似値

```
v ← 最初のノード
w_1 ← v
i ← 1
t1 ← 0
t2 ← 1/k_v
while i < r do
  v の隣接ノードから一様ランダムにノード w を選ぶ
  if i == 1 then
    v ← w
    i ← i + 1
    w_i ← v
    t2 ← t2 + 1/k_v
  else
    if k_v ≦ 1 or w ≠ w_{i−1} then
      v ← w
      i ← i + 1
      w_i ← v
      if i ≧ 3 then
        if w_{i−2} ∈ N(w_i) then
          t1 ← t1 + 1/k_{w_{i−1}}
        end if
      end if
      t2 ← t2 + 1/k_v
    else
      ノード v に滞在する
    end if
  end if
end while
t1 ← t1/(r − 2)
t2 ← t2/r
return t1/t2
```

のネットワークに対して既存手法と提案手法で計算機実験を行った．ここでの既存手法とは，Hardiman らが提案した SRW によるクラスタ係数推定手法である．

5.1 データセット

Stanford Network Analysis Project (SNAP) のデータセット [24] で公開されているソーシャルネットワーク・引用ネットワークを用いて実験を行った．表 2 に各データセットの概要を示す．

表 2　データセット

ネットワーク	全ノード数 n	平均次数	平均クラスタ係数
Amazon	334,863	5.530	0.3967
DBLP	317,080	6.622	0.6324
Gowalla	196,591	9.668	0.2367
Live Journal	3,997,962	17.35	0.2843

5.1.1 Amazon ネットワーク

Amazon Web サイト (注1) をクローリングして集められたネットワークである．このネットワークは購入者がどの商品とどの商品が一緒に買われやすいかに注目している．ノードは商品であり，頻繁に一緒に購入される商品（ノード）に対しては無向エッジが張られる．

5.1.2 DBLP ネットワーク

Dblp computer science bibliography (注2) が提供する，コンピュータ科学の研究論文の共著者ネットワークである．ノードは論文の著者であり，一度でも共著関係になると無向エッジが張られる．

5.1.3 Gowalla ネットワーク

Gowalla は位置情報サービスを利用したソーシャルネットワーク Web サイトで，ユーザはチェックインを行うことで自分がいる場所を共有する．公開 API によって友達関係のネットワークを集めることができる．しかし 2016 年 1 月 20 日現在サービスは終了している．

5.1.4 LiveJournal ネットワーク

LiveJournal(注3) は他のユーザと友達関係を持つことができるオンラインブログコミュニティである．

5.2 実験環境

既存手法と提案手法を Python と NetworkX(注4) を用いて実装した．本実験は MacBook Pro (Retina, 13-inch, Early 2015)，プロセッサ 3.1 GHz Intel Core i7，メモリ 16 GB で行った．

5.3 実験内容

与えられたグラフに対して既存手法と提案手法を適用し誤差を計測する．既存手法には Hardiman ら [12] の SRW による手法を与える．他の手法と比較しないのは [12, 13] より，他の既存の手法より優れていることが実験によりすでに示されているからである．

まず予備実験として，既存手法と提案手法で 10000 ノードサンプルするのに必要なステップ数を計測し，10 回測定した平均値を求めた．その結果を表 3 に記した．

続いて誤差についての実験について説明する．10000 ノードサンプルするまでランダムウォークを続けさせ，各手法のクラスタ係数推定値を 1000 回測定し，その**正規化平均二乗誤差** (NBRSE) [21] を計算した．その結果を図 5，6，7，8 に示した．

5.4 実験結果

表 3 と正規化平均二乗誤差の実験から，提案手法が既存手法よりステップ数が少ないのにも関わらず高い精度でクラスタ係数を推定できていることがわかる．ゆえに提案手法の方が既存手法より優位といえる．

6. まとめと今後の課題

本論文では，NBRW と三角形の数えるアイデアを組み合わせた手法を提案した．実験から既存手法より効率的かつ高精度であり，クラスタ係数の推定手法として優れていることがわか

(注1)：http://www.amazon.co.jp/
(注2)：http://dblp.uni-trier.de/
(注3)：http://www.livejournal.com/
(注4)：https://networkx.github.io/

表 3　10000 ノード収集するのに必要なステップ数の平均

ネットワーク	既存手法	提案手法
Amazon	20308	14981
DBLP	16144	14232
Gowalla	14758	13406
Live Journal	11417	10616

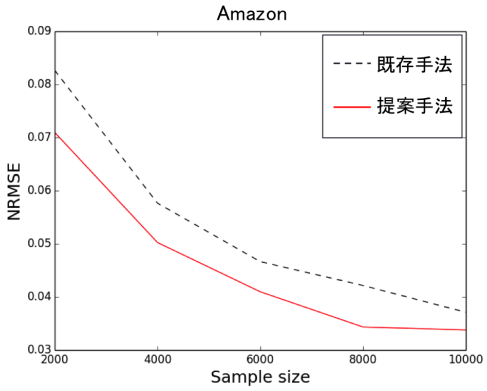

図 5　Amazon ネットワーク

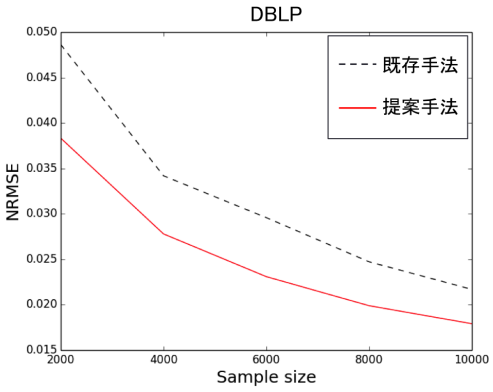

図 6　DBLP ネットワーク

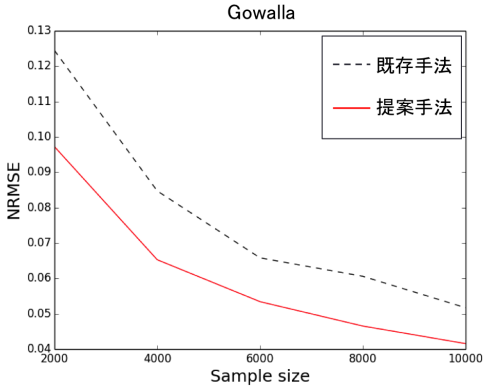

図 7　Gowalla ネットワーク

る．その理由として，NBRW が SRW の一つ前のノードに遷移する可能性があるためサンプリング結果に偏りが出やすいという欠点を解決できている点と，式 (4) から NBRW と三角形の数え上げのアイデアによるクラスタ係数推定が相性が良いという点が考えられる．今後の課題として，提案手法が既存手法よりステップ数が少ないのにも関わらず精度が良くなる理由を

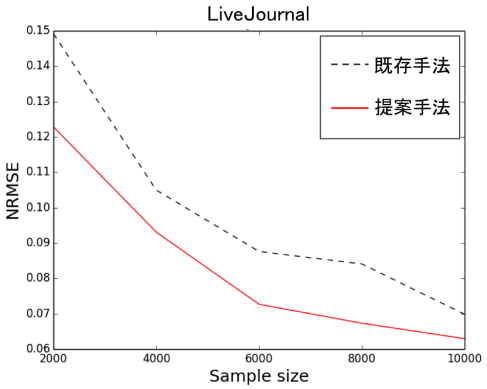

図 8　LiveJournal ネットワーク

定量的に評価することとする．

謝辞

本研究は JSPS 科研費 25700008，26540161 の助成を受けたものである．

文　献

[1] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th international conference on World Wide Web*, pp. 835–844. ACM, 2007.

[2] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM, 2010 Proceedings IEEE*, pp. 1–9. IEEE, 2010.

[3] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 29–42. ACM, 2007.

[4] Thomas Schank and Dorothea Wagner. *Approximating clustering-coefficient and transitivity*. Universität Karlsruhe, Fakultät für Informatik, 2004.

[5] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. Practical recommendations on crawling online social networks. *Selected Areas in Communications, IEEE Journal on*, Vol. 29, No. 9, pp. 1872–1892, 2011.

[6] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 631–636. ACM, 2006.

[7] Shaozhi Ye, Juan Lang, and Felix Wu. Crawling online social graphs. In *Web Conference (APWEB), 2010 12th International Asia-Pacific*, pp. 236–242. IEEE, 2010.

[8] Maciej Kurant, Athina Markopoulou, and Patrick Thiran. Towards unbiased BFS sampling. *Selected Areas in Communications, IEEE Journal on*, Vol. 29, No. 9, pp. 1799–1809, 2011.

[9] Daniel Stutzbach, Reza Rejaie, Nick Duffield, Subhabrata Sen, and Walter Willinger. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Transactions on Networking (TON)*, Vol. 17, No. 2, pp. 377–390, 2009.

[10] Amir H Rasti, Mojtaba Torkjazi, Reza Rejaie, and D Stutzbach. Evaluating sampling techniques for large dynamic graphs. *Univ. Oregon, Tech. Rep. CIS-TR-08*, Vol. 1,

, 2008.
[11] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks*, pp. 19–24. ACM, 2008.
[12] 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. International World Wide Web Conferences Steering Committee, 2013.
[13] Rong-Hua Li, Jeffrey Xu Yu, Lu Qin, Rui Mao, and Tan Jin. On random walk based graph sampling. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*, pp. 927–938. IEEE, 2015.
[14] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th international conference on World Wide Web*, pp. 835–844. ACM, 2007.
[15] Tianyi Wang, Yang Chen, Zengbin Zhang, Tianyin Xu, Long Jin, Pan Hui, Beixing Deng, and Xing Li. Understanding graph sampling algorithms for social network analysis. In *Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on*, pp. 123–128. IEEE, 2011.
[16] Ziv Bar-Yossef, Alexander Berg, Steve Chien, Jittat Fakcharoenphol, and Dror Weitz. Approximating aggregate queries about web pages via random walks. 2000.
[17] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Physical Review E*, Vol. 73, No. 1, p. 016102, 2006.
[18] Maciej Kurant, Minas Gjoka, Carter T Butts, and Athina Markopoulou. Walking on a graph with a magnifying glass: stratified sampling via weighted random walks. In *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pp. 281–292. ACM, 2011.
[19] David Aldous and Jim Fill. Reversible markov chains and random walks on graphs, 2002.
[20] Minas Gjoka. *Measurement of Online Social Networks*. PhD thesis, UNIVERSITY OF CALIFORNIA, 2010.
[21] Chul-Ho Lee, Xin Xu, and Do Young Eun. Beyond random walk and metropolis-hastings samplers: why you should not backtrack for unbiased graph sampling. In *ACM SIGMETRICS Performance Evaluation Review*, Vol. 40, pp. 319–330. ACM, 2012.
[22] Bruno Ribeiro and Don Towsley. Estimating and sampling graphs with multidimensional random walks. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pp. 390–403. ACM, 2010.
[23] L da F Costa, Francisco A Rodrigues, Gonzalo Travieso, and Paulino Ribeiro Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, Vol. 56, No. 1, pp. 167–242, 2007.
[24] Stanford large network dataset collection. https://snap.stanford.edu/data/.
[25] Galin L Jones, et al. On the markov chain central limit theorem. *Probability surveys*, Vol. 1, pp. 299–320, 2004.

付　　録

1. NBRW の定常分布

グラフ G 上の一般的なランダムウォーク，もしくは可逆性のある既約な有限マルコフ連鎖 $\{X_t \in V, t = 0, 1...\}$ は遷移確率行列 $\mathbf{P} = \{P(i,j)\}_{i,j \in V}$，定常分布 $\boldsymbol{\pi} = [\pi(i), i \in V]$ とする．任意の関数 $f : V \to \mathbb{R}$ に対して次のような推定量を定義する．

$$\hat{\mu_t}(f) \stackrel{\text{def}}{=} \frac{1}{t}\sum_{s=1}^{t} f(X_s)$$

定常分布 $\boldsymbol{\pi}$ に関する関数 f の期待値は次のように与えられる．

$$\mathbb{E}_{\boldsymbol{\pi}}(f) \stackrel{\text{def}}{=} \sum_{i \in V} \pi(i) f(i).$$

[25] では $\{X_t\}$ が定常分布 $\boldsymbol{\pi}$ の有限で既約なマルコフ連鎖であるとき，任意の初期分布 $\mathbb{P}\{X_0 = v\}, v \in V, (t \to \infty)$ に対して

$$\hat{\mu_t}(f) \to \mathbb{E}_{\boldsymbol{\pi}}(f)\ almost\,surely\,(a.s.)$$

が成立つ．ただし $\mathbb{E}_{\boldsymbol{\pi}}(|f|) < \infty$ とする．

この可逆なマルコフ連鎖から，NBRW によるランダムウォーク，つまり同じ定常分布 $\boldsymbol{\pi}$ の非可逆で既約な有限マルコフ連鎖を作る [21].

NBRW によるランダムウォークを $\{X'_t \in V, t = 0, 1...\}$ とする．現在のノードが X'_t のとき，次のノード X'_{t+1} の決定のされ方は現在のノード X'_t だけでなく，一つ前のノード X'_{t-1} にも依存する．このため，$\{X'_t\}_{t \geqq 0}$ 自体は V の状態空間ではマルコフ連鎖になり得ない．したがって，マルコフ連鎖を作れるように次のような状態空間を定義する．

$$\Omega \stackrel{\text{def}}{=} \{(i,j) : i, v \in V \,\text{s.t.}\, P(i,j) > 0\} \subseteqq V \times V$$

$|\Omega| < \infty$ であり，$Z'_t \stackrel{\text{def}}{=} (X'_{t-1}, X'_t) \in \Omega\,(t \geqq 1)$ とする．簡単に表記するために e_{ij}は状態 $(i,j) \in \Omega$ を表すとする．ただし $e_{ij} \neq e_{ji}$である．$\mathbf{P'} \stackrel{\text{def}}{=} \{P'(e_{ij}, e_{lk})\}_{e_{ij}, e_{lk} \in \Omega}$ を状態空間 Ω 上の既約なマルコフ連鎖 $\{Z'_t \in \Omega, t = 1, 2, ...\}$ の遷移確率行列とする．この時定義より $P'(e_{ij}, e_{lk}) = 0\,(\forall j \neq l)$．マルコフ連鎖 $\{Z'_t\}$ の定常分布 $\boldsymbol{\pi'} \stackrel{\text{def}}{=} [\pi'(e_{ij}), e_{ij} \in \Omega]$ が次のように与えられたとする．

$$\pi'(e_{ij}) = \pi(i)P(i,j),\ \ e_{ij} \in \Omega$$

この時元のランダムウォーク $\{X_t\}$ の可逆性から $\pi'(e_{ij}) = \pi'(e_{ji})$ である．また，定常状態中のノード j にいる NBRW によるランダムウォーク $\{X'_t\}$ の確率は，元の可逆マルコフ連鎖 $\{X_t\}$ の定常分布の $\pi(j)\,(\forall j \in V)$ と同じである．

$$\sum_{i \in V : e_{ij} \in \Omega} \pi'(e_{ij}) = \sum_{i \in V} \pi(i)P(i,j) = \pi(j), \forall j \in V$$

最初の等号は，$P(v,w) = 0, \forall (v,w) \notin \Omega$ によって導かれる．特に任意の関数 $f : V \to \mathbb{R}$ に対して，もう一つの関数 $g : \Omega \to \mathbb{R}$，$g(e_{ij}) = f(j)$ としたとき，

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\pi}'}(g) &= \sum_{e_{ij} \in \Omega} g(e_{ij})\pi'(e_{ij}) = \sum_{j \in V}\sum_{i \in V} f(i)\pi(i)P(i,j) \\ &= \sum_{j \in V} f(j)\pi(j) = \mathbb{E}_{\boldsymbol{\pi}}(f)\end{aligned}$$

大数の法則より

$$\frac{1}{t}\sum_{s=1}^{t} g(Z'_s) = \frac{1}{t}\sum_{s=1}^{t} f(X'_s) \to \mathbb{E}_{\boldsymbol{\pi}'}(g) = \mathbb{E}_{\boldsymbol{\pi}}(f)\,a.s.,$$

すなわち，$\sum_{s=1}^{t} g(Z'_s)/t = \sum_{s=1}^{t} f(X'_s)/t$ は $\mathbb{E}_{\boldsymbol{\pi}}(f)$ の不偏推定量である．