Vol.2013-DBS-158 No.9
2013/11/26

複雑ネットワークの生成モデルを反映したグラフサンプリング手法

宇都宮健太[1,a)]　首藤一幸[1,b)]

概要：グラフの性質は次数分布や直径，クラスタ係数といった特徴量で表され，それらはグラフ全体の構造から算出できる．しかし，実世界の大規模グラフには全体の把握が困難なものも多い．例えば，SNS によっては友人関係の収集が許されていなかったり，そもそも全体の収集が規模や時間的に困難な場合が多い．その場合，サンプリングによってグラフの一部から全体の特徴量を推定する．従来のサンプリング手法では，グラフの性質について何も仮定を置かずに特徴量を推定する．しかし，グラフの性質について何らかの予測が立つ場合，その予測を踏まえた特徴量推定を行うことで，より正確な推定を行える可能性がある．そこで本研究では，対象のグラフが複雑ネットワークであるという予測・前提の基に，複雑ネットワークの生成モデルを反映したサンプリングを行う．実際のソーシャルネットワーク上のグラフを対象に提案手法でサンプリングを行った結果，特徴量の種類によっては従来手法より真値に近い推定結果を得られた．

キーワード：グラフ，複雑ネットワーク

A Graph Sampling Technique Reflecting A Complex Network Generation Algorithm

Kenta Utsunomiya[1,a)]　Kazuyuki Shudo[1,b)]

Abstract: Graph sampling is techniques to extract a subgraph of a large graph to analyze characteristics of the whole graph. Existing techniques do not introduce any preconception about the characteristics of the target graph. It is natural because the purpose of the sampling is investigation of them. But sampling based on prior knowledge or estimation on the target graph possibly yields a subgraph that reflects the whole graph better. Especially in this paper, we present a sampling technique supposing that the whole graph is a complex graph. The sampling algorithm is based on a complex network generation algorithm. It showed better results on number of characteristics.

Keywords: Graph , Complex Network

1. はじめに

ソーシャルネットワーク解析の対象は Facebook, Twitter などの SNS のグラフデータのように大規模である．しかしながらそういった巨大なグラフ全体を直接詳細に解析することは，莫大なコストがかかるため非常に困難である．そのためグラフ全体を直接解析するのではなく，グラフ全体から部分グラフを抽出し，解析を行い，全体の特徴量の推定を行う．この手法を，グラフサンプリングと呼ぶ．部分グラフを解析することでグラフ全体の特徴を知ることができ，そのグラフがどういった構造を持つのかを理解できるようになる．

サンプリングの代表的な戦略に，一様分布にしたがってノードを選択し，部分グラフを得る方法がある．例えば

[1] 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
Dept. of Mathematical and Computing Sciences , Tokyo Institute of Technology
2-12-1 Ookayama , Meguro , Tokyo ,152-8552 , Japan
[a)] utsunomiya.k.aa@m.titech.ac.jp
[b)] shudo@is.titech.ac.jp

SNS のグラフデータで全てのノードに連番の ID が振られているような状況では，乱数でノード ID を指定して一様分布にしたがってノードを得ることはできる，しかしこのような状況はグラフの一般的な性質とは外れてしまっているため特殊であり，一般的なグラフでは適用出来るとは言えない．そのため，一般的にはノードを一様に選択することは現実的ではない．また，連番の ID が振られていたとしても，ID を生成したそのノードが存在するとは限らない．そこで，別の戦略として，ある始点のノードから隣接関係をたどっていく探索的な方法でノードをサンプルする方法が行われる．

一方でソーシャルネットワークの多くは，ランダムグラフや規則的なグラフではなく，複雑ネットワークである．従来のサンプリング手法ではこういったソーシャルネットワークに特徴的な性質については何も仮定をおかずに特徴量を推定している．そこで，グラフの性質に対してある程度の予測が出来る場合その予測を踏まえて特徴量の推定を行うことでより少ないサンプルでより正確な推定を行える可能性がある．本研究では対象のグラフが複雑ネットワークであるという前提のもとで，複雑ネットワークの生成モデルを反映したサンプリングを行う．具体的には、複雑ネットワークの生成アルゴリズムの 1 つである BA モデル [1] をサンプリングに適用する．本手法で実際のソーシャルグラフでのサンプリングを行った結果，特徴量の種類によっては従来手法よりも真値に近い推定結果が得られた．

2. グラフサンプリング

グラフサンプリングとはグラフの頂点，辺の一部を抽出し，部分グラフを作成し，そして部分グラフから元の特徴量の概算を行うことである．巨大で全てを分析する際に多大なコストが掛かるグラフや，全てのノード，辺を網羅することが難しいグラフに対して有効である．

2.1 グラフサンプリングの目標

グラフサンプリングとはグラフの特徴を保ったまま，グラフのサイズを小さくする手法だといえる．抽出した小さなグラフでももとのグラフの特徴を保ったグラフを得ることが目標である．Keskovec [4] はグラフサンプリングにおいて，どのような特徴量を正確に算出出来ればいいか基準を挙げた．以下にその基準を列挙する．

次数の分布

次数の分布はグラフの特徴を端的に表している大域的特徴量である．複雑ネットワークの場合は次数はべき乗則に従う．また格子グラフやランダムグラフも次数の分布は特徴的である．そのため，サンプルされた部分グラフの次数の分布が元のグラフのものに近ければ，そのグラフの特徴を保っているといえる．

任意の 2 ノード間のホップ数

ある 2 ノード間のホップ数が正確であると，その 2 ノード間の最短経路がわかっているということである．

グラフの直径

グラフの直径はグラフの大域的特徴量である．サンプル済のノード数が増えるに連れてある値に収束すると，グラフを十分に大域的に網羅しているといえる．

グラフのクラスタ係数の分布

あるノードのクラスタ係数とはそのノードがどれくらいの数の三角形を構成するかを示す指標である．ノード v の次数を d，ノード v を含む三角形の数を N_v とすると，クラスタ係数 C_v は以下のようにして定義される．

$$C_v = \frac{2N_v}{d(d-1)}$$

ある一つのノードに対してクラスタ係数が正確であると，ノードの周りをうまく探索しているという事がいえる．算出されたクラスタ係数が近いノードが多いほど，全体のノードを探索できているといえる．

2.2 サンプリングの戦略

ノードをどういった手順で選んでいくかがすなわちグラフサンプリングのアルゴリズムである．ノード選択の戦略としては主にノード指向のサンプリング，辺指向のサンプリング，そして探索的手法が存在する．

ノード指向のサンプリング戦略

ノード指向のサンプリングとはノードを隣接関係は一切考慮せずにランダムに選択していく手法である．例えば Facebook や mixi ではユーザーごとに連番の ID が振られている．ID の最大値と最小値を知ることが出来れば，各ノードが選択される確率が一様なノード指向のサンプリングが可能である．各ノードが選択される確率が一様であると，次数分布を正確に求めることができる．

また各ノードが選択される確率に重み付けを行う手法も提案されている．例えば，ノードのページランク [3] やノードの次数などの特徴量で重み付けを行う手法が提案されている．ノードのページランクで重み付けを行う手法では，連結性とクラスタ係数の分布は少数サンプルで正確に求まる．

探索的手法

ノード指向のサンプリングを行うことができるということは，グラフ内の全てのノードをランダムに選択できることを前提としている．しかしこのような状況は一般的なグラフにはいえない．また連番の ID が振られていたとしても，ID を生成したノードが存在するとは限らない．そのため，ノード指向のサンプリングを行うことは実世界のネットワークおいては難しい．そのため，実際の大規模なソーシャルネットワークをサンプリングする際には，ノード間の隣接関係をたどっていく手法が現実的である．ランダム

ウォークは，あるノードから隣接する辺を一つランダムに選択し，この操作を繰り返し行い辺をたどっていく手法である．ランダムウォークではある確率でジャンプして始点に戻る．

また，ランダムジャンプは始点から繰り返し辺をたどっていきある確率でジャンプするという点ではランダムウォークと同じである．しかし，ランダムウォークはジャンプ先が必ず始点であるのに対し，ランダムジャンプはジャンプするときにグラフ内の任意のノードへジャンプする．そのため，ランダムジャンプはノード指向のサンプリングと同じく，全ての頂点を一様に選択できることを仮定している．

ランダムウォーク，ランダムジャンプともに十分大きな数だけサンプリングを行うと，あるノードがサンプリング結果に含まれる確率はノードの入次数に比例する．そのため，それらの手法でグラフサンプリングを行うとサンプルされるノードは次数が高いものに偏ってしまう．そこでランダムウォークを改良し，隣接するノードを選択する際に次数の高いノードへ偏ってしまわないように重みをつけたMonteCaroSampling [2] といった手法が提案されている．ただ，この手法は隣接ノード群の入次数を必要とするため，一つのノードをサンプルに含める際にそのノードに隣接する全てのノードの入次数を調べなければならない．

Forest Fire 法 [5] はグラフの成長モデルを反映した手法で，あるノードと、出ている辺を経由して隣接しているノードをサンプル済のノードに含めるかどうかを確率 p_f で決定する．そして各サンプル済のノードに対して同じ操作を行っていく手法である．この手法は少ないサンプル数でクラスタ係数の分布を正確に求められることが知られている．しかし，この手法は大きなサイズのサンプルが必要な場合は，ランダムジャンプを必要とする．つまり，全ての頂点を一様に選択できることを仮定している．

その他の手法

他の探索手法として幅優先探索や深さ優先探索を行う手法もある．しかし，両方ともサンプルされた部分グラフがもとのグラフの局所部分になってしまうため，グラフの大域的特徴量を導出するには不適切である．

3. 複雑ネットワークの性質と生成モデル

3.1 複雑ネットワークの性質

複雑ネットワークは実世界における巨大で複雑な構造をしているグラフ構造をしたものと定義されている [8]．SNSの人間関係，交通網，論文の参照関係など実世界におけるグラフ構造をしたもののの一部が複雑ネットワークであるといえる．複雑ネットワークには以下に述べるようにランダムグラフや規則的なグラフにはない特徴的な性質がある．

3.1.1 スモールワールド性

複雑ネットワークの特徴の一つとして，グラフの任意の二頂点間のホップ数がノード数と比較して小さいことが挙げられる．この性質をスモールワールド性と呼ぶ．友達関係のグラフは 6 次の隔たりと呼ばれ，世界中の任意の人間へ友達関係を辿って行くと高々 6 人を経由するだけで全ての人間へたどりつくことができるといわれている [6]．

3.1.2 クラスタ性

クラスタ性とは，クラスタ係数の値が十分大きな値をとることである．例えば，ランダムグラフではクラスタ係数は辺の生成確率 p に比例する．しかし，複雑ネットワークでは辺と頂点の数が同じランダムグラフと比較すると，クラスタ係数は高いという特徴を持つ．

3.1.3 スケールフリー性

複雑ネットワークのもう一つの特徴として，スケールフリー性が挙げられる．スケールフリー性とは次数分布がべき乗則に従うことをいう．つまり次数 k の頂点の割合 $p(k)$ が $k^{-\gamma}$ に比例するということである．複雑ネットワークのほとんどのノードは次数が少ないものであるが，次数が極端に大きいノードもある程度は存在することがこの性質からわかる．

3.2 BA モデル

BA モデル [1] はスケールフリー性を満たす複雑ネットワークのモデルとして考案されたものである．また，このモデルはネットワークの成長という特徴を持っており，以下のアルゴリズムで生成される．

(1) n 個のノードからなる完全グラフをつくる
(2) 新しいノードを追加するときに，すでに存在するグラフの中で n 個のノードに対して辺を張る．このときに辺が張られる確率はそのノードの次数に比例する．
(3) 2. をグラフのサイズが任意の大きさになるまで続ける．

以上の手法でグラフを生成すると $p(k) \propto k^{-3}$ となるスケールフリー性をもつグラフができる．BA モデルで生成されたグラフの任意の二頂点間の平均経路長は $\log N$ となり，スモールワールド性も満たす．しかし，クラスタ係数の分布は $O(N^{-0.75})$ となり，グラフのサイズが大きくなるとクラスタ係数は 0 に近い値になる．そのため，BA モデルにクラスタ性をもたせる改良を加えたモデルも様々考案されている．

4. 提案手法

既存のサンプリング手法ではグラフの性質には何も仮定をおかずに特徴量を推定してる．しかし解析対象のグラフが複雑ネットワークである，など，グラフの性質について何かしらの事前知識や予測が立つ場合がある．その予測を踏まえて特徴量の推定を行うことでより少ないサンプルで，より正確な特徴量の推定ができる可能性がある．本研究では解析対象のグラフが複雑ネットワークの性質を持つグラフと仮定して複雑ネットワークの生成モデルを反映したサンプリング手法を提案する．提案手法は，サンプリン

グ結果であるサブグラフを複雑ネットワークとすることを狙う．そこで代表的な複雑ネットワークの生成モデルである，BA モデルを利用する．BA モデルのノードを足していくという操作を，ノードをサンプル済のノード集合に含めるという操作に置き換えることで複雑ネットワークのサンプリングに適用する．

4.1 アルゴリズム

以下に BA モデルに基づくサンプリングアルゴリズムを示す．基本的に，BA モデルのグラフ生成過程をサンプリングでのノード選択に置き換えたものである．

(1) グラフの中から，サイズ N のクリークを発見し，サンプル済のノードに含める．
(2) サンプル済ノードの集合からサンプルされていないノード N 個と隣接しているノードを見つける．
(3) （2）で見つけたサンプル候補ノードを $v_1, v_2, ...$ とする．BA モデルにならい，以下の通り，ノードを 1 つ選択する．
  (a) 各 v_i に隣接しているサンプル済みノード N 個について，サンプル済みグラフ内での次数を求める．それぞれ $d_1, d_2, ..., d_N$ とする．
  (b) 重み $w(v_i)$ を $w(v_i) = \prod_{j=1}^{N} d_j$ とする．
  (c) 各ノード対し重み $w(v_i)$ を計算し，重み $w(v_i)$ に比例する確率で次のノードを選択する．
(4) サンプル済みグラフが目標の大きさになるまで（2）〜（3）の操作を繰り返す．

ノードの重みに次数の総積を使う理由は，BA モデルの優先的選択を真似て，次にサンプルされたノードがサンプル済のノード内のどこのノードと連結させるかを次数に比例させるようにするためである．

上記の手順 1 ではサイズが N のクリークを発見するとしている．あるグラフから任意のサイズのクリークを見つけることは，N の値が大きいと困難になる．そのためパラメータ N は小さい値にすることが望ましい．$N = 1, 2$ の場合はグラフの中からクリークを見つける操作は容易である．$N = 3$ の場合はグラフの中から三角形を見つける操作を行う必要がある．しかし，本研究でサンプリングの対象としているグラフは複雑ネットワークであり，クラスタ性を満たすモデルを前提としている．そのため，隣接している任意の二ノード間の辺を含む三角形が存在する確率は比較的高い．つまり，$N = 3$ の場合でも最初のクリークを見つける操作は容易である．

5. 実験

提案手法を評価するために実データでのシミュレーションを行った．提案手法と比較するために，ランダムウォーク Forest Fire 法 [5] で同様の実験を行いそれらの結果を比較した．

5.1 実験データ

本研究では Stanford Network Analysis Project [7] で提供されているソーシャルグラフのグラフデータを用いてシミュレーションを行った．主な特徴量ははそれぞれ表 1 の通りである．

表 1 グラフデータの特徴量

データセット名	総ノード数	総辺数	直径	平均クラスタ係数	平均ホップ数
LiveJournal	4,847,571	42,851,240	9	0.167	4.6
okut	3,072,441	117,185,083	12	0.121	3.6
Facebook	1,563,931	26,415,672	13	0.131	5.1

5.2 評価の尺度

本研究では次数分布，ホップ数の分布，クラスタ係数，そして直径を評価の尺度とする．ただし，サンプルされた部分グラフのどの情報を利用するかは尺度によって異なる．以下ににその尺度の詳細と，利用する情報について述べる．

次数分布

次数分布に関してはサンプルされたノードのみで行う手法と，サンプルされたノードに隣接しているノードも含める手法が存在する．本研究では扱う対象のグラフはすべて無向グラフであり，入次数と出次数がそれぞれ等しく，サンプルされたノードの入出次数は正確に求められる．サンプルされたノードに隣接しているノードを含めると，次数が実際の値より極端に低いノードが多く出現してしまう．そのため，次数分布はサンプルされたノードのみで行うことが適切である．

ホップ数の分布

ホップ数はサンプルされたノード間のものを評価する．ホップ数を導出するにあたって，サンプルされた 2 つのノード間の辺のみを経由する方法と，サンプルされていないノードも経由する方法が考えられる．前者の方法は後者の方法よりも大きなホップ数を導出してしまうことがある．そのため，サンプルされたノードに隣接している全ての辺を経由し，経路に含める手法でホップ数を導出する．

クラスタ係数

クラスタ係数はホップ数と同じく導出するにあたって，サンプルされた 3 ノード間の辺の三角形の数を数える方法と，サンプルされたノードにつながっている辺全ての集合で構成される三角形を含める手法が考えられる．しかし，後者の手法は三角形の一つの頂点しかサンプルされてない場合に導出された値は，実際のクラスタ係数より極端に小さな値になってしまう．そのため，評価に用いるサンプルされたノード v に対するクラスタ係数 C_v は以下のように

して定義する．

$$C_v = \frac{N'}{\frac{d(d-1)}{2} + dd'}$$

- C': v を含む三角形のうち，サンプル済のノードが少なくとも二つ含まれる三角形の数．
- d: v に隣接している辺のうち，サンプル済のノードへ隣接している辺の数．
- d': v に隣接している辺のうち，サンプル済のノードではないノードへ隣接している辺の数．

直径

直径はサンプルされた任意の二ノード間のホップ数の中で最大のものとする．

5.3 評価方法

各特徴量の性質によって評価の方法は異なる．本項ではそれぞれの特徴量の評価方法について述べる．

5.3.1 次数分布の評価方法

表 2 グラフデータの γ 値

データセット名	γ
LiveJournal	0.75
orkut	1.6
Facebook	2.43

次数分布は元のデータがスケールフリーのグラフのモデルである，つまり次数 k の確率分布が $p(k) = ak^{-\gamma}$ であると仮定し，非線形回帰を行う．この指標で元のグラフの次数分布を再現しているかどうかがわかる．実験データの次数分布で非線形回帰を行った時の，γ 値は表 2 の通りである．

5.3.2 クラスタ係数, ホップ数の評価方法

分布の誤差の評価を求めるため二乗誤差和を用いる．二乗誤差和 RSS は以下の式で求まる．

$$RSS = \frac{\sum_{i=1}^{n}(e_i - e_i')^2}{\sum_{i=1}^{n} e_i^2}$$

- n:クラスタ係数の場合はサンプルされたノードの数．ホップ数の場合は ${}_nC_2$
- e_i:各ノード，二ノード間のクラスタ係数・ホップ数の真値
- e_i':サンプルリングを行なって上記の方法でクラスタ係数やホップ数を求めた時の概算値

5.4 シミュレーション結果

本節では以上の評価尺度，評価手法で評価を行ったグラフデータで，0 - 10 %のノードのサンプリングを行った．また本手法との比較を行うために既存のサンプリング手法である，Forest Fire 法，ランダムウォークとの比較を行った．

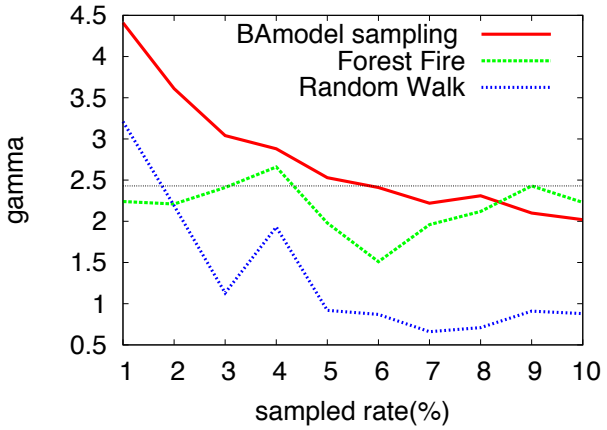

図 1 Facebook データの次数の頻度分布の γ 値との比較

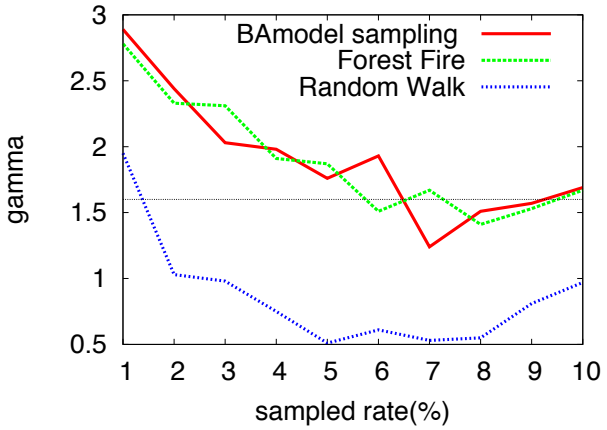

図 2 orkut データの次数の頻度分布の γ 値との比較

Forest Fire 法はパラメータ $p_f = 0.2$ とし，始点はランダムに決定する．

ランダムウォークはジャンプ確率は 0.15 とする．

また本提案手法でのシミュレーションは最初の始点はランダムに決定し，サンプリングの際のパラメータは $N = 3$ としてある．

5.4.1 次数分布

次数分布に関する真値との誤差は図 1，図 2，図 3 である．中央の黒い横線に近ければ近いほど元のグラフの特徴を表すよいサンプルであると言える．

どのデータセットに対しても同じ傾向がである．ランダムウォークの場合は次数の高いノードにサンプル済のノードが集中してしまう傾向があるため，γ の値は実際より高くなっている．提案手法と Forest Fire 法は正しい γ の値へ収束している．そのため次数の高いノードへのサンプル済のノードの集中は抑えられている．

5.4.2 ホップ数

ホップ数誤差の収束データは図 4，図 5，図 6 である．それぞれの値が 0 に近ければ近いほどよい．

Forest Fire 法は始点を選びなおす手法であるため，サン

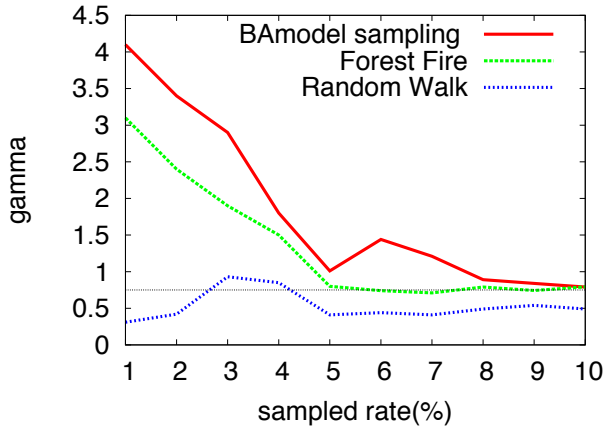

図 3　LiveJournal データの次数の頻度分布の γ 値との比較

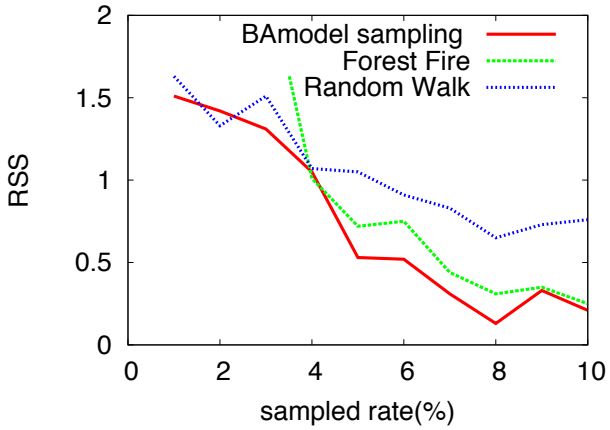

図 5　orkut データのホップ数の誤差

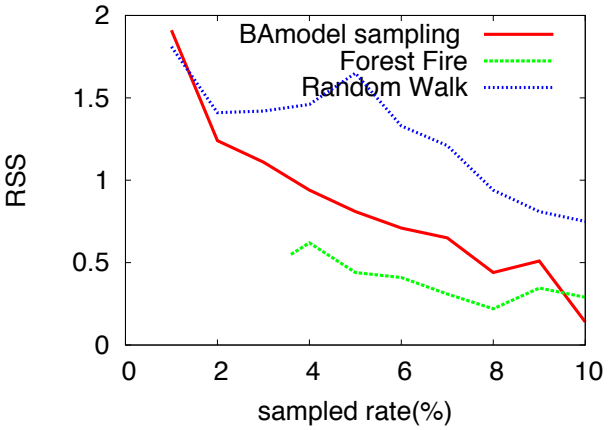

図 4　Facebook データのホップ数の誤差

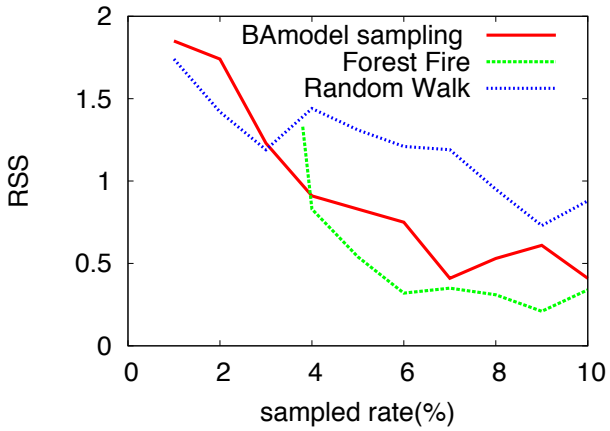

図 6　LiveJournal データのホップ数の誤差

プル済ノードの数が少ないと連結でなくなってしまう．そのため，サンプル数が少ない場合はホップ数が定義できなくなる場合があるので，連結でない場合はプロットをおこなっていない．提案手法とランダムウォークは単一始点でのサンプリングを行っているためサンプル数が少なくても非連結なグラフが生成されてしまうことはない．

データセットとも本提案手法はランダムウォークと比較して良い結果になっている提案手法と Forest Fire 法では非連結なグラフが出現しないという面では提案手法の方が勝っているが，Forest Fire 法で連結なグラフを作成できた後の結果は Forest Fire 法と提案手法ではさほど差はない．

5.4.3　クラスタ係数の分布

クラスタ係数の分布の誤差は図 7，図 8，図 9 である．ホップ数と同じく 0 に近いほど元のクラスタ係数の分布と近い値になっているということである．orkut,Facebook のデータは少ないサンプリング数では本手法はランダムウォーク，ForestFire 法ともに劣っていることがわかる．LiveJournal のデータでは ForestFire 法と提案手法が同じく収束している．しかし他のデータセットには見られない傾向であるため，これはデータセット依存である可能性がある．

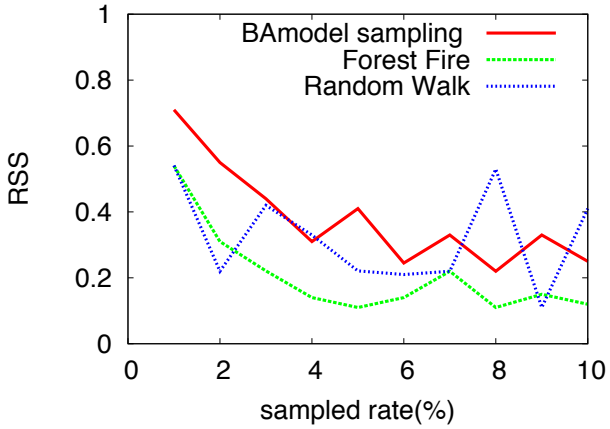

図 7　Facebook データのクラスタ係数の誤差

5.4.4　直径

各データセットにおいて直径とサンプリングで求めた値を比較したのが図 10，図 11，図 12 である．次数分布と同じく黒い横線が直径の真値であり，収束が早いほどより少ないサンプルでグラフ全体を網羅していることになる．サンプルされたグラフの直径の値が整数でないのは，各サン

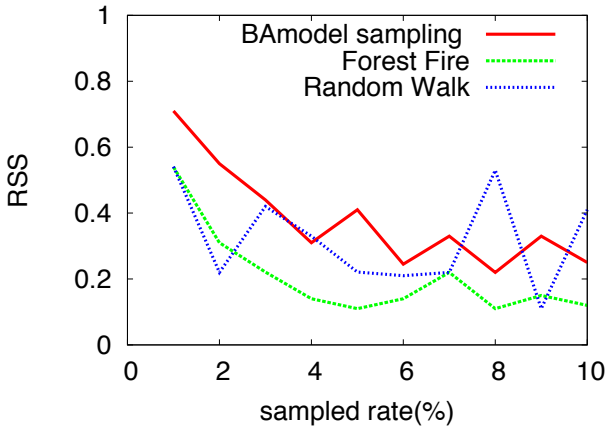

図 8　orkut データのクラスタ係数の誤差

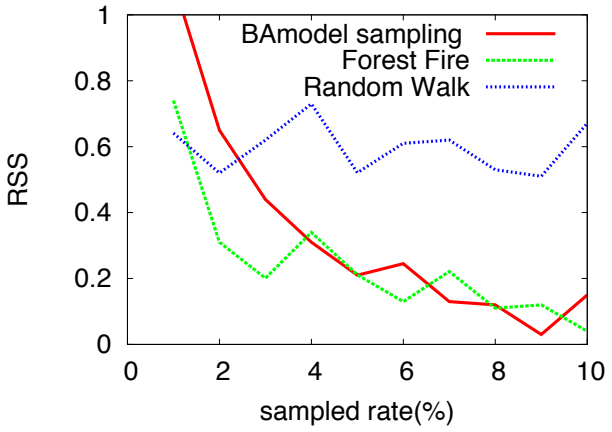

図 9　LiveJournal データのクラスタ係数の誤差

プル率において複数回シミュレーションを行い，それらの値の平均をとっているためである．また，サンプル数が少ない時の Forest Fire 法の直径の値が定義できないのはホップ数の時と同じく，連結でないグラフになってしまうからである．

ランダムウォークはでは，どのデータセットともサンプルのサイズが大きくなっても収束していない．しかし，どのデータセットでも，Forest Fire 法と提案手法の BA モデルサンプリングはサンプルのサイズが 6%を超えた辺りから一定の値に収束しつつあることがわかる．

5.5　考察

シミュレーションによって，提案手法の BA モデルサンプリングは各特徴量導出は，クラスタ係数以外はどのデータセットとも単一始点のランダムウォークと比較してよいサンプリング手法であるということがわかった．既存手法の Forest Fire 法と比較したところ，収束速度や精度といった面では差は見られなかった．しかし Forest Fire 法はサンプル済のノードの個数が大きくなるたびに，複数の始点をランダムに選びなおすことを必要としている．それに対

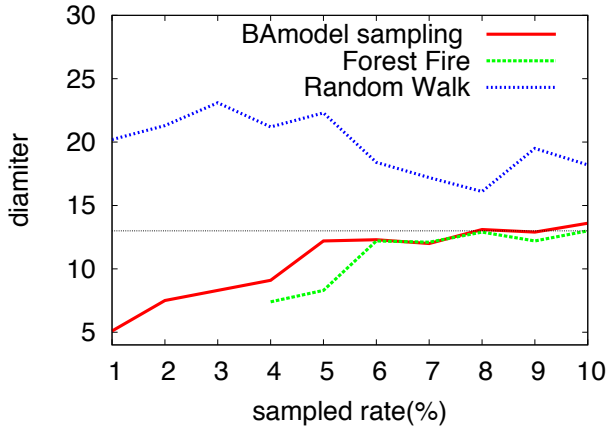

図 10　Facebook データの直径のサンプリングで求めた値と実際の値との比較

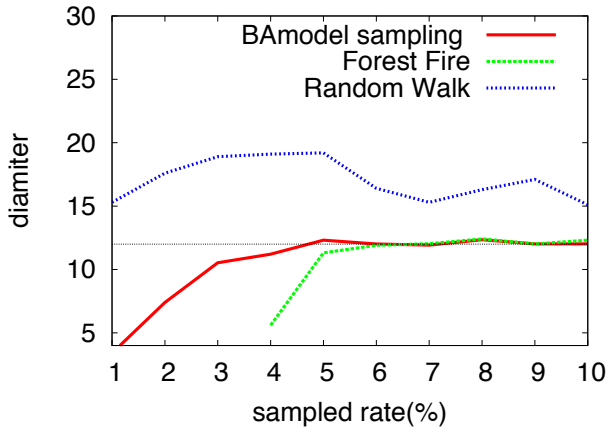

図 11　orkut データの直径のサンプリングで求めた値と実際の値との比較

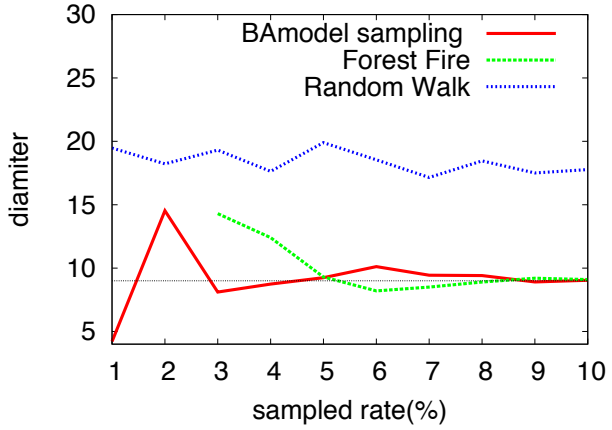

図 12　LiveJournal データの直径のサンプリングで求めた値と実際の値との比較

して本手法はサンプルのサイズに関わらず単一の始点だけでも利用できる手法であるである．そのため，複数始点でのサンプリングが難しい状況では本提案手法を利用するべきである．

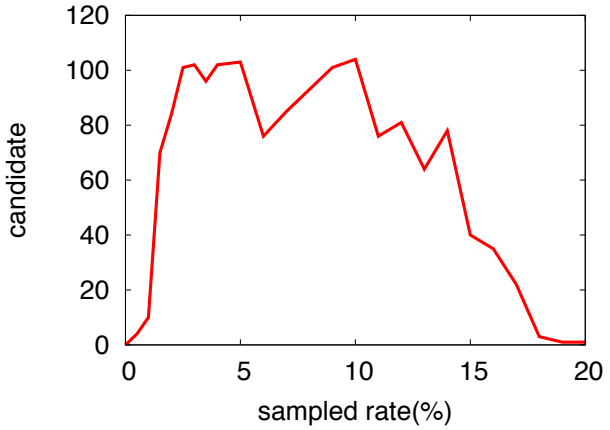

図 13　サンプルサイズに対するサンプルされるノードの候補の数 (orkut)

5.6　サンプルのサイズが大きい場合

提案手法のサンプリング手法はサンプルされたノードの集合からサンプルされていないノードへちょうど N 本の辺が存在するものを新たにサンプルするアルゴリズムである．しかし，この手法はちょうど N 本の辺がつながっているノードが存在しない場合は次にサンプルされるノードは存在しなくなってしまう．この現象はサンプル済のノード数によらず，理論上はどの時点でも起こりうる．

しかしパラメータ $N = 3$ とし，サンプル済のノードを最大でも全体の 10%とした本シミュレーションにおいて次へ進むべきノードが存在しないといった現象は，3 つのデータセット全てにおいては起こらなかった．

しかし，orkut のデータでサンプル済のノードを 10%以上にすると，図 13 のようにサンプルのサイズが大きくなるに連れて次にサンプルされるノードの候補が少なくなっていき，20%を超えた辺りからほとんどなくなってしまっている．この理由は，サンプル済のノードの個数が大きくなれば，サンプルされていないノードからサンプル済のノードへ隣接しているノードの数が増えていってしまい，サンプル済ノードへの辺の数が N となるサンプルされていないノードが少なくなってしまうからである．そのため，本手法はサンプルのサイズを大きくする場合では不適切である．そのため，サンプル済のノードの割合を高くしてもサンプリングを行わせるために，工夫を行う必要がある．

6.　まとめ

本研究ではグラフのサンプリングを行う際の戦略として，複雑ネットワークの生成モデルである BA モデルでの，グラフの成長を反映したサンプリング手法を提案した．提案手法でサンプルされたグラフは，スケールフリー性やスモールワールド性を満たすような部分グラフになる．

またシミュレーションによって，BA モデルサンプリングでもサンプルされたグラフの平均クラスタ係数が高くなるようなグラフになることを発見した．提案手法のシミュレーション結果とランダムウォークでのサンプリングを比較すると，特徴量は本手法のほうが全て少ないサンプルで正確に求められるということを発見した．Forest Fire 法でのサンプリング法と比較した際のシミュレーション結果はほとんど同じものであることも発見した．しかし，Forest Fire 法はサンプル済のノードの数が増えるほど始点を複数とる必要がある手法である．そのため，始点を複数ランダムに選択することが難しいようなネットワークの解析においては本手法は明らかに有用である．

今後の課題としては手法の理論的裏付けや，BA モデル以外へのサンプリングアルゴリズムの適用が考えられる．

参考文献

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.
[2] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
[3] Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. On near-uniform url sampling. *Comput. Netw.*, 33(1-6):295–308, June 2000.
[4] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 631–636, Philadelphia, PA, USA, 2006. ACM.
[5] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 177–187, Chicago, Illinois, USA, 2005. ACM.
[6] Stanley Milgram. The small world problem. *Psychology Today*, 1(1):61–67, 1967.
[7] Stanford University. Stanford large network dataset collection. http://snap.stanford.edu/data/index.html.
[8] 増田直紀，今野紀雄. 複雑ネットワーク: 基礎から応用まで. 近代科学社, 2010.