DEIM Forum 2013 E9-5

一様独立サンプリングとランダムウォークを組み合わせたグラフのサンプリング手法

宇都宮健太† 首藤一幸††

†, †† 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
〒 152-8550 東京都目黒区大岡山 2-12-1-W8-43
E-mail: †utsunomiya.k.aa@m.titech.ac.jp, ††shudo@is.titech.ac.jp

あらまし ソーシャルグラフ等，大規模なグラフを対象としてグラフの特徴量を得るためのサンプリング手法として，ランダムウォーク（RW）と一様独立サンプリング（UIS）がある．一般的に UIS はノード ID が既知であることなどを前提とするために適用できないケースが多く，RW の方が現実的であると考えられている．しかし，ノード ID が既知でなくともある程度密であれば，存在する ID を一定の成功率で推測することは可能である．以下，存在する ID の推測成功率が低いために UIS を効率的に行えない状況を想定する．そこで我々は，UIS と RW を組み合わせるサンプリング手法を提案する．これにより，ある種の特徴量をより正確に算出できるという UIS の利点と，ノード ID が既知でなくともよいという RW の利点を両立させる．また，UIS と RW の混合割合によって，トレードオフの関係にある特徴量の正確さとサンプリング効率がどうなるかを調べる．

キーワード ノードサンプリング，ソーシャルグラフ

A Graph Sampling Technique combining Random Walk and Uniform Independent Sampling

Kenta UTSUNOMIYA† and Kazuyuki SHUDO ††

†, †† Dept. of Mathematical and Computing Sciences, Tokyo Institute of Technology
2-12-1-W8-43 Ookayama, Meguro, Tokyo, 152-8552 Japan
E-mail: †utsunomiya.k.aa@m.titech.ac.jp, ††shudo@is.titech.ac.jp

Key words Node sampling, social graph

1. 導　　入

グラフ解析の対象は、インターネット、航空網などから Facebook、Twitter といった SNS まで、より大規模なものとなってきた。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りすぎるため非常に困難である．そのため，現在は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，その部分グラフを解析する手法を用いる．このような方法はグラフサンプリングと呼ばれる．

グラフサンプリングでは大規模なグラフからどのノードを抽出するかが特に重要であり，代表的な抽出方針としてにランダムウォーク（RW）と一様独立サンプリング（Uniform Independent Sampling，UIS）がある．

ランダムウォークはあるノードから隣接する他のノードへと次々とノードをたどっていく抽出手法である．ランダムウォークは各ノードがどのノードと連結しているかという情報さえあればよく，ノード ID を知る必要がないため，サンプリングがしやすいという利点がある．また解析対象のグラフが連結でなかった場合は，抽出されないノードが必ず存在してしまうため，グラフ全体にわたる特徴を見つけることは困難である．

一方，UIS は全ノードからランダムに，ノードを選択していく抽出手法である．UIS では各ノードの ID を知る必要があるためサンプリングは比較的難しい．しかし各ノードが選択され

る確率は全ノードで等しいため，平均次数といった特徴量はランダムウォークと比べて正確になるという利点がある．しかし，UIS で求めた部分グラフは連結でない可能性が高いためクラスタ係数といったノード同士のつながりに関する特徴量の計算には向かない．

そこで我々はある種の特徴量をより正確に算出できるという UIS の利点と，ノード ID が既知でなくともよいという RW の利点を両立させるため，ランダムウォークと UIS を組み合わせたグラフのサンプリング手法を提案する．そして，その手法によって各ノードが選択される確率を算出する．さらに，この手法を実際のソーシャルグラフのデータセットに適用し，グラフのサイズや次数分布といったグラフの特徴量ををどれだけ正確に算出できるかを評価した．また RW と UIS の比をチューニングすることでサンプリング全体に対して，どのような作用が起こるかどうかということも考察した．

2. 関 連 研 究

Metropolized Random Walk (MRW) [] はランダムウォークで遷移先の各ノードに重みを設定し重みが軽い方が遷移する確率が高くなるランダムウォークを行なっている．そのため，各ランダムウォークの遷移手順では入次数の高いノードへ遷移する確率は低く抑えることができる．

ForestFire 法 [] は UIS と幅優先探索を組み合わせた手法である．幅優先探索を一定の確率で別スタート地点から行うことで，様々な部分のサンプルを取得することができる．

グラフの特徴量を導出する際に，グラフのノード数が必要になってくることがある．グラフのサンプリングにおいてグラフ全体のサイズを求める手法は，Katzir [] がランダムウォーク中にサンプルノードが衝突する回数を利用してノード数 N を

$$N = \frac{\Psi \cdot \Psi^{-1}}{2C}$$

- Ψ　　サンプルノードの入次数の和
- Ψ^{-1}　　サンプルノードの入次数の逆数の和
- C　　サンプルノードの衝突回数

として概算できるとした．

3. RW と UIS を組み合わせたサンプリング手法

我々の提案手法では UIS が高い確率で可能であるという仮定の下で，UIS でサンプリングするノード数と RW でサンプリングするノード数の比を最初にパラメータとして与える．

入力：

（1） $G = (V, E)$　　サンプリング対象のグラフ

（2） r　　サンプリング回数

（3） c　　サンプルノードのうち UIS で選択するノードの割合

出力：

（1） $G' = (V', E')$：サンプリングされた部分グラフ

（2） C：サンプルノードの衝突回数

アルゴリズム：

（1） $count[v] = 0$ *for all* $v \in V$

$V' = \phi$

$E' = \phi$ と初期化

（2） グラフ内のノードを一つ選択し now_node とする

（3） 以下の作業（a, b, c, d）を r 回繰り返す

（a） 次のノードへジャンプする．
- 確率 c でグラフ内の別のノードへジャンプする
- 確率 $(1-c)$ で now_node と隣接しているノードへジャンプする

（b） ジャンプ先のノード v に対し $count[v_i] += 1$

（c） v を V' に加える．また v と隣接しているエッジを全て E' に加える．

（d） $now\_node = v$

（4） $C = \frac{1}{2}\sum_{i=1}^{|V'|} (count[v_i] \times (count[v_i] - 1))$ を計算

（5） G', C を出力して終了

このアルゴリズムでは，$c = 0$ の場合は単一始点ランダムウォークとなり，$c = 1$ の場合は UIS となる．

入力パラメータ c については，連続してランダムウォークで選択される数は幾何分布に従うため，以下のように決めることもできる

$$c = \frac{1}{k}$$

- k：ランダムウォークで連続してノードがサンプリングされる個数の平均

4. 特徴量分析

4.1 ノード選択確率

ランダムウォークにおいて各ノードが選択される確率はノードの入次数に比例する．UIS ではノードの入次数にかかわらず全ノードで一定である．そのため，UIS において各ノードが選択される確率は以下のようになる．

$$P\left(deg(v) = d\right) = \frac{d}{D}(1-c) + \frac{c}{N}$$

- D　全ノードの入次数の和
- d　ノードの入次数
- N　グラフのノード数

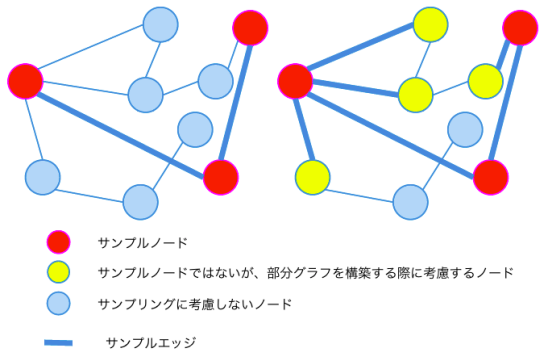

図 1　Induced Subgraph Sampling (左) と Star Sampling (右)

ノードが選択される確率を求めるにはグラフ全体のサイズと全体の入次数の和を知る必要がある．

Katzir の手法 [] を用いると，グラフサンプリング中のノード数は各ノードが選択される確率がノードに比例する．そのためこの手法はランダムウォークにおいてにしか成り立たない．そのため，ランダムジャンプを含む我々の提案手法でノードのサイズの概算を行う場合は以下の式となる．

$$N = \frac{1}{2C}\left(\Psi\Psi^{-1}(1-c) + cr^2\right)$$

本研究では解析するグラフの対象を複雑ネットワークに限定している．そのため次数に関しては，複雑ネットワーク特有の性質であるスケールフリー性を利用して解析を行う事ができる．スケールフリー性とは，ノードの入次数の確率分布がべき乗則に比例するという性質である．

つまり，次数の分布が以下のようになるということである．

$$p(deg(v) = k) \propto k^{-\alpha}$$

この事実を利用して，次数の和や次数分布の概算ができる．次数の和の概算値は以下の通りである．

$$D \cong N \frac{\sum_{k=1}^{d_{max}} k^{-\alpha+1}}{\sum_{k=1}^{d_{max}} k^{-\alpha}}$$

一般的なスケールフリーネットワークでは $0.5 \leqq \alpha \leqq 3$ である．定数倍の誤差を許容することになるが，仮に $\alpha = 2$ とすると辺の本数は以下のように概算できる．

$$D \cong N \frac{6 \ln d_{max}}{\pi^2}$$

またサンプリングの際に，より正確な α も求めるならばサンプリングされたノードの次数分布に対して非線形回帰分析を行うことで，α を求めることができる．

4.2　部分グラフの構築手法

サンプリングされたノードから部分グラフを構築する方法として, **Star Sampling** と **Induced Subgraph Sampling** の二つがある

Star Sampling はサンプリングされた部分グラフのエッジのみに注目し, サンプリングされていないノードは一切含めない方法である．この方法では，サンプルグラフのノード同士がつながっていることが正確にわかる．しかし，部分グラフのサイズはサンプルノード数以下となる．

Induced Subgraph Sampling はサンプリングされたノードだけではなく，サンプルノードに直接連結しているノードとその間のエッジまで含める方法である．この方法を用いると，サンプリングされたグラフはサンプルノード数よりも大きくなる．

表 1　各データセットのグラフの詳細

	orkut	LiveJournal	Facebook
ノード数	3,072,441	4,847,575	1,567,931
総エッジ数	117,185,083	42,851,240	117,185,083
最大次数	33,313	20,333	4,929
α	1.6	0.75	2.43

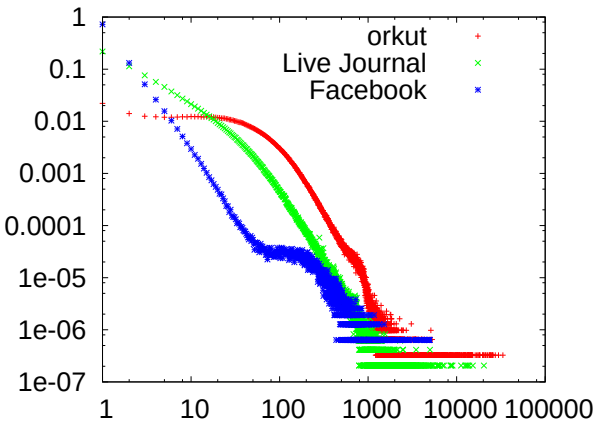

図 2　各 データセット全体の次数分布

5.　実　　験

提案手法がどれだけ正確に特徴量を算出できるかを，実際のソーシャルグラフを用いて検証した．

5.1　実験内容

本実験ではウェブページ [5] で提供されている orkut のソーシャルグラフと，Facebook のグラフと LiveJournal のグラフを用いた．各特徴量の真値は表 である．

各データセットの友人関係のソーシャルグラフは無向グラフであり，入次数と出次数が一致するため，出次数に関してのみ次数分析を行えば十分である．

提案手法の評価のためにジャンプ確率 c を変えてノードのサンプリング行った．このとき，$c = 0$ の場合はランダムウォークになり，$c = 1$ の場合は UIS となる．サンプル数のサイズは $\frac{N}{10000} \leqq r \leqq \frac{N}{100}$ でグラフのサンプリングを行いサンプリングされたグラフにおいて以下の特徴量を求め，正しい値との比較を行った

- ノードサイズの概算

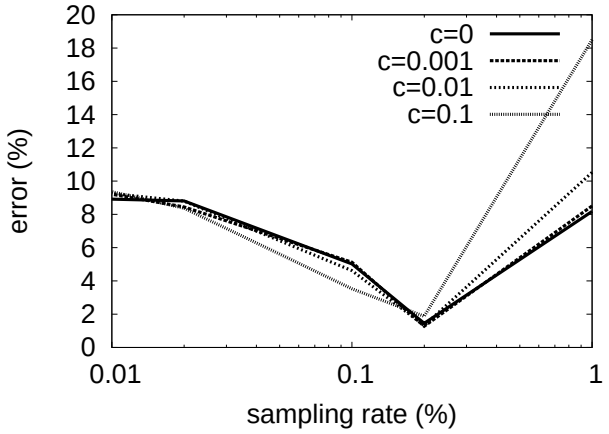

図 3　ノード数の概算

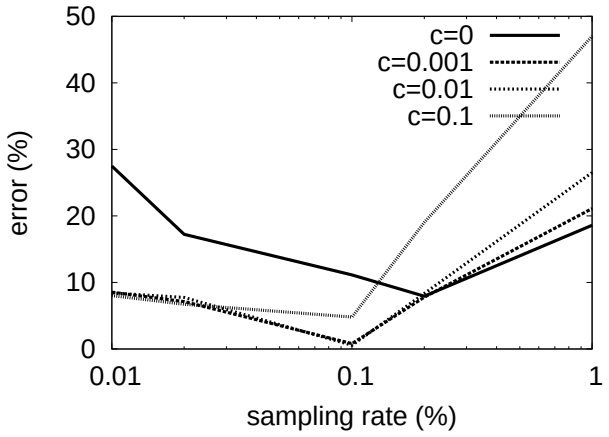

図 4　ノード数の概算 (LiveJournal)

- グラフ次数の和の概算
- ノード次数分布

グラフ次数の和の概算は計算を簡略化するため $\alpha = 2$ と仮定して概算を行い，以下の式を利用して評価を行った．

$$\frac{|\bar{x} - x|}{x}$$

- x：真値
- $\bar{x}$：概算値

またノードの次数分布については Star Sampling と Induced Subgraph Sampling の 2 種類の部分グラフでの評価を行った．

5.2 実験結果

5.2.1 ノードサイズの概算

orkut グラフと LiveJournal のグラフでは，グラフのサイズが十分でないときの誤差は大きかったが，c=0.1 の場合が最も誤差が小さくなっている部分があった．しかし，サンプルのサイズが 0.1%を超えたあたりから $c = 0$ のときを除いて誤差が極端に大きくなっている．このデータセットにおいて，サンプルサイズが大きな時に誤差はが大きくなる原因は，$c = 0$ の時を除いて概算値の r^2 の項の影響が大きくなってしまうためだと思われる．

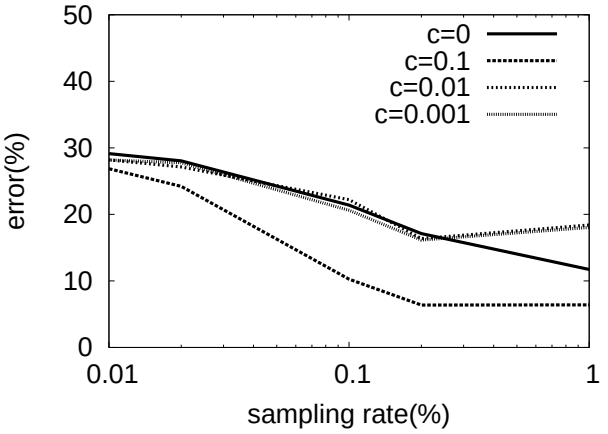

図 5　ノード数の概算 (Facebook)

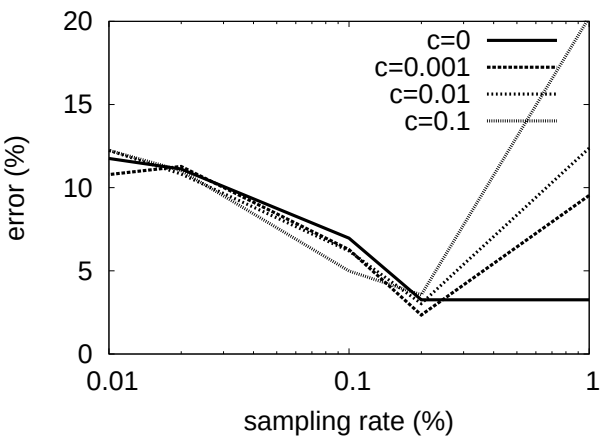

図 6　次数和のの概算 (orkut)

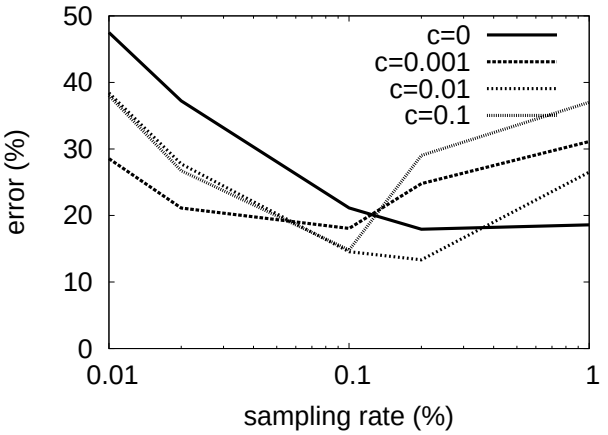

図 7　次数和のの概算 (LliveJournal)

対して Facebook のサンプルでは他のものと比較してグラフのサイズが小さいにもかかわらず，正確な値を取得するに至っていない．

5.2.2 グラフの次数の和の概算

次数和の精度はノードサイズの精度と相関があるが，α の値の取り方によって定数倍の誤差が生じる．そのためノードサイズの概算よりも誤差が大きくなっている．

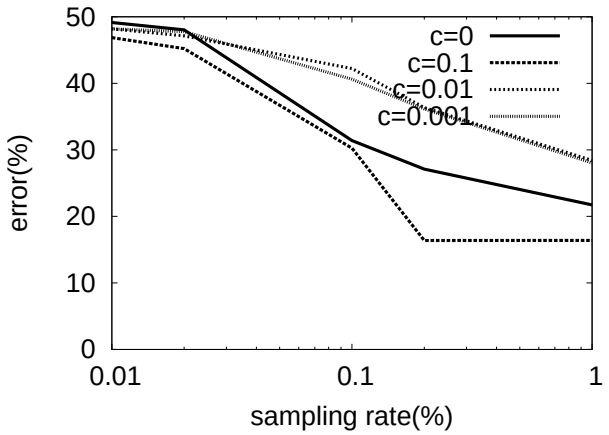

図 8　次数和のの概算 (Facebook)

表 2　InducedSubgraphSampling での α の概算値の平均

	真値	c=0	c=0.001	c=0.01	c=0.1
orkut	1.6	0.71	0.65	0.53	0.41
LiveJournal	0.75	0.41	0.44	0.42	0.44
Facebook	2.43	1.8	1.65	1.34	0.9

表 3　StarSampling での α の概算値の平均

	真値	c=0	c=0.001	c=0.01	c=0.1
orkut	1.6	1.54	1.60	1.55	1.56
LiveJournal	0.75	0.71	0.63	1.43	1.1
Facebook	2.43	1.86	1.21	1.53	1.43

5.2.3　次 数 分 布

各分布において回帰分析を行い，α の値を概算したところ表 2, 3 の通りになった．

何れのサンプルも c の値によらず Induced Subgraph Sampling にでは α の値はやや低めに出てしまっている．

Star Sampling では次数の低いノードの部分は確率分布はほぼ一致しているのに対して，次数の高いノードが極端に少ない．しかしながら，InducedSubgraphSampling よりも正確な α の値を得ることができている．

6. 考　　察

orkut, LiveJournal, Facebook のデータセットといった複雑ネットワークにおいてグラフのサイズといった特徴量分析を行った．その結果，orkut と LiveJournal のデータセットでは，0.1%程度のサンプル数で最も正確な値を得られることができた．しかし Facebook のデータセットでは，最もノード数が少ないにもかかわらず，正確な値を得るにはより多くのノードをサンプリングする必要がある．

次数分布の解析においては Induced Subgraph Sampling はサンプルサイズに対して多くのノードを部分グラフの中に構築することができる．しかし，サンプリングされていないノードが大量に部分グラフに残ってしまうため次数分布においては StarSampling と比べて精度が劣っている．

7. ま と め

本研究では UIS と RW を組み合わせた手法を提案し, その手法で各ノードが選択される確率を算出した．また, 提案手法を用いて実際のソーシャルグラフの特徴量を求め, 真値との比較を行った．次数分布を求めることにおいては，本研究では非線形回帰を用いて概算を行った．その結果 RW も UIS も適切に計算を行えば，どちらも同程度の精度でグラフの特徴量を求められることがわかった．しかしながら，データセットの性質によって精度の違いが露呈することもわかった．データセットの性質によらず常に正確な値を得る評価手法の提案を今後の課題とする．

謝　　辞

本研究は科研費 (24650025) および独立行政法人情報通信機構（NICT）委託研究「新世代ネットワークを支えるネットワーク仮想化基盤技術の研究開発」の助成を受けたものである．

文　　献

[1] W. Hastings , Monte Carlo Sampling Methods Using Markov Chains and Their Applications , Biometrika, 1970.
[2] J Leskovec , C Faloutsos , Sampling from Large Graphs , Proc. KDD'06, 2006.
[3] Liran Katzir, Edo Liberty, Oren Somekh, Estimating Sizes of Social Networks via Biased Sampling, Proc. WWW'11, 2011.
[4] Stanford Large Network Dataset Collection, http://konect.uni-koblenz.de/networks/orkut-links
[5] orkut Network Dataset-KONECT, http://snap.stanford.edu/data/index.html