様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ５ 月 ２５ 日現在

機関番号：１２６０８
研究種目：挑戦的萌芽研究
研究期間：2014～2015
課題番号：２６５４０１６１

研究課題名（和文）特徴量の高精度な推定を可能にする大規模グラフのサンプリング手法

研究課題名（英文）Graph sampling techniques for precise estimation of large-scale graph measures

研究代表者
首藤　一幸（Shudo, Kazuyuki）

東京工業大学・情報理工学（系）研究科・准教授

研究者番号：９０３０８２７１

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：オンラインソーシャルネットワークといった、規模や入手性を理由として全体の解析が現実的でないグラフに対して、全体の特徴量を推定するために、グラフサンプリングが行われる。我々は2通りのアプローチで、推定精度の向上を達成した。第一のアプローチは、対象のグラフが複雑ネットワークであることを仮定してその仮定を活用することである。第二のアプローチは、通常のランダムウォークをnon-backtrackingランダムウォークに置き換えることである。後者の手法は、既存手法と比較して、同一のサンプル頂点数を収集するために必要なステップ数を減らし、なおかつ、同一のサンプル頂点数で比較してもより高い精度を達成した。

研究成果の概要（英文）：Graph sampling is an effective approach to estimate measures of graphs in case it is not possible to analyze an etire graph for reasons such as difficulty in obtaining and its great magnitude. We took two approaches to improve precision of the estimation. An approach is assuming a target graph to be a complex network and utilizing the assumption. Another approach is replacing normal random walk with non-backtracking random walk. The latter approach reduced the number of sampling steps to collect a certain number of vertexes in comparison to the existing best technique and improved the precision even with the same number of sampled vertexes.

研究分野：大規模データ処理

キーワード：グラフサンプリング　大規模グラフ

１．研究開始当初の背景

交通網や流通網、電力網、ウェブ、友人関係、タンパク質相互作用など、社会構造や自然現象の多くがグラフとしてモデル化され、解析される。グラフの性質は、次数分布、平均距離、クラスタ係数といった特徴量として表され、それによって、性質の直感的な把握やグラフ間の比較が可能となる。

昨今、ソーシャルネットワーキングサービス（SNS）上の友人関係など、10^9（10 億）～規模のグラフが解析の対象として浮上してきたことなどを受け、大規模なグラフを扱える高速・省メモリなアルゴリズムや処理システムの研究が盛んに行われている [1-4]。その一方、グラフ全体を解析の対象とできない場合は多い。例えば SNS であれば、Facebook の 10^9～ユーザと友人関係をスクレイピングして手元に持ってくることは、時間的にとても現実的ではない。SNS 運営側が許可していないことも多い。また、もし手元にグラフ全体があったとしても、10^9 といった規模のグラフは現実的な時間での解析は困難かまたは不可能である（ゆえに盛んに研究されている）。

全体の入手や解析が現実的でない規模のグラフに対しては、一部の頂点・辺を抽出してそれを解析するというアプローチがある。つまり、グラフサンプリングである（図１）。例えば、サンプリングによって解析対象をグラフ全体の 0.1%とできたなら、時間計算量が O(n^4)（n は頂点数）である特徴量（例：δ-hyperbolicity）の場合、解析に要する時間は１兆分の１で済むことになる。また、同じ時間をかけるのであれば、大きさが１兆倍のグラフを解析できることになる。このとき、グラフサンプリングによる解析の課題は、精度、つまり結果の正確さということになる。

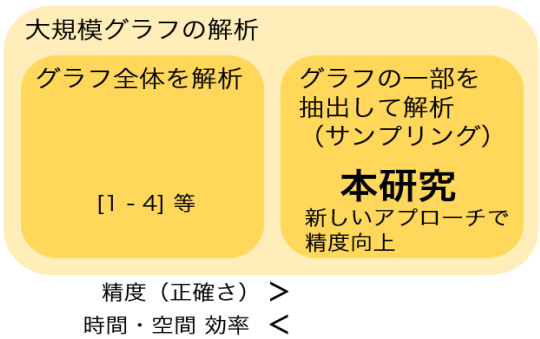

図１　グラフサンプリングの位置付け

２．研究の目的

グラフサンプリングの結果として得る特徴量の精度を向上させる。サンプル対象をグラフ全体、100%とすれば、得られる特徴量が真値となることは当然である。そこで、いかにサンプル対象の頂点・辺の数、またはサンプルの手間（データ取得回数など）を抑えつつ精度を向上させるか、が課題となる。

３．研究の方法

[方法 1]

サンプリングの手法は２つに大別される。グラフ全体からランダムに任意の点・頂点を選んでいく手法と、ランダムまたは次数等に応じた偏りをもって頂点・辺を辿り集めていく手法である。どちらにしても、従来手法では、グラフの性質について何の仮定も置かずにサンプリングを行う。そもそも性質を調べるためにサンプリングを行うのだから、当然のことである。しかしここで、グラフの性質について事前知識がある場合、または、何らかの予測が立つ場合、それを考慮に入れたサンプリングを行うことで、全体の性質により近い性質を持つ頂点・辺群を抽出できる可能性がある。つまり、より高い精度で特徴量を推定できる可能性がある。

[方法 2]

頂点・辺を辿って集めていく方法には、幅優先探索（BFS）に基づく方法や、ランダムウォークに基づく方法がある。ランダムウォークに基づく手法では、明に防がない限りは、すでにサンプルした点・頂点を再びサンプルしてしまうということが起こる。これが起こると、手間をかけたにも関わらず新たなサンプルを得られないこととなり、その手間が無駄となってしまう。こうした再訪 / 再収集を防ぐことで、手間に対して収集結果のグラフ規模を大きくできる。

４．研究成果

[方法 1]

複雑ネットワークを対象としたグラフサンプリング手法 [5] の問題を解決し、精度を向上させた。

我々の提案手法 [5] は、複雑ネットワーク生成モデルであるバラバシ・アルバートモデル（BA モデル）に倣った手順でサンプリングを行うというものであり、一部の特徴量において従来手法より高い精度を達成した。しかし一方で、サンプリングを始める頂点の選び方や、乱数を用いたサンプリングの進み方によっては、サンプリングの継続ができなくなることがある、という問題があった。つまり、サンプル済み頂点群との間の辺の数がちょうど k となる未サンプル頂点を収集する、という、BA モデルに忠実な手法であるため、そうした、ちょうど k となる頂点がなくなってしまうと、そこでサンプリングを止めざるを得なかった。

本研究ではこの問題を解決した。まず、サンプル済み頂点群に隣接するすべての未サンプル頂点を次のサンプル候補とする。次に、サンプル候補の各頂点について、隣接するサンプル済み頂点を数え、それと k に基いてサンプル候補の各頂点に確率（合計 1）を割り

当てる。こうして決めた確率に応じて、次のサンプル頂点を選ぶ。

こうして得たグラフに対して特徴量を求めたところ、その精度は、我々の元手法 [5] と遜色ないものであった。

本成果は、今後、発表予定である。

[方法 2]

我々は、オンラインソーシャルネットワークに対するクラスタ係数推定手法として、現在、もっとも効率的である Katzir らの手法 [6] を発展させ、同一の手間でのより多くのサンプル頂点数の収集、さらに、同一サンプル頂点数でのより高い精度を達成した [7]。

オンラインソーシャルネットワークとは、例えば Facebook といった SNS を指す語であり、ソーシャルグラフ全体の入手は不可能または極めて困難であってネットワーク越しにユーザ情報（頂点 + 隣接する辺&頂点）1 つずつにアクセスしていく他ないことを強調した語である。オンラインソーシャルネットワークでは、ランダムに生成したユーザ名やユーザ ID が存在するとは限らず、ということは、ランダムサンプリングは現実的ではない。例えば、ランダムに生成したユーザ名や ID 1,000 のうち 1 つが実際に存在するとして、1 つのユーザ名/ID を試すためにかかる時間が 10 ミリ秒だとすると、1 つのユーザ情報を得るために 10 秒を要することとなる。これは現実的ではない。オンラインソーシャルネットワークに対しては、BFS やランダムウォークといった頂点・辺を辿っていく手法を採るしかない。

Katzir らの手法 [6] は、頂点・辺を辿ることでクラスタ係数を推定する手法として、現在、もっとも効率的である。効率的とは、同一のサンプル頂点数で比較して精度がよいことや、同一の精度をより少ないサンプル頂点数で達成することを指す。我々は、この Katzir らの手法を発展させ、次を達成した。

(1) 同一の手間でより多くのサンプル頂点を収集できる。サンプル頂点数が多いということは、推定の精度も高くなる。
(2) さらに、同一のサンプル頂点数で比較して、より高い精度で推定できる。

Katzir らの手法は、グラフをランダムウォークで辿りながら、2 ステップ前の頂点、1 ステップ前の頂点、現頂点が三角形をなした場合の数を数えていく。この数に基いてクラスタ係数の推定値を算出する。

これに対して我々の手法 [7] は、通常のランダムウォークではなく、1 ステップ前の頂点には戻らない non-backtracking ランダムウォークを用いる。Katzir らの手法では、もともと、三角形をなすか否かを判断するために 2 ステップ前までの頂点を記憶する。そのため、1 ステップ前の頂点に戻ることを防ぐための新たな記憶、手間は発生しない。1 ステップ前の頂点はサンプル済み頂点であるため、そこに戻らないことで、ステップ数に対するサンプル頂点数を増やすことができる。

ここまでは、効果的ではあるが、工夫はごく単純であり、効率化は当然の結果であるように見えるだろう。我々の手法は、以上に加えて、同一のサンプル頂点数（少ないステップ数）で比較して、Katzir らの手法より高い精度を達成する。

Katzir らの手法と我々の手法を、Stanford Network Analysis Project（SNAP）[8] が提供するグラフ群に適用した。表 1 は、10000 頂点の収集に要したステップ数、および、Katzir らの手法でのステップ数に対する我々の手法で要したステップ数の比である。Katzir らの手法に対して、73.8% ～ 93.0% のステップ数で同一のサンプル頂点数を収集できている。

ネットワーク	Katzir らの手法	我々の手法
Amazon	20308	14981 73.8%
DBLP	16144	14232 88.2%
Gowalla	14758	13406 90.1%
Live Journal	11417	10616 93.0%

表 1 10000 頂点の収集に要したステップ数の平均

図 2 ～ 5 は、Katzir らの手法、および、我々の手法における、サンプル頂点数に対する精度（NRMSE）である。同一のサンプル頂点数で比較して、我々の手法は精度を大きく改善していることを見てとれる。

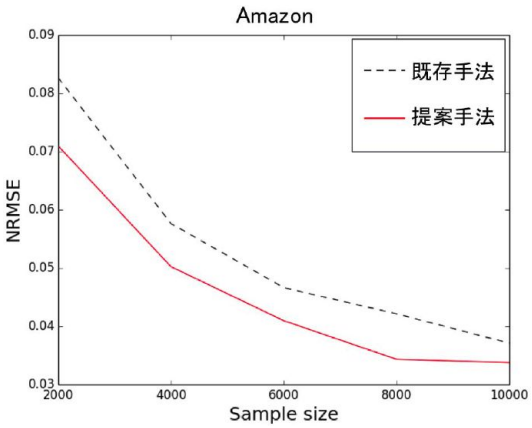

図 2 サンプル頂点数に対する精度：Amazon

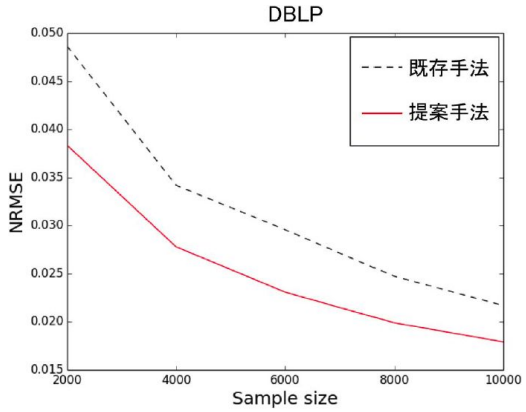

図３　サンプル頂点数に対する精度：DBLP

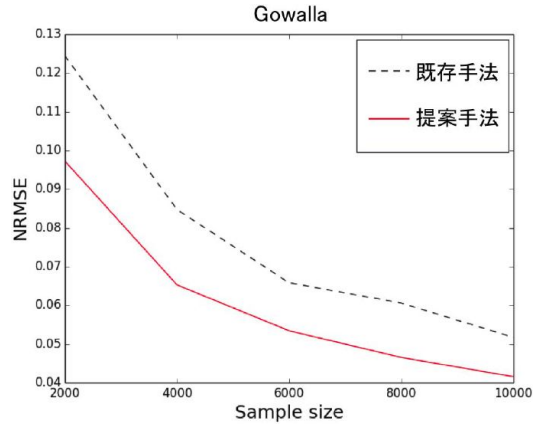

図４　サンプル頂点数に対する精度:Gowalla

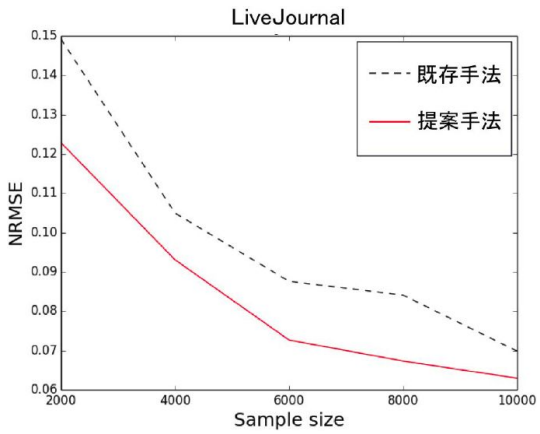

図５　サンプル頂点数に対する精度：Live Journal

[1] 科学技術振興機構 ERATO 河原林巨大グラフプロジェクト，2012 年 10 月，http://www.jst.go.jp/erato/kawarabayashi/
[2] 科学技術振興機構 CREST ポストペタスケールにおける超大規模グラフ最適化基盤，2011 年，http://www.graphcrest.jp/
[3] G. Malewicz, M. Austern, A. Bik, J. Dehnert, and I. Horn: Pregel: A System for Large-Scale Graph Processing, Proc. OSDI'10, October 2010.
[4] U. Kang, C. E. Tsourakakis, and C. Faloutsos: PEGASUS: A Peta-Scale Graph Mining System Implementation and Observations, Proc. ICDM'09, December 2009.
[5] 宇都宮健太，首藤一幸：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，情報処理学会 研究報告，2013-DBS-158(-9),2013 年 11 月 26 日.
[6] Liran Katzir, Stephen J. Hardiman: Estimating Clustering Coefficients and Size of Social Networks via Random Walk, ACM Transactions on the Web, Vol.9, Issue 4, pp.19:1-19:20, October 2015.
[7] 岩﨑謙汰，華井雅俊，首藤一幸：ソーシャルグラフ向けクラスタ係数推定手法の効率化，第 8 回データ工学と情報マネジメントに関するフォーラム（DEIM2016），2016 年 2 月 29 日～3 月 2 日.
[8] Stanford Large Network Dataset Collection. http://snap.stanford.edu/data/

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計 2 件）

岩﨑謙汰、華井雅俊、<u>首藤一幸</u>、ソーシャルグラフ向けクラスタ係数推定手法の効率化、第 8 回データ工学と情報マネジメントに関するフォーラム（DEIM2016）、福岡、2016 年 2 月 29 日～3 月 2 日

岩﨑謙汰、華井雅俊、<u>首藤一幸</u>、ソーシャルグラフ向けクラスタ係数推定手法の効率化、第 8 回広域センサネットワークとオーバレイネットワークに関するワークショップ、東京、2016 年 3 月 22 日～23 日

６．研究組織

(1)研究代表者

首藤　一幸（SHUDO KAZUYUKI）
東京工業大学・大学院情報理工学研究科・准教授
研究者番号：90308271

(2)研究分担者

秋岡　明香（AKIOKA SAYAKA）
明治大学・総合数理学部・准教授
研究者番号：90333533

(3)連携研究者