DEIM2020 C3-2(day1 p18)

ランダムウォークサンプリングによるソーシャルグラフの復元

中嶋　一貴†　　首藤　一幸†

† 東京工業大学 情報理工学院 数理・計算科学系

あらまし　ソーシャルネットワークのグラフ構造を正確に分析する主要な目的は，社会構造の性質の発見や理解および現実的なグラフデータを公に利用可能にすることである．しかし，大規模なグラフデータのアクセス制限のために，その正確な構造分析は依然として挑戦的な課題である．我々は，この課題を克服するために，ランダムウォークサンプリングによるソーシャルグラフの復元アルゴリズムを提案する．提案するアルゴリズムは，まずランダムウォークでサンプリングされたグラフデータで構成される部分グラフを生成し，次に，サンプルから推定した局所的な統計量が満たされるように，サンプリングされていないグラフデータを補間する．実際のソーシャルネットワークデータセットを使用した実験により，提案するアルゴリズムは，既存のアルゴリズムより誤差の小さい複数の統計量を持つグラフを生成することを示す．

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，ランダムウォーク，グラフ復元

1 はじめに

ソーシャルネットワークのグラフ構造を正確に分析する主要な目的は，社会構造の特性の発見や理解および現実的なグラフデータを公に利用可能にすることである．これまで，様々なソーシャルネットワークの次数分布やクラスタ係数などのグラフ統計量が分析され，その構造的特性が明らかになってきた [1–7]．構造分析による特性の発見は，その特性を満たすグラフを生成するモデルの研究 [2,7] に発展し，ネットワークの生成原理の理解につながる．ソーシャルグラフ分析のために収集されたグラフデータはよくデータセットとして公開される [8–10]．公開されたデータセットは，ソーシャルネットワーク上の影響拡散 [11] や表現学習 [12] などのグラフアルゴリズムの研究分野で活用されている．

しかしながら，ソーシャルグラフの正確な構造分析は，そのグラフデータの規模の大きさと厳しいアクセス制限のために依然として挑戦的な課題である．いくつかの研究 [1,13,14] は，完全なソーシャルグラフのデータを使用してその構造を分析しているが，研究者などの第三者が全てのデータを取得することは現実的ではない．実用的なシナリオでは，そのアプリケーション・プログラミング・インターフェース (API) を通じて一部のグラフデータをサンプリングして，そのグラフ構造を分析する [1,3–5]．

サンプリングを通じたソーシャルグラフの基本的な分析手法は，幅優先探索などのクローリング手法によるサンプルから得られる部分グラフを分析することである [1,5,6]．多くのソーシャルネットワークは，あるユーザの隣接情報を返す API を提供してきており，隣人を辿るクローリング手法を適用できる．特に幅優先探索は，一部の範囲の完全なグラフデータをサンプルでき，その部分グラフはしばしば元のグラフの代表的なサンプルとして考えられている [6]．一方で，幅優先探索による少量のサンプルから生成される部分グラフには，修正が困難な統計量のバイアスが発生する [3]．このバイアスを取り除くために，通常ほとんど全てのグラフデータをサンプルする必要がある．

再重み付けランダムウォーク法 [3] は，グラフ統計量の不偏推定量を得るための有効なアプローチである．ソーシャルグラフ上のランダムウォークによって得られる各サンプルに対して，マルコフ性に基づくそのサンプリングのバイアスを修正するための重み付けを施し，グラフ統計量の不偏推定量を得る．これまで，様々な統計量に対して再重み付けランダムウォーク法による推定アルゴリズムが提案されてきた [15–23]．一方で，再重み付けランダムウォーク法は，グラフ内で大域的に定義される統計量の推定やグラフ生成に適用できないという問題点がある．再重み付けランダムウォーク法は，サンプルした各ノードに対して各統計量特有の重み関数を計算する必要がある．統計量の定義範囲が広くなるにつれて，サンプルしたグラフデータだけで計算できる重み関数を定義することが困難となり，追加のサンプリングを許容せざるを得ない．さらに，再重み付けランダムウォーク法は，特定の統計量の推定値を返すアルゴリズムであり，サンプルからグラフを生成する目的には適用できない．

Gjoka らは，ランダムウォークによるサンプルからグラフ構造を分析するアプローチとして，dK シリーズという枠組みに基づく $2.5K$ グラフ生成アルゴリズムを提案した [24]．dK シリーズとは，対象とするグラフ内の d 個のノードから成る部分グラフ集合の次数分布を満たす dK グラフを生成する枠組みである．d の値を大きくするにつれて，元のグラフをより正確に模倣するグラフを生成することができる．現実のグラフでは，d の値が 2, 3 程度で複数の統計量を正確に模倣するグラフを生成できることが知られている [25]．2.5K グラフアルゴリズムは，はじめに，再重み付けランダムウォークでジョイント次数分布とクラスタ係数の 2 つの統計量を推定する．次に，推定

したジョイント次数分布を入力として，dK シリーズの枠組みを用いて $2K$ グラフを生成する．最後に，生成グラフのクラスタ係数が推定値の許容誤差範囲内になるようにエッジをスワップする．2.5K アルゴリズムは，局所的な統計量を再重み付けランダムウォーク法によって推定し，大域的な統計量を dK シリーズの枠組みによって正確に模倣するグラフを生成する．一方で，dK シリーズによる生成グラフの統計量は一致性を持たない，つまりサンプル数を十分に増やしても元のグラフに一致しない．これは，dK シリーズによる生成グラフはあくまで元グラフを模倣しているだけで，元のグラフデータが含まれていないためである．

本研究では，ランダムウォークのサンプルから元のグラフに似たグラフを生成する，すなわち復元することを目標として，部分グラフ，再重み付けランダムウォーク法，そして dK シリーズを組み合わせたグラフ生成アルゴリズムを設計する．はじめに，ランダムウォークで得たサンプルから部分グラフを生成する．次に，再重み付けランダムウォークでグラフ生成に必要な統計量を推定する．最後に，推定した統計量を満たすように，部分グラフにグラフデータを補間する．提案する生成アルゴリズムは，既存の両者のアルゴリズムの利点を持つ．すなわち，提案アルゴリズムは，サンプル数を増やすにつれて元のグラフに近づいていく性質を持ち，少ないサンプル数で複数の統計量を正確に捉えるグラフを生成する最初のアルゴリズムである．我々は，本研究の目標達成に向けて，部分グラフと $1K$ グラフを組み合わせた生成アルゴリズムを提案する．実際のソーシャルネットワークのデータセットを用いた実験により，提案アルゴリズムが少ないサンプル数で複数の統計量を既存アルゴリズムより正確に捉えるグラフを生成し，かつサンプル数を増やしていくにつれて統計量の平均誤差が減少し，元のグラフに収束していくことを確認した．

2 準　　備

本研究では，ソーシャルネットワークを連結で重み無しの無向グラフ $G = (V, E)$ で表す．ここで，$V = \{v_1, ..., v_n\}$ を n 個のノード (ユーザ) の集合，E をエッジ (ユーザ間の友好関係) の集合とする．ノード v_i の隣接ノード集合を $N(i) = \{v_j \in V | (v_i, v_j) \in E\}$, ノード v_i の次数を $d_i = |N(i)|$, 次数の総和を $D = \sum_{v_i \in V} d_i$ と表す．次数 d を持つノードの集合を $V(d) = \{v_i \in V | d_i = d\}$, $V(d)$ の要素数を $n(d) = |V(d)|$ とする．グラフ G の次数分布を $\{P(d) = \frac{n(d)}{n}\}_d$ と定義する．既存研究 [3] に基づいて以下の 2 つを仮定する: (1) ノード v_i のインデックス i をクエリして v_i の隣接ノード集合を取得できる, (2) グラフデータをサンプリングする間はグラフは静的である．

2.1 問 題 定 義

本研究は，グラフ G 上のランダムウォークによってサンプリングされたグラフデータから，可能な限り多くの統計量が G の統計量と小さい誤差を持つグラフ $\tilde{G}$ を生成する問題を扱う．この問題は，以下の 2 つの部分に分かれる．

（1） 統計量推定: ランダムウォークのサンプルからグラフ生成に必要な統計量を推定する．

（2） グラフ生成: ランダムウォークのサンプルと統計量の推定値を入力としてグラフ $\tilde{G}$ を生成する．

2.2 着目する統計量

以下の 9 個の統計量に着目する．

（1） ノード数 ($n = |V|$)

（2） エッジ数 ($m = |E|$)

（3） 次数分布 (degree distribution, DD)

（4） クラスタ係数 (Average Clustering Coefficient, ACC): 各ノードに接続されている三角形の割合の平均値 [7].

（5） 次数ごとのクラスタ係数 (Degree-Dependent Clustering Coefficient, $DDCC$): 次数 k のノードのクラスタ係数の平均値 [1, 24].

（6） 次数相関 (degree correlation, Knn): 次数 d を持つノードの隣接ノードの平均次数 [26].

（7） 平均最短距離 (Average Shortest Path Length, APL): ノード間の最短距離の平均値．

（8） 直径 (Diameter, D): ノード間の最短距離の最大値．

（9） 最短距離分布 (Shortest Path Distribution, SPD): ノード間の最短距離の分布．

2.3 サンプリングと統計量の推定

はじめに，グラフ G 上でランダムウォークを実行してグラフデータをサンプリングする．グラフ G 上の r ステップのランダムウォークは以下のように実行される: 任意の初期ノードを選び，隣接ノード集合からランダムに 1 つ選び遷移することを $r-1$ 回繰り返す．r 個のサンプルノードのインデックスと隣接ノード集合の列 $R = \{(x_k, N(x_k))\}_{k=1}^{r}$ を得る．ここで，x_k は k 番目に遷移したノードのインデックスである．サンプル v_{x_k} の次数は $d_{x_k} = |N(x_k)|$ と計算できる．

次に，サンプル列 R からグラフ生成に必要な統計量を再重み付けランダムウォークアルゴリズムによって推定する．再重み付けランダムウォークアルゴリズムは，推定する統計量に応じた重み付けを各サンプルに施してサンプリングの偏りを除去し，統計量の不偏推定値を得る [3]．我々は，ノード数 n と次数分布 $P(d)$ を推定するための既存アルゴリズム [3, 19] を適用する．

ノード衝突アルゴリズム [18, 19] は，ランダムウォークのサンプルからグラフのノード数 n を推定するための有効な推定アルゴリズムである．このアルゴリズムは，十分なステップ離れたサンプルのペアのインデックスの一致数からノード数を推定する．1 以上 r 以下で m 以上離れた整数のペアの集合を $I = \{(k, l) \mid m \leqq |k - l| \wedge 1 \leqq k, l \leqq r\}$ とおく．$\phi_{k,l} = 1_{\{x_k = x_l\}}$ を k 番目と l 番目のサンプルのインデックスが一致するときに 1, 一致しないときに 0 を返す変数とする．サンプルペアのインデックスの一致数の平均値 Φ_n，サンプリングの偏りを除去するための重みの平均値 Ψ_n，ノード数の推定

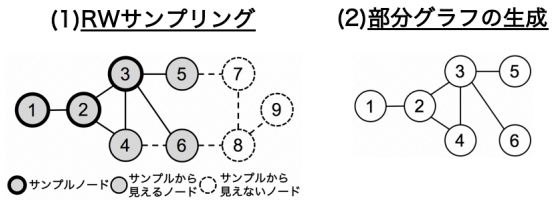


図 1　ノード 1, 2, 3 をサンプルした時に生成される部分グラフ.

値 $\tilde{n}$ はそれぞれ以下のように定義される:

$$\Phi_n = \frac{1}{|I|}\sum_{(k,l)\in I}\phi_{k,l},\ \Psi_n = \frac{1}{|I|}\sum_{(k,l)\in I}\frac{d_{x_k}}{d_{x_l}},\ \tilde{n} \triangleq \frac{\Psi_n}{\Phi_n^{NC}}.$$

既存研究 [18,19] より以下の補題が成り立つ.

補題 1. *[18, 19]* $r \to \infty$ のとき，推定値 $\tilde{n}$ は漸近的に真値 n に収束する.

次数分布の推定アルゴリズム [3] を述べる．次数 d を持つサンプルの個数とサンプリングの偏りを除去するための重みの平均値である 2 つの変数 Φ_d, Ψ_d と次数 d の分布 $P(d)$ の推定値 $\tilde{P}(d)$ はそれぞれ以下のように定義される:

$$\Phi_d = \frac{1}{r}\sum_{i}^{r}\frac{1}{d_{x_k}}1_{\{d_{x_k}=d\}},\ \Psi_d = \frac{1}{r}\sum_{i}^{r}\frac{1}{d_{x_k}},\ \tilde{P}(d) \triangleq \frac{\Phi_d}{\Psi_d}.$$

既存研究 [27] より以下の補題が成り立つ.

補題 2. *[27]* $r \to \infty$ のとき，推定値 $\tilde{P}(d)$ は漸近的に真値 $P(d)$ に収束する.

2.4　コンフィグレーション・モデル

コンフィグレーション・モデルは，与えられた n 個の次数の列 $\{d_1, d_2, ..., d_n\}$ からランダムにネットワークを生成する [28,29]. 入力の次数列を調整することで，生成されたグラフが任意の次数分布を満たすようにできる．アルゴリズムは以下の通りである.

（1） n 個の各ノード v_i に次数 d_i を割り当て，d_i 個のエッジの片割れを持つようにする.

（2） 2 つのエッジの片割れを一様ランダムに選び接続する.

（3） エッジの片割れがなくなるまで (2) を続ける.

上記のアルゴリズムが正常に終了するために，次数列の和 $\sum_{i=1}^{n} d_i$ が偶数である必要がある [29].

dK シリーズの枠組みに倣うと，コンフィグレーション・モデルは 1K グラフに相当する．以後，コンフィグレーション・モデルを 1K グラフと呼ぶ.

3　ソーシャルグラフ生成

3.1　部分グラフ生成アルゴリズム

サンプリングしたグラフデータからグラフを復元する素朴なアプローチは，サンプルしたノードとその隣接ノード集合のグラフデータの和集合をとる部分グラフ生成アルゴリズムである [1,5,30,31]．サンプル列 $R = \{(x_k, N(x_k))\}_{k=1}^{r}$ から $\tilde{V}_{sub} = \bigcup_{k=1}^{r}\{v_{x_k} \cup N(x_k)\}$, $\tilde{E}_{sub} = \bigcup_{k=1}^{r}\bigcup_{w\in N(x_k)}(v_{x_k}, w)$ とするグラフ $\tilde{G}_{sub} = (\tilde{V}_{sub}, \tilde{E}_{sub})$ を生成する.

図 1 は，あるグラフ上のランダムウォークでノード 1, 2, 3 をサンプルした場合に生成される部分グラフを示している．サンプルしたノードの隣接ノード集合 (図 1 のノード 4, 5, 6) はランダムウォークで遷移していなくとも観測することができるため，それらのグラフデータも部分グラフに含めることができる.

部分グラフ生成アルゴリズムの利点は，生成されたグラフが実際のグラフデータから成ることである．部分グラフであるという性質から，サンプルを十分に増やすと生成されるグラフは元のグラフと一致する.

部分グラフ生成アルゴリズムの問題点は，サンプル数が少ないときに，サンプルされていないグラフデータの欠如によって生成されるグラフの統計量に比較的大きな誤差が生じることである．例えば，ノード数が想像に容易い．図 1 の例では，サンプルされていないノード 7, 8, 9 を補填することができず，大きい誤差を伴う.

Algorithm 1 Generating the realizable sequence of number of nodes with each degree of d.

Input: Estimates $\tilde{n}$ and $\{\tilde{P}(d)\}_d$
Output: Sequence $\{\hat{n}(d)\}_d$
1: **for** each degree d such that $\tilde{P}(d) > 0$ **do**
2:　$\hat{n}(d) \leftarrow max(round(\tilde{n} \times \tilde{P}(d)), 1)$
3: **end for**
4: **if** $\sum_d d \times \hat{n}(d)$ is not even **then**
5:　add 1 to $\hat{n}(1)$
6: **end if**
7: **return** $\{\hat{n}(d)\}_d$

3.2　1K グラフによる生成アルゴリズム

誤差の小さいノード数と次数分布を持つグラフを生成するように，1K グラフを導入する．1K グラフは，入力の次数列を満たすグラフを生成する．2.3 章のノード数と次数分布の不偏推定量 $\tilde{n}, \{\tilde{P}(d)\}_d$ から次数列を作成し，それを入力とするコンフィグレーション・モデルを用いてグラフを生成する.

次数列を作成するために，ノード数と次数分布の不偏推定量 $\tilde{n}, \{\tilde{P}(d)\}_d$ から，実現可能な各次数 d のノード数の列 $\{n(d)\}_d$ を作成する．ここで，実現可能な $\{n(d)\}_d$ は以下の条件を満たす時に言う:

（1） 各次数 d に対して，$n(d)$ は非負な整数である.

（2） $\sum_d d \times n(d)$ は偶数である.

ノード数と次数分布の不偏推定量 $\tilde{n}, \{\tilde{P}(d)\}_d$ から，実現可能な各次数 d のノード数の列 $\{\hat{n}(d)\}_d$ を作成するアルゴリズムを Algorithm 1 に示す．*round* 関数は端数処理関数で，偶数への丸め (round to even) を採用する．$max(a, b)$ は a, b の大きい値を返す関数である.

条件 1 は 2 行目の処理によって満たされる．$\hat{n}(d)$ を 1 以上の整数とするのは，2.3 章の次数分布の推定値の定義から，正の推定値 $\tilde{P}(d)$ が得られたとき，グラフ G 内に少なくとも 1 つ

の次数 d を持つノードが存在することを考慮するためである．

条件 2 は 6 行目の処理によって満たされる．次数和が奇数であるとき，和が偶数となるようにいくつかの次数の個数を調整する必要がある．次数 d の個数を 1 増やす (減らす) と，ノード数とエッジ数はそれぞれ 1, $\frac{d}{2}$ だけ増える (減る)．可能な限り，ノード数とエッジ数増減を小さくして条件 2 を満たすようにする．ここでは，次数 1 のノードが存在しない場合を考慮して，次数 1 のノードを 1 個だけ増やす調整を採用する．結果として，ノード数とエッジ数はそれぞれ 1 増加する．

実現可能な各次数 d のノード数列 $\{\hat{n}(d)\}_d$ を作成し，各次数 d を $\hat{n}(d)$ 個ずつ含めた次数列を作成する．そして，2.4 章のアルゴリズムにしたがって，グラフを生成する．

1K グラフによる生成アルゴリズムの利点は，誤差の小さいノード数と次数分布を持つグラフを生成することである．再重み付けランダムウォークアルゴリズムは，少ないサンプル数でノード数と次数分布の正確な推定値を提供する．コンフィグレーション・モデルによる生成アルゴリズムは，それらの推定値をほとんど正確に満たすグラフを生成する．

1K グラフによる生成アルゴリズムの問題点は，2 つある．1 つ目の問題点は，ノード数と次数分布以外の統計量がモデルに大きく依存することである．例えば 1K グラフは，ランダムにエッジを張るため，次数相関はほとんど現れない [32]．一方で，ソーシャルネットワークは次数相関がしばしばあられる [1,3,4]．また，クラスタ係数は $C = \frac{\left(\frac{1}{n}\sum_{v_i \in V}(d_i^2 - d_i)\right)^2}{n\left(\frac{1}{n}\sum_{v_i \in V} d_i\right)^3}$ と導出され [33]，現実のネットワークを常に模倣できるわけではない [29]．2 つ目の問題点は，サンプル数を十分に増やしても，元のグラフに近づいていくとは限らないことである．あくまで 1K グラフによって模倣される生成グラフであり，実際のグラフデータを含んでいないためである．サンプル数を増やすと，補題 1, 2 より，ノード数と次数分布の厳密値 $n, \{P(d)\}_d$ を入力とする 1K グラフによる生成グラフに近づく．

3.3 提案アルゴリズム

我々は，部分グラフ生成アルゴリズムと 1K グラフによる生成アルゴリズムを組み合わせた生成アルゴリズムを提案する．はじめに，ランダムウォークのサンプルから部分グラフを生成し，次に，ノード数と次数分布の推定値が満たされるように部分グラフにノードとエッジを補間する．

提案アルゴリズムは，2 つの利点を持つ．1 つ目の利点は，サンプル数を十分に増やしていくと元のグラフに一致することである．これは，生成されるグラフにサンプルされた実際のグラフデータが埋め込まれているためである．2 つ目の利点は，少ないサンプル数で，ノード数と次数分布以外の統計量の誤差を改善することである．これは，サンプルされていないグラフデータによる誤差の発生という部分グラフ生成アルゴリズムの問題点と，実際のグラフデータが埋め込まれていないというコンフィグレーション・モデルによる生成アルゴリズムの両者の問題点を解消するためである．

提案アルゴリズムの生成手順を述べる．提案アルゴリズムによる生成グラフを $\hat{G} = (\hat{V}, \hat{E})$ と表す．

はじめに，部分グラフ生成アルゴリズムによって部分グラフを生成する．生成する部分グラフの各ノードに「サンプルしたノード」と「サンプルから見えるノード」のラベルをつける．サンプルしたノードとサンプルから見えるノードの集合をそれぞれ $\tilde{V}_{sampled}, \tilde{V}_{visible}$ とおく．このとき，$\tilde{V}_{sub} = \tilde{V}_{sampled} \cup \tilde{V}_{visible}$ である．部分グラフ $\tilde{G}_{sub}$ における $v_i \in \tilde{V}_{sub}$ の次数を d'_i とおく．図 1 の例では，$V_{sampled} = \{1,2,3\}, V_{visible} = \{4,5,6\}, d'_1 = 1, d'_2 = 3, d'_3 = 4, d'_4 = 2, d'_5 = 1, d'_6 = 1$ である．また，各ノードに割り当てられる次数を $\hat{d}_i$ と表す．

第 2 に，Algorithm 1 に従って，実現可能な各次数 d のノード数の列 $\{\hat{n}(d)\}_d$ を作成する．ここで，ノード数の推定値は，$max(\tilde{n}, |\tilde{V}_{sub}|)$ を入力する．少なくとも，$|\tilde{V}_{sub}|$ 個のノードが存在するからである．

第 3 に，サンプルした各ノード $v_i \in \tilde{V}_{sampled}$ に次数を割り当て，$\{n(d)\}_d$ を更新する．サンプルしたノード v_i に接続する隣接ノードは全て取得できるため，部分グラフ内の次数は元グラフ内の次数と等しい，つまり $d'_i = d_i$ である．このため，次数 $\hat{d}_i = d'_i$ を割り当て，$\hat{n}(d) = max(\hat{n}(d) - 1, 0)$ として更新する．$\hat{n}(d)$ を 0 以上とする更新方法により，条件 1 が満たされる．

第 4 に，サンプルから見える各ノード $v_i \in \tilde{V}_{visible}$ の次数を割り当てる．サンプルから見えるノード v_i に接続されているエッジは実際のグラフデータであるため，部分グラフ内の次数は元のグラフ内の次数以下，つまり $d'_i \leqq d_i$ である．このため，$d'_i \leqq d$ かつ $n(d) > 0$ であるような次数 $\hat{d}_i = d$ を割り当て，$\hat{n}(d) = max(\hat{n}(d) - 1, 0)$ として更新する．

サンプルから見えるノードの次数を割り当てる順番は 1 つの問題である．後方に次数が割り当てられるノードは，条件 $d'_i \leqq d$ かつ $n(d) > 0$ を満たす次数 d が存在しない場合がある．この場合，次数 d'_i の個数 $n(d'_i)$ を 1 増やし，次数 $\hat{d}_i = d'_i$ を割り当てる．増やす次数の個数は割り当てる順番に依存する．

我々は，部分グラフ内の次数 d'_i が大きい次数順に次数を割り当てる方法を採用する．一般的に，ソーシャルネットワークの次数分布は低次数に偏っており [1,3–5]，大きい次数の個数は極めて少ない．このため，部分グラフ内で大きい次数を持つノードほど，割り当ての候補である次数が少なくなる．

第 5 に，追加される次数の和 $\sum_{v_i \in V_{sampled} \cup V_{visible}} (\hat{d}_i - d'_i) + \sum_d d \times n(d)$ が偶数であるようにする．3.2 章と同様に，次数 1 のノードが存在しない場合を考慮して，次数 1 のノードを 1 個だけ増やす調整を採用する．

第 6 に，更新された $\{\hat{n}(d)\}_d$ が満たされるように，ノードを追加して次数を割り当てる．$\sum_d \hat{n}(d)$ 個のノードを追加し，各次数 d が $\hat{n}(d)$ 個含む次数列を作成して，各追加ノードに次数を割り当てる．追加されたノード集合を V_{added} とおく．この時点で追加された各ノード $v_i \in V_{added}$ の次数は $d'_i = 0$ である．

最後に，各ノード $v_i \in V_{sampled} \cup V_{visible} \cup V_{added}$ に $\hat{d}_i - d'_i$ 個のエッジの片割れを持たせ，エッジの片割れがなくなるまでエッジを接続する．

表 1　データセット.

Network	ノード数	エッジ数	クラスタ係数
Anybeat [9]	12,645	49,132	0.204
Enron [10]	33,696	180,811	0.509
Facebook New Orleans [24]	63,392	816,884	0.222
Epinions [10]	75,877	405,739	0.138

4　実　　験

実際のソーシャルネットワークのデータセットを用いて，提案アルゴリズムを評価する．各生成アルゴリズムによる生成グラフと元のグラフの統計量の誤差を比較する．対象とする統計量は 2.2 章で述べた統計量である．

4.1　実験準備

データセット: 公開されているソーシャルネットワークの 4 個のデータセットを用いる．本研究の実験では，4 個の元のデータセットに以下の前処理を加えた連結な向こうグラフを用いる: (1) 元のグラフが有向グラフであれば，エッジの向きを削除する, (2) 元のグラフの最大連結成分に含まれないノードを削除する．表 4 に各グラフのノード数，エッジ数，クラスタ係数を示す.

誤差指標: 統計量の誤差指標は，正規化平方二乗誤差 (Normalized Root Mean Square Error, NRMSE) を用いる．NRMSE は既存研究 [24] に倣って，以下のように定義する: $NRMSE(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{\sum_{i=1}^{k}(x_i-\hat{x}_i)^2}{\sum_{i=1}^{k}x_i^2}$，ここで $\boldsymbol{x} \in \boldsymbol{R}^l, \hat{\boldsymbol{x}} \in \boldsymbol{R}^m$ はそれぞれ元のグラフと生成されたグラフの統計量の l 次元, m 次元ベクトルを表し，$k = max(l,m)$ とする．一方のベクトルのみに存在する要素は，もう一方の要素を 0 として誤差を計算する．全ての実験を独立に 100 回行い，NRMSE の平均値を計算する.

初期ノード選択: 全てのデータセットにおける以後の実験では，グラフ上のノードからランダムにランダムウォークの初期ノードを選択する．初期ノードは実験ごとに独立に選択される.

比較するアルゴリズム: 以下の 4 つのアルゴリズムを比較して，提案アルゴリズムの有効性を評価する:

- 幅優先探索サンプリングによる部分グラフ生成アルゴリズム: 幅優先探索は，ソーシャルネットワークのグラフデータをサンプリングする基本的な手法である [1,3,5,6].
- ランダムウォークサンプリングによる部分グラフ生成アルゴリズム (3.1 章).
- ランダムウォークサンプリングによる 1K グラフに基づく生成アルゴリズム (3.2 章).
- 提案アルゴリズム (3.3 章)

公正な比較のために，幅優先探索とランダムウォークの初期ノードを同じにし，ランダムウォークに基づく生成アルゴリズムは，同一のランダムウォークによるサンプルから生成する.

4.2　生成グラフの統計量精度

2 つの実験により，以下の設問に答える:

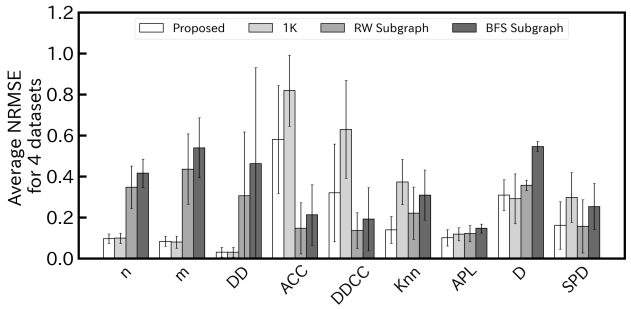


図 2　各統計量の NRMSE の 4 つのデータセット間の平均値 (サンプル率: 5%).

（1） 提案アルゴリズムによる生成グラフは，小さいサンプル率で各統計量の誤差を改善するか.

（2） 提案アルゴリズムによる生成グラフは，サンプル率を増加させると元のグラフに収束していくか.

4.3　小さいサンプル率における統計量の精度

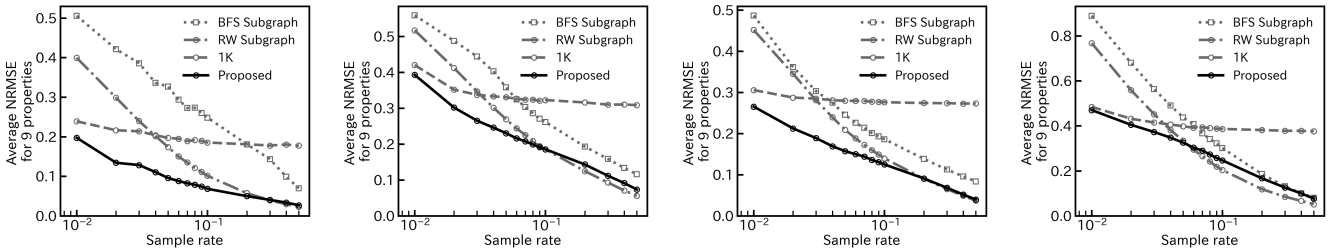
図 2 は，各生成グラフの 9 個の統計量の NRMSE の 4 つのデータセット間の平均値を示す．サンプル率は 5%である.

まず，提案アルゴリズムのノード数 (n)，エッジ数 (m)，次数分布 (DD) は，部分グラフのそれらより誤差と分散が十分に小さい．これは，提案アルゴリズムがノード数，エッジ数，次数分布の不偏推定量をほとんど満たすように生成するためである．さらに，提案アルゴリズムは，1K グラフと比較してノード数，エッジ数，次数分布の推定精度をほとんど維持していることがわかる．これは，提案アルゴリズムが生成過程でノード数とエッジ数の増減をできる限り小さくするように設計されているためであると考える.

次に，提案アルゴリズムのクラスタ係数 (ACC) と次数ごとのクラスタ係数 ($DDCC$) の誤差は，1K グラフのそれより小さいものの，部分グラフのそれより大きい．現実のグラフのクラスタ係数は，1K グラフによる生成グラフのクラスタ係数のそれより十分に大きいことが知られている [7]．サンプル率が小さい場合，提案アルゴリズムによる生成グラフには 1K グラフによるノードとエッジの追加数が多くなるため，生成グラフのクラスタ係数の誤差が発生すると考える.

提案アルゴリズムの次数相関 (Knn) の誤差は，部分グラフと 1K グラフのそれより小さい．1K グラフでは，ランダムに全てのエッジを繋ぐため，通常は次数相関が現れない [32]．部分グラフでは，実際のグラフデータを含むため元のグラフの隣接関係を反映するが，グラフデータの欠如によって接続されているノード間の次数相関を正確に反映することはできない．提案アルゴリズムは，部分グラフ内のノード間の隣接関係を保ったまま，ノードやエッジを追加していくため，より正確な次数相関を反映できると考える.

最後に，提案アルゴリズムの平均最短距離 (APL)，直径 (D)，最短距離分布 (SPD) の誤差は，部分グラフと 1K グラフのそれらと比較してほとんど改善が見られない．提案アルゴリズムは，1K グラフに倣ってノード数と次数分布をほとんど満たす



(a) Anybeat (b) Enron (c) Epinions (d) Facebook New Orleans

図 3　サンプル率を変化させたときの各生成グラフの 9 個の統計量の NRMSE の平均値.

ようなグラフを生成する．最短距離といった大域的な統計量をより正確に反映するためには，より多くの局所的な統計量を満たすグラフ生成モデルを適用する必要があると考える．

4.4　サンプル率を増やしたときの統計量の精度

図 3 は，4 つのデータセットにおいてサンプル率を変化させたときの各生成グラフの 9 個の統計量の NRMSE の平均値を示す. サンプル率は，1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%と変化させた．

まず，提案アルゴリズムの統計量の平均誤差は，1%から 5%程度の少ないサンプル率で全ての既存アルゴリズムの平均誤差より小さい．部分グラフと 1K グラフと比較して，部分グラフを埋め込み 1K グラフによってグラフデータを補間するアプローチの有効性が確認できる．

さらに，提案アルゴリズムの統計量の平均誤差は，サンプル率が増加するにつれて減少しており，直感的に提案アルゴリズムによる生成グラフが元のグラフに収束している．部分グラフは同様に平均誤差が減少しているが，1K グラフは，サンプル率が増加してもある一定の誤差から減少しない．

Enron と Facebook New Orleans のデータセットにおいて，提案アルゴリズムの平均誤差は，約 10%以上の大きいサンプル率で部分グラフの平均誤差より大きい．これは，提案アルゴリズムの生成過程が最適化されていないためであると考える．例えば，部分グラフ内のサンプルから見えるノードに割り当てることができる次数が存在しない場合に部分グラフ内の次数を割り当てているが，部分グラフ内の次数にパラーメタ k だけ追加する，といった方法も考えられる．提案アルゴリズムの生成過程をより最適化し，サンプル率が増加した場合に平均誤差の増加を改善することは今後の課題である．

5　関 連 研 究

隣接ノードを辿るクローリング手法は，クエリされたノードの隣接情報が得られるソーシャルネットワークにおいて，そのグラフデータをサンプリングをするために有効である [1,3,5,6]．特に幅優先探索は基本的なサンプリング手法であり，既存研究でも適用されている [1,3,5,6]．幅優先探索によるサンプルから生成された部分グラフは，データセットとして公開されており [5,8–10]，グラフ分野のアルゴリズム研究で活用されている．

Gjoka らは，ランダムウォークによるサンプルからソーシャルグラフ統計量の不偏推定量を得る実用的なフレームワークを設計した [3]．彼らは Facebook 上のクローリングの事例研究を通じて，ランダムウォークによる各サンプルにサンプリングのバイアスを修正する重み付けを施して統計量の不偏推定量を得る「再重み付けランダムウォーク法」が有効であると述べた．これまで，様々な統計量に対する再重み付けランダムウォークによる推定アルゴリズムが提案されている [15–23]．

一方で，再重み付けランダムウォーク法は，各サンプルの隣接情報から重み関数を計算する必要があるため，次数分布，クラスタ係数といった，各ノードの局所的な情報で定義される統計量しか適用できない．例えば，グラフの最短距離の分布を再重み付けランダムウォークで推定する場合，重み関数を計算するために，各サンプルから他の全ノードへの最短距離の厳密値が必要となり，膨大な追加のサンプルを必要とする．

Gjoka らは，この課題に対処するために，ランダムウォークのサンプルから元のグラフと似たグラフを生成する 2.5K グラフアルゴリズムを提案した [24]．このアルゴリズムは，ノード数 d の部分グラフ集合の次数分布を揃えた dK グラフという概念を採用している．2.5K グラフアルゴリズムは，はじめに再重み付けランダムウォーク法でジョイント次数分布とクラスタ係数の 2 つの統計量を推定する．次に，推定したジョイント次数分布から $2K$ グラフを生成する．最後に，生成されたグラフのクラスタ係数が推定値と許容誤差範囲内となるようにエッジをスワップする．生成されたグラフは，最短距離分布などの大域的な統計量も正確に模倣できる．今後は，部分グラフを埋め込んだ 2K, 2.5K グラフを生成するアルゴリズムを設計し，より正確なソーシャルグラフを復元することを目指す．

6　ま　と　め

本研究は，ランダムウォークサンプリングによるソーシャルグラフの復元に取り組み，新しい復元アルゴリズムを提案する．提案するアルゴリズムは，まずランダムウォークでサンプリングされたグラフデータで構成される部分グラフを生成し，次に，再重み付けランダムウォーク法によってサンプルから局所的な統計量を推定する．最後に，dK シリーズの枠組みに基づいて，推定した局所的な統計量が満たされるように，サンプリングされていないグラフデータを補間する．実際のソーシャルネットワークデータセットを使用した実験は，提案するアルゴリズムが，既存のアルゴリズムより誤差の小さい複数の統計量を持つソーシャルグラフを生成することを示している．

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．

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 *WWW*, pp. 835–844, 2007.

[2] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, Vol. 286, No. 5439, pp. 509–512, 1999.

[3] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *INFOCOM*, pp. 1–9, 2010.

[4] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a Social Network or a News Media? In *WWW*, pp. 591–600, 2010.

[5] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and Analysis of Online Social Networks. In *IMC*, pp. 29–42, 2007.

[6] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna PN Puttaswamy, and Ben Y Zhao. User interactions in social networks and their implications. In *EuroSys*, pp. 205–218. ACM, 2009.

[7] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, Vol. 393, No. 6684, p. 440, 1998.

[8] Jérôme Kunegis. KONECT - The Koblenz Network Collection. In *WWW*, pp. 1343–1350, 2013.

[9] Ryan A. Rossi and Nesreen K. Ahmed. The Network Data Repository with Interactive Graph Analytics and Visualization. In *AAAI*, 2015.

[10] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[11] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD*, pp. 137–146. ACM, 2003.

[12] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *KDD*, pp. 701–710. ACM, 2014.

[13] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four Degrees of Separation. In *WebSci*, pp. 33–42, 2012.

[14] Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information Network or Social Network? The Structure of the Twitter Follow Graph. In *WWW*, pp. 493–498, 2014.

[15] Xiaowei Chen, Yongkun Li, Pinghui Wang, and John Lui. A General Framework for Estimating Graphlet Statistics via Random Walk. *VLDB*, Vol. 10, No. 3, pp. 253–264, 2016.

[16] Anirban Dasgupta, Ravi Kumar, and Tamas Sarlos. On Estimating the Average Degree. In *WWW*, pp. 795–806, 2014.

[17] Stephen James Hardiman, Peter Richmond, and Stefan Hutzler. Calculating statistics of complex networks through random walks with an application to the on-line social network Bebo. *The European Physical Journal B*, Vol. 71, pp. 611–622, 2009.

[18] Stephen J Hardiman and Liran Katzir. Estimating clustering coefficients and size of social networks via random walk. In *WWW*, pp. 539–550, 2013.

[19] Liran Katzir, Edo Liberty, and Oren Somekh. Estimating sizes of social networks via biased sampling. In *WWW*, pp. 597–606, 2011.

[20] Kazuki Nakajima, Kenta Iwasaki, Toshiki Matsumura, and Kazuyuki Shudo. Estimating Top-k Betweenness Centrality Nodes in Online Social Networks. In *SocialCom*, pp. 1128–1135, 2018.

[21] Kirill Paramonov, Dmitry Shemetov, and James Sharpnack. Estimating graphlet statistics via lifting. In *KDD*, pp. 587–595, 2019.

[22] Bruno Ribeiro and Don Towsley. Estimating and Sampling Graphs with Multidimensional Random Walks. In *IMC*, pp. 390–403, 2010.

[23] Pinghui Wang, John Lui, Bruno Ribeiro, Don Towsley, Junzhou Zhao, and Xiaohong Guan. Efficiently Estimating Motif Statistics of Large Networks. *TKDD*, Vol. 9, No. 2, 2014.

[24] Minas Gjoka, Maciej Kurant, and Athina Markopoulou. 2.5 k-graphs: from sampling to generation. In *INFOCOM*, pp. 1968–1976, 2013.

[25] Priya Mahadevan, Dmitri Krioukov, Kevin Fall, and Amin Vahdat. Systematic topology analysis and generation using degree correlations. *ACM SIGCOMM Computer Communication Review*, Vol. 36, No. 4, pp. 135–146, 2006.

[26] Romualdo Pastor-Satorras, Alexei Vázquez, and Alessandro Vespignani. Dynamical and correlation properties of the internet. *Physical review letters*, Vol. 87, No. 25, p. 258701, 2001.

[27] 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 *SIGMETRICS*, pp. 319–330, 2012.

[28] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, Vol. 6, No. 2-3, pp. 161–180, 1995.

[29] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical review E*, Vol. 64, No. 2, p. 026118, 2001.

[30] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Physical review E*, Vol. 73, No. 1, p. 016102, 2006.

[31] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *KDD*, pp. 631–636, 2006.

[32] Michele Catanzaro, Marián Boguná, and Romualdo Pastor-Satorras. Generation of uncorrelated random scale-free networks. *Physical review e*, Vol. 71, No. 2, p. 027103, 2005.

[33] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, Vol. 45, No. 2, pp. 167–256, 2003.