コミュニケーションネットワークの
次数分布を保存する探索的サンプリング

数理情報学専攻 48-176225　宮崎雄貴
指導教員　田中久美子 教授

1　はじめに

大規模なグラフ構造である複雑ネットワークを探求する際、サンプリングにより小規模な部分ネットワークを得ることが不可欠である。多くの場合、サンプリングにより得られるネットワークには、元のネットワークにはない特徴が付与される。このことをバイアスがかかると言う。本研究で扱う Twitter 等のコミュニケーションネットワークは、幅優先サンプリング (BFSS) 等の単純な方法でサンプルされることが多い。この手法は高次数の頂点がサンプルされやすく、次数分布にバイアスがかかることが知られている [1]。そのため、バイアスのないサンプリング手法が望まれる。本研究では元ネットワークの次数分布をより保存するような改良手法を提案する。図 1、2 は本研究での BFSS の改良概要を示したものである。左側のグラフが BFSS によるサンプル前後の次数分布であり、ずれがあることがわかる。それを右側のように一致させることが本研究の目的である。

2　関連研究

2.1　幅優先サンプリング (BFSS)

幅優先サンプリング (BFSS) とは、ネットワーク上で幅優先探索を行い、通った頂点と枝を全てサンプルする手法である。

2.2　次数分布へのバイアス考察の枠組み

サンプリングにおける次数分布の変化を考察する際、サンプリングの過程を図 3 のように頂点選択、枝削減の 2 段階に分けて考察する方法が用いられている [2]。step 1 は、サンプルネットワーク (step 2) の頂点全てに対し、元ネットワーク (original) でその頂点から出ている枝全てを残したものである。また、本稿で使用する記号を図 3 に示す。

2.3　全枝選択

全枝選択は、サンプルネットワークの枝不足を解決するための手段である [2]。予めサンプルした頂点集合に対して、その中の 2 頂点同士を繋ぐ枝が元ネットワークにあった場合、その枝を全てサンプルすることで枝を増やす。

3　提案手法

3.0.1　BFSS の次数分布へのバイアス考察

$p_{org}(k)$ とサンプルされた頂点の元ネットワークでの次数の分布 $p_1(k)$ を比較する。つまり頂点選択での次数分布の変化を考える。BFSS では、ある頂点から見た隣接頂点が全てサンプルされるため、以下が成立する。

$$p_1(k) = p_{nei}(k) \tag{1}$$

この p_{nei} はネットワーク全体で一様と言え、それを求める。隣接頂点はネットワーク中の頂点から出ている枝をランダム

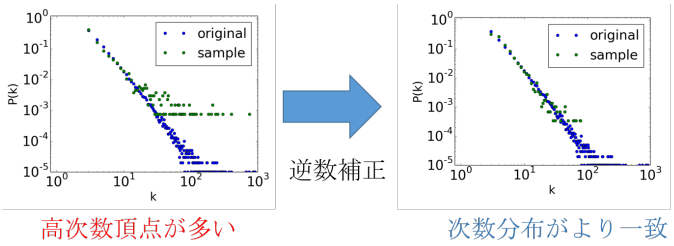

図 1. BA ネットワークへの BFSS の改良概要

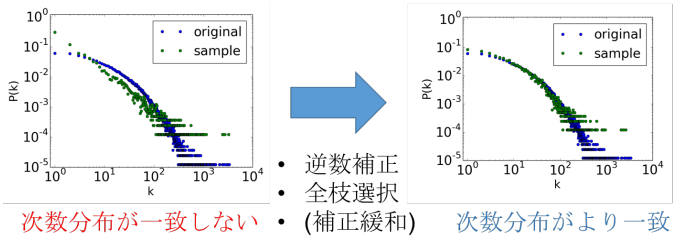

図 2. 実ネットワークへの BFSS の改良概要

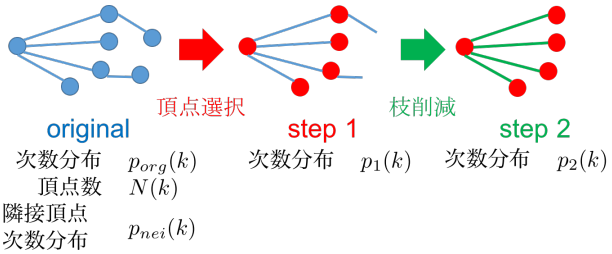

図 3. サンプリングの各段階と記法

に選ぶことで決まると考えることができ、以下が成立する。

$$p_{nei}(k) \propto k \cdot N(k) \tag{2}$$
$$\therefore p_1(k) \propto k \cdot p_{org}(k) \tag{3}$$

右辺の k により、高次頂点ほどサンプルされやすいと言える。これがバイアスを引き起こすと考えられ、k を打ち消す必要がある。本研究では次数 k の隣接頂点を k の逆数確率でサンプルする逆数補正を提案する。

3.1　逆数補正幅優先サンプリング (ID-BFSS)

逆数補正幅優先サンプリングは、次数の逆数確率で探索を行うサンプリング手法である。BFSS では隣接頂点を全てサンプルしたが、ID-BFSS では次数 k の隣接頂点を確率 $\min(c \cdot f(k), 1)$(c は定数、$f(k) = 1/k$) でサンプルする。ここで $f(k)$ を補正関数と呼ぶ。

3.2　全枝選択逆数補正幅優先サンプリング (ID-BFSS-i)

ID-BFSS は、実ネットワーク上でのサンプリング実験では枝削減において枝数が大きく減少し、結果低次頂点の割合が大きくなるという問題がある。ここでは ID-BFSS の枝削減における枝数減少を緩和するため、全枝選択を ID-BFSS に

表 1. 用いたネットワーク

ネットワーク	頂点数	平均次数	クラスター係数
BA	100000	6.00	8.93×10^{-4}
Pokec	1632803	27.32	0.12
Google+	107614	227.45	0.52
LiveJournal	4847571	17.68	0.35

表 2. 実ネットワークのサンプル後の平均次数

	BFSS	ID-BFSS	ID-BFSS-i
Pokec	7.37	2.51	14.32
Google+	20.45	2.47	75.14
LiveJournal	19.69	5.68	25.01

導入する。この方法を全枝選択逆数補正幅優先サンプリング (ID-BFSS-i) と呼ぶ。

4 実験条件

各サンプリング手法をネットワーク上で用いる実験を行なった。BFSS、ID-BFSS、ID-BFSS-i の 3 手法を比較する。サンプリングは頂点数が元ネットワークの 0.1 倍になるまで行い、$c = 3$ とした。サンプル前後の次数分布類似度を KS D-statistic で評価し、バイアスを数値化した。ネットワークは BA ネットワークと 3 種類の実ネットワークを用いた。BA ネットワークは初期頂点数 $m = 3$ で生成したものであり、このネットワークには次数 $m = 3$ 以上の頂点のみが存在するため、次数 3 以上の条件付き分布を用いた。実ネットワークは大規模なコミュニケーションネットワークである Pokec、Google+、LiveJournal[SNAP] を使用した。各ネットワークの基本情報を表 1 に示す。

5 実験結果

5.1 BA ネットワーク

図 4 のように、ID-BFSS で次数分布が最も元ネットワークに近くなった。また、step 1 時点では逆数補正を用いる ID-BFSS、ID-BFSS-i の次数分布が元ネットワークに類似しており、逆数補正が理論通りに働いたと言える。

5.2 実ネットワーク

図 5 のように、ID-BFSS では step 1 から step 2 で大幅に次数が減少し、次数分布が元ネットワークから離れた。全枝選択を用いた ID-BFSS-i では図 7 のように次数減少が抑えられ、サンプル前後で次数分布が一致した。このことは表 2 に示すサンプル前後の平均次数からもわかる。

一方で Pokec に関して、表 2 のように ID-BFSS-i を用いても元ネットワークの平均次数との間に差がある。図 6 のように次数分布そのものも元ネットワークとは乖離している。ここから Pokec での ID-BFSS-i では、全枝選択を用いても枝不足が十分に改善されていないと言える。

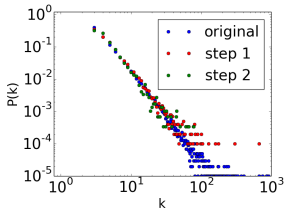

図 4. BA の ID-BFSS でのサンプル前後の次数分布

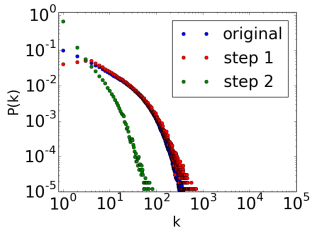

図 5. Pokec の ID-BFSS でのサンプル前後の次数分布

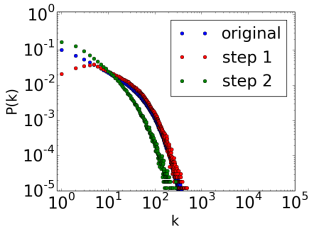

図 6. LiveJournal の ID-BFSS-i でのサンプル前後の次数分布

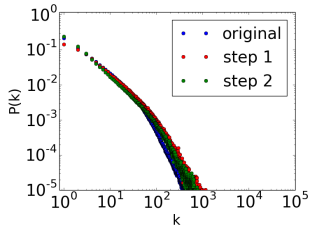

図 7. LiveJournal の ID-BFSS-i でのサンプル前後の次数分布

以上から、実ネットワークでは逆数補正、全枝選択が両方機能すれば ID-BFSS-i が最も優れた手法であるが、ネットワークの種類によっては全枝選択が十分に機能しない可能性があると言える。

6 考察

実ネットワークの多くに対し、ID-BFSS-i が最も優れた手法と言える。一方で ID-BFSS-i では全枝選択部分に関して、Pokec で枝不足が十分に改善されなかった。こうしたネットワークで最終的な枝不足を緩和する方法として、step 1 で高次数頂点が増えるように補正関数 $f(k)$ を設定し、補正を緩和する方法が挙げられる。補正緩和が必要なネットワークの選別、適切な補正関数の設定方法等が今後の課題と言える。

7 結論と今後の課題

本研究では BFSS の次数分布へのバイアスという問題点に着目し、BA ネットワークに有効な ID-BFSS および実ネットワークに有効な ID-BFSS-i という手法を提案した。ID-BFSS-i により枝不足が生じる実ネットワークもあり、それへの対応が今後の課題と言える。

参考文献

[1] M. Kurant, A. Markopoulou, and P. Thiran. Towards unbiased BFS sampling. *IEEE Journal on Selected Areas in Communications*, Vol. 29, pp. 1799–1809, 2011.

[2] N. K. Ahmed, J. Neville, and R. Kompella. Network Sampling: From Static to Streaming Graphs. *ACM Transactions on Knowledge Discovery from Data*, Vol. 8(2), No. 7, pp. 1–56, 2014.