3C-03

Bottom-k Sketch による可到達ノード数推定に基づくネットワーク構造分析

山岸　祐己 †　　斉藤　和巳 †
† 静岡県立大学院 経営情報イノベーション研究科

1　はじめに

ソーシャルメディアの普及などにより大規模なソーシャルネットワークが構築されるようになり，その特徴的な構造特性を見出すため，中心性分析 [1] など多様な研究が展開されている．本稿では，情報拡散トレースなどに代表される有向非巡回 (DAG: Directed Acyclic Graph) 性の高い構造を持つネットワークを対象に，1) 異なる特性を有する人工ネットワーク構成法，2) Bottom-k Sketch による可到達ノード数推定法 [2]，及び，3) 到達ノード数分布のプロットにより，ネットワーク構造を分析する方法を提案する．大規模な実ネットワークを用いた評価実験では，構造特性が到達ノード数分布に明確に反映されることを確認し提案法の有望性を示す．

2　ネットワーク構造分析法

与えられたネットワークを $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ とする．ここで，$\mathcal{V} = \{u, v, w, \cdots\}$ と $\mathcal{E} = \{(u, v), \cdots\}$ は，それぞれノード集合とリンク集合を表す．いま，ノード $u \in \mathcal{V}$ からリンク集合 $\mathcal{E}$ を順方向に辿って到達可能なノード集合を $\mathcal{R}(u)$ とする．本稿で提案するネットワーク構造分析法は次の 3 ステップで構成される．

1. 与えられたネットワークと同程度規模の BA DAG と CNN DAG を DCNN 法と DBA 法で構築する．
2. これらネットワークの各ノード u の可到達ノード数 $|\mathcal{R}(u)|$ を Bottom-k Sketch で求める．
3. それぞれのネットワークでノードを $|\mathcal{R}(u)|$ で昇順にソートし，そのランクと $|\mathcal{R}(u)|$ のペアをプロットした可到達ノード数分布を出力する．

以下では，DCNN 法と DBA 法による DAG 性の高い構造を持つネットワーク構成法，及び，Bottom-k Sketch による可到達ノード数推定法について説明する．

まず，CNN モデル [3] を土台にする DCNN 法について説明する．いま，ノードのペア $\{v, w\}$ 間に直接リンクはないが，少なくとも一つの共通な隣接ノードを持つとき，これらノードは潜在ペア（potential pair）であると言う．DCNN 法は，単一孤立ノード集合 $\{v\}$ を設定し，以下の処理を L 回繰り返すアルゴリズムとなる: 1) 確率 $(1-\epsilon)$ で新ノード v を生成し，任意の既存ノード $v \in \mathcal{V}$ を任意に選択し，新たなリンク (u, v) または (v, u) を生成する; 2) 確率 ϵ で任意の潜在ペア $\{v, w\}$ を選択し，ネットワークが非巡回グラフとなるように新たなリンク (u, v) または (v, u) を随意に生成する．ここで，ϵ はノード数とリンク数の比をコントロールするパラメータであり，上記ステップ 2 で非巡回グラフとならないリンク生成も許容すれば，ネットワークの非巡回性も調整できる．

次に，BA モデル [4] を土台にする DBA 法について説明する．いま，ノードがその隣接ノード数に比例する確率で選択されるとき，優先性アタッチメント（preferential attachment）による選択と言う．DBA 法は，リンク数 H の任意の初期グラフ（例えば，DCNN 法で構築）を設定し，以下の処理を $L - H$ 回繰り返すアルゴリズムとなる: 新ノード v を生成し，異なる J 個のノードを $\{v_1, \cdots, v_J\} \subset \mathcal{V}$ を優先性アタッチメントで選択し，リンク集合 $\{(u, v_1), \cdots, (u, v_J)\}$ または $\{(v_1, u), \cdots, (v_J, u)\}$ を随意に生成する．ここで，J はノード数とリンク数の比をコントロールするパラメータであり，上記リンク生成で一部のリンク方向を反転させれば，ネットワークの非巡回性も調整できる．なお，DCNN 法でネットワーク生成すれば，DBA 法での生成と比較して，潜在ペア間にリンクが生成されるケースが多くなるので，フィードフォアワード型トライアド（feedforward triads）パターンが多くなるのが特徴と言える．以下では，DCNN 法で生成されるネットワークを CNN DAG と呼び，DBA 法でのネットワークを BAN DAG と呼ぶ．

最後に，Bottom-k Sketch [2] による可到達ノード数推定法について述べる．まず，与えられたネットワークの任意のノード $v \in \mathcal{V}$ に対し，$(0, 1)$ の一様乱数で値 r_v を付与する．ノード部分集合 $\mathcal{U} \subset \mathcal{V}$ の Bottom-k Sketch とは，付与した乱数値の小さい順に k 個の要素を選び構成したボトム部分集合 $\mathcal{B}_k(\mathcal{U}) \subset \{r_u \mid u \in \mathcal{U}\}$ のことである．いま，ボトム集合 $\mathcal{B}_k(\mathcal{U})$ の最大値を $\max \mathcal{B}_k(\mathcal{U})$ で表せば，$(k-1)/\max \mathcal{B}_k(\mathcal{U})$ により，元の集合のノード数 $\mathcal{U}$ を高精度で不偏推定できる．一方，任意のノード $u \in \mathcal{V}$ の可到達ノード数 $|\mathcal{R}(u)|$ を求めるには，各ボトム集合を $\mathcal{B}_k(\mathcal{R}(u)) = \emptyset$ を初期化して，r_v が最小のノー

Network Structure Analysis Based on Reachable Nodes Estimation by Bottom-k Sketch
†Yuki YAMAGISHI　†Kazumi SAITO
†University of Shizuoka

ド v から順に，リンクの逆向き方向で到達可能なノード u を求め，$\mathcal{B}_k(\mathcal{R}(u)) \leftarrow \mathcal{B}_k(\mathcal{R}(u)) \cup \{r_v\}$ の処理を繰り返す．任意の $|\mathcal{R}(u)|$ を求める計算量は $O(k \times |\mathcal{E}|)$ である．

3 実験による評価

ベンチマークとして，ここでは Stanford Large Network Dataset Collection * から得た実ネットワークの cit-Patents ($|\mathcal{V}| = 3,774,768$, $|\mathcal{E}| = 16,518,948$) と wiki-Talk ($|\mathcal{V}| = 2,394,385$, $|\mathcal{E}| = 5,021,410$) を用いる．今回，構築されるネットワークがおおよそ $|\mathcal{V}| = 1,000,000$, $|\mathcal{E}| = 10,000,000$ となるよう，$L = 10,000,000$, $\epsilon = 0.1$ と設定した．図 1 は，DCNN 法と DBA 法により構築した CNN DAG と BA DAG における可到達ノード数とランクの分布を示したものである．図より，CNN DAG に比べ，BA DAG には可到達ノード数が極めて多いノードが多数存在していることが見て取れる．つまり，同程度の BA DAG タイプのネットワークでは，多くのノードが爆発的な情報拡散を起こし得るということが自然と推察される．更に，可到達ノード数推定法は，与えられたネットワークの構造特性を特徴付ける指標として有用であることも示唆される．図 2 は，cit-Patents と wiki-Talk における可到達ノード数とランクの分布を示したものである．図より，cit-Patents に比べ，wiki-Talk には可到達ノード数が極めて多いノードが多数存在していることが先程と同様に見て取れる．即ち，wiki-Talk と cit-Patents は，それぞれ BA DAG タイプと CNN DAG タイプに分類されると考えることができる．

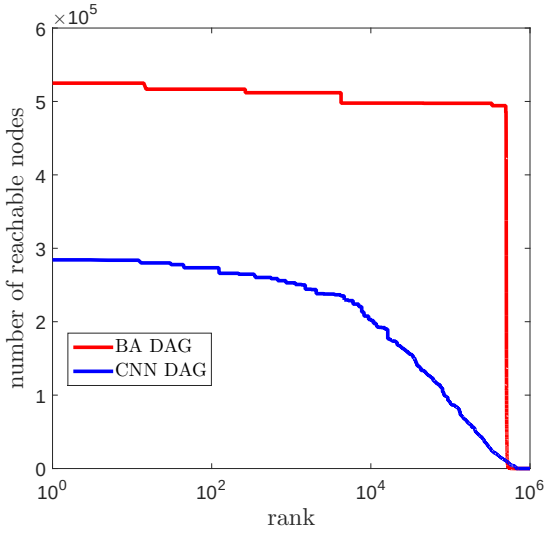

図 1: 人工データでの比較

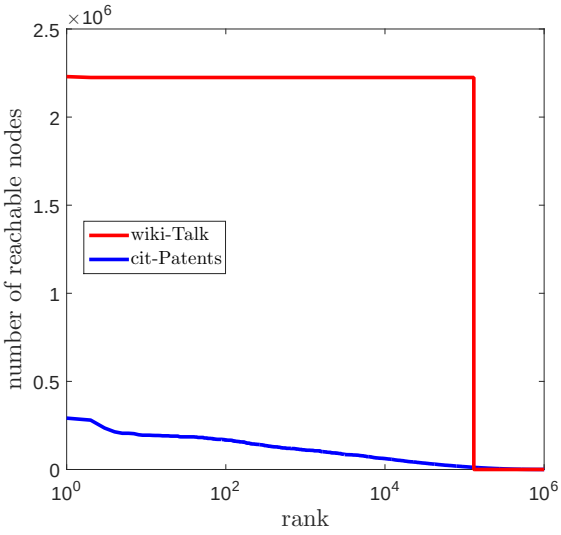

図 2: 現実データでの比較

4 おわりに

本稿では，情報拡散トレースなどに代表される DAG 性の高い構造を持つネットワークを対象に，1) 異なる特性を有する人工ネットワーク構成法，2) Bottom-k Sketch による可到達ノード数推定法，及び，3) 到達ノード数分布のプロットにより，ネットワーク構造を分析する方法を提案した．大規模な実ネットワークを用いた評価実験では，構造特性が到達ノード数分布に明確に反映されることを確認することで提案法の有望性を示した．今後は，多様なネットワークへの適用を通して提案法の有用性検証する．

謝辞 本研究は，総務省 SCOPE(No.142306004) 及び，科研費 (No.25330635) の補助を受けた．

参考文献

[1] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[2] E. Cohen. All-distances sketches, revisited: Hip estimators for massive graphs analysis. In *Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 88–99, 2014.

[3] A. Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review*, Vol. 67, No. 5, p. 056104, 2003.

[4] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, Vol. 286, pp. 509–512, 1999.

*https://snap.stanford.edu/data/