学術論文の引用ネットワークに対する生成モデル

安井 雄一郎　総合研究大学院大学 複合科学研究科 統計科学専攻 D5

1　はじめに

学術論文の引用関係のより深い理解を目的として, 引用ネットワークの特徴的な構造を表現する生成モデルを構築する. 本研究では主にクラリベイト・アナリティクス社のオンライン研究文献データベースである Web of Science (WoS) に格納されている, 1981 年から 2016 年まで (35 年分) の確率・統計カテゴリに該当する 179,483 の文献と 1,106,622 の引用関係を対象とする. 引用関係から構成されるネットワーク構造は, 文献を点に, 文献間の引用を枝に対応させた有向グラフ $G = (V, E)$ で表現され, 各点 $v \in V$ には非負整数値となる公開時刻 $\tau : v \to Z_+$ をもつものとする.

図 1は各時刻ごとの文献 v_i が引用している文献 v_j との相対時刻 $\tau(v_i) - \tau(v_j)$ の頻度分布を示しており, 各文献はより古い文献を引用していること, 引用は文献間の相対時刻で変化し 3–5年後にピークとなることを確認できる.

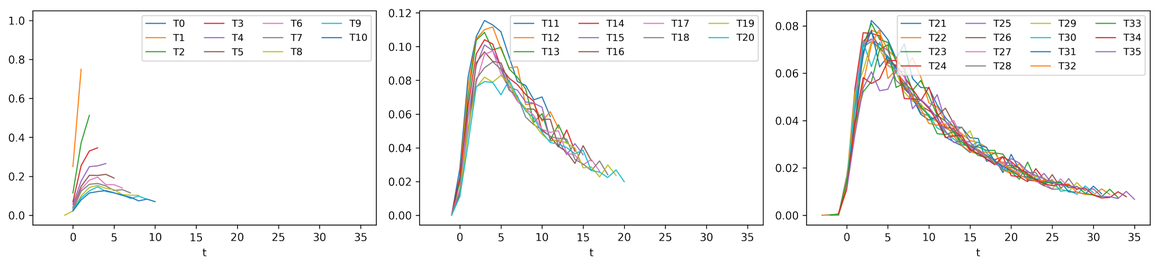

図 1: 引用割合の経年変化 (正規化した時刻ごと. 1981年は T0 に対応.)

2　ネットワーク生成モデル

本研究の生成モデルは Barabási and Albert [1] (以後, Barabási–Albert モデルとする) により提案された Preferential attachment (PA 処理), Holme and Kim [2] (以後, Holme–Kim モデルとする) により提案された Triad formation (TF 処理), Wu and Holme [3] (以後, Wu–Holme モデルとする) により考慮された引用割合の経年変化を採用している. Wu–Holme モデルでは時刻の考慮は点 ID を用いて行われており, 時刻順にソートされた 0, 1, .., $n-1$ (n は点数) となるような点IDを用意することが必要となる. しかしながら WoS の文献データでは, 長期間の文献データを扱う際に用いることができる時刻の粒度は発表年となるため, 必要な点ID体系を正確に用意することはできない.

そこで本研究では PA 処理と TF 処理をベースとしつつ, (点ID よりもしばしば粒度の粗い) 時刻を用いた生成モデルを構成した. 前述のように WoS で得られる時刻粒度は年となるため, 文献数や引用関係数に比べて, 粗い 35 期間となる時刻を用いることになる.

本研究のモデルでは, 以下の手順で時刻 t ごとに処理を行うことで, ネットワーク構造を成長させていく. 開始時点のネットワーク構造は空とし, 時刻 $t = 0$ から順に処理していく. **(a)** 時刻 t の点集合 $V^{(t)}$ を生成する. このとき各点 $v \in V^{(t)}$ には時刻 $\tau(v) = t$ と,推定された Out-degree $k'_{\mathrm{out}}(v)$ が割り当てられる. **(b)** 生成された各点 $v_i \in V^{(t)}$ に対し PA 処理もしくは TF 処理により枝を生成する. このとき PA 処理と TF 処理はパラメータ $\beta \in [0, 1]$ を用いて確率的に選択される. PA 処理を用いた枝生成は, まず経過時刻 $\tau(v_i) - \tau(v_j)$ に基づいた引用割合を用いて点 v_j の時刻 $\tau(v_j)$ を決定する. このとき $\tau(v_j)$ が範囲外であればその枝生成は行わない. 続いて, 時刻 t での次数 $k'^{(t)}_{\mathrm{in}}(v)$ を用いて, 1点 v_j を選択し, 枝 (v_i, v_j) を生成する. このとき, 点 v の選択確率は $k'^{(t)}_{\mathrm{in}}(v) / \sum_{w \in V^{(0)} \cup \ldots \cup V^{(t)}} k'^{(t)}_{\mathrm{in}}(w)$ である. TF 処理を用いた枝生成は, 生成した点 v_i と直前のPA処理で選択した点 v_j の隣接点集合 $\left\{v_k \mid v_k \neq v_i, (v_i, v_j) \in E, (v_j, v_k) \in E\right\}$ の中から, 経過時刻に基づく引用割合とその時刻における In-degree $k'^{(t)}_{\mathrm{in}}(v)$ を用いて1点 v_k を選択し, 枝 (v_i, v_k) を生成する. β は先行研究と同じ 0.99 を用いた.

2.1　パラメータの推定

本研究の生成モデルで用いる入力パラメータを推定する. **(1)** 各点の生成時に割り振られる Out-degree は点 $v_i \in V$ の Out-degree $k_{\mathrm{out}}(v_i) = \left|\left\{v_j \mid (v_i, v_j) \in E\right\}\right|$ をGeneralized Pareto distribution を用いて推定した. **(2)** 各枝の生成時に経過時刻に基づく引用割合は, 各枝 $(v_i, v_j) \in E$ の引用の経過年 $\tau(v_i) - \tau(v_j)$ の頻度分布を Johnson's S_B distribution を用いて推定した. **(3)** 各時刻で生成するべき点数は, 各時刻 t の点集合 $V^{(t)}$ の要素数 $n^{(t)}$ を Generalized logistic function を用いて推定した.

図 2はパラメータごとの元データとシミュレーション結果を比較したものである. いずれもあてはまりに問題がないことが確認できる.

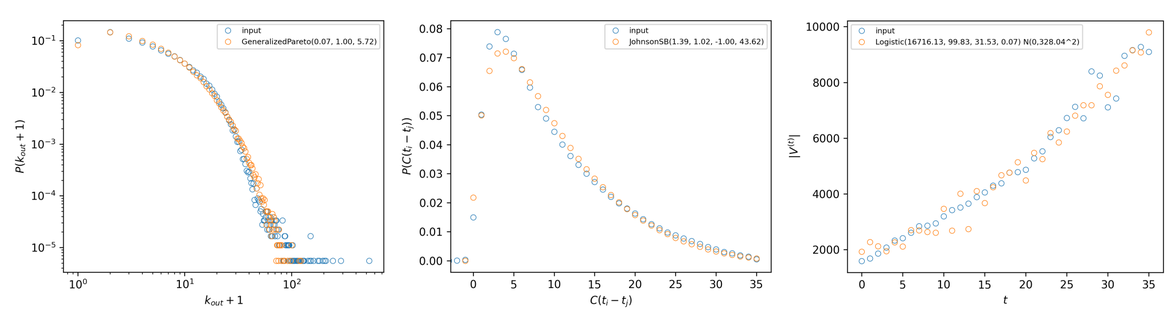

図 2: パラメータごとの元データとシミュレーション結果

3　生成モデルごとのネットワーク特徴量の比較

本研究と先行研究の生成モデルを比較する. 図 3にモデルごとのネットワーク指標をまとめる. 用いた In-degree, Out-degree, Triangle participation はそれぞれ被引用数の分布, 引用数の分布, 各点が参加する三角形数の分布 (無向グラフとして算出) を表している. まず Barabási–Albert モデルは PA 処理により, In-degree を捉えられている. 他のアルゴリズムもこの PA 処理を踏襲しており In-degree へのあてはまりがよい. 一方で Barabási–Albert モデルと Holme–Kim モデルは点を追加した際に接続する枝数 (Out-degree) を定数 (入力グラフの Out-degree の平均) としているため, Out-degree や Triangle participation の性質が十分に表現できていない. また Wu-Holme model は対象グラフの Out-degree そのものを入力とし, 時間経過による引用率の変化を考慮しているため, Out-degree と Triangle participation へのあてはまりが改善している.

我々のモデルは点 ID の順序よりも粒度の粗い時刻をもとした生成モデルとなり, 用いたデータ WoS の引用ネットワークのように, 長期間の文献データをまとめて扱う際に用いることができる時刻の粒度が発表年 (35期間) 程度であっても問題なく適用できる. 図 3より In-degree, Out-degree, Triangle participation で最もあてはまりがよいことを確認できる.

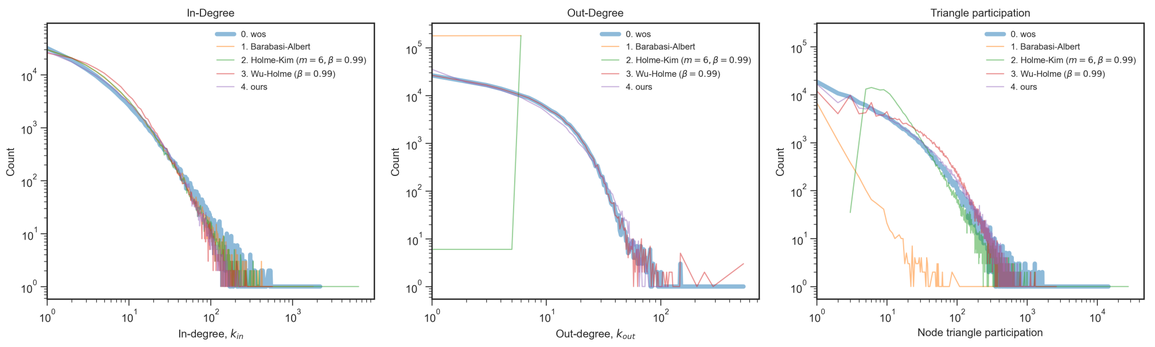

図 3: 生成モデルごとのネットワーク特徴量の比較

参考文献

[1] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[2] Petter Holme and Beom Jun Kim. Growing scale-free networks with tunable clustering. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 65(2):2–5, 2002.

[3] Zhi Xi Wu and Petter Holme. Modeling scientific-citation patterns and other triangle-rich acyclic networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 80(3), 2009.