学術論文の引用ネットワークに対する時刻差を考慮した生成モデル

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　概要

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. 本研究では主に論文アップロードサイトである arXiv における11 年分 (1992–2002 年) の高エネルギー物理理論に関する論文 26,792 件との引用関係 333,973 件を用いた. 本研究では文献の発表時刻を四半期の粒度 (**33期分**) で用いた.

学術論文の引用関係の大部分は新しい文献が古い文献を引用する. そのためある時間範囲に含まれる文献データから引用ネットワークを構築すると, 古い文献は引用する文献がデータの範囲外となりやすい一方で, 十分な時間範囲を確保すれば, 新しい文献では引用する文献がデータに含まれている可能性が高い. これらの性質を考慮するため, 我々は時刻差を考慮した引用ネットワークの生成モデルの構築を行った.

2　時刻差を考慮した引用ネットワークの特徴

引用関係から生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. さらに時刻差 w を導入し, 一定の時刻差 w をもつ部分グラフ構造を (V_w, E_w) を次のように定義する.

$$V_w = \{v \mid v \in V, \tau(v) \geq w\}, \tag{1}$$
$$E_w = \left\{(v_i, v_j) \mid v_i \in V_w, (v_i, v_j) \in E, 0 \leq \tau(v_i) - \tau(v_j) \leq w\right\}. \tag{2}$$

図 1, 2 は時刻差 $w = 25$ を設定した引用ネットワーク (V_w, E_w) に対し, Out-degree distribution と時刻差ごとの引用率を示したものである. いずれも引用関係 (v_i, v_j) における引用元 v_i の時刻 $\tau(v_i)$ ごとのプロットであるが, 時刻により変化しにくい構造をもつことが確認できる.

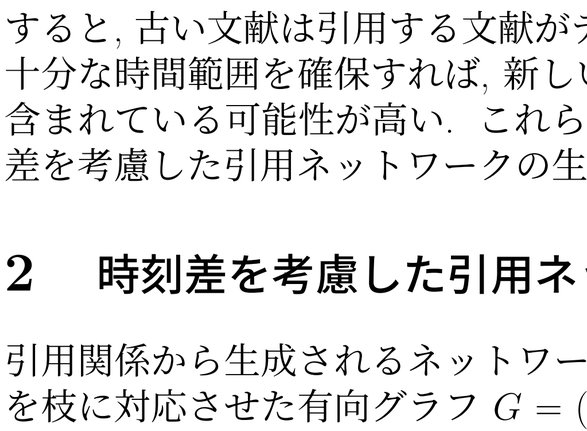

図 1: 引用時刻ごとの出次数分布

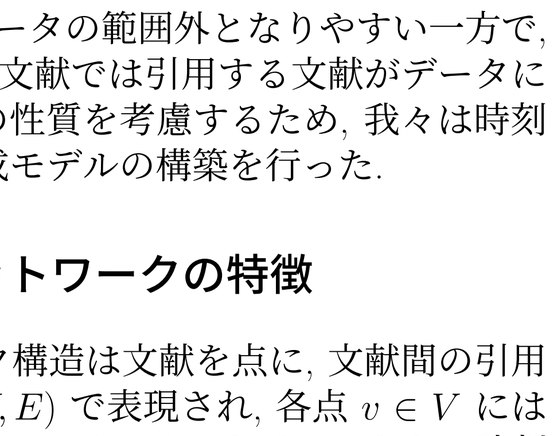

図 2: 時刻ごとの時刻差と引用率

3　提案のネットワーク生成モデル

本研究のモデルではまず初期化した文献集合 V, 引用関係集合 E を用意する. 各時刻 t では $f_n(t)$ 件の文献集合 U を生成し V に追加する. 生成された各文献 $v_i \in U$ には $f_o(k)$ を従い引用文献の件数 k が割り当てられ, PA (Preferential attachment) 処理 [1] か TF (Triad formation) 処理 [2] により引用先文献 v を決定し, 引用関係 (v_i, v) を E に追加する.

PA 処理では引用文献 v_j をすでに存在する文献の中から, 重要度 $Imp(v_j)$ と時刻差における引用率 $f_c(\tau(v_i) - \tau(v_j))$ を考慮した確率 $P_{PA}(v_j)$ で選択する. 一方, TF 処理は引用文献 v_k を, 直前のPA処理で選択した文献 v_j の周辺の文献集合 $\{v_k \mid v_k \in A(v_j), v_i \neq v_k, (v_i, v_k) \notin E\}$ から, 確率 $P_{TF}(v_k)$ で選択する.

$$P_{PA}(v_j) \sim f_c(\tau(v_i) - \tau(v_j)) \cdot Imp(v_j), \tag{3}$$
$$P_{TF}(v_k) \sim f_c(\tau(v_i) - \tau(v_k)) \cdot Imp(v_k). \tag{4}$$

TF処理により引用関係の三角形 (v_i, v_j, v_k) が形成される. ここで $A(v)$ は v の隣接ノード集合を示す. また, 重要度 $Imp(v)$ は In-degree $k_{\text{in}}(v)+1$ で近似した. TF 処理が選択される確率はパラメータ $\beta \in [0, 1]$ で与える.

各 f_n, f_c, f_o の推定方法について説明する. 各時刻 t の文献数 $f_n(t)$ は Generalized logistic function を, 引用関係における時刻差 t により変化する引用率 $f_c(t)$ は Inverse Gaussian distribution を, 各文献が取り得る引用文献数 k となる確率 $f_o(k)$ は Generalized Pareto distribution を, それぞれ用いて推定した. f_c と $f_o(k)$ については設定した時刻幅 $w = 25$ で設定した部分グラフ (V_w, E_w) を用いて推定を行った. パラメータ β はシミュレーションにより 0.99 に決定し以後の評価に用いた.

4　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. In-degree (図 3), Out-degree (図 4), Node triangle participation (図 5) 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（引用関係間の密さ）の分布を表している. また Scree plot (図 6) は隣接行列の上位固有値をプロットであり, ネットワーク構造の理解に広く用いられている指標である.

まず In-degree については各モデルがもつ PA 処理により特徴を捉えられている. 続いて Out-degree についてはモデルの特徴が現れており, Barabási–Albert モデル [1] や Holme–Kim モデル [2] は各文献の引用文献数が定数と仮定しているため特徴が捉えられていない. Wu–Holme モデル [3] は Out-degree そのものを用いるため実データと一致している. 一方, 我々のモデルは時刻差を考慮したモデル化 (f_o) により特徴を捉えることに成功している. また我々のモデルはWu–Holme モデルと比較して, Node triangle participation は同程度, Scree plot では同程度以上のあてはまりを確認できた.

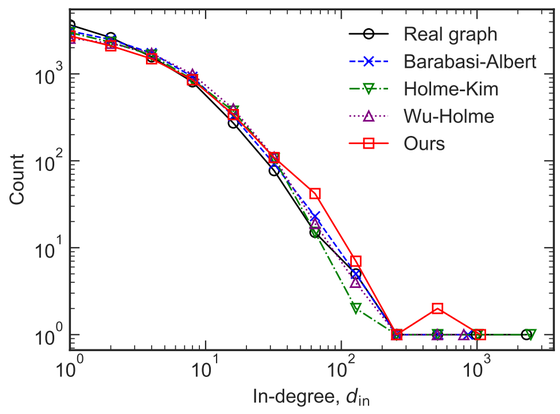

図 3: In-degree distribution

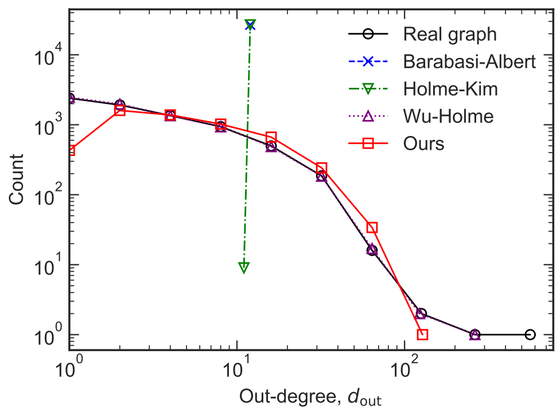

図 4: Out-degree distribution

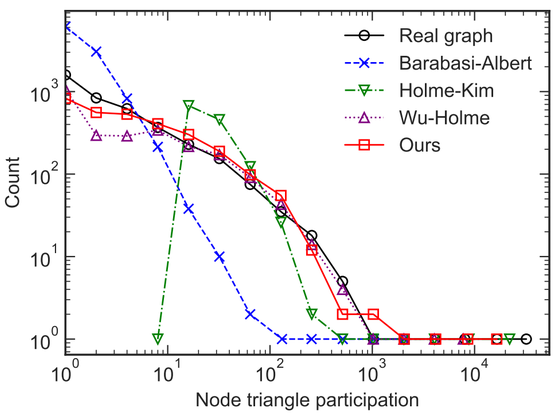

図 5: Node triangle participation

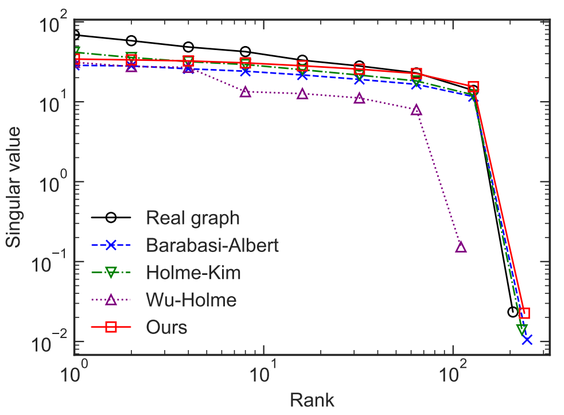

図 6: Scree plot

5　まとめ

提案モデルは先行研究 [3] と比較し, 次のような特徴をもつ. (a) 粒度の粗い（年や四半期など数十期間ほどの）時刻があれば適用可能である（先行研究では発表時間でソートされた文献IDが必要となる）(b) いくつかのネットワーク特徴量での評価で同等以上のあてはまりを示した. (c) 時刻を明示的に扱うことで生成結果の解釈性が向上した.

参考文献

[1] Albert-László Barabási and Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, oct 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.