特許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

概要

特許の引用ネットワークに対する確率生成モデルについて報告する. 我々が構築した学術論文の引用関係に対する確率生成モデルを, 米国特許の引用ネットワークに適用しモデルの汎用性を検証した.

1　学術文献の引用ネットワークに対する確率生成モデル

引用ネットワークは文献を点に, 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 $1, 2, ..., T$ と正規化された公開時刻 $\tau(v)$ をもつものとする.

我々の確率生成モデル (Yasui and Nakano 2022) は, 引用ネットワーク上の (a) 時刻ごとの文献数 $n(t)$, (b) 引用の時刻差 $s = \tau(v_i) - \tau(v_j)$ ごとの引用率 $c(s)$, (c) T 期遡ったときの出次数分布, に着目しており, (a) はロジスティクス関数 $f_n(t)$, (b) は逆ガウス分布の確率密度関数の定数倍 $f_c(s)$, (c) は一般化パレート分布（もしくは指数分布）f_o, にあてはまることを確認している.

ネットワーク構造の生成には, まず空集合で初期化した文献点集合 V', 引用枝集合 E' を用意する. そして各時刻 t で $f_n(t)$ 件の文献点集合 U を生成し V' に追加する. 各文献 $v_i \in U$ は $f_o(k)$ を従い割り当てられた参考文献数 k の回数だけ, PA (Preferential attachment) 処理 (Barabási and Albert 1999) かTF (Triad formation) 処理 (Holme and Kim 2002) を実施し, T 期以内の引用先文献 v を決定し引用枝 (v_i, v) を E' に追加する. PA 処理では引用先文献 v_j をすでに存在する文献点集合の中から, 重要度と, 時刻差における引用率 $f_c(\tau(v_i) - \tau(v_j))$ を考慮し確率的に選択する. 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 の隣接文献点集合から PA と同様に確率的に選択する. 重要度は入次数で近似した. V' と E' から対象の時刻の範囲外を除外して V と E として出力する.

図 1は Web of Science の統計・確率分野の引用データを用いて生成した学術論文の引用ネットワークにおける推定結果であり, 黒丸がデータを, 赤線が推定結果を表している.

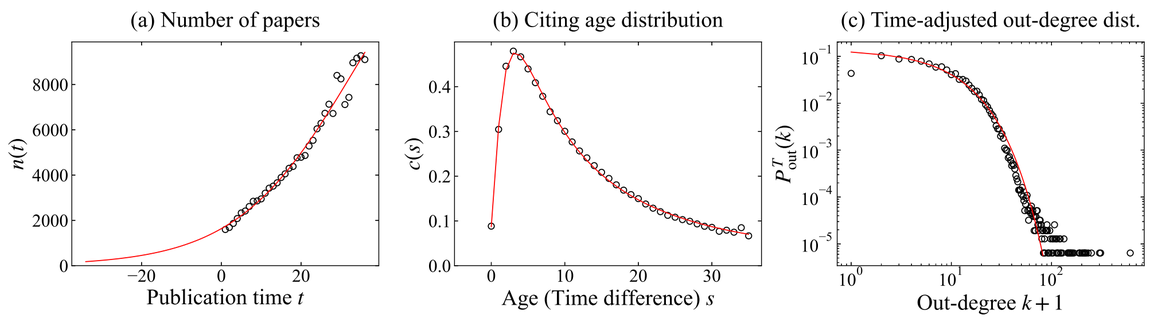


図 1: 学術論文の引用ネットワークにおける推定結果

2　特許の引用ネットワーク

NBER (National Bureau of Economic Research) で公開されている米国特許の引用情報 (`https://www.nber.org/research/data/us-patents`) を用いる. 公開時刻などのメタ情報を持たない文献を対象外とし, 1963年から1999年までの36年分の特許番号, 公開時刻, 引用情報から 2,745,762 ノード (＝特許数), 13,965,411 エッジ (＝引用数) の引用ネットワークを構築した. 各特許は 6カテゴリ (Chemical, Computers & Communications, Drugs & Medical, Electrical & Electronic, Mechanical, Others), 36サブカテゴリのいずれかに割り当てられている.

図 2, ネットワーク全体で得られた推定結果である. (b) 引用時刻の分布は問題がない一方で, (c) で出次数分布であてはまりが十分でない.

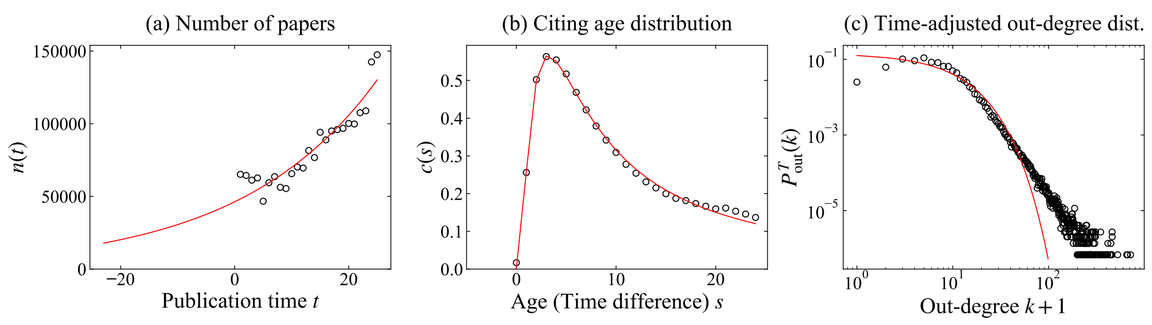
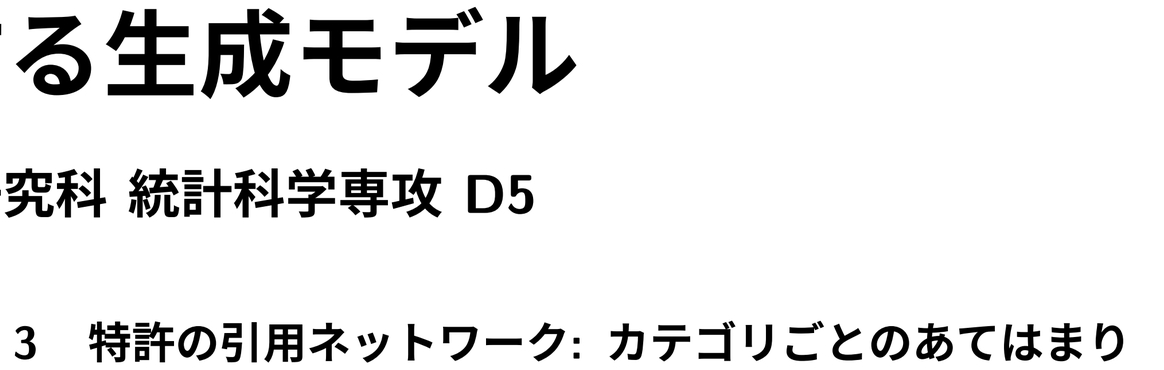


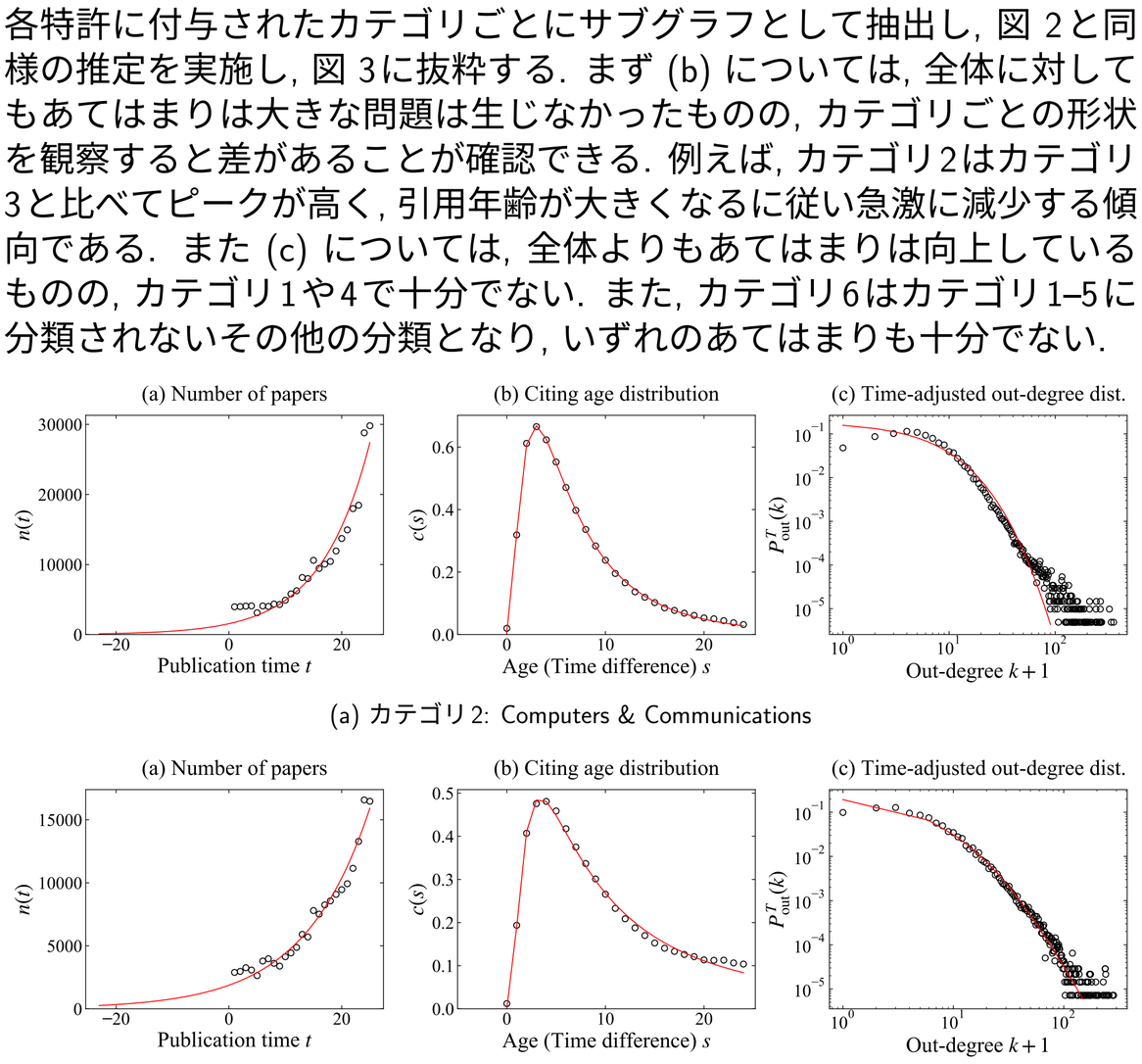
図 2: 米国特許の引用ネットワークにおける推定結果

3　特許の引用ネットワーク: カテゴリごとのあてはまり

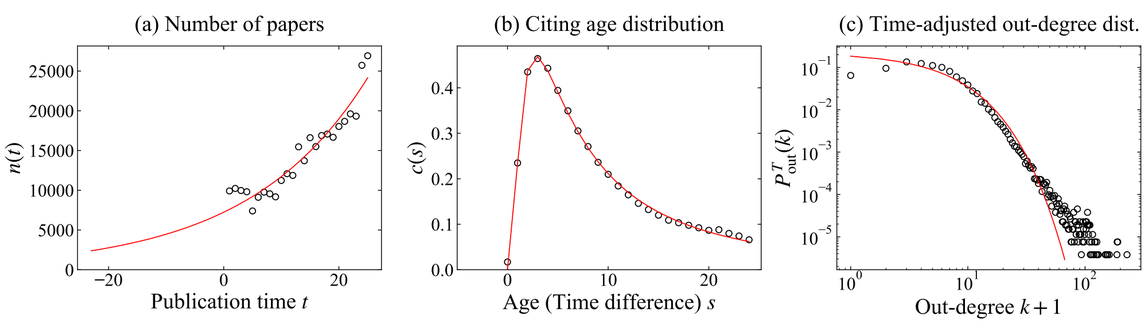
各特許に付与されたカテゴリごとにサブグラフとして抽出し, 図 2と同様の推定を実施し, 図 3に抜粋する. まず (b) については, 全体に対してもあてはまりは大きな問題は生じなかったものの, カテゴリごとの形状を観察すると差があることが確認できる. 例えば, カテゴリ2はカテゴリ3と比べてピークが高く, 引用年齢が大きくなるに従い急激に減少する傾向である. また (c) については, 全体よりもあてはまりは向上しているものの, カテゴリ1や4で十分でない. また, カテゴリ6はカテゴリ1–5に分類されないその他の分類となり, いずれのあてはまりも十分でない.



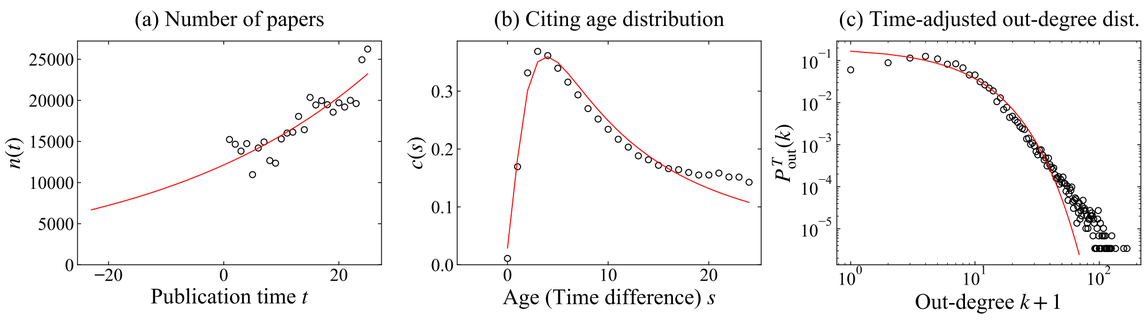
(a) カテゴリ 2: Computers & Communications



(b) カテゴリ 3: Drugs & Medical



(c) カテゴリ 4: Electrical & Electronic



(d) カテゴリ 6: Others

図 3: カテゴリごとの推定結果

4　まとめ

米国特許の引用ネットワークに対して, 我々の確率生成モデルで用いる3つの性質に対するあてはまりを検証した. 対象の引用ネットワークは全体としては6のカテゴリが混在しており, あてはまりが十分でない. その一方で, いくつかのカテゴリに対しては, 学術文献の引用ネットワークと同様のあてはまりを確認することができた.

参考文献

Barabási, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks, *Science*, **286** (5439), 509–512.

Holme, P. and Kim, B. J. (2002). Growing scale-free networks with tunable clustering, *Physical Review E*, **65** (2), 2–5.

Yasui, Y. and Nakano, J. (2022). A stochastic generative model for citation networks among academic papers, *PLOS ONE*, (accepted).