引用ネットワークの特徴に関する考察

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 博士課程(5年一貫制)5年

1　Web of science 引用ネットワーク

本研究の目的は Thomson Reuters 社のオンライン研究文献データベースである Web of science (WoS) の 2005 年から 2014 年までの統計・確率カテゴリの 65,750 の文献と 189,321 の引用関係から構成される**ネットワーク構造の理解**である. ネットワークは (1) 各文献はより古い文献を引用している, (2) 引用は3–5年後に集中している (図 1), (3) 時刻 t における点数 $N(t)$ と枝数 $E(t)$ に $E(t) \propto N(t)^a$ (a は密性を示す定数)が成り立つ (Densification Power Law [1], 図 2), などといった特徴を有する.

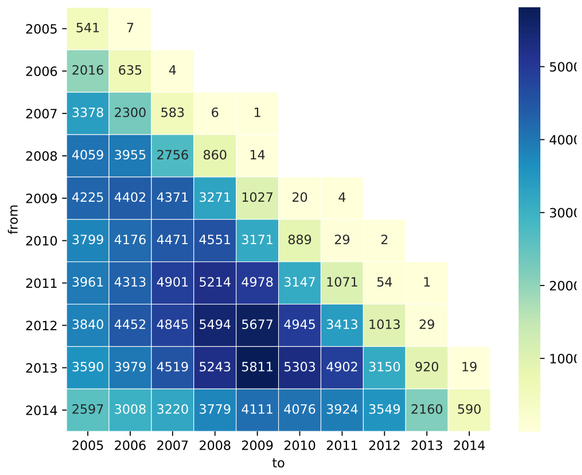

図 1: 年ごとの引用数

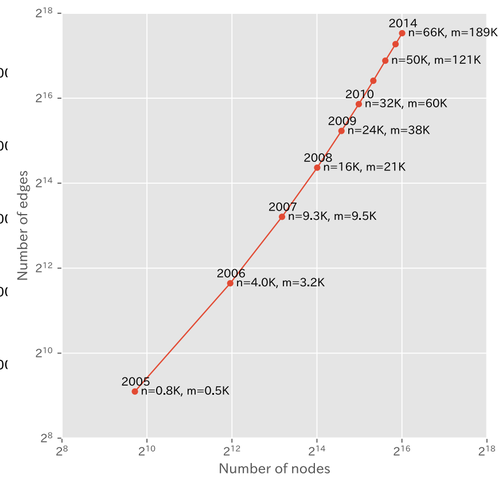

図 2: ネットワークの成長

2　確率的クロネッカーグラフモデルへのあてはめ

まずは Kronecker product を用いたランダムグラフモデル Stochastic Kronecker graph (SKG) [1] を検討した. SKG [1] は少ないパラメータ (巨大なネットワークでも概ね 4 つのパラメータで表現可能) 先行研究の数値実験で様々な複数の実ネットワークへのあてはまりの良さを主張している. SKG モデルの利用は Leskovec らが開発している SNAP (Stanford Network Analysis Platform) (`https://snap.stanford.edu/snap/index.html`) の `kronfit` と `krongen` を用いる.

SKG では n 次正方行列で各要素が $[0,1]$ の実数を取る初期パラメータ行列 $\Theta^{[0]}$ を入力とし, $\Theta^{[0]}$ 自身への k 回の Kronecker product を適用し得られた行列 $\Theta^{[k]}$ の各 (i,j)–成分を, 枝 (i,j) の生成確率に対応させる.

- パラメータ推定 `kronfit` は最適化問題 $\arg\max_\Theta P(G \mid \Theta^{[k]})$ からなる最尤推定により実現している. 枝の生成確率を独立したベルヌーイ確率変数としてモデル化し, 尤度関数を以下のように定義される. ここで σ は点の順列を表し, i 番目の要素は点番号 σ_i に対応する.

$$P(G \mid \Theta^{[k]}, \sigma) = \prod_{(u,v)\in G} \Theta^{[k]}[\sigma_u, \sigma_v] \prod_{(u,v)\notin G} (1 - \Theta^{[k]}[\sigma_u, \sigma_v])$$

- ネットワーク生成 `krongen` は得られたパラメータ Θ を元に必要な回数分 Kronecker product を乗じた $\Theta^{[k]}$ の各 (i,j)-成分を枝 (i,j) の生成確率に対応させ, 各枝を独立に評価して生成するかどうかを決定する.

図 3は WoS 引用ネットワークをクロネッカーグラフへあてはめ, 推定された初期行列から生成したネットワークと元のネットワーク構造とをグラフ特徴量による比較である. 用いた3種類の初期パラメータ行列 (a) $\begin{pmatrix}0.7 & 0.4\\ 0.4 & 0.5\end{pmatrix}$, (b) $\begin{pmatrix}0.8 & 0.5\\ 0.4 & 0.3\end{pmatrix}$, (c) $\begin{pmatrix}0.9 & 0.4\\ 0.3 & 0.4\end{pmatrix}$ は異なる初期パラメータを与えて20回の `kronfit` を実施し, 典型的なパターンを選出した. また対象の規模になるように Kronecker product を17回適用している. 結果より Out-degree, Hop plot, Scree plot, Network plot においては比較的あてはまりの良さを確認できたものの, In-degree, Triangle participation においては改善の余地があることが確認できる. (b) と (c) など, 異なる初期パラメータにより異なる推定値を得たものの特徴量の類似性が高いことから, 少ないパラメータで表現することには制限があると考えられる. また先行研究で行われた **高エネルギー物理分野の論文引用ネットワーク**への検証に関して, 本研究での追実験は WoS 引用ネットワークと同様に良いあてはまりとは言いがたい結果であった.

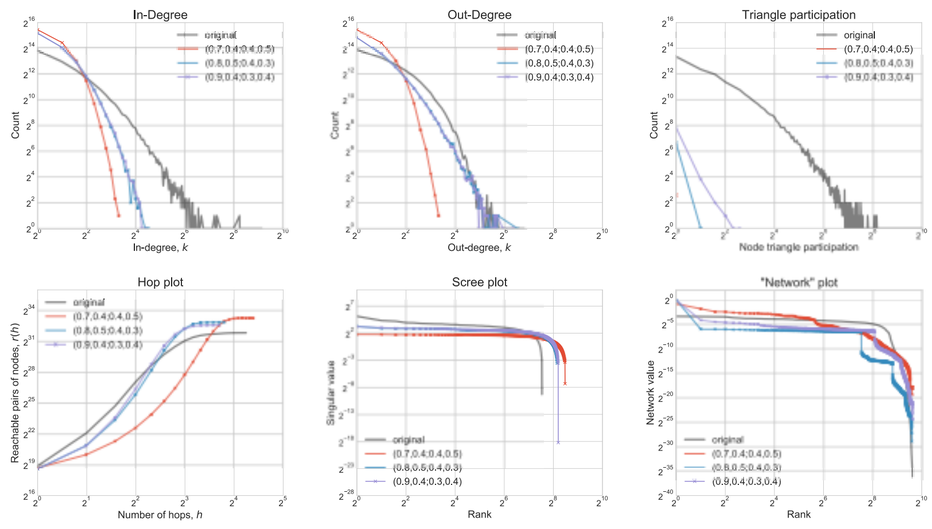

図 3: 初期行列ごとのグラフのネットワーク特徴量

3　中心性指標の時系列変化

モデルのあてはまりを評価するために, 典型的なグラフ特徴量である中心性指標を用いた時間変化ごとの構造変化に関する検証を進めた. 文献に対応するネットワーク上の各点には公開年が与えられており, $G_t = (V_t, E_t)$ を t 年までの引用関係からなるネットワーク構造を抽出することが可能である. このとき時点 t の点 $v \in V_t$ の媒介中心性 $BC_t(v) = \dfrac{1}{(|V_t|-1)(|V_t|-2)/2} \displaystyle\sum_{s,t\in V_t} \frac{\sigma(s,t \mid v)}{\sigma(s,t)}$ の時間変化を考える ($\sigma(s,t)$ は s–t 間の最短路数, $\sigma(s,t \mid v)$ は s–t 間の最短路のうち v を通る数).

図 4 は時点 t を $2005, 2006, ..., 2014$ と推移した引用ネットワーク G_{2005}, G_{2005}, ..., G_{2014} 上の媒介中心性 $BC_t(v), v \in V_t$ を, 2005年に発表された論文に着目して典型的な変化ごとにプロットした結果である. クラスタ 1 は媒介中心性の変化がない, クラスタ 2, 3 は上昇傾向もしくはその後一定値を保持, クラスタ 4, 6 は上昇後の下降を確認できる.

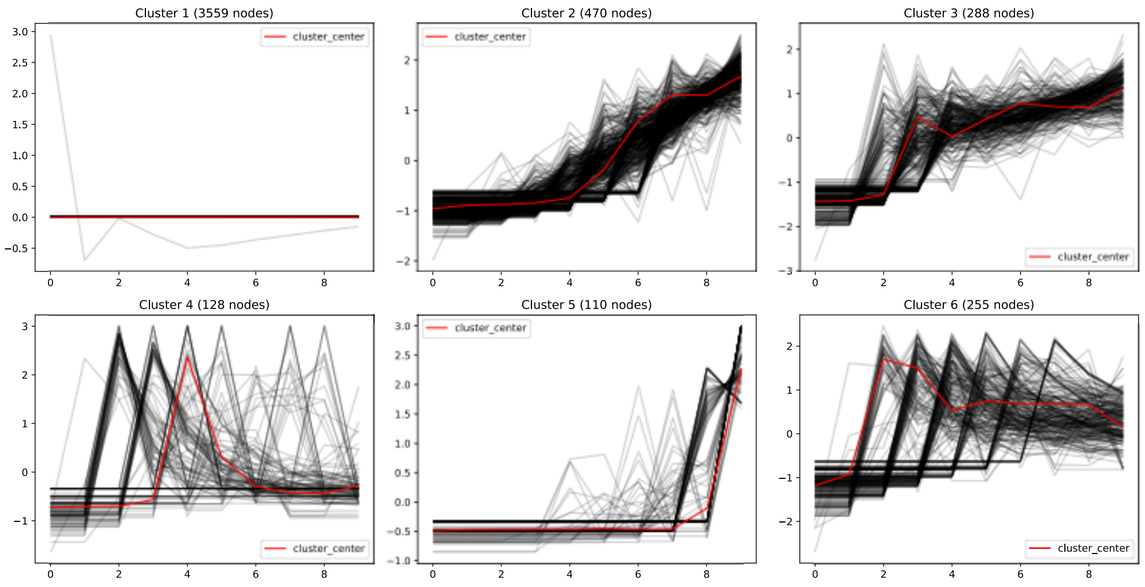

図 4: 初期行列ごとのグラフのネットワーク特徴量

4　まとめ

2. の結果より SKG は必ずしもあてはまりが良いとはいえないが, そもそも基本なネットワーク特徴量ではネットワークのあてはまりの評価を行うことは難しいともいえる. そこで 3. のように, 各点に対する特徴量の評価を傾斜して実施するために, 同様の特徴を持つノード集合 (クラスタ) ごとに評価を実施することを検討している. なお, 媒介中心性は疎密で定義されるコミュニティ間につなぐ点を重要と評価する指標であり, 引用ネットワーク上のコミュニティ (文献集合) をあるトピックと仮定すると, 同様の時系列変化をするノードはトピック間をつなぐ同様の役割を担っていると考えることができる. 引き続き, 引用ネットワークの構造を明らかにする手法について探索する予定である.

参考文献

[1] J. Leskovec et al., *Kronecker graphs: An approach to modeling networks*, *The Journal of Machine Learning Research*, **11**, 985–1042, 2010.