3F3-3

部分情報からの連結性を保証したネットワーク構造推定

Connectivity-Guaranteed Estimation of Network Structure from Partial Information

伏見 卓恭[*1*2] 斉藤 和巳[*1] 風間 一洋[*3]
Takayasu FUSHIMI Kazumi SAITO Kazuhiro KAZAMA

[*1]静岡県立大学 経営情報イノベーション研究科
Graduate School of Management and Information of Innovation, University of Shizuoka

[*2]日本学術振興会特別研究員
JSPS Research Fellow

[*3]和歌山大学 システム工学部
Faculty of Systems Engineering, Wakayama University

There is a research issue about estimating whole network structure from only ego-centric information obtained by social surveys such as qusetionnaires. For this problem, we proposed a method that estimates an adjacency matrix whose elements mean existing probability of links between corresponding nodes and extracts high probablity links in k-NN like manner. However, this method offers no guarantee that all nodes in a estimated network are weakly connected. In this paper, we propose link selection method based on the Minimum Spanning Tree algorithm and compare to the above-mentioned method. From experimental results using two real networks, we confirm that our proposed method shows somewhat higer precision than the k-NN based method.

1. はじめに

近年様々な分野において，大規模・複雑な事象をネットワークとしてとらえ，ノード間の相互関係やネットワーク構造，ネットワーク上での現象を分析する研究が盛んに行われている．これらの研究の多くは，マーケティングなどの多様な経営問題や公共政策問題の解決において重要な役割を果たすと考えらえるが，ネットワーク構造を既知として分析されることが多い．ところが，現実社会においてはプライバシーの問題や取得データ量制限などの理由から，完全な全体ネットワーク構造を知ることが困難な場合がある．従って，ネットワークに関するさまざまな種類の断片情報や統計情報を収集することにより，ネットワークの全体構造をできるだけ精緻に推定することは重要な研究課題である．

ネットワーク構造推定の既存研究として [Nowell 03, Hasan 06] などがある．これらの研究では，一部のノード間のリンク有無が既知である状況で，教師あり学習のアプローチでリンク有無が未知のノード間にリンクが存在する確からしさを推定する．すなわち，リンク構造が既知である部分から未知である部分を推定する．

本研究では，社会調査などから得られるエゴセントリック情報（自身の友人数や所属，趣味など）から，ネットワークの全体構造を推定する方法を考える．エゴセントリック（Ego-Centric）情報とは，アンケートの回答者とその人と直接交流がある人物の属性のような，自分の直接の近傍に関する局所的な情報である．しかし，アンケートではプライバシーを保護するために，実名のような個人を特定できる情報が含まれないことがある．そのような場合にはアンケートから得られる複数のエゴセントリック情報を集計しても，プライバシー保護のために匿名化されているため，ネットワーク構造は再現できないが，どの属性値のノード間にリンクが存在する傾向にあるかは把握できる．つまり，ネットワークを構成する全ノードの次数や属性が得られれば，これらの情報をもとに，次数制約を満たし，集計したリンク傾向になるように，ノード間のリンクを推定できる．著者らは，文献 [伏見 13] で，人工的に生成した属性から，属性間のリンク傾向を表す Mixing Matrix を構築し，推定ネットワークの Mixing Matrix ができるだけ一致するように，ネットワークを推定する手法を提案した．しかしこの手法では，推定されるネットワークに連結性が保証されず，平均ノード間距離などのマクロ指標の観点では異なる性質のネットワークが推定されてしまう問題があった．

そこで本稿では，有向ネットワークを対象に，推定ネットワークの全ノードが 1 つの弱連結成分になることを保証した推定法に拡張する．拡張前の推定法と比較して，提案法によりどの程度の推定精度が得られるか評価する．

連絡先: 伏見卓恭，静岡県立大学経営情報イノベーション研究科，静岡県静岡市駿河区谷田５２－１，054-264-5436

2. 構造推定法

エゴセントリック情報として，各ノードの次数およびカテゴリカル属性値が与えられた場合に，それらの情報に基づきネットワーク全体のリンク構造を推定する枠組みを以下に示す．

1. 属性情報から Mixing Matrix を構築；
2. Mixing Matrix に合うように隣接行列を推定；
3. 推定隣接行列から次数制約を満たすようにリンク選択；

2 および 3 について次節以降で詳細に説明する．

2.1 隣接行列の推定

ネットワークを構成する N 個のノードの集合 V, 各ノード $u \in V$ に対して，K 個の属性値および次数 d_u が与えられる．K 個の属性のうち k 番目の属性は $S^{(k)}$ 個の値を取りうる．$S^{(k)}$ のことを属性 k のカテゴリ数とよぶ．

ノード u の 各属性の属性値を要素とする K 次元の特徴ベクトルを $\mathbf{f}_u = (f_u^{(1)}, \ldots, f_u^{(K)})$ と表す．ただし，$f_u^{(k)} \in \{1, \ldots, S^{(k)}\}$ である．計算の便宜上，属性 k に対して，各ノードの属性値を 2 値に射影したベクトルを集めた $(N \times S^{(k)})$ の行列 $\mathbf{W}^{(k)}$ を以下のように構築する．

$$\mathbf{W}^{(k)} = (w_{u,s}^{(k)}) = \begin{cases} 1 & \text{if} \quad f_u^{(k)} = s \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

これらの情報を入力とする．

そして，属性 k に関する $S^{(k)} \times S^{(k)}$ の Mixing Matrix $\hat{\mathbf{M}}^{(k)} = (\hat{m}^{(k)}_{s,t})$ を以下のように計算する．

$$\hat{m}^{(k)}_{s,t} = \frac{1}{L} \sum_{u \in V} \sum_{v \in V \setminus \{u\}} a_{u,v} w^{(k)}_{u,s} w^{(k)}_{v,t} \tag{2}$$

ここで，ノード $u, v \in V$ 間にリンクがある場合 $a_{u,v} = 1$，そうでなければ $a_{u,v} = 0$ であり，$L = |E|$ は総リンク数を表す．

エゴセントリック情報を集計し得られた真の Mixing Matrix $\mathbf{M}^{(k)}$ と推定した Mixing Matrix $\hat{\mathbf{M}}^{(k)}$ の KL ダイバージェンスが最小になるように，ネットワーク全体の $N \times N$ の隣接行列 $\mathbf{A} = (a_{u,v})$ を推定する．

$$\begin{aligned} \hat{\mathbf{A}} &= \arg\min_{\mathbf{A}} \left\{ \sum_{k=1}^{K} \mathrm{KL}\left(\mathbf{M}^{(k)} || \hat{\mathbf{M}}^{(k)}\right) \right\} \\ &= \arg\max_{\mathbf{A}} \left\{ \sum_{k=1}^{K} \sum_{s} \sum_{t} \right. \\ &\qquad \left. m^{(k)}_{s,t} \log \sum_{u \in V} \sum_{v \in V \setminus \{u\}} a_{u,v} w^{(k)}_{u,s} w^{(k)}_{v,t} \right\} \end{aligned} \tag{3}$$

ここで，$\mathrm{KL}(P||Q)$ は，確率分布 P と Q 間の KL ダイバージェンスを表す．

式 3 に示す隣接行列 $\hat{\mathbf{A}}$ を EM アルゴリズムにより反復的に推定する．E ステップでは，属性 k の属性値 s を有するノードと t を有するノード間リンクが，ノード u, v 間のリンクである事後確率を以下のように計算する．

$$\bar{q}^{(k)}_{s,t,u,v} = \frac{\bar{a}_{u,v} w^{(k)}_{u,s} w^{(k)}_{v,t}}{\sum_{x \in V} \sum_{y \in V} \bar{a}_{x,y} w^{(k)}_{x,s} w^{(k)}_{y,t}} \tag{4}$$

ここで，$\bar{a}_{u,v}$ は現在推定されている u, v 間のリンク存在確率を表す．

また，本研究で取り扱う問題は，各ノードの次数は既知であるため，各ノードの次数を制約条件とする（$\forall u \in V, \sum_{v \in V} a_{u,v} = d_u$）．ただし，実際の隣接行列の各要素の値は $a_{u,v} \in \{0, 1\}$ であるが，ラグランジュ緩和により非負実数値（ノード間のリンク存在確率）に問題を緩和する．ゆえに，完全データの対数尤度の事後確率に関する期待値である Q 関数を以下のように計算する．

$$\begin{aligned} \mathrm{Q}\left(\mathbf{A}|\bar{\mathbf{A}}\right) &= \sum_{k=1}^{K} \sum_{s=1}^{S^{(k)}} \sum_{t=1}^{S^{(k)}} m^{(k)}_{s,t} \sum_{u \in V} \sum_{v \in V \setminus \{u\}} \bar{q}^{(k)}_{s,t,u,v} \log a_{u,v} \\ &\quad + \sum_{u \in V} \lambda_u (d_u - \sum_{v \in V \setminus \{u\}} a_{u,v}) \end{aligned} \tag{5}$$

ここで，λ_u はノード u に関するラグランジュ乗数である．

次に，M ステップでは，ラグランジュ乗数を考慮し Q 関数を $a_{u,v}$ について微分し，Q 関数を最大にする隣接行列の推定値を以下のように更新する．

$$a_{u,v} = \frac{d_u \sum_{k=1}^{K} \sum_{s=1}^{S^{(k)}} \sum_{t=1}^{S^{(k)}} m^{(k)}_{s,t} \bar{q}^{(k)}_{s,t,u,v}}{\sum_{k=1}^{K} \sum_{s=1}^{S^{(k)}} \sum_{t=1}^{S^{(k)}} m^{(k)}_{s,t} \sum_{v \in V \setminus \{u\}} \bar{q}^{(k)}_{s,t,u,v}} \tag{6}$$

上述した E ステップと M ステップを推定パラメータである隣接行列が収束するまで繰り返し，ノード間のリンク存在確率を要素とした隣接行列を推定する．そして，各ノードに関して確率が高い順に次数分だけリンク先を決定する．

非負実数値に緩和したこの問題を解くことは，凸集合上での凸関数の最適化であるため，初期値に依存しない大域的最適解を求めることができる．

2.2 リンク付与

推定した隣接行列の要素の値 $a_{u,v}$ は，ノード u, v 間のリンク存在確率を意味する．したがって，各ノードに対して，隣接相手ノードとしてリンク存在確率の高いノードを次数分選択することで，推定ネットワークを得る．$R = \{(u_x, v_x); \forall x < y, \hat{a}(u_x, v_x)\hat{a}(u_y, v_y)\}$ は，リンク存在確率による降順ノードペアリストであり，$R(u) = \{(u, v_x); \forall x < y, \hat{a}(u, v_x)\hat{a}(u, v_y)\} \subset R$ は，ノード u と隣接相手ノード間のリンク存在確率の降順リストである．$R_d(u) = \{(u, v_x); xd\} \subset R(u)$ はノードペア降順リストのうち上位 d 件のリストである．

2.2.1 k-NN ベースリンク付与法

文献 [伏見 13] では，任意のノード u に対して，リンク存在確率が高い順に d_u ノードを選ぶことで，隣接相手ノードを選定する．したがって，k-NN グラフの構築法をベースにリンクを選択・付与する手法である．すなわち，推定リンク集合は

$$\hat{E} = \{(u, v); \forall u \in V, v \in R_{d_u}(u)\}$$

となる．各ノードがリンク存在確率の高いノードを次数分選択するだけなため，推定したネットワーク $\hat{G} = (V, \hat{E})$ が 1 つの連結成分になることは保証されない．

2.2.2 MST ベースリンク付与法

本稿では，推定ネットワーク全体が 1 つの連結成分になることを保証するために，MST（Maximum Spanning Tree）をベースとしたリンク選択・付与法を提案する．まず，R からリンク存在確率を重みとした MST を生成する．ここで MST とは，確率値を重みとした全ノードペアの中から，リンクの重み和が最大となり，かつ，全ノードが連結になるようにリンクを選択し生成された木である．ただし本稿では，連結とは弱連結成分を意味する．具体的には，ノードペアリスト R から順にノードペア間にリンクを付与する．このとき，通常の MST と同様に，同一の木に属するノードペアにはリンクを付与しない．さらに，既知の情報である各ノードの出次数を超えない範囲内でノードペアを選択する．全ノードが一つの木に属するまで R からのノードペア選択を繰り返し，MST（$T = (V, \hat{E})$）を出力する．この時点では，出次数制約を満たさない，すなわち，出次数が不足するノードも存在する．

従って，次に各ノードに対して不足分のリンクを $R \backslash \hat{E}$ から順に選択し，付与する．

3. 評価実験

全体構造が既知なネットワークに対して，次数のみを抽出し，人工属性を割り当てる．次数および属性のみからネットワーク全体のリンク構造を推定する．

3.1 ネットワークデータ

評価実験に用いるネットワークデータについて述べる．

1 つ目のネットワークは，Web のハイパーリンクネットワークである．大学のウェブサイト（http://cis.k.hosei.ac.jp/）内のページを 2010 年 8 月に収集し，ウェブサイトのハイパーリンク構造からハイパーリンクネットワークを構築した．ノード数は 600，有向リンク数は 1,833 である．本稿では Hosei ネットワークと呼ぶ．

2 つ目のネットワークは，国際会議 NIPS(Neural Information Processing Systems) の第 1 から 12 回に発表された論文の共著者ネットワークである．各著者をノードとし，二人の著者が少なくとも一つの共著論文を発表していれば，その著者間にリンクを付与する．このように構築したネットワークにおいて，最大の連結成分を抽出した．ノード数は 1,036，リンク数は 2,044 である．本稿では Nips ネットワークと呼ぶ．

3.2 人工属性

本研究では，上記ネットワークを用いた評価に際して，Voter Model [Liggett 99] により Assortativity [Newman 02] の高い属性値を各ノードに割り当てる．

Voter Model とは，ネットワーク上での意見形成過程をモデル化したもので，PageRank と同様に，離散タイムステップで展開する確率モデルである．各ノード $u \in V$ は，自身の親ノード（無向ネットワークの場合隣接ノード）からランダムに選択した親ノードを選択し，その親ノードが時刻 t で有する意見を時刻 $t+1$ で採用する．この試行を繰り返すことで，同一の意見を有するノードが近傍に存在するようになる．すなわち，隣接するノードどうしは同一の意見を持つ確率が高くなり，Assortative な状態になる．この意見をノード属性とすれば，ネットワーク全体で Assortative になるようなノード属性を割り当てることができる．具体的には，タイムステップ $t=0$ に，各ノードに S 個の値をとりうるランダムな属性値を割り当てる．そして，上述したように，タイムステップが進むにつれ，各ノードは親ノードの有する属性値の 1 つを選択し，次のタイムステップに自身の属性として割り当てることを繰り返す．実際には，10 ステップほど繰り返すことで，Assortative な属性値を割り当てることができる．

3.3 評価法

推定したネットワークの推定精度をリンク集合の F 値とコミュニティ正解率により定量的に評価する．

真のネットワークのリンク集合を E^*, 推定したネットワークのリンク集合を $\hat{E}$ と表記する．以下のように F 値を計算する．

$$F(E^*, \hat{E}) = \frac{2|E^* \cap \hat{E}|}{|E^*| + |\hat{E}|} = \frac{|E^* \cap \hat{E}|}{|E^*|} \tag{7}$$

ここで，次数制約のため，ネットワーク全体のリンク数は真のネットワークと推定ネットワークとで等しいことに注意する ($|E^*| = |\hat{E}|$)．すなわち，再現率と適合率は等しくなる．

推定法の特性を評価するために，推定したリンク相手ノード (子ノード) が属するコミュニティに注目する．真のネットワークにおけるノード u の子ノード集合を $F(u) = \{v; (u,v) \in E\}$, 推定ネットワークにおけるノード u の子ノード集合を $\hat{F}(u) = \{v; (u,v) \in \hat{E}\}$ とする．また，各ノードは H 個のコミュニティのいずれかに属するとし，ノード u の属するコミュニティ番号を $c(u) \in \{1, \ldots, H\}$ と表記する．この時，真のネットワーク，および，推定ネットワークにおけるノード u の子ノードが属するコミュニティ分布をそれぞれ $y_u(h) = |\{v; c(v) = h, v \in F(u)\}|$, $\hat{y}_u(h) = |\{v; c(v) = h, v \in \hat{F}(u)\}|$ と表す．ノード u のリンク相手である子ノードの属するコミュニティの正解率として，

$$ca(u) = \frac{1}{|F(u)|} \sum_{h=1}^{H} \min(y_u(h), \hat{y}_u(h)) \tag{8}$$

を定義する．式 8 を全ノードで平均した値 $CA = 1/|V| \sum_{u \in V} ca(u)$ をコミュニティ正解率として定義する．

3.4 実験設定

上述したネットワークに独立に生成した人工属性を複数割り当てる．抽出した次数と割り当てた属性のみからネットワーク全体の隣接行列を推定する．人工属性の生成はそれぞれ独立に 10 回実行し，それらの属性に基づきネットワークを推定し，10 回の F 値の平均値およびコミュニティ正解率をもって定量的に評価する．比較のため，属性を用いず，次数制約のみを付与しランダムにリンクを推定するランダム法を用いる．また，コミュニティ正解率の算出には，CNM コミュニティ [Clauset 04] を用いる．

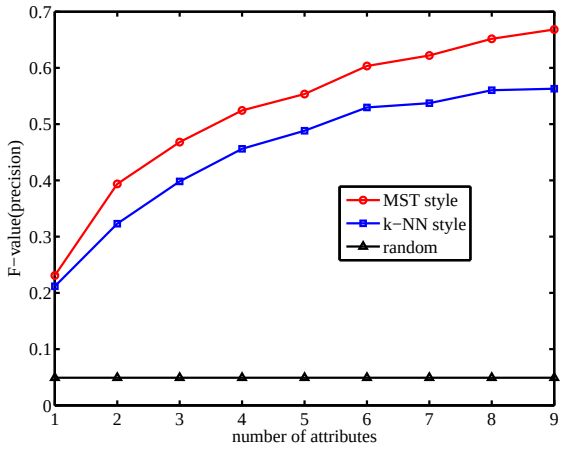
3.5 評価結果と考察

推定に使用する属性の数と推定精度の関係を評価する．図 1(a) と (b) に各ネットワークの属性数を変えた際の推定精度を図示する．経験上，実際のアンケートなどで得られる属性数に合わせるため，属性数 K を $1 \leq K \leq 9$ とした．使用する属性のカテゴリ数は $S^{(k)} = 5$ とした．各図において，横軸に属性数，縦軸に F 値をプロットした．それぞれ，赤○線は MST を，青□線は k-NN をベースとしたリンク付与法によるもの，黒△線はランダム法を用いた推定精度である．

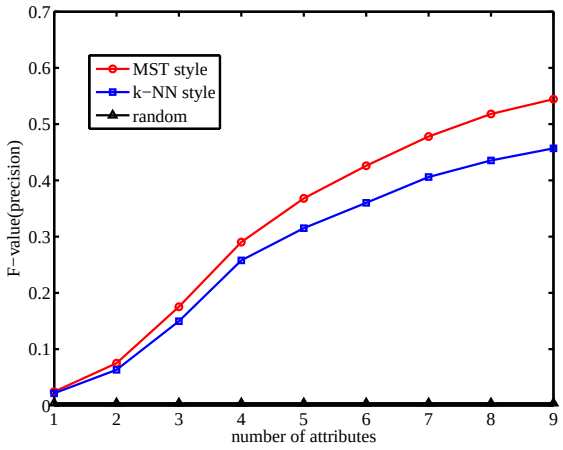
図 1 を見ると，どちらのネットワークにおいても，使用する属性数が多いほど推定精度が高くなることがわかる．これは，独立に生成した人工属性が多いほど，どの属性のノード間にリンクが存在するかという条件が絞られてくることが原因だと考えらえる．さらに，k-NN ベースより MST ベースのリンク付与法の方が幾分か推定精度が高くなっていることがわかる．Assortative な属性を用いた推定隣接行列は，同一コミュニティ内のノードペア間のリンク存在確率が高くなる傾向にあり，k-NN ベースでは，同一コミュニティ内のノード同士がリンクし合うようにリンクが付与される．一方，MST ベースでは，ツリーの制約から，閉路が出現せず，同一コミュニティ内の過度のつながりを軽減している．ゆえに，相対的に異なるコミュニティのノードとのリンクを再現でき，F 値が高くなっていると考えられる．また，両手法とも属性数を増やすに従い，推定精度の向上率は小さくなる傾向になる．これは，独立に生成した人工属性ではあるが，使用する属性数が多いほど少なからず相関を持つ属性が現れることが原因だと考えられる．また，どちらのリンク付与法を用いても，ランダム法より高い推定精度が得られることがわかる．

図 2(a) と (b) に各ネットワークのコミュニティ数を変えた際のコミュニティ正解率を図示する．各図において，横軸にコミュニティ数，縦軸にコミュニティ正解率をプロットした．それぞれ，赤○線は MST を，青□線は k-NN をベースとしたリンク付与法によるもの，黒△線はランダム法を用いたコミュニティ正解率である．

図 2 を見ると，どちらのネットワークも，k-NN ベースのリンク付与法，MST ベースのリンク付与法共に高い正解率を示していることがわかる．すなわち，精度（F 値）では 6 割未満であるが，推定リンク先が同一コミュニティである場合を許容した場合は，高い精度で推定できていることを意味する．特に，Nips ネットワークでは，MST ベースのリンク付与法の方がやや正解率が高いこともわかる．MST をベースにしていることで，異なるコミュニティへのリンクが推定しやすくなったことが影響し，コミュニティを跨ぐリンクを再現できたことによるものと考えられる．さらに，CNM コミュニティ数を変化させても，正解率に大きな変化が見られなく，頑健な結果が得られた．また，どちらのリンク付与法を用いても，ランダム法より高いコミュニティ正解率が得られることがわかる．

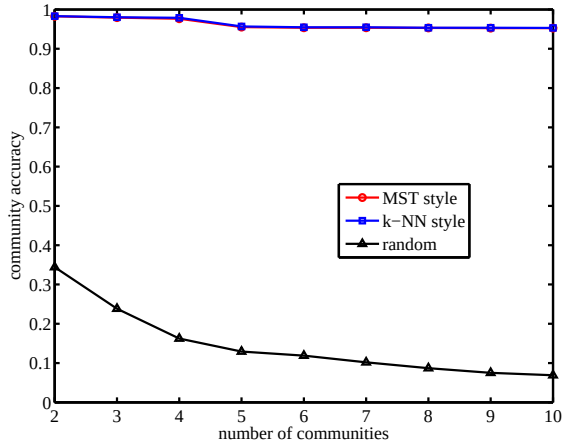


(a) Hosei ネットワーク

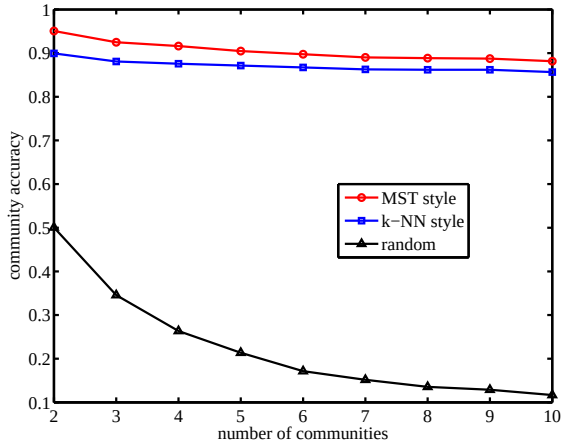


(b) Nips ネットワーク

図 1: 属性数の変化と推定精度



(a) Hosei ネットワーク



(b) Nips ネットワーク

図 2: 属性数の変化とコミュニティ正解率

4. おわりに

本研究では，アンケートなどから得られるエゴセントリック情報のみを用いてネットワーク全体構造を推定する問題について，EM アルゴリズムにより推定した隣接行列から，リンクを選択・付与する方法を比較・検討した．複数のネットワークを用いた結果，本稿で提案した MST をベースとした手法の方が従来の k-NN をベースとした手法やランダム法と比較して高い推定精度を得られることを確認した．またコミュニティ正解率の観点からも，安定した推定結果が得られることを確認した．

本稿では Assortativity が高い人工属性を用いたが，Assortativity の高さではなく Mixing Matrix の偏りが大きい属性を用いた評価も進めていきたい．さらに，どの程度の推定精度が得られれば現実問題への応用が可能かを評価していきたい．

謝辞 本研究は，科学研究費補助金 (No.25・10411) の補助を受けた．

参考文献

[Clauset 04] Clauset, A., Newman, M. E. J., and Moore, C.: Finding community structure in very large networks, *Physical Review E*, Vol. 70, No. 6, pp. 066111+ (2004)

[Hasan 06] Hasan, M. A., Chaoji, V., Salem, S., and Zaki, M.: Link prediction using supervised learning, in *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security* (2006)

[Liggett 99] Liggett, T. M.: *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes (Grundlehren der mathematischen Wissenschaften)*, Springer, 1 edition (1999)

[Newman 02] Newman, M. E. J.: Assortative mixing in networks, *Structure*, Vol. 2, No. 4, p. 5 (2002)

[Nowell 03] Nowell, D. L. and Kleinberg, J.: The link prediction problem for social networks, in *CIKM '03: Proc. of the 12th international conf. on Information and knowledge management*, pp. 556–559, (2003)

[伏見 13] 伏見 卓恭, 斉藤 和巳, 風間 一洋：エゴセントリック情報からのネットワーク構造推定, 第 6 回 Web とデータベースに関するフォーラム (WebDB Forum2013) (2013)