ノード属性情報からのリンク存在確率の推定
Estimating Link Existence Probabilities from Node Attributes Information

伏見 卓恭[♡] 斉藤 和巳[◇] 風間 一洋[♠]

Takayasu FUSHIMI Kazumi SAITO
Kazuhiro KAZAMA

本稿では，各ノードの属性が既知だが隣接関係が未知であるネットワークに対して各ノードペアのリンク存在確率を推定することを試みる．著者らは，社会調査などから得られるエゴセントリック情報のみを用いて，ノード間のリンク存在確率を推定する手法を提案している．エゴセントリック情報は，アンケートの回答者自身の属性や友人数，回答者と直接交流がある人物の属性情報のみからなる．アンケートではプライバシーの問題から個人を特定できる実名などを記入できないが，エゴセントリック情報を集計することで，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を **Mixing Matrix** として知ることができる．我々の手法は，社会調査集計結果から得られた観測 **Mixing Matrix** と推定 **Mixing Matrix** の **KL** ダイバージェンスを最小化するように対象ネットワークの隣接行列を推定する．本稿では，既存研究で用いられているモデルを本稿の問題設定に合わせた推定手法と，我々の提案手法を定量的に比較する．

In this paper, we address a problem of estimating the link existence probabilities of an actual social network under the situation that node attributes are given but adjacency relationships are unknown. For this problem, we have proposed an estimation method from only ego-centric information which consists of two types of attributes, personal attributes like sex and relational ones like the numbers of female and male friends. From these attribute values, we can obtain some macroscopic information about the network, like the ratio of number of links between female and male to the total number of links, as mixing matrices. However, we cannot directly know the actual connections between two nodes only from these observed mixing matrices. Thus, our method estimates the whole structure of the hidden network by minimizing the Kullback-Leibler divergence between the observed and estimated mixing matrices. In this paper, we compare our method to an existing model and show that the our method can produce much better estimation results.

♡ 正会員 静岡県立大学大学院経営情報イノベーション研究科，日本学術振興会特別研究員 takayasu.fushimi@gmail.com
◇ 正会員 静岡県立大学大学院経営情報イノベーション研究科 k-saito@u-shizuoka-ken.ac.jp
♠ 正会員 和歌山大学システム工学部 kazama@ingrid.org

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．良い推定精度のネットワークが得られれば，部分的な情報しか得られない場合でも情報拡散モデルや意見形成モデルのシミュレーションが可能となり，高影響度ノードを抽出できるようになる．

ネットワーク構造推定の既存研究として [1, 2, 3, 4, 5, 6, 7] 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，既知のリンクとノードの関係から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．

著者らは，社会調査などから得られるエゴセントリック情報のみから，ネットワークの全体構造を推定する方法を提案した [8]．エゴセントリック情報は，回答者自身の性別や世代などのパーソナル属性と，回答者の属性ごとの友人数などのリレーショナル属性からなると想定する．これらの情報を集計することで，男性と女性のつながる比率などのマクロ情報を Mixing Matrix として得ることができる．この前提をもとに，提案手法は社会調査集計結果から得られた観測 Mixing Matrix と推定 Mixing Matrix の KL ダイバージェンスを最小化するように対象ネットワークの隣接行列を推定する．全てのリンクが未観測である点で，既存研究とは異なる．

本稿では，既存研究 [5, 7] で用いられているモデルを次章で述べる問題設定に合わせて，ノード間のリンク存在確率を推定する．そして，我々の提案手法と比較する．評価実験では，推定したネットワークの精度を評価するために，全体構造が既知である 3 つのネットワークデータを用いる．これらのネットワークに対して，Voter Model により人工的に属性を割り当て，推定精度を定量的に評価する．推定精度は，推定に使用する属性の Assortativity に依存すると考えられる．従って，Assortativity のレベルと使用する属性数，属性のカテゴリ数を変化させて推定精度を比較，評価する．

2. 問題設定

本研究の対象とする問題は，社会調査などの N 人の回答者をノードとする．ノード集合を V，カテゴリカル属性数を K とする．カテゴリカル属性の例として，性別や所属などがあるが，このような回答者自身に関する情報をパーソナル属性と呼ぶことにする．属性 k のカテゴリ数を $S^{(k)}$ とし，各カテゴリに 1 から $S^{(k)}$ の整数値を割り当てる．パーソナル属性から，各ノード u に対し，属性値を要素とした K 次元属性ベクトル $\mathbf{f}_u = (f_u^{(1)}, \ldots, f_u^{(K)})$ を構築できる．ここで，k 番目の要素は $f_u^{(k)} \in \{1, \ldots, S^{(k)}\}$ である．

また，社会調査からノード u の友人の中で，属性 k の値が t である人数 $g_u^{(k)}(t)$ を得ることができる．これをリレーショナル属性と呼ぶことにする．すべての属性値で和をとると，$d_u = \sum_{t=1}^{S^{(k)}} g_u^{(k)}(t)$ となり，ノード u の次数が得られる．このとき，属性 k に対して，以下のように Mixing Matrix の要素 $m_{s,t}^{(k)}$ を集計できる．

$$m_{s,t}^{(k)} = \frac{1}{L} \sum_{\{u: f_u^{(k)} = s\}} g_u^{(k)}(t) \qquad k = 1, \ldots, K.$$

ここで，$L = \sum_{u \in V} d_u$ は総リンク数を表す．$m^{(k)}_{s,t}$ は，属性 k の値が s であるノードと t であるノードがつながっている割合を意味する．このようにして，観測 Mixing Matrix $\mathbf{M}^{(k)} = [m^{(k)}_{s,t}]$ を得ることができる．Mixing Matrix を構築できれば，Newman の文献 [9] に則り Assortative 係数を計算できるため，属性が Assortative であるのかどうかを判別できる．

ここで，目的関数を導出するため，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．

$$\mathbf{W}^{(k)} = [w^{(k)}_{u,s}] = \begin{cases} 1 & \text{if} \quad f^{(k)}_u = s \\ 0 & \text{otherwise.} \end{cases}$$

隣接行列 $\mathbf{A} = [a_{u,v}]$ が既知の場合は，属性 k に関する $S^{(k)} \times S^{(k)}$ の Mixing Matrix $\hat{\mathbf{M}}^{(k)}(\mathbf{A}) = [\hat{m}^{(k)}_{s,t}(\mathbf{A})]$ を以下のように計算する．

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

ここで，ノード $u, v \in V$ 間にリンクがある場合 $a_{u,v} = 1$ であり，そうでなければ $a_{u,v} = 0$ である．しかし，本稿の問題設定では隣接行列は未知であるため，観測 Mixing Matrix $\mathbf{M}^{(k)}$ と推定隣接行列により計算した推定 Mixing Matrix $\hat{\mathbf{M}}^{(k)}(\mathbf{A})$ の KL ダイバージェンスが最小になるように，隣接行列 $\mathbf{A} = [a_{u,v}]$ を推定する．

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \left\{ \sum_{k=1}^{K} \mathrm{KL}\left(\mathbf{M}^{(k)} \| \hat{\mathbf{M}}^{(k)}(\mathbf{A}) \right) \right\}. \quad (1)$$

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 (1) を以下のように変形する．

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

従って，以下の目的関数を解くことにより，観測 Mixing Matrix に合うような最適な隣接行列 $\mathbf{A}$ を得ることができる．

$$J(\mathbf{A}) = \sum_{k=1}^{K} \sum_{s=1}^{S^{(k)}} \sum_{t=1}^{S^{(k)}} m^{(k)}_{s,t} \log \sum_{u \in V} \sum_{v \in V \setminus \{u\}} a_{u,v} w^{(k)}_{u,s} w^{(k)}_{v,t}. \quad (2)$$

本稿で取り扱う問題は，各ノードの次数は既知であるため，各ノード $u \in V$ に対して以下のような次数制約を付与する．

$$\sum_{v \in V \setminus \{u\}} a_{u,v} = d_u.$$

3. 解法アルゴリズム

EM アルゴリズムを用いて，目的関数 (2) を最大化する．まず，属性 k の属性値 s を有するノードと t を有するノードが，ノード u, v 間のリンクをもつ事後確率を以下のように計算する．

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

ここで，$\bar{a}_{u,v}$ は隣接行列の現在の推定値を表す．さらに，完全データの対数尤度の事後確率に関する期待値を Q 関数として以下のように定義する．

$$Q\left(\mathbf{A}|\bar{\mathbf{A}}\right) = \sum_{k=1}^{K} \sum_{s=1}^{S^{(k)}} \sum_{t=1}^{S^{(k)}} m^{(k)}_{s,t} \sum_{u \in V} \sum_{v \in V \setminus \{u\}} \bar{q}^{(k)}_{s,t,u,v} \log a_{u,v}.$$

本稿の問題設定から，各ノードの次数は既知であるため，以下のようにラグランジュ項により制約条件を付与する．

$$\tilde{Q}\left(\mathbf{A}|\bar{\mathbf{A}}\right) = Q\left(\mathbf{A}|\bar{\mathbf{A}}\right) + \sum_{u \in V} \lambda_u \left(d_u - \sum_{v \in V \setminus \{u\}} a_{u,v} \right).$$

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．本来，隣接行列の値は $a_{u,v} \in \{0, 1\}$ であるが，ラグランジュ緩和により実数値 $a_{u,v} \in [0, 1]$ とした問題を解く．緩和した要素を持つ隣接行列を緩和隣接行列と呼ぶことにする．

EM アルゴリズムの E ステップにおいて，$\tilde{Q}$ 関数を推定し，M ステップでは，$\tilde{Q}$ 関数をパラメータ $a_{u,v}$ と λ_u について偏微分し，$\tilde{Q}$ 関数を最大にする緩和隣接行列の推定値を以下のように更新する．

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

そして，緩和隣接行列の推定値が収束するまで，E ステップと M ステップを繰り返す．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．

ここで，提案手法をまとめる．ノード集合 V と K 個のカテゴリカル属性を入力とし，ネットワークの全リンクを以下のステップで推定する．

STEP1

- 各ノード u に対して, K 個のカテゴリカル属性値を抽出し，K 次元の属性ベクトル [パーソナル属性] $\mathbf{f}_u$ を構築する；
- 各ノード u に対して，友人数 d_u と各属性 k の属性値 t ごとの内訳人数 [リレーショナル属性] $g^{(k)}_u(t)$ を抽出する；
- 各属性 k に対して，観測 Mixing Matrix $\mathbf{M}^{(k)}$ を構築する;

STEP2

- 観測 Mixing Matrix $\mathbf{M}^{(k)}$ と推定 Mixing Matrix $\hat{\mathbf{M}}^{(k)}(\mathbf{A})$ の KL ダイバージェンスが最小になるように緩和隣接行列 $\hat{\mathbf{A}}$ を推定する；

アルゴリズム全体として，推定 Mixing Matrix と緩和隣接行列の更新に時間計算量 $O(K|V|^2)$ を要する．

4. 関連研究

ネットワークのリンク構造を推定する既存手法では，リンク構造の既知の部分が与えられ，未知の部分を予測するというものが多い．既知の情報として，リンク構造に関する情報，ノード属性に関する情報がある．文献 [1] では，共通隣接ノード指標，Katzβ 指標などいくつかのリンク指標を用いて既知のリンク集合からノード間のリンク存在の確からしさを定量化し，未知のリンク集合を予測する手法を提案しており，高い精度が得られている．文献 [2] では，類似性質を有するノード間にリンクが存在するという直観に基づき，ノード属性の特徴ベクトルの内積からカーネルを構築し，リンク構造を推定する手法を提案している．さらに，文献 [3] では，ノード属性と既知のリンク構造の情報の両方を組み合わせる手法として，既知構造のグラフラプラシアンの拡散カーネルから未知のリンク集合を推定する手法を提案している．文献 [4] では，既知構造からネットワークの階層性を表すデンドログラムを MH アルゴリズムにより構築し，未知のノードペア間のリンク存在確率を推定している．文献 [6] では，既知構造から行列分解を利用して潜在属性を学習し，協調フィルタリングと同様の枠組みで未知ノードペア間のリンクを推定している．これらの既存研究は，ノードの属性に着目しリンクの傾向を用いる点で本研究と関

連している．しかし，既存研究は既知のリンク構造を陽に用いるが，本稿の問題設定は各ノード自身の属性と隣接相手の属性のみの情報からなるエゴセントリック情報のみを用いる点で異なる．また，属性が類似するノード間にリンクが存在すると推定するものが多いが，我々の手法では Mixing Matrix の値を考慮しているため，同一属性間だけでなく，異なる属性間の関連性の強弱も適切に設定することができる．

文献 [5, 7] では，ノンパラメトリックベイズの手法を用いて潜在属性の数を限定しない枠組みで，ノード間のリンク存在確率を推定するモデルが提案されている．これらの手法では，既知のリンク構造を用いる点で上述の既存研究と同じであるが，ノード間のリンクモデルに関しては，Mixing Matrix 同様の行列と各ノードの属性情報を用いてモデリングしている．

$$\hat{a}_{u,v} = \sigma\left(\sum_{k=1}^{K} \mathbf{w}_u^{(k)} \mathbf{M}^{(k)} \mathbf{w}_v^{(k)}\right) = \sigma\left(\sum_{k=1}^{K} \sum_{s=1}^{S^{(k)}} \sum_{t=1}^{S^{(k)}} m_{s,t}^{(k)} w_{u,s}^{(k)} w_{v,t}^{(k)}\right). \quad (3)$$

ここで，$\sigma(\cdot)$ はシグモイド関数を表す．本稿では，社会調査から得られるリレーショナル属性から構築する Mixing Matrix とパーソナル属性をこのモデルに適用し推定した結果と比較する．

5. 評価実験

上述したように，本研究の目的は，社会調査から得られるエゴセントリック情報のみからネットワークの全体構造を推定することである．しかし本稿では，提案手法による推定精度を定量的に評価するために，全体構造が既知なネットワークを用いて評価する．さらに，推定精度は使用する属性の属性数やカテゴリ数，Assortativity などによって変化すると考えられる．従って，Assortative な属性を用いる場合，Non-Assortative（ランダム）な属性を用いる場合，推定に使用する属性数を変化させた場合，各属性のカテゴリ数を変化させた場合をそれぞれ評価するために，人工的に属性値を割り当てる．そして，使用する属性と提案手法の推定精度の関係を評価する．さらに，ネットワーク統計量とコミュニティの観点から提案手法の特性を評価する．

5.1 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，Ravasz らによる階層性のあるネットワークを生成するモデルにより構築した人工ネットワークである [10]．ノード数は 125，リンク数は 410 である．本稿では Hierarchical ネットワークと呼ぶ．

2 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無向化した．ノード数は 600，リンク数は 1,299 である．本稿では Hosei ネットワークと呼ぶ．

3 つ目のネットワークは，国際会議 NIPS(Neural Information Processing Systems) の第 1〜12 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．このようにして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ものである．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．

5.2 人工属性

この節では，Voter Model [11] を用いて，各ノードに Assortativity の高い属性値を割り当てる方法を説明する．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．$h_t(u)$ を時刻 t においてノード u が有している意見，$\Gamma(u)$ をノード u の親／隣接ノード集合とする．ネットワーク $G = (V, E)$ に対して，S 個の整数値を割り当てる：

VM1 初期化：各ノードに初期値を一様ランダムに割り当てる $h_0(u) \leftarrow s \in \{1, \ldots, S\}$，$t \leftarrow 1$；

VM2 更新：各ノードの時刻 t の意見として，親ノードを一様ランダムに選択し，その親ノードが時刻 $t-1$ に有する意見を割り当てる $h_t(u) \leftarrow h_{t-1}(v), v \in \Gamma(u)$；

VM3 終了条件：Assortative 係数が $r > 0.8$ となったら終了；さもなくば，$t \leftarrow t+1$ とし，VM2 へ；

評価実験では，およそ 10 反復ほどで所望の意見分布が得られた．Voter Model により得られた整数値を各ノードの属性値とする．

5.3 実験設定

提案手法による推定精度と属性の関係を評価するために，Voter Model による Assortative 属性とランダムに割り当てた Non-Assortative 属性の 2 つのタイプの属性群を用意する．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 値の平均値をもって定量的に評価する．

具体的には，Voter Model の反復終了ステップを T と表すと，ノード u の属性 k の属性値を $f_u^{(k)} \leftarrow h_T(u)$ とすることで Assortative な属性値を割り当て，$f_u^{(k)} \leftarrow h_0(u)$ とすることで Non-Assortative な属性値を割り当て，提案手法（以下，proposed 法）により推定する．さらに，属性を用いず次数制約のみでランダムにリンクを推定する random 法，モデル (3) を用いた existing 法と比較する．

5.4 属性数と推定精度に関する評価結果

推定したネットワークの推定精度をリンク集合の F 値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．

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

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．

図 1 に属性数 K を 1 から 9 の範囲で変化させた際の，それぞれの手法での推定精度 F 値を示す．各属性のカテゴリ数を $S^{(k)} = 5$ とした．各図において，proposed 法 × Assortative 属性を赤●，proposed 法 × Non-Assortative 属性を青■，existing 法（× Assortative 属性）を緑▲，random 法を黒線でプロットした．

図 1 より，各ネットワークにおいて以下の 2 つの事柄が言える；1)proposed 法 × Assortative 属性は，推定に使用する属性数が多いほど推定精度が安定して改善される．2)proposed 法 × Non-Assortative 属性では，使用する属性数を増やしても，十分な推定精度の改善がみられない．特に，比較的大きなサイズの Nips ネットワークでは，ほとんど改善がみられない．これらのことから，高い Assortative 係数の属性情報が十分な数得られる状況において，提案手法により良い推定精度が得られることが示唆された．特に，現実のソーシャルネットワークなどでは，人々は同一属性値を持つ人と繋がる傾向にあるため，そのような性質のネットワークに対して有効であることが期待できる．一方で，Hierarchical ネットワーク程度の小さいネットワークに対しては，Assortative 係数が低い属性情報からでも，独立な複数の属性を用いればリンク相手ノードを絞り込むことができ，ある程度の推定精度が得られることが確認できた．また，Assortative な属性を用いた existing 法や属性情報を用いない random 法と比較しても，proposed 法 × Assortative 属性の結果が良いことが確認できた．

existing 法では，Mixing Matrix の値を類似度として扱っているため，Mixing Matrix の要素の値が大きい属性値を有するノードペア間のリンクを過剰に推定しており (over-estimate)，Assortative 係数が真のネットワークより大きくなっていることから，観測

Mixing Matrix と幾分か離れた推定結果が得られたと考えられる．一方提案手法では，観測 Mixing Matrix と隣接行列の現在の値により計算する推定 Mixing Matrix の KL ダイバージェンスを反復計算することで，より観測値に近い推定結果が得られたと考えられる[1]．

5.5 カテゴリ数と推定精度に関する評価結果

図 2 に属性のカテゴリ数 $S^{(k)}$ を 2 から 10 の範囲で変化させた際の，それぞれの手法での F 値を示す．属性数を $K = 1$ とした．

図 2 より，各ネットワークにおいて属性数を変化させた実験と同様な 2 つの事柄が言える；1)proposed 法 × Assortative 属性は，推定に使用する属性のカテゴリ数が多いほど推定精度が安定して改善される．2)proposed 法 × Non-Assortative 属性では，カテゴリ数を増やしても，十分な推定精度の改善がみられない．これらのことから，高い Assortative 係数の属性情報が 1 つしか得られない状況においては，カテゴリ数が多い属性であれば提案手法により良い推定精度が得られることが示唆された．一方で，Hierarchical ネットワーク程度の小さいネットワーク以外では，Assortative 係数が低い属性情報からでは，高い推定精度が得られないことが確認できた．また，Assortative な属性を用いた existing 法や属性情報を用いない random 法と比較しても，proposed 法 × Assortative 属性の結果が良いことが確認できた．

図 1 と図 2 の属性数とカテゴリ数を変化させた実験結果から，カテゴリ数の多い Assortative な属性が十分得られれば，エゴセントリック情報のみからネットワークの全体構造を推定できることが示唆された．

5.6 ネットワーク統計量の観点での評価結果

各手法により推定したネットワーク構造を，ネットワーク統計量の観点から評価する．本稿では，平均クラスタ係数 $\bar{C}$, クラスタ係数の標準偏差 $C_{S.D.}$, 平均ノード間距離 $\bar{D}$ を採用する．ここで，平均ノード間距離はネットワークの非連結性に対処するために，最短パス長の調和平均により計算する．すなわち，到達できないノードへの距離は無限大とし，その逆数は 0 となる．

表 1 は，真のネットワーク (true)，先ほどの評価で用いた $K = 9$, $S^{(k)} = 5$ の Assortative 属性を用いて proposed 法と existing 法により推定したネットワーク，random 法により推定したネットワークの統計量を比較している．表 1 から，各ネットワークにおいて以下の 2 つの事柄が言える；1) 平均クラスタ係数 $\bar{C}$ は，比較した 2 つの手法によるネットワークと比べて，proposed 法による推定ネットワークが有意に真のネットワークと近い．2) クラスタ係数の標準偏差 $C_{S.D.}$ に関しても，proposed 法による推定ネットワークが妥当な近さを示している．属性がネットワーク上で Assortative な状態であるとは，隣接するノード同士が同一の属性値を有することを意味する．すなわち，局所構造を定量化するクラスタ係数の比較結果から，proposed 法は局所的な構造をより正確に推定できることが示唆された．3) 小規模なネットワークにおいては，平均ノード間距離 $\bar{D}$ もある程度妥当な近さの値を示している．反対に，比較的大規模なネットワークにおいては，平均ノード間距離は真のネットワークと離れた値を示している．これは，大規模なネットワークになるほど，推定ネットワークが非連結になる可能性が高くなることが原因だと考えられる．一方 random 法は，スモールワールドモデル [12] のように，小さな平均ノード間距離を実現できている．existing 法に関しては，F 値による評価の考察と同様に，Assortative 係数が非常に高くなるように over-estimate されていることから，平均クラスタ係数では幾分か，平均ノード間距離では非常に，真のネットワークと離れた値を示していると考えられる．これらの結果から，提案手法である proposed 法は局所的な構造をより正確に推定できることが示唆された．

[1]他手法と比べて，Assortative 係数が真のネットワークと非常に近いことは言うまでもない．

5.7 コミュニティ構造の観点での評価結果

proposed 法 × Assortative 属性は，Assortativity が高い属性を用いるため，同じ属性値を持つノードが集中し同一コミュニティに含まれやすい．それゆえ，同一コミュニティ内のノードでつながりあい，平均クラスタ係数をより正確に再現できたと考えられる．この節では，誤推定した場合の推定リンク先ノードが同一コミュニティかどうかを分析することで，その誤推定が同一属性を持つノードに対して許容できるものかどうかを評価する．

ノード u の真のネットワークと推定ネットワークにおける子ノード集合をそれぞれ $F(u) = \{v; (u, v) \in E\}$，$\hat{F}(u) = \{v; (u, v) \in \hat{E}\}$ とする．また，真のネットワークを CNM 法により H 個のコミュニティに分割すると，各ノードは H 個のコミュニティのいずれかに属し，ノード u の属するコミュニティ番号を $c(u) \in \{1, \ldots, H\}$ とする．この時，真のネットワークと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 のリンク相手が属するコミュニティ正解率 (Community Accuracy) として，

$$ca(u) = \frac{1}{|F(u)|} \sum_{h=1}^{H} \min(y_u(h), \hat{y}_u(h)) \quad (4)$$

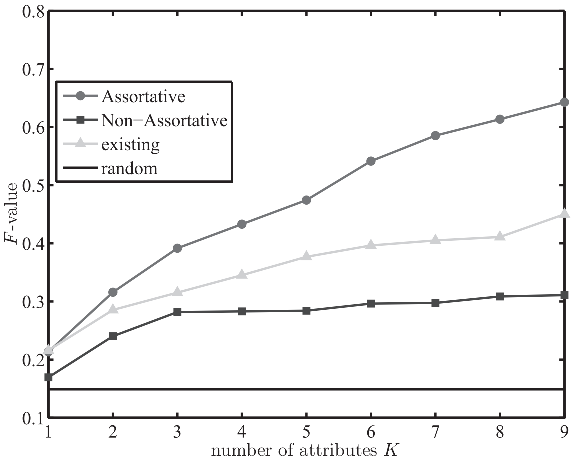
を定義する．ノード u のコミュニティ正解率 $ca(u)$ は，ノード u のリンク相手として真の相手ノードと異なる（F 値では不正解）場合でも，同一コミュニティのノード同士を間違えた場合は許容される指標になっている．式 4 を全ノードで平均した値 $CA = 1/|V| \sum_{u \in V} ca(u)$ を全体のコミュニティ正解率として定義する．

図 3 に，CNM 法 [13] の閾値を変え抽出されたコミュニティ数とコミュニティ正解率の関係を図示する．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 proposed 法は赤●，existing 法は緑▲，random 法は黒線でプロットした．proposed 法と existing 法は，Assortative な属性の数を $K = 9$，各属性のカテゴリ数を $S^{(k)} = 5$ とした．

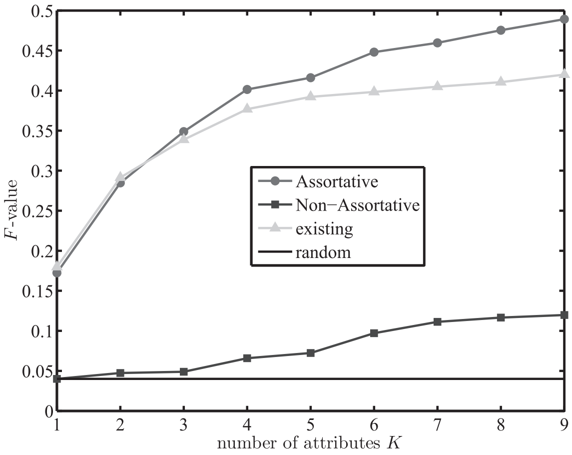
図 3 から，proposed 法は Hierarchical ネットワークでは 8 割以上，Hosei，Nips ネットワークでは 9 割以上の正解率を示している．F 値では 5 割や 6 割程度であるが，推定リンク先が同一コミュニティである場合を許容したこの指標では，高い精度で推定できていることを意味する．また，同じく属性情報を用いる existing 法と比べても，幾分か高い正解率を示しており，random 法と比べても有意に高い正解率を示している．すなわち，proposed 法は Assortative な属性を用いているため，コミュニティを跨ってリンク相手ノードを間違えることは少ないが，コミュニティ内の同値なノードを判別するのが困難であることを意味する．ネットワーク統計量による評価の考察と同様に，コミュニティなどの局所的な構造を正確に推定できることが示唆された．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，コミュニティ分割の粒度に対して頑健な結果が得られた．

6. おわりに

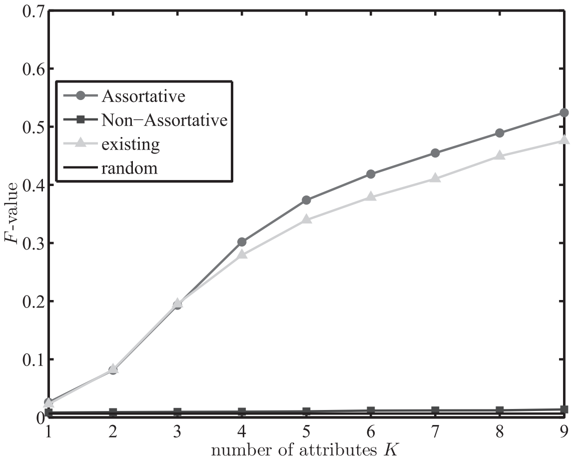
本稿では，社会調査などから得られるエゴセントリック情報（パーソナル属性，リレーショナル属性）のみを用いてネットワーク全体構造を推定する問題を定式化し，EM アルゴリズムにより推定する方法を提案し評価した．複数のネットワークを用いた結果，Assortative な属性を複数用いるほうが，Non-Assortative な属性を用いる場合や，ランダムにリンクを推定する方法，比較に用いた手法よりも高い推定精度を得られることを確認した．特に，コミュニティなどの局所的な構造を正確に推定できることが示唆されたが，random 法と異なり小さな平均ノード間距離を実現することが課題として残った．今後は，提案手法にランダム性を導入することで，random 法のようにスモールワールド性を実現できる推定法へと拡張する予定である．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



(a) Hierarchical ネットワーク

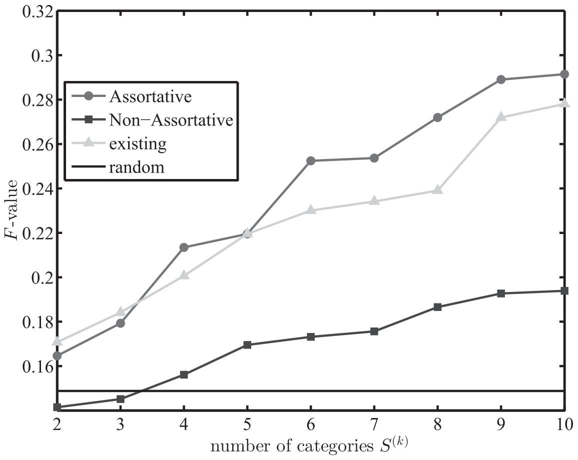


(b) Hosei ネットワーク

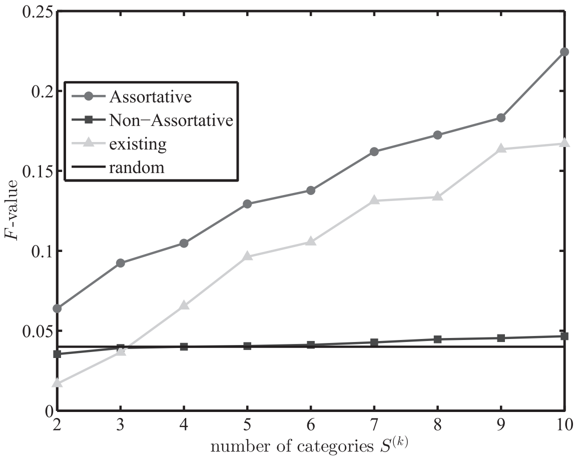


(c) Nips ネットワーク

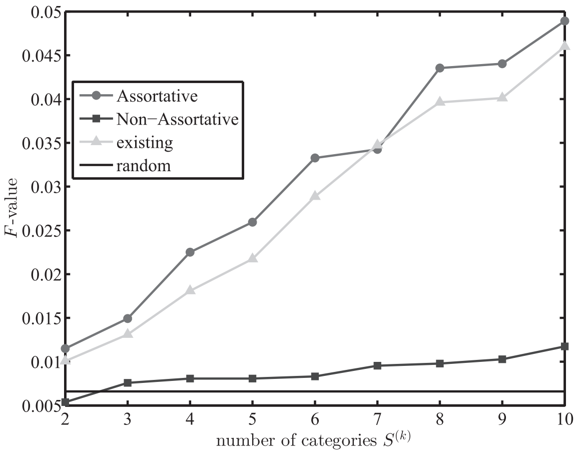
図 1: 属性数の変化と推定精度



(a) Hierarchical ネットワーク



(b) Hosei ネットワーク



(c) Nips ネットワーク

図 2: カテゴリ数の変化と推定精度

表 1: ネットワーク統計量の比較

(a) Hierarchical ネットワーク

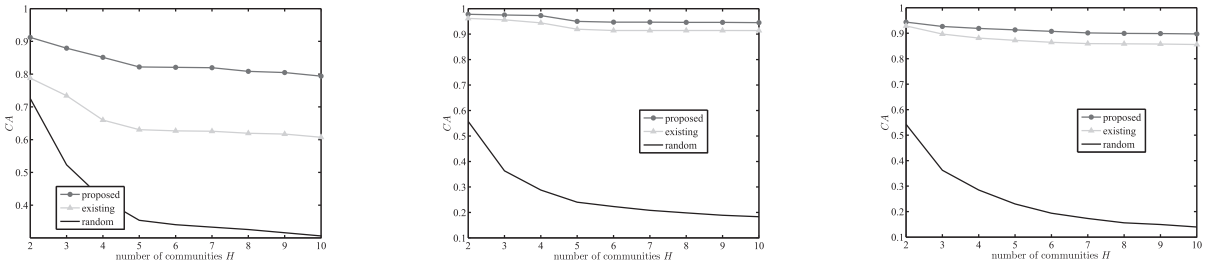
	$\bar{C}$	$C_{\mathrm{S.D.}}$	$\bar{D}$
true	0.84	0.16	2.13
proposed	0.62	0.21	2.83
existing	0.61	0.25	2.96
random	0.06	0.09	2.50

(b) Hosei ネットワーク

	$\bar{C}$	$C_{\mathrm{S.D.}}$	$\bar{D}$
true	0.54	0.41	4.22
proposed	0.34	0.39	14.43
existing	0.21	0.39	20.51
random	0.00	0.03	4.65

(c) Nips ネットワーク

	$\bar{C}$	$C_{\mathrm{S.D.}}$	$\bar{D}$
true	0.58	0.43	6.21
proposed	0.42	0.39	15.92
existing	0.29	0.39	22.09
random	0.00	0.01	5.06



(a) Hierarchical ネットワーク　(b) Hosei ネットワーク　(c) Nips ネットワーク

図 3: コミュニティ正解率

Matrix の偏りが大きい属性を用いた評価も進めていきたい．また，Assortative な属性と Non-Assortative な属性のみを用いて評価したが，これらの混合比を変化させた際の精度についても評価していくつもりである．

[謝辞]

本研究は科学研究費補助金 (No.25・10411) の補助を受けた．

[文献]

[1] Nowell, D. L. and Kleinberg, J.: The link prediction problem for social networks, *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, New York, NY, USA, ACM, pp. 556–559 (2003).

[2] Oyama, S. and Manning, C. D.: Using feature conjunctions across examples for learning pairwise classifiers, *15th European Conference on Machine Learning (ECML2004)*, pp. 322–333 (2004).

[3] Kato, T., Tsuda, K. and Asai, K.: Selective integration of multiple biological data for supervised network inference., *Bioinformatics (Oxford, England)*, Vol. 21, No. 10, pp. 2488–2495 (2005).

[4] Clauset, A., Moore, C. and Newman, M. E. J.: Hierarchical structure and the prediction of missing links in networks, *Nature*, Vol. 453, pp. 98–101 (2008).

[5] Miller, K., Jordan, I. M. and Griffiths, L. T.: Nonparametric latent feature models for link prediction, *Advances in Neural Information Processing Systems 22*, Curran Associates, Inc., pp. 1276–1284 (2009).

[6] Menon, A. K. and Elkan, C.: Link prediction via matrix factorization, *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, ECML PKDD'11, Berlin, Heidelberg, Springer-Verlag, pp. 437–452 (2011).

[7] Palla, K., Ghahramani, Z. and Knowles, D. A.: An infinite latent attribute model for network data, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, New York, NY, USA, ACM, pp. 1607–1614 (2012).

[8] Fushimi, T., Saito, K. and Kazama, K.: Estimating network structure from anonymous ego-centric information, *Proceedings of the 13th Pacific Rim Knowledge Acquisition Workshop (PKAW2014)*, Berlin, Heidelberg, Springer-Verlag, pp. 236–245 (2014).

[9] Newman, M. E. J.: Mixing patterns in networks, *Physical Review E*, Vol. 67, No. 2, pp. 026126+ (2003).

[10] Ravasz, E. and Barabási, A. L.: Hierarchical organization in complex networks, *Physical Review E*, Vol. 67, No. 2, pp. 026112+ (2003).

[11] Even-Dar, E. and Shapira, A.: A note on maximizing the spread of influence in social networks, *Internet and Network Economics*, LNCS, Vol. 4858, Springer Berlin / Heidelberg, pp. 281–286 (2007).

[12] Watts, D. J. and Strogatz, S. H.: Collective dynamics of 'small-world' networks, *Nature*, Vol. 393, No. 6684, pp. 440–442 (1998).

[13] 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).

伏見 卓恭　Takayasu FUSHIMI

静岡県立大学経営情報イノベーション研究科客員研究員．日本学術振興会特別研究員（PD）．2014 年静岡県立大学大学院経営情報イノベーション研究科博士後期課程修了．複雑ネットワークの研究に従事．電子情報通信学会，人工知能学会，情報処理学会，日本データベース学会各正会員．

斉藤 和巳　Kazumi SAITO

静岡県立大学経営情報学部教授．1985 年慶応義塾大学理工学部数理科学科数学専攻卒業，1998 年東京大学博士（工学）．複雑ネットワークの研究に従事．情報処理学会，電子情報通信学会，人工知能学会，日本神経回路学会，日本応用数理学会，日本行動計量学会，日本データベース学会各会員．

風間 一洋　Kazuhiro KAZAMA

和歌山大学システム工学部教授．1988 年京都大学大学院工学研究科精密工学専攻修士課程修了．2005 年京都大学大学院情報学研究科システム科学専攻博士課程修了．Web 情報検索，Web マイニングの研究に従事．人工知能学会，日本ソフトウェア科学会，日本データベース学会，ACM 各会員．