# ＰＳＤ推定の適用範囲拡大と精度向上手法の検討

- Application of an extended PSD estimation method having wide applicability -

グエン　ホン　ハ [*1] Nguyen Hong Ha　鷲尾　隆 [*1] Washio Takashi　宇野 毅明 [*2] Uno Takeaki　エー　ペング　リン [*3] Ee-Peng Lim　桑島 洋 [*4] Kuwajima Hiroshi

[*1]大阪大学産業科学研究所　[*2]国立情報学研究所　[*3]Singapore Management University
[*4](株) マイクロソフト・ディベロップメント

This paper applies PSD estimation method named "MAXPEC" to estimate potential friendships among users in community network. Its performance has been compared with the estimation by SVM. The characteristic of MAXPEC has been clarified through its application to real world data and the comparison with SVM.

## 1. はじめに

近年の社会におけるインタネットの爆発的な普及ともに, ネットワークを介したコミュニケーションに基づくコミュニティ（コミュニティ・ネットワーク）が急速に増加している．これらの中には，メーリングリスト，ネットニュース，チャットなど様々なコミュニケーション形態に基づくものがあり，これらのコミュニティ・ネットワークの活動を記録したデータは，内容が様々でかつ大規模であるが，多くの場合，参加者間のコミュニケーションパスやリンク，知人，友人関係などの構造に関する情報を含んでいる．

コミュニティ・ネットワークの構造は，そのコミュニティに参加する人々の関係を反映しており，逆にコミュニティ・ネットワークの構造を把握し，かつそれを適切に誘導，刺激することができれば，そのコミュニティ活動の活性化や発展を促すことができると期待される．例えば，オンライン・ショッピングのコミュニティ・ネットワークの記録データを解析して，各参加者のショッピング傾向と参加者間の友人関係を把握した上で，現状は友人関係ではないが，潜在的に友人となる可能性の高い参加者同士の紹介を，適切に支援することを考える．これによって，オンライン・ショッピングサイトで，人間関係や興味を共有する参加者間での製品に関するコミュニケーションが促進されば，購買活動が盛んになると期待される．また，同時に当該オンライン・ショッピングサイトのコミュニティ・サイトとしての発展も一層期待できる．しかし，このようなことができるためには，先に述べたように今後友人となる可能性の高い参加者ペアを推定することが求められる．

一方，コミュニティ・ネットワークにおける参加者間の友人関係は，全参加者を頂点集合 $V$ と考え，全参加者の友人関係を辺集合 $E$ と考えることによって, グラフ $G(V, E)$ で表現できる．従って，上述した参加者間の現状の友人関係から，今後潜在的に友人となる可能性の高い参加者ペアを推定する問題は，グラフ G(V,E) において，現状の辺集合で表わされる頂点間の関係から，今後，新たに辺が付与される可能性の高い頂点ペアを推定する問題に置き換えることができる．このようなグラフ上で存在する可能性の高い辺を推定する問題を,「リンク予測」という．[Getoor 05] によると，リンク予測の方法には大別して 2 種類ある．一方はリンク予測にネットワークの構造的性質を用いる方法であり，他方はリンク予測に属性情報を用いる方法である．

以上のリンク予測研究と並行して，我々は半正定性（PSD; Positive Semi-Definiteness）を満たす行列について，その数学的許容性に基づいて行列の既存要素値から他の要素値を推定する手法（PSD 推定手法）の研究に取り組んで来た [Kido 08]. グラフは行列で表現することができる．グラフを表す行列には，隣接行列や Signless Laplacian 行列 (SL 行列) などがある．グラフの SL 行列は半正定行列（PSD 行列）[Cvetkovic 07] であるので，PSD 推定手法を適用して, グラフの SL 行列の既存要素値から他の要素値を推定すること，すなわち，グラフの既存の辺接続構造から他に存在する可能性の高い辺を推定することができる．それによって，参加者間に存在する可能性の高い友人関係も推定できる．この方法は，上述した区別によれば，ネットワークの構造的性質を用いるものと考えられる．

この論文では，我々がこれまで開発した PSD 推定手法の 1 つである MAXimal Psd Estimation based on clique enumeration and incomplete Cholesky decomposition (MAXPEC)[Nguyen 08] を適用して, SL 行列を推定することによって，コミュニティ・ネットワークの参加者間に存在する可能性の高い友人関係を推定することを試みる．さらに，既存の代表的な機械学習方法である Support Vector Machine（SVM）[Burges 98] を適用した方法と，その性能を比較する．

## 2. 解析対象データ

### 2.1 対象とするコミュニティ・ネットワーク

本研究では，オンライン・ショッピングサイトの www.epinions.com のデータに MAXPEC と SVM を適用して，参加者の潜在的友人関係を推定比較する．www.epinions.com サイトでは，自動車，本，音楽，パソコン，電子製品など様々な商品が販売されている．このオンライン・ショッピングサイトは会員制を採用しており，会員である参加者同士が商品を紹介し合うコミュニティ・サイトとなっている．本研究の潜在的友人関係推定では，このコミュニティ・サイトにおいて，どの会員参加者が他のどの会員参加者の友人（知り合い）であるかを表すデータを用いる．

### 2.2 対象データ

対象とするデータは，このコミュニティ・ネットワークにおける参加者間の友人関係を表現する 1 枚の大規模なグラ

連絡先: 氏名: Nguyen Hong Ha
所属: 大阪大学 産業研究所 知能推論研究分野
住所: 567-0047 大阪府茨木市美穂ヶ丘 8-1
Mail: hongha@ar.sanken.osaka-u.ac.jp

フ $G(V,E)$ である．グラフデータ $G(V,E)$ の頂点集合 $V = \{v_1, v_2, \ldots, v_N\}$ の各要素はコミュニティ・ネットワークにおける各参加者であり，辺集合 $E$ の各要素は参加者同士の個別の友人関係を表している．即ち，参加者 $v_i$ と $v_j$ が友人関係にあれば，それを表す辺 $(v_i, v_j)$ が E に含まれる．E は $E \subseteq V \times V$ を満たす．参加者 $v_i$ から見て参加者 $v_j$ が友人であれば，参加者 $v_j$ から見ても参加者 $v_i$ は友人である．従って，グラフ G は無向グラフである．このデータは www.epinions.com の参加者 16230 人の友人関係を記録しており，その間に 165592 の友人関係が存在する．即ち，グラフデータ $G(V, E)$ には，16230 の頂点があり，165592 の辺がある．1 つの頂点は平均で他の 10 頂点と結ばれ，言い換えれば，各頂点の平均次数は 10 である．

今回は，MAXPEC と SVM を適用して，コミュニティ・ネットワークににおける各参加者間の潜在的友人関係を推定する．MAXPEC や SVM の実装では，計算時間制約により最大でも 120 × 120 程度まで PSD 行列推定しか推定できない．また，MAXPEC は推定対象とする PSD 行列中で推定に使用する要素の割合が高ければ，推定精度は高くなる．すなわち，グラフの頂点数に対して相対的に辺数が多い dense なグラフを推定対象とする方が，推定精度が向上する．今回は，対象とするコミュニティ・ネットワークの中でも，ような性質を有する部分のみを切りだして，潜在的友人関係を推定することとする．そのため，グラフ G の全頂点をその次数が高い順番でソートして，上位 90 頂点を選んで，$G(V,E)$ から部分グラフ G'(V', E') を切り出した．ここで，グラフ G'(V',E') の頂点集合 $V' = \{v_1, v_2, \ldots, v_{90}\} \subset V$, 辺集合 $E' \subseteq V' \times V'$ である．この部分グラフ $G'(V', E')$ には，90 頂点および 1678 個の辺が含まれ，その平均次数は 18 であった．すなわち，対象コミュニティ・ネットワークにおいて選択した部分コミュニティ・ネットワークは，最も多くの友人を持つ 90 人の参加者間の友人関係を表し，一人の参加者は平均で 18 人の友人を持つ．実験では MAXPEC と SVM をこの部分グラフ G'(V',E') に適用し，潜在的友人関係の推定を行う．

## 3. 解析手法

### 3.1 MAXPEC の概要

MAXPEC は対象とする PSD 行列 X の一部の要素集合 $X_s \subset X$ とその PSD 行列 X の数学的な性質に基づいて，それ以外の $\bar{X}_s = X - X_s$ の要素値とその存在許容区間を推定する手法である．MAXPEC を適用する際には，対象 PSD 行列 X の中で推定に使用する各要素 $x_{ij} \in X_s$ に対応する要素 $r_{ij}$ を 1 とし，それ以外の推定対象要素 $x_{ij} \in \bar{X}_s$ に対応する要素 $r_{ij}$ を 0 として得られる隣接行列 R が表すグラフが，Joined Complete Split Graph with Multiple Independent Vertices (Joined CSGMIV) になるように，推定に使用する要素を選択しなければならない [Nguyen 08]．Joined CSGMIV とは，図 1 に示す Compete Split Graph with Multiple Independent Vertices (CSGMIV) に分解することが可能なグラフである．個々の CSGMIV はクリークと独立頂点集合から成り，クリークの頂点と独立集合の全ての頂点が接続されたグラフである．Joined CSGMIV は複数の CSGMIV が組み合わされたグラフである．MAXPEC のアルゴリズムは主に 2 つのステップから成る．ステップ 1 は，New Algorithms for Enumerating All Maximal Cliques (MACE) [Makino 08] を，上述の Joined CSGMIV を表す隣接行列 R に適用し，各 CSGMIV を表す各主小行列 $R_s^i$ $(i = 1, ..., I)$ に分解する．各 $R_s^i$ 中で

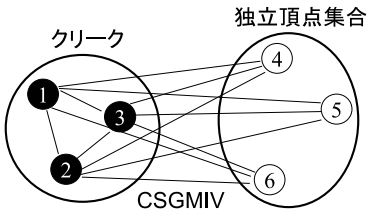


図 1: CSGMIV

1 である要素 $r_{ij}$ に対応する X 中の要素 $x_{ij}$ は推定に用いる要素集合 $X_s$ に属し，0 である要素 $r_{ij}$ に対応する X 中の要素 $x_{ij}$ は推定対象要素の集合 $\bar{X}_s$ に属する．ステップ 2 は分解された各主小行列 $R_s^i$ に対応する元の対象 PSD 行列部分 $X_s^i$ に Large PSD Matrix Estimation from Partical Elements (PERCH)[Kuwajima 07] を適用して，各々の主小行列 $X_s^i$ 内の推定対象要素集合 $\bar{X}_s$ に属する各要素値とその存在許容区間を推定する．そして，すべての主小行列からこのようにして推定された $\bar{X}_s$ 中の各々の推定対象要素 $x_{ij}$ の存在許容区間の重なり区間を，その推定区間とし，その中央値を推定値とする．推定対象とする PSD 行列 X のサイズを N とすれば，ステップ 1 の計算量は O $(N^{2.376})$ である．ステップ 2 の 1 回の PERCH の計算量は O $(k^2 n)$，そしてその結果を用いる要素推定の計算量は O $(kn^2)$ であり，これらの合計は O $(k^2 n + kn^2)$ である．ここで，n はステップ 1 で分解された各主小行列 $R_s^i$ が表す CSGMIV の平均頂点数であり，k はそれらの CSGMIV のクリークが含む平均頂点数である．ステップ 2 ではステップ 1 で分解された全主小行列 $R_s^i$ $(i = 1, ..., I)$ に PERCH を適用するため，ステップ 2 の計算量は O $(I(k^2 n + kn^2))$ となる．従って，MAXPEC 全体の計算量は O $(N^{2.376} + I(k^2 n + kn^2))$ である．一般に I は対象 PSD 行列のサイズ N に対して指数的に増加するため，MAXPEC で取扱可能な対象 PSD 行列のサイズには限界がある．

### 3.2 潜在的友人関係の推定

MAXPEC をグラフ G'(V',E') に適用し，上記の 90 人のコミュニティ・ネットワークの参加者間の潜在的友人関係を推定する．はじめにグラフ G'(V',E') のデータを，前節で説明した MAXPEC が適用可能な PSD 行列 X に変換する必要がある．ここでは，X をグラフ G'(V',E') を表す PSD 行列の一種である正規化 Signless Laplacian 行列とする．グラフ G'(V',E') を表す Signless Laplacian 行列 SL は，その定義より，$A$ と $D$ をそれぞれのグラフ G'(V',E') の隣接行列と次数行列とすれば，$SL = A + D$ と表される．A の各要素は

$$a_{ij} = \begin{cases} 1 & (v_i, v_j) \in E';\ \ v_i, v_j \in V' \\ 0 & (v_i, v_j) \notin E';\ \ v_i, v_j \in V' \end{cases}$$

であり，D の各要素は

$$d_{ij} = \begin{cases} 0 & i \neq j \\ \sum_{k=1}^{90} a_{ik} & i = j \end{cases}$$

である．次に SL の各対角成分を 1 に正規化し，それを X とする．SL の要素を $sl_{ij}$ とすれば，X の要素 $x_{ij}$ は $x_{ij} = \frac{sl_{ij}}{\sqrt{sl_{ii}}\sqrt{sl_{jj}}}$ で計算される．

本研究では，上記 G'(V',E') の実データを表す X から，90 人間の友人関係の一部を $X_s$ として取り出して，それから残りの友人関係 $\bar{X}_s$ を MAXPEC で推定することによって，その 90 人の間の潜在的友人関係を推定する．実データから取り出

した一部の友人関係 $X_s$ の構造に基づいて，残りの友人関係 $\bar{X}_s$ の構造を推定すれば，推定された参加者間の関係 $\bar{X}_s$ は，推定に用いた元の参加者間の関係 $X_s$ の性質を反映したものとなる．従って，実データ X において実際には友人ではない参加者同士に $\bar{X}_s$ の推定結果では友人関係が推定されれば，コミュニティ・ネットワークの他部分の友人関係構造 $X_s$ から見れば，友人であるべき参加者同士がまだ友人ではないことになり，それを潜在的友人関係と見なすことができる．

前節で述べたように，MAXPEC を推定に用いるためには，推定に用いる要素集合 $X_s$ の各要素 $x_{ij}$ に対応する要素 $r_{ij}$ が 1 で，推定対象の要素集合 $\bar{X}_s$ の各要素 $x_{ij}$ に対応する要素 $r_{ij}$ が 0 である隣接行列 R が，Joined CSGMIV を表していなければならない．更に，証明は省略するが，R が Joined CSGMIV を表す必要十分条件は”$\forall r_{ij}=0, \exists k \in \{1,...,N\} \quad s.t \quad r_{ik}=r_{jk}=1$” である．そこで，この条件を基づいて，$\frac{|X_s|}{|X|}=\beta$ ($|X_s|$ と $|X|$ はそれぞれ $X_s$ と $X$ の要素数である）になるように $X_s$ を乱数で選ぶ．実データ X のちょうど半分の $X_s$ を推定に用いる β=0.5 の場合，詳細は省略するが X に対応する R に含まれる CSGMIV の数 I は最小になる．従って，前節の計算量評価から明らかなように MAXPEC の計算時間は最も少なくなる．一方，$\beta$ が 1 に近くなる，すなわち，推定に使用する要素数 $|X_s|$ が大きくなると，R を表す Joined CSGMIV はより dense になり，R が含む CSGMIV の数が多くなって，より多くの主小行列 $R_s^i$ から推定対象要素 $x_{ij}$ の存在許容区間を得ることができる．これにより，これら存在許容区間の重なり区間である推定対象要素 $x_{ij}$ の存在許容区間は狭くなり，推定精度は高くなる．以上のような推定精度と計算時間のバランスを考え，本計算では $\beta=0.7$ を用いる．

以上のようにして X から選択した $X_s$ から MAXPEC で $\bar{X}_s$ を推定し，これによって得られた行列 X の推定行列を X’ とする．ただし，X’ の各要素 $x'_{ij}$ の値は $0 \leq x'_{ij} \leq 1$ であるので，各要素が 0 と 1 のみで構成される隣接行列 A とは異なり，明確に各参加者間の友人関係を読み取ることはできない．従って，X’ をそれに対応する隣接行列 A’ に変換する必要がある．X’ から A’ を得るためには，次数行列 D’ の計算が必要である．しかし，MAXPEC で推定された X’ は対角成分が正規化されており，それから，次数行列 D’ は計算できない．そこで，近似的に D’ を推定することを考える．グラフ G’(V’,E’) において，推定に用いる $X_s$ に対応する部分で参加者同士が友人関係であるペアの割合を $\alpha$ とすれば，推定対象とする $\bar{X}_s$ においても関係 $(v_i, v_j)$ が友人関係である確率は $\alpha$ と考えられる．これに基づき，近似隣接行列 $\hat{A}$ と近似次数行列 $\hat{D}$ を以下のように計算する．$\hat{A}$ の各要素を

$$\hat{a}_{ij} = \begin{cases} a_{ij} & x_{ij} \in X_s \\ \alpha & x_{ij} \in \bar{X}_s \end{cases}$$

とし，$\hat{D}$ の各要素を

$$\hat{d}_{ij} = \begin{cases} 0 & i \neq j \\ \sum_{k=1}^{90} a'_{ik} & i = j \end{cases}$$

とする．この近似を，正規化 Signless Laplacian 行列の定義に基づく以下の関係式に代入し，X’ から SL の推定行列 SL’ を得る．

$sl'_{ij} = x_{ij}(\sqrt{\hat{sl}_{ii}}\sqrt{\hat{sl}_{jj}}) = x_{ij}(\sqrt{\hat{a}_{ii}+\hat{d}_{ii}}\sqrt{\hat{a}_{jj}+\hat{d}_{jj}})$
$(i,j=1,...,N)$

更に，推定対象である隣接行列 A’ の近似を，以下により求める．

$A' = SL' - \hat{D}$

ただし，この近似により得られた A’ の各要素 $a'_{ij}$ は，必ずしも 0 や 1 にはならない．しかし，各 $a'_{ij}$ は推定対象とする関係 $(v_i, v_j)$ が友人である可能性を表すと見なすことができるので，$a'_{ij}$ が友人関係の平均割合 α 以上 ($a'_{ij} \geq \alpha$) であれば 1，すなわち，$(v_i, v_j)$ が友人関係であると推定し，一方，$a'_{ij}$ が友人関係の平均割合 α 未満 ($a'_{ij} < \alpha$) であれば 0，すなわち，$(v_i, v_j)$ が友人関係ではないと推定する．隣接行列 A は 90 人間のコミュニティ・ネットワークの現状の友人関係を表すので，上記で推定された A’ と A を比較することによって，90 人間のコミュニティ・ネットワークの潜在的友人関係を推定できる．例えば，$a_{ij}=0, a'_{ij}=1$ の場合，$(v_i, v_j)$ は現状友人関係ではないが，コミュニティ・ネットワークの友人関係の部分構造 $X_s$ から見れば，$(v_i, v_j)$ は友人関係であるべきと判断される．すなわち，この $(v_i, v_j)$ は今後潜在的に友人になる可能性が高いと判断される．このようにして，実際のコミュニティ・ネットワークの構造と MAXPEC の推定構造の比較を通じて，90 人のコミュニティ・ネットワークの潜在的友人関係を推定できる．

## 4. SVM を用いた潜在的友人関係の推定

本章では，MAXPEC による推定との比較対象とする SVM による推定方法を説明する．SVM は，現在知られている多くの手法の中でも，最も分類性能が優れた学習モデルの 1 つであるといわれている [Burges 98]．SVM は，線形入力素子を利用して 2 クラス分類器を構成する手法である．各訓練データ点との距離が最大となる分離平面（超平面）を求めて，予測データを分類する．

はじめに，グラフ G’(V’,E’) のデータを，SVM が適用可能なデータに変換する必要がある．前節で述べたように，グラフ G’(V’,E’) の隣接行列 A は 90 人間の友人関係を表す．そして，A の行ベクトル $a_i$ は参加者 $v_i$ の現状の全ての友人を表す．更に，参加者 $v_i$ と $v_j$ が友人になる可能性は，参加者 $v_i$ と $v_j$ の現状の全ての友人の情報と関係があると考えられる．例えば，参加者 $v_i$ と $v_j$ は共通の友人が多ければ，友人になる可能性が高いと考えられる．そこで，A の 2 つの行ベクトル $a_i$ と $a_j$ $(i<j)$ を縦につなげて，$v_i$ と $v_j$ の現状の全ての友人の情報を持つ 180 次元ベクトル $p_{i,j}=(a_i, a_j)$ を作成する．$p_{i,j}$ は $v_i$ と $v_j$ の友人になる可能性を表すベクトルと考えられる．さらに，$v_i$ と $v_j$ の友人関係を表す $a_{ij}$ を $p_{i,j}$ のクラスとする．これにより，グラフ G’(V’,E’) をベクトル集合 $P=\{(p_{i,j}, a_{ij})\}$ で表し，これを SVM に適用する．

ここでは，グラフ G’(V’,E’) の実データを表す隣接行列 A から，MAXPEC で推定に使用した要素集合 $A_s \subset A$ を SVM による推定に使用し，それ以外の $\bar{A}_s = A - A_s$ を推定する．しかし，$A_s$ は隣接行列 A の一部に過ぎないため，SVM が適用可能な上記ベクトル集合 P を構成することができない．そこで，節 3.2 で述べたように，$A_s$ から友人関係の存在割合 α を計算し，それを用いて近似隣接行列 $\hat{A}$ を計算する．$\hat{A}$ の i 番目の行ベクトルを $\hat{a}_i$ とすれば，この $\hat{A}$ から $\hat{p}_{i,j}=(\hat{a}_i, \hat{a}_j)$ を作成し，$\hat{a}_{ij}$ =0 または 1 であれば，$\hat{a}_{ij}$ を $\hat{p}_{i,j}$ のクラスとする．一方，$\hat{a}_{ij}=\alpha$ であれば $\hat{p}_{i,j}$ のクラスは”未確定”である．クラスが確定した $(\hat{p}_{i,j}, a_{ij})$ の集合を $\hat{P}_s$ として，クラスが”未確定”の $\hat{p}_{i,j}$ の 集合を $\hat{Q}_s$ とする．

SVM は，$\hat{P}_s$ を訓練データとして，$\hat{Q}_s$ の各ベクトル $p_{i,j}$ のクラス $\hat{a}_{i,j}$ を予測する．これにより，全ての $\hat{a}_{ij} \in \bar{A}_s$ の値を推定できる．MAXPEC による推定と同様に，$a_{i,j}(\in A)=0$ であり，かつ $\hat{a}_{i,j}(\in \bar{A}_s)=1$ であれば，推定に用いた $A_s$ の構

造から見て，参加者 $v_i$ と $v_j$ は潜在的に友人となる可能性が高いと判断される．従って，SVM によっても，90 人間のコミュニティ・ネットワークの潜在的友人関係を推定できる．

$\hat{p}_{i,j}$ の次元 d は 2N であり，訓練データ $\hat{P}_s$ のベクトル数 $\ell$ は $\beta N(N-1)/2$ である．一般に，SVM の計算量は O $(d\ell^2)$ [Burges 98] であるので，本推定の計算量は O $(d\ell^2)$ = O $(2N\times(\beta N(N-1)/2)^2)$ = O $(\frac{\beta^2 N^5}{2})$ となる．

## 5. 性能評価

### 5.1 推定性能指標

参加者ペアが友人関係である場合を 1, そうでない場合を 0 で表わした時，実際の G'(V',E') において友人関係が $f \in \{0,1\}$ で，推定された友人関係が $g \in \{0,1\}$ である参加者ペアの数を $m(f,g)$ とする．推定性能評価の統計的安定性を確保するため，10 回に亘るランダムな $X_s$ (従って $A_s$) の選択による推定結果の平均を $m(f,g)$ に用いる．これにより，以下の 5 つの推定性能指標を定義する．

正例（友人である例）の正答率: $\frac{m(1,1)}{m(1,1)+m(1,0)}$

負例（友人でない例）の正答率: $\frac{m(0,0)}{m(0,1)+m(0,0)}$

正例（友人である例）の再現率: $\frac{m(1,1)}{m(1,1)+m(0,1)}$

負例（友人でない例）の再現率: $\frac{m(0,0)}{m(1,0)+m(0,0)}$

MCC：

$$\frac{m(1,1)m(0,0)-m(1,0)m(0,1)}{\sqrt{(m(1,1)+m(1,0))(m(0,1)+m(0,0))(m(1,1)+m(0,1))(m(1,0)+m(0,0))}}$$

各正答率や再現率が 1 に近ければ，各条件における推定の信頼性は高いと言える．一方，MCC は Matthews Correlation Coefficient であり，全体の推定品質が高ければ MCC の絶対値は 1 に近づき，低ければ 0 に近づくことが知られている [Matthews 75].

### 5.2 性能比較

表 2，表 3 は，それぞれ MAXPEC 及び SVM を適用した場合の $m(f,g)$ の結果及び上記 5 つの性能指標を示す．各々，友人関係である場合を 1 で表し，友人関係でない場合を 0 で表す．左列の 1 と 0 は現状の関係を表し，上行の 1 と 0 は推定された関係を表す．

表 2 と表 3 より，MAXPEC の MCC は 0.404063 であり，SVM の MCC は 0.485061 である．従って，SVM による推定の品質の方が MAXPEC より良い．表 2 と表 3 を見ると，MAXPEC の正例の正答率は SVM より小さいが，負例の正答率は SVM より高い．そして，MAXPEC の正例の再現率と負例の再現率の両方とも SVM より小さい．全般的には，SVM の推定の信頼性は MAXPEC より高いといえる．

負例の正答率は，現状は友人関係にはなく，かつ友人関係がないと推定される割合である．表 2 と表 3 より，SVM の負例の正答率は MAXPEC より小さいので，SVM で推定する方が，潜在的友人関係を見つけやすいことが判る．

上記の結果より，MAXPEC より SVM の方が推定の信頼性もより高く，潜在的友人関係もより見つけやすいことが判る．SVM では，参加者ペアの全友人関係とそのペアが友人であるか否かの情報から分類学習が行われる．従って，コミュニティ・ネットワークの広範囲の構造から局所的な構造を比較的高い精度で予測することができる．これに対して，MAXPEC は，グラフのある部分構造から，そのグラフを表すある種の行列が PSD となる他の部分構造の数学的許容範囲を推定する．従って，直接にコミュニティ・ネットワークの構造同士の関係を学習するわけではない．この違いにより，潜在的友人関係の推定という目的には，SVM の方が高い性能を示したと考えられる．逆に，グラフの部分構造から，他の部分構造の許容範囲を推定する目的には MAXPEC が向いていると言える．

| 本当＼推定 | 1 | 0 | 正答率 |
|---|---|---|---|
| 1 | m(1,1) =254.9 | m(1,0) =193.5 | 0.568466 |
| 0 | m(0,1) =133.7 | m(0,0) =618.9 | 0.822349 |
| 再現率 | 0.655944 | 0.761817 | |
| MCC | | 0.404063 | |

表 2：MAXPEC を適用した場合の実験結果

| 本当＼推定 | 1 | 0 | 正答率 |
|---|---|---|---|
| 1 | m(1, 1) =315.8 | m(1,0) =132.6 | 0.704282 |
| 0 | m(0,1) =160.8 | m(0,0) =591.8 | 0.786341 |
| 再現率 | 0.66261 | 0.816952 | |
| MCC | | 0.485061 | |

表 3：SVM を適用した場合の実験結果

## 6. まとめ

コミュニティ・ネットワークにおける潜在的友人関係の推定という目的には，PSD などの数学的許容制約を用いる推定方法よりも，ネットワーク構造を学習する方法の方が向いていると考えられる．今後は，数学的制約を加えることで推定精度の向上がもたらされる学習方法の確立の可能性などについて検討していきたい．

## 参考文献


[Getoor 05] Getoor.L & Diehi, C. P.: Link Mining: A survey, SIGKDD Explorations, Vol.7, No.2,pp3-12(2005).

[Kido 08] K.Kido, H.Kuwajima & T. Washio: A Range Query Approach for High Dimensional Euclidean Space Based on EDM Estimation, SIAM International Conference on Data Mining, SDM 2008, 387-398(2008).

[Cvetkovic 07] D. Cvetkovic, P. Rowlinson & S. K. Simic: Liniear Algebra and its Applications, Volume 423, Issue 1, 155-171(2007).

[Burges 98] C. J. C. Burges: A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 2, 121-167(1998).

[Nguyen 08] H. H. Nguyen, T. Washio, T. Uno & H. Kuwajima: PSD 推定の適用範囲拡大と精度向上に関する研究, The 22nd Annual Conference of the Japanese Society for Artificial Intelligence, 45-46(2008).

[Makino 08] K.Makino & T.Uno: New Algorithms for Enumerating All Maximal Cliques, proc. of SWAT2004. Scandinavia Workshop on Algorithm Theory, LNCS 3111, 260-272 (2007)

[Kuwajima 07] H.Kuwajima & T.Washio: Large PSD Matrix Estimation from Partical Elements, Workingnotes of Seventh IEEE International COnference on Data Mining - Workshop, ICDMW.2007.24, 337-342(2007).

[Matthews 75] B. M. Matthews: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 1975, 405, 442-451,(1985).