情報処理学会論文誌　数理モデル化と応用　Vol.5　No.1　65–73 (Mar. 2012)

# ネットワークのノード情報を考慮した正則化モジュラリティ固有空間法

吉田 哲也[1,a)]

受付日 2011年8月19日，再受付日 2011年9月22日，
採録日 2011年10月20日

**概要**：ネットワークに存在する少数のリンクしか観測できない状況下でのコミュニティ発見に対し，本稿ではノード情報を考慮した正則化モジュラリティ固有空間法を提案する．従来の研究ではノード間の接続関係に基づくアプローチが多かったが，ネットワークの接続関係は時間とともに変遷する可能性があり，また部分的にしか観測できない場合もある．接続関係とあわせてノードに関する情報が得られる場合には，前者を補うために後者を副次的な情報として活用してコミュニティ発見を行うことが考えられる．両者を統一的に扱うために本稿ではノード情報に基づくグラフ構造を構築し，コミュニティの評価指標であるモジュラリティを最大化する際，構築したグラフを正則化項として活用してコミュニティを発見する手法を提案する．提案法を共著関係ネットワークに適用し，他手法との比較を通じて提案法の有効性を示す．

**キーワード**：モジュラリティ，ノード情報，固有空間法

# Regularized Modularity Eigenmap for Community Discovery with Side Information

Tetsuya Yoshida[1,a)]

Received: August 19, 2011, Revised: September 22, 2011,
Accepted: October 20, 2011

***Abstract:*** We propose a method for community discovery with side information based on a regularized modularity eigenmap. Even when the connectivity relation in a network is only partially available, if other information about the network is available, it can be exploited as auxiliary or side information for community discovery. The proposed approach constructs a graph structure based on the side information so that both link information and side information can be uniformly dealt with in terms of graph representation. In addition, based the modularity matrix of a network, an objective function with a regularization term for the side information is proposed in this paper. Finally, a partitioning based clustering of the constructed eigenmap is conducted for community discovery. Extensive experiments are conducted over social network datasets and comparison with several state-of-the-art methods is reported. The results are encouraging to exploit side information for community discovery.

***Keywords:*** modularity, node profile, eigenmap

## 1. はじめに

近年のネットワーク環境の進展にともない，Facebook や Twitter などのソーシャルネットワークサービスが新しいコミュニケーションメディアとして広く用いられ，ネットワークを介して様々な情報が広く流通されるようになっている [12]．ネットワークからのコミュニティ発見はこれまで主に社会科学などの分野で研究されてきたが，近年の計算資源の発展と普及にともない，計算機科学の立場からの研究も活発に行われている．たとえばネットワークの構造的な性質の探求などが活発に行われている [18], [19]．

[1] 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060–0814, Japan
[a)] yoshida@meme.hokudai.ac.jp

© 2012 Information Processing Society of Japan

ネットワークからのコミュニティ発見とはノード（頂点）どうしのリンク（辺）が多く密な部分ネットワークをコミュニティとして同定する問題であり，様々な研究が行われてきた [13], [14], [16], [17]．従来の研究ではノード間の接続関係に基づくアプローチが多かったが，ノード間のリンクは時間とともに変遷する可能性があり，部分的にしか観測できない場合もある．ネットワークに存在する少数のリンクしか観測できない場合にコミュニティを同定することは困難な問題である．

ネットワークに存在する少数のリンクしか観測できない状況においても，接続情報とあわせてノード（頂点）に関する情報が得られる場合には，前者を補うために後者を副次的な情報として活用してコミュニティ発見を行うことが考えられる．本稿では，観測できるネットワークの接続情報が少ない状況におけるコミュニティ発見を半教師あり学習の観点からとらえ [1]，リンク情報とノード情報を統一的に扱うためにノード情報に基づくグラフ構造を構築し，構築したグラフを正則化項として活用してコミュニティ発見を行う手法を提案する．

実データから構築した 2 つの共著関係ネットワークに提案法を適用して評価し，他手法との比較を行い，観測できるリンク数が少ない状況下での提案法の有効性を確認した．ノード情報に基づき構築したグラフと観測される接続情報を混合モデルとして活用することは困難であったが，ネットワークのリンク情報に対する正則化項として活用することにより，少ないリンク情報からでもネットワークにおける密な部分グラフをコミュニティとして同定できることを確認した．

2 章で関連研究を紹介し，3 章で提案法の詳細を説明する．4 章で他手法との評価実験を報告し，提案法の有効性を議論する．5 章でまとめと今後の展望を述べる．

## 2. ネットワークからのコミュニティ発見

### 2.1 準備

本稿では，行列は太字の大文字，ベクトルは太字のイタリック小文字で表記し，$\mathbf{A}_{ij}$ で行列 $\mathbf{A}$ の第 $ij$ 要素を表す．tr は行列のトレースを表し，$\mathbf{A}$ の転置を $\mathbf{A}^T$ で表す．要素がすべて 1 である $n$ 次元ベクトルを $\mathbf{1}_n$ と表記する．

ネットワーク（グラフ）中のノード（頂点）の数を $n$ とし，リンク（辺）の数を $m$ とする．ノード数が $n$ であるネットワークにおいて，ノード間のリンクに基づいてネットワークの接続関係を表す表現行列を $\mathbf{A}$ と表記する．ネットワークの表現行列 $\mathbf{A}$ に対して $\boldsymbol{k} = \mathbf{A}\mathbf{1}_n$ を次数ベクトルと呼び，要素 $k_i$ は $i$ 番目のノードの次数を表す．本稿ではネットワークのリンクが無向の場合を扱い，自己ループはないものとする．

### 2.2 コミュニティ発見指標

ネットワークからのコミュニティ発見とは，ネットワークのノード集合をもれがなく重複のない部分集合に分割することであり，それぞれの部分集合がコミュニティ（クラスタとも呼ぶ）に対応する．ネットワークからのコミュニティ発見ではコミュニティにおけるノードどうしのリンクが多く密な部分ネットワークとなっていることが望ましい．これまでコミュニティ発見に対して様々な指標が提案されてきたが [10]，モジュラリティと呼ばれる指標 [14] がコミュニティ発見においてしばしば用いられている．ネットワークの分割 $\mathcal{P}$ に対し，モジュラリティはその表現行列 $\mathbf{A}$ に基づいて下記で定義される．

$$
\begin{aligned}
Q &= \frac{1}{2m}\sum_{C\in\mathcal{P}}\sum_{i,j\in C}(\mathbf{A}-\mathbf{P})_{ij} \\
&= \frac{1}{2m}\sum_{C\in\mathcal{P}}\sum_{i,j\in C}\left(\mathbf{A}_{ij}-\frac{k_i k_j}{2m}\right) \qquad (1)
\end{aligned}
$$

分割 $\mathcal{P}$ の要素である $C$ はコミュニティに対応し，$i$，$j$ はコミュニティ $C$ に属するノードを表す．行列 $\mathbf{P}$ はランダムな分割に対応する null モデルを表現し，式 (1) のように $\mathbf{P}_{ij} = k_i k_j/2m$ とすることが多い．モジュラリティは分割 $\mathcal{P}$ と null モデルとの差を表現し，$2m$ は正規化係数の役割を果たす．式 (1) の値が大きいほどネットワークの分割が良いとされる．

式 (1) のモジュラリティは，ネットワークの表現行列をノード間のリンク数に対応する整数値から類似度などのリンクの重みに対応する非負の実数値に拡張した場合でも同様に定義される．その場合，式 (1) での正規化係数の役割を果たす $2m$ は重みの総和（$\sum_{i,j}\mathbf{A}_{ij}$）となる．

ノード集合を分割することはノードのクラスタリングを行うことに対応するが，ラベルありデータに基づく評価と異なり，コミュニティ発見では正解に対応するラベル（ノードの割当てに対する正解）がないことが多い．このため，上記のモジュラリティはあくまでネットワークの構造に基づいた評価指標にすぎず，モジュラリティが高い分割における部分集合が本当にコミュニティと呼ばれるものに対応するかは議論が分かれる[*1]．

### 2.3 関連研究

ネットワークからのコミュニティ発見に対して従来から様々な研究が提案されてきた [13], [14], [16], [17]．Label-Propagation [17] はノード数 $n$ に対して線形に近い計算量で動作する非常に高速な手法である．まず最初に各ノードごとに異なるラベルを割り当て，その後はノード間でリンクに沿ってラベルを伝播しあい，最近傍法と同様に各ノードごとに近傍ノードにおける最頻ラベルへの更新を行う．

[*1] ラベルに基づく評価ではなく，データ表現に基づく cluster validity [3] での評価を行うことに対応する．

© 2012 Information Processing Society of Japan

ラベルの変化がなくなるまでリンクを通じた更新を行い，最後に同じラベルを持つノードを同一のコミュニティに割り当てる．

Walktrap [16] は PageRank$^{TM}$ [15] と同様にネットワーク上でのランダムウォークに基づく手法である．短いランダムウォークから計算される遷移確率に基づいてノード間の距離を計算するランダムウォークにより求めたノード間距離を用いてボトムアップな階層的クラスタリングを行い，モジュラリティが最大となるコミュニティへの分割を行う．

LeadingEigenvector [13] は，2 つのクラスタ（コミュニティ）への分割に対する緩和問題をグラフカット [6], [10] と同様に式 (1) のモジュラリティを行列の固有値として定式化し，最大の固有値に対応する固有ベクトルを用いてコミュニティの分割を行う．ネットワークの 2 分割を再帰的に適用することでトップダウンに階層的クラスタリングを行い，モジュラリティが最大となるコミュニティへの分割を行う．

リンク構造が完全には観測できない場合に対し，文献 [11] は将来のネットワーク構造を予測するというリンク予測問題を定義し，ノードの近接性に基づいてリンク予測を行う手法を提案した．また，リンク構造以外の情報の活用に関しては，文献 [9] は social rating におけるユーザからの評価を活用することを考え，観測されたリンクに基づくネットワークとユーザ（ノード）間の類似性に基づくネットワークにおける近傍ノードの影響をそれぞれ評価して比較した．

## 3. ノード情報を考慮した正則化モジュラリティ固有空間法

### 3.1 ネットワークのノード情報の活用

ネットワークに存在する少数のリンクしか観測できない状況においては，ノードの接続情報（リンク）とあわせてノードに関する情報を副次的な情報として活用してコミュニティ発見を行うことが考えられる．本稿では，ノード情報に基づいて非負の実数値を持つ重み付きグラフとしてノード対の関係を表現することにより，リンク情報とノード情報を統一的に活用するアプローチを提案する．以下ではノード対の関係を表現するグラフをノードグラフと呼び，その重み行列を $\mathbf{W}$ と表記し，$\mathbf{W}$ の各要素はノード対の類似度を表すものとする．ノードグラフを定義するためにはノード対の類似度とノード間の接続関係を定義する必要があるが，それぞれ 3.1.1 項と 3.1.2 項で述べる．

#### 3.1.1 ノード対の類似度

ノード情報がテキスト形式で表現されると想定して，i) 文書処理で標準的なコサイン類似度，ii) 相互情報量に基づく類似度，を考える．後者は下記で定義される [20]．

$$\mathbf{K}_{ij} = exp(-\beta D_{KL}[p(y|x_i)||p(y|x_j)]) \tag{2}$$

$$\mathbf{W}_{ij} = \frac{1}{2}(\mathbf{K}_{ij} + \mathbf{K}_{ji}) \tag{3}$$

ここで $p(\cdot|\cdot)$ は条件付き確率，$D_{KL}[\cdot||\cdot]$ は Kullback-Leibler（KL）ダイバージェンスである [7] であり，$\beta$ はパラメータである．ノード情報がテキスト形式で記述されている場合，$x_i$ は $i$ 番目のノードに対応し，$y$ はそのノード情報の記述に使用された単語の頻度に対応する．以下では，式 (3) を KL 類似度と呼ぶ．

#### 3.1.2 接続関係

ノード間の接続関係としては，密なグラフを用いる場合と疎なグラフ[*2]を用いる場合が考えられる．密なグラフの例としては，すべてのノードがお互いに接続される完全グラフがある．完全グラフではノード対の類似度に関する情報をすべて用いることになる．他方，疎なグラフとしては，ノードごとに類似度の降順に他の $k$ 個のノードを選択して接続する $k$-近傍グラフがある．ネットワーク分析で用いられるデータは自己ループのない単純グラフとして表現されることが多いため，本稿ではどちらの場合も自己ループはないものとする．

### 3.2 混合モデル

近年，ラベルありデータとラベルなしデータを活用する半教師あり学習が注目を集めている [1], [5]．この理由として，半教師あり学習では少量のラベルありデータと大量のラベルなしデータを用いることにより，学習に必要なラベルありデータを準備する手間を抑えながら性能を格段に向上させることができることがあげられる．本稿ではノード情報から得られる情報を半教師あり学習における教師情報 [1] ととらえ，少数のリンクしか観測できない状況下でのコミュニティ発見に活用することを考える．以下では観測されたリンクに基づいたネットワークの表現行列を $\mathbf{A}_r$ と表す．

ネットワークにおけるリンク情報とノード情報をグラフ構造の観点から統一的に扱う際，統計的推測における混合分布モデルのように両者を混合モデルとして表現して活用することが考えられる．観測されたリンクに基づく表現行列 $\mathbf{A}_r$ とノードに基づく表現行列 $\mathbf{W}$ における各要素値が [0,1] であるという正規化条件のもとで，以下の重み付きグラフとしての混合モデルを考える．

$$\mathbf{M}_r(\alpha) = \alpha\mathbf{A}_r + (1-\alpha)\mathbf{W} \tag{4}$$

ここで $\alpha \in [0,1]$ は混合時の重み係数である．

### 3.3 正則化モジュラリティ固有空間法

#### 3.3.1 モジュラリティ行列

ネットワークをコミュニティに分割することはグラフの

[*2] 辺数 $m$ が頂点数 $n$ に対し $O(n)$ の場合は疎なグラフと呼ばれる [8]．

© 2012 Information Processing Society of Japan

辺を除去して部分グラフに分割することに対応する．部分グラフへの分割に対してはこれまで機械学習の分野で様々な手法が提案されているが，グラフカットに基づくグラフラプシアン [6] を用いた固有空間法では固有空間への非線形射影を活用して分割を行う [2]．しかし，機械学習とコミュニティ発見では一般に評価指標（最適化規準）が異なるため，この手法を単にネットワークに適用するだけではモジュラリティの高いコミュニティを同定することは困難である [13]．

式 (1) で定義されるモジュラリティの最大化は，以下の行列の最大固有値に対する固有ベクトルを求めることに対応することが示されている [13]．

$$\mathbf{B} = \mathbf{A} - \mathbf{P} \tag{5}$$

ここで行列 $\mathbf{B}$ はモジュラリティ行列と呼ばれる．少数のリンクしか観測できない場合であっても，その表現行列と対応する式 (1) の null モデルに基づいてモジュラリティ行列 $\mathbf{B}_r$ を定義することができる．

### 3.3.2　正則化モジュラリティ固有空間法

式 (5) のモジュラリティ行列に基づき，ノード情報をコミュニティ発見に対する副次的な情報ととらえて正則化に活用することを考え，観測されたリンク情報とあわせてノード情報を考慮したコミュニティ発見を下記の目的関数の最大化として定式化する．

$$J_1 = \mathrm{tr}(\mathbf{F}^T\mathbf{B}_r\mathbf{F}) + \nu\,\mathrm{tr}(\mathbf{F}^T\mathbf{W}\mathbf{F}) \tag{6}$$

ここで $\nu \geq 0$ はパラメータであり，$\mathbf{F} \in \mathbb{R}^{n\times l}$ の列ベクトルは式 (6) を最大化するために固有値の降順に選択した固有ベクトルである．式 (6) の第 1 項は文献 [13] と同様にモジュラリティの最大化のために式 (5) の固有ベクトルを求めることに対応する．他方，第 2 項はノード間での類似性を固有空間においてどの程度保存するかを表す正則化項であり，類似性を保存するほど大きな値となる．このため，提案法ではノード情報に基づく類似性を保存しながらモジュラリティを最大化する固有ベクトルを活用することになる．

半教師あり学習，特にノード対に関する制約のもとでの半教師ありクラスタリング [1], [5] においても，グラフ構造から導出されるグラフラプラシアン $\mathbf{L}$ を正則化した以下の目的関数を最小化する行列 $\mathbf{F}$ が利用されている [4], [21]．

$$J_2 = \mathrm{tr}(\mathbf{F}^t\mathbf{L}\mathbf{F}) + \lambda\,\mathrm{tr}(\mathbf{F}^t\mathbf{S}\mathbf{F}) \quad \text{s.t.}\ \ \mathbf{F}^t\mathbf{D}\mathbf{F} = \mathbf{I} \tag{7}$$

ここで $\mathbf{D}$ はグラフの次数ベクトル $\boldsymbol{k}$ を対角要素とする対角行列，実数 $\lambda \geq 0$ はパラメータであり，行列 $\mathbf{S}$ は与えられた制約を表現するラベル行列と呼ばれ，第 2 項が教師情報（ノード対に関する制約）を反映した正則化項となる．

式 (6) と式 (7) を比べると，両方ともそれぞれの処理の目的を反映する第 1 項に正則化項を加えた形となっていることが分かる．このため，式 (6) はコミュニティ発見における目的関数（モジュラリティの最大化）を副次的に得られる情報（ノードグラフ）で正則化したものと解釈することができる．

**Require:** $\mathbf{A}_r$; *//an adjacency matrix*
**Require:** $\mathbf{W}$; *//a similarity matrix*
**Require:** $\nu$; *//regularization parameter*
**Require:** $l$; *//the number of dimensions of the subspace*
**Require:** $c$; *//the number of clusters*

1: $\boldsymbol{k} = \mathbf{A}_r\mathbf{1}_n$　　*//*$\mathbf{1}_n$ *= (1,...,1)*$^T$
2: $\mathbf{P}_r = \boldsymbol{k}\boldsymbol{k}^T/\mathbf{1}_n^T\mathbf{A}_r\mathbf{1}_n$
3: $\mathbf{B}_r = \mathbf{A}_r - \mathbf{P}_r$
4: Find $l$ eigenvectors $\mathbf{F} = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_l\}$ for $\mathbf{B}_r + \nu\mathbf{W}$ with the largest eigenvalues.
5: Apply a clustering algorithm over $\mathbf{F}$ and construct $c$ clusters.
6: **return** clusters

**図 1**　正則化モジュラリティ固有空間法に基づくコミュニティ発見
**Fig. 1**　Community discovery based on regularized modularity eigenmap.

### 3.3.3　混合モデルとの対応

本稿では半教師あり学習の観点からノード情報を活用することを考え，3.3.2 項ではモジュラリティ行列の正則化に基づくアプローチを提案した．式 (4) と式 (6) ではノードグラフの用い方は異なるが，式 (6) も行列の混合モデルとしての解釈が可能である．

式 (6) の目的関数は $0 < \alpha \leq 1$ である $\alpha$ に対して以下に書き換えられる[*3]．

$$\begin{aligned} J &= \mathrm{tr}(\mathbf{F}^T\mathbf{B}_r\mathbf{F} + \nu\mathbf{F}^T\mathbf{W}\mathbf{F}) \qquad (8)\\ &= \mathrm{tr}\left(\frac{1}{\alpha}\mathbf{F}^T(\alpha\mathbf{B}_r + \alpha\nu\mathbf{W})\mathbf{F}\right)\\ &= \frac{1}{\alpha}\,\mathrm{tr}\left(\mathbf{F}^T(\alpha\mathbf{B}_r + (1-\alpha)\mathbf{W})\mathbf{F}\right) \qquad (9)\end{aligned}$$

ここで $\nu = (1-\alpha)/\alpha$ である．このため，3.3.2 項の提案法は 3.2 節と同様に混合和と見なすことができる．固有ベクトルから構成される $\mathbf{F}$ を求める際には式 (9) の係数 $1/\alpha$ に不変であることに注意されたい．

### 3.3.4　アルゴリズム

提案法に基づくコミュニティ発見アルゴリズムを**図 1** に示す．図 1 の 1 行目から 3 行目で生成したモジュラリティ行列に対し，4 行目で提案法を適用する．5 行目で固有空間上で既存のクラスタリング手法を適用してノード集合の分割を行いコミュニティを同定する．

[*3] $\alpha = 0$ は $\mathbf{W}$ を用いない場合に対応するため，式 (6) で正則化を用いない場合と解釈できる．

© 2012 Information Processing Society of Japan

## 4. 評価

### 4.1 実験設定

#### 4.1.1 対象データ

実験では，1) Pascal データセット（Pascal）[*4]，2) InfoVis2004 データセット（IV'04）[*5]，を対象とし，共著関係ネットワークからのコミュニティ発見を行った．Pascal は 313 人の著者が Pascal EPrints サーバにアップした約 720 の論文における共著関係ネットワークであり，接続行列は 381 個の非零要素を含む．IV'04 は情報可視化分野において 1974 年から 2004 年に発表された 614 の論文に関するデータであり，2004 年の可視化コンテストで使用された．このデータから 735 人の著者を抽出して共著関係ネットワークを構築したものであり，接続行列は 1360 個の非零要素を含む．以下ではそれぞれの共著関係ネットワークの表現行列を $\mathbf{A}$ とするが，要素値 $\mathbf{A}_{ij}$ は著者 $i$，$j$ の共著論文数を各データセットでの最大共著数で割り正規化したものとした[*6]．

ノード情報としては，ノードに対応する著者が発表した論文のタイトルとアブストラクトは著者の専門分野を反映すると考え，そこで使用された単語を抽出した．抽出した単語に porter stemmer [*7]を用いて stemming を行い，MontyTagger [*8]を用いて品詞に分解し，stop word を除去して相互情報量で上位 2,000 語の単語を選択し，ベクトル空間モデルに基づいてノード（著者）情報を単語頻度として表現した．

#### 4.1.2 比較手法

提案法を 2.3 節で述べた，1) labelPropagation [17]，2) leadingEigenvector [13]，3) walktrap [16]，と比較した．それぞれの手法による分割に対し，オリジナルの共著関係ネットワークの表現行列に対して式 (1) のモジュラリティを評価した．なお，著者間での関係は共著論文数を正規化した値として表現されるため，2.2 節で述べたように非負の重み行列に対するモジュラリティを評価した．

#### 4.1.3 実験手順

観測できるリンク数が少ない状況をシミュレートするため，オリジナルの共著関係ネットワークからランダムにノード対を選択し，選択されたノード間でのリンクのみを含むネットワークを構築した．構築したネットワークに各手法を適用してネットワークを分割した際，オリジナルの共著関係ネットワークに対してその分割のモジュラリティを評価したため，その値はオリジナルのネットワークにおけるコミュニティ構造を観測されたリンクからどの程度同定できるかを表すことになる．なお，同定されるコミュニティはノード対のランダムな選択に影響されるため，100 回試行に対する平均を計算した．

コミュニティ発見では一般にコミュニティ（クラスタ）の数が未知であるため，2.3 節で述べたように Walktrap [16] や LeadingEigenvector [13] では階層的クラスタリングを行う際にクラスタ数を変えてモジュラリティが最大となる分割を求める．提案法で行列 $\mathbf{F}$ に対しクラスタリングを行う際も同様にクラスタ数を変えてモジュラリティ（ただし，観測された $\mathbf{A}_r$ に対するもの）が最大となる分割を求めた．

#### 4.1.4 実験パラメータ

4.1.3 項で述べた選択するノード対の割合を [0,1] の範囲で変え，観測できるリンク数の影響を調べた．また，3.3.3 項で述べたように提案法におけるパラメータ $\nu$ は 3.2 節における重み係数 $\alpha$ と対応づけられるため，以下では $\alpha$ の影響を調べた．

3.1.2 項で述べたように，ノードグラフにおける接続関係として密なグラフと疎なグラフを用いることが考えられる．密なグラフの例として完全グラフ，疎なグラフの例として小さな近傍数での $k$-近傍グラフに対する結果を報告する．ただし，ノードグラフはリンクが部分的にしか観測できない場合を補うためのものであるが，近傍数 $k$ を小さくすることはノード情報から得ることができる情報をある意味で捨ててしまうことに対応する．

提案法では固有ベクトルの数 $l$ を指定する必要がある．グラフカット [6], [10] に基づくスペクトルクラスタリングにおいてはクラスタ数が既知の場合には $l$ をクラスタ数とすることが多いが，コミュニティ発見ではクラスタ数は事前に指定できない．スペクトルクラスタリングと同様，予備実験の際に提案法の挙動は本数にあまり影響されなかったため，以下の実験では $l$=10 とした．

### 4.2 結果

#### 4.2.1 密なグラフに対する結果

ノードグラフとして密なグラフを用いた場合（完全グラフ）を**図 2**，**図 3** に示す[*9]．図中で縦軸はモジュラリティであり，横軸は選択したノード対の割合[*10]である．凡例中の値は式 (4) での重み係数 $\alpha \in [0,1]$ であり，丸付き黒線は $\alpha$=0（リンク情報を用いず，ノードグラフのみを用いる場合），三角赤破線は $\alpha$=1（観測されたリンク情報のみを用いる場合）に対応する．各図で上段は Pascal，下段は IV'04 の結果であり，図 2 ではコサイン類似度，図 3 では KL 類似度を用いた．

図 2，図 3 の結果より，観測できるリンクが少ない状

[*4] http://analytics.ijs.si/~blazf/pvc/data.html
[*5] http://iv.slis.indiana.edu/ref/iv04contest/
[*6] このため，2.2 節で述べたように共著関係ネットワークは非負の実数値を持つ表現行列として表現される．
[*7] http://www.tartarus.org/~martin/PorterStemmer
[*8] http://web.media.mit.edu/~hugo/montytagger
[*9] KL 類似度を用いた場合に walktrap は動作しなかったため図 3 に示していない．
[*10] 観測されるリンク数の割合に対応し，大きいほど共著関係ネットワークにおけるより多くのリンクが使用されることになる．

© 2012 Information Processing Society of Japan

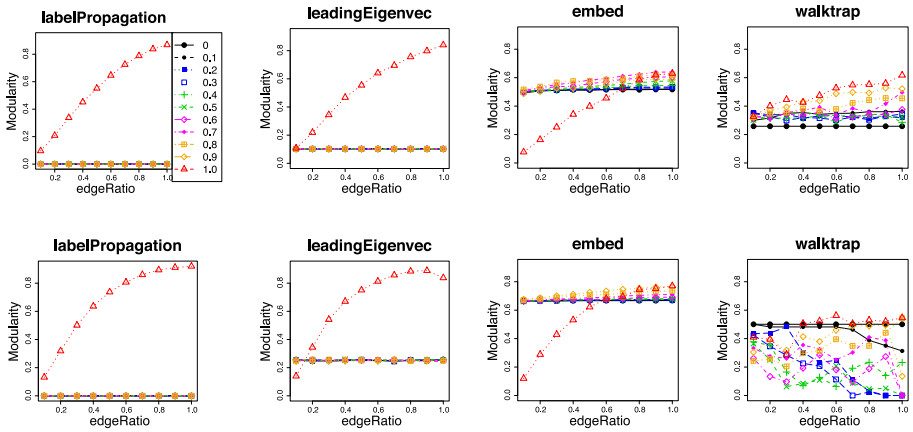



図 2　密なグラフの場合（完全グラフ，コサイン類似度）（上段：Pascal，下段：IV'04）

**Fig. 2**　Dense Graph (complete graph, cosine similarity) (first row: Pascal, second row: IV'04).

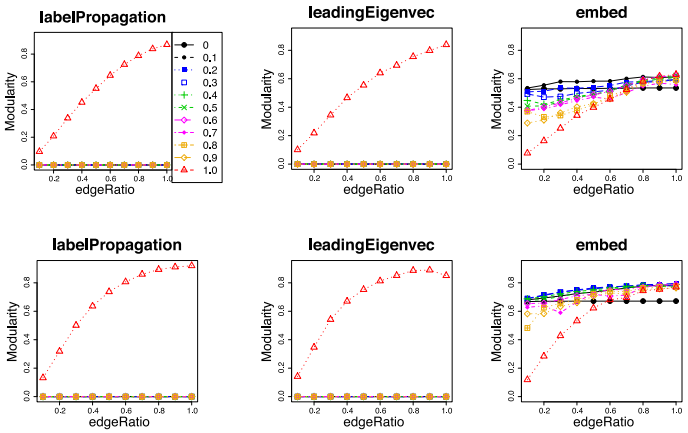



図 3　密なグラフの場合（完全グラフ，KL 類似度）（上段：Pascal，下段：IV'04）

**Fig. 3**　Dense Graph (complete graph, KL similarity) (first row: Pascal, second row: IV'04).

況を補うためにノードグラフを活用する場合（$\alpha < 1$ の場合）には，コサイン類似度，KL 類似度を用いた場合もともに提案法（図の左から 3 列目，embed と表記）は他手法を上回った．特に，リンクが少ない状況ではノードグラフを用いない場合（$\alpha = 1$，三角赤破線）よりも用いた場合（$\alpha < 1$）のほうがモジュラリティが大きくなり，ノードグラフを活用する効果が見られた．式 (6) の第 2 項を通じてノード情報に基づく類似性を保存することが，観測されたリンクが少ない場合にはノード（著者）間での関係をとらえることに効果的であるためと考えられる．他方，提案法では観測されるリンクが多くなるほどモジュラリティは大きくなったが，観測できるリンクが多い状況下でリンク情報のみを用いる場合（$\alpha = 1$）での他手法と比較すると，labelPropagation，leadingEigenvector には及ばなかった．式 (1) のモジュラリティを評価指標とする場合，式 (6) の第 1 項はモジュラリティの最大化に対応するが，第 2 項は最大化に寄与するとは限らないためと考えられる．

labelPropagation [17]（図で最左の列）と leadingEigenvector [13]（図で左から第 2 列目）においては，$\alpha < 1$ としてノードグラフを用いるとリンク数を変えてもモジュラリティはほぼ一定となり，$\alpha$ に対する変化もみられなかった．ノードグラフを活用することでモジュラリティが非常に小さな値となったため，これらの手法では 3.2 節の混合モデルとしてノード情報を活用することはできなかったと考えられる．特に，labelPropagation では近傍ノードの数が多い場合にサイズが大きい（ノード数の多い）コミュニティの影響が強くなってしまい，観測されたリンクから得られる情報があまり活用できないためと考えられる．他方，walktrap [16]（図で最右の列）では $\alpha$ に応じてモジュラリティは変化し，基本的には $\alpha$ が大きいほど（リンク情報をより使用するほど）モジュラリティは大きくなった（ただし，図 2 で IV'04 に対しては非常にばらついた）．ネットワーク上でのランダムウォークはノードの次数（近傍ノードの数）が多いほど辺に沿った遷移確率が一般に小さくなり，他手法と比べて walktrap では混合比の影響を受けやすいためと考えられる．

密なグラフとしては $k$-近傍グラフを大きな近傍数に対して構築することも考えられるが，近傍数 $k$ が 40 以上となると完全グラフの場合とほぼ同様な結果となった．このため，ノードグラフはある程度次数が大きくなると密なグラ

© 2012 Information Processing Society of Japan

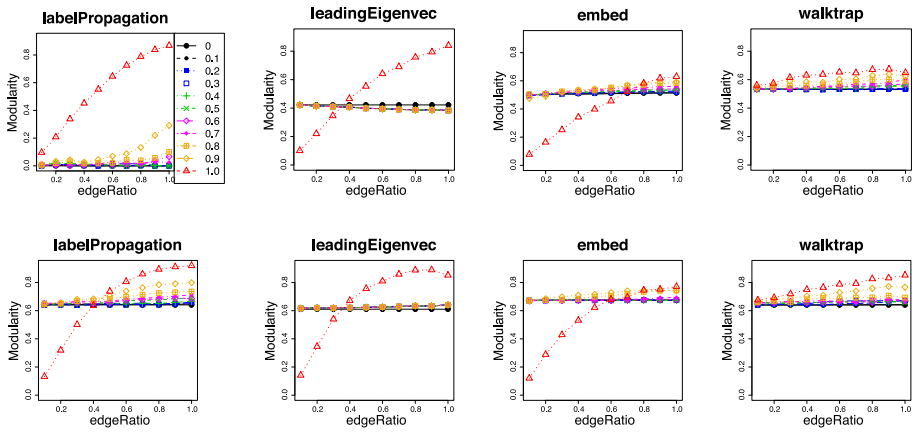



図 4　疎なグラフの場合（$k$=20，コサイン類似度）（上段：Pascal，下段：IV'04）

**Fig. 4**　Sparse Graph ($k$=20, cosine similarity) (first row: Pascal, second row: IV'04).

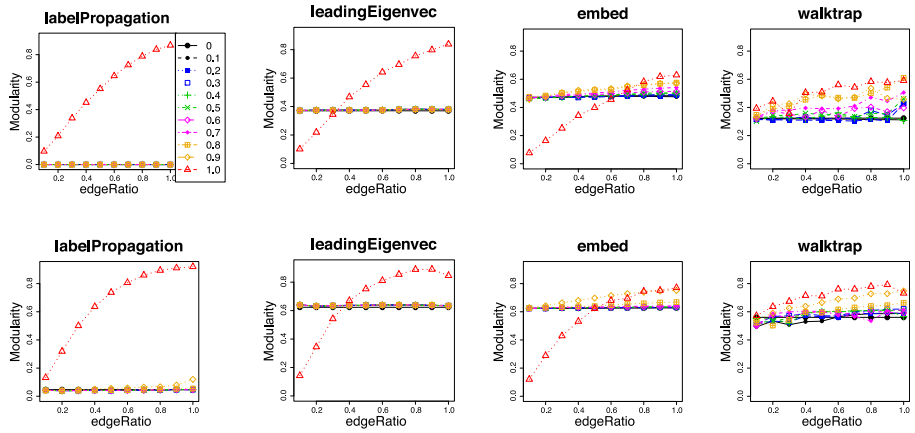



図 5　疎なグラフでの結果（$k$=20，KL 類似度）（上段：Pascal，下段：IV'04）

**Fig. 5**　Sparse Graph ($k$=20, KL similarity) (first row: Pascal, second row: IV'04).

フと見なせると考えられる．

### 4.2.2　疎なグラフに対する結果

疎なグラフの例として，近傍数 $k$ が少ない $k$-近傍グラフを用いた結果を示す．以下では近傍数 $k$ が 10，20 の場合を示す[*11]．

近傍数 $k$ を 20 とした場合を**図 4**，**図 5** に示す．4.2.1 項での密なグラフに対する結果と同様な傾向がみられ，観測できるリンクが少ない状況下においては提案法はノードグラフを活用することで他手法を上回った．しかし，観測できるリンクが増えた際には他手法でノードグラフを用いない場合（$\alpha = 1$）には及ばなかった．提案法と同様にモジュラリティ行列に対する固有ベクトルを用いる leadingEigenvector では，ノードグラフを用いた場合（$\alpha < 1$）にはリンク数や $\alpha$ に応じたモジュラリティの変化がみられなかったが，疎なグラフを用いることでモジュラリティは大きくなった．共著関係を表現するオリジナルのネットワークは疎なグラフであるが，ノードグラフでも近傍数を減少させて類似度の高いノードとのみ接続させることにより，前者に類似した構造がノードグラフにも表現されやすくなるためではないかと考えられる．labelPropagation も leadingEigenvector と同様に $\alpha < 1$ ではモジュラリティは小さかったが，コサイン類似度に対して疎なグラフを用いることで $\alpha$ に応じてモジュラリティが多少変動した（図 4 参照）．walktrap は疎なグラフでも完全グラフの場合とほぼ同様な挙動を示した．

さらに近傍数 $k$ を減らしてより疎なグラフを用いた場合（$k$=10）を**図 6**，**図 7** に示す[*12]．基本的には上述の $k$=20 の場合と同様な結果となり，labelPropagation, leadingEigenvector ではモジュラリティはほぼ一定であったが，近傍数 $k$ が少なくなるほどモジュラリティが大きくなり，また labelPropagation では $\alpha$ に応じた変化が多少みられた（図 6 参照）．walktrap においても近傍数 $k$ が少ない方がモジュラリティが大きくなった．他方，提案法では逆に近傍数 $k$ が少なくなるほど $\alpha$ に応じた変化がみられなくなった．近傍数が減ることで式 (6) の第 2 項で考慮するノード対が少なくなり，その影響が小さくなるためと考えられる．

[*11] $k$ が小さすぎるとネットワーク全体が非連結になるため比較した手法（walktrap）は動作しなかった．

[*12] IV'04 に対して KL 類似度を用いた場合に walktrap は動作しなかったために示していない．

© 2012 Information Processing Society of Japan

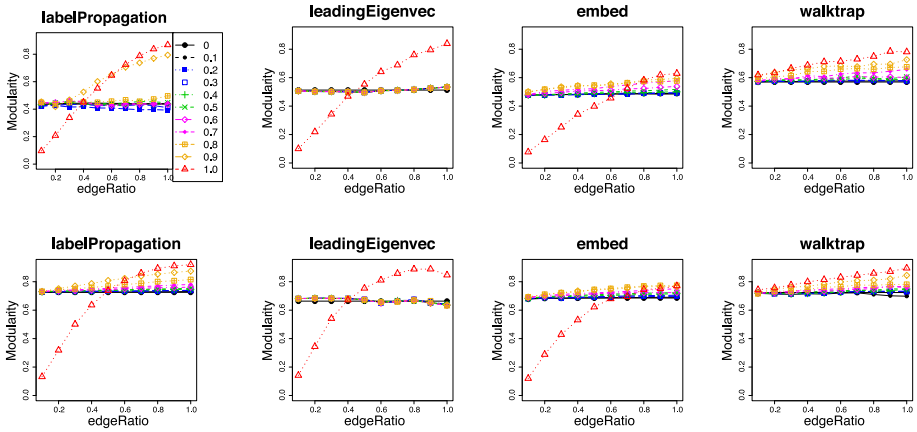



図 6　疎なグラフの場合（$k$=10，コサイン類似度）（上段：Pascal，下段：IV'04）

**Fig. 6**　Sparse Graph ($k$=10, cosine similarity) (first row: Pascal, second row: IV'04).

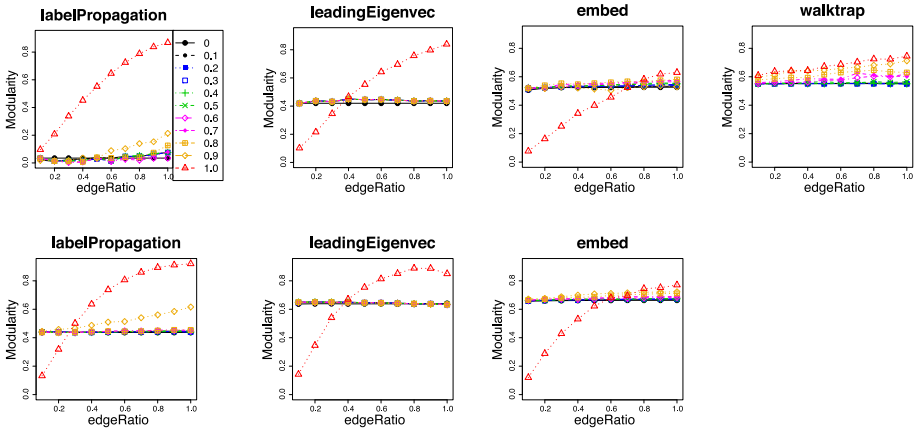



図 7　疎なグラフの場合（$k$=10，KL 類似度）（上段：Pascal，下段：IV'04）

**Fig. 7**　Sparse Graph ($k$=10, KL similarity) (first row: Pascal, second row: IV'04).

### 4.3　考察

観測できるネットワークの接続情報が少ない状況下におけるコミュニティ発見に対し，本稿ではネットワークのノードに関する情報を副次的な情報として活用してコミュニティ発見を行うアプローチを提案した．提案法ではノード情報に基づいて構築したグラフ（ノードグラフ）を正則化項として固有ベクトルを求め，固有空間においてノード集合を分割してコミュニティを同定する．ネットワークにおける少数のリンクしか観測できない状況下でノードグラフを活用した場合には提案法は他手法を上回った．しかし観測できるリンクが増えた際には他手法でリンク情報のみを用いた場合には及ばなかった．より積極的なリンク情報の活用を通じた性能向上は今後の課題である．

ノード情報を非負の重み付きグラフとして活用する際には，グラフにおける辺の重みを定義する類似度だけではなく，その接続関係（次数）も同定されるコミュニティに影響を与えることが分かった．提案法は密なグラフとしてノードグラフを用いる際には他手法と比較して効果的であったが，逆に（非常に）疎なグラフに対しては他手法に及ばなかった．他方，4.1.4 項で述べたように疎なグラフを用いることはノード情報から得られる情報をある意味で捨ててしまうことに対応するが，比較した手法では疎なグラフとしてノードグラフを活用した場合のほうがモジュラリティが大きくなった．

4.1.2 項の比較手法では 3.2 節の混合モデルにより構築したネットワークに対してコミュニティ発見を行った．これは観測されたリンク情報をノードグラフで補ったネットワーク上でコミュニティ発見行うことに対応する．他方，提案法ではネットワークの重み行列をノードグラフで補うのではなく，観測されたネットワークのモジュラリティ行列に対する正則化にノードグラフを用いてコミュニティを同定した．

## 5.　おわりに

観測されるネットワークの接続情報が少ない状況下でのコミュニティ発見に対し，本稿ではノード情報を考慮した正則化モジュラリティ固有空間法を提案した．ネットワークにおける接続関係とあわせてノード（頂点）に関する情報が得られる場合には，前者を補うために後者を副次的な情報として活用してコミュニティ発見を行うことが考えら

© 2012 Information Processing Society of Japan

れる．提案法ではリンク情報とノード情報を統一的に扱うためにノード情報に基づくグラフ構造を構築し，構築したグラフを正則化項として活用してコミュニティ発見を行う．

提案法を共著関係ネットワークに適用し，他手法との比較を通じて観測できるリンクが少ない状況下での提案法の有効性を確認した．ノード情報に基づくグラフ構造を正則化項として活用することにより，少ないリンクからでもネットワークにおける密な部分グラフをコミュニティとして同定することができた．しかし，観測できるリンクが増えた場合にはモジュラリティの観点からは課題が残ることも判明した．今後はより積極的なリンクの活用を通じた性能向上や，次数が少ない疎なグラフの効果的な活用に取り組んでいく予定である．

## 参考文献


[1] Basu, S., Davidson, I. and Wagstaff, K.(Eds.): *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, Chapman & Hall/CRC Press (2008).

[2] Belkin, M. and Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, *Neural Computation*, Vol.15, pp.1373–1396 (2002).

[3] Bezdek, J. and Pal, N.: Some new indexes of cluster validity, *IEEE Trans. Systems, Man and Cybernetics*, Vol.28, No.3, pp.301–315 (1998).

[4] Bie, T.D., Suykens, J. and Moor, B.D.: Learning from General Label Constraints, LNCS 3138, pp.671–679 (2004).

[5] Chapelle, O., Schölkopf, B. and Zien, A.(Eds.): *Semi-Supervised Learning*, MIT Press (2006).

[6] Chung, F.: *Spectral Graph Theory*, American Mathematical Society (1997).

[7] Cover, T. and Thomas, J.: *Elements of Information Theory*, Wiley (2006).

[8] Diestel, R.: *Graph Theory*, Springer (2006).

[9] Jamali, M. and Ester, M.: Modeling and Comparing the Influence of Neighbors on the Behavior of Users in Social and Similarity Networks, *Proc. ICDM Workshops*, pp.336–343 (2010).

[10] Kannan, R., Vempala, S. and Vetta, A.: On Clusterings: Good, Bad and Spectral, *J. ACM*, Vol.51, No.3, pp.497–515 (2004).

[11] Liben-Nowell, D. and Kleinberg, J.: The Link Prediction Problem for Social Networks, *Proc. CIKM'03*, pp.556–559 (2003).

[12] Mika, P.: *Social Networks and the Semantic Web*, Springer (2007).

[13] Newman, M.: Finding community structure using the eigenvectors of matrices, *Physical Review E*, Vol.76, No.3, p.036104 (2006).

[14] Newman, M.: *Networks: An Introduction*, Oxford University Press (2010).

[15] Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web, Technical Report SIDL-WP-1999-0120, Stanford University (1999).

[16] Pons, P. and Latapy, M.: Computing communities in large networks using random walks, *Journal of Graph Algorithms*, Vol.10, No.2, pp.191–218 (2006).

[17] Raghavan, U., Albert, R. and Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks, *Physical Review E*, Vol.76, p.036106 (2007).

[18] Watts, D.J.: *Small Worlds: The Dynamics of Networks Between Order and Randomness*, Princeton Univ Press (2003).

[19] Kempe, D., Kleinberg, J. and Tardos, E.: Maximizing the Spread of Influence through a Social Network, *Proc. KDD-2003*, pp.137–146 (2003).

[20] 吉田哲也：相互情報量に基づくクラスタリングに対するグラフモデルとその評価，情報処理学会論文誌：数理モデル化と応用，Vol.3, No.3, pp.1–11 (2010).

[21] 吉田哲也，岡谷一宏：制約を反映するグラフ表現に基づく射影に基づく半教師ありクラスタリング，情報処理学会論文誌：数理モデル化と応用，Vol.4, No.1, pp.62–71 (2011).





**吉田 哲也**　（正会員）

1968 年生．1991 年東京大学工学部航空工学科卒業．1997 年東京大学大学院博士課程修了．工学博士．同年大阪大学大学院基礎工学研究科助手．2001 年大阪大学産業科学研究所助手．2004 年北海道大学大学院情報科学研究科助教授．現在，同大学准教授．主に機械学習，知識獲得，データマイニング等の研究に興味を持つ．人工知能学会会員．


© 2012 Information Processing Society of Japan