情報処理学会論文誌　データベース　Vol.5　No.3　26–35 (Sep. 2012)

機能性に基づくコミュニティ抽出法の比較

伏見 卓恭[1,a)]　斉藤 和巳[1]　風間 一洋[2]

受付日 2012年3月20日, 採録日 2012年7月7日

概要：本稿では，ネットワークに対する各ノードの役割・機能・立場が類似したノードからなるコミュニティを抽出することを主題とする．周辺ノードとのリンク関係の類似性，すなわち同値性を同定するための Versim 法，Simrank 法と呼ぶ従来の手法，および機能の類似するノード群を抽出する Randwalk 法の 3 手法に着目する．Randwalk 法は，ネットワーク全体でのランダムウォークにより類似経路構造を探す方法であり，PageRank 反復計算時のスコア収束曲線の類似性を用いる手法である．結果より，局所的なリンク構造に着目する Versim 法と Simrank 法では，手法の問題点が顕著に現れる人工ネットワークや表出する構造にバラつきのある現実ネットワークへの適用に限界があることを示す．一方，大域的な構造上での現象の類似性による Randwalk 法は，現実ネットワークに対しても機能が類似するノード群を抽出可能であることを示す．

キーワード：機能コミュニティ，正則同値，ランダムウォーク

Comparison of Community Extraction Method Based on Functionality Equivalence

Takayasu Fushimi[1,a)]　Kazumi Saito[1]　Kazuhiro Kazama[2]

Received: March 20, 2012, Accepted: July 7, 2012

Abstract: In this paper, we attempt to extract communities in a given network, each of which consists of nodes with a homogeneous role or function. To this end, we focus on three methods using respective node similarity based on regular equivalence, i.e., two conventional methods refered to as Versim and Simrank, and a recently proposed method refered to as Randwalk, which calculates the node similarity based on the convergence process of a PageRank score. By our experimental results using some artificial and real networks, we show that the Versim and Simrank methods have some real limitations to apply to real networks because of directly focusing on local link structures. On the other hand, the Randwalk method more explicitly extracted fucntional community, charcterized by similarity from a relative location, role, or hierarchical status. This is because the Randwalk method focuses on the similarity of phenomena on the global link structure.

Keywords: functional community, regular equivalence, random walk

1. はじめに

現実社会における人と人とのつながりや，Web サービスにおけるユーザ間のつながりなどのソーシャル・ネットワークに対する関心が高まっている．人間関係にとどまらず，Web ページのハイパーリンクネットワークや道路網など，あらゆるところで複雑ネットワークが見受けられるようになっている．現実ネットワークにおいて，すべてのノードは均質ではなく，各ノードは固有の立場や役割，機能を有しており，これらに基づき，多大なノード群をクラスタリングしたり，重要ノードを抽出したりするための手法が提案されている．ネットワーク構造に関しても，全体が均質ではなく，リンクが密な部分もあれば疎な部分もあり，コミュニティ構造を有することが指摘されている [1].

[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, Shizuoka 422–8526, Japan
[2] 日本電信電話株式会社未来ねっと研究所
Network Innovation Laboratories, Nippon Telegraph and Telephone Corporation, Musashino, Tokyo 180–8585, Japan
[a)] j11507@u-shizuoka-ken.ac.jp

既存のコミュニティ抽出手法として，Newman による Modularity というネットワーク分割指標を用いたコミュニティ抽出手法が高速で大規模ネットワークに対しても有効であり注目を浴びている [2]．さらに，スペクトラルグラフ分析の手法である Normalized Cut 法 [3] や Ratio Cut 法 [4] などもあげられる．これらは，クラスタ内リンクを多く，クラスタ間リンクを少なくする，すなわち，ノードどうしの結合が疎な部分を切断し，いくつかのノード集合（サブネットワーク）に分割する方法である．一方，ネットワーク上でのノードどうしが密結合したサブネットワークをコミュニティと見なして，クリーク（clique）やクリークの条件を緩めたサブネットワークをみつけるための様々な手法が提案されている [5], [6], [7], [8]．これらを代表とする既存のコミュニティ抽出手法の多くは，無向ネットワークにおいてリンク構造の粗密に着目し，全ノード集合をいくつかの部分集合に分割することに主眼を置いている．

また，ネットワークにおいて構造上類似した立場にあるノードの概念として同値性がある．同値なノード群を同定する代表手法である REGE，CATREGE アルゴリズムは，計算量の点で大規模なネットワークには対応できない [9]．同値性は社会学において古典的な考え方であるため，文献 [9] は 20 年前の研究であるが，近年では同値性の概念を発展させ，近似的に同定する手法として Leicht ら [10] や Jeh ら [11] の手法がある．Leicht らは，“類似ノードの周囲のリンク関係は類似する” という仮定のもとで，ノード間の類似度を定義している．Jeh らは “共通の隣接ノードを有するノードは互いに類似している” という直観に基づき，ノード間の類似度を定義している．隣接行列やラプラシアン行列によるカーネルを用いて，ノード間の類似度を計算する手法も提案されている [12], [13]．ソーシャル・ネットワーク上での情報拡散において，類似した立場や役割のノード，すなわち正則同値なノードどうしは，同様の情報を保持する可能性があるという知見 [14] もある．ノードの同値性により，いくつかのグループにクラスタリングすることは，ネットワーク上でのノードの動向を調査する点で重要と考えられている [15]．

類似の研究として，著者らは文献 [16] で「機能・役割が類似するノードであれば，PageRank スコアの収束パターンも類似する」と考え，機能コミュニティと呼ぶ類似した機能を持つノード集合を抽出する方法を提案している．文献 [16] では，人工ネットワークおよびハイパーリンクネットワークを対象に，Newman らによるクラスタリング結果と提案手法による機能コミュニティの違いを定性的に評価し，従来のコミュニティ抽出とは異なる概念のコミュニティを抽出できることを示した．本稿では，各ノード間に類似度を定義し，類似度に基づいてノードをクラスタリングし，近似的に正則同値なノード集合を抽出する手法および我々の機能コミュニティを抽出する手法について比較，考察する．具体的には，前述した Leicht らと Jeh らの手法および我々の手法により，ノード間類似度行列を計算する．大規模ネットワークへの適用を視野に入れ，計算量により定量的に，可視化により定性的に評価する．評価の結果，我々の手法は大規模なネットワークに対しても適用可能であり，正則同値なノード群を抽出する 2 手法と比べ，機能・役割が類似したノード群をより明示的に抽出できることを示す．

本稿は以下のような構成である．2 章で本稿で重要となる概念について簡単に説明する．3 章でリンク密度に基づいて，ノードをクラスタリングしコミュニティを抽出する手法の代表例である Newman クラスタリング法について簡単に触れ，4 章では本稿で着目するノード間類似度の計算法について述べる．5 章で複数の異なる構造を持つネットワークを用いて，各手法を評価・比較する．6 章で我々の手法に対するいくつかの考察をし，最後に本稿のまとめと今後の展望を 7 章で述べる．

2. 同値性と機能コミュニティ

この章では，本稿で重要となる同値性の概念および機能コミュニティについて説明する．

2.1 同値性

ネットワークにおける同値性とは，隣接ノードとの局所的なリンク傾向の類似性であり，定義により構造同値（strucutal equivalence）や正則同値（regular equivalence）などに分類される．

ノード u と v が構造同値であるとは，u，v が隣接するすべてのノードとの関係が同一であることをいう．すなわち構造同値なノードを入れ替えても，ネットワーク構造はまったく変わらない．

ノード u と v が正則同値であるとは，u，v が正則同値な任意のノードペア x，y に対する接続関係が同一である場合である．構造同値と異なり，リンク相手どうしが同値なノードであれば，必ずしも相手ノードが同一ノードである必要はない．たとえば，ある会社組織において，部長らは上司として社長と，部下として何人かの課長と一般社員とリンクしていると仮定する．どの部長も，1 人の社長と何人かの異なる部下とリンクしているが，結合パターンは社長ノード，課長ノード，一般社員ノードとリンクしている点において，部長という正則同値なノード群が同定される．

2.2 機能コミュニティ

機能コミュニティとは，ネットワーク内での各ノードの機能が類似するノード群により構成されるコミュニティである．ネットワーク内での相対的位置や階層的地位，周辺ノードとの関係のパターンや次数などが類似すれば，ノー

ドが提供する機能が類似するという考えに基づいた概念である．同値性のように局所的なリンク傾向に主眼を置くのではなく，ネットワーク全体に対する個々のノードの役割に着目したものである．たとえば，社長，部長，課長，一般社員のような局所的なリンク構造から得られる形式的に定められた立場ではなく，部門内に外部からの情報を伝達する役割の媒介度の高い社員や，小グループ内でのハブ的役割の社員など，非公式に定まる機能が類似するノードを同定することを目的としている．

3. リンク密度に基づくコミュニティ抽出

リンク密度に基づくコミュニティ抽出法の代表例である，Newman クラスタリング法（以下 Newman 法）について簡単に触れる．Newman 法では，コミュニティ抽出の度合いを Modularity という定量的指標により評価している．K 個のコミュニティに対する Modularity Q は，コミュニティ i と j 間のリンク数の総リンク数に対する割合 e_{ij} を要素とする $K \times K$ の対称行列 $\mathbf{E}$ を定義し，$Q = \sum_{i=1}^{K}(e_{ii} - a_i^2) = \mathrm{Tr}(\mathbf{E}) - ||\mathbf{E}^2||$ で計算される．ここで $a_i = \sum_{j=1}^{K} e_{ij}$ であり，$||\mathbf{B}||$ は行列 $\mathbf{B}$ の要素の和（L1 ノルム）である．この値が高ければ，同一コミュニティ内のノード間にリンクが相対的に多いことになる．コミュニティの具体的な抽出法は，階層的クラスタリングと同様にデンドログラムを用いて，Modularity が最も増加するノードどうしを結合するステップを繰り返す．Modularity が最も高くなるステップ数でコミュニティを出力する．

4. ノード間類似度に基づくコミュニティ抽出

この章では，本稿で着目するノード間類似度の各種計算手法について述べる．正則同値性は，周辺リンク構造の一致を調べるものであるが，現実ネットワークの構造のばらつきに適応できるように，構造的なノード間類似度として扱うことが多い．機能コミュニティに関しても，ノードの有する機能の類似度を計算する．以下 で述べる類似度行列に基づきクラスタリングし，コミュニティを抽出する．

無向ネットワーク $G = (V, E)$ の各ノードに 1 から $|V|$ までの整数値を一意に割り振る．ここで $(u, v) \in E$ のとき $a(u, v) = 1$，それ以外のとき $a(u, v) = 0$ とし隣接行列 $\mathbf{A} \in \{0, 1\}^{|V| \times |V|}$ を定義する．自己リンクを持つノードもあり，その場合 $a(v, v) = 1$ となる．各ノード $u \in V$ に対して，$\Gamma(u)$ をノード u の隣接ノード集合とする．すなわち，$\Gamma(u) = \{v \in V;\ (u, v) \in E\}$ となる．$|\Gamma(u)|$ をノード u の次数という．自己リンク付きノード u は $u \in \Gamma(u)$ である．

4.1 Vertex Similarity 法

Leicht らの Vertex Similarity 法（以下 Versim 法）は，正則同値性の概念を拡張し，“類似ノードの周囲のリンク関係は類似する” という仮定の下で類似度行列を計算している．Versim 法による類似度行列の計算法を以下に示す．

ノード u，v 間の類似度 $s(u, v)$ を再帰的に計算する：

$$s(u, v) = \phi \sum_{w \in V} a(u, w) \cdot s(w, v) + \delta(u, v). \tag{1}$$

ここで，$0 < \phi < 1$ は減衰係数で，$\delta(u, v)$ はクロネッカーのデルタである．式 (1) を行列表記し整理すると，

$$\begin{aligned} \mathbf{S} &= \phi\mathbf{AS} + \mathbf{I} \\ &= [\mathbf{I} - \phi\mathbf{A}]^{-1} \\ &\simeq \mathbf{I} + \phi\mathbf{A} + \phi^2\mathbf{A}^2 + \cdots \end{aligned} \tag{2}$$

となる．ここで $\mathbf{I}$ は単位行列を表す．隣接行列 $\mathbf{A}$ の l 乗の u，v 要素 $a^l(u, v)$ は，ノード u からノード v への距離 l のパス数を表す．すなわち，ノード u からノード v へのパス数が多ければ多いほど，ノード u とノード v は類似度が高くなる．また減衰係数により，距離が短いパスに大きな重みが付くことになる．各項に対して，パス数の期待値で正規化し，

$$s(u, v) = \sum_{l=0}^{\infty} C_l^{uv} a^l(u, v) \tag{3}$$

とする．ここで，$C_l^{uv} = \frac{2|E|}{|\Gamma(u)| \cdot ||\Gamma(v)|} \lambda^{-l+1}$ であり，λ は隣接行列 $\mathbf{A}$ の最大固有値である．各ノードの次数を対角要素に持つ行列 $\mathbf{D}$ を用いて，式 (3) を行列表記し，式 (2) にならい整理すると，

$$\begin{aligned} \mathbf{S} &= 2\lambda|E|\mathbf{D}^{-1}\left(\mathbf{I} - \frac{\alpha}{\lambda}\mathbf{A}\right)\mathbf{D}^{-1} \\ \mathbf{DSD} &= \frac{\alpha}{\lambda}\mathbf{A}(\mathbf{DSD}) + \mathbf{I} \end{aligned} \tag{4}$$

ここで，$0 < \alpha < 1$ は減衰係数で，初期値 $(\mathbf{DSD})_0 = \mathbf{I}$ とし，式 (4) を $||(\mathbf{DSD})_t - (\mathbf{DSD})_{t-1}|| < \varepsilon$ となるか，所定の回数まで繰り返し計算することで類似度行列 $\mathbf{S}$ を得る．

Versim 法の主たる時間計算量は，行列 $\mathbf{DSD}$ の収束までの反復回数 T とし，各ノードペアに対して．隣接するノードの類似度を足し合わせるため，$O(T \times |V|^2 \times \bar{d})$ である．ここで，$\bar{d}$ は平均次数を表す．

4.2 SimRank 法

Jeh らの SimRank 法（以下 Simrank 法）は，“共通の隣接ノードを有するノードは互いに類似している” という仮定の下で類似度行列を計算している．Simrank 法による類似度行列の計算法を以下に示す．

ノード u，v 間の類似度 $s(u, v)$ を再帰的に計算する：

$$s(u, v) = \frac{\phi}{|\Gamma(u)| \cdot |\Gamma(v)|} \sum_{i \in \Gamma(u)} \sum_{j \in \Gamma(v)} s(i, j). \tag{5}$$

ここで ϕ は減衰係数であり，$u = v$ の場合は $s(u, v) = 1$ とする．初期値 $\mathbf{S}_0 = \mathbf{I}$ とし，式 (5) を $||\mathbf{S}_t - \mathbf{S}_{t-1}|| < \varepsilon$ となるか，所定の回数まで繰り返し計算することで類似度行

列 $\mathbf{S}$ を得る．

Simrank 法の主たる時間計算量は，行列 $\mathbf{S}$ の収束までの反復回数 T とし，各ノードペアに対して，その隣接ノード数の積のペアの類似度を足し合わせるため，$O(T \times |V|^2 \times \bar{d}^2)$ である．Versim 法より平均次数 $\bar{d}$ 倍計算量がかかることが分かる．

4.3 Random Walk 法

我々の Random Walk 法（以下 Randwalk 法）は，ネットワーク全体でのランダムウォークにより類似経路構造を探索する方法で，PageRank の反復計算時のスコアの収束曲線を特徴ベクトルとし，ベクトル間のコサイン類似度により類似度行列を定義する [16]．ノードの機能，地位，階層や役割は，周辺ノードとの隣接関係，周辺ノードの次数，ネットワーク内での相対的な位置などの影響を受ける．同様にランダムウォークも，任意のノードからスタートし，各ステップでそのノードに到達する期待値を計算している．

Randwalk 法による類似度行列の計算法を以下に示す．行推移確率行列 $\mathbf{P}$ は，各要素を $p(u,v) = a(u,v)/|\Gamma(u)|$ とする．各ノードのランダムウォークにおける到達期待値ベクトル $\mathbf{y}$ は，$y(v) \geq 0$ で $\sum_{v \in V} y(v) = 1$ となる．繰返しステップ数 t を用い，ランダムウォーク期待値ベクトル $\mathbf{y}$ は以下の更新式の極限分布として定義される：

$$\mathbf{y}_t^T = \mathbf{y}_{t-1}^T \mathbf{P} \tag{6}$$

ここで $\mathbf{b}^T$ は $\mathbf{b}$ ベクトルの転置を表す．このモデルは，Web ページのランキングアルゴリズムとして有名な PageRank [17] の大域ジャンプを除いたものと等価である．単一コンポーネントの自己ループ付き無向ネットワークを対象とすれば，推移確率行列 $\mathbf{P}$ は非周期かつ既約であるため，初期ベクトルによらない唯一の最大固有値を有し，極限分布が定常ベクトルに収束することがペロン・フロベニウスの定理により保証される．

また，ノード u に注目すると，

$$\begin{aligned} y_t(u) &= \sum_{v \in \Gamma(u)} y_{t-1}(v) \cdot p(v,u) \\ &= \sum_{v \in \Gamma(u)} \frac{y_{t-1}(v)}{|\Gamma(v)|} \end{aligned} \tag{7}$$

で計算される．ノード u の値の極限値は，ノード u の次数 $|\Gamma(u)|$ により決定される [18]．

$$y_\infty(u) = \frac{|\Gamma(u)|}{\sum_{v \in V} |\Gamma(v)|}. \tag{8}$$

反復を繰り返し，各ノードの値は式 (8) に収束する．$||\mathbf{y}_t - \mathbf{y}_{t-1}|| < \varepsilon$ となるか，所定の回数 T まで繰り返し，各反復回数でのノード u の値を要素としたベクトルを $\mathbf{x}_u = (y_1(u), y_2(u), \ldots, y_T(u))^T$ と定義する．このベクトル $\mathbf{x}_u$ をノード u の収束曲線と呼ぶ．各ノードの収束する値は，各ノードの次数のみで決まるが，一般に収束曲線は次数のみでは決まらない．周辺ノードの影響や周辺ノードとの相対的な位置関係，ネットワーク構造の影響を受ける．Randwalk 法では，初期ベクトル $\mathbf{y}_0 = (1/|V|, \ldots, 1/|V|)^T$ で収束曲線を計算する．

各ノードの収束曲線間のコサイン類似度により，ノード間の類似度を計算する．ノード u と v の類似度 $s(u,v)$ は，

$$s(u,v) = \frac{\mathbf{x}_u^T}{||\mathbf{x}_u||} \cdot \frac{\mathbf{x}_v^T}{||\mathbf{x}_v||} \tag{9}$$

で計算される．コサイン類似度は，ノルムが 1 となるように正規化するため，最終的な収束した値の高低は影響せず，収束までの変化パターンの類似性による．以上のようにしてノード間類似度 $s(u,v)$ を要素とする類似度行列 $\mathbf{S} = [s(u,v)]$ を得る．

Randwalk 法の時間計算量は，PageRank スコアの収束までの反復回数 T とし，PageRank スコア計算に $T \times |E|$，コサイン類似度計算に $T \times |V|^2$ であり，主たる計算量は $O(T \times |V|^2)$ である．Versim 法より平均次数 $\bar{d}$ 倍，Simrank 法より平均次数の 2 乗 $\bar{d}^2$ 倍計算量が少ないことが分かる．**表 1** に各手法の時間計算量をまとめる．

表 1　各種法の時間計算量
Table 1　Time complexities.

手法	Versim 法	Simrank 法	Randwalk 法
時間計算量	$O(T \times \|V\|^2 \times \bar{d})$	$O(T \times \|V\|^2 \times \bar{d}^2)$	$O(T \times \|V\|^2)$

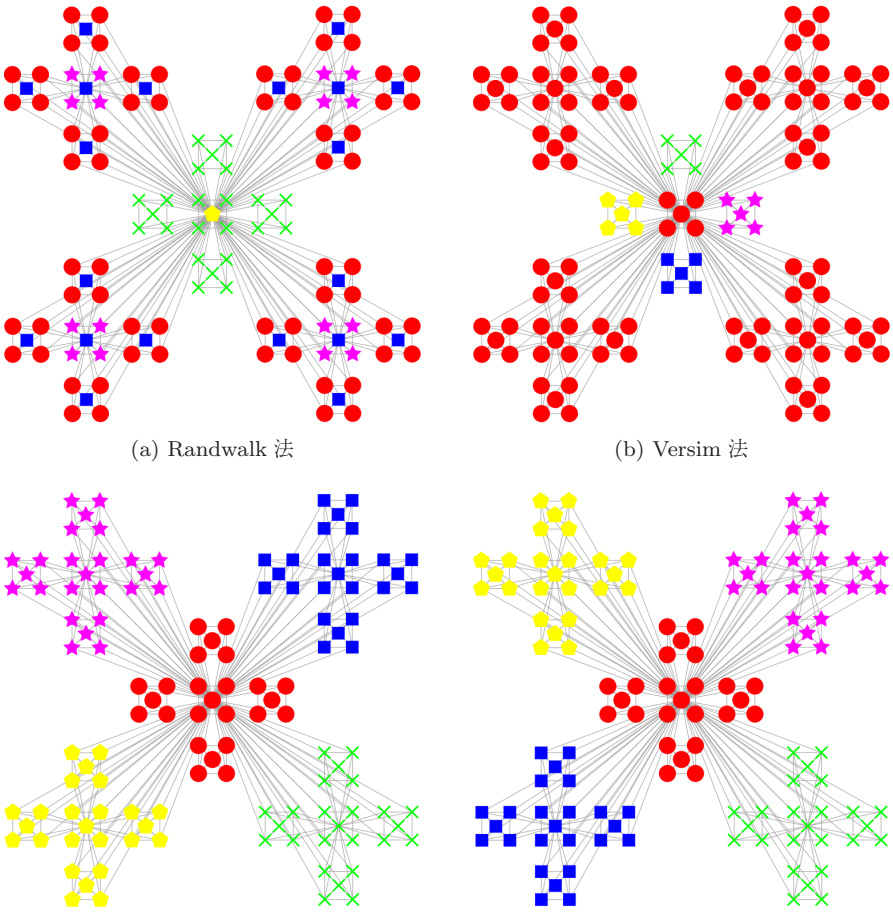
5. 評価実験

本章では，現実の Web ネットワークデータ，ソーシャル・ネットワークおよび人工ネットワークを対象に，3 章で述べた手法によりコミュニティを抽出する．コミュニティ抽出結果を可視化により定性的に，また実行時間により定量的に評価する．本稿ではクラスタリングの方法として，K-median 法を採用する．

5.1 ネットワークデータ

実験では，4 つのネットワークを用いる．

1 つ目のネットワークは，Ravasz らによって提案された階層性のあるネットワークモデルにより生成した人工ネットワークである [19]．階層性のあるネットワークとは，企業内の社員のネットワークや Web サイトのハイパーリンクネットワークのようにトップノードと他のすべてのノード間にはリンクが張られているが，その他のノードどうしは限られた範囲でのみリンクが張られている構造を持っている．すなわちトップノード（社長やトップページほか）は高い次数を有しているが，クラスタ係数が非常に小さい．一方，その他のノード（一般社員や普通のページほか）は低い次数を有しているが，狭い範囲内で密につながっているためクラスタ係数が大きくなる．このような性質を有



(a) Randwalk 法　(b) Versim 法

(c) Simrank 法　(d) Newman 法

図 1　Hierarchical ネットワーク（$K = 5$）．色とマーカの種類は各クラスタを意味し，リンクの長さに本質的な意味はない

Fig. 1　Hierarchical Network ($K = 5$). Each functional community is indicated by a different color marker, and the length of links have no special meanings.

するネットワークを HN モデルにより生成し，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，Lattice ネットワークである．2 次元平面上の正方格子であり，縦に 10，横に 10 でノード数は 100 のネットワークを作る．本稿では Lattice ネットワークと呼ぶ．このネットワークは，上下左右に同じ構造が連続しており，クラスタの判別が難しい事例である．

3 つ目のネットワークは，ネットワーク分析のベンチマークとして広く用いられている，空手クラブ内の友人関係ネットワークである．社会ネットワークの特徴であるスケールフリー性とスモールワールド性を有する [20]．本稿では Karate ネットワークと呼ぶ．

4 つ目のネットワークは，複数の国公立大学のウェブサイト内のページを 2010 年 8 月に収集し，各ウェブサイトのハイパーリンク構造から構築したハイパーリンクネットワークである．本稿ではスペースの都合上，法政大学情報科学部のホームページ[*1] のネットワーク（以下 Hosei ネットワーク）に対する結果を示す．この例は Web サイトの構造を分析するために用いる．

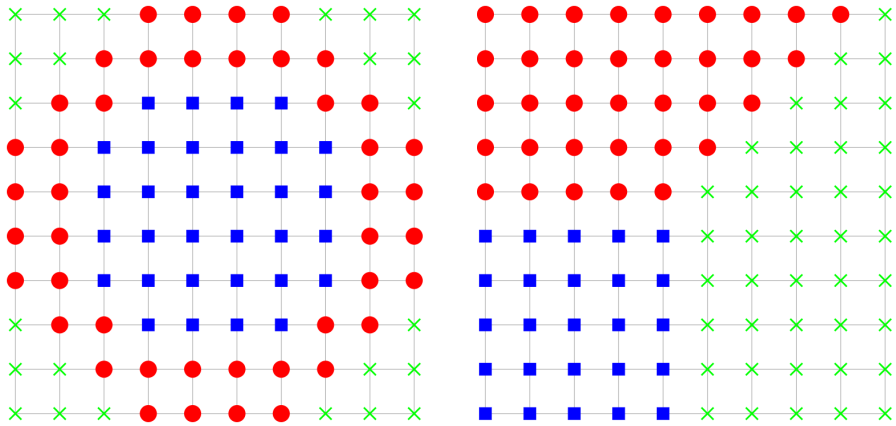
[*1] 法政大学情報科学部 http://cis.k.hosei.ac.jp/

5.2　可視化による定性的評価

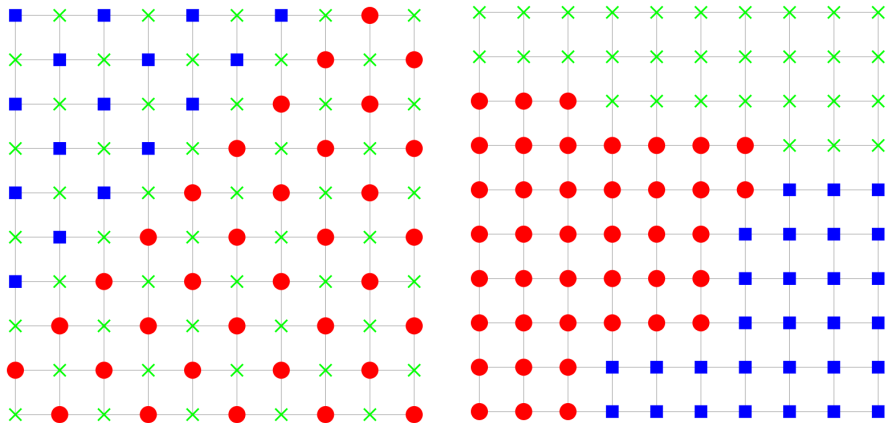
上述したネットワークに対する結果を**図 1**，**図 2**，**図 3** にそれぞれ示す．なお説明の便宜上，適切なクラスタ数 K を図示しているが，他の K の場合でも我々の実験の範囲では，ほぼ同様の結果が得られた．各手法の収束判定は $\varepsilon = 10^{-12}$ とした．Hierarchical ネットワークは文献 [19] に従い，Karate ネットワークおよび Hosei ネットワークはクロスエントロピー法により可視化した [21]．クロスエントロピー法は，ノード間の距離関係ではなく隣接関係によりノード座標を計算しており，可視化結果のリンクの長さに意味はないことに注意する．各可視化結果において，同一の色・マーカのノードは同一のコミュニティに属することを意味する．

5.2.1　Hierarchical ネットワーク

Hierarchical ネットワークの結果（図 1）を比較すると，Randwalk 法 (a) では，階層上の同質（同一階層）のノード，すなわち，同質の機能・役割を持つノード群が同一の



(a) Randwalk 法　(b) Versim 法



(c) Simrank 法　(d) Newman 法

図 2　Lattice ネットワーク（$K = 3$）．色とマーカの種類は各クラスタを意味し，リンクの長さに本質的な意味はない

Fig. 2　Lattice Network ($K = 3$). Each functional community is indicated by a different color marker, and the length of links have no special meanings.

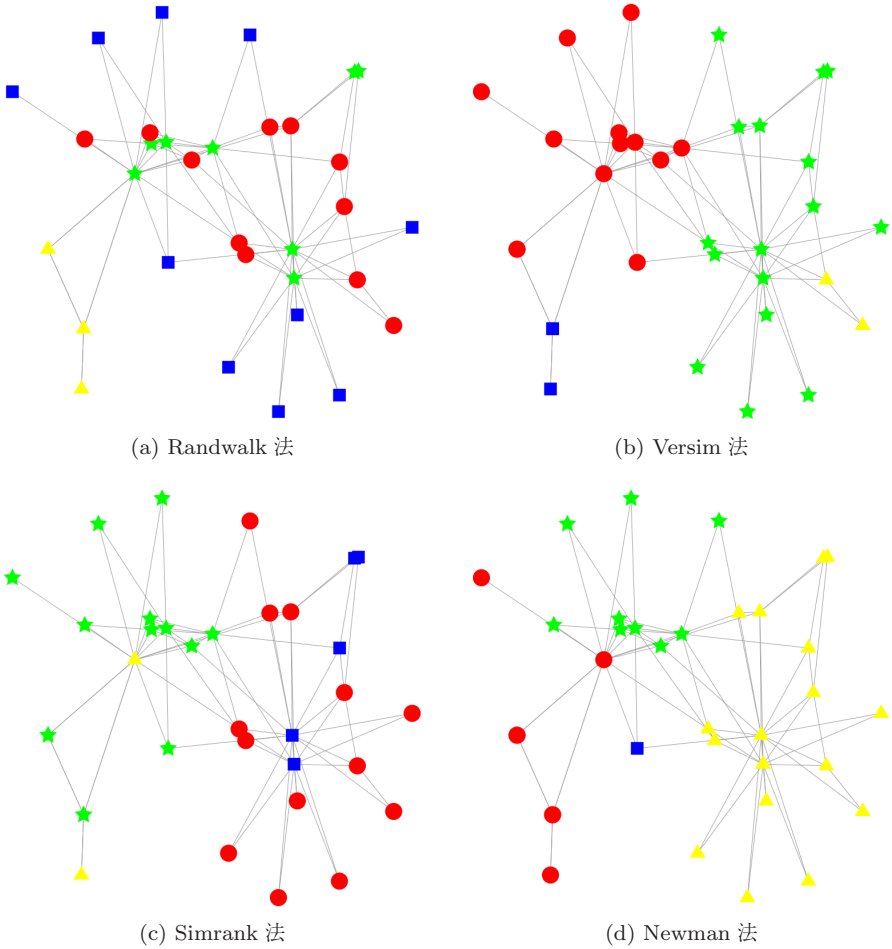
コミュニティとして抽出されている．会社組織でたとえるならば，部長，課長のような役割別にコミュニティが抽出されている．一方 Newman 法 (d) の結果は，リンク密度によるコミュニティ抽出のため，密に隣接するノードどうしが同一のコミュニティとして抽出されている．会社組織でたとえるならば，営業部，人事部のような部門別にコミュニティが抽出されている．Simrank 法 (c) でも同様に，密結合するノード群が同一のコミュニティとして抽出されている．また Versim 法 (b) では，隣接度により結合するノード群（●）と外側の（●）と直接結合しないノード群（●以外）に分割されている．

5.2.2　Lattice ネットワーク

Lattice ネットワークの結果（図 2）を比較すると，Randwalk 法では，ネットワークの全体に対する相対的な位置関係の違いで，中心部（■），末端部（×），中間部（●）に分割されている．一方 Newman 法および Versim 法の結果は，リンク密度や隣接度により近傍ノードを群をまとめて分割している様子がうかがえる．Simrank 法の結果は，共通隣接ノードを有するノードペアの類似度が高くなるため，2 部グラフのような強い周期性（●→×，▲→×）がみられる結果となった．Lattice ネットワークは正方格子で規則正しい構造をしているため，局所的に見るとどの部分も等しい構造をしており，局所的な構造しか考えない Versim 法や Simrank 法は，たとえば同じ機能を果たしているはずの四隅が同一のコミュニティに分類されないように，ノードを立場・役割で分類することができない．Randwalk 法はネットワーク全体を見ているので，全体における立場・役割でノードを分類できている．

5.2.3　Karate ネットワーク

Karate ネットワークの結果（図 3）を比較すると，Randwalk 法では，ハブ的な存在のノード（★），ハブ間の橋渡し的役割かつ互いに結合するノード（●），ハブとだけ友人関係にあるノード（■），ハブとつながりがありかつ小グループを形成するノード（▲）に分類されている．抽出されたコミュニティ内のノードどうしは隣接していない場合もあるが，同様な立場にあり役割・機能が類似したノードを同一のコミュニティとして抽出している．一方 Versim 法や Newman 法では，隣接性を強く考慮しているため，密



(a) Randwalk 法　(b) Versim 法

(c) Simrank 法　(d) Newman 法

図 3　Karate ネットワーク（$K = 4$）．色とマーカの種類は各クラスタを意味し，リンクの長さに本質的な意味はない

Fig. 3　Karate Network ($K = 4$). Each functional community is indicated by a different color marker, and the length of links have no special meanings.

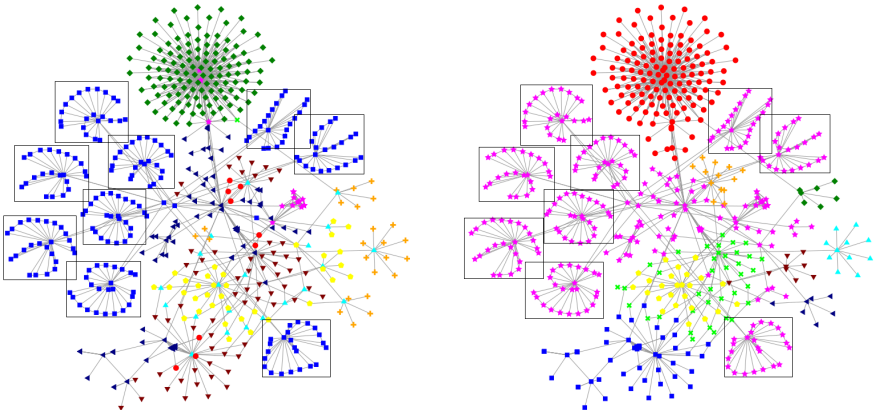
結合するノード群を同一のコミュニティとして抽出している．Simrank 法では，共通隣接ノードを有するノードペアの類似度が高くなるため周期性（▲→★→●→■）がみられる（2 部グラフのように完全に分かれていないため，完全な周期性ではない）．

5.2.4　Hosei ネットワーク

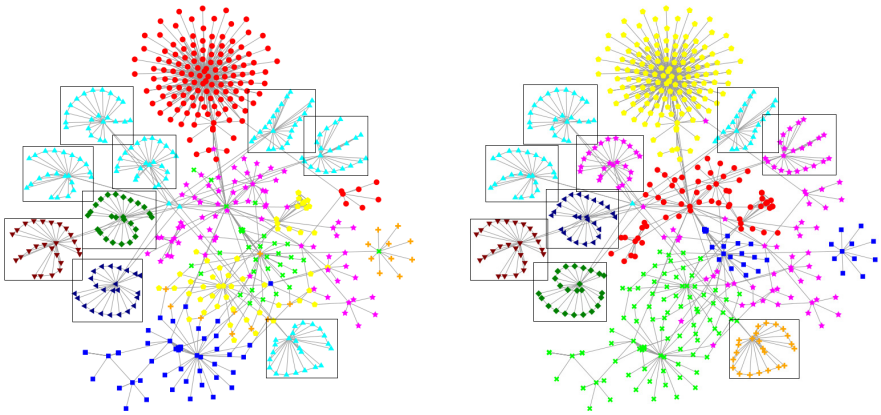
Hosei ネットワークの特徴は，教員の成果報告ページが年度ごとに別のディレクトリにまとめて整理されて公開されていることである．なお，インデックスページからどの年度にもたどれるが，年度間のリンクは存在しない．Hosei ネットワークの結果（図 4）を比較すると，Randwalk 法では，可視化結果の左側部分の 6 つのノード群や右上 2 つ，右下 1 つのノード群のノード（■）は同じコミュニティに分割されている．このノード群は，上述した対象大学の各年度の教員の成果報告ページであり，ノードの機能としては同質であると考えられ，同一のコミュニティとして抽出できている．Versim 法でも同様に，研究成果ページ群（★）を同一コミュニティとして抽出しているが，それらの間にあるページも同一のコミュニティとして抽出している点で Randwalk 法と異なる．これは，定義式から隣接性を強く考慮しているためと考えられる．一方 Newman 法は，各年度の教員研究成果ページ間に直接リンクが存在しないため，異なるコミュニティとして抽出している Simrank 法では，教員研究成果ページ群を同一コミュニティとして抽出しているが，抽出漏れがあることが分かる．

5.2.5　定性的評価のまとめ

これらの異なる構造のネットワークに対する実験結果より，Randwalk 法では，トップノードからの深さやネットワーク内での相対的位置，周辺リンク構造の類似性など，ネットワークに対する機能が類似するノード群をコミュニティとして抽出できることが示された．Versim 法や Simrank 法は，Randwalk 法と近い結果が得られる場合があった．Versim 法は隣接性を強く意識しているため，Randwalk 法で同一と判定されたコミュニティ間をつなぐ部分（異なるコミュニティ）も一緒に抽出される傾向があった．Simrank 法では，直接隣接することより共通隣接ノードを有するかに焦点を当てているため，コミュニティ抽出結果に周期性がみられた．Versim 法や Simrank 法は，



(a) Randwalk 法　(b) Versim 法



(c) Simrank 法　(d) Newman 法

図 4　Hosei ネットワーク（$K = 10$）．年度ごとの教員成果報告ページ群を白抜きの四角で囲っている．色とマーカの種類は各クラスタを意味し，リンクの長さに意味はない

Fig. 4　Hosei Network ($K = 10$). The annual reports pages are surrounded with large transparent squares. Each functional community is indicated by a different color marker, and the length of links have no special meanings.

正則同値なノードを近似的に同定するための手法であるが，局所的なリンク構造の類似性に直接着目することから，ノードの機能を考慮しない Newman 法と近い結果が得られる場合もあった．

また，ある種の類似機能を有するノード群があった場合，それらのノードはある特有の構造を有することが観測できる．しかし，現実ネットワークのように表出する構造にバラつきがある場合，構造だけで機能の類似性を判定することは困難である場合がある．一方，ネットワーク全体でのランダムウォークにおいて，各ノードへの到達確率という期待値の収束曲線の類似性を扱う Randwalk 法では，ある程度のバラつきがある場合においても正しく同定できたと考えられる．

5.3　時間計算量による定量的評価

ノード間類似度に基づく 3 手法を時間計算量の点から定量的に評価する．各手法の収束判定は $\varepsilon = 10^{-12}$ とした．評価実験には，ノード数・リンク数の異なる 6 つのランダムなネットワークを用いる（**表 2**）．理論的には，Versim 法は Randwalk 法の平均次数 $\bar{d}$ 倍，Simrank 法は平均次数の 2 乗 $\bar{d}$ 倍の計算量を必要とする．Versim 法と Simrank 法は収束するまで類似度行列を反復させるが，Randwalk 法は特徴ベクトルが収束するまで反復させるだけなため，反復回数において Randwalk 法が有利となる．また，Versim 法と Simrank 法は局所的なノード間の関係を見ている一方で，Randwalk 法はネットワーク全体を考慮しているが，**図 5** から実際の計算時間を比較すると，Randwalk は他の 2 手法より高速に計算できており，大規模なネットワークに対しても有効な手法であるといえる．

表 2　ネットワークの統計量

Table 2　Network statistics.

	NW1	NW2	NW3	NW4	NW5	NW6
$\|V\|$	100	500	1000	5000	10000	50000
$\|E\|$	500	2500	5000	25000	50000	250000

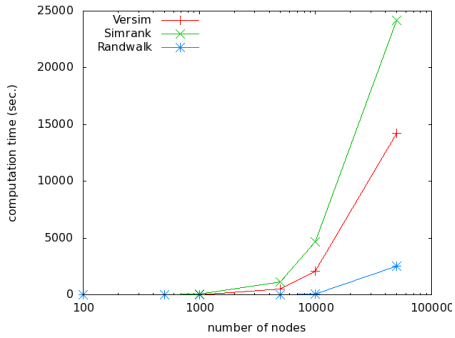


図 5　実行時間
Fig. 5　Computation time.

6. 考察

Randwalk 法は，比較した他の手法と異なり，機能コミュニティをより明示的に抽出できることが分かった．評価実験より，あるノードと類似した機能のノードを有する部分ネットワークを抽出可能であることが示唆された．Randwalk 法における特徴ベクトルの要素は，ランダムウォークの各ステップにおける期待値を意味する．会社組織でたとえるならば，任意の社員からランダムに隣接社員へ E-mail を送信する試行を繰り返したとき，各ステップにおける E-mail を受け取る確率ということができる．周辺社員とのリンクパターンが類似する社員同士は，各ステップで E-mail を受け取る確率の推移が類似すると自然に想定できる．このようにして，Randwalk 法では，機能（性質・役割）の類似するノードを同定できたと考えられる．

ノードの性質として，ソーシャル・ネットワーク分析の分野で有名な中心性指標 [22] や PageRank [17]，HITS [23] などがあげられる．これらの指標は単一尺度であるため，次数や近接度，媒介度，ハブ度といった特定の性質を判定することしかできない．一方，Randwalk 法による機能コミュニティは，機能ごとにクラスタリングすることから，ハブやオーソリティ，ゲートキーパなどのグループを抽出可能であり，多次元的な分析が可能となる．実際に，図 3 (a) におけるハブノード（★）とハブ間の橋渡しノード（●）はともに，HITS ランキング（ハブ度）上位となり分類できない．クラスタ係数に関していうと，ハブノード（★）はきわめて小さく，ハブ間の橋渡しノード（●）は相対的に大きいという違いで Randwalk 法は分類できている．このように，中心性概念にはない性質のノード群も分類することができた．

Randwalk 法は，初期ベクトル $\mathbf{y}_0 = (1/|V|, \ldots, 1/|V|)^T$ により収束曲線を計算している [16]．しかし，ノードの役割や機能は，視点となるノードを変えれば異なるものになるというのが直感的である．すなわち，ノード u に対するノード w の役割は，ノード v に対するノード w の役割とは異なるということである．Randwalk 法は，初期ベクトル $\mathbf{y}_0 = (0, \ldots, 1, \ldots, 0)^T$ のようにノード u に対応する要素のみ 1 でその他の要素は 0 としたベクトルとすることにより，ノード u の視点から，他のノードの機能・役割をクラスタリングすることができる．我々はパーソナライズ機能コミュニティと呼ぶ．収束する値は初期ベクトルに依存しないが，収束曲線のパターンは変化するため，通常の Randwalk 法とは異なる結果が得られる．個々のノードからの視点によるコミュニティ抽出法は，今までにない新たなパラダイムであり，今後の発展性に大いに期待できると考えられる．

7. おわりに

本稿では，従来のリンク密度に基づくコミュニティとは異なり，ノード間の類似性に着目したコミュニティ抽出法に焦点を当てた．ノード間の同値性を近似的に計算する手法として Versim 法と Simrank 法，類似機能を有するノードを同定する Randwalk 法を取り上げ，3 手法に基づいたコミュニティ抽出の結果を可視化により定性的に，時間計算量により定量的に評価した．Versim 法，Simrank 法ともに近似的に正則同値を同定するための手法であるため，Randwalk 法と近い結果が得られる場合があった．しかし Versim 法は，隣接性が強く考慮されているため，Randwalk 法で異なると判定されたコミュニティも同一コミュニティとして抽出する傾向があった．また Versim 法，Simrank 法は，局所的なリンク構造の類似性に直接着目することから，ノードの機能を考慮しない Newman 法と近い結果が得られる場合もあった．Randwalk 法は，特徴ベクトルとして PageRank スコアの収束曲線を用いて，大域的な構造上における現象の類似性を見ており，かつスケーラビリティのある方法である．その結果，局所構造をみる Versim 法や Simrank 法と比較して，全体構造をみる Randwalk 法は同質の機能・役割を有するノード群を要素とするコミュニティをより適切に抽出できることが示された．また，中心性指標とは異なり，一義的な性質の判別ではなく，性質ごとのコミュニティを抽出可能であることを示した．さらに，個々のノードからの視点によるコミュニティ抽出といった，発展性のある手法であることも示唆された．今後は，有向ネットワークや多重ネットワーク，2 部グラフなどの一般的なネットワークを対象としたコミュニティ抽出などへの拡張を検討していくつもりである．

謝辞　本研究は，NTT 未来ねっと研究所との共同研究，および，科研費（23500128）の支援を受けて行ったものである．

参考文献

[1] Newman, M.E.J. and Park, J.: Why social networks are different from other types of networks, *Phys. Rev. E*, Vol.68, No.3, p.036122 (online), DOI: 10.1103/Phys-

RevE.68.036122 (2003).
[2] Newman, M.E.J.: Detecting community structure in networks, *The European Physical Journal B – Condensed Matter and Complex Systems*, Vol.38, No.2, pp.321–330 (online), DOI: 10.1140/epjb/e2004-00124-y (2004).
[3] Shi, J. and Malik, J.: Normalized Cuts and Image Segmentation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.22, No.8, pp.888–905 (2000).
[4] Hagen, L. and Kahng, A.B.: New spectral methods for ratio cut partitioning and clustering, *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, Vol.11, No.9, pp.1074–1085 (online), DOI: 10.1109/43.159993 (1992).
[5] Palla, G., Derényi, I., Farkas, I. and Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society, *Nature*, Vol.435, pp.814–818 (2005).
[6] Saito, K., Yamada, T. and Kazama, K.: The k-Dense Method to Extract Communities from Complex Networks, *Mining Complex Data*, Zighed, D., Tsumoto, S., Ras, Z. and Hacid, H. (Eds.), Studies in Computational Intelligence, Vol.165, pp.243–257, Springer Berlin/Heidelberg (2009).
[7] Seidman, S.B.: Network structure and minimum degree, *Social Networks*, Vol.5, No.3, pp.269–287 (online), DOI: 10.1016/0378-8733(83)90028-X (1983).
[8] 風間一洋，佐藤進也，斉藤和巳，山田武士：人間関係の重なりを持つコミュニティ構造の抽出（特集ネットワークが創発する知能），コンピュータソフトウェア，Vol.24, No.1, pp.81–90 (2007-01-26).
[9] Borgatti, S.: Two algorithms for computing regular equivalence, *Social Networks*, Vol.15, No.4, pp.361–376 (1993).
[10] Leicht, E.A., Holme, P. and Newman, M.E.J.: Vertex similarity in networks, *Physical Review E*, Vol.73, No.2, pp.1–10 (2005).
[11] Jeh, G. and Widom, J.: SimRank: A measure of structural-context similarity, *Proc. 8th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '02*, pp.538–543, ACM (2002).
[12] Smola, A.J. and Kondor, R.: Kernels and Regularization on Graphs, *Machine Learning*, Vol.2777, No.212938, pp.1–15 (2003).
[13] Higham, N.J.: The scaling and squaring method for the matrix exponential revisited, *SIAM J. Matrix Anal. Appl*, Vol.26, p.2005 (2005).
[14] Christakis, N.A. and Fowler, J.H.: *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*, Little, Brown and Company (2009).
[15] Borgatti, S.P. and Everett, M.G.: Notions of Position in Social Network Analysis, *Sociological Methodology*, Vol.22, No.1992, pp.1–35 (online), available from ⟨http://www.jstor.org/stable/270991?origin=crossref⟩ (1992).
[16] 伏見卓恭，斉藤和巳，風間一洋：ネットワーク機能コミュニティ抽出法，日本データベース学会論文誌，Vol.10, No.3, pp.13–18 (2012).
[17] Langville, A.N. and Meyer, C.D.: Deeper inside pagerank, *Internet Mathematics*, Vol.1, No.3, pp.335–380 (2004).
[18] Even-Dar, E. and Shapira, A.: A Note on Maximizing the Spread of Influence in Social Networks, *Internet and Network Economics*, Deng, X. and Graham, F. (Eds.), Lecture Notes in Computer Science, Vol.4858, pp.281–286, Springer Berlin/Heidelberg (2007).
[19] Ravasz, E. and Barabási, A.L.: Hierarchical organization in complex networks, *Physical Review E*, Vol.67, No.2, pp.026112+ (online), DOI: 10.1103/PhysRevE.67.026112 (2003).
[20] Zachary, W.: An information flow model for conflict and fission in small groups, *Journal of Anthropological Research*, Vol.33, pp.452–473 (1977).
[21] Yamada, T., Saito, K. and Ueda, N.: Cross-entropy directed embedding of network data, *Proc. 20th International Conference on Machine Learning* (*ICML03*), pp.832–839 (2003).
[22] Freeman, L.: Centrality in social networks: Conceptual clarification, *Social Networks*, Vol.1, No.3, pp.215–239 (online), DOI: 10.1016/0378-8733(78)90021-7 (1979).
[23] Kleinberg, J.M.: Authoritative sources in a hyperlinked environment, *J. ACM*, Vol.46, pp.604–632 (1999).

伏見 卓恭

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在学中．2011 静岡県立大学大学院経営情報学研究科修士課程修了．複雑ネットワークの研究に従事．電子情報通信学会，日本データベース学会，人工知能学会各学生会員．

斉藤 和巳　（正会員）

静岡県立大学経営情報学部教授．1985 慶応義塾大学理工学部数理科学科数学専攻卒業，1998 東京大学博士（工学）．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，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．著書に「ウェブサイエンス入門—インターネットの構造を解き明かす」（NTT 出版）．

風間 一洋　（正会員）

NTT 未来ねっと研究所主任研究員．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．同年日本電信電話（株）入社．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．博士（情報学）．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．

（担当編集委員　小山 聡）