

PageRank 収束曲線を用いたコミュニティ特性の定量化

Quantifying Characteristics of Communities using PageRank Convergence Curves

伏見 卓恭^{*1*2}

Takayasu FUSHIMI

斉藤 和巳^{*3}

Kazumi SAITO

風間 一洋^{*4}

Kazuhiro KAZAMA

佐藤 哲司^{*1}

Tetsuji SATOH

^{*1}筑波大学 図書館情報メディア系

Faculty of Library, Information and Media Science, University of Tsukuba

^{*2}日本学術振興会特別研究員

JSPS Research Fellow

^{*3}静岡県立大学 経営情報学部

School of Management and Information, University of Shizuoka

^{*4}和歌山大学 システム工学部

Faculty of Systems Engineering, Wakayama University

In this paper, we attempt to represent characteristics of communities in vector space. We have proposed a novel method in order to represent node's function which is designed by the convergence curve of the PageRank score of a node. When we try to represent characteristics of arbitrary node groups, it is natural to construct a synthetic vector of the convergence curves of these nodes belonging to the community. From experimental evaluation using several networks, we confirm that the synthetic vectors represent characteristics of communities.

1. はじめに

Facebook や Twitter などの SNS や口コミサイト上のユーザ関係など、今やソーシャルネットワークは Web 上で盛んにみられるようになってきている。これらを対象とした研究が注目され、ネットワーク成長過程やネットワーク上での現象のメカニズムが明らかにされつつある。こうした多くの知見を得るためには、大規模なノード群とそれらの間を複雑につなぎ合わせるリンク群から形成されるネットワーク構造を分析することから始まる。しかし、このようなネットワークを対象に分析するには、複雑性と大規模性のため困難な場合がある。それゆえ、重要なノードを抽出したり、ノード群を類似性に基づきクラスタリングするなどして、ネットワーク構造を要約し分析することが望まれる。

ネットワーク構造からリンクが密なサブネットワークを抽出するコミュニティ抽出が広く研究されている。密につながるノード群は、同じ属性や傾向をもつ集団であることが多いため有用な情報となる。よく知られたコミュニティ抽出手法として、Clauset らによる 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 [Clauset 04]。この手法は、ノード同士の結合が疎な部分を切断し、いくつかのサブネットワークに分割する方法である。また、ネットワーク上でのノード同士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) の条件を緩めたサブネットワークを見つけるための手法も提案されている [Palla 05]。これらの手法は、リンク構造の粗密に着目し、ネットワークをいくつかのサブネットワークに分割する。しかしこれらのコミュニティの概念は、局所的な類似性に基づくため、大規模・複雑な場合はさらに大局的な見地から見るという新たなステップが必要になる。

一方、著者らは、各ノードの特性を機能ベクトルと呼ぶベクトルで表現し、ベクトル間の類似度に基づきクラスタリングすることで、ネットワーク内に点在する機能的類似ノード群を抽出する手法を提案した [伏見 12]。この手法は、ネット

ワーク上でのランダムウォークに着目し、パワー法を利用した PageRank スコア計算の各ステップで得られるスコアを要素としたベクトルを用いる。複数のネットワークを用いた評価実験により、ネットワーク構造から窺い知ることのできる「機能」が類似するノード群を同定できることを検証した。この手法の自然な拡張として、複数ノードの合成ベクトルを構築することにより、対象ノード群の特性をベクトル表現することがあげられる。前述したように、機能ベクトルの各次元は収束までの PageRank スコアを意味する。PageRank スコアは加算性が成り立つので、各次元の和をとること、すなわち、合成ベクトルを構築することは自然である。全ノードの機能ベクトルを合成すると、PageRank スコアの性質上、すべての要素の値が 1 であるベクトルとなるが、任意のノード群の合成ベクトルはそうはならず、多様なベクトルが得られる。本稿では、ネットワークのコミュニティに着目し、コミュニティ内のノードの機能ベクトルを合成することで、コミュニティのどのような特性を表現できるかを検証する。

2. 提案手法

この章では、各コミュニティの特性をベクトル表現する方法を説明する。ノード集合 V とリンク集合 E からなるネットワーク $G = (V, E)$ が与えられ、各ノードは H 個のコミュニティ $C = \{C_1, \dots, C_H\}$ のうちいずれかに所属している。

$$V = \bigcup_{h=1}^H C_h, \quad C_h \cap C_k = \emptyset, \quad h \neq k$$

以下の手順で、コミュニティ特性を表すベクトルを構築する。

1. パワー法により PageRank スコアベクトル群を $\{y_1, \dots, y_S\}$ 構築;
2. 各ノード $u \in V$ に対して、機能ベクトル $x_u = (y_1(u), \dots, y_S(u))^T$ を計算;
3. 各コミュニティ $C \in \mathcal{C}$ に対して、所属するノードの機能ベクトルを合成;

連絡先: 伏見卓恭, 筑波大学図書館情報メディア系, 茨城県つくば市春日 1-2, 029-859-1391

2.1 機能ベクトル

文献 [伏見 12] において、各ノードの機能・役割はネットワーク構造に埋め込まれていると仮定している。ノードの機能として、ネットワーク内での階層的地位や相対的な位置、次数や周辺ノードの次数、周辺ノードとのつながり方などを意図しているが、これらが類似するノードどうしは、PageRank スコアの収束過程も類似すると推測できる。従って、ネットワーク構造上でのランダムウォークのモデルである PageRank [Langville 04] を用いて、各ノードの機能を表す特徴ベクトルを計算する。

以下に、ネットワーク構造から窺い知ることのできるノードの機能を表す収束曲線の計算法を以下に示す。収束曲線は、大域ジャンプなしの PageRank を用いて計算する。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。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。ここで、行推移確率行列 \mathbf{P} の各要素を $p(u, v) = a(u, v)/|\Gamma(u)|$ とする。各ノードの PageRank スコアを要素としたベクトル \mathbf{y} は、 $y(v) \geq 0$ で $\sum_{v \in V} y(v) = 1$ となる。初期ベクトルを $\mathbf{y}_0 = (1/|V|, \dots, 1/|V|)^T$ とし、繰り返しステップ数 s を用い、PageRank スコアベクトル \mathbf{y} は以下の更新式の極限分布として定義される：

$$\mathbf{y}_s^T = \mathbf{y}_{s-1}^T \mathbf{P} \quad (1)$$

ここで \mathbf{b}^T は \mathbf{b} ベクトルの転置を表わす。式 1 は、推移確率行列 \mathbf{P} の左固有ベクトルをパワー法により求めていることに等しい。また、ノード u に注目すると、

$$\begin{aligned} y_s(u) &= \sum_{v \in \Gamma(u)} y_{s-1}(v) \cdot p(v, u) \\ &= \sum_{v \in \Gamma(u)} \frac{y_{s-1}(v)}{|\Gamma(v)|} \end{aligned} \quad (2)$$

で計算される。ノード u の値の極限值は、ノード u の次数 $|\Gamma(u)|$ により決定される。

$$y_\infty(u) = \frac{|\Gamma(u)|}{\sum_{v \in V} |\Gamma(v)|}. \quad (3)$$

このことは、式 3 を式 2 の $y_{s-1}(u)$ に代入すると、

$$\begin{aligned} y_s(u) &= \sum_{v \in \Gamma(u)} \left\{ \frac{1}{|\Gamma(v)|} \cdot \frac{|\Gamma(v)|}{\sum_{w \in V} |\Gamma(w)|} \right\} \\ &= \frac{|\Gamma(u)|}{\sum_{w \in V} |\Gamma(w)|} \end{aligned} \quad (4)$$

となり、明らかである。反復を繰り返し、各ノードの値は式 3 に収束する。パワー法の反復計算を所定の回数 S まで繰り返し、各反復回数でのノード u のスコアを要素としたベクトルを

$$\mathbf{x}_u = (y_1(u), y_2(u), \dots, y_S(u))^T \quad (5)$$

と定義する。このベクトル \mathbf{x}_u をノード u の機能ベクトル（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。

2.2 機能ベクトルの合成

各ノードの機能ベクトルの要素は、PageRank スコアを表す。従って、機能ベクトルの任意の要素（次元） s に対する和は PageRank スコアの和であるため、全ノードでの和は必ず $\sum_{u \in V} x_u(s) = 1$ となる。

さらに、PageRank スコアは加算性が成り立つため、任意のノード群の PageRank スコア和は、そのノード群の PageRank スコアといえる。ゆえに、機能ベクトルの合成ベクトルは、対象ノード群の PageRank スコアの推移を表すとも言える。コミュニティ C に対し、所属するノード群の機能ベクトルの合成ベクトルを以下のように計算する。

$$\mathbf{z}_C = \sum_{u \in C} \mathbf{x}_u \quad (6)$$

この合成ベクトルにより、対象コミュニティの特性を定量化する。

3. 評価実験

3.1 ネットワークデータ

評価に用いるネットワークについて述べる。1 つ目のネットワークは、1 つのハブノードと 8 個の周辺ノード（以下、リーフノードと呼ぶ）からなるコミュニティが 4 つ隣接するネットワークである。各コミュニティにおいて、リーフノードはそれぞれ 0 近傍、1 近傍、2 近傍、3 近傍ノードと隣接する構造である。本稿では人工ネットワークと呼ぶ。

2 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、複雑ネットワーク研究分野の共著関係ネットワークである [Newman 06]。ノード数は 379、リンク数は 914 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。本稿では Coauthor ネットワークと呼ぶ。

3 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 Hyperlink ネットワークと呼ぶ。

3.2 実験結果

本稿では、CNM 法 [Clauset 04] により、対象ネットワークを H 個のコミュニティに分割する。コミュニティ数 H は、CNM 法によりモジュラリティが最大となるような H を選んだ。また、図 1(a),(b),(c) は各ネットワークの可視化結果であり、コミュニティごとに色を分けてノードを着色した。図 2, 3, 4 は各ネットワークの対応するコミュニティの合成ベクトルを表す。

図 2 に、人工ネットワークに対する処理結果を示す。この人工ネットワークの構造はコミュニティ内部の構造は異なるが、コミュニティ間の関係は均等であることを念頭に置いておく。多少の変動はあるものの、大きく分けて減少傾向か増加傾向の 2 種類のコミュニティが存在する。初期の値は 0.25 とほぼ均等であるが、相対的にリンク密度の高いコミュニティは、コミュニティ内部でスコアを貯めやすく、増加傾向にある。逆に、相対的に密度が低いコミュニティは減少傾向にある。さらに、ハブノードとリーフノードの次数差が大きいコミュニティの合成ベクトルは、初期の振幅が大きくなるのも見て取れる。

図 3 に、Coauthor ネットワークに対する処理結果を示す。減少傾向と増加傾向に加え、増加した後に減少する「上に凸」

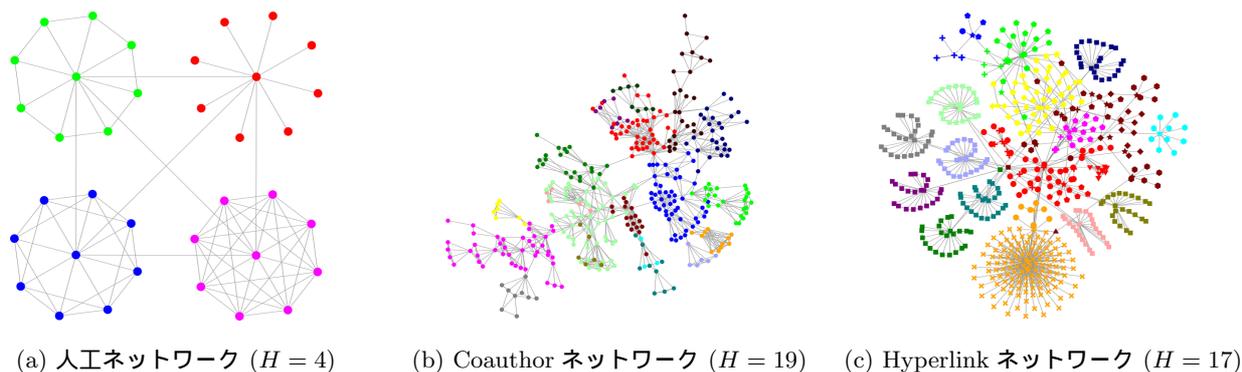


図 1: 対象ネットワークの可視化結果

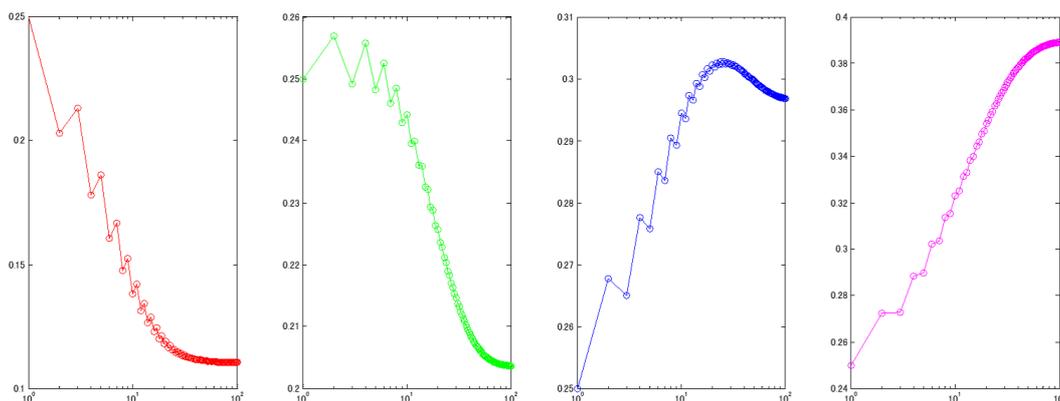


図 2: 人工ネットワークの合成ベクトル群

型のコミュニティ，減少した後に増加する「下に凸」型のコミュニティも見受けられる．減少傾向のコミュニティの共通特徴として，他のコミュニティとあまり接続関係になく，いわゆる，ぶら下がりコミュニティのものが多い．増加傾向のコミュニティの共通特徴として，多くのコミュニティと接続関係にあり，可視化結果でも比較的中心に配置されている．上に凸型のコミュニティの共通特徴として，コミュニティ内部と外部のパイプが複数あることがあげられる．つながるコミュニティの影響を受けるタイミングが異なるため，単調性がなく，中盤で盛り上がる点が特徴的である．

図 4 に，Hyperlink ネットワークに対する処理結果を示す．Coauthor ネットワークとほぼ同様で，増加傾向，減少傾向，上に凸，下に凸などのコミュニティが見受けられた．Coauthor ネットワークと比べて，初期の振幅が比較的大きめなコミュニティが存在するが，共著関係よりハイパーリンク構造の方がノード間の次数差が大きくなる傾向にあるため，その点を反映した結果と考えられる．

図示はしていないが，これらのネットワークにおいて，多くのノードの機能ベクトルも属するコミュニティの合成ベクトルと類似の傾向を示している．すなわち，ベクトルの大域的な傾向はコミュニティそのものの特性を反映し，初期に発生する振幅などがコミュニティ内のローカルな構造を反映していると考えられる．

4. おわりに

本研究では，各ノードの機能を表すベクトルをコミュニティごとに合成することで，コミュニティの特性を表すベクトルを構築し，評価実験によりその有効性を検証した．今後はさらに

多様なネットワークを用いて，提案手法の有効性を検証していきたい．

謝辞 本研究は，科学研究費補助金基盤研究 (B)(No.25280110) 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, p. 066111 (6 pages) (2004)
- [Langville 04] Langville, A. N. and Meyer, C. D.: Deeper inside pagerank, *Internet Mathematics*, Vol. 1, pp. 335–380 (2004)
- [Newman 06] Newman, M. E. J.: Finding community structure in networks using the eigenvectors of matrices, *Physical Review E*, Vol. 74, No. 3, p. 036104 (22 pages) (2006)
- [Palla 05] 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)
- [伏見 12] 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)

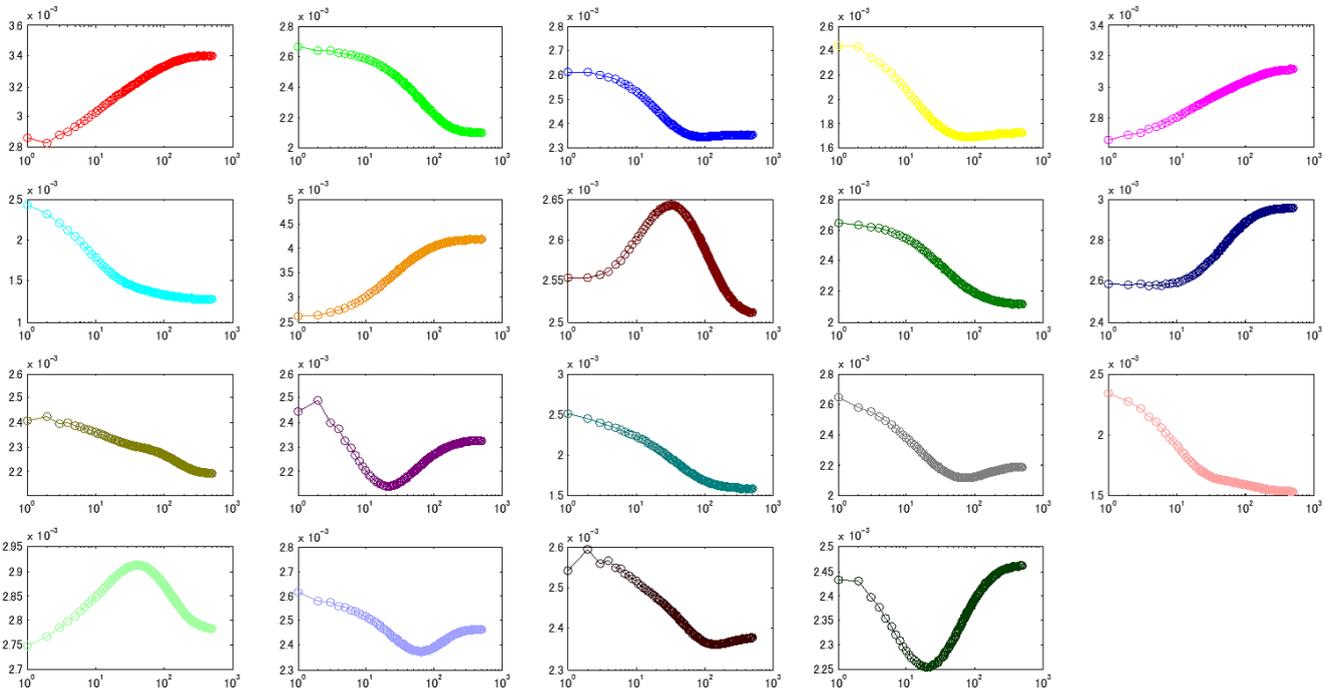


図 3: Coauthor ネットワークの合成ベクトル群

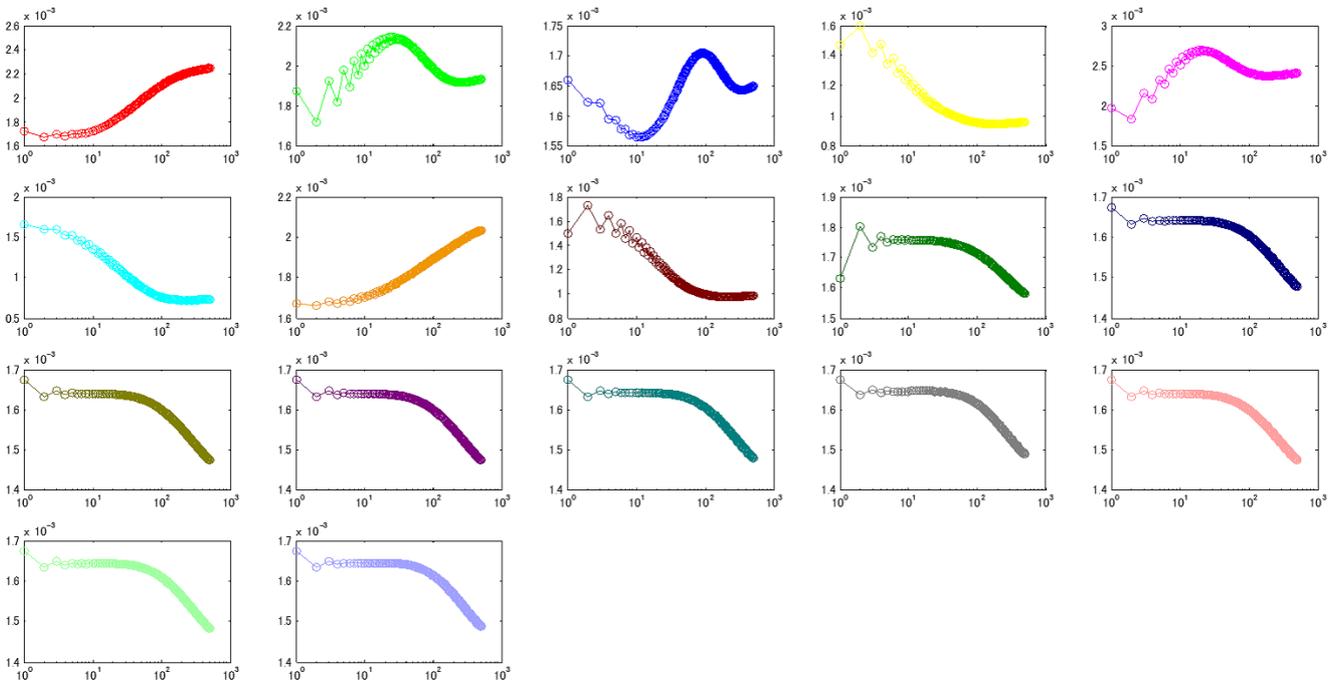


図 4: Hyperlink ネットワークの合成ベクトル群