2J1-5

情報拡散影響度に基づく機能コミュニティ抽出法

Functional Community Extraction Method Based on Influence Degree of Information Diffusion

伏見 卓恭[*1] 斉藤 和巳[*1] 池田 哲夫[*1] 風間 一洋[*2]
Takayasu FUSHIMI Kazumi SAITO Tetsuo IKEDA Kazuhiro KAZAMA

[*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

[*2]和歌山大学　システム工学部
Faculty of Systems Engineering, Wakayama University

In this paper, we propose a method for extracting functionally similar nodes in information diffusion processes. A conventional method for extracting functional communities bases on random walk processes over a network, and calculates the similarities of PageRank score convergence curves which corresponding to the function of each node. From our experimental results using three networks, our method can extract functionally similar nodes each of which distant from important nodes such as hub nodes through a perspective of information diffusion, and also extract highly clustered hub nodes which can not be distinguish from other nodes by the conventional method.

1. はじめに

SNS やレビューサイト内でのユーザ間のつながりなど，複雑ネットワークがあらゆる場面で見受けられるようになり，それらを対象とした分析が盛んに行われている．また，近年ではソーシャルグラフ上でのクチコミ（情報拡散）を利用した情報推薦や広告などの技術が多くみられる．このように，身近に存在するネットワーク上での情報拡散現象を分析・モデル化・予測することで，Web マーケティングなどへの貢献が期待できる．

ネットワーク上での情報拡散過程において，発信された情報をすぐに受信できるノードもいれば，拡散過程の終わりまで情報を受信できないノードもいる．さらに，一般に情報拡散の経路は複数考えられ，多くノードとつながるノードは情報が集まる可能性も高くなる．このように，ネットワーク構造などの様々な要因が重なり合い，情報受信の期待値は各ノードにより異なる．情報受信の期待値が類似するノードは，ネットワーク内で類似の機能・役割を有すると考えられる [Christakis 09]．

類似機能を有するノードを抽出する手法として，機能コミュニティ抽出法がある [伏見 12a, 伏見 12b]．この手法は，ネットワーク上でのランダムウォークにおける各タイムステップの期待値の変化の類似性によりノードを分類する．各タイムステップでの期待値変化が類似するノードはネットワーク内で類似の立場にあると仮定し，相対的位置や階層的地位などネットワーク内での大域的機能によりノードを分類する．

しかし，現実のソーシャルネットワークを対象とする場合，ランダムウォークモデルに基づき抽出された機能コミュニティは，抽出されたノードがもつさまざまな属性を考慮しないと具体的な機能性が判明しないことから，その意味付けが困難な場合がある．本稿では，情報拡散モデルに基づき，情報受信の期待値変化の類似性によりノードを分類する手法を提案する．複数のネットワークを用いた評価実験より，従来法では区別が困難なノードを抽出可能なことなどを示す．

連絡先: 伏見卓恭，静岡県立大学経営情報イノベーション研究科，静岡県静岡市駿河区谷田５２－１，054-264-5436，j11507@u-shizuoka-ken.ac.jp

2. 情報拡散機能コミュニティ抽出法

代表的な情報拡散モデルを用い，情報受信の期待値変化の類似性に基づきノードを分類する提案法について説明する．

提案法は，ノード集合 V，リンク集合 E からなるネットワーク $G = (V, E)$ とコミュニティ数 K を入力とし，以下のようなアルゴリズムにより情報拡散機能コミュニティを抽出する．

1. 情報拡散モデルにより，各タイムステップでの情報受信期待値ベクトル $\{\mathbf{y}_1, \cdots, \mathbf{y}_S\}$ を計算;
2. 各ノードの機能ベクトルとして期待値変化ベクトル $\mathbf{x}_v$ を構築;
3. 各ノードペアの機能ベクトル $\mathbf{x}_u$ と $\mathbf{x}_v$ のコサイン類似度 $\rho(u, v)$ を計算;
4. K-median 法により全ノードを K 個のグループに分割;
5. 情報拡散機能コミュニティ $\{\mathcal{C}_1, \cdots, \mathcal{C}_K\}$ を出力;

以下に，アルゴリズムの主要技術に関する詳細を説明する．

2.1 情報拡散モデル

情報拡散過程において，情報を受信した状態をアクティブな状態，それ以外の状態を非アクティブな状態と定義する，本稿では離散同期型かつ SIR 型のモデルを用いる．すなわち，情報拡散過程は，離散時刻 $t \geq 0$ で展開し，一度アクティブになったノードは二度と非アクティブにはならない．提案法で用いる代表的な情報拡散モデルである，独立カスケード（Independent Cascade，以下 IC モデル）モデルと線形閾値（Linear Threshold，以下 LT モデル）モデルについて説明する [Kempe 03, Kimura 10]．

IC モデルでは，各リンク (u, v) に対して，実数値 $p_{u,v} \in [0, 1]$ を前もって指定する．ここで $p_{u,v}$ はリンク (u, v) を通しての拡散確率である．IC モデルの情報拡散過程は，アクティブノードの初期集合 A が与えられたとき，次のように進んでいく．ノード u が時刻 s でアクティブになったとき， u は非

アクティブな子ノード v をアクティブにする唯一の機会が与えられ，その試行は確率 $p_{u,v}$ で成功する．成功した場合，ノード v は時刻 $s+1$ でアクティブになる．この成功・失敗に関わらず，ノード u がノード v をアクティブにする機会はこの時のみである．ネットワーク中で試行対象となるノードがなくなったとき，情報拡散過程は終了する．

LT モデルでは，任意のノード $v \in V$ に対して，その親ノード u からの重み $w_{u,v}(>0)$ を，$\sum_{u\in\Gamma(v)} w_{u,v} \leq 1$ となるように，前もって指定する．各ノードの v の閾値 θ_v を区間 $[0,1]$ から一様ランダムに設定する．LT モデルの情報拡散過程は，アクティブノードの初期集合 A が与えられたとき，次のように決定論的に進んでいく．時刻 s で非アクティブなノード v は，時刻 s でアクティブな親ノード u から重み $w_{u,v}$ の影響を受ける．アクティブな親ノードからの重みの和が，自身の閾値を超えたときノード v はアクティブになる．すなわち，時刻 s でアクティブなノード v の親ノード集合を $\Gamma_t(v)$ とすると，$\sum_{u\in\Gamma_t(v)} w_{u,v} \geq \theta_v$ であれば，ノード v は時刻 $s+1$ でアクティブになる．ネットワーク中で試行対象となるノードがなくなったとき，情報拡散過程は終了する．

2.2 情報受信期待値変化ベクトルの計算

上述した情報拡散モデルを用いたシミュレーションにより，各離散時刻で各ノード v がアクティブになる期待値を計算する．情報源ノードを u とした情報拡散過程の時刻 s でアクティブになったノード集合を $A_s(u)$ としたとき，

$$f_s^u(v) = \begin{cases} 1 & \text{if } \ v \in A_s(u) \\ 0 & \text{otherwise} \end{cases}$$

のように $\mathbf{f}_s^u$ を構築する．そして，各ノードを情報源とした際の期待値を計算する：

$$y_s(v) = \frac{1}{|V|} \sum_{u\in V} f_s^u(v). \tag{1}$$

ここで，これらを要素とするベクトル $\mathbf{y}_s$ を時刻 s での情報受信期待値ベクトルと呼ぶ．シミュレーションを M 回実施した場合，各シミュレーションで得られる情報受信期待値ベクトルを平均して期待値を計算する．

次に，各ノードの機能を表す期待値変化ベクトルを以下のように構築する．ノード v の機能ベクトルである期待値変化ベクトルは，

$$\mathbf{x}_v = \{y_1(v), \ldots, y_S(v)\} \tag{2}$$

と定義する．各ノードの期待値変化ベクトル間の類似度に基づき，ノードをクラスタリングする．

3. 評価実験

提案法により抽出されたコミュニティの性質を評価するために，前述したランダムウォークに基づく機能コミュニティ抽出法と比較する．以下ランダムウォークモデル，IC モデル，LT モデルに基づく機能コミュニティ抽出法を，それぞれ RWFC 法，ICFC 法，LTFC 法と呼ぶ．

3.1 ネットワークデータ

1 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．本稿では Hosei ネットワークと呼ぶ[*1]．

2 つ目のネットワークは，Hosei ネットワークと同様に，大学のウェブサイト内のページを 2012 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ものである．本稿では Yamaguchi ネットワークと呼ぶ[*2]．

3 つ目のネットワークは，ブログのトラックバックネットワークである．2005 年 5 月に“ goo ”[*3]というサイトの「JR 福知山線脱線事故」というテーマからトラックバックを 10 段辿ることにより収集し無向化した．本稿では Trackback ネットワークと呼ぶ．

3.2 実験設定

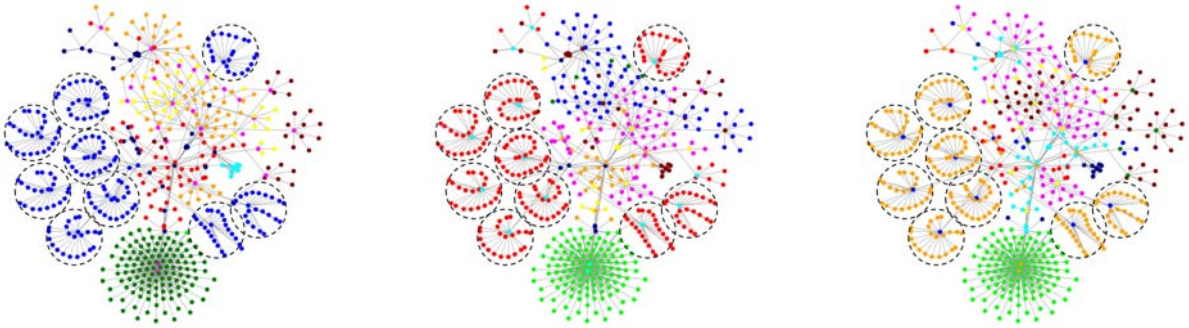
提案法において，シミュレーション回数を $M = 10{,}000$，任意のリンク (u, v) における拡散確率を各ネットワークの平均次数の逆数 $p_{u,v} = p \simeq 0.5 \cdot |V|/|E|$ と設定する．各ノードが平均して 1 ノード程度をアクティブ化できるように設定した．コミュニティ数 K を $2 \leq K \leq 10$ の範囲で最も適切な値を選んだ時の抽出結果を図 1 から図 3 に示す．なお他の K の場合でも本稿の主張と大きな矛盾のない結果が得られている．可視化にはノード間の隣接関係が視認できるように，クロスエントロピー法を用いて [Yamada 03]，コミュニティごとにノードを異なる色で着色した．

3.3 実験結果

図 1 は Hosei ネットワークの処理結果である．RWFC 法の結果を見ると，図中下部のシラバスページ群や，点線で囲っている各年度の成果報告ページなど，同一の機能を有するノードが同一のコミュニティとして抽出されていることがわかる．ICFC 法，LTFC 法の結果を見ると，全体の傾向は RWFC 法の結果と類似している．しかし，点線で囲っている各年度の成果報告ページに関しては，RWFC 法では中心のインデックスページとその周辺の各教員の成果ページの区別ができていないが，ICFC 法，LTFC 法では情報拡散の点で重要ノードとなるハブノードとその他のノードを区別できている．ランダムウォークモデルにおいては，共通隣接ノードを有するクラスタ（三角形）を構成するノードを識別できないため [Gfeller 07]，インデックスページと連続する 2 つの教員成果ページが作り出すクラスタにおいて中心となるハブノード（インデックスページ）が同一のコミュニティと判定されると考えられる．情報拡散モデルにおいて各ノードがアクティブになるタイミングは，多くのノードへ情報を拡散させるハブノードからの距離が大きな要因であり，クラスタ係数は大きな影響はないと考えられる．

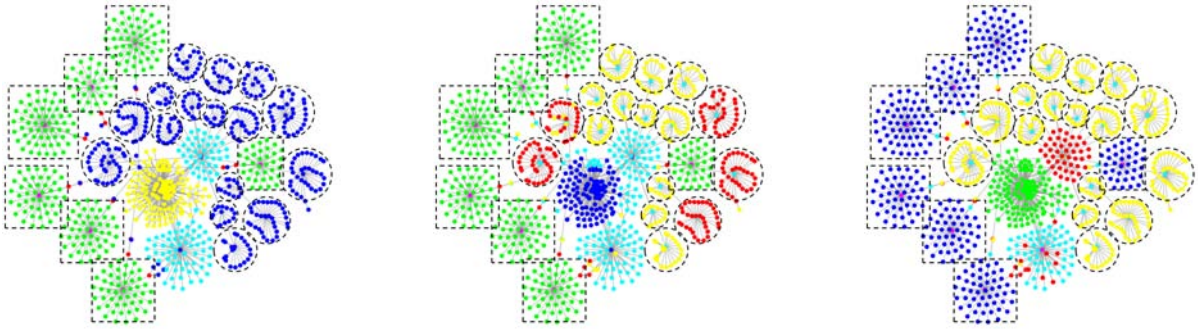
図 2 は Yamaguchi ネットワークの処理結果である．RWFC 法の結果を見ると，中心部分のプレスリリースページや，点線で丸く囲っている月報の広報ページ，点線で四角く囲っている大学歴史紹介ページなど，同一の機能を有するノードが同一のコミュニティとして抽出されている．ICFC 法，LTFC 法の結果を見ると，全体の傾向は RWFC 法の結果と類似している．しかし，点線で丸く囲っている月報広報ページに関しては，RWFC 法では中心のインデックスページとその周辺のコンテンツページの区別ができていないが，ICFC 法，LTFC 法では情報拡散の点で重要ノードとなるハブノードとその他のノードを区別できている．一方，同じハブノードでもクラスタ

∗1 法政大学情報科学部 http://cis.k.hosei.ac.jp/
∗2 山口大学 http://www.yamaguchi-u.ac.jp/
∗3 `http://blog.goo.ne.jp/usertheme/`



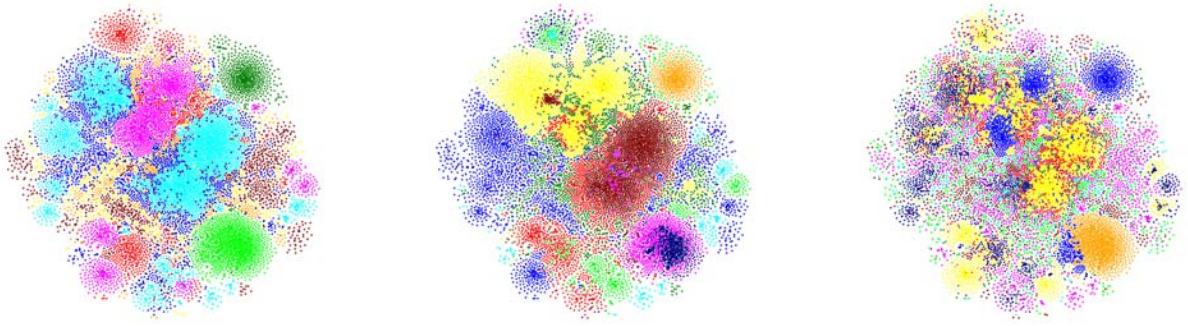
(a) RWFC 法　(b) ICFC 法（$p = 0.3$）　(c) LTFC 法

図 1: Hosei ネットワーク（$K = 10$）



(a) RWFC 法　(b) ICFC 法（$p = 0.2$）　(c) LTFC 法

図 2: Yamaguchi ネットワーク（$K = 6$）



(a) RWFC 法　(b) ICFC 法（$p = 0.2$）　(c) LTFC 法

図 3: Trackback ネットワーク（$K = 10$）

を形成していない四角く囲っている大学歴史紹介ページは，どの手法でもハブノードとリーフノードを識別できている．

図 3 は Trackback ネットワークの処理結果である．可視化に用いたクロスエントロピー法では，密に結合するコアな部分は可視化結果のより中心に集まるようにノードの座標を決定する．RWFC 法の結果を見ると，ネットワーク内のコアな部分を抽出できているが，近傍に位置するノードを同一のクラスタに分類する傾向がある．すなわち，比較的密につながるノード集合（いわゆる CNM 的コミュニティ）内のノードを同一のコミュニティとして抽出している．これは上記のクラスタの議論と同様に，ランダムウォークモデルにおいては，クラスタを構成するノードの識別が困難であることが起因していると考えられる．ICFC 法と LTFC 法の結果を見ると，可視化結果で随所に見られる各ノード集合において，コアな部分とその周辺部分，さらに末端部分というように，コアなノード群からの距離や，近傍に位置するノードの数に基づいて層状にクラスタリングされていることが窺える．図示はしていないが，クラスタリング結果の機能ベクトルを見ると，ハブノードなどの影響度の大きいノード群からの距離によってクラスタリングしている傾向にあった．

以上の 3 つのネットワークに対する結果より，提案法では RWFC 法が比較的苦手としていたクラスタ係数の高いハブノードを抽出できることが示唆された．

3.4 コミュニティ抽出結果の類似性分析

各ネットワークのコミュニティ抽出結果の類似性評価として，正規化相互情報量 [Cheng 08] を表 1 に示す．正規化相互情報量は，0 から 1 の値をとり，値が大きいほどクラスタ集合が類似していることを示す．表中 IC-RW の列は ICFC 法と RWFC 法によるコミュニティ抽出結果の類似性，LT-RW の列は LTFC 法と RWFC 法によるコミュニティ抽出結果の類似性を示している．最終列は提案法である ICFC 法と LTFC 法それぞれによる抽出結果の類似性を示している．表 1 より，LTFC 法の方が RWFC 法と類似する結果が得られていることがわかる．

表 1: 正規化相互情報量

ネットワーク	IC-RW	LT-RW	IC-LT
Hosei	0.63368	0.76294	0.61365
Yamaguchi	0.71809	0.85440	0.75614
Trackback	0.33061	0.42887	0.23388

ICFC 法と LTFC 法は仮定しているモデルが違うため，両手法の類似度は比較的高いが異なる抽出結果が得られた．両手法の大きな違いは，ノードの機能を表す機能ベクトルが意味するところにある．IC モデルによるシミュレーションでは，各リンクが与えられた拡散確率でそのリンクにおける子ノードをアクティブにするため，各ノードは複数の親ノードから情報を受け取れる可能性がある．すなわち，自分に情報が到達する経路は多く存在する．一方 LT モデルによるシミュレーションでは，各ノードが持つ閾値を超した親ノードのみが子ノードをアクティブにするため，各ノードは唯一の親ノードから情報を受け取る．すなわち，自分に情報が到達する経路は限られる．したがって，距離 d 離れたノードを d 近傍ノードとすると，ICFC 法では情報拡散において有効な d 近傍ノード数により，LTFC 法では自分に情報が到達する経路長の期待値により分類していると考えられる．機能ベクトルのクラスタリング結果を見ると，近傍ノード数あるいはハブノードからの距離などに基づいて分類される傾向にある．

4. おわりに

本稿では，ランダムウォークモデルを仮定した従来の機能コミュニティ抽出法とは異なり，ネットワーク上での情報拡散モデルを仮定した機能コミュニティ抽出法を提案した．提案した ICFC 法と LTFC 法による結果の基本的な分析をし，提案法の有効性及び従来法との相違点を明らかにした．今回用いた情報拡散モデルは，ノードが他のノードをアクティブにするタイミングが同期的であり，かつ，一度アクティブになったノードが再び非アクティブさらにアクティブになることができない SIR 型である．今後は，情報拡散のタイミングに時間遅れを導入した非同期型や何度でもアクティブになることができる SIS 型の情報拡散モデルによる検証を進めていくつもりである．

謝辞 本研究は，科学研究費補助金基盤研究 (C) (No. 23500128) の補助を受けた．

参考文献

[Cheng 08] Cheng, H., Hua, K. A., and Vu, K.: Constrained locally weighted clustering, *PVLDB*, Vol. 1, No. 1, pp. 90–101 (2008)

[Christakis 09] 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)

[Gfeller 07] Gfeller, D. and Paolo, : Spectral Coarse Graining of Complex Networks, *Physical Review Letters*, Vol. 99, No. 3 (2007)

[Kempe 03] Kempe, D., Kleinberg, J., and Tardos, E.: Maximizing the spread of influence through a social network, in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pp. 137–146, New York, NY, USA (2003), ACM

[Kimura 10] Kimura, M., Saito, K., Nakano, R., and Motoda, H.: Extracting influential nodes on a social network for information diffusion, *Data Min. Knowl. Discov.*, Vol. 20, No. 1, pp. 70–97 (2010)

[Yamada 03] Yamada, T., Saito, K., and Ueda, N.: Cross-entropy directed embedding of network data, in *Proceedings of the 20th International Conference on Machine Learning (ICML03)*, pp. 832–839 (2003)

[伏見 12a] 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)

[伏見 12b] 伏見 卓恭, 斉藤 和巳, 風間 一洋：機能性に基づくコミュニティ抽出法の比較, 情報処理学会論文誌 データベース, Vol. 5, No. 2, pp. 1–10 (2012)