3C1-OS14-2

ノードの類似性が情報伝播に与える影響について

Influence of the similarity of nodes on information diffusion

白山晋[*1] 永田勝也[*1] 亀山周明[*2]
Shirayama Susumu Nagata Katsuya Kameyama Shumei

[*1]東京大学大学院工学系研究科 [*2]東京大学大学院工学系研究科（研究当時）
School of Enginnering, the University of Tokyo School of Enginnering, the University of Tokyo

The effects of "word-of-mouth", "trust-marketing", etc. based on personal relationships have been considered to play important roles in the acquisition and utilization of information in a society or a group. Although it has been pointed out and examined that characteristics of the propagation strongly depend on the network structure, it is considered that there will be many unknown relationships. In this paper, the relationships between the diffusion of information and network structure are studied. Especially, we focus on the similarity of nodes in the structure. First, a method to divide the nodes into some groups based on the similarity is proposed. Second, some information diffusion model is introduced. Finally, the effects of the similarity of nodes are examined by numerical simulations.

1. はじめに

クチコミ等の情報伝播は人を介して行われるため，社会的に繋がりのある人々が同様の情報を保持する傾向がある．そして，その背後にあるネットワーク構造が情報伝播に影響を与えていることが指摘されている [1][2]．また，同様の意見を持つ人々が繋がりを形成する傾向があることもわかっている [3]．このような中で，ネットワーク構造としてみた立場の類似している人々が，同様の情報を保持する可能性が指摘されてきた [3]．ネットワークの特定の位置にいる人のみが頻繁に情報を取得し，それ以外の人は情報を得られないという可能性の指摘でもある．集団での情報の活用を考えるとき，この不均一性を考慮する必要がある．

このような状況は，社会ネットワーク分析においては，中心性や同値性という観点で議論され，ネットワークにおける特定の位置の役割として分析が進められてきた [4]．ところが，ネットワークの大規模化にともない，ブロックモデルを代表とするネットワーク構造に基づいたノード間の類似性評価が難しくなり，平均頂点間距離，クラスタリング係数などの統計的指標や中心性を主とした分析に移行している．ノード間の類似性に関しては全体から得られる指標によって考慮される．

一方，大規模ネットワークの多くがコミュニティ構造などの階層的構造を持つことや，複数のネットワークが合体したものであることが明らかにされると [5]，局所構造からの，ある部分のノードと別の部分のノードの類似性が改めて注目されるようになる．Leicht らは，同値性の考え方を発展させる形で，類似するノードの周囲のリンク構造は類似するという仮定の下でノードの類似度指標を提案し，特定の位置の役割をノードの類似性から分析している [6]．彼女らの手法によって算出された類似度が高いノード対は，性別・学年・人種などの属性においても類似していることが示されている．また，隣接していないが類似度の高いノード対においても同様の傾向が表れることを示している．しかしながら，そのようなノードの類似性が情報伝播に与える影響については調べられていない．

本研究では，類似度の高いノード群が情報伝播に与える影響を調べる．はじめに，Leicht ら [6] の手法を拡張し，類似度の高いノードの集団をクラスタリングによって求める手法を提案する．また，いくつかのネットワークモデルによって生成したネットワークに対して提案手法を適用し，クラスタリングの結果を考察する．次に，情報伝播シミュレーションを行い，類似度が高いノード集団が情報伝播に与える影響を調べる．

連絡先: 永田勝也，東京大学大学院工学系研究科，〒 113-8656
東京都文京区本郷 7-3-1，nagata@nakl.t.u-tokyo.ac.jp

2. 提案手法

はじめにネットワークの類似度行列を求める Leicht ら [6] の手法を説明する．次に，類似度行列を用いて，類似度の高いノード集団をクラスタリングによって求めるという手法を示す．

2.1 類似度行列の算出法

類似度行列は，Leicht らの方法 [6] によって求める．後述するクラスタリングの方法，さらに結果の解釈のために，本節でその方法を詳述する．

類似度行列を $\mathbf{S}$ で表す．その ij 成分を S_{ij} とし，S_{ij} をノード i と j の類似度とする．Leicht らは，「頂点 i が隣接する頂点 v を持ち，v が j に似ているのであれば，i は j に似ている」という定義から，S_{ij} を，隣接ノードの類似度の総和と，自分自身との類似度で再帰的な形式で表現した [6]．式で表すと，

$$S_{ij} = \phi \sum_{v} A_{iv} S_{vj} + \psi \delta_{ij} \tag{1}$$

となる．ここで，δ_{ij} はクロネッカーのデルタ，A_{ij} は隣接行列の ij 成分，ϕ，ψ はパラメータ定数である．この式を行列で表現すると，

$$\mathbf{S} = \phi \mathbf{A}\mathbf{S} + \psi \mathbf{I} \tag{2}$$

となる．Leicht らは，類似度を，その絶対値に意味があるのではなく各ノード間の相対的な関係に意味があることを指摘し，$\psi = 1$ とした．その後，式 (2) を整理し，

$$\mathbf{S} = [\mathbf{I} - \phi \mathbf{A}]^{-1} \tag{3}$$

$$\simeq \mathbf{I} + \phi \mathbf{A} + \phi^2 \mathbf{A}^2 + \cdots \tag{4}$$

を導いた．$[\mathbf{A}^l]_{ij}$ はノード i から j へのグラフ距離 l での経路を表すので，式 (4) の右辺第 1 項はノードが自分自身と類似し

ていること，第2項は距離1のノードの類似度が ϕ，第3項は距離2の頂点の類似度が ϕ^2，… というように，距離が増えるごとに減衰係数 ϕ をかけ，その和をとったものになっている．Leicht らは，経路を多く持つノードの組の類似度が高く出る傾向を指摘し，これらの項を経路数の期待値で正規化し，式 (4) を次式のように一般化した．

$$S_{ij} = \sum_{l=0}^{\infty} C_l^{ij} [\mathbf{A}^l]_{ij} \tag{5}$$

ここで，C_l^{ij} は類似度を経路の数の期待値で正規化する係数で，$C_l^{ij} = \frac{2m}{k_i k_j} \lambda_1^{-l+1}$ によって算出する．また，k_i および k_j は，i と j の次数，λ_1 は，$\mathbf{A}$ の最大固有値である．さらに，いくつかの考察の後，以下の類似度行列 $\mathbf{S}$ の算出式を導いた [6]．

$$\mathbf{DSD} = \frac{\alpha}{\lambda_1} \mathbf{A}(\mathbf{DSD}) + \mathbf{I} \tag{6}$$

ここで，$\mathbf{D}$ は，対角成分に次数をもち，他の成分が 0 となる行列である ($D_{ij} = k_i \delta_{ij}$)．また，α は，$0 < \alpha < 1$ となる定数とする．

Leicht らは定式化を示したが，類似度行列 $\mathbf{S}$ の具体的な求め方は示していない．本稿では，式 (6) から反復法によって類似度行列 $\mathbf{S}$ を近似的に求めるにした．$\mathbf{B} = \mathbf{DSD}$ とすると，

$$\mathbf{B}^{n+1} = \frac{\alpha}{\lambda_1} \mathbf{A}\mathbf{B}^n + \mathbf{I} \tag{7}$$

を，$|\mathbf{B}^{n+1} - \mathbf{B}^n| < \epsilon$ となるか，所定の回数まで繰り返す．ここで，n は繰り返しのステップ数，ϵ は小さな正数である．また，$\mathbf{B}^1 = \mathbf{I}$ とする．$\mathbf{B}$ が求まれば，$\mathbf{S} = \mathbf{D}^{-1}\mathbf{B}\mathbf{D}^{-1}$ によって類似度行列が算出できる．数値実験の結果からは，このような単純な反復法であっても収束性は悪くないことがわかっている．また，$\mathbf{S}$ は密行列になるために，計算速度よりは，記憶容量が問題になる．

このようにして求めた類似度行列を用いて，次節で述べるアルゴリズムによるクラスタリング手法を提案する．

2.2 類似度行列を利用したクラスタリング

はじめに，$\mathbf{S}$ を以下の手順に従って変形する．

1. 行列 $\mathbf{S}(S_{ij})$ の行和を s_i=$\sum_j S_{ij}$ とする．行和 s_i の最小値を s_l としたとき，$\mathbf{S}$ の l 行目と，1 行目を入れ替える．

2. 1 行目の要素と，j 行目の要素 ($j > 1$) の差の 2 乗の和 r_j:
$$r_j = \sum_{n \neq 1, n \neq j} (S_{1,n} - S_{j,n})^2 \tag{8}$$
を全ての行にわたって求める．この際，式に示すように，対角成分に相当する成分を除く．次に，最も小さな j 行目を 2 行目と入れ替える．

3. 2 行目と，j 行目 ($j > 2$) の 2 乗和の中で最も小さなものを求め 3 行目と入れ替える．4 行目，5 行目,…,n 行目を決めていき，$\mathbf{S}$ を変形する．

ここで，i 行目と j 行目を入れ替える際に，i 列目と j 列目も同時に入れ替えることにする．類似度行列の i 行は，ノード i と他のノードとの類似度を表すものである．行を基本とすれば，i 行の行和は，ノード i と他のノードとの類似度の総和を示すものである．1 番目の処理において，行和の最小値が s_l であるというのは，ノード l が他のノードとの類似度が最も小さいノードであることを示している．このノードを 1 行目として，それに近いノードを求め，2 行目にするのが 2 番目の処理である．このように類似しているノード順に行を並び替えて類似度行列を変形する．変形後の行列を $\mathbf{S}'$ とする．

次に，変形後の行列 $\mathbf{S}'$ の要素を用いて，ノードのクラスタリングを行う．

第一に，行列 $\mathbf{S}'$ の隣接する 2 行の差の二乗の和 G_j（これをギャップと呼ぶ) を下式 (9) によって求める．

$$G_j = \sum_{n \neq j, n \neq j+1} (S'_{j,n} - S'_{j+1,n})^2 \tag{9}$$

第二に，G_j の値の大きい順にクラスタを分割していく．クラスタをどこまで分割するかという指標については，式 (10) によって全ギャップの偏差値 d_j を求め，偏差値 β 以上のギャップまでを切断するという方法をとっている（$\bar{G}$ は G の平均，σ_G は G の標準偏差）．

$$d_j = \frac{10(G_j - \bar{G})}{\sigma_G} + 50 \tag{10}$$

本研究では，このクラスタ分割の方法を，SMC（Similarity Clustering Matrix）法と呼ぶことにする．

なお，本稿では以下特に断りのない限り，$\alpha = 0.95$,$\beta = 60$ としている．また，クラスタをノードのグループと呼ぶことにする．

2.3 ネットワークモデル

本稿では，実ネットワークの性質を表すネットワークモデルとして，BA モデル [8]，KE モデル [9]，CNN モデル [10]，ファミリーネット [7] を利用する．

BA モデル [8] は,「成長」と「優先的選択接続」の概念を導入したモデルである．このモデルでは，以下のルールによってネットワークを生成する．

- あらかじめ m_0 個の互いに全て接続された完全グラフを用意する．
- 各ステップごとに 1 つ新たにノードを生成し，m 個の既存のノードを選択し，そのノードとの間にエッジを生成する．ただし，既存のノード v_i はその次数 k_i に応じて次の確率 π_i で接続先として選ばれる．

$$\pi_i = \frac{k_i}{\sum_j k_j} \tag{11}$$

KE モデル [9] は,「成長」と「優先的選択接続」の概念に加えて，頂点の非活性化を考慮するモデルである．以下のプロセスを繰り返すことによってネットワークは成長する．

- あらかじめ $m = m_0$ 個の互いに全て接続された完全グラフを用意し，各ノードを活性化状態
- 枝を m 本持ったノードを 1 つ活性化状態で生成し，新たに生成したノードと既存の活性ノードを全て接続する
- $m+1$ 個の活性ノードのうち 1 つを非活性化する．このとき，ノード v_i が非活性化される確率は，正の定数 a を用いて以下の式で表される

$$P(k_i) = \frac{\frac{1}{k+a}}{\sum_{j=1}^{m+1} \frac{1}{k_j+a}} \tag{12}$$

ただし，本稿では，これらのプロセスに加えて，確率 p で優先的選択による成長のプロセスを加える．p が小さい時には，スケールフリー性とスモールワールド性を有するネットワークとなる．

CNN モデル [10] は,「成長」と「潜在リンクのリンク化」という概念によって構成されるモデルである．以下のプロセスを繰り返すことによってネットワークは成長する．

- 確率 $1-u$ で新規ノードを作成．その際，新規ノードとランダムに選択されたノード j との間にリンクを生成（その際に，新規ノードとノード j の全ての隣接ノードとの間に潜在リンクが形成される）．
- 確率 u で，(i) で形成された潜在リンクのうち 1 つをランダムに選択し，実リンク化．

ファミリーネット [7] は，小関によって提案されたネットワークモデルである．BA モデルのノードの追加プロセスを，M ノードのグループ単位でネットワークに加入し，それぞれのノードから一本ずつリンクを優先的選択で接続するというものにしている．この際，グループ内で M 個のノードは互いに全て接続された状態で加入する．このモデルでは十分大きな k に対して次数分布のベキ指数 $\gamma = M + 2$ が成立する．

2.4 情報伝播モデル

情報伝播には，ある地点，ある人から移流拡散的に伝わっていくというモデルと，全体の合意形成のような，ある状態からある状態へと至る遷移を扱うものがある．

本研究では後者に注目し，ノードに+1,-1 の 2 状態を与え，それが時間的に変化する下記の情報伝播モデルを扱う．隣接行列を A_{ij} とし，$\sigma_i(t)$ をノード i の時間 t での状態を示すものとする．

はじめに，局所的な場 $h_i(t)$ を次式で定める．

$$h_i(t) = \sum_{j}^{N} A_{ij}\sigma_j(t) \tag{13}$$

ここで，N はノードの総数である．次に，ノード i の状態を式 14 で表されるルールによって変化させる．

$$\sigma_i(t+1) = \begin{cases} \mathrm{sgn}\{h_i(t)\} & \text{if} \quad h_i(t) \neq 0 \\ \sigma_i(t) & \text{if} \quad h_i(t) = 0 \end{cases} \tag{14}$$

時間 t の+1 ($\sigma_i(t) = 1$) の割合を $r(t)$ とすると，

$$r(t) = \frac{1}{N}\sum_{i=1}^{N} \frac{(\sigma_i(t) + |\sigma_i(t)|)}{2} \tag{15}$$

となる．r_0 は $r_0 = r(0)$ である．本研究においては情報伝播する際，十分時間が経過した際の $r(t)$ を $r(\infty)$ として，これを収束値と呼び，r_∞ で示す．

2.5 観測値

本稿では，提案手法によるクラスタリングの結果として生じるグループに対して，同じグループ内で同じ情報を保持する傾向があるかを調べる．この傾向を定量化する指標として以下の 2 つのものを用いる．

1 つは，全体を 1 つのグループとみた場合，そのグループ内での情報の偏りを表す指標である．これを IS_W とし，

$$IS_W = 1 - \frac{\min\{r_\infty, 1-r_\infty\}}{\max\{r_\infty, 1-r_\infty\}} \tag{16}$$

によって求める．この指標は，収束状態において，それぞれの情報を保持する割合が半分のとき，0 になり，1 種類に偏った場合に 1 となる．

もう 1 つは，各グループにおいて同じ情報を持つ傾向がどの程度あるかを表す指標である．これを IS_G とする．

クラスタリングの結果，ノードが N_{sc} 個に分割されたとする．また，グループ i のサイズを $K_{sc}(i)$，収束状態における+1 の割合を $r_\infty^{sc}(i)$ $(i = 1, 2, ..., N_{sc})$ とする．指標 IS_G を加重平均を用いて次式で求める．

$$IS_G = \frac{1}{N}\sum_{i=1}^{N_{sc}} K_{sc}(i)\left(1 - \frac{\min\{r_\infty^{sc}(i), 1-r_\infty^{sc}(i)\}}{\max\{r_\infty^{sc}(i), 1-r_\infty^{sc}(i)\}}\right) \tag{17}$$

2 つの指標からは，IS_G が IS_W に比べて大きいほど，グループ内のノードは同じ情報を持つ傾向が強いことがわかる．

3. 実験結果

3.1 ネットワーク生成のパラメータ

ノード数は 1000 とし, 平均次数は 6 となるように，各ネットワークモデルのパラメータを調整する．このために，BA モデルのパラメータを $m0 = m = 3$ とした．また，KE モデルのパラメータを $m = m0 = 3$, $a = 3$ とし，優先的選択による成長の確率 $p = 0.1$ とした．CNN モデルのパラメータは $u = \frac{1}{3}$ である．ファミリーネットでは $M = 5$ を用いた．

3.2 情報伝播モデルのパラメータ

伝播モデルのパラメータは，初期状態における+1 の割合を表す r_0 と+1 の与え方である．本稿では，r_0 は 0.6 に固定する．また，+1 をランダムに与えることにした．

3.3 類似度行列によるクラスタリング

はじめに，提案手法によるクラスタリングの結果を示す．クラスタリングの結果，各ネットワークのノードはいくつかのグループに分かれる．図 1 に，各グループの大きさ K_{sc} の確率分布 $p(K_{sc})$ を示す．

BA モデルに対してはスケールフリー性を有することがわかる．KE モデルにおいても区分的なスケールフリー性が見受けられる．また，BA モデルと比較すると，KE モデルでは，規模が 10〜100 の中規模のグループが多く存在することがわかる．

CNN モデルにおいては，規模が 100 のところまではスケールフリー性が現れる．また，規模が非常に大きなグループの存在も確認できる．

ファミリーネットに対しては特殊な分布となる．図からは，5 の倍数の規模のグループが他の規模のグループより多く存在していることがわかる．本稿で用いた M が 5 であることから，サイズ分布の生成パラメータ M への依存性が大きいと考えられる．

3.4 情報伝播の傾向

表 1 に，各ネットワークにおける IS_W と IS_G を示す．各ネットワークに対して 100 回の試行を行い，その平均によって 2 つの指標を求めた．

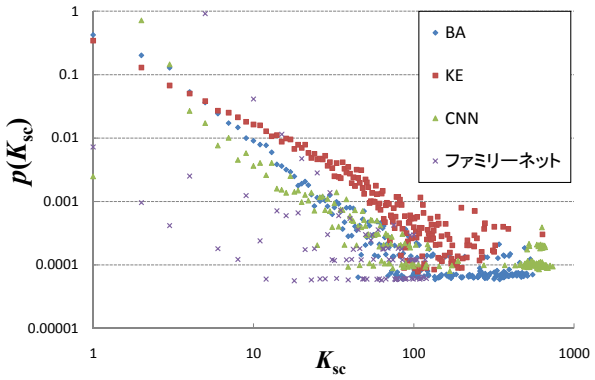

図 1: グループの規模の確率分布

BA モデルにおいては，いずれの場合も 1 となっていることがわかる．情報伝播の際，すべてのノードが+1 を保持したことによる．

KE モデル，CNN モデル，ファミリーネットにおいては，いずれも IS_W よりも IS_G の方が大きい．特に，ファミリーネットにおいては大きな差が存在する．したがって，同じグループは同じ情報を保持する傾向が強いことがわかる．ファミリーネットの場合，ネットワークの形成過程からは類似度の高いノードの集団がより明確に区別されているものと考えられる．このため，情報はグループ毎に分離して存在しやすく，IS_W よりも IS_G の方がより大きい値になっていると考えられる．

	BA モデル	KE モデル	CNN モデル	ファミリーネット
IS_W	1	0.840	0.845	0.787
IS_G	1	0.870	0.889	0.957

表 1: 各ネットワークにおける評価値

4. まとめ

本稿では，類似度の高いノードの集団が情報伝播に与える影響について考察した．

はじめに，Leicht[6] らの手法を拡張し，類似度の高いノードの集団をグループ化する手法（SMC 法）を提案した．そして，いくつかのネットワークモデルによって生成したネットワークに対して SMC 法を適用し，グループサイズの分布を求めた．

次に，得られたグループと情報伝播の関係性について，同じグループ内で同じ情報を保持する傾向を測る指標を提案した．ネットワーク上で情報伝播シミュレーションを行い，SMC 法によって得られたグループ内において，同じ情報を保持する傾向があることを確認した．この結果は，ノードの類似性が情報伝播に影響を与えることを示唆している．しかしながら，個々のグループに注目した場合，そのグループにおけるノード群が情報伝播の際にネットワーク全体に与える影響については調べられていない．また，グループサイズとの関係性についても明らかになっていない．これらが今後の課題である．

参考文献

[1] Boccaletti, S., Latora, V., Moreno, Y., Chavez,M., and Hwang,D.U.: Complex networks:Structure and dynamics, *Physics Reports*, Vol.424, pp. 175–308 (2006)

[2] 林幸雄：噂の拡がり方－ネットワーク科学で世界を読み解く, 化学同人 (2007)

[3] 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 (2010)

[4] 金光淳：社会ネットワーク分析の基礎，勁草書房 (2003)

[5] Sales-Pardo, M., Guimera,R., Moreira,A.A. and Amaral,L.A.N.: Extracting the hierarchical organization of complex systems, *PNAS*, Vol.104, No.39, pp. 15224–15229 (2007)

[6] Leicht, E.A., Holme,P., and Newman,M.E.J.: Vertex similarity in networks, *Physical Review E 73*, No. 026129 (2006)

[7] Ozeki, T.: Evolutional Family networks generated by group-entry growth mechanism with preferential attachment and their features, *The International Conference on Complex Systems*, ID 405, Boston (2006)

[8] Barabasi, A.-L. and Albert, R.: Emergence of scaling in random networks, *Science*, 286, 509-512 (2005)

[9] Klemm, K. and Eguíluz, V.M.: Growing scale-free networks with small-world behavior, *Physical Review E 65*, No.057102 (2002)

[10] Vaźquez, A.: Growing network with local rules:Preferential attachment,clustering hierarchy, and degree correlations, *Physical Review E 67*, No. 056104 (2003)