4C1-R-6-4

ネットワーク粗視化による情報拡散過程の可視化法

Visualization method of information diffusion processes by network coarse graining

小出 明弘[*1] 斉藤 和巳[*1] 長屋 隆之[*2] 伊藤 健二[*2]
Akihiro Koide Kazumi Saito Takayuki Nagaya Kenji Ito

[*1]静岡県立大学大学院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]株式会社豊田中央研究所
Toyota Central R&D Labs., Inc.

We propose a method to visualize the information diffusion processes by coarse-graining of large-scale networks. Specifically, we attempt to visualize real information diffusion phenomena over large-scale networks such as Twitter mention networks by our coarse-graining method which focuses on users, who are regarded as main role players in Twitter. In addition, in order to clarify the characteristics of real information diffusion phenomena, we contrast real information diffusion processes to artificial ones generated from the IC（Independent Cascade） and LT（Liner Threshold） models.

1. はじめに

近年マイクロブログと呼ばれる，ブログと SNS を組み合わせたサービスが増加している．その中でも，短文投稿サービスである Twitter [*1]は，急激にユーザ数を伸ばしており，研究対象として注目され，様々な知見が得られている [Kwak 2010, Huberman 2009]．その中でも我々は，人間の交友関係により構成されたネットワーク内で，あるユーザによって情報が発信されるとき，その情報が様々なユーザへ拡散する情報拡散現象に関心がある．

このようなネットワークの有する特徴や構造を理解する 1 つの有効なアプローチは，情報を可視化することである．可視化することにより，対象ネットワーク内のノードの相互関係や内在する特徴など，多くの情報をわかりやすく，直感的に把握することが期待できる．そのため，ネットワークの可視化は重要であり，様々なネットワーク可視化法が提案されている [Kamada 1989, Torgerson 1958]．しかし，ネットワークの大規模化・複雑化に伴い，ノードが密に配置されることにより，ネットワークの構造を把握することが困難になる．

本論文では，大規模なネットワークに対し粗視化を施すことにより，情報拡散過程を可視化する手法を提案する．具体的には，Twitter の mention ネットワークを次数に着目した手法により粗視化し，現実の拡散現象を粗視化したネットワークに流すことによって，情報拡散現象を視覚的に明らかにする．また，IC モデルや LT モデルを利用した人工の拡散データを作成し，現実の拡散現象に見られる特徴を明らかにする．

第 2 章では，本論文で用いるネットワークの粗視化手法について述べる．第 3 章で使用データの詳細と実験設定について説明し，第 4 章で分析結果について述べる．最後に，第 5 章で本論文をまとめ，今後の課題について述べる．

2. ネットワーク粗視化法

本章では，大規模なネットワークをノードの次数に基づいて粗視化する手法について述べる．具体的には，以下の 3 ステップの処理を施す．

連絡先: 小出明弘，静岡県立大学 大学院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 ，054-264-5102，j11103@u-shizuoka-ken.ac.jp

*1 http://twitter.com/

(i) 次数中心性を拡張した集合次数中心性の概念に基づき，ノードを選定する．

(ii) 選定した各ノードに対し，K-近傍法を用いてリンクを結び，ネットワークを生成する．

(iii) 作成されたネットワークを，各ノードの類似度に基づいて配置し，粗視化する．

また，今後の共通的な設定として，ネットワークを $G = (V, E)$ と定義し，ノード集合を $V = \{v_1, v_2, \ldots, v_N\}$ ，エッジ集合 E を $V \times V$ の部分集合とする．なお， N はネットワークの全ノード数を表す．各ノード $v_i \in V$ に対して，v_i の子ノード集合 $F(v_i)$ を $F(v_i) = \{v_j \in V; (v_i, v_j) \in E\}$，親ノード集合 $B(v_i)$ を $B(v_i) = \{v_j \in V; (v_j, v_i) \in E\}$ と定義する．

2.1 集合次数中心性

ノード v_i の次数とは，自身とつながっているノード数である．$A(v_i)$ を $A(v_i) = F(v_i) \cup B(v_i)$ と定義したとき，ノード v_i の次数は $deg(v_i) = |A(v_i)|$ と表すことができる．次数中心性とは，次数が高いノードほど重要ノードであるという直観に基づいた中心性指標である．本分析では，この指標をノード集合に対する概念に拡張した，集合次数中心性 [Fushimi 2012] を利用する．選定するノード集合 R の要素数を $M = |R|$ とすると，集合内のいずれかのノードと隣接するノード数によって定義できる．

$$SD(R) = \Big| \bigcup_{v_i \in R} A(v_i) \Big|. \tag{1}$$

個々のノードに対する次数中心性，すなわち“ 多くの隣接ノードを持つノードは重要である ”という概念の自然な拡張となっており，式 1 を最大にするような集合 R を求める問題は，M 個のノードで出来るだけ多くの隣接ノードを被覆する問題 M-Vertex Covering 問題とみなすことができる．また，求める集合の要素数 $M = 1$ のとき，集合次数中心性により選定されるノードは，次数中心性トップノードと等しくなる．

式 1 を最大にするような集合 R を求めるための具体的なアルゴリズムを以下に示す．求める集合 R の要素数を $M = |R|$ とする．

(i) 初期化：$R = \emptyset, m = 1$ とする

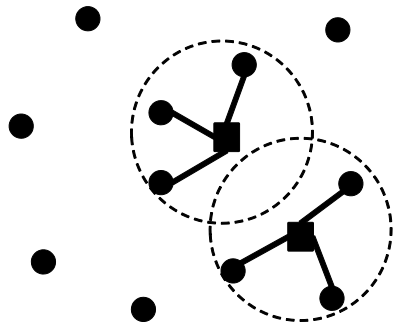

図 1: K-NN 法によるネットワーク構築

(ii) 選定：追加した際の増分が最大となるノード

$$r_m = \arg\max_{v_i \in V} SD(R \cup v_i) - SD(R) \text{ を求める}$$

(iii) 反復： $R \leftarrow R \cup r_m$, $m = M$ なら終了，さもなければ $m \leftarrow m+1$ とし（ii）へ

2.2 K-近傍法

本分析では，粗視化したノード集合をネットワークにするために，K-近傍法（以下 K-NN 法）を利用する．K-NN 法は，パターン認識の分野でしばしば用いられる単純な機械学習アルゴリズムである．本分析では，K-NN 法を応用した無向 K-NN ネットワークを作成する．詳細な手順を以下に示す．

(i) 全ノードペア間の類似度を計算する

(ii) 各ノードに対して，自身との類似度で降順に相手ノードを並び替え，$k \leftarrow 1$ とする

(iii) 各ノードは，自身との類似度が k 位のノードとリンクを結ぶ

(iv) 全ノードが 1 つの連結成分になるまで $k \leftarrow k+1$ とし，(iv) を繰り返す

K-NN 法によるネットワーク構築のモデル図を図 1 に示す．図 1 では，$K = 3$ とし，ノード間の類似度を 2 次元のユークリッド距離で表している．図 1 中の四角いノードはそれぞれ，自身の最近傍である 3 つのノードにリンクを張っている．この処理をすべてのノードに対して行う．K-NN 法により構築したネットワークを K-NN ネットワークと呼ぶ．

本分析では，粗視化により選定されたノード集合を $R = \{r_1, r_2, \ldots, r_M\}$ の，任意の 2 つのノード間の類似度 $\rho(r_i, r_j)$ を以下のように定義し，K-NN ネットワークを作成した．

$$\rho(r_i, r_j) = \frac{\left|A(r_i) \cap A(r_j)\right|}{\left|A(r_i) \cup A(r_j)\right|}. \quad (2)$$

3. 使用データと実験設定

3.1 使用データ

本論文で粗視化対象とするネットワークは，Twitter の mention ネットワークである．2011 年 3 月 7 日から 2011 年 3 月 23 日までの日本語で投稿された @user を含むツイートから，各ユーザをノードとし，@user で指定されたユーザにリンクしてネットワークを作成した．さらに，ネットワークに弱連結成分分解を施し，最大の連結成分を分析対象とした．最終的に作成されたネットワークのノード数は 4,548,304 ，リンク数は 50,945,956 である．なお，最大連結成分は元のネットワークの 99.6 %をカバーしている．

また，情報拡散データとして公式リツイートを利用する．先述の mention ネットワークの収集期間と同様の期間内のツイートから，公式リツイートされたツイートを抽出する．各ツイートからは，リツイートしたユーザ，リツイート元のユーザ，リツイートした時刻，を得ることができ，これらの情報を利用することにより，あるツイートが時間の経過に伴いどのようにユーザを介して伝搬していくのか把握することができる．本分析では，多くのユーザからリツイートされているユーザである@tsuda, @masason のツイートを分析対象とする．なお，本分析で使用するデータは東日本大震災の期間を含んでおり，この期間内では通常の利用形態とは異なる利用がなされていたと考えられる．そこで，震災前後でデータを分割し，震災前後での情報の拡散現象の変化を解析する．

3.2 基本的な情報拡散モデル

本節では，実データと比較するための基本的な情報拡散モデルである IC（独立カスケード）モデルと LT （線形閾値）モデルについて述べる．

3.2.1 IC モデル

IC モデルは感染病の広がり方などを表すとされる基本的な確率モデルである．このモデルでは，各リンク (v_i, v_j) に対して前もって実数値 $p_{v_i,v_j} (0 < p_{v_i,v_j} < 1)$ を割り当てる．ここで，p_{v_i,v_j} をリンク (v_i, v_j) における拡散確率と呼ぶ．IC モデルでの拡散過程は離散時間 $t \geq 0$ で展開され，情報源ノードから以下の方法によって広がっていく．ノード v_i が時刻 t でアクティブになったとき，ノード v_i には現在非アクティブの子ノード v_j に対して一度だけアクティブにさせるチャンスが与えられ，それは拡散確率 p_{v_i,v_j} で成功する．成功したら，ノード v_j は時刻 $t+1$ でアクティブになる．もし，ノード v_j の複数の親ノードが時刻 t で同時にアクティブになった場合には，任意の順番で拡散試行が行われるとする．このプロセスが反復して行われ，次の時刻でアクティブになるノードが無くなったとき，情報拡散は終了する．

3.2.2 LT モデル

LT モデルは，すべてのノード $v_i \in V$ に対して，$\sum_{v_j \in B(v_i)} \omega_{v_j,v_i} \leq 1$ となるように前もって重み $\omega_{v_j,v_i} (> 0)$ を割り当てる．LT モデルでの拡散過程は，初期アクティブ集合 S が与えられた上でランダムルールに従って行われる．まず，全てのノード $v_i \in V$ に対して，閾値 θ_{v_i} が区間 [0, 1] から一様ランダムに選ばれる．時刻 t で非アクティブノード v_i は各親ノード $v_j \in B(v_i)$ から ω_{v_j,v_i} の影響を受ける．もし，ノード v_i のアクティブな親ノードから受けた重みの合計が θ_{v_i} 以上になった場合，ノード v_i は時刻 $t+1$ でアクティブになる．このプロセスが反復して行われ，次の時刻でアクティブになるノードが無くなったとき，情報拡散は終了する．

3.3 実験設定

本分析では，4,548,304 ノードを有する mention ネットワークを，集合次数中心性に基づき $M = 10000$ で選定したネットワークに粗視化した．その後，粗視化したノード集合のノード間の類似度に基づき K-NN ネットワークを作成し，2 次元平

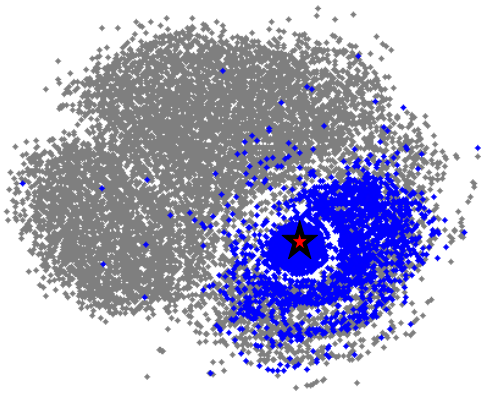

図 2: 震災前の @tsuda のツイート拡散

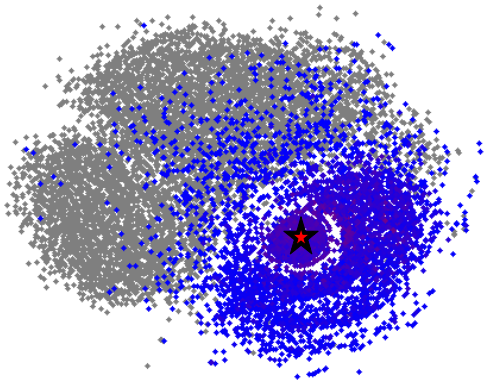

図 3: 震災後の @tsuda のツイート拡散

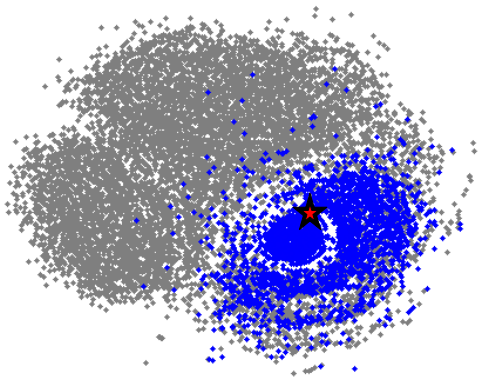

図 4: 震災前の @masason のツイート拡散

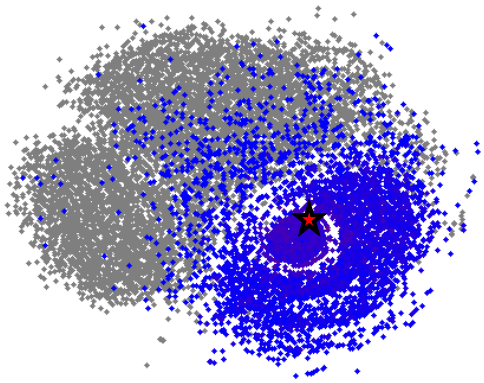

図 5: 震災後の @masason のツイート拡散

面上にネットワークを配置する．各ノードの配置法にはばねモデルを利用した．また，$\frac{SD(R)}{N} = 0.456$ である．

また，表 1 に本分析で情報拡散現象として利用する実データでの拡散系列（公式リツイート）数と，人工データの拡散系列数をそれぞれ示す．較として利用する情報拡散データは，@tsuda と @masason を情報源として設定し，拡散確率は先行研究に基づき，IC モデルで $p_{v_i,v_j} = 0.02$，LT モデルで $p_{v_i,v_j} = 1.0$ 設定した．なお，拡散確率は $p_{v_i,v_j} = p$ とし，各ノード間で一定とする．

表 1: 比較用データの拡散系列数

	震災前	震災後	IC,p=0.02	LT,p=1.0
@tsuda	674	89,430	27,227	4,963
@masason	852	58,057	26,836	2,721

4. 分析結果

mention ネットワークを 10,000 ノードに粗視化したネットワークに対してリツイートデータを可視化したものを図 2 から図 5 にそれぞれ示す．ここで，情報を拡散していないノードを灰色のノード，粗視化した各ノード集合の中で一つでも情報を拡散したノードが存在するとき，青色のノード，粗視化した各ノード集合の中で多くのノードが情報を拡散したとき，赤色のノードで配色している．また，各図において情報源を星で表している．震災の前後での情報拡散を比較すると，震災前は左側の塊の中でのみ情報が伝わっているが，震災後は多少ではあるが右側のノード群へと情報が拡散している．また，震災後には左側の中央付近に赤いノードが現れ，この付近で多くの情報の伝達がなされていたと考えられる．また，@tsuda と @masason の可視化結果を比較すると，配色がほとんど変化しない．このことから，両者の情報は似たようなユーザを介して伝達されている一方，ネットワーク内の限られたユーザにのみ拡散していると考えられる．

続いて，@tsuda , @masason をそれぞれ情報源としたときの，IC モデルと LT モデルでの情報拡散結果を図 6 から図 9 にそれぞれ示す．まず，@tsuda と @masason の可視化結果を比較すると，どちらを情報源においても配色結果にはほとんど違いが見られない．また，実データと人工データを比較すると，震災後の実データと IC モデルでの拡散結果は，左側のノード群の配色傾向はほぼ一致しているが，IC モデルでは，右側のノード群にも広く情報が拡散している．IC モデルによって生成された拡散系列は，実データに比べ 50% 以下の大きさであるにもかかわらず，広範囲のノードに情報が拡散しており，現実の拡散現象は，IC モデルによって人工的に生成された拡散現象には見られない特性を有していると考えられる．一方，LT モデルでの可視化結果をみると，拡散系列は LT モデルにより生成された人工データの方が 3〜8 倍大きいが，震災前の実データとほぼ一致した可視化結果が得られている．これらの結果から，現実の情報拡散現象は，既存の情報拡散モデルでは

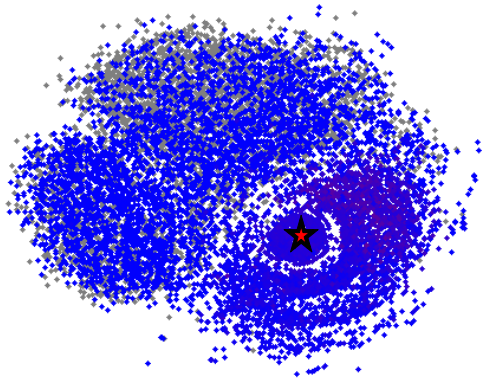

図 6: 情報源：@tsuda，IC モデル，$p = 0.02$

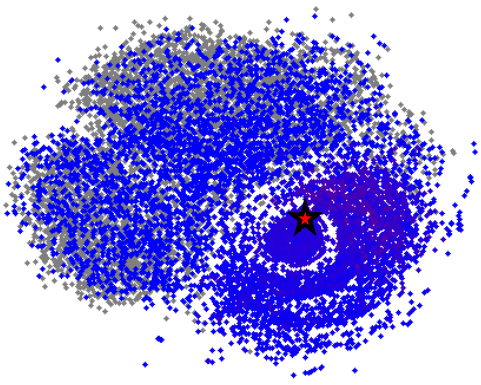

図 8: 情報源：@masason，IC モデル，$p = 0.02$

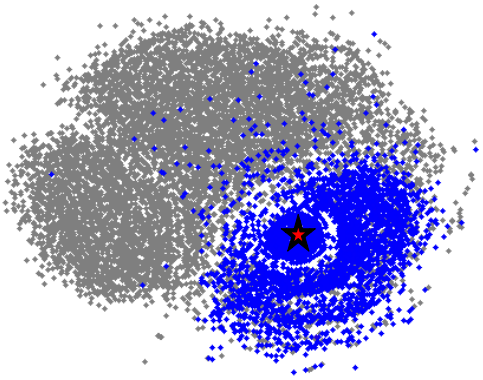

図 7: 情報源：@tsuda, LT モデル，$p = 1.0$

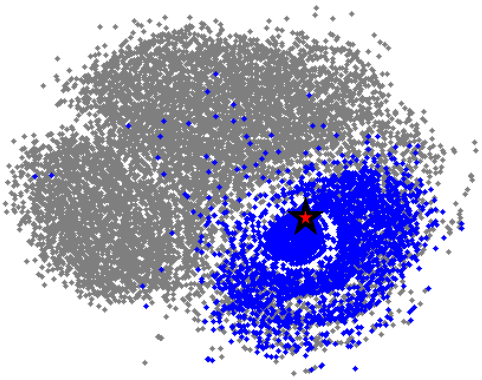

図 9: 情報源：@masason，LT モデル，$p = 1.0$

表現できないような構造を持つことが推測される．

5. おわりに

本論文では，大規模なネットワーク上での情報拡散現象を視覚的に明らかにするため，集合次数中心性を利用したネットワークの粗視化手法を提案した．さらに，公式リツイートを情報拡散現象と仮定し，粗視化したネットワークに拡散データを流し込むことにより，可視化を行った．また，基本的な情報拡散モデルである IC モデル，LT モデルを用いて拡散系列を作成し，現実の情報拡散現象と比較を行うことにより，現実の拡散系列は既存のモデルとは異なる特徴を持つことを示した．

今後の課題として，本分析で利用した集合中心性を利用した粗視化では，元のネットワークを完全に再現することができていないため，ネットワークのすべてのノードに対して適用することが可能な粗視化手法を検討していく．また，本論文では本提案手法の有効性を定量的に評価することができていないため，本分析手法を定量的に評価するための新たな指標に関しても検討していく．

謝辞 本研究は株式会社豊田中央研究所との共同研究および，科研費 (23500312) の助成を受けた．

参考文献

[Kwak 2010] H.Kwak, C.Lee, H.Park, and S.Moon, What is Twitter, a social network or a news media? In Proceedings of the 19th international conference on World wide web, pp.591-600. ACM, (2010).

[Huberman 2009] B.A.Huberman, D.M.Romero and F.Wu, Social networks that matter: Twitter under the microscope, First Monday, Volume 14. Number 1. January 5 (2009).

[Kamada 1989] T.Kamada and S.Kawai, An algorithm for drawing general undirected graph. Information Processing Letters, 32, pp.7-15, (1989).

[Torgerson 1958] W.Torgerson, Theory and Methods of Scaling. Wiley, New York (1958).

[Goldenberg 2001] J. Goldenberg, B. Libai, and E. Muller, Talk of the network: A complex systems look at the underlying process of word-of-mouth, Marketing Letters, vol.12, pp.211-223, (2001).

[Watts 2002] D.J. Watts, A simple model of global cascades on random networks, Proceedings of National Academy of Science, USA, vol.99, pp.5766-5771, (2002).

[Fushimi 2012] 伏見卓恭，斉藤和巳，武藤伸明，池田哲夫，ノード集合に対する媒介中心性の提案，第 4 回データ工学と情報マネジメントに関するフォーラム,，(2012).