社会ネットワークの情報拡散モデルとコミュニティ構造の分析

What Does an Information Diffusion Model Tell about Social Network Structure?

伏見卓恭[*1] 水本嗣留[*1] 斉藤和巳[*1] 元田浩[*2] 木村昌弘[*3]
Takayasu Fushimi Tsuguru Mizumoto Kazumi Saito Hiroshi Motoda Masahiro Kimura

[*1]静岡県立大学 University of Shizuoka
[*2]大阪大学 Osaka University
[*3]龍谷大学 Ryukoku University

In social network, there are two representative information diffusion models such as the Independent Cascade (IC) model and the Linear Threshold (LT) model. Using these two models, we analyzed how the structure of the network affects the diffusion phenomena. For this purpose, we generated the Generalized Random (GR) network to destroy community structure by rewiring links without changing the degree of each node, and compared the results by plotting the influence degree against the node degree of the information source. There is a clear difference between the IC and the LT models as well as between the original and the GR networks.

1. はじめに

「クチコミ」による情報伝播が起こる代表的な社会ネットワークには，ウェブ上でのブログのトラックバックネットワークなどが考えられる．すなわち，各人や各ブログなどをノードとし，ブログ間トラックバックなどその繋がりをリンクとする社会ネットワークで情報伝播が起こる．このような社会ネットワークでは，どのノードに最初に情報を与えれば情報がより多くまたはより広く伝播するかということが重要になってくる．すなわち影響度の高いノードを調べる必要がある．そこで我々は，情報伝播の基本的なモデルである独立カスケードモデル（Independent Cascade model）と線形閾値モデル（Linear Threshold model）[?] を用いて分析を行う．特に本論文では，ネットワークに存在するコミュニティ構造が情報伝播の影響度にどのように関係しているかを分析する．そのために，一般ランダムネットワークを作り，オリジナルネットワークと比較する方法論を提案し，現実世界のネットワークでの評価を行う．

2. 情報伝播の基本的な確率モデル

本論文ではこの分析を行うにあたって，基本的な情報伝播の確率モデルを適用する．使用する確率モデルは有向リンクのネットワーク $G = (V, E)$ を想定する．ネットワーク G での V と E $(\subset V \times V)$ はそれぞれノード集合とリンク集合を意味している．この社会ネットワークにおいて，ある情報が伝播していく現象について分析する．情報が伝播し，その情報を得たノードを“ アクティブ ”な状態であると呼び，そうでないノードを“ 非アクティブ ”な状態であると呼ぶ．“ アクティブ ”なノードが“ 非アクティブ ”に変化することはないと仮定する．情報伝播のプロセスは離散時刻 t で展開される．情報伝播の試行が成功しようと失敗しようと，非アクティブな子ノードをアクティブにできる機会は一度きりである．このような前提のもとで，最も代表的な独立カスケードモデルと線形閾値モデル [?] を用いて分析を行う．

連絡先: 伏見卓恭，静岡県立大学経営情報学研究科，〒 422-8526 静岡市駿河区谷田 52 番 1 号，054-264-5436，j09118@u-shizuoka-ken.ac.jp

3. 分析の手法と枠組み

今回の実験を行うに際して提案した分析手法は，２つのステップから構成される．１つ目は，一般ランダムネットワークを作成することである [?]．一般ランダムネットワークは，解析の対象とするオリジナルのネットワークからそれぞれのノードの次数すなわちリンク数を変更せずにランダムにリンクを張り替えて生成する．このリンク張り替え時には，自分から自分へリンク (self-link) を張らないようにまた，同じ相手ノードに対し複数のリンク (multiple-link) を張らないように注意しなければならない．具体的な生成方法は，以下の通りである．オリジナルのネットワークからノードリスト $L_E = (e_1, \cdots, e_{|E|})$ を準備する．それぞれのリンクは，例えば $e = (u, v)$ のように順序対になっており，$from-part$ と $to-part$ から成っている．これから，２つのノードリスト L_F と L_T を作り出す．そして，注意事項である，self-link と mutiple-link に気をつけ， L_T をランダムに並び替える．その後， L_F と L_T を結合させ新しいリンクリストを作り，これに基づき一般ランダムネットワークを作成する．

２つ目は，情報伝播モデルに基づいて，シミュレーションを行った結果を，横軸にノードの次数をとり，縦軸に各ノードの期待影響度をプロットしたグラフを描画することにより，ネットワークの構造を分析することである．

4. 評価実験

本論文で取り扱う実験は，２つのネットワークデータを用いて行ったものである．

4.1 ブログのトラックバックネットワーク

１つ目のネットワークデータは，ブログのトラックバックネットワークのデータである．ブログすなわち Weblog のトラックバック機能とは，あるブログの内容が自身のブログの内容と関連性が高い時，またはあるブログの内容を参照・引用した場合などに，そのブログにトラックバックというリンクを張るが，その際にリンクを張った相手に通知する機能のことである．このリンク関係によるネットワークを利用し，あるブログ著者から別のブログ著者へと情報が伝播しうると考えられるので，ブログのトラックバックネットワークを用いて実験を行った．本論文ではこのネットワークを，ブログネットワークと呼

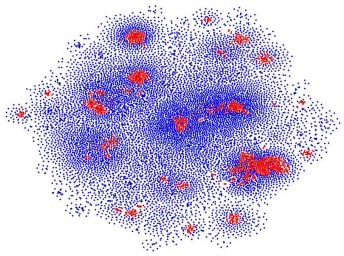


図 1 ブログネットワーク オリジナル

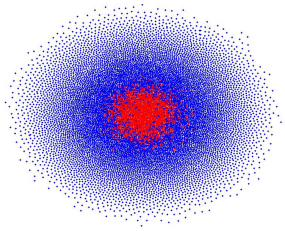


図 2 ブログネットワーク 一般ランダム

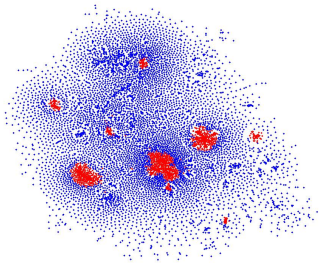


図 3 ウィキペディアネットワーク オリジナル

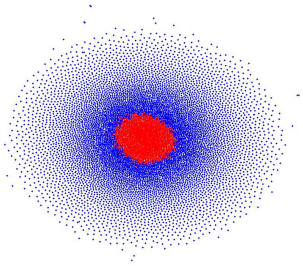


図 4 ウィキペディアネットワーク 一般ランダム

ぶことにする．ブログネットワークのデータは,「goo ブログ」(http://blog.goo.ne.jp/usertheme/) 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，2005 年 5 月に収集したものである．このネットワークは，12,047 ノードと 79,920 リンクをもつ有向ネットワークである．多くの大規模なネットワークと同なじように，自分に向かうリンク次数の分布（入次数分布）も自分から向かうリンク次数の分布（出次数分布）もべき則分布に従うという特徴を有す．

ブログネットワークの構造を直感的に理解するために，我々はクロスエントロピー法 [?] によりネットワークの可視化を試みた．可視化の方法としては，k-core [?] の値が平均次数より高いノードを赤い点でプロットし，それ以外のノードを青い点でプロットしている．オリジナルのブログネットワークの構造は，図 1 から読み取れるように，いくつかのコミュニティ構造が存在する． k-core の値が高いすなわち赤いノードの周りに，多くの青いノードが散布している．これらをコミュニティとすると，いくつかのコミュニティが点在していることが見て取れる．一方，一般ランダムネットワークの可視化結果の図 2 では，全体が一つのコミュニティのようになっており，オリジナルのネットワークで存在していた，コミュニティ構造が壊れてしまっていることがわかる．これらのコミュニティ構造の違いが，情報伝播にどのように影響するのかを，我々は本論文で検証していく．

4.2 ウィキペディアの人名ネットワーク

2 つ目のネットワークデータは，日本の「ウィキペディア」内の「人名一覧」から人物ネットワークを用いて評価実験を行った．実際に,「人名一覧」に登場する人物において，ウィキペディア内の記事中に 6 回以上共に載せられている 2 人の人物をリンクで結ぶことによって無向グラフのネットワークを構築する．無向グラフの最大連結成分を抽出し，無向リンクを双方向リンクとして，有向グラフのネットワークに変換する．本論文ではこのネットワークをウィキペディアネットワークと呼ぶことにする．ノード数は 9,481 であり，有向リンク数は 245,044 である．図 3 の可視化の結果を見てみると，ウィキペディアネットワークもブログネットワークと同様に，オリジナルのネットワークではいくつかのコミュニティ構造が存在することがわかる．さらに一般ランダムネットワークの可視化結果の図 4 では，全体が一つのコミュニティになっている．すなわち，コミュニティ構造が壊れてしまっていることが見て取れる．

4.3 ネットワークデータの特徴

無向社会ネットワークのデータには，そうでないネットワークのデータとは異なった 2 つの統計的な特徴があることが知られている．1 つ目は，隣接しているノードどうしの次数には高い相関関係があるということ．2 つ目は，一般にスモールワールド性を有することである．スモールワールド性を有するネットワークでは，クラスタ係数が比較的大きく，最短ノード間距離の平均が比較的小さくなる．ここで，無向ネットワークのクラスタ係数は以下のように定義される：

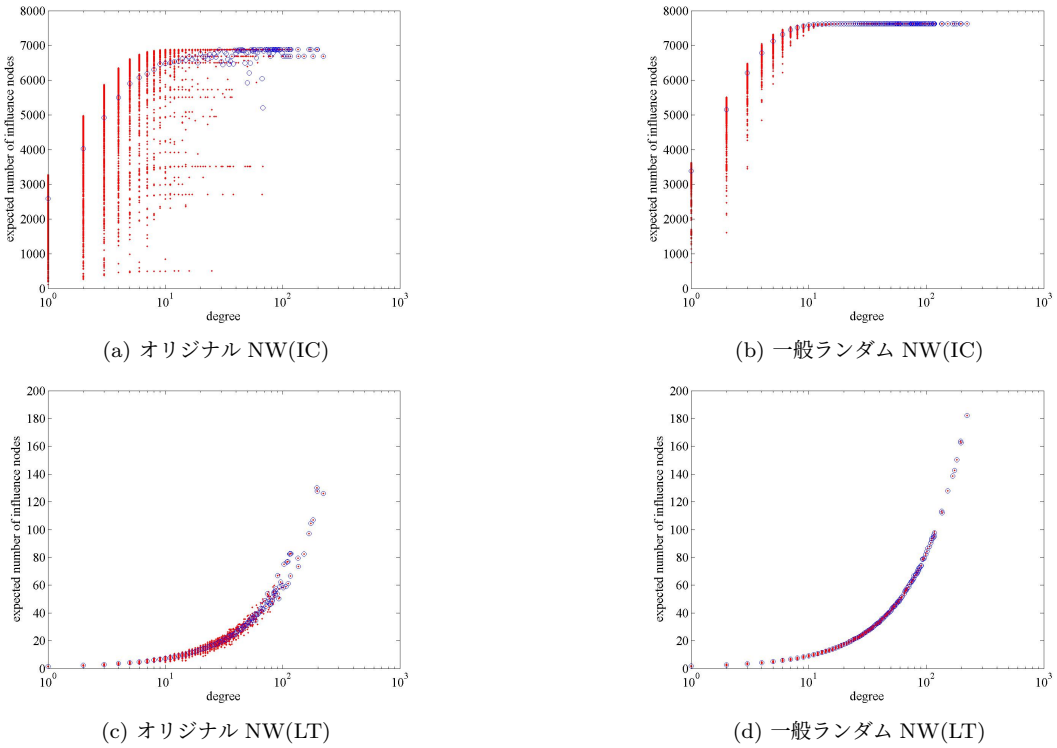
$$C = \frac{1}{|V|}\sum_{u\in V}\frac{|\{(v\in V, w\in V) : v\neq w, w\in A_G(v)\}|}{|A_G(u)|(|A_G(u)|-1)}.$$

ここで,$A_G(u)$ はノード u の近傍ノード集合を意味している．さらに，ネットワークにおける平均ノード間距離は以下のように定義される：

$$L = \frac{1}{|V|(|V|-1)}\sum_{u\neq v} l(u,v).$$

ここで $l(u,v)$ は，ノード u とノード v の最短ノード間距離を意味している．このような基本的なネットワーク統計量の視点から見ると，本論文で取り扱っている 2 つのネットワークデータは表 **??**のようになる．

表 **??**を見てみると，どちらのネットワークデータも，一般ランダムネットワークに比べるとオリジナルのネットワーク



(a) オリジナル NW(IC)　(b) 一般ランダム NW(IC)

(c) オリジナル NW(LT)　(d) 一般ランダム NW(LT)

図 5 ブログネットワークにおける情報伝播過程の比較

表 1: ネットワークの基本統計量

ネットワーク	C	L
オリジナル ブログ	0.26197	8.17456
一般ランダム ブログ	0.00523	4.24140
オリジナル ウィキペディア	0.55182	4.69761
一般ランダム ウィキペディア	0.04061	3.12848

は，クラスタ係数が非常に大きくなっている．平均ノード間距離も幾分か大きくなっているが同程度のスケールである．このことからも，オリジナルのネットワークには，スモールワールド性を有することが数値的にも見て取ることができる．

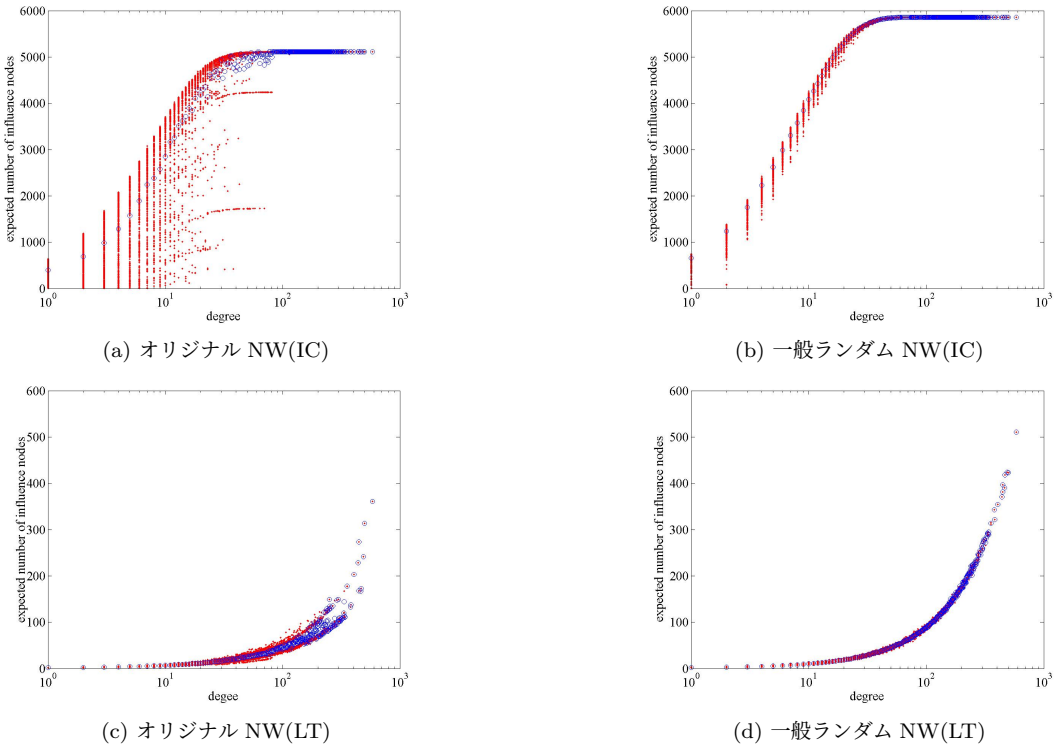
4.4 実験設定

本論文での実験におけるＩＣモデルとＬＴモデルの設定内容を説明する．まずＩＣモデルでは，全ての有向リンク (u, v) に伝播確率 $\beta_{u,v}$ を付与する．平均次数の逆数である $\beta = |V|/|E|$ を用いる．実際に計算したところ，ブログネットワークにおける伝播確率は $\beta = 0.1507$ であり，ウィキペディアネットワークにおける伝播確率は $\beta = 0.0387$ になった．これよりＩＣモデルでは，ブログネットワークには $\beta_{u,v} = \beta = 0.2$ を設定し，ウィキペディアネットワークには $\beta_{u,v} = \beta = 0.03$ を設定する．次にＬＴモデルでは，全てのノード v に重みを付与する．ノード v の親ノード $u \in \Gamma(v)$ から受ける重みは次のように定義した: $\omega_{u,v} = 1/|\Gamma(v)|$．以上のように定義されたパラメータを用いて実験を行う．

4.5 ２つのネットワークでの実験結果

ブログネットワークを用いた実験の結果を図 5 に示す．図 5(a) はブログネットワークのオリジナルネットワークを用いたＩＣモデルの実験結果である．同様に図 5(b) は一般ランダムネットワークにＩＣモデルを適用したものである．図 5(c) はオリジナルネットワークにＬＴモデルを適用したものであり，図 5(d) は一般ランダムネットワークにＬＴモデルを適用したものである．図中の赤い点は，各ノードの影響度をプロットしたものであり，青い丸は，各次数における影響度の平均をプロットしたものである．情報伝播モデル間の特徴を比較してみる．ＩＣモデルとＬＴモデルの間に見られる顕著な特徴は，どちらのモデルを適用した場合であっても，期待影響度は次数が高い情報発信源ノードの方が大きくなっている．すなわち図は，右肩上がりとなっている．しかしその曲率は逆向きであり，ＩＣモデルでは上に凸でＬＴモデルでは下に凸になっている．さらに全体的な期待影響度がＬＴモデルより，ＩＣモデルの方が非常に大きいことが見てとれる．

次にオリジナルネットワークと一般ランダムネットワークの特徴を比較してみる．どちらの情報伝播モデルを適用しても，オリジナルネットワークより一般ランダムネットワークの方が期待影響度が幾分か大きくなっている．これは一般ランダムネットワークの平均ノード間距離 L が，オリジナルネットワークの平均ノード間距離 L より小さくなっていることによるものだと考えられる．特に，ＩＣモデルではその特徴が顕著に出ている．また，一般ランダムネットワークにＬＴモデルを適用した結果の図 5(d) では，期待影響度がほぼ各ノードの次数により一意に決まっている．すなわち図は，赤い点が散らばらずに，平均である青い丸に近いところにプロットされていて分散が小さくなっている．最も顕著な違いは，ＩＣモデルを適用した場合である．オリジナルネットワークの図 5(a) では，赤い点が多くの左右に延びる筋のような線になっているのが見受けられる．しかし一般ランダムネットワークでは，その筋のような線が消えている．



(a) オリジナル NW(IC)　(b) 一般ランダム NW(IC)

(c) オリジナル NW(LT)　(d) 一般ランダム NW(LT)

図 6 ウィキペディアネットワークにおける情報伝播過程の比較

ウィキペディアネットワークを用いた実験でも，前述したブログネットワークを用いた実験とほぼ同様な結果が得られた．

以上の２つの実験結果より以下のことがわかった．1) 次数が高いノードの方がより多くのノードに影響を与えることができるが，ＩＣモデルとＬＴモデルではその曲線の曲率は異なる；2) どちらの情報伝播モデルを適用しても，一般ランダムネットワークの方がオリジナルネットワークより幾分か期待影響度が大きくなっている；3) 一般ランダムネットワークにＬＴモデルを適用した時に，期待影響度の値はほぼ一意に決まっている；4) オリジナルネットワークにＩＣモデルを適用した場合，多くの左右に広がる筋のような線が存在する．

5. おわりに

本論文では，オリジナルのネットワークから一般ランダムネットワークを生成することにより，異なるコミュニティ構造のネットワークデータを実験に使用した．異なるコミュニティ構造をもつ大規模なネットワークに，ＩＣモデルとＬＴモデルという情報伝播モデルを適用し，情報伝播のシミュレーションを行った．その実験結果をより正確に理解するために，次数ごとの期待影響度の値をプロットした図を描画した．これらの実ネットワークを使用することにより，今後の研究で重要なステップになり得るいくつかの興味深い知見を見出せたと考えている．この実験でわかったことでもっとも重要なことは，コミュニティ構造の影響はＬＴモデルに比べて，ＩＣモデルの方がより受けやすいということである．我々の今後の研究としては，さまざまなネットワークデータを用いることにより，コミュニティ構造と情報伝播モデルとの関連性を探求することで，影響最大化問題などに活かせるようしていくつもりである．

参考文献

[Kempe 03] Kempe, D., leinberg, J. and Tardos, E. (2003) Maximizing the spread of influence through a social network, Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 137–146).

[Newman 03] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167–256.

[Seidman 03] S.B. Seidman, S. B. (1983). Network Structure and Minimum Degree, *Social Networks*, *5*, 269–287.

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