拡散データからのモデル推定による期待影響度の予測*

吉川　友也†　　斉藤　和巳†　　元田　浩‡　　木村　昌弘§　　大原　剛三※
静岡県立大学†　　大阪大学‡　　龍谷大学§　　青山学院大学※

1. はじめに

現実の社会ネットワークの情報拡散分析は，効果的なクチコミマーケティングに役立つと考えられる．具体的には，情報は社会ネットワーク上をどのように拡散し，どれだけの人に知られるのかということを知ることで，マーケティング戦略をより効果的にすることができると考える．

私たちは情報拡散の基本モデルである IC モデルを拡張した CTIC モデルによって情報拡散分析を行っている[1]．観測された拡散データが CTIC モデル上で情報拡散したと仮定すると，拡散確率 κ と時間遅れパラメータ r を推定することができる．この推定値と実際のネットワーク情報があれば，私たちは期待影響度を予測することが可能である．しかし，観測できる拡散データは 1 つである．故に，偶然にそのトピックがよく拡散したり，逆に全く拡散しなかったりすることが考えられる．

本論文では，観測された拡散データより推定した拡散確率 κ と時間遅れパラメータ r の頑健性について分析する．詳細には，人工的に拡散データを作り，その拡散データよりパラメータ推定を行う．求められたパラメータである κ と r を用いて期待影響度を求めたとき，それが観測された拡散データ（今回はシミュレーションデータ）とどれだけ離れているかを調査する．

今回の分析により，パラメータの頑健性が保証されれば，現実の拡散データを用いた情報拡散分析において，確からしい期待影響度を求めることが可能であることが示唆される．

2. 分析方法

2.1. 連続時間独立カスケード CTIC モデル

まず，使用する情報拡散モデル CTIC モデルを定義する．有向ネットワークを $G = (V, E)$ で定義する．$V = \{u, v, w, \cdots\}$ はノード集合を，$E = \{(u, v), (v, w), \cdots\}$ はリンク集合を表す．ここで，ノード v がリンクする子ノード集合を $F(v) = \{w; (v, w) \in E\}$ とし，ノード v へリンクする親ノード集合を $B(v) = \{u; (u, v) \in E\}$ で表す．今，ノードが情報を保持している状態をアクティブと呼び，そうでない状態を非アクティブと呼ぶ．CTIC モデルでは，非アクティブからアクティブへ状態は変わるが，逆は起こらない．アクティブなノード v は，各出リンクを通し独立に子ノード集合 $F(v)$ の各ノードを確率 κ $(0 \leq \kappa \leq 1)$ でアクティブにすることができる．この情報拡散試行が行われるのは一度限りで，時刻 t_v で子ノード w をアクティブにする試行に成功したとき，w がアクティブになる時刻は指数分布 $p(t) = r \exp(- r(t - t_v))$ で与えられるとする．r は指数分布のパラメータを表す．なお，リンク毎に，拡散確率や指数分布パラメータが異なるように一般化したモデルも同様に定義できる．

2.2. 分析手法

分析は現実のブログネットワークを用いた拡散シミュレーションによって行う．詳細には，最初に情報源ノードと各リンクの拡散確率 κ と時間遅れパラメータ r を任意に決定し，これを真のパラメータとして拡散シミュレーションを行う．この結果得られた拡散データを観測データとする．この観測データより平均の影響度を求め，これを真の影響度と仮定する．私たちはこの真の影響度からの近さをもって，パラメータの頑健性を主張する．次に，先の拡散データより，パラメータ(κ,r)を推定する．推定の方法は，先行研究の推定法[1]に従う．最後に，推定値(κ,r)より，再び拡散シミュレーションを行い，各推定値(κ,r)における期待影響度を求める．ここで求められた期待影響度と最初に仮定した真の影響度との差を評価し，頑健性を分析する．

3. 実験

3.1. 実験データと設定

今回の実験には，ブログのトラックバックネットワークのデータを利用した．ノード数は 12,047，リンク数は 79,920 であった．

今回は真のパラメータを拡散確率 κ=0.1，時間遅れパラメータ r=1.0 とし，情報源をこのネットワークにおける最も期待影響度の高いノードに設定し，実験を行った．

3.2. 実験結果と考察

図 1 は κ=0.1，r=1.0 で拡散シミュレーションした結果である．横軸は時刻を表し，縦軸は時刻 t までにアクティブになったノードの総数を表す．私たち

---

*The Prediction of The Expected Number of Influential Nodes by Estimating Parameters from Diffusion Data
† Yuya Yoshikawa, Kazumi Saito・University of Shizuoka
‡ Hiroshi Motoda・Osaka University
§ Masahiro Kimura・Ryukoku University
※Kouzou Ohara・Aoyama Gakuin University

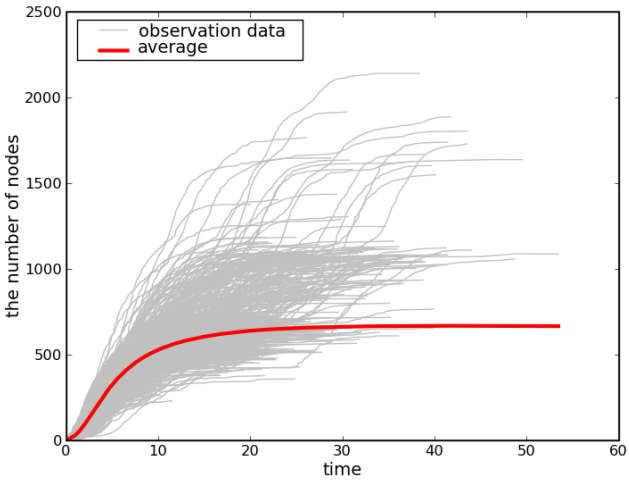

図 1　観測データと真の影響度

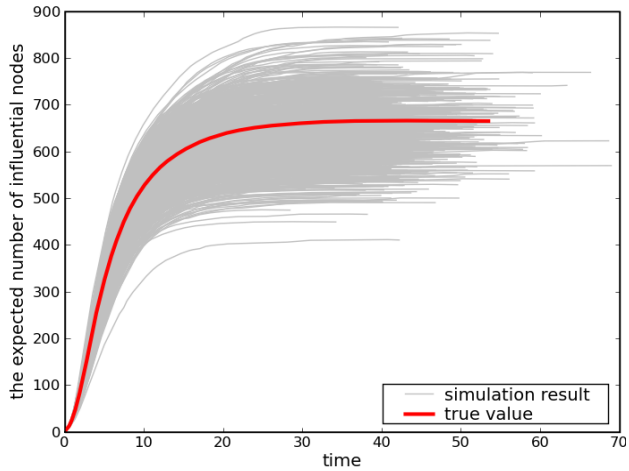

図 3　推定パラメータによる拡散シミュレーション

は 1 本の細線を観測データと仮定する．観測データを見ると，最終的に 300 程度のノードしかアクティブにならないときもあれば，2000 以上のノードがアクティブになるときもあり，同じパラメータからでも幅のあるデータが得られることが分かる．そして，太線は細線の平均値であり，これが真の影響度曲線だと仮定する．

図 2 は各観測データより推定したパラメータの散布図で，縦軸に拡散確率 κ，横軸に時間遅れパラメータ r としている．また，図の中央に見えるバツ印は真のパラメータである κ=0.1，r=1.0 の点である．パラメータ推定により，ほとんどの点は真のパラメータである κ=0.1，r=1.0 の点の周りに散らばることが分かる．

図 3 は推定値のパラメータにより，拡散シミュレーションを行った結果である．細線は図 1 同様にアクティブノードの分布を表すが，今回は 1 本の線が期待影響度曲線（100 回の平均値）である．最終的な期待影響度の最大値が 800 程度，最小値が 400 程度であり，この結果から，図 3 は図 1 に比べて，真の影響度曲線である太線から非常に近いところに分布していることが分かる．

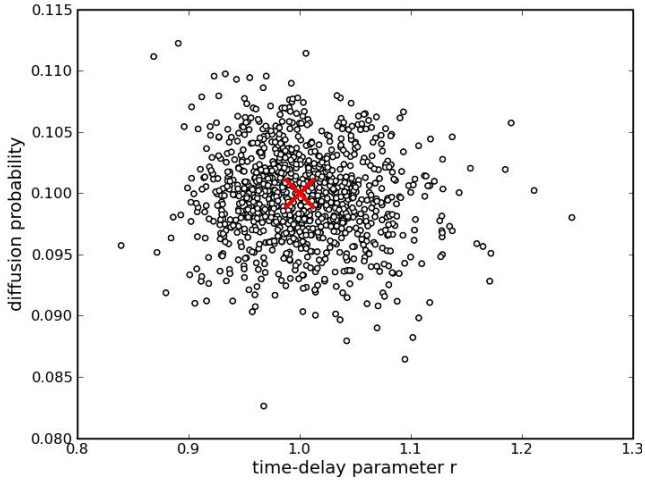

図 4　真の影響度からの誤差

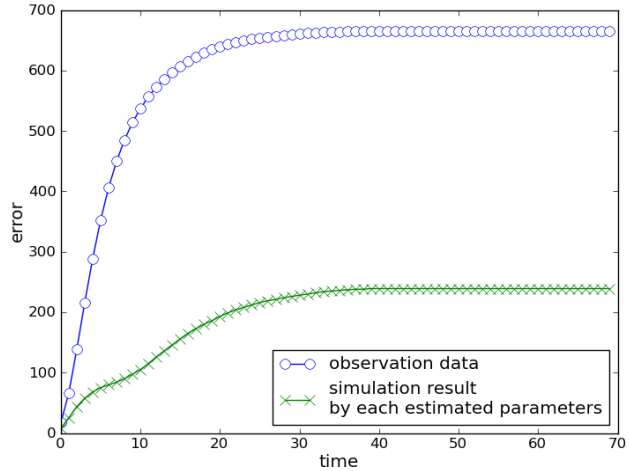

図 2　観測データによるパラメータ推定値

図 4 は，図 1 と図 3 の拡散データの結果と真の影響度曲線との誤差を表した図である．横軸は時刻，縦軸は各時刻における真の影響度との RMSE（Root Mean Squared Error）を表す．丸印は図 1 の観測データ，×印は図 3 の拡散データであるが，これを見ると明らかに図 3 の拡散データの方が真の影響度から近いことが分かる．

4. おわりに

本論文では，実際の社会ネットワーク分析で確からしい期待影響度を得られるかどうかの検証として，人工的に作った観測された拡散データより推定したパラメータの頑健性について分析した．今回の実験では，パラメータ推定により期待影響度は真の影響度により近づくことが示唆された．

参考文献

[1] Kazumi Saito, Masahiro Kimura, Kouzou Ohara, Hiroshi Motoda,“Learning Continuous-Time Information Diffusion Model for Social Behavioral Data Analysis”, First Asian Conference on Machine Learning, pp.322-337,(2009)