論　文

# 情報拡散モデルに基づくソーシャルネットワーク上でのノードの期待影響度曲線推定法

吉川　友也[†]　斉藤　和巳[†]　元田　浩[††]　大原　剛三[†††]
木村　昌弘[††††]

Estimating Method of Expected Influence Curve from Single Diffusion Sequence on Social Networks

Yuya YOSHIKAWA[†], Kazumi SAITO[†], Hiroshi MOTODA[††], Kouzou OHARA[†††], and Masahiro KIMURA[††††]

**あらまし**　本論文では，非同期時間遅れ付き独立カスケード (AsIC) モデルと非同期時間遅れ付き線形閾値 (AsLT) モデルのそれぞれの場合を仮定して，観測した単一の拡散系列から各時刻における期待影響度 (期待影響度曲線) を高精度で推定する問題に取り組む．単純な方法として，観測した拡散系列のアクティブノード数を数えて期待影響度曲線とすることが考えられるが，拡散系列は情報拡散の確率的な動作によって多様な結果になるため，この方法での期待影響度曲線推定には本質的な限界がある．本論文の提案法では，観測した拡散系列から各モデルのパラメータを EM アルゴリズムによって学習し，学習したモデルパラメータを使って，シミュレーションによって期待影響度曲線を推定する．提案法を評価するために，現実のソーシャルネットワーク構造データを用いて人工的に拡散系列を生成して評価実験を行う．生成される拡散系列の長さは，同じ条件であっても多様な長さになる．我々は，提案法を使うことによって，多様な長さの拡散系列からでも期待影響度曲線を高精度で推定できることを示す．

**キーワード**　社会ネットワーク分析，情報拡散モデル，ソーシャルネットワーク，期待影響度曲線

## 1. まえがき

インターネットや World Wide Web の興隆は，大規模なソーシャルネットワークの発生を加速させており，情報を普及させるための重要メディアとして，最近ソーシャルネットワークが注目されている [1] 〜 [5]．噂やトピック，イノベーションなどの情報は，クチコミによってソーシャルネットワークを介して人から人へと伝播する．このような現象の下で，情報がどのように広がり，何人の人に情報が伝わるのかを知ることは重要であり，特に，情報が伝わった人数を表す影響度に関する予測技術を創ることは，効果的なクチコミマーケティングのヒントになり得る．本論文では，ソーシャルネットワーク上の情報拡散の予測技術の 1 つとして，情報拡散モデルを使ったノードの期待影響度曲線の獲得に焦点をあてて研究を行う．

最近の研究において広く使われる情報拡散モデルは，独立カスケード (IC: Independent Cascade) モデル [6] 〜 [8] と線形閾値 (LT: Linear Threshold) モデル [9], [10] である．これらは影響最大化問題 [7], [11] などの解決のために使われている．IC モデルと LT モデルは，それぞれ異なった情報拡散の解釈を与えている．IC モデルは送信者中心型であり，アクティブノードが隣接する非アクティブなノードに対して，リンクに割り当てられた拡散確率に従って独立に影響を与える．他方，LT モデルは受信者中心型で，受け取った重みの合計がそのノードに割り当てられた閾値を超え

† 静岡県立大学経営情報学部，静岡県
School of Management and Information, University of Shizuoka
†† 大阪大学産業科学研究所，大阪府
Institute of Scientific and Industrial Research, Osaka University
††† 青山学院大学理工学部，神奈川県
College of Science and Engineering, Aoyama Gakuin University
†††† 龍谷大学理工学部，京都府
Faculty of Science and Technology, Ryukoku University

たときアクティブになるようにモデル化されている．これらのモデルは事前にパラメータを割り当てる必要があり，これは IC モデルでは拡散確率，LT モデルでは重みである．しかし，これらの真のパラメータ値は実際には分からない．よって，情報拡散結果の集合からパラメータを推定する別の問題をもたらすことになる.IC モデルやそれの情報伝播の非同期時間遅れを考慮する改良型モデル (以下，AsIC モデル) でのパラメータ推定法は既に提案されており，LT モデルにおいても同様に，情報伝播の非同期時間遅れを考慮する改良型モデル (以下，AsLT モデル) でのパラメータ推定法が存在する [3], [12] 〜 [14] ．本論文では, [13], [14] のパラメータ推定法を用いる．

ある 1 つの情報源ノードから情報が拡散し，我々がそれを観測したとする．この観測したデータのことを，拡散系列と呼ぶことにする．この拡散系列よりどのようにして期待影響度を推定すればいいのだろうか？これこそがまさに我々が解決したい問題である．まず，観測した拡散系列は期待影響度の粗い知識とみなすことができる．なぜなら，この拡散系列における時刻 $t$ までに影響を受けたノードの数を数えることができるからである．しかし，情報拡散モデルは確率的な動作をするため，この拡散系列は生成されたモデルパラメータの値が既知の場合でさえ，シミュレーション (試行) ごとに，影響を受けるノード数が大きく異なる.従って，観測した単一の拡散系列をそのまま期待影響度とすることは望ましくない．

本論文では，我々は AsIC モデルと AsLT モデルのどちらかによってネットワーク上を情報が拡散すると仮定し，最初に，観測した単一の拡散系列からモデルのパラメータを学習し，学習したモデルを使って期待影響度を推定する方法を提案する．評価実験では，提案法によって精度良く期待影響度曲線を推定できているかを確かめるために，現実のソーシャルネットワークの構造データを用いたシミュレーション実験を行う．

## 2. 情報拡散モデル

IC モデルについて簡単に触れた後，本論文で使用する AsIC モデルについて説明する．同様に，LT モデルと AsLT モデルについて説明する．まず，$G = (V, E)$ を自己リンクなしの有向ネットワークとする．$V$ はノード集合，$E(\subset V \times V)$ はリンク集合を表す．ノードが情報を保持している状態をアクティブ，そうでない状態を非アクティブと呼ぶ．ノードは非アクティブからアクティブへ一度だけ状態が変わることができ，アクティブから非アクティブへ状態が変わることはないものとする．今，初期値としてアクティブノード集合 $S$ が与えられたとし，その他のノードはすべて非アクティブとする．$(u, v) \in E$ のとき，ノード $u$ はノード $v$ の親ノードと呼び，ノード $v$ はノード $u$ の子ノードと呼ぶ．各ノード $v \in V$ に対して，$v$ の子ノード集合 $F(v)$ と親ノード集合 $B(v)$ を以下のように定義する．

$$F(v)=\{w\in V;(v,w)\in E\},\ B(v)=\{u\in V;(u,v)\in E\}$$

### 2.1 独立カスケード (IC) モデル

IC モデルは感染病の広がり方を示す基本的な確率モデルである．このモデルでは，各リンク $(u, v)$ に対して前もって実数値 $\kappa_{u,v}(0 < \kappa_{u,v} < 1)$ を割り当てる.ここで，$\kappa_{u,v}$ をリンク $(u, v)$ 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 $t \geqq 0$ 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 $u$ が時刻 $t$ でアクティブになったとき，ノード $u$ には現在非アクティブの子ノード $v$ 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 $\kappa_{u,v}$ で成功する．成功したら，ノード $v$ は時刻 $t + 1$ でアクティブになる．もし，ノード $v$ の複数の親ノードが時刻 $t$ で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．この拡散過程は，新たにアクティブになるノードがなくなったときに終了する．

### 2.2 非同期時間遅れ付き独立カスケード (AsIC) モデル

非同期時間遅れを考慮するように拡張した IC モデルを，非同期時間遅れ付き独立カスケード (AsIC) モデルと呼ぶ．AsIC モデルでは，各リンクには IC モデルと同様の拡散確率 $\kappa_{u,v}$ と，新たに時間遅れパラメータ $r_{u,v}(> 0)$ が割り当てられる．

AsIC モデルでの拡散過程は連続時間 $t$ 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．今，ノード $u$ が時刻 $t$ でアクティブになったとする．その後，ノード $u$ には現在非アクティブの子ノード $v$ 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る．ここでパラメータ $r_{u,v}$ の指数分布から情報到達の時間間隔 $\delta$ を決定する．もしノード $v$ が時刻 $t+\delta$ までにアクティブになっていなかったら，ノード $u$ はノード $v$ をアクティブにする試行を行う．ノード $u$ が拡散に成功した場合，ノード $v$ は時刻 $t + \delta$ でアクテ

ィブになる．連続時間モデルのもとではノード $v$ の複数の親ノードが同じ時刻にノード $v$ をアクティブにすることは考えにくい．よって，ここではこの可能性は無視する．この拡散過程は，新たにアクティブになるノードがなくなったときに終了する．

情報源ノード $v$ に対して，ある時刻 $t$ までのアクティブノードの総数を $\phi(t;v)$ と表す．また，$\phi(t;v)$ の期待値を $\sigma(t;v)$ と表し，$\sigma(t;v)$ のことを AsIC モデルにおけるノード $v$ の期待影響度曲線と呼ぶ．

### 2.3 線形閾値 (LT) モデル

LT モデルはイノベーションの広がりを表す基本的な確率モデルである．このモデルでは，すべてのノード $v \in V$ に対して，$\sum_{u\in B(v)} \omega_{u,v} \leqq 1$ となるように前もってリンク $(u,v)$ に重み $\omega_{u,v}(>0)$ 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 $v \in V$ に対して，閾値 $\theta_v$ 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 $v$ は各親ノード $u \in B(v)$ から $\omega_{u,v}$ の影響を受ける．もし，ノード $v$ のアクティブな親ノードから受けた重みの合計が $\theta_v$ 以上になった場合，ノード $v$ は時刻 $t+1$ でアクティブになる．この拡散過程は，新たにアクティブになるノードがなくなったときに終了する．

### 2.4 非同期時間遅れ付き線形閾値 (AsLT) モデル

非同期時間遅れを考慮した LT モデルを考える．このモデルを非同期時間遅れ付き線形閾値 (AsLT) モデルと呼ぶ．AsLT モデルでは，重み集合 $\{\omega_{u,v}\}$ に加えて，各ノード $v \in V$ に対して実数値 $r_v(>0)$ を前もって割り当てる．ここで $r_v$ をノード $v$ の時間遅れパラメータと呼び，$r_v$ はノード $v$ にのみ依存する．これはすなわち，アクティブになる条件を満足した時点でいつ情報を受け取るかはノード $v$ が決定するということである．

AsLT モデルでの拡散過程は連続時間 $t$ で展開され，初期アクティブノード集合 $S$ が与えられた上でランダムルールに従って行われる．最初，ノード $v$ のアクティブな親ノードの重みの合計は時刻 $t$ において閾値 $\theta_v$ 以下になっているとする．その後，ノード $v$ は時刻 $t+\delta$ でアクティブになる．ここで，$\delta$ はパラメータ $r_v$ の指数分布より計算される．また，たとえノード $v$ の非アクティブな親ノードが時刻 $t$ と $t+\delta$ の間にアクティブになった場合でも，ノード $v$ のアクティブになった時刻に影響は与えないとする．その他の拡散メカニズムは LT モデルと同様である．AsIC と同様に，AsLT モデルでの情報源ノード $v$ の期待影響度曲線を $\sigma(t;v)$ と定義する．

## 3. パラメータ学習アルゴリズム

AsIC モデルにおいて，時間遅れパラメータベクトル $\boldsymbol{r} = (r_{u,v})_{(u,v)\in E}$ と拡散パラメータベクトル $\boldsymbol{\kappa} = (\kappa_{u,v})_{(u,v)\in E}$ を定義する．AsLT モデルにおいても同様に，重みベクトル $\boldsymbol{\omega} = (\omega_{u,v})_{(u,v)\in E}$ と $\boldsymbol{r} = (r_v)_{v\in V}$ を定義する．実際にはこれらパラメータの真の値は分からない．したがって，情報拡散の結果よりパラメータを学習しなければならない．

まず，$M$ 個の独立な情報拡散結果が得られたとし，これを $\{D_m; m = 1,\cdots,M\}$ とする．ここで,各 $D_m$ は $m$ 回目の情報拡散結果におけるアクティブになったノードとその時刻のペアの集合で，$D_m = \{(u,t_{m,u}),(v,t_{m,v}),...\}$ である．各 $D_m$ に対して，初期時刻 $t_m = \min\{t_{m,v};(v,t_{m,v}) \in D_m\}$ と，最終時刻 $T_m \geqq \max\{t_{m,v};(v,t_{m,v}) \in D_m\}$ を計算する．ここで，$T_m$ は必ずしも観測した最後のアクティブ時刻と同じである必要はない．今後，観測データのことを $\mathcal{D}_M = \{(D_m,T_m);\ m = 1,\cdots,M\}$ と表現する．任意の $t \in [t_m,T_m]$ に対して，$C_m(t) = \{v;(v,t_{m,v}) \in D_m, t_{m,v} < t\}$ と定義する．これはつまり，$m$ 回目の情報拡散結果における時刻 $t$ までのアクティブノード集合である．また，簡便のために，$C_m$ を $m$ 回目の情報拡散結果の全てのアクティブノード集合とする．その上で，最低一つのアクティブな親ノードを持った非アクティブノード集合を $\partial C_m = \{v;\ (u,v) \in E,\ u \in C_m,\ v \notin C_m\}$ として表す．各ノード $v \in C_m \cup \partial C_m$ に対して，ノード $v$ をアクティブにするチャンスがあった親ノードの部分集合を以下で定義する．

$$\mathcal{B}_{m,v} = \begin{cases} B(v) \cap C_m(t_{m,v}) & \text{if } v \in C_m \\ B(v) \cap C_m & \text{if } v \in \partial C_m \end{cases}$$

与えられた観測データ $\mathcal{D}_M$ より AsIC モデルのパラメータベクトルの $\boldsymbol{r}$ と $\boldsymbol{\kappa}$，または AsLT モデルの $\boldsymbol{r}$ と $\boldsymbol{\omega}$ の値を学習するために，我々は [13] と [14] で提案された学習法を採用する．以下で簡単に方法について説明する．

### 3.1 AsIC モデルのパラメータ学習法

観測データ $\mathcal{D}_m$ から AsIC モデルの $\boldsymbol{r}$ と $\boldsymbol{\kappa}$ の値を学習するために，目的関数である $\boldsymbol{r}$ と $\boldsymbol{\kappa}$ の尤度関数

$\mathcal{L}(\boldsymbol{r}, \boldsymbol{\kappa}; \mathcal{D}_m)$ を導出する．最初に，$m$ 回目の情報拡散結果に対して $t_{m,v} > t_m$ であるノード $v \in C_m$ について考える．ノード $u \in B(v) \cap C_m(t_{m,v})$ が時刻 $t_{m,v}$ にノード $v$ をアクティブにする確率密度 $\Phi_{m,u,v}$ は

$$\Phi_{m,u,v} = \kappa_{u,v} r_{u,v} \exp(-r_{u,v}(t_{m,v} - t_{m,u})) \quad (1)$$

である．また，期間 $[t_{m,u}, t_{m,v}]$ の間にノード $v$ がノード $u \in B(v) \cap C_m(t_{m,v})$ からアクティブにされない確率 $\Psi_{m,u,v}$ は

$$\Psi_{m,u,v} = 1 - \kappa_{u,v} \int_{t_{m,u}}^{t_{m,v}} r_{u,v} \exp(-r_{u,v}(t - t_{m,u}))dt \quad (2)$$

である．2.2 節で説明したように，たとえ $\eta = |B(v) \cap C_m(t_{m,v})| > 1$ であっても複数のアクティブな親ノードによって同時にアクティブにされることを考える必要はない．したがって，ノード $v$ が時刻 $t_{m,v}$ でアクティブにされる確率密度 $h_{m,v}^{(IC)}$ は以下のように表せる．

$$h_{m,v}^{(IC)} = \sum_{u \in H(m,v)} \Phi_{m,u,v} \left( \prod_{x \in H(m,v) \setminus \{u\}} \Psi_{m,x,v} \right) \quad (3)$$

ただし，$H(m,v) = B(v) \cap C_m(t_{m,v})$ とする．ここで，どのノードがノード $v$ をアクティブにしたかを知ることは出来ない．これは隠れ変数とみなすことができる．

次に，$m$ 回目の情報拡散結果において $v \in C_m$ かつ $w \notin C_m$ であるリンク $(v,w) \in E$ について考える．ノード $w$ が期間 $[t_m, T_m]$ 内でノード $v$ によってアクティブにされない確率 $g_{m,v,w}^{(IC)}$ は以下の式で表せる．

$$g_{m,v,w}^{(IC)} = \kappa_{v,w} \exp(-r_{v,w}(T_m - t_{m,v})) + (1 - \kappa_{v,w}) \quad (4)$$

その結果，式 (3) と式 (4) を使って，$\boldsymbol{r}$ と $\boldsymbol{\kappa}$ の尤度関数 $\mathcal{L}(\boldsymbol{r}, \boldsymbol{\kappa}; \mathcal{D}_m)$ を以下のように定義する．

$$\mathcal{L}(\boldsymbol{r}, \boldsymbol{\kappa}; \mathcal{D}_M) = \prod_{m=1}^{M} \prod_{v \in C_m} \left( h_{m,v}^{(IC)} \prod_{w \in F(v) \setminus C_m} g_{m,v,w}^{(IC)} \right) \quad (5)$$

したがって，ここで解決すべき問題は，式 (5) を最大化する時間遅れパラメータベクトル $\boldsymbol{r}$ と拡散確率ベクトル $\boldsymbol{\kappa}$ を見つけることとなる．ここで我々は $\boldsymbol{r}$ と $\boldsymbol{\kappa}$ を得るために，安定した解を得ることが出来る EM アルゴリズムに基づくパラメータ学習法を用いる [13] .

### 3.2 AsLT モデルのパラメータ学習法

観測データ $\mathcal{D}_m$ から AsLT モデルの $\boldsymbol{r}$ と $\boldsymbol{\omega}$ の値を学習するために，目的関数である $\boldsymbol{r}$ と $\boldsymbol{\omega}$ の尤度関数 $\mathcal{L}(\boldsymbol{r}, \boldsymbol{\omega}; \mathcal{D}_m)$ を導出する．簡便のために，$\omega_{v,v} + \sum_{u \in B(v)} \omega_{u,v} = 1$ となるような値調整の重み $\omega_{v,v}$ を導入する．ここで，重み $\omega_{v,v}$ はノード $v$ のアクティベーションに貢献することはないことに注意する．さらに，閾値 $\theta_v$ は区間 $[0,1]$ から一様ランダムに選ばれるため，各重み $\omega_{*,v}$ は多項確率とみなすことができる．

$m$ 回目の情報拡散結果において，ノード $v$ が時刻 $t_{m,v}$ でアクティブになったと仮定する．そのとき,ノード $v$ のある親ノード $u \in \mathcal{B}_{m,v}$ がアクティブになった時刻で，ノード $v$ のアクティブな親ノードから受け取る重みの合計が $\theta_v$ 以上になったということがわかる．しかし，$|\mathcal{B}_{m,v}| > 1$ の場合，非同期時間遅れがあるためにどのノードから影響を受けたかを正確に知る術はない．また，ノード $\xi \in \mathcal{B}_{m,v}$ がアクティブになったとき，ノード $v$ はアクティブになったとする．その場合，$\theta_v$ は $\sum_{u \in B(v) \cap C_m(t_{m,\xi})} \omega_{u,v}$ と $\omega_{\xi,v} + \sum_{u \in B(v) \cap C_m(t_{m,\xi})} \omega_{u,v}$ の範囲の値となる．これは要するに，$\theta_v$ がこの範囲から選ばれる確率が $\omega_{\xi,v}$ ということである．ここで，異なるアクティブな親ノードに関するそれぞれの事象は互いに素である．したがって，ノード $v$ が時刻 $t_{m,v}$ でアクティブにされる確率密度 $h_{m,v}^{(LT)}$ は以下の式で表せる．

$$h_{m,v}^{(LT)} = \sum_{u \in \mathcal{B}_{m,v}} \omega_{u,v} r_v \exp(-r_v(t_{m,v} - t_{m,u})) \quad (6)$$

ただし，$t_{m,v} = t_m$ の場合は，$h_{m,v}^{(LT)} = 1$ となる．

次に，$m$ 回目の情報拡散結果において，$\partial C_m = \{w; (v,w) \in E \wedge v \in C_m(T_m) \wedge w \notin C_m(T_m)\}$ に属するノード $w \in V$ について考える．ノード $v$ が区間 $[t_m, T_m]$ でアクティブにされない確率を $g_{m,v}$ で表し，以下のように計算できる．

$$g_{m,v}^{(LT)} = 1 - \sum_{u \in \mathcal{B}_{m,v}} \omega_{u,v} \int_{t_{m,v}}^{T_m} r_v \exp(-r_v(t - t_{m,v}))dt \quad (7)$$

よって，式 (6) と式 (7) を使って，$\boldsymbol{r}$ と $\boldsymbol{\omega}$ の目的関数 $\mathcal{L}(\boldsymbol{r}, \boldsymbol{\omega}; \mathcal{D}_M)$ を以下のように定義する．

$$\mathcal{L}(\boldsymbol{r},\boldsymbol{\omega};\mathcal{D}_M) = \prod_{m=1}^{M}\left(\prod_{v\in C_m} h_{m,v}^{(LT)}\right)\left(\prod_{v\in \partial C_m} g_{m,v}^{(LT)}\right) \quad (8)$$

ここでの問題は，式 (8) を最大化する時間遅れパラメータベクトル $\boldsymbol{r}$ と重みパラメータベクトル $\boldsymbol{\omega}$ を見つけることであり，AsIC モデルの場合と同様に，EM アルゴリズムに基づく学習法によって計算する [14] .

## 4. 期待影響度推定法

これまで，時間遅れや拡散のパラメータはノードやリンク毎に様々であると仮定してきた．しかし，我々は単一の拡散系列から期待影響度曲線を推定する問題に取り組んでいるため，観測できるデータの数が圧倒的に少ない．よって，観測データに対する過学習を避けるために，ネットワーク $G$ の全てのノードやリンクで同じパラメータを使うという制約を置くことにする．したがって，AsIC モデルの場合には全てのリンク $(u,v)\in E$ に対して，$r_{u,v}=r$，$\kappa_{u,v}=\kappa$ とし，AsLT モデルの場合には全てのノード $v\in V$ とリンク $(u,v)\in E$ について，$r_v=r$，$\omega_{u,v}=\kappa|B(v)|^{-1}$ と設定する．ここで，$\kappa$ の定義域は $0<\kappa<1$ であり，$\omega_{v,v}=1-\kappa$ となる．すなわち，AsLT モデルのパラメータ $\kappa$ は拡散確率の一種として解釈できるため,AsIC モデルと同じ記号を使うことにする．この制約は，観測できる唯一の系列がネットワーク $G$ のごく一部のリンクで構成されるため，パラメータを学習する上で必要不可欠である．

AsIC モデルと AsLT モデルの下での期待影響度曲線を推定する方法について説明する．最初に，以下の時刻 $t_0$ で情報源ノード $v_0$ の単一の拡散系列 $d$ が観測されたとする．

$$d=\{(v_0,t_0),(v_1,t_1),\cdots,(v_T,t_T)\}$$

まず，3.1 節と 3.2 節で説明した方法を使い，観測した単一の拡散系列 $d$ からモデルパラメータ $r$ と $\kappa$ のペアを学習する．次に，2.2 節と 2.4 節で説明したモデルを使い，以下の $K$ 個の人工拡散系列群を生成する．

$$s_k=\{(v_0,t_0),(v_{k,1},t_{k,1}),\cdots,(v_{k,T},t_{k,T})\} \quad (k=1,\cdots,K)$$

ここで，情報源ノード $v_0$ とその時刻 $t_0$ は全ての系列で同じである．しかし，アクティブノードの最終時刻 $\{t_{k,T}\}$ やアクティブノード数 $\{|s_k|\}$ については，後の実験で示すように，広範囲で多様な値になる．最後に，生成した人工拡散系列群 $\mathcal{S}=\{s_1,...,s_K\}$ を用いて，式 (9) で期待影響度曲線 $\sigma(t,v_0,d)$ を推定する．

$$\sigma(t;v_0,d)=\frac{1}{K}\sum_{k=1}^{K}|\{(v,\tau)\in s_k;\tau\leqq t\}| \quad (9)$$

この推定法は最初に，一つの観測した拡散系列 $d$ とその拡散が起きたネットワーク構造 $G$，拡散シミュレーション回数 $K$ の 3 種類の値を入力する．そして最終的に，期待影響度曲線 $\sigma(t;v_0,d)$ を出力する．以上で説明したアルゴリズムの内容をまとめると，以下のようになる．

期待影響度推定法

**step.1** $r$ と $\kappa$ を拡散系列 $d$ より学習する
**step.2** 推定した $r$ と $\kappa$ を使って $K$ 回拡散シミュレーションを行い，$\mathcal{S}=\{s_1,\cdots,s_K\}$ を生成する
**step.3** $\mathcal{S}$ の平均値である期待影響度曲線 $\sigma(t;v_0,d)$ を計算する

今回の実験では，拡散シミュレーション回数 $K$ を 100 に設定した．

## 5. 実　　験

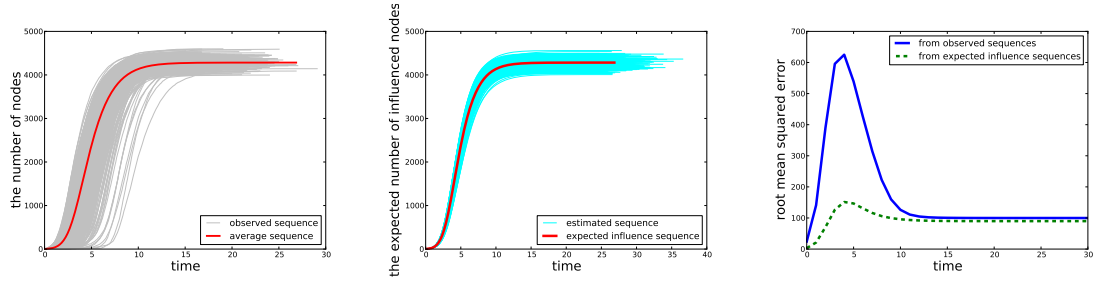
現実の大規模なソーシャルネットワーク構造を用いて，提案した推定法の実用性を評価する．

### 5.1 実験手順

今回行った実験手順は以下の通りである．

実験手順

**proc.1** 情報拡散モデル (AsIC または AsLT )，真のパラメータのペア $(r^*,\kappa^*)$，時刻 $t_0$ の情報源ノード $v_0$ の 3 つの値を決定する
**proc.2** **proc.1** の設定の下で，$N$ 個の拡散系列 $d_n$ を生成し，拡散系列群 $D=\{d_1,\cdots,d_N\}$ とする．
**proc.3** 真の期待影響度曲線 $\sigma^*(t;v_0)$ と，各 $d_n\in D$ の影響度曲線 $\phi(t;v_0,d_n)$ を計算する
**proc.4** 4 章の提案法を使い，各 $d_n\in D$ ごとに期待影響度曲線 $\sigma(t;v_0,d_n)$ を推定する
**proc.5** 推定した期待影響度曲線を評価するために，RMSE 曲線 $E_C$，$E_D$ を計算する




(a) 生成した拡散系列の影響度曲線　(b) 推定した期待影響度曲線　(c) 推定精度比較

図 1　AsIC モデルの下でのコスメネットワークの結果 ($\kappa^* = 0.05$)
Fig. 1　Result set of Cosme network under the AsIC model($\kappa^* = 0.05$)

現実には，観測によって真の期待影響度曲線を取得することは殆どの場合不可能である．したがって今回は，**proc.1** で仮定した情報拡散モデル (AsIC または AsLT) と真のパラメータ $(r^*, \kappa^*)$ を使った人工データによって実験を行う．ここで，真のパラメータとは,Kempe ら [7] などの研究で行われているように，人工的に拡散データを生成するために我々が真と仮定して設定したパラメータのことである．次に，真のパラメータを設定したモデルで拡散シミュレーションを行い，$N$ 個の人工拡散系列群 $D = \{d_1, \cdots, d_N\}$ を生成する (**proc.2**)．$D$ に対して式 (10) を適用することで，経験的な真の期待影響度曲線 $\sigma^*(t; v_0)$ を計算する．

$$\sigma^*(t; v_0) = \frac{1}{N} \sum_{n=1}^{N} |\{(v, \tau) \in d_n; \tau \leqq t\}| \qquad (10)$$

ここで，各 $d_n \in D$ に対して，影響度曲線 $\phi(t; v_0, d_n) = |\{(v, \tau) \in d_n; \tau \leqq t\}|$ も計算しておく (**proc.3**)．各 $d_n \in D$ を一つの観測した拡散系列とみなして，4 章の提案法を用い期待影響度曲線 $\sigma(t; v_0, d_n)$ を推定する (**proc.4**)．最後に，**proc.5** で二乗平均平方根誤差曲線 (RMSE)$E_C(t)$ によって提案法で推定した期待影響度曲線の平均の精度を計算し，さらに，影響度曲線 $\phi(t; v_0, d_n)$ における RMSE 曲線 $E_D(t)$ と $E_C(t)$ を比較する．ここで，RMSE 曲線 $E_C(t)$，$E_D(t)$ は以下のように定義する．

$$E_C(t) = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \big(\sigma(t; v_0, d_n) - \sigma^*(t; v_0)\big)^2}$$

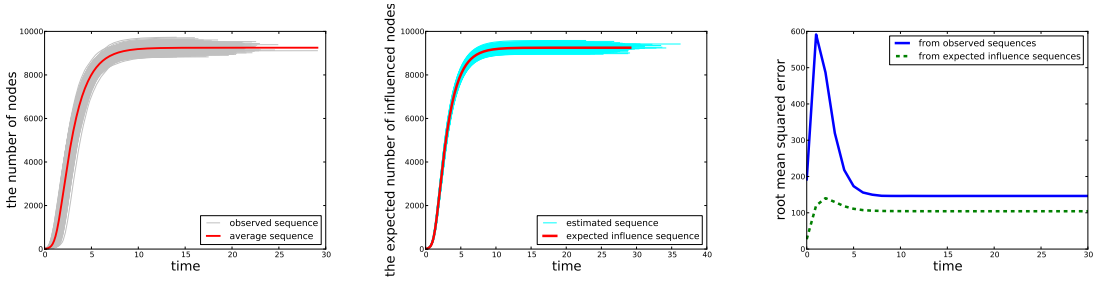
$$E_D(t) = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \big(\phi(t; v_0, d_n) - \sigma^*(t; v_0)\big)^2}$$

### 5.2　実験データと設定

今回の実験では，ソーシャルネットワークの主要な特性を持つ 2 つの大規模なネットワークデータを用いる．1 つ目は，化粧品の口コミサイト「*@cosme*(注1)」におけるユーザ間のお気に入り関係のネットワークで,あるユーザ X がユーザ Y をお気に入り登録をしたとき，X から Y へ有向リンクを張って構築した．このネットワークは，2009 年 12 月にランダムに選択したユーザから 10 段辿って収集し，ノード数は 45,024 でリンク数は 351,299 である．以下では，このネットワークをコスメネットワークと呼ぶことにする．2 つ目は，ブログサービス「*Ameba* ブログ(注2)」の読者ネットワークである．*Ameba* ブログには，ブロガーが他のブログをお気に入りブログとしてハイパーリンクを張る機能があり，ブロガーがお気に入りブログを定期的に見ると仮定すれば，これはブロガー間のリンクとみなすことができる．2006 年 6 月に *Ameba* ブログの 117,374 ブログのお気に入りブログを取得し，連結成分を抽出した．このネットワークのノード数は 56,604，リンク数は 734,737 である．以下では，このネットワークをアメーバブログネットワークと呼ぶことにする．

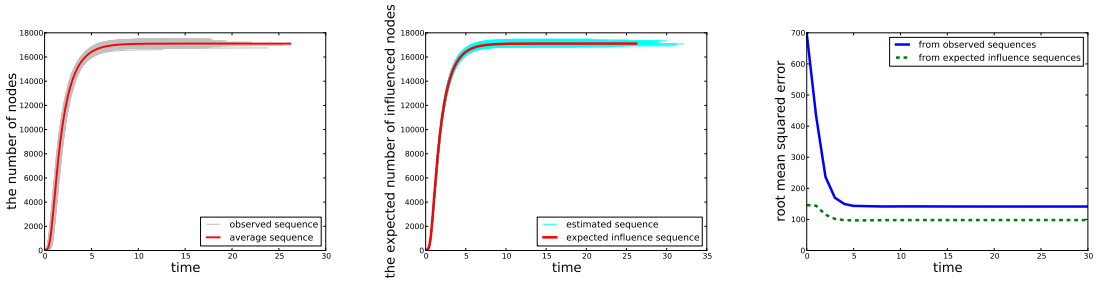
真のパラメータ $(r^*, \kappa^*)$ は，AsIC モデルの拡散確率の設定については，Kempe ら [7] が行った実験の設定である $\kappa^* = 0.1$ と 0.01 を参考にする．しかし,本評価実験に用いるネットワークでは，$\kappa^* = 0.01$ とすると，情報源ノードから情報が殆ど拡散しない場合が見られた．そこで以下の評価実験では，統一的に $\kappa^* = 0.1$ と 0.05 の設定を採用する．時間遅れパラメータ $r^*$ については，$r^*$ を変更することは時間のスケールを変える

(注1)：http://www.cosme.net/
(注2)：http://ameblo.jp/




(a) 生成した拡散系列の影響度曲線　(b) 推定した期待影響度曲線　(c) 推定精度比較

図 2　AsIC モデルの下でのアメーバブログネットワークの結果 ($\kappa^* = 0.05$)
Fig. 2　Result set of Ameba blog network under the AsIC model($\kappa^* = 0.05$)




(a) 生成した拡散系列の影響度曲線　(b) 推定した期待影響度曲線　(c) 推定精度比較

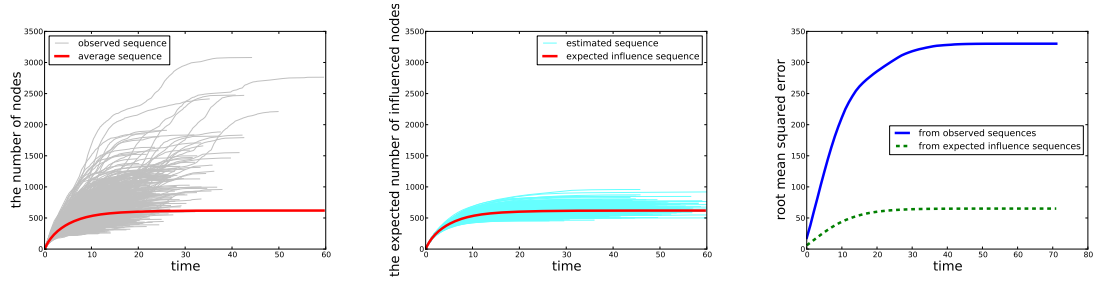
図 3　AsIC モデルの下でのアメーバブログネットワークの結果 ($\kappa^* = 0.1$)
Fig. 3　Result set of Ameba blog network under the AsIC model($\kappa^* = 0.1$)

ことと同等なので，今回は簡単に 1.0 に設定した.次に AsLT モデルの真のパラメータ値を設定する．AsLT モデルは情報がなかなか広く伝わらなく，$\kappa^*$ が大きな値でないと適当な量の拡散が起きないため，$\kappa^*$ は 0.9 に設定した．また，時間遅れパラメータ $r^*$ は，AsIC モデルの場合と同様で 1.0 に設定した．

### 5.3　実 験 結 果

#### 5.3.1　コスメネットワーク上の AsIC モデル

図 1 は，パラメータ $(r^*, \kappa^*) = (1.0, 0.05)$ と設定した AsIC モデルにおいて，コスメネットワークを使った実験の結果である．最初に，図 1(a) は，同一の情報源ノードから拡散をスタートさせたときの結果で，$N = 1000$ 回分の拡散系列 $d_1, \cdots, d_N$ を表示している.横軸は時刻，縦軸はその時刻までのアクティブノード数を表す．この図より，各拡散系列を表す影響度曲線 (灰色) は，AsIC モデルの確率的な動作によって広範囲で多様な値を取ることが分かる．ここで我々の目的は，一つの拡散系列から，$N = 1000$ 本の灰色線の平均によって近似した期待影響度曲線 (赤 (黒)) を推定することである．以下では，赤色で示した期待影響度曲線のことを，真の期待影響度と呼ぶことにする．図 1(b) は，推定した期待影響度曲線 (シアン (灰色)) の結果である．具体的には，拡散系列からパラメータ $(r, \kappa)$ を推定し，その推定したパラメータを使って同じ情報源ノードから $K = 100$ 回の拡散シミュレーションを行うことによって推定した，各時刻における期待影響度 $\sigma(t; v_0, d_n)$ である．ここで，真の期待影響度曲線 (赤 (黒)) は図 1(a) のものと同じである．図 1(c) は，真の期待影響度曲線に対する拡散系列の影響度曲線 $\phi(t; v_0, d_n)$(図 1(a)) と推定した期待影響度曲線 $\sigma(t; v_0, d_n)$(図 1(c)) の二乗平均平方根誤差 (RMSE) の結果である．この結果を見ると，推定した期待影響度曲線の RMSE は，どの時刻を見ても安定して 100 ノード程度の誤差で推定できていることが分かる．また，時刻 $t = 4$ 付近では，推定した期待影響度曲線の RMSE は拡散系列の影響度曲線の RMSE の 1/6 程度であり，各時刻においても，平均すると推定した期待影響度曲線の RMSE の方が小さいことが見て取れる．したがって，推定した期待影響度曲線は拡散系列の影響度曲線よりも真の期待影響度曲線に近いと言える．




(a) 生成した拡散系列の影響度曲線　(b) 推定した期待影響度曲線　(c) 推定精度比較

図 4　AsLT モデルの下でのコスメネットワークの結果 ($\kappa^* = 0.9$)
Fig. 4　Result set of Cosme network under the AsLT model($\kappa^* = 0.9$)

| | $\kappa^*$ | $\varepsilon_\kappa$ | $r^*$ | $\varepsilon_r$ |
|---|---|---|---|---|
| アメーバブログ | 0.05 | 0.0004 | 1.0 | 0.0111 |
| | 0.1 | 0.0006 | 1.0 | 0.0081 |
| コスメ | 0.05 | 0.0006 | 1.0 | 0.0159 |
| | 0.1 | 0.0008 | 1.0 | 0.0104 |

表 1　AsIC モデルのパラメータの推定誤差
Table 1　The error values of parameter estimator under the AsIC model

| | $\kappa^*$ | $\varepsilon_\kappa$ | $r^*$ | $\varepsilon_r$ |
|---|---|---|---|---|
| アメーバブログ | 0.9 | 0.0133 | 1.0 | 0.0347 |
| コスメ | 0.9 | 0.0100 | 1.0 | 0.0199 |

表 2　AsLT モデルのパラメータの推定誤差
Table 2　The error values of parameter estimator under the AsLT model

$\kappa^* = 0.1$ の場合にも同様の結果が得られた．

**5.3.2**　アメーバブログネットワーク上の AsIC モデル

図 2 と図 3 は，それぞれ $\kappa^* = 0.05$ と $\kappa^* = 0.1$ と設定した AsIC モデルにおいて，アメーバブログネットワークを使った実験結果である．このネットワークによる実験でも，図 1 で見た結果と同様に，提案法による期待影響度曲線は時刻の早い段階で特に効果があることが見て取れる．また，提案法で推定した期待影響度曲線の RMSE は拡散系列の影響度曲線の RMSE に比べて小さくなることがわかる．

**5.3.3**　コスメネットワーク上の AsLT モデル

図 4 は $\kappa^* = 0.9$ に設定した AsLT モデルの下でのコスメネットワークの結果である．AsIC モデルとは異なり，情報は狭い範囲で伝わり，拡散系列の長さも短い．相応して，アクティブなノード数も 1000 以下と少なく，その結果，推定したパラメータの誤差は AsIC モデルに比べて大きくなった．それでも，提案法で推定した期待影響度曲線の RMSE は拡散系列の影響度曲線の RMSE と比べて約 1/6 であり，提案法を使った場合の方が期待影響度をより正確に推定できている．同様の結果は，アメーバブログネットワークでも観測された．

### 5.4　パラメータ学習精度の評価

実験の最後に，本提案法の中核的な技術となるパラメータ学習の精度について見ていく．具体的には，提案法である期待影響度推定法の **step.1** で各拡散系列 $d_n$ に対して学習したパラメータ $(r_n, \kappa_n)$ と，実験で最初に設定した真のパラメータ $(r^*, \kappa^*)$ の誤差を計算する．上で行った実験では，ネットワークと真のパラメータ設定の各組に対して $N = 1000$ 本の拡散系列を生成しているので，誤差の計算も $N$ 回分行うことができる．よって，ここでは以下の式を使って平均誤差を計算し評価に使う．

$$\varepsilon_\kappa = \frac{1}{N}\sum_{i=1}^{N}|\kappa^* - \kappa_n|, \quad \varepsilon_r = \frac{1}{N}\sum_{i=1}^{N}|r^* - r_n|$$

表 1 は，上で示した AsIC モデルを仮定した場合の実験の中で生成した拡散系列に対して，パラメータ学習をさせた際の推定値 $(r, \kappa)$ の平均誤差を表している．表 2 も同様に，AsLT モデルを仮定した場合の推定値の推定誤差を表す．表 1，表 2 を見て分かるように，モデルやネットワークデータに関わらず推定誤差はとても小さい．よって，このパラメータ学習法によって学習したパラメータ推定値は十分に信用できるものであると言える．

## 6.　議　　論

本論文では，観測する拡散系列はモデルの確率的な動作によって多様になることを前提としている．これ

は情報拡散モデルの性質というだけではなく，現実にも観測される自然な現象である．ブログを例に説明しよう．ブログの投稿者 (ブロガー) をノード，読者関係をリンクとしたソーシャルネットワークでは，ブロガーが書いた記事がリンクを介してトラックバックという形で伝播していく．今，ある一人のブロガーを情報源ノード $v_0$ とするとき，ノード $v_0$ からは複数の記事 (情報) が投稿される．ノード $v_0$ から流れる情報はいつも同じ人数に情報が伝わるわけではなく，ある日の記事は多くの人に伝わり，またある日の記事はあまり伝わらないということがある．実験で生成した多様な拡散系列はまさにこの現象を表しており，今回採用した情報拡散の確率モデルは現実的であると言える．また，本提案法によって期待影響度を推定することで，偶発的に起きた大きな拡散や小さな拡散の影響をあまり受けずに，ノードの持つ影響度に関する知見を得ることができる点で有用である．

本論文内で行った実験は，全てのリンクやノードに対して $r$ と $\kappa$ を一つの値に設定するという，簡単なケースを想定して行った．[7], [15] の研究では，実データを用いた実験により同じようなトピックの情報拡散においては，ネットワークを構成する人々は同じような振る舞いをすることを示している．このような知見より，類似したトピックとの制約はつくものの，本論文の提案法でやってきたようなパラメータの想定でも現実データの分析に十分貢献できると考えている．

また本論文では，AsIC モデルと AsLT モデルを仮定した場合の期待影響度曲線の推定法について研究した．しかし，本提案法の枠組みは他の情報拡散モデルにも適用可能な汎用性を持っていると言える．例えば，複数の意見がソーシャルネットワーク上をどのように広がるか，ということに興味がある場合には，この枠組みの下で Voter Model を使い，意見シェア曲線の期待値を推定すればいい [16]．どのモデルを使うかは解決したい問題によって決め，その性能評価はそのタスクに依存する尺度に基づいて決めるべきである．我々は，本論文での基本的なモデルにおけるこれらの結果は，より現実的な拡散モデルに対しても適用可能であると確信している．

## 7. あとがき

ソーシャルネットワーク分析における 1 つの挑戦として，ある時刻において期待される影響度 (期待影響度曲線) を推定することが挙げられる．しかし，情報拡散の確率的な動作のために，単一の観測した拡散系列は期待影響度曲線の近似値として使うことはできない．本論文では，我々は非同期時間遅れ付き独立カスケード (AsIC) モデルと非同期時間遅れ付き線形閾値 (AsLT) モデルの 2 つの場合を仮定し，単一の観測した拡散系列から高精度で期待影響度を推定する新しい手法を提案した．提案法では，まず，観測した一つの拡散系列からモデルパラメータを学習し，その後，学習したモデルパラメータを使い情報拡散シミュレーションによって期待影響度曲線を推定した．評価実験では，現実のソーシャルネットワークの構造から人工的に生成した拡散系列を用いて，提案法により推定した期待影響度曲線は，観測した拡散系列を期待影響度の近似値として使った場合よりも精度が高いことを示した．今後は，今回実験に使ったネットワークデータ以外にも，人工ネットワーク [17], [18] による評価も積極的に進めたいと考えている．

**謝辞**　本研究の一部は科研費 (20500147) の助成を受けて行なったものである．

## 文　献

[1] M.E.J. Newman, S. Forrest, and J. Balthrop, “Email networks and the spread of computer viruses,” Physical Review E, vol.66, p.035101, 2002.

[2] M.E.J. Newman, “The structure and function of complex networks,” SIAM Review, vol.45, pp.167–256, 2003.

[3] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, “Information diffusion through blogspace,” SIGKDD Explorations, vol.6, pp.43–52, 2004.

[4] P. Domingos, “Mining social networks for viral marketing,” IEEE Intelligent Systems, vol.20, pp.80–82, 2005.

[5] J. Leskovec, L.A. Adamic, and B.A. Huberman, “The dynamics of viral marketing,” Proceedings of the 7th ACM Conference on Electronic Commerce (EC’06), pp.228–237, 2006.

[6] 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.

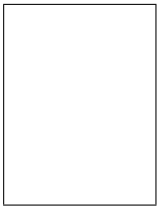
[7] D. Kempe, J. Kleinberg, and E. Tardos, “Maximizing the spread of influence through a social network,” Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003), pp.137–146, 2003.

[8] M. Kimura, K. Saito, and H. Motoda, “Blocking links to minimize contamination spread in a social network,” ACM Transactions on Knowledge Discovery from Data, vol.3, pp.9:1–9:23, 2009.

[9] D.J. Watts, “A simple model of global cascades on

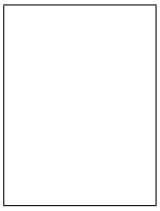
random networks," Proceedings of National Academy of Science, USA, vol.99, pp.5766–5771, 2002.
[10] D.J. Watts and P.S. Dodds, "Influence, networks, and public opinion formation," Journal of Consumer Research, vol.34, pp.441–458, 2007.
[11] M. Kimura, K. Saito, and R. Nakano, "Extracting influential nodes for information diffusion on a social network," Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI-07), pp.1371–1376, 2007.
[12] M. Kimura, K. Saito, R. Nakano, and H. Motoda, "Finding influential nodes in a social network from information diffusion data," Proceedings of the International Workshop on Social Computing and Behavioral Modeling (SBP09), pp.138–145, 2009.
[13] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Learning continuous-time information diffusion model for social behavioral data analysis," Proceedings of the 1st Asian Conference on Machine Learning (ACML2009), pp.322–337, 2009.
[14] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Behavioral analyses of information diffusion models by observed data of social network," Proceedings of the the 2010 International Conference on Social Computing, Behavioral Modeling, and Prediction (SBP 2010), pp.149–158, 2010.
[15] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Selecting Information Diffusion Models over Social Networks for Behavioral Analysis," Machine Learning and Knowledge Discovery in Databases, pp.180–195, 2010.
[16] M. Kimura, K. Saito, H. Motoda, and K. Ohara, "Learning to predict opinion share in social networks," Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI-10), pp.1364–1370, 2010.
[17] 内田誠，白山晋，"SNS のネットワーク構造の分析とモデル推定，" 情報処理学会論文誌，vol.47，no.9，pp.2840–2849，2006．
[18] 鳥海不二夫，石田健，石井健一郎，"SNS におけるネットワーク成長モデルの提案，" 電子情報通信学会論文誌. D，情報・システム，vol.93，no.7，pp.1135–1143，2010．

（平成 xx 年 xx 月 xx 日受付）



**吉川　友也**　（学生員）

2011 年静岡県立大学経営情報学部卒業．同年より，奈良先端科学技術大学院大学情報科学研究科在学中．社会ネットワーク解析，機械学習に関する研究に興味を持つ．電子情報通信学会学生員．

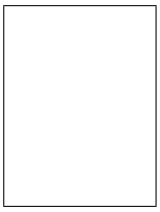


**斉藤　和巳**　（正員）

1985 年慶応義塾大学理工学部数理科学科卒業．同年日本電信電話株式会社入社．1991 年より 1 年間オタワ大学客員研究員．2007 年より，静岡県立大学経営情報学部教授．機械学習，複雑ネットワーク等の研究に従事．博士 (工学)．情報処理学会，電子情報通信学会，日本神経回路学会，日本応用数理学会各会員．

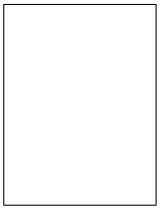
**元田　浩**

1965 年東京大学工学部原子力工学科卒業．1967 年同大学院原子力工学専攻修士課程修了．同年，日立製作所に入社．同社中央研究所，原子力研究所，エネルギー研究所，基礎研究所を経て 1995 年退社．1996 年大阪大学産業科学研究所教授（知能システム科学研究部門，高次推論研究分野），2006 年定年退職し，現在米国空軍科学技術局アジア宇宙航空研究開発研究事務所 (AFOSR/AOARD) 科学顧問．大阪大学名誉教授．大阪大学産業科学研究所招聘教授．原子力システムの設計，運用，診断，制御に関する研究を経て，機械学習，知識獲得，知識発見，データマイニング，社会ネットワーク解析の研究に従事．工学博士．



**大原　剛三**　（正員）

1995 年大阪大学大学院基礎工学研究科前期課程修了．1996 年日本学術振興会特別研究員．1997 年より大阪大学産業科学研究所助手，同助教を経て，2009 年より青山学院大学理工学部情報テクノロジー学科准教授．博士（工学）．データマイニング，機械学習，社会ネットワーク解析に関する研究に従事．IEEE ,AAAI，情報処理学会，人工知能学会各会員．



**木村　昌弘**　（正員）

1987 年大阪大学理学部数学科卒業．1989 年同大学理学研究科数学専攻修士課程修了．同年，日本電信電話株式会社入社．NTT コミュニケーション科学基礎研究所を経て，現在，龍谷大学理工学部電子情報学科教授．複雑ネットワーク科学，データマイニングおよび機械学習の研究と教育に従事．博士 (理学)．日本数学会，日本応用数理学会，日本神経回路学会，電子情報通信学会各会員．

**Abstract** We address the problem of estimating the expected influence curves with good accuracy from a single observed diffusion sequence, for both the asynchronous independent cascade model and the asynchronous linear threshold model. We solve this problem by first learning the model parameters and then estimating the influence curve using the learned model. Since the length of the observed diffusion sequence may vary from a long one to a short one, we evaluate the proposed method by simulation using artificial diffusion sequence and show that the proposed method can estimate the expected influence curve robustly from a single diffusion sequence.

**Key words** social network analysis, information diffusion model, social network, expected influence curve