Vol.2013-MPS-96 No.10
Vol.2013-BIO-36 No.10
2013/12/11

最短パス長と時間遅れを用いたネットワーク構造推定法

本條　貴裕[1]　小野　景子[2]　熊野　雅仁[2]　木村　昌弘[2]

概要：ソーシャルネットワークは，レコメンデーションシステムやバイラルマーケティングなどにおいて，重要な役割を果たし得る．ところで，ある情報をどのユーザがいつ発信したかというデータは観測可能であることが多いので，情報拡散系列を収集することは比較的容易であるが，そのような情報を拡散させたユーザ集合におけるソーシャルネットワーク構造を知ることは，プライバシーの問題などから一般に困難である．本研究では，与えられたユーザ集合に対して，観測された情報拡散系列から，その背後にあるソーシャルネットワーク構造を最短パス長と時間遅れを用いて推定する手法を提案する．そして、大規模な実データを用いた実験により，代表的な既存法よりも提案法が有効であることを示す．

1. はじめに

近年，Web 空間に登場したブログや Facebook や Twitter などのソーシャルメディアサービスがポピュラーとなり，これらのサービスの利用は我々の日常活動の一部にまでになってきている．そして，これらのサービスを通じて形成される大規模ソーシャルネットワークは，ニュース，アイデア，オピニオンから悪意のある噂まで，多種多様な情報の伝搬を媒介し [6]，我々の意思決定プロセスや行動などにも大きな影響を及ぼしつつある．また，バイラルマーケティングと呼ばれる，ソーシャルネットワーク上の情報拡散を用いたマーケティング戦略も注目されるようにもなってきた．したがって近年，多くの研究が，ソーシャルネットワークにおける情報拡散 [1], [3], [8], [9], [10], [16], [17], [18], [19], [20] や意見形成 [2], [4], [12], [13], [21], [22] に対して行われている．これらの研究では，情報拡散や意見形成を媒介するソーシャルネットワーク構造は基本的には既知とされていたが，しかしながら，単なる儀礼的なお友達リンクと異なり，情報拡散や意見形成を媒介し，バイラルマーケティングなどにおいても有用となるようなソーシャルネットワーク構造は，プライバシー問題などにより一般には獲得が困難である．

ところで，「ある種の情報（例えば，ニュース記事の URL）を，どのユーザがそれを見て，いつ自分のページに記述して発信したか」というデータは観測可能であることが多いので，情報拡散系列（その情報を誰がいつポストしたかの時系列）を収集することは比較的容易である．従って，与えられたユーザ集合に対して，観測された複数の情報拡散系列から，その背後にあるソーシャルネットワーク構造を推定する研究が注目されている [5], [15]．本研究では，情報拡散におけるパス長を考慮した，Mannila と Terzi による情報拡散系列からのリンク推定法 [15] を拡張して，情報拡散における時間遅れの効果を組み込んだ新たな手法を提案する．Leskovec [14] らによって構築されたソーシャルネットワークと情報拡散系列の大規模な実データを用いた実験により，代表的な既存手法である Mannila と Terzi の手法 [15] および Gomez-Rodriguez らの手法 [5] と比較して，提案手法が有効であることを実証する．

2. ネットワーク構造推定問題

与えられたノード集合を $\mathcal{I} = \{i \,|\, i = 1, \ldots, n\}$，その中で拡散した情報の集合を $\mathcal{U} = \{u \,|\, u = 1, \ldots, m\}$，観測時刻の集合を $\mathcal{T} = \{t \,|\, t = 1, \ldots, T\}$ とする．

本論文では，時刻 $t = 1$ から時刻 $t = T$ 間にノード集合 $\mathcal{I}$ 内に拡散した m 種類の情報に対して，観測された情報拡散系列データ $M_1, \ldots, M_T$ から，その背後にあるネットワーク構造（これら n 個のノードの間のリンク構造）G を求める問題を論じる．ここに，$t = 1, \ldots, T$ に対して，M_t は n 行 m 列の行列であり，$i = 1, \ldots, n$, $u = 1, \ldots, m$ に対して，$M_t(i, u)$ は行列 M_t の (i, u) 成分であり，時刻 t にノード i が情報 u を持っているときは 1，持っていないときは 0 である．ただし，$M_\tau(i, u) = 1$ ならば，$t \geq \tau$ において $M_t(i, u) = 1$ である．G は n 行 n 列の行列であり，$i, j = 1, \ldots, n$ に対して，$G(i, j)$ は行列 G の (i, j) 成分であり，ノード i とノード j の間にリンクがある場合は 1，ない場合は 0 である．

[1] 龍谷大学 大学院 理工学研究科　電子情報学専攻
[2] 龍谷大学 理工学部 電子情報学科

Vol.2013-MPS-96 No.10
Vol.2013-BIO-36 No.10
2013/12/11

3. ネットワーク構造推定法

まず，Mannila と Terzi によるリンク推定法 [15] を説明し，次に，それを拡張した我々の提案法を述べる．

3.1 Mannila-Terzi 法

3.1.1 拡散モデル

Mannila-Terzi 法では，情報拡散に対し次のようなモデル化を行う．$t = 1, \ldots, T$ に対して，ネットワーク構造 G，時刻 t でのイニシエータ行列 N_t および，時刻 $t-1$ までの情報拡散データ $M_1, \ldots, M_{t-1}$ が与えられたとき，時刻 t での情報拡散データ M_t を観測する確率を，

$$Pr(M_t \,|\, M_1, \ldots, M_{t-1}, N_t, G) = \prod_{i=1}^{n} \prod_{u=1}^{m} Pr(M_t(i,u) \,|\, M_1, \ldots, M_{t-1}, N_t, G) \quad (1)$$

とする．ここに，ネットワーク構造 G，時刻 t でのイニシエータ行列 N_t および，時刻 1 から時刻 $t-1$ までの情報拡散データ $M_1, \ldots, M_{t-1}$ が与えられたとき，時刻 t でノード i が情報 u を持たない確率を，情報拡散におけるパス長を考慮して，

$$\begin{aligned} &Pr(M_t(i,u) = 0 \,|\, M_1, \ldots, M_{t-1}, N_t, G) \\ &= (1 - M_{t-1}(i,u)) \\ &\quad \times \prod_{j \neq i} \{1 - M_{t-1}(j,u) \exp(-\alpha d_G(j,i) - C) \\ &\qquad - N_t(j,u) \exp(-\alpha d_G(j,i) - C)\} \end{aligned} \quad (2)$$

とする．ただし、$d_G(j,i)$ はネットワーク G においてノード j からノード i への最短パス長であり，α は $0 < \alpha < 1$ なる定数，C は正定数である．イニシエータ行列 N_t は n 行 m 列の行列であり，その (j,u) 成分の $N_t(j,u)$ はノード j が情報 u を持っているときは 1，持っていないときは 0 である．N_t は，M_t を生成する情報源のうち M_{t-1} で観測されていないものを表現している．

3.1.2 推定アルゴリズム

観測された情報拡散系列データ $M_1, \ldots, M_T$ からその背後にあるネットワーク構造 G を，事後確率 $Pr(G, N_1, \ldots, N_T \,|\, M_1, \ldots, M_T)$ を最大にする G, $N_1, \ldots, N_T$ を求めることによって推定する。ベイズの定理より，

$$\begin{aligned} &Pr(G, N_1, \ldots, N_T \,|\, M_1, \ldots, M_T) \\ &\propto Pr(G) \prod_{t=1}^{T} Pr(M_t \,|\, M_1, \ldots, M_{t-1}, N_t, G) Pr(N_t) \\ &\stackrel{\text{def}}{=} \pi(G, N_1, \ldots, N_T \,|\, M_1, \ldots, M_T) \end{aligned} \quad (3)$$

であるので，G と N_t の事前確率を，

$$Pr(G) \propto \exp(-C_1|G|), \; Pr(N_t) \propto \exp(-C_2|N_t|) \quad (4)$$

として，$\pi(G, N_1, \ldots, N_T \,|\, M_1, \ldots, M_T)$ を最大にする G と $N_1, \ldots, N_T$ を，Metropolis-Hasting 法を用いて求める。ここに，C_1, C_2 は，$0 < C_1, C_2 < 1$ なる定数であり，$|G|$ と $|N_t|$ はそれぞれ，ネットワーク G のリンク総数と時刻 t までに情報を持ったノードの総数である．推定アルゴリズムを以下に示す．

(1) ネットワーク構造 G とイニシエータ行列の系列 $N_1, \ldots, N_T$ のペア $(G, N_1, \ldots, N_T)$ を初期化し，最大ステップ数 s を指定する．
(2) $s > 0$ ならば次の処理 (3) に進み，$s = 0$ ならば $(G, N_1, \ldots, N_T)$ を出力して処理を終了する．
(3) G の次の部分のみを変更して，新たなネットワーク構造 G' を構築する．ノードペア (i,j)(ただし，$i \neq j$) を一様ランダムに選択し，$G(i,j) = 1$ ならば $G'(i,j) = 0$ とし，$G(i,j) = 0$ ならば $G'(i,j) = 1$ とする．
(4) すべての $t = 1, \ldots, T$ に対して，N_t の次の部分のみを変更して，新たなイニシエータ行列 N'_t を構築する．ノード i と情報 u のペア (i,u) を一様ランダムに選択し，$N_t(i,u) = 1$ ならば $N'_t(i,u) = 0$ とし，$N_t(i,u) = 0$ ならば $N'_t(i,u) = 1$ とする．
(5) 確率

$$\min\left\{1, \frac{\pi(G', N'_1, \ldots, N'_T \,|\, M_1, \ldots, M_T)}{\pi(G, N_1, \ldots, N_T \,|\, M_1, \ldots, M_T)}\right\}$$

で，$(G, N_1, \ldots, N_T)$ を $(G', N'_1, \ldots, N'_T)$ に変更する．
(6) $s \leftarrow s - 1$ として処理 (2) に戻る．

3.2 提案法

本論文では, 情報拡散におけるパス長を考慮した Mannila-Terzi 法を拡張し，情報拡散における時間遅れの効果を新たに組み込む手法を提案する．

まず，Mannila-Terzi 法と同様，$t = 1, \ldots, T$ に対して，ネットワーク構造 G，時刻 t でのイニシエータ行列 N_t および，時刻 $t-1$ までの情報拡散データ $M_1, \ldots, M_{t-1}$ が与えられたとき，時刻 t での情報拡散データ M_t を観測する確率 $Pr(M_t \,|\, M_1, \ldots, M_{t-1}, N_t, G)$ を，式 (1) でモデル化する．ただし，ネットワーク構造 G，時刻 t でのイニシエータ行列 N_t および，時刻 1 から時刻 $t-1$ までの情報拡散データ $M_1, \ldots, M_{t-1}$ が与えられたとき，時刻 t でノード i が情報 u を持たない確率を，情報拡散におけるパス長だけでなく時間遅れの効果も考慮して，

$$\begin{aligned} &Pr(M_t(i,u) = 0 \,|\, M_1, \ldots, M_{t-1}, N_t, G) \\ &= (1 - M_{t-1}(i,u)) \\ &\quad \times \prod_{j \neq i} \{1 - M_{t-1}(j,u) \exp(-\alpha d_G(j,i) - \lambda(t - \tau_{j,u}) - C) \\ &\qquad - N_t(j,u) \exp(-\alpha d_G(j,i) - C')\} \end{aligned} \quad (5)$$

とする．ここに、$d_G(j,i)$ はネットワーク G においてノード

j からノード i への最短パス長であり，$\tau_{j,u}$ はノード j が情報 u をもった最初の時刻である．また，α, λ は $0 < \alpha, \lambda < 1$ なる定数であり，C, C' は正定数である．

提案法では，観測された情報拡散系列データ $M_1, \ldots, M_T$ が与えられたとき，ネットワーク構造が G であり，イニシエータ行列が $N_1, \ldots, N_T$ である事後確率 $\pi(G, N_1, \ldots, N_T \,|\, M_1, \ldots, M_T)$ を，式 (1), (4), (5) に基づいて，式 (3) からを計算する．そして，$\pi(G, N_1, \ldots, N_T \,|\, M_1, \ldots, M_T)$ を最大にする G と $N_1, \ldots, N_T$ を，Mannila-Terzi 法と同様，Metropolis-Hasting 法に基づいて，3.1.2 節に示した推定アルゴリズムにより求める．提案法ではさらに，G の探索において，同じ情報を一度も共有しなかったノード間のリンクを，あらかじめ除外することにより効率化を図っている．このようにして求められた G を，観測された情報拡散系列 $M_1, \ldots, M_T$ の背後にあるネットワーク構造の推定結果とする．

4. 実験

提案法の有効性を示すために，実ソーシャル・ネットワークデータを用いて実験を行った．まず，実験に用いたネットワークデータを説明し，次に，そのネットワークデータの統計分析の結果について述べる．さらに，代表的な既存法として Maniila-Terzi 法と Gomez-Rodriguez らの手法を取り上げ，それら 2 つの手法と提案法の性能を比較する．

4.1 実験データの作成

実験では，Leskovec [14] らによって構築されたソーシャルネットワークと情報拡散系列の大規模な実データを用いた．ブログ記事総数が 96,608,034，ミーム総数が 210,999,824，リンク総数が 418,237,269 であった．このデータに対し次の手順で性能評価用の 5 つのデータセットを構築した．

(1) 伝わったノード数が多い順にミームを次のようにソートする (図 1 参照)，

$$u_1, u_2, u_3, u_4, u_5.$$

(2) $k = 1, \ldots, 5$ に対して $u \leftarrow u_k$ として以下の処理を行う．

(3) ミーム u_k が伝わったノード群のネットワークを作成する (図 2 参照)．

(4) (3) で作成したネットワーク内で，最も多くの種類のミームをもったノード i^* を選択する (図 3 参照)．

(5) ノード i^* に伝わった各ミームを u として，処理 (3) に戻る．

ミーム u_1, u_2, u_3, u_4, u_5 に対して構築したデータセットをそれぞれデータセット A, B, C, D, E とする．それらの基本統計量を表 1 に示す．

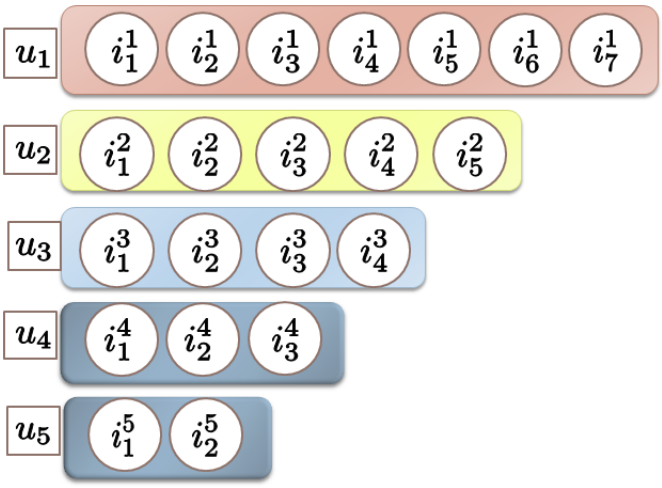


図 1 ミームのソーティングの例
(u_1, u_2, u_3, u_4, u_5 はミームを表し，$i_1^1, i_2^1, i_3^1, \ldots$ はミーム u_1 が伝わったノードを表す)．
Fig. 1 Example of sorting memes.

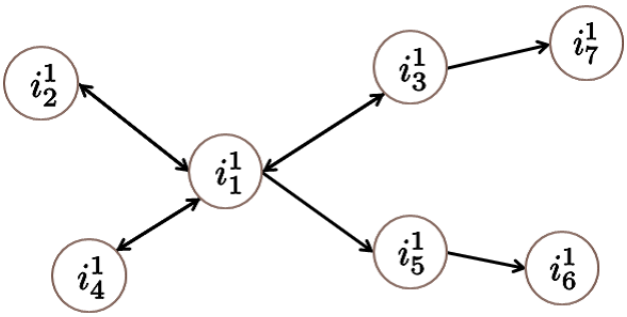


図 2 ミーム u_1 が伝わったネットワーク構造．
Fig. 2 Meme u_1 network.

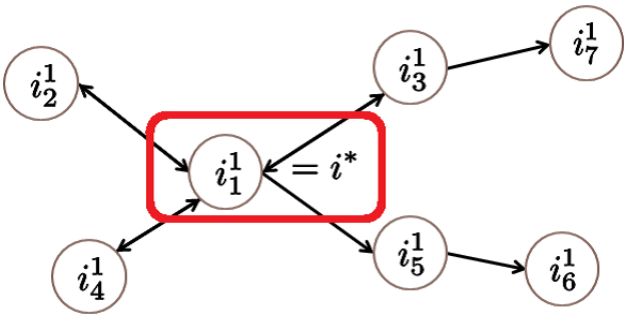


図 3 1 番目に多くのミームを伝搬したノード i^* の選択の例．
Fig. 3 Example of selecting the node i^* having the most memes in meme u_1 network.

4.2 統計分析

任意の有向ノードペア (i, j) に対して，時間遅れ $\Delta t(i, j)$ と拡散確率 $q(i, j)$ の関係を次の様に計算する．

(1) ノード i がもったミーム u の集合 U_i を抽出し，各 $u \in U_i$ に対して i が u をもった時刻 $t_{u,i}$ を求める．

(2) 各 $u \in U_i$ に対して，$U_{i,j} = \{u \in U_i \,|\, \text{ノード } i \text{ が } t_{u,i} \text{ よりも後に } u \text{ をもった}\}$ を抽出する．

(3) $\forall u \in U_{i,j}$ に対して，j が u をもった時刻 $t_{u,j}$ を求める．

表 1　データセットの基本統計量
Table 1　Basic statistics of the test dataset A, B, C, D, E.

データセット	ノード数	リンク数	拡散ミーム数
A	1287	3846	2389
B	1145	3945	2184
C	1079	3421	2069
D	856	2254	1050
E	686	1958	974

(4) 拡散確率

$$q(i,j) = \frac{|U_{i,j}|}{|U_i|}$$

を求める．

(5) $\forall u \in U_{i,j}$ に対する $t_{u,j} - t_{u,i}$ の平均値である時間遅れ $\Delta t(i,j)$ を求める．

データセット A, B, C, D, E における時間遅れ $\Delta t(i,j)$ と拡散確率 $q(i,j)$ の関係をプロットしたものを図 4 に示す．時間遅れが大きくなるにつれ，拡散確率が急激に下がっていることが分かる．図 4 における黒色の曲線は，

$$q(i,j) = e^{-\lambda \Delta t(i,j)} + \mu$$

（ただし，$\lambda > 0, \mu$ は定数である）をフィッティングした結果である．各データセットに対する λ の推定結果は表 2 の通りである．

表 2　λ の推定結果.
Table 2　Results of estimating λ.

データセット	λ
A	0.23
B	0.262
C	0.265
D	0.198
E	0.178

4.3　実験結果

実験では提案法を既存法である，Maniila-Terzi 法 [16] と Gomez 法 [5] と比較した．Maniila-Terzi 法では論文 [16] に従って $C = 0.5$，$\alpha = 0.2$ とした．それに従い，提案法でも $C = C' = 0.5$，$\alpha = 0.2$ を用いた．また，λ に関しては各データの統計分析結果に基づいて表 2 の結果を用いた．

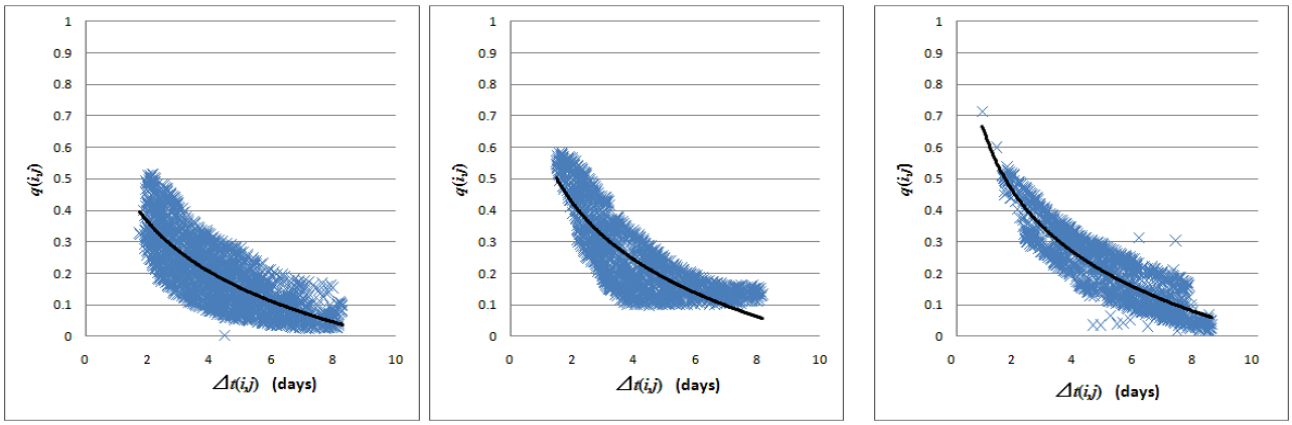
図 5 はデータセット A, B, C, D, E に対する Precision-Recall 曲線の結果を表している．図中の実線は提案法の結果を，一点鎖線は Maniila-Terzi 法の結果を，鎖線は Gomez 法の結果を示している．これらの結果より，提案法は他の２種類の手法より性能が上回っていることが確認できた．

5.　まとめ

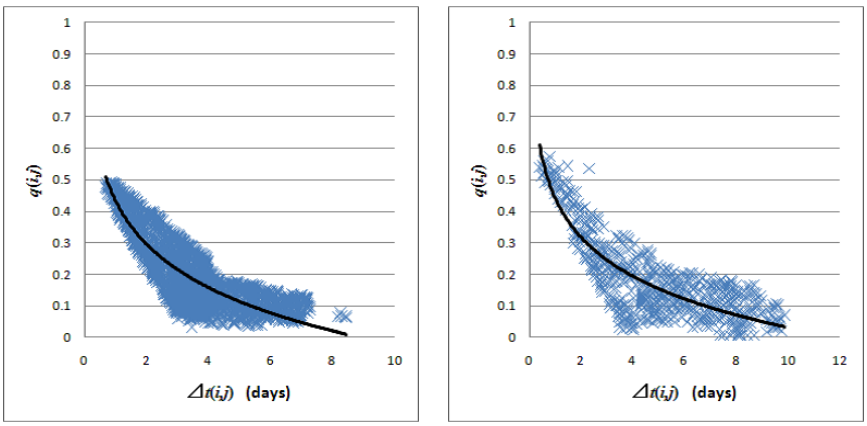
本論文では，与えられたユーザ集合において，複数の情報拡散系列の観測データから，その背後にあるソーシャルネットワーク構造を推定する手法を提案した．情報拡散モデルに最短パス長とともに時間遅れの効果も組み込むことにより，Mannila と Terzi による手法を拡張した．また，提案法には複数のパラメータが存在するが，Mannila-Terzi 法と異なり，評価実験では，それらパラメータの値をデータの統計分析結果に基づいて決定した．代表的な既存手法として，Mannila-Terzi 法と Gomez-Rodriguez らの手法を取り上げ，Leskovec らによって構築されたソーシャルネットワークと情報拡散系列の大規模な実データを用いて，それら２つのの手法との性能を比較することにより，提案法の有効性を実証した．

参考文献

[1] E. Bakshy, B. Karrer, and L.A. Adamic, Social influence and the diffusion of user-created content, *Proceedings of the 10th ACM Conference on Electronic Commerce (EC'09)*, pp.325–334, ACM, 2009.
[2] D. Bindel, J. Kleinberg, and S. Oren, How bad is forming your own opinion?, *Proceedings of the 52nd IEEE Annual Symposium on Foundations of Computer Science (FOCS'11)*, pp.57–66, IEEE, 2011.
[3] W. Chen, C. Wang, and Y. Wang, Scalable influence maximization for prevalent viral marketing in large-scale social networks, *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10)*, pp.1029–1038, ACM, 2010.
[4] E. Even-Dar and A. Shapira, A note on maximizing the spread of influence in social networks, *Proceedings of the 3rd International Workshop on Internet and Network Economics (WINE'07)*, LNCS 4858, pp.281–286, Springer, 2007.
[5] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, Inferring networks of diffusion and influence, *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10)*, pp.1019–1028, 2010.
[6] M.S. Granovetter, The strength of weak ties, *The American Journal of Sociology*, Vol.78, pp.1360–1380, 1973.
[7] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, Information diffusion through blogspace, *SIGKDD Explorations*, Vol.6, pp.43–52, 2004.
[8] 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'03)*, pp.137–146, ACM, 2003.
[9] 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.
[10] M. Kimura, K. Saito, R. Nakano, and H. Motoda, Extracting influential nodes on a social network for information diffusion, *Data Mining and Knowledge Discovery*, Vol.20, pp.70–97, 2010.
[11] M. Kimura, K. Saito, K. Ohara, and H. Motoda, Learning information diffusion model in a social network for predicting influence of nodes, *Intelligent Data Analysis*, Vol.15, pp.633-652, 2011.
[12] M. Kimura, K. Saito, K. Ohara, and H. Motoda, Opin-



(a) データセット A　(b) データセット B　(c) データセット C



(d) データセット D　(e) データセット E

図 4　時間遅れと情報拡散確率の結果.

Fig. 4　Results of the relation between $q(i,j)$ and $\Delta t(i,j)$.

ion formation by voter model with temporal decay dynamics, *Proceedings of 2012 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD'12)*, LNCS 7524, pp.565–580, Springer, 2012.

[13] M. Kimura, K. Saito, K. Ohara, and H. Motoda, Learning to predict opinion share and detect anti-majority opinionists in social networks, *Journal of Intelligent Information Systems*, Vol.41, pp.5–37, 2013.

[14] J. Leskovec, L. Backstrom, and J. M. Kleinberg, Meme-tracking and the dynamics of the news cycle, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, pp.497–506, ACM, 2009.

[15] H. Mannila and E. Terzi, Finding links and initiators: a graph-reconstruction problem, *Proceedings of the 2009 SIAM International Conference on Data Mining (SDM'09)*, pp.1207–1217, SIAM, 2009.

[16] M.E.J. Newman, S. Forrest, and J. Balthrop, Email networks and the spread of computer viruses, *Physical Review E*, Vol.66, pp.035101:1–035101:4, 2002.

[17] M. Richardson and P. Domingos, Mining knowledge-sharing sites for viral marketing, *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*, pp.61–70, ACM, 2002.

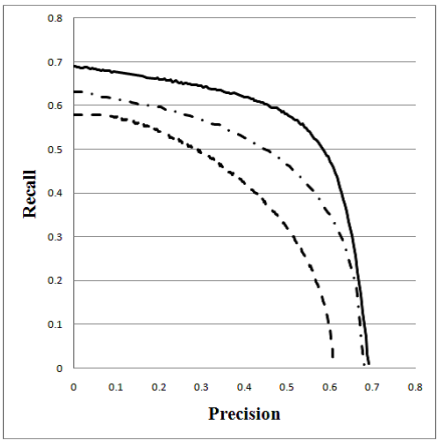
[18] K. Saito, M. Kimura, K. Ohara, and H. Motoda, Efficient discovery of influential nodes for SIS models in social networks, *Knowledge and Information Systems*, Vol.30, pp.613–635, 2012.

[19] K. Saito, M. Kimura, K. Ohara, and H. Motoda, Learning asynchronous-time information diffusion models and its application to behavioral data analysis over social networks, *Journal of Computer Engineering and Informatics*, Vol.1, pp.30-57, 2013.

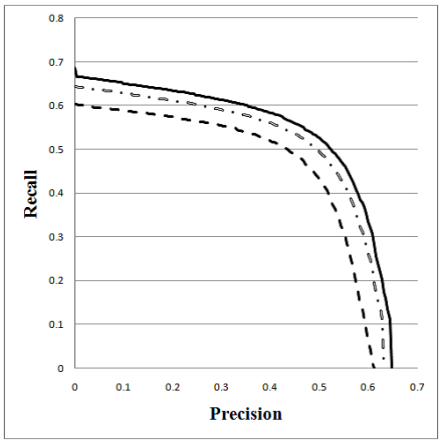
[20] K. Saito, M. Kimura, K. Ohara, and H. Motoda, Detecting changes in information diffusion pattern over social network, *ACM Transactions on Intelligent Systems and Technology*, Vol.4, pp.55:1–5:23, 2013.

[21] F. Wu and B.A. Huberman, How public opinion forms, *Proceedings of the 4th International Workshop on Internet and Network Economics (WINE'08)*, LNCS 5385, pp.334–341, Springer, 2008.

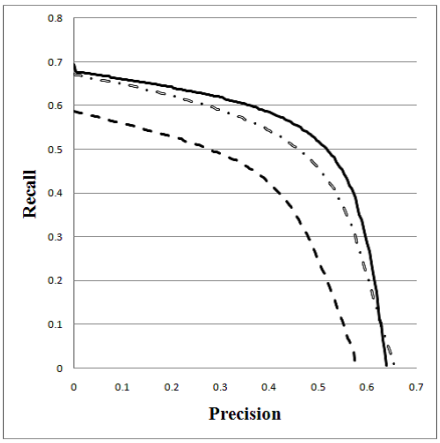
[22] H. Yang, Z. Wu, C. Zhou, T. Zhou, and B. Wang, Effects of social diversity on the emergence of global consensus in opinion dynamics, *Physical Review E*, Vol.80, pp.046108:1–046108:5, 2009.



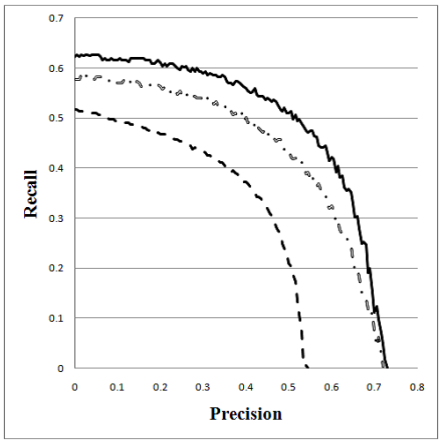
(a) データセット A



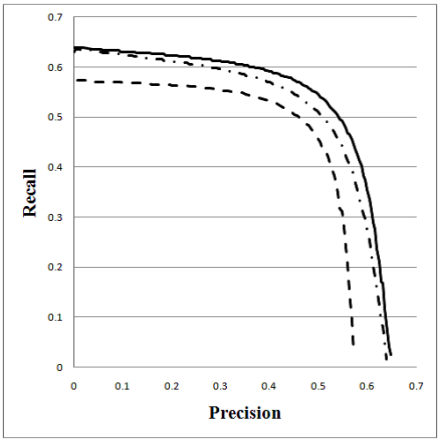
(b) データセット B



(c) データセット C



(d) データセット D



(e) データセット E

図 5 Precision-Recall 曲線.
Fig. 5 Precision-Recall curve.