論　文

社会ネットワーク上の情報伝搬における強影響力ノード抽出の効率化

木村　昌弘[†a)]　　斉藤　和巳[††]　　中野　良平[†††]

Efficient Finding of Influential Nodes from a Social Network for Information Diffusion

Masahiro KIMURA[†a)], Kazumi SAITO[††], and Ryohei NAKANO[†††]

あらまし　広く用いられている二つの確率的情報伝搬の基本モデルに関して，大規模社会ネットワーク上で最も影響力が強いノード群を見つけるという，組合せ最適化問題を考察する．本最適化問題に関しては，貪欲戦略が高性能な近似解を提供できることが知られている．しかしながら，貪欲アルゴリズムに基づく従来手法では，与えられたノード集合から影響を受けるノード数の期待値の各周辺ゲインを推定する際，モデルのランダム過程を多数回シミュレーションするため，計算負荷が問題となっていた．本論文では，ボンドパーコレーションとグラフの強連結成分分解に基づいて，それらすべての周辺ゲインを効率的に推定する手法を提案し，貪欲アルゴリズムのもとで本最適化問題を近似的に解くことに適用する．そして，大規模な実世界ネットワークを用いた実験により，提案法は従来法よりも効率的であることを実証する．

キーワード　社会ネットワーク分析，情報伝搬モデル，影響最大化問題，ボンドパーコレーション

1. まえがき

社会ネットワークとは，個人やグループ，組織などのような社会的実在の間の関係や相互作用を表すネットワークのことである．近年，インターネットや World Wide Web の普及により，大規模な実社会ネットワークを調べることが可能になってきたため，社会ネットワーク分析への関心が高まりつつある [1]～[4].

新たな価値あるサービスを提供する Web サイトの URL 情報など，様々な情報が社会ネットワーク上を所謂“口コミ”という形で，あるノードから別のノードへとリンクを通して伝搬し得る．例えば，Microsoft 社の Hotmail や Google 社の Gmail のようなフリー E メールサービス情報は，主として個人から個人への E メールを通しての紹介，すなわち，E メールネットワークという社会ネットワークを通して広がった．したがって，社会ネットワークは新商品や新思想などを普及させる上で重要な役割を果たし得るといえ，このようなネットワーク効果を利用したマーケティングとして“バイラルマーケティング”が注目されている [5]～[8]．ある情報を社会ネットワーク上で広く普及させたいとき，バイラルマーケティングでは,「影響力が強いと予想される少数のノード群をターゲットとし，最初にこれらノード群にその情報を伝えることにより，社会ネットワーク上でのノードからノードへの情報伝搬を引き起こし，そして，それら情報伝搬の結果として，より多くのノード群にその情報を伝える」という戦略が用られる．ここに，最初にその情報を伝えるノード集合を“ターゲット集合”と呼ぶ．したがって，社会ネットワーク上の情報伝搬の確率モデルが与えられ，ターゲット集合の要素数が指定されたとき，その情報が伝わるノード数の期待値を最大にするにはどのノード集合をターゲット集合とするかという組合せ最適化問題は，重要な研究課題となっている [5]～[8]．ここに，与えられたターゲット集合からその情報が伝わるノード数の期待値を，このターゲット集合の“影響度”と呼び，本最適化問題を“影響最大化問題”と呼ぶ．

Kempe らは，社会ネットワーク上の情報伝搬の基本

† 龍谷大学理工学部電子情報学科，大津市
Department of Electronics and Informatics, Ryukoku University, Otsu-shi, 520–2194 Japan
†† 静岡県立大学経営情報学部，静岡市
School of Administration and Informatics, University of Shizuoka, Shizuoka-shi, 422–8526 Japan
††† 名古屋工業大学大学院工学研究科情報工学専攻，名古屋市
Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan
a) E-mail: kimura@rins.ryukoku.ac.jp

確率モデルとして広く用いられている，“Independent Cascade (IC) モデル” [7], [9], [10] と “Linear Threshold (LT) モデル” [7], [11] に基づいて，影響最大化問題を研究した [7]．そして，貪欲アルゴリズムによる解が，影響力が強いと予想されるノード群を抽出するための手法として従来の社会ネットワーク分析で用いられている，高次数性及び中心性ヒューリスティクスによる解よりも，高性能であることを大規模な共著ネットワークを用いた実験で示した．ここに，高次数性ヒューリスティクスとは，次数が高い順にノード群を抽出する手法であり，中心性ヒューリスティクスとは，ネットワーク内の他ノードとの平均グラフ間距離が小さい（すなわち，中心性が高い）順にノード群を抽出する手法である．また，彼らは，貪欲アルゴリズムによる解の性能保証を，劣モジュラー関数に基づく解析を用いて数学的に証明している．

ところで，貪欲アルゴリズムにおいては，要素数 $k-1$ のターゲット集合 A_{k-1} が求められたとき，要素数 k のターゲット集合 A_k を計算するには，A_{k-1} の影響度のすべての “周辺ゲイン” を計算する必要がある．ここに，ノード集合 A が与えられたとき，A に属さない任意のノード v に対して，ターゲット集合 $A \cup \{v\}$ の影響度を，“A の影響度の v における周辺ゲイン” と呼ぶ．しかしながら，IC 及び LT モデルにおいては，ターゲット集合の影響度の厳密値を効率的に計算する手法はいまだ知られておらず，モデルのランダム過程を多数回シミュレーションすることでその推定値を得ていた．したがって，貪欲アルゴリズムのもとで大規模ネットワーク上での影響最大化問題を解くには，計算負荷が問題となっていた．

本論文では，IC 及び LT モデルに関する影響最大化問題の近似解を，貪欲アルゴリズムに基づいて効率的に求める手法を提案する．提案法では，IC 及び LT モデルをそれぞれ，あるボンドパーコレーションモデルと同一視することにより，与えられたノード集合 A の影響度のノード v ($\notin A$) における周辺ゲインを推定する問題を，対応するボンドパーコレーション過程により生成されるグラフにおいて，$A \cup \{v\}$ から到達可能なノードの総数の期待値を推定する問題に変換する．そして，グラフの強連結成分分解を用いて，各強連結成分ごとに，その中のすべてのノード u に対しては，$A \cup \{u\}$ から到達可能なノード集合を同時に計算することにより，A のすべての周辺ゲインを効率的に推定することを目指す．二つの大規模な実世界ネットワークを用いた実験により，提案法の有効性を検証する．

2. 情報伝搬モデルと影響最大化問題の定義

本論文では，IC 及び LT モデルに基づいて，有向グラフ $G = (V, E)$ により表現される社会ネットワーク上の影響最大化問題を考える．ノード集合 V の要素数を N とし，リンク集合 E の要素数を L とする．任意の $u, v \in V$ に対して，ノード u からノード v への有向リンク (u, v) が存在するならば，ノード v をノード u の “子ノード” と呼び，ノード u をノード v の “親ノード” と呼ぶ．任意の $v \in V$ に対して，ノード v の親ノード全体の集合を $\Gamma(v)$ とする．

本章では，Kempe らの研究 [7] に従って，G 上の IC 及び LT モデルを定義し，それらに基づいて影響最大化問題を定義する．

2.1 情報伝搬モデルの定義

ある情報が社会ネットワーク上で広がっていく現象の数理モデルである IC 及び LT モデルでは，次が仮定されている．

- ノードは “アクティブ” か “非アクティブ” のどちらかの状態しかとらない．
- その情報が伝わったノードをアクティブノードとし，そうでないノードを非アクティブノードとする．
- ノードは非アクティブからアクティブには変化するが，その逆には変化しない．
- ネットワーク上でのその情報の広がり（拡散）は，アクティブノードの広がり（拡散）として表現する．
- アクティブノードの初期集合 A が与えられたとき，A に属するノードは時刻 0 で初めてアクティブになったとし，その他のノードは非アクティブとする．そして，アクティブノード拡散過程は離散時間 $t \geq 0$ で展開していく．

2.1.1 Independent Cascade モデル

まず，IC モデルを定義する．本モデルでは，各有向リンク (u, v) に対して，実数値 $p_{u,v} \in [0, 1]$ を前もって指定する．ここに，$p_{u,v}$ はリンク (u, v) を通しての “伝搬確率” と呼ばれる．本モデルにおけるアクティブノード拡散過程は，アクティブノードの初期集合が与えられたとき，次のように進んでいく．ノード u が，時刻 t で初めてアクティブになったとする．このとき，u は，非アクティブであるその各子ノード v をアクティブにする試行を時刻 t で行う．ただし，その試行は，“成功” か “失敗” のどちらかであり，確率 $p_{u,v}$ で

成功する．もし，v の複数の親ノードが時刻 t で初めてアクティブになった場合は，それら親ノードが v をアクティブにする試行は任意の順序で独立に順々に行われることになるが，これらの試行はすべて時刻 t で行われる．そして，v をアクティブにする試行のうち，少なくとも一つの試行が成功したとき，v は時刻 $t+1$ においてアクティブとなる．ところで，u が時刻 t で v をアクティブにするのに成功したか失敗したかにかかわらず，時刻 $t+1$ 以降では，u はもはや v をアクティブにする試行を行うことはできない．すなわち，初めてアクティブになったノードのみが，その非アクティブ子ノードをアクティブにする試行を行うことができる．新たにアクティブとなるノードが存在しなくなったとき，本アクティブノード拡散過程は終了する．

A をアクティブノードの初期集合とする．ω を，IC モデルのアクティブノード拡散過程において，アクティブノードが初期集合 A からリンクを通して広がっていった経路とし，$P(\omega)$ を経路 ω の確率，$g(\omega)$ を経路 ω によるアクティブノードの最終集合の要素数とする．このとき，最終アクティブノード数の期待値 $\sigma(A)$,

$$\sigma(A) \;=\; \sum_{\omega} g(\omega)\, P(\omega)$$

を，A の "影響度" と呼ぶ．ただし，ω に関する和は，IC モデルのアクティブノード拡散過程における A からの可能な経路全体でとる．

2.1.2 Linear Threshold モデル

次に，LT モデルを定義する．本モデルでは，任意のリンク $(u,v) \in E$ に対して，"重み" と呼ばれる正数 $w_{u,v}$ を，

$$\sum_{u\in\Gamma(v)} w_{u,v} \;\le\; 1$$

となるように前もって指定する．アクティブノードの初期集合が与えられたとき，本モデルにおけるアクティブノード拡散過程は，次のように進んでいく．まず，各ノード v に対して，"しきい値" と呼ばれる実数 θ_v を区間 $[0,1]$ から一様ランダムに選ぶ．ノード v を時刻 t での非アクティブノードとする．このとき，v は，時刻 t でアクティブな親ノード u から，重み $w_{u,v}$ に従って影響を受ける．$\Gamma_t(v)$ を，時刻 t でアクティブである v の親ノード全体の集合とする．もし，アクティブな親ノードからの重みの合計がしきい値 θ_v 以上であれば，すなわち，

$$\sum_{u\in\Gamma_t(v)} w_{u,v} \;\ge\; \theta_v$$

であれば，v は時刻 $t+1$ でアクティブとなる．新たにアクティブとなるノードが存在しなくなったとき，本アクティブノード拡散過程は終了する．

A をアクティブノードの初期集合とし，LT モデルでのアクティブノード拡散過程を考える．各ノード v に与えられるしきい値 θ_v をまとめて，$\theta = (\theta_v)_{v\in V}$ とおく．ここに，N 次元ベクトル θ は，$[0,1]^N$ 上の一様分布に従う確率ベクトルとみなされることに注意．$\omega(\theta, A)$ を，このアクティブノード拡散過程において，アクティブノードが初期集合 A からリンクを通して広がっていった経路とし，$g(\omega(\theta, A))$ を経路 $\omega(\theta, A)$ によるアクティブノードの最終集合の要素数とする．このとき，最終アクティブノード数の期待値 $\sigma(A)$,

$$\sigma(A) \;=\; \int_{[0,1]^N} g(\omega(\theta, A))\, d\theta$$

を，A の "影響度" と呼ぶ．ただし，$d\theta = \prod_{v\in V} d\theta_v$ である．

2.2 影響最大化問題の定義

IC 及び LT モデルに基づいて，グラフ $G=(V,E)$ 上での影響最大化問題を数学的に定義する．k を N 未満の正整数とする．このとき，IC 及び LT モデルに関する G 上での "ターゲット集合サイズが k の影響最大化問題" とは，「G 上の IC 及び LT モデルに関して，要素数が k の V の部分集合のうち，影響度の最大値を実現するもの，すなわち，

$$A_k^* \;=\; \mathrm{argmax}_{A\in\{S\subset V;\,|S|=k\}}\; \sigma(A) \qquad (1)$$

を求めよ」という問題である．ここに，$|S|$ はノード集合 S の要素数を表す．

3. 従 来 法

Kempe らは，IC 及び LT モデルに関する影響最大化問題においては，貪欲アルゴリズムに基づく解法が有効であることを示した [7]．本章では，まず，貪欲アルゴリズムを述べ，次に，貪欲アルゴリズムに基づく Kempe らによる解法を述べて，その計算量を見積もる．

3.1 貪欲アルゴリズム

IC 及び LT モデルに関する，グラフ G 上でのターゲット集合サイズが k の影響最大化問題の近似解 A_k を，次の貪欲アルゴリズムに基づいて求める．

（1） Set $A \leftarrow \emptyset$.
（2） **for** $i = 1$ to k **do**
（3） Choose a node $v_i \in V$ that maximizes $\sigma(A \cup \{v\}), (v \in V \setminus A)$.
（4） Set $A \leftarrow A \cup \{v_i\}$.
（5） **end for**

このとき，近似解 A_k の性能保証

$$\sigma(A_k) \ \geq \ \left(1 - \frac{1}{e}\right) \sigma(A_k^*)$$

が数学的に証明されている [7]．ここに，A_k^* は式 (1) で定義される真の解である．

任意の $S \subset V$ に対して，$N - |S|$ 次元ベクトル $\nabla\sigma(S)$ を，

$$\nabla\sigma(S) \ = \ (\sigma(S \cup \{v\}))_{v \in V \setminus S} \in \mathbf{R}^{N-|S|}$$

で定義し，S の“影響度周辺ゲインベクトル”と呼ぶ．ここに，$\nabla\sigma(S)$ は，S の影響度 $\sigma(S)$ のすべての周辺ゲインから構成されている．本貪欲アルゴリズムのステップ (3) では，$\nabla\sigma(A)$ を計算する手法が必要であることに注意．

3.2 従来法の影響度周辺ゲインベクトル推定

3.1 の貪欲アルゴリズムのステップ (3) において，$\nabla\sigma(A)$ の厳密値を効率的に計算する手法は明らかではない．そこで Kempe らは，$\nabla\sigma(A)$ を次のように推定していた [7]．まず，十分大きな正整数 M を指定する．そして，各 $v \in V \setminus A$ に対して独立に，初期アクティブ集合 $A \cup \{v\}$ から，IC 及び LT モデルにおけるアクティブノード拡散過程を M 回シミュレーションして，それらの経験平均を用いて $\sigma(A \cup \{v\})$ を推定することにより，$\nabla\sigma(A)$ の推定値を計算する．

すなわち，従来法では，すべての $v \in V \setminus A$ に対して独立に，$\sigma(A \cup \{v\})$ の値を，次のアルゴリズムにより推定していた．

（1） **for** $m = 1$ to M **do**
（2） Compute $a(A \cup \{v\})$.
（3） Set $x_m \leftarrow a(A \cup \{v\})$.
（4） **end for**
（5） Set $\sigma(A \cup \{v\}) \leftarrow (1/M) \sum_{m=1}^{M} x_m$.

ここに，$a(A \cup \{v\})$ は，初期アクティブ集合 $A \cup \{v\}$ から，IC 及び LT モデルにおけるアクティブノード拡散過程をシミュレーションしたときの最終アクティブノード数を表している．これらの拡散過程はランダム過程であるので，$a(A \cup \{v\})$ の値はシミュレーションごとに異なることに注意．

ところで，任意の $v \in V \setminus A$ に対して，$a(A \cup \{v\})$ の値は，IC 及び LT モデルに基づいて，次のアルゴリズムにより計算される．

（1） Set $H_0 \leftarrow A \cup \{v\}$.
（2） Set $t \leftarrow 0$.
（3） **while** $H_t \neq \emptyset$ **do**
（4） Set $H_{t+1} \leftarrow$ {activated nodes at time $t+1$}.
（5） Set $t \leftarrow t + 1$.
（6） **end while**
（7） Set $a(A \cup \{v\}) \leftarrow \sum_{j=0}^{t-1} |H_j|$

3.3 従来法の計算量

我々は，提案法と Kempe らの手法の計算量を，“探索ノード数”の期待値として見積もることにより比較することを考える．ここに，“探索ノード”とは，各手法において，そのノードの，すべての出リンクまたはすべての入リンクを，対象とするグラフ上でたどる必要があるノードを意味している．ここでは，Kempe らの手法において，**3.1** の貪欲アルゴリズムのステップ (3) で，$\sigma(A \cup \{v\}), (v \in V \setminus A)$，すなわち $\nabla\sigma(A)$ を推定する計算量を，探索ノード数の期待値により見積もる．

Kempe らの手法では，すべての $v \in V \setminus A$ に対して，グラフ G 上で IC 及び LT モデルを，初期アクティブ集合 $A \cup \{v\}$ から M 回シミュレーションする必要がある．そして，各シミュレーションでは，グラフ G 上において，そのシミュレーションでのアクティブノードからのすべての出リンクを辿る必要がある．ゆえに，Kempe らの手法において，$\nabla\sigma(A)$ を推定するときの探索ノード数の期待値は，

$$M \sum_{v \in V \setminus A} \sigma(A \cup \{v\}) \qquad (2)$$

と考えられる．

4. 提　案　法

IC 及び LT モデルに関する有向グラフ $G = (V, E)$ 上での影響最大化問題の近似解を，貪欲アルゴリズムに基づいて計算する新たな手法を提案し，その計算量を見積もる．そして，**3.3** で見積もった Kempe ら手法の計算量と比較し，提案法により計算効率の改善が期待できることを見る．

4.1 提案法の概要

提案法では，Kempe らの手法と同様，**3.1** の貪欲

アルゴリズムに基づいて，ターゲット集合サイズが k の影響最大化問題の近似解 A_k を求める．ただし，その貪欲アルゴリズムのステップ (3) では，影響度周辺ゲインベクトル $\nabla\sigma(A)$ を以下のようにして推定する．

まず，**4.3** で詳説するように，IC 及び LT モデルをそれぞれ，あるボンドパーコレーションモデルと同一視する．そして，$\sigma(A \cup \{v\}), (v \in V \setminus A)$ を，対応するボンドパーコレーション過程から生成されたグラフにおける，$A \cup \{v\}$ から到達可能なノードの総数の期待値として，推定することを考える．すなわち，十分大きな正整数 M を指定し，対応するボンドパーコレーション過程を M 回行い，M 個のグラフを生成する．そして，生成された各グラフ上で，$A \cup \{v\}$ から到達可能なノードの総数を計算し，それらの平均値として，$\sigma(A \cup \{v\})$ を推定する．ただし，**4.4** で詳説するように，すべての $v \in V \setminus A$ に対して独立に，$A \cup \{v\}$ から到達可能なノードの総数を計算するのではなく，グラフの強連結成分分解を用いた次のような手法により効率化を図る．

- まず，A から到達可能なノード集合を計算し，そのノード集合内のすべてのノード v に対して同時に，$A \cup \{v\}$ から到達可能なノードの総数を計算する．
- 次に，A から到達可能なノード集合を削除することにより得られるグラフを計算し，そのグラフを強連結成分分解する．そして，各強連結成分ごとに，その中のすべてのノード v に対して，$A \cup \{v\}$ から到達可能なノードの総数を同時に計算する．

4.2 用語と記号法の定義

提案法を詳説するために必要となる，グラフに関しての用語と記号法を定義する．

$G' = (V', E')$ を任意の有向グラフとする．V' の部分集合 V_0 に対して，$E_0 = E' \cap (V_0 \times V_0)$ とするとき，グラフ $G_0 = (V_0, E_0)$ をグラフ G' の V_0 への "誘導グラフ" と呼ぶ．$u_0, \cdots, u_\ell \in V'$ に対して，$(u_{i-1}, u_i) \in E', (i = 1, \cdots, \ell)$ であるとき，$(u_0, \cdots, u_\ell)$ を "ノード u_0 からノード u_ℓ への道" と呼ぶ．ノード u からノード v への道が存在するとき，"ノード u はノード v に到達可能" であると呼び，"ノード v はノード u から到達可能" であると呼ぶ．グラフ G' のノード v に対して，ノード v から到達可能なノード全体の集合を $F(v; G')$ と定義し，ノード v に到達可能なノード全体の集合を $B(v; G')$ と定義する．また，任意の部分集合 $S \subset V'$ に対して，

$$F(S; G') = \bigcup_{v \in S} F(v; G'), \ B(S; G') = \bigcup_{v \in S} B(v; G')$$

と定義し，$F(S; G')$ を "S から到達可能なノード集合" と呼び，$B(S; G')$ を "S に到達可能なノード集合" と呼ぶ．グラフ G' のノード v に対して，v を含む強連結成分を $SCC(v; G')$ と定義する．$SCC(v; G') = F(v; G') \cap B(v; G')$ であることに注意．

4.3 ボンドパーコレーション

G 上の "ボンドパーコレーション過程" とは，ある確率分布に従って，G の各リンクを "占領" か "不占領" かを宣言することである．ここに，ネットワーク上の情報伝搬という観点においては，占領リンクは情報伝達経路となるリンクを表しており，不占領リンクは情報伝達経路とならないリンクを表している．次のような L 次元ベクトルの集合

$$R_G = \left\{ r = (r_{u,v})_{(u,v) \in E} \in \{0, 1\}^L \right\}$$

を考える．G 上のボンドパーコレーション過程は，R_G 上の確率分布 q により決定される．すなわち，q から生成されるランダムベクトル $r \in R_G$ に対して，各リンク $(u, v) \in E$ を，$r_{u,v} = 1$ なら "占領" と宣言し，$r_{u,v} = 0$ なら "不占領" と宣言する．各 $r \in R_G$ に対して，占領リンク全体の集合を E_r とし，グラフ (V, E_r) を G_r とする．このとき，G_r 上の決定論的情報伝搬モデル $\mathcal{M}_r$ を，A がアクティブノードの初期集合ならば，$F(A; G_r)$ がアクティブノードの最終集合となるものとして定義する．G_r 上の情報伝搬モデル $\mathcal{M}_r$ を，R_G 上の確率分布 q に随伴させることにより，G 上の確率的情報伝搬モデルを定義する．この情報伝搬モデルを G 上の "ボンドパーコレーションモデル" と呼び，その確率分布 q をモデルの "占領確率分布" と呼ぶ．

G 上の IC モデルは，G 上での病気の蔓延に関しての所謂 "susceptible/infective/recovered (SIR) モデル" と同一視することができる [12]．ここに，IC モデルにおいて時刻 t で初めてアクティブになったノードというのは，SIR モデルにおける時刻 t での infective ノードに対応している．ネットワーク上の SIR モデルは，同じネットワーク上のあるボンドパーコレーションモデルと同値であることが知られている [12], [13]．ゆえに，G 上の IC モデルは，G 上のあるボンドパーコレーションモデルと同値であることがわかる．すなわち，これら二つの情報伝搬モデルは，初期ターゲット集合が与えられたとき，アクティブノードの最終集

合に対して同じ確率分布を与えることになる．ここに，G 上の IC モデルに対して，対応するボンドパーコレーションモデルの占領確率分布 $q(r)$ は，

$$q(r) = \prod_{(u,v)\in E} \left\{ (p_{u,v})^{r_{u,v}} (1-p_{u,v})^{1-r_{u,v}} \right\}$$

で与えられる．すなわち，G の各リンク (u,v) を，独立に確率 $p_{u,v}$ で“占領”と宣言することにより，$q(r)$ は生成される．ここに，$p_{u,v}$ は IC モデルにおけるリンク (u,v) を通しての伝搬確率である．

一方，LT モデルにおいて影響度関数 σ が劣モジュラーであることを導くために，Kempe らは，G 上の LT モデルが G 上のあるボンドパーコレーションモデルと同値であることを証明した [7]．ここに，G 上の LT モデルに対して，対応するボンドパーコレーションモデルの占領確率分布 q は，次のように占領リンクと不占領リンクを宣言することで生成される．まず，任意の $v \in V$ に対して，v への入リンクを次のようにしてせいぜい一つ選び取る．すなわち，確率 $w_{u,v}$ でリンク (u,v) を選択し，確率 $1-\sum_{u\in\Gamma(v)} w_{u,v}$ でどのリンクも選択しない．本試行を行った後，選び取ったリンクを“占領”と宣言し，他のリンクを“不占領”と宣言する．ただし，$w_{u,v}$ は LT モデルにおけるリンク (u,v) の重みである．ここに，$q(r)$ は具体的には，

$$q(r) = \prod_{v\in V} \prod_{u\in\Gamma(v)} \left\{ (w_{u,v})^{r_{u,v}} \cdot \left(1 - \sum_{u\in\Gamma(v)} w_{u,v}\right)^{\left(1-\sum_{u\in\Gamma(v)} r_{u,v}\right)} \right\}$$

で与えられる．ただし，$\sum_{u\in\Gamma(v)} w_{u,v} < 1$ ならば，$\sum_{u\in\Gamma(v)} r_{u,v} \leq 1$ であり，$\sum_{u\in\Gamma(v)} w_{u,v} = 1$ ならば，$\sum_{u\in\Gamma(v)} r_{u,v} = 1$ である．

4.4 提案法の影響度周辺ゲインベクトル推定

3.1 の貪欲アルゴリズムに基づいて，IC 及び LT モデルに関するターゲット集合サイズが k の影響最大化問題の近似解 A_k を求める際には，影響度周辺ゲインベクトル $\nabla\sigma(A)$ の推定が必要であった．提案法における，$\nabla\sigma(A)$ の推定法について詳説する．

4.3 で示したように，G 上の IC 及び LT モデルは，それぞれ G 上のあるボンドパーコレーションモデルと同一視することができる．したがって，どちらのモデルの場合でも，q を対応する占領確率分布とするとき，任意の $v \in V \setminus A$ に対して，

$$\sigma(A\cup\{v\}) = \sum_{r\in R_G} q(r)\, |F(A\cup\{v\}; G_r)|$$

が成り立つ．

提案法では，指定された十分大きい正整数 M に対して，対応する占領確率分布 $q(r)$ から，独立に M 個の R_G 上のベクトル $\{r_1, \cdots, r_M\}$ を生成する．すなわち，独立に M 個のグラフ $\{G_{r_m};\ m = 1, \cdots, M\}$ を生成する．そして，任意の $v \in V \setminus A$ に対して，

$$\sigma(A\cup\{v\}) \simeq \frac{1}{M}\sum_{m=1}^{M} |F(A\cup\{v\}; G_{r_m})| \qquad (3)$$

により，$\nabla\sigma(A)$ を推定する．すなわち，次のアルゴリズムにより，$\{\sigma(A\cup\{v\});\ v \in V\setminus A\}$ を推定する．

（1） **for** $m = 1$ to M **do**
（2） Generate graph G_{r_m}.
（3） Compute $\{|F(A\cup\{v\}; G_{r_m})|;\ v \in V\setminus A\}$.
（4） Set $x_{v,m} \leftarrow |F(A\cup\{v\}; G_{r_m})|$ for all $v \in V\setminus A$.
（5） **end for**
（6） Set $\sigma(A\cup\{v\}) \leftarrow (1/M)\sum_{m=1}^{M} x_{v,m}$ for all $v \in V\setminus A$.

ただし，占領確率分布 $q(r)$ から生成されたグラフ G_r に対して，$\{|F(A\cup\{v\}; G_r)|;\ v \in V\setminus A\}$ を次のアルゴリズムにより計算する．

（1） Compute $F(A; G_r)$.
（2） Set $|F(A\cup\{v\}; G_r)| \leftarrow |F(A; G_r)|$ for all $v \in F(A; G_r)$.
（3） Compute the subset $V_r^A = V \setminus F(A; G_r)$ of V, and the induced graph G_r^A of G_r to V_r^A.
（4） Set $U \leftarrow \emptyset$.
（5） **while** $V_r^A \setminus U \neq \emptyset$ **do**
（6） Pick a node $u \in V_r^A \setminus U$.
（7） Compute $F(u; G_r^A)$.
（8） Compute the subset $C(u; G_r^A) = B(u; G_r^A) \cap F(u; G_r^A)$ of $F(u; G_r^A)$.
（9） Set $|F(A\cup\{v\}; G_r)| \leftarrow |F(u; G_r^A)| + |F(A; G_r)|$ for all $v \in C(u; G_r^A)$.
（10） Set $U \leftarrow U \cup C(u; G_r^A)$.
（11） **end while**

ここで，本アルゴリズムついて説明する．まず，ステッ

プ (1) で，グラフ G_r 上で A から到達可能なノード集合 $F(A;G_r)$ を計算している．ステップ (2) では，$v \in F(A;G_r)$ ならば，$A \cup \{v\}$ から到達可能なノード集合 $F(A \cup \{v\};G_r)$ は $F(A;G_r)$ に等しいという事実を用いて，すべての $v \in F(A;G_r)$ に対して同時に $|F(A \cup \{v\};G_r)|$ を計算している．ステップ (3) では，V から $F(A;G_r)$ を取り除いたノード集合 V_r^A を計算し，グラフ G_r の V_r^A への誘導グラフ G_r^A を計算している．ステップ (4) 以降では，$v \notin F(A;G_r)$ ならば，$|F(A \cup \{v\};G_r)|$ は $|F(A;G_r)|$ と $|F(v;G_r^A)|$ の和であるという事実を用いて，対象グラフを G_r から G_r^A に縮小し，グラフ G_r^A を強連結成分分解している．ステップ (6) と (7) では，グラフ G_r^A において，既に抽出された強連結成分に属さないノード u をとり，u から到達可能なノード集合 $F(u;G_r^A)$ を計算している．ステップ (8) では，G_r^A の $F(u;G_r^A)$ への誘導グラフ内で u からリンクを逆向きにたどることより，ノード集合 $C(u;G_r^A) = B(u;G_r^A) \cap F(u;G_r^A)$ を計算している．ここに，$C(u;G_r^A) = SCC(u;G_r^A)$ であることに注意．ステップ (9) では，$v \in C(u;G_r^A)$ ならば，$|F(v;G_r^A)| = |F(u;G_r^A)|$ という事実を用いて，すべての $v \in C(u;G_r^A)$ に対して同時に，$|F(A \cup \{v\};G_r)|$ を計算している．

4.5 提案法の計算量

3.3 において述べたように，我々は，提案法と Kempe らの手法の計算量を，探索ノード数の期待値という観点から比較することを考える．ここでは，提案法において，**3.1** の貪欲アルゴリズムのステップ (3) で，$\sigma(A \cup \{v\})$, $(v \in V \setminus A)$，すなわち $\nabla\sigma(A)$ を推定する計算量を，探索ノード数の期待値により見積もる．

提案法において，占領確率分布 $q(r)$ から生成されたグラフ G_r に対し，$\{|F(A \cup \{v\};G_r)|;\ v \in V \setminus A\}$ を求めるときの探索ノード数の期待値について考える．まず，$F(A;G_r)$ とその要素数を求めるときの探索ノード数は $|F(A;G_r)|$ であるので，このときの探索ノード数の期待値は $\sigma(A)$ と考えられる．さて，誘導グラフ G_r^A の強連結成分分解を，

$$V_r^A = \bigcup_{u \in U_r^A} SCC(u;G_r^A)$$

とする．ここに，U_r^A は G_r^A の強連結成分の代表ノード全体を表す．任意の $u \in U_r^A$ に対し $F(u;G_r^A)$ を求めるとき，その探索ノード数は $|F(u;G_r^A)|$ である．更に，$F(u;G_r^A)$ が求められたもとで u の強連結成分 $SCC(u;G_r^A)$ を計算するとき，$B(u;G_r^A) \cap F(u;G_r^A)$ を計算するので，その探索ノード数は $|F(u;G_r^A)|$ 以下である．したがって，グラフ G_r^A を強連結成分分解して，すべての $u \in U_r^A$ に対し，$|F(u;G_r^A)|$ を求めるときの探索ノード数の期待値は，$|F(A \cup \{u\};G_r)| = |F(A;G_r)| + |F(u;G_r^A)|$ に注意すれば，

$$\alpha_r^A \sum_{u \in U_r^A} (\sigma(A \cup \{u\}) - \sigma(A))$$

と考えられる．ただし，α_r^A は，

$$1 \le \alpha_r^A \le 2$$

となる A と r に依存する定数である．よって，$\{|F(A \cup \{v\};G_r)|;\ v \in V \setminus A\}$ を求めるときの探索ノード数の期待値は，

$$\sigma(A) + \alpha_r^A \sum_{u \in U_r^A} (\sigma(A \cup \{u\}) - \sigma(A))$$

と考えられる．

ゆえに，提案法において，$\nabla\sigma(A)$ を推定するときの探索ノード数の期待値は，

$$M\left\{ \sigma(A) + \left\langle \alpha_r^A \sum_{u \in U_r^A} (\sigma(A \cup \{u\}) - \sigma(A)) \right\rangle_r \right\} \quad (4)$$

と考えられる．ただし，$\langle\ \rangle_r$ は $q(r)$ のもとで r に関して平均をとる演算を表す．

4.6 提案法と従来法の計算量の比較

提案法と Kempe らの手法の計算量を，**3.1** の貪欲アルゴリズムのステップ (3) において，$\sigma(A \cup \{v\})$, $(v \in V \setminus A)$ を推定するときの探索ノード数の期待値により比較する．

3.3 と **4.5** の結果を用いる．まず，$k \ll N$ であるので，式 (2) における $\sigma(A \cup \{v\})$ の v に関する和は，ほとんどすべての $v \in V$ に対してとられることになる．一方，式 (4) の第 2 項の u に関する和においては，一般に $|U_r^A| \ll N$ と期待できる．また，一般に，$u \in V \setminus A$ に対し $(\sigma(A \cup \{u\}) - \sigma(A))$ は，$|A|$ が増加すると減少し，そして，$|A|$ がある程度以上大きくなると，$\sigma(A \cup \{u\})$ の数パーセント以下になると期待できる．ゆえに，式 (2)，(4) より，提案法は一般に，Kempe らの手法に比べて探索ノード数が少なくなる

と期待できる．

ところで，各探索ノードにおいてたどる必要があるリンク数は，Kempe らの手法の方が提案法よりも多いことに注意．実際，Kempe らの手法ではもとのグラフ G を対象としているが，提案法では，ボンドパーコレーションにより G からリンクを削除したグラフ G_r を対象としているからである．

以上の結果より，提案法は，Kempe らの手法に比べて，計算効率の改善が期待できると考えられる．

5. 実験評価

IC モデルと LT モデルにおける影響最大化問題に対して，貪欲アルゴリズムに基づいた近似解法としての提案法の性能を，大規模な実世界ネットワークを用いて実験評価する．

5.1 ネットワークデータセット

評価実験においては，実社会ネットワークの顕著な特徴を多くもつ大規模ネットワークの利用が望ましいと考えられる．本論文では，そのような実世界ネットワークの二つのデータセットを用いた実験結果を報告する．

まず，ある種の情報は，トラックバックを通してあるブログ著者から別のブログ著者へと伝搬し得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評価実験を行った．トラックバックネットワークデータは，「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 段たどることにより，2005 年 5 月に収集した．本ネットワークは，$12,047$ ノードと $53,315$ リンクをもつ連結有向グラフであり，たいていの大規模な実ネットワークと同様，入次数分布も出次数分布もいわゆるべき乗則に従っていた．以降，本ネットワークデータをブログデータセットと呼ぶ．

次に，「ウィキペディア」内の「人名一覧」から導かれる人物ネットワークを用いて評価実験を行った．具体的には，「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 回以上共起した 2 人の人物をリンクすることから得られる無向グラフの最大連結成分を抽出し，それら無向リンクを双方向リンクとみなすことにより有向グラフを構築した．以降，本ネットワークデータをウィキペディアデータセットと呼ぶ．ここに，ノード数は $9,481$ であり，有向リンク数は $245,044$ であった．

Newman と Park は，無向グラフとして表現される社会ネットワークは，社会ネットワーク以外の実ネットワークとは異なり，一般に次の二つの統計的性質をもつということを観察している [14]．まず，そのような社会ネットワークでは，次数相関が正である．次に，クラスタ係数の値が，対応する “configuration モデル”（ランダムネットワークモデル）におけるその値に比べて非常に大きい．ただし，無向グラフのクラスタ係数 C は，

$$C = \frac{3 \times \text{number of triangles on the graph}}{\text{number of connected triples of nodes}},$$

で定義する．ここに，“triangle” とはノードの三つ組であり，その各ノードが他の各ノードとリンクで結ばれているものである．そして，“connected triple” とは，ノードの二つ組にリンクをもつノードを表している．また，次数分布が与えられたとき，対応するランダムネットワークの configuration モデルとは，その次数分布をもつすべての可能なグラフを，すべて等しい重み付けで集めたもの全体である．ところで，configuration モデルにおける C の値は，厳密に計算できることが知られている [12]．ウィキペディアデータセットの無向グラフにおいては，C の値は，対応する configuration モデルでは 0.046 であり，実測値では 0.39 であった．更に，そのグラフの次数相関は正であった．したがって，ウィキペディアデータセットは，社会ネットワーク上での影響最大化問題を解くことに対し，提案法を評価するネットワークデータとして利用可能と考える．

5.2 実験設定

貪欲アルゴリズムのもとで影響最大化問題を解くことに対して，提案法と従来法を比較した．

従来法において $\nabla\sigma(A)$ を推定する際には，計算時間の問題から，100 回シミュレーションと 1000 回シミュレーション ($M = 100, 1000$) を本実験では主に用いた．IC モデルにおいて，$M = 100$ を用いる手法を “IC100”，$M = 1000$ を用いる手法を “IC1000” と，それぞれ定義する．LT モデルに対しても同様に，“LT100”，“LT1000”，“LT10000” を定義する．

ボンドパーコレーションに基づいた提案法においては，式 (3) におけるサンプルベクトル数 M を指定する必要がある．そこで，IC モデルにおいて，$M = 100$ を使う手法を “ICBP100”，$M = 1000$ を使う手法を “ICBP1000”，$M = 10000$ を使う手法を “ICBP10000” と，それぞれ定義する．LT モデ

表 1 ブログデータセットにおける $p = 10\%$ の IC モデルのもとでの影響最大化問題に対する近似解の性能
Table 1 Performance of approximate solutions for the influence maximization problem under the IC model with $p = 10\%$ in the blog dataset.

k	IC100	IC1000	ICBP100	ICBP1000
1	173.9	173.9	173.9	173.9
10	661.0	693.4	693.1	701.8
20	743.1	858.1	869.0	874.3
30	831.7	959.1	983.8	990.7

表 2 ブログデータセットにおける LT モデルのもとでの影響最大化問題に対する近似解の性能
Table 2 Performance of approximate solutions for the influence maximization problem under the LT model in the blog dataset.

k	LT100	LT1000	LTBP100	LTBP1000
1	275.6	285.6	285.6	285.6
10	1543.8	1592.4	1590.5	1603.5
20	2126.2	2412.0	2428.0	2436.5
30	2649.9	3023.5	3049.6	3065.3

ルにおいても同様に，“LTBP100”，“LTBP1000”，“LTBP10000” を定義する．

一方，IC モデルにも LT モデルにも，前もって指定すべきパラメータがある．IC モデルでは，一様な確率 p を，任意の有向リンク (u, v) に対する伝搬確率 $p_{u,v}$ に割り当てた．すなわち，$p_{u,v} = p$ とした．LT モデルにおいては，重みを次のように一様に設定した．任意のノード v に対して親ノード $u \in \Gamma(v)$ からの重み $w_{u,v}$ を，$w_{u,v} = 1/|\Gamma(v)|$ で与えた．

5.3 実験結果

提案法と従来法を，ターゲット集合サイズ k に対して得られた近似解 A_k の性能と，その処理時間の観点から比較した．近似解 A_k の性能は，その影響度 $\sigma(A_k)$ で測定した．実験では $\sigma(A_k)$ の値を，Kempe らの研究 [7] に従い 300,000 回シミュレーションを用いて推定した．また，実験はすべて 1 台の Dell 社 PC (Intel 3.4 Ghz Xeon プロセッサー，メモリ 2 GByte, Linux 環境）で行った．

表 1，表 2 に，ブログデータセットにおける，各手法によるサイズ k での近似解 A_k の性能を示す．ここに，表 1 は $p = 10\%$ の IC モデルでの結果であり，表 2 は LT モデルでの結果である．ただし，数値は小数第 1 位までに丸められている．予想どおり，IC1000, ICBP1000, LT1000 及び LTBP1000 による解は，それぞれ，IC100, ICBP100, LT100 及び LTBP100 による解よりも，高性能であった．更に，ICBP1000 及び LTBP1000 による解は，それぞれ，IC1000 及び LT1000 による解よりも高性能であった．特に，$k = 30$ においては，ICBP1000 による解は IC1000 による解よりも約 3.3%，LTBP1000 による解は LT1000 による解よりも約 1.4%，それぞれ性能が向上していた．

表 3 ブログデータセットにおける処理時間（秒）
Table 3 Processing time (s) in the blog dataset.

k	IC1000	ICBP1000
1	3.70×10^2	7.07
10	4.69×10^4	5.68×10^1
20	1.24×10^5	1.09×10^2
30	2.13×10^5	1.60×10^2

k	LT1000	LTBP1000
1	6.57×10^2	3.19
10	4.24×10^4	2.96×10^1
20	1.25×10^5	5.64×10^1
30	2.32×10^5	8.20×10^1

表 3 に，ブログデータセット上で，IC1000, ICBP1000, LT1000 及び LTBP1000 により，サイズ k での近似解 A_k を得るのに要した時間を示す．ただし，数値は有効数字 3 けたに丸められている．予想どおり，IC1000, ICBP1000, LT1000 及び LTBP1000 は，それぞれ，IC100, ICBP100, LT100 及び LTBP100 の約 10 倍の処理時間を要していた．また，ICBP1000 及び LTBP1000 は，それぞれ，IC1000 及び LT1000 よりも効率的であった．特に，$k = 30$ に対する近似解 A_{30} を得るのに，IC1000 も LT1000 もともに約 2.5 日を要したが，ICBP1000 は約 2.5 分，LTBP1000 は約 1.5 分しか要しなかった．すなわち，ターゲット集合サイズが $k = 30$ の影響最大化問題の近似解を計算するのに，ICBP1000 は IC1000 の約 0.08%，LTBP1000 は LT1000 の約 0.04% の処理時間しか要しなかった．

ところで，ブログデータセット上で LT10000 を調べたところ，$k = 30$ において，その解の性能は 3059.0 であり，LT1000 による解よりもまだ約 1.2% も性能向上していた．一方，ICBP10000 及び LTBP10000 を調べたところ，$k = 30$ において，ICBP10000 及び LTBP10000 による解の性能は，それぞれ，991.6 及び 3066.3 であった．すなわち，ICBP1000 及び LTBP1000 による解よりも，それぞれ，約 0.09% 及び 0.03% しか性能向上していなかった．更に，近似解 A_{30} を得るのに，LT10000 は約 27 日を要したが，

表 4　ウィキペディアデータセットにおける $p = 1\%$ の IC モデルのもとでの影響最大化問題に対する近似解の性能
Table 4　Performance of approximate solutions for the influence maximization problem under the IC model with $p = 1\%$ in the Wikipedia dataset.

k	IC100	IC1000	ICBP100	ICBP1000
1	122.0	138.6	137.1	138.6
10	371.1	390.6	396.6	405.3
20	410.8	455.7	469.3	475.1
30	449.5	497.0	509.8	516.0

表 5　ウィキペディアデータセットにおける LT モデルのもとでの影響最大化問題に対する近似解の性能
Table 5　Performance of approximate solutions for the influence maximization problem under the LT model in the Wikipedia dataset.

k	LT100	LT1000	LTBP100	LTBP1000
1	340.8	340.8	293.4	340.8
10	1237.2	1715.5	1669.3	1718.0
20	1991.8	2554.8	2496.3	2581.6
30	2214.4	3117.2	3054.8	3181.0

LTBP10000 は約 14 分しか要しなかった．

表 4，表 5，表 6 に，ウィキペディアデータセットにおける実験結果を示す．ブログデータセットの場合と同様な結果が観察される．例えば，ICBP1000 及び LTBP1000 による解は，それぞれ，IC1000 及び LT1000 による解よりも高性能であり，特に，$k = 30$ においては，ICBP1000 による解は IC1000 による解よりも約 3.8%，LTBP1000 による解は LT1000 による解よりも約 2.0%，それぞれ性能が向上していた．また，ICBP1000 及び LTBP1000 は，それぞれ，IC1000 及び LT1000 よりも効率的であり，特に，$k = 30$ での近似解を計算するのに，ICBP1000 は IC1000 の約 0.06%，LTBP1000 は LT1000 の約 0.02% の処理時間しか要しなかった．

5.4 考　　察

これらの実験結果より，提案法は従来法よりも効率的であると考えられる．実際，$M = 1000$ の場合，ターゲット集合サイズが $k = 30$ の影響最大化問題において，提案法による解は従来法による解よりも 1% 以上性能が向上したにもかかわらず，提案法は従来法の 0.1% 未満の処理時間しか要しなかった．

ここで，$M = 100, 1000$ の場合，提案法による解が従来法による解よりも，なぜ性能が良かったかの理由を調べる．$\nabla\sigma(A_k)$ の値を推定し，そして，$\sigma(A_k \cup \{v\})$

表 6　ウィキペディアデータセットにおける処理時間（秒）
Table 6　Processing time (s) in the Wikipedia dataset.

k	IC1000	ICBP1000
1	6.63×10^2	1.91×10^1
10	1.94×10^5	1.74×10^2
20	4.82×10^5	3.42×10^2
30	8.03×10^5	5.10×10^2

k	LT1000	LTBP1000
1	5.41×10^2	5.17
10	9.60×10^4	4.64×10^1
20	3.03×10^5	8.57×10^1
30	5.69×10^5	1.21×10^2

$(v \in V)$ を最大にするノード v_{k+1} を選択することを考えてみよう．このとき，従来法では，シミュレーションでノード v ごとに独立に $\sigma(A_k \cup \{v\})$ の値を推定しているので，ノード v ごとに A_k の影響の数値化が異なりうる．すなわち，すべての $v \in V$ において A_k の影響を等しく評価していないということに注意する．また，IC モデルでも LT モデルでも，与えられたターゲット集合に対する最終アクティブノード数は，シミュレーションごとに非常に大きく変動していたということに注意する（付録を参照）．これらの事実より，従来法で十分な回数のシミュレーションを行わないような場合には，v_{k+1} の選択は，v ごとに A_k の影響が偶々どのように評価されたのかに左右されてしまうと考えられる．そして実験結果より，$M = 100, 1000$ ではシミュレーション回数が十分ではないと推測される．一方，提案法では，すべての $v \in V$ において A_k の影響を等しく評価している．実際，式 (3) を用いて $\sigma(A_k \cup \{v\})$ の値を推定するとき，各 $|F(A_k \cup \{v\}; G_{r_m})|$ を基本的には，

$$|F(A_k \cup \{v\}; G_{r_m})| = \left|F(v; G_{r_m}^{A_k})\right| + |F(A_k; G_{r_m})|$$

と計算しているからである．それがゆえに，我々は，$M = 100, 1000$ の場合，提案法による解の方が従来法による解よりも性能が良いと考える．

6. む す び

IC 及び LT モデルに関して，有向グラフで表現される大規模社会ネットワーク上での影響最大化問題を考察した．そして，ターゲット集合サイズが k の影響最大化問題の近似解 A_k を，貪欲アルゴリズムに基づ

いて効率的に計算する手法を提案した．提案法では，貪欲アルゴリズムに基づく従来法において計算負荷が問題となっていた，与えられたノード集合 A の影響度周辺ゲインベクトル $\nabla\sigma(A)$ の推定を，以下のように行うことでその効率化を図った．すなわち，提案法では，IC 及び LT モデルをそれぞれ，あるボンドパーコレーションモデルと同一視し，A の影響度の周辺ゲイン $\sigma(A \cup \{v\})$, $(v \in V \setminus A)$ を，対応する占領確率分布 $q(r)$ から生成されたグラフ G_r における，$A \cup \{v\}$ から到達可能なノードの総数 $|F(A \cup \{v\}; G_r)|$ の経験平均として推定した．ただし，各グラフ G_r 上において，$\{|F(A \cup \{v\}; G_r)|;\ v \in V \setminus A)\}$ を次のようにして計算した．まず，A から到達可能なノード集合 $F(A; G_r)$ を計算し，すべてのノード $v \in F(A; G_r)$ に対して同時に $|F(A \cup \{v\}; G_r)|$ を計算した．次に，ノード集合 $F(A; G_r)$ を削除することにより得られる，グラフ G_r の誘導グラフ G_r^A に対して，その強連結成分分解を行った．そして，各強連結成分ごとに，その中のすべてのノード v に対して，$|F(A \cup \{v\}; G_r)|$ を同時に計算した．

ノード数が 10,000 前後の実世界ネットワークデータである，ブログデータセットとウィキペディアデータセットを用いた実験により，提案法は従来法よりも効率的であることを実証した．特に，ターゲット集合サイズが $k = 30$ の影響最大化問題において，IC モデルの場合には，提案法による解は従来法による解よりも 3.3% 以上性能が向上したが，提案法は従来法の 0.08% 以下の処理時間しか要しなかった．また，LT モデルの場合には，提案法による解は従来法による解よりも 1.4% 以上性能が向上したが，提案法は従来法の 0.04% 以下の処理時間しか要しなかった．

謝辞 本研究は，文部科学省科学研究費補助金基盤研究 (C)(No.18500113) の補助を受けた．

文　　献

[1] M.E.J. Newman, "The structure of scientific collaboration networks," Proc. National Academy of Science, vol.98, pp.404–409, USA, 2001.

[2] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "Topic and role discovery in social networks," Proc. 19th International Joint Conference on Artificial Intelligence, pp.786–791, 2005.

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

[4] J. Leskovec, L.A. Adamic, and B.A. Huberman, "The dynamics of viral marketing," Proc. 7th ACM Conference on Electronic Commerce, pp.228–237, 2006.

[5] P. Domingos and M. Richardson, "Mining the network value of customers," Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.57–66, 2001.

[6] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.61–70, 2002.

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

[8] D. Kempe, J. Kleinberg, and E. Tardos, "Influential nodes in a diffusion model for social networks," Proc. 32nd International Colloquium on Automata, Languages and Programming, pp.1127–1138, 2005.

[9] 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.

[10] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," Proc. 7th International World Wide Web Conference, pp.107–117, 2004.

[11] D.J. Watts, "A simple model of global cascades on random networks," Proc. National Academy of Science, vol.99, pp.5766–5771, USA, 2002.

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

[13] P. Grassberger, "On the critical behavior of the genral epidemic process and dynamical percolation," Mathematical Bioscience, vol.63, pp.157–172, 1983.

[14] M.E.J. Newman and J. Park, "Why social networks are different from other types of networks," Phys. Rev. E, vol.68, 036122, 2003.

付　　録

情報伝搬モデルのシミュレーション結果の変動

IC モデル及び LT モデルにおいて，与えられたターゲット集合に対する最終アクティブノードの数は，シミュレーションごとに非常に大きく変動していた．例えば，ネットワーク内の各ノード $v \in V$ をターゲット集合とし，情報伝搬モデル（IC モデルまたは LT モデル）を 1000 回シミュレーションしたとき，最終アクティブノード数 $a(v)$ のシミュレーションごとの変動は，ブログデータセットとウィキペディアデータセットでは次のとおりであった．各ノード v に対して $a(v)$ の平均値と標準偏差をそれぞれ $m(v)$ と $s(v)$ と

し，ネットワーク全体での $m(v)$ と $s(v)$ の平均値をそれぞれ $\bar{m}$ と $\bar{s}$ とする．このとき，ブログデータセットでは，

IC モデル ($p = 10\%$): $\bar{m} = 8.6, \bar{s} = 14.3,$

LT モデル： $\bar{m} = 6.8, \bar{s} = 14.9,$

であり，ウィキペディアデータセットでは，

IC モデル ($p = 1\%$): $\bar{m} = 8.1, \bar{s} = 16.1,$

LT モデル： $\bar{m} = 12.6, \bar{s} = 42.4,$

であった．ただし，数値は小数第 1 位までに丸められている．これらの結果から，平均値（の平均値）$\bar{m}$ に対して標準偏差（の平均値）$\bar{s}$ が非常に大きいことが観察される．したがって，最終アクティブノード数は，シミュレーションごとに大きく変動していたことが見て取れる．

（平成 19 年 6 月 29 日受付，10 月 2 日再受付）

木村　昌弘　（正員）

昭 62 阪大・理・数学卒．平元同大大学院修士課程了．同年日本電信電話（株）入社．平 17 龍谷大学助教授．現在，龍谷大学理工学部電子情報学科准教授．博士（理学）．複雑系の数理，機械学習，Web マイニング，ニューラルコンピュテーションの研究に興味をもつ．平 16 年度人工知能学会研究会優秀賞受賞．人工知能学会，日本神経回路学会，日本応用数理学会，日本数学会各会員．

斉藤　和巳　（正員）

昭 60 慶大・理工・数理卒．同年 NTT 電気通信研究所入所．以来，機械学習，神経回路網，複雑ネットワークなどの研究に従事．平 19 より静岡県立大学経営情報学部教授．工博．学習アルゴリズムの研究に興味をもつ．平 3〜4 カナダ Ottawa 大学客員研究員．平 8 年度情報処理学会論文賞，平 10 年度人工知能学会論文賞など受賞．情報処理学会，人工知能学会，日本神経回路学会各会員．

中野　良平　（正員）

1971 東大・工・計数卒．同年電電公社電気通信研究所入所．以来，統計解析，データベース，人工知能，遺伝的アルゴリズム，ニューラル情報処理の研究に従事．1998〜1999 奈良先端科学技術大学院大学客員教授．1999 より名古屋工業大学知能情報システム学科教授，2003 より同大学大学院工学研究科教授．工博．人工知能，最適化，ニューラル情報処理の研究に興味をもつ．1996 情報処理学会論文賞，1997 電気通信普及財団テレコムシステム技術賞，1998 人工知能学会論文賞，1999 本会論文賞各受賞．人工知能学会，日本神経回路学会，情報処理学会各会員．