情報拡散モデルに基づく社会ネットワーク上の影響度分析

大原　剛三，斉藤　和巳，木村　昌弘，元田　浩

近年，インターネット上で大規模な社会ネットワークが構築され，さまざまな情報を急速，かつ広範囲に拡散させる媒体として注目を集めている．そのような社会ネットワークに関する研究の対象問題の一つとして，情報拡散力の高い一定数のノードの組合せを見つける影響最大化問題がある．影響最大化問題はNP-困難な最適化問題であるため，一般には，貪欲法に基づき近似解を求める．本稿では，その近似解を効率的に求めるボンドパーコレーション法の基本技術を概説する．また，情報拡散モデルのパラメータ学習，より現実的な情報拡散を再現するモデルについても紹介する．

キーワード：社会ネットワーク，影響度，情報拡散，確率モデル

1. はじめに

近年，Facebook や Twitter などのソーシャルメディアの急速な普及に伴い，大規模な社会ネットワークがインターネット上に構築されている．ここでいう社会ネットワークとは，個人，もしくは組織などの社会的主体をノードとし，それらを友人関係などの関係性に基づいてつなげたネットワークを指す．そのような社会ネットワークを介して，アイディアや意見，デマに至るまで非常に多様な情報が急速，かつ大規模に拡散し，われわれの日常生活に多大な影響を与えつつある．そのため，情報拡散という観点から社会ネットワークを分析する研究が近年数多く報告されている [1〜9]．このような既存研究では，独立カスケード (IC: Independent Cascade) モデルや線形閾値 (LT: Linear Threshold) モデル [10] などの確率に基づく基本的な情報拡散モデルが多用されている．

一方，最も多く研究されている問題の一つに影響最大化問題 [10] がある．これは，情報を効果的に拡散することができるという意味で影響度の高い一定数のノードの組合せを，社会ネットワークの中から見つけ出す問題である．この問題は，NP-困難な最適化問題となるため，一般にはその近似解を効率よく求めることが目的となり，これまで多くの取り組みが報告されている [11〜16]．しかしながら，これらの多くは，たとえば，ネットワークが DAG (Directed Acyclic Graph) でないといけないなど，対象とするモデルなどに何らかの近似，もしくは仮定が導入されている．

これに対して，IC モデルや LT モデルなどの一般的な情報拡散モデルに何の制約も課さず，貪欲法の下で影響最大化問題の近似解を効率的に求める手法として，われわれはこれまでにボンドパーコレーション法 [7, 17〜20] を提案している．本稿では，ボンドパーコレーション法の基本技術を解説するとともに，そこで用いる情報拡散モデルのパラメータ学習法 [21] についても概説する．また，より現実的な情報拡散を再現するいくつかの新しい情報拡散モデル [8, 9, 22, 23] も紹介する．

2. 情報拡散モデルと影響最大化問題

本節では，基本的な情報拡散モデルとして IC モデルと LT モデルを概説した後，影響最大化問題の形式的な定義を与える [7, 10]．以下，V を全ノード集合，E $(\subset V \times V)$ を全リンク集合とする有向ネットワーク $G = (V, E)$ を用いて社会ネットワークを表現するものとする．ここで，リンク $(u, v) \in E$ において，u をノード v の親ノード，v をノード u の子ノードと呼び，$B(v) = \{u \in V; (u, v) \in E\}$ を v の親ノードの集合，$F(u) = \{v \in V; (u, v) \in E\}$ を u の子ノードの集合とする．また，各ノードが情報の受信に成功した状態をアクティブと呼び，両モデルとも，その情報拡散過程は初期アクティブノードを起点に離散時間 $t \geq 0$ で進行し，ノードの状態は非アクティブからアクティブに変化するが，その逆には変化しないものとする．

おおはら　こうぞう
青山学院大学理工学部情報テクノロジー学科
ohara@it.aoyama.ac.jp
さいとう　かずみ
静岡県立大学経営情報学部
k-saito@u-shizuoka-ken.ac.jp
きむら　まさひろ
龍谷大学理工学部電子情報学科
kimura@rins.ryukoku.ac.jp
もとだ　ひろし
大阪大学産業科学研究所
motoda@ar.sanken.osaka-u.ac.jp

2.1 IC モデル

IC モデルでは，各リンク (u,v) はパラメータとして拡散確率 $p_{u,v}$ $(0<p_{u,v}<1)$ をもつ．そして，ノード u が時刻 t にてアクティブになった場合，u はその時点で非アクティブな子ノード v をアクティブにする機会を一度だけ与えられ，その試行は確率 $p_{u,v}$ で成功する．その試行が成功した場合，v は時刻 $t+1$ でアクティブとなる．v の複数の親ノードが時刻 t に同時にアクティブとなった場合，それらの親ノードは任意の順序で v をアクティブにすることを試みるが，いずれの試行も時刻 t で実行される．一方，親ノード u はその試行が成功するかどうかにかかわらず，それ以降，v をアクティブにすることを試みることはできない．この情報拡散過程は，いずれの非アクティブノードに対してもアクティブにする試行が実行できなくなった時点で終了する．

このモデルは，情報送信者主導のモデルであり，たとえば，Twitter におけるリツイートの連鎖による情報拡散をモデル化することができる．

2.2 LT モデル

LT モデルにおいては，各リンク (u,v) はパラメータとして重み $q_{u,v}(>0)$ をもち，その重みは $\sum_{u\in B(v)} q_{u,v} \leq 1$ という関係を満たす．LT モデルでは，まずすべてのノード $v\in V$ に対して，区間 $[0,1]$ から一様ランダムに閾値 θ_v を選択し，割り当てる．そして，時刻 t で非アクティブであるノード v は，その時点でアクティブである親ノードとの間のリンクのもつ重みの総和が閾値 θ_v 以上となった場合，すなわち $\sum_{u\in B_t(v)} q_{u,v} \geq \theta_v$ が満たされた場合に，親ノードの影響を受け，時刻 $t+1$ にアクティブとなる．ここで，$B_t(v)$ は v の親ノードのうち時刻 t の時点でアクティブであるものの集合を表す．この情報拡散過程は，いずれの非アクティブノードもそれ以上アクティブになることができなくなった時点で終了する．

このモデルは，情報受信者主導のモデルであり，たとえば，一定数の友人がある特定のトピックに関するブログ記事を投稿した時点で，それを読んだユーザ v がその影響を受けて同じトピックに関するブログ記事を投稿するような情報拡散をモデル化する．

2.3 影響最大化問題

前述のような情報拡散モデルに基づき，社会ネットワーク G 上をある情報が拡散する状況を考える．いま，時刻 $t=0$ における初期情報源（アクティブ）ノード集合 $W(\subset V)$ に対し，IC モデル，もしくは LT モデルの下での情報拡散過程が時刻 $t\geq 0$ で終了し，その時点までにアクティブとなったノード数を $\varphi_G(W)$ とする．$\varphi_G(W)$ は確率変数となるため，その期待値 $\sigma_G(W)$ を定義でき，以下，$\sigma_G(W)$ をノード集合 W のネットワーク G における影響度と呼ぶ．このとき，影響最大化問題は，与えられたネットワーク $G=(V,E)$ と定数 K に対して，影響度 $\sigma_G(W_K)$ を最大化する K 個のノード集合 $W_K(\subset V)$ を求める問題であり，次のように定式化される．

$$\operatorname*{argmax}_{W_K\subset V} \sigma_G(W_K) \tag{1}$$

3. ボンドパーコレーションに基づく影響度推定

3.1 貪欲法による影響度推定

前述の影響度 $\sigma_G(W)$ は，IC モデル，LT モデルいずれの場合も劣モジュラ関数となることが知られている [10]．すなわち，ノード集合 $W, W'(\subset V)$ が $W'\subseteq W$ という関係を満たす場合，ノード $v\in V$ に対して，$\sigma_G(W'\cup\{v\})-\sigma_G(W') \geq \sigma_G(W\cup\{v\})-\sigma_G(W)$ が成り立つ．このことから，すでに選定した $k-1$ 個のノード集合 W_{k-1} に $\sigma_G(W_{k-1}\cup\{v\})$ を最大化するノード v を追加して新たな W_k を求める再帰的な貪欲法により妥当な精度の近似解を求めることができる．式 (1) で定義される影響最大化問題の真の解を W_K^* としたとき，その貪欲法で得られる近似解 W_K の性能は，

$$\sigma_G(W_K) \geq \left(1-\frac{1}{e}\right)\sigma_G(W_K^*) \tag{2}$$

となることが数学的に証明されている [10]．ここで，$W_0=\emptyset$ とする．

上記の貪欲法において，$\sigma_G(W_{k-1}\cup\{v\})$ を最大化するようなノード v を求めるナイーブな方法は，各ノード $v\in V\setminus W_{k-1}$ に対して，IC モデル，もしくは LT モデルの下で $W_{k-1}\cup\{v\}$ を初期アクティブノード集合としたシミュレーションを M 回試行し，得られる $\varphi_G(W_{k-1}\cup\{v\})$ の平均を比較するというものである．ここで，$A\setminus B$ は集合 A から集合 B を引いた差集合を表す．しかしながら，M として十分大きな値を取らなければ一定の精度で $\sigma_G(W_{k-1}\cup\{v\})$ を近似できないため，対象とするネットワークが大規模になった場合，各 $v\in V\setminus W_{k-1}$ に対して M 回の試行が必要なこの方法では現実的な時間内で影響最大化問題を解くことは困難である．これに対して，われわれはボンドパーコレーションに基づく影響度推定法 [7, 17] とその効率化手法を提案してきた [19, 20]．以下では，それ

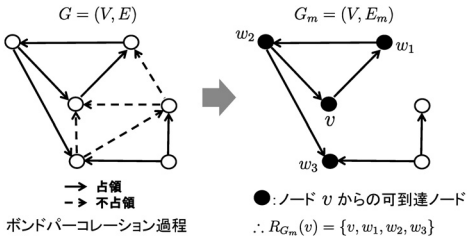

図 1　ボンドパーコレーション法における 1 回分のシミュレーション

らの技術の概要を説明する．

3.2　ボンドパーコレーションモデル

ボンドパーコレーションは確率モデルの一つであり，ネットワーク 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 の各リンクに対して“占領”(occupied) か“不占領”(unoccupied) かを宣言することである．図 1 左にボンドパーコレーション過程の例を示す．ここでは，占領と宣言されたリンクを実線，不占領と宣言されたリンクを破線で表している．このとき，ネットワーク上の情報拡散という観点から，占領リンクは情報が伝播するリンク，不占領リンクは情報が伝播しないリンクを表すと解釈する．そして，ある初期アクティブノード集合 W から占領リンクのみを辿って到達可能なノード集合 $R_G(W)$ を W から始まった情報拡散過程によりアクティブになったノード集合 $\varphi_G(W)$ と見なすモデルをボンドパーコレーションモデルと呼ぶ．IC モデル，および LT モデルは，対象とするネットワーク G 上のあるボンドパーコレーションモデルと同一視できることが知られている [10]．対応するボンドパーコレーションモデルのリンクの占領・不占領を決定する確率分布に関しては，仮定される情報拡散モデルとそのパラメータによって定まる．たとえば，IC モデルを仮定した場合，確率 $p_{u,v}$ で各リンク (u, v) を独立に“占領”と宣言する．

3.3　ボンドパーコレーション法

ここでは，ボンドパーコレーションモデルの下で影響度 $\sigma_G(W)$ を推定するボンドパーコレーション法 [7, 17] の概要について述べる．以下，ボンドパーコレーション過程を M 回試行し，そのうち m 回目の試行において占領と宣言されたリンクの集合を E_m，E_m から構成される G の部分ネットワーク (V, E_m) を G_m とする．また，G_m 上でノード集合 $W \subset V$ からリンクを辿って到達可能なノードの集合 $R_{G_m}(W)$ の要素数を $|R_{G_m}(W)|$ とする．なお，以下では W が単一のノード $v \in V$ のみから構成される場合，$R_{G_m}(\{v\})$ を単に $R_{G_m}(v)$ と記述する．たとえば，図 1 右は，図 1 左に示すボンドパーコレーション過程に対応するネットワーク G_m を表しており，その中のノード v に対して $R_{G_m}(v)$ は $\{v, w_1, w_2, w_3\}$ となる．

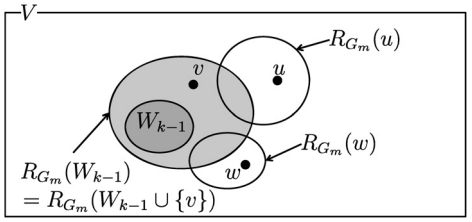

図 2　可到達ノード集合間の関係

このとき，ボンドパーコレーション法では，次式で定義される $\bar{\sigma}_G(W)$ により $\sigma_G(W)$ を近似する．

$$\bar{\sigma}_G(W) = \frac{1}{M} \sum_{m=1}^{M} |R_{G_m}(W)| \tag{3}$$

1 回のボンドパーコレーション過程により任意の $v \in V$ について $R_{G_m}(v)$ を得ることができることから，十分な近似精度を得るために M を大きくしたとしても，前述のナイーブなアプローチより効率的に $\sigma_G(v)$ の近似を得ることができる．

前述の貪欲法の枠組みにボンドパーコレーション法を適用して影響最大化問題の近似解を求める場合，G_m 上のノード集合 $W_{k-1} \cup \{v\}$ に対する可到達ノード集合 $R_{G_m}(W_{k-1} \cup \{v\})$ は，可到達ノード集合間の関係性に着目することにより効率的に数え上げることができる [7, 19]．まず，$v \in R_{G_m}(W_{k-1})$ であるならば，$R_{G_m}(W_{k-1} \cup \{v\}) = R_{G_m}(W_{k-1})$ が成り立つことに着目する．これは，図 2 に示すように $R_{G_m}(v) \subseteq R_{G_m}(W_{k-1})$ となるためである．たとえば，図 1 では，$W_1 = \{v\}$ としたとき，$R_{G_m}(w_1) = \{v, w_1, w_2, w_3\} \subseteq R_{G_m}(W_1)$ である．これにより，$v \in R_{G_m}(W_{k-1})$ であるようなノード $v \in V \setminus W_{k-1}$ に対しては，実際には $R_{G_m}(W_{k-1} \cup \{v\})$ を数え上げることなく $|R_{G_m}(W_{k-1} \cup \{v\})|$ を求めることができる．

次に，$v \notin R_{G_m}(W_{k-1})$ であるようなすべてのノード $v \in V \setminus W_{k-1}$ に関しては，図 2 におけるノード u や w のように，$R_{G_m}(W_{k-1} \cup \{v\}) = R_{G_m}(W_{k-1}) \cup R_{G_m}(v)$ となり，いずれも $R_{G_m}(W_{k-1})$ を共通に含むことに着目する．言い換えると，$|R_{G_m}(W_{k-1} \cup \{v\})|$ を最大化することは，この共通部分を除いた $R_{G_m}(v) \setminus R_{G_m}(W_{k-1})$ が最大となるような v を選択することに等しい．このことから，$R_{G_m}(W_{k-1})$ さえわかっていれば，$R_{G_m}(W_{k-1} \cup \{v\})$ ではなく $R_{G_m}(v)$ のみを計算すればよく，かつその数え上げ対象としては V から

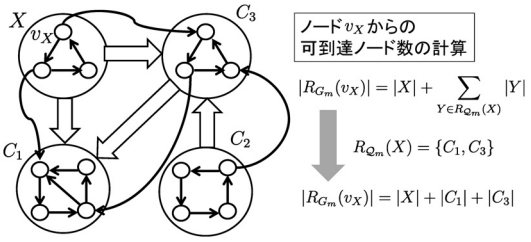

図 3　強連結成分分解に基づく商グラフ

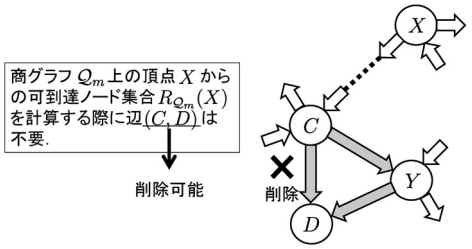

図 4　REP による冗長リンクの削除

$R_{G_m}(W_{k-1})$ を除いた $V \setminus R_{G_m}(W_{k-1})$ をノード集合とする G_m の誘導部分グラフのみを考えればよいことがわかる．

3.4　強連結成分分解に基づく可到達ノード計算の効率化

ボンドパーコレーション法では，ネットワーク G_m 上のノード $v \in V$ に対する可到達ノード数 $|R_{G_m}(v)|$ の計算が基本となる．実際には，G_m を強連結成分に分解することで $|R_{G_m}(v)|$ を効率よく計算することができる．ここで，G_m の強連結成分とは，任意のノード $v, w \in C$ に対して G_m 上で v から w への経路が存在するような V の極大部分集合 C により構成される G_m の誘導部分グラフのことである．以下，簡単のため強連結成分をそのノード集合 C で表現する．このとき，ある強連結成分 C 中の任意のノード $v, w \in C$ に対して，$R_{G_m}(v) = R_{G_m}(w)$ が成り立つ．ゆえに，各強連結成分 C に関しては，任意に選んだ一つの代表ノード $v_C \in C$ についてのみ $|R_{G_m}(v_C)|$ を計算すれば，他の $v \in C \setminus \{v_C\}$ に対する $|R_{G_m}(v)|$ を得ることができる．図 3 に強連結成分分解の例を示す．この図では，元のネットワークが四つの強連結成分 X，C_1，C_2，C_3 に分解されており，X 中の三つのノードに関しては，そこから到達可能なノードの集合はいずれも同じであることがわかる．

実際には，ノード $v \in V$ の可到達ノード集合 $R_{G_m}(v)$ は，$G_m = (V, E_m)$ の強連結成分を頂点とする商グラフ $\mathcal{Q}_m = (\mathcal{C}_m, \mathcal{E}_m)$ 上で計算できる[1]．ここで，$\mathcal{C}_m$ は G_m 中のすべての強連結成分の集合であり，$\mathcal{E}_m$ $(\subset \mathcal{C}_m \times \mathcal{C}_m)$ は $\mathcal{Q}_m$ の辺集合である．すなわち，強連結成分 C，$D \in \mathcal{C}_m$ に対して，$(v, w) \in E_m$ であるようなノードペア $v \in C$ と $w \in D$ が存在するとき，$(C, D) \in \mathcal{E}_m$ となる．図 3 は，四つの強連結成分を頂点とし，それらを結ぶ 4 本の辺（ブロック矢印）をもつ商グラフを表している．ここで，商グラフ $\mathcal{Q}_m$ の各頂点は元のネットワークにおける強連結成分であるため，$\mathcal{Q}_m$ 自体は DAG になることに注意されたい．

$\mathcal{Q}_m$ 上の頂点 $C \in \mathcal{C}_m$ に対して，C から到達可能な頂点の集合を $R_{\mathcal{Q}_m}(C)$ とする．すなわち，$D \in R_{\mathcal{Q}_m}(C)$ であるなら，商グラフ $\mathcal{Q}_m$ において C から D への経路が存在する．このとき，任意のノード $v \in C$ に対して，ネットワーク G_m における v からの可到達ノード数は以下のように求めることができる．

$$|R_{G_m}(v)| = |C| + \sum_{D \in R_{\mathcal{Q}_m}(C)} |D| \qquad (4)$$

たとえば，図 3 では，強連結成分 X 中のノード v_X の可到達ノード数 $|R_{G_m}(v_X)|$ は，商グラフ $\mathcal{Q}_m$ 上で頂点 X から到達可能な頂点が $R_{\mathcal{Q}_m}(X) = \{C_1, C_3\}$ であることから，$|R_{G_m}(v_X)| = |X| + |C_1| + |C_3|$ となる．

以上をまとめると，ボンドパーコレーション法では，1) 各強連結成分 $C \in \mathcal{C}_m$ に対して $\mathcal{C}_m$ の部分集合 $R_{\mathcal{Q}_m}(C)$ を計算し，2) 式 (4) に従い C 中の一つのノード $v_C \in C$ について $|R_{G_m}(v_C)|$ を計算し，3) ノード $v \in C \setminus \{v_C\}$ の可到達ノード数を $|R_{G_m}(v)| \leftarrow |R_{G_m}(v_C)|$ とする．以下，商グラフ $\mathcal{Q}_m$ 上での $R_{G_m}(v)$ の計算をさらに効率化する二つの技術 [20] を概説する．

まず，商グラフ $\mathcal{Q}_m$ 上での可到達ノード数の計算に不要な辺を削除する REP (Redundant-Edge Pruning) について述べる．図 4 に示すような状況を考えた場合，頂点 C から頂点 D へは頂点 Y を経由して到達可能であるため，C と D を直接結ぶ辺 (C, D) は可到達ノード数の計算においては不要であり，実際に削除しても，任意のノード $v \in G_m$ についてその可到達ノード数 $|R_{G_m}(v)|$ は影響を受けない．このような冗長な辺の削除により，同一頂点の重複探索を回避できる．

次に，商グラフ $\mathcal{Q}_m$ 上で次数（接続する辺の数）が 1 である頂点とその頂点に接続する唯一の辺を削除する MCP (Marginal-Component Pruning) について説明

[1] 以下では，元のネットワークにおけるノード，リンクと区別するために商グラフにおけるノード，リンクをそれぞれ頂点，辺と表記する．

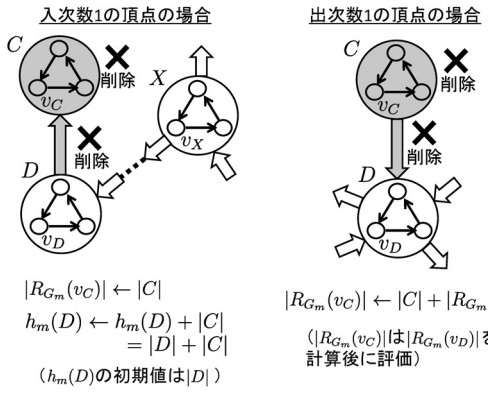

図 5 MCP による商グラフ中の頂点と辺の削除

する．具体的には，次数 1 の頂点としては，図 5 左に示す頂点 C のような入次数が 1 の場合と，同図右に示すような出次数が 1 の場合の 2 通りが考えられる．まず，図 5 左に示す入次数が 1 であるような頂点 C について考えると，ノード $v_C \in C$ の可到達ノード数 $|R_{G_m}(v_C)|$ は明らかに $|C|$ である．一方，$\mathcal{Q}_m$ 上で頂点 C に到達可能な任意の頂点 X は必ず C の唯一の親頂点である D を経由して C に到達する．このことから，仮に頂点 C，および C に接続する唯一の辺 (D, C) を削除したとしても，頂点 D に C の要素数 $|C|$ の情報をもたせておけば，任意のノード $v \in V$ の可到達ノード数 $|R_{G_m}(v)|$ は正しく計算することができる．

具体的には，$\mathcal{Q}_m$ 上の任意の頂点 $X \in \mathcal{C}_m$ について $h_m(X) \leftarrow |X|$ として $h_m(X)$ を初期化し，入次数 1，かつ出次数 0 の頂点 C の唯一の親頂点 D について，$h_m(D) \leftarrow h_m(D) + |C|$ としたうえで，C，および C に接続する唯一の辺 (D, C) を削除する．このとき，任意の頂点 $X \in \mathcal{C}_m \setminus \{C\}$ に対して，その代表ノード $v_X \in X$ の可到達ノード数 $|R_{G_m}(v_X)|$ は次式により計算できる．

$$|R_{G_m}(v_X)| = h_m(X) + \sum_{Y \in R_{\mathcal{Q}_m}(X) \setminus \{C\}} h_m(Y)$$

たとえば，図 5 左では，頂点 C の削除時には $h_m(D) = |D| + |C|$ となり，C を削除した後でも $|R_{G_m}(v_X)|$ を正しく計算できる．

次に，図 5 右に示すような出次数が 1，かつ入次数が 0 であるような頂点 C について考える．このとき，任意の頂点 $X \in \mathcal{Q}_m \setminus C$ は C には到達できないため，C を削除しても任意のノード $v \in V$ の G_m における可到達ノード数は影響を受けない．一方，C 中の代表ノード $v_C \in C$ は，C 中のノードに加え，$\mathcal{Q}_m$ における C の唯一の子頂点 D 中のノード v_D が到達可能なすべてのノードに到達可能である．したがって，v_C の可到達ノード数 $|R_{G_m}(v_C)|$ は次式で与えられる．

$$|R_{G_m}(v_C)| = |C| + |R_{G_m}(v_D)|$$

言い換えると，$|C|$ の値さえ保持しておけば，C と C に接続する唯一の辺 (C, D) を削除しても，$|R_{G_m}(v_D)|$ を計算した時点で $|R_{G_m}(v_C)|$ も正しく計算できる．

以上のように，商グラフ $\mathcal{Q}_m$ 上の次数 1 の任意の頂点は可到達ノード数の計算に影響を与えることなく事前に削除できる．ここで，図 4 において冗長な辺 (C, D) を削除した場合，頂点 D が新たに次数 1 の頂点になることに注意されたい．一般に，REP により冗長な辺を $\mathcal{Q}_m$ から削除した場合，新たな次数 1 の頂点が生じうるため，MCP より先に REP を実行する必要がある．一方，図 5 の左において頂点 C と辺 (D, C) を削除した場合，頂点 D が新たに次数 1 になる．このように，MCP の適用も新たな次数 1 の頂点を生じさせるため，MCP は再帰的に適用する必要がある．

4. 情報拡散モデルの学習

IC モデルにおける拡散確率 $p_{u,v}$ や，LT モデルにおけるリンク重み $q_{u,v}$ などのモデルパラメータは，事前にその値を指定する必要がある．しかし，これらのパラメータの真の値を知ることは実際には不可能である．そのため，過去の情報拡散系列に基づいてそれらの値を学習することが現実的なアプローチとなる [8, 21, 23]．以下では，最尤推定の枠組みで IC モデルの拡散確率 $p_{u,v}$ を学習するための目的関数について説明する [21]．なお，同様の枠組みは LT モデル，およびこれらの基本モデルを拡張したものにも適用可能である [8, 22, 23]．

いま，ネットワーク $G = (V, E)$ における IC モデルの拡散確率ベクトルを $\Theta = (p_{u,v})_{(u,v) \in E}$ とし，過去に観測した M 個の独立な情報拡散系列 $D_1, \cdots, D_M$ からその推定値 $\hat{\Theta}$ を学習することを考える．各情報拡散系列 D_m は時刻 t で初めてアクティブになったノード全体の集合を $D_m(t)$ としたとき，次のような時系列として与えられるものとする．

$$D_m = \langle D_m(0), D_m(1), \ldots, D_m(T_m) \rangle$$

ここで，T_m は m 番目の情報拡散系列の最終時刻を表し，$D_m(T_m + 1) = \emptyset$ とする．このとき，Θ に関する一つの情報拡散系列 D_m の尤度関数 $\mathcal{L}(\Theta; D_m)$ を考える．いま，あるノード $v \in D_m(t)$ に対して，リンク $(v, w) \in E$ が存在し，$w \in D_m(t+1) \cap F(v)$ であると

する．これは，ノード v がリンク (v,w) を介してノード w をアクティブにした可能性を示唆するものであるが，w の別の親ノード v' が時刻 t で同様にアクティブになっていた場合，すなわち $(D_m(t)\cap B(w))\backslash\{v\}\neq\emptyset$ である場合，$v'\in D_m(t)\cap B(w)$ が w をアクティブにした可能性もある．このことから，w が時刻 $t+1$ で初めてアクティブとなる確率 $P_{m,t+1}(w;\Theta)$ は次式で与えられる．

$$P_{m,t+1}(w;\Theta)=1-\prod_{v\in B(w)\cap D_m(t)}(1-p_{v,w}) \quad (5)$$

この式の右辺の第 2 項は，時刻 t でアクティブとなった w のすべての親ノードが w をアクティブにするのに失敗する確率を表している．

一方，時刻 t でのアクティブノード全体の集合を $S_m(t)=D_m(0)\cup\cdots\cup D_m(t)$ としたとき，ノード $v\in D_m(t)$ に対してその子ノード w が時刻 $t+1$ でアクティブでなかった場合，すなわち $w\in F(v)\backslash S_m(t+1)$ である場合，v がリンク (v,w) を介して w をアクティブにすることに失敗したことは確かであると言える．これらのことから，尤度関数 $\mathcal{L}(\Theta;D_m)$ は次のように定義できる．

$$\mathcal{L}(\Theta;D_m)=\left(\prod_{t=0}^{T_m-1}P_t^+(D_m;\Theta)\right)\left(\prod_{t=0}^{T_m}P_t^-(D_m;\Theta)\right) \quad (6)$$

ただし，$P_t^+(D_m;\Theta)$，および $P_t^-(D_m;\Theta)$ は次式で与えられるものとする．

$$P_t^+(D_m;\Theta)=\prod_{w\in D_m(t+1)}P_{m,t+1}(w;\Theta)$$

$$P_t^-(D_m;\Theta)=\prod_{v\in D_m(t)}\prod_{w\in F(v)\backslash S_m(t+1)}(1-p_{v,w})$$

直観的には，$P_t^+(D_m;\Theta)$ は時刻 t にアクティブとなったノードにより $D_m(t+1)$ 中のノードがアクティブにされる確率を表し，$P_t^-(D_m;\Theta)$ は時刻 t にアクティブとなったノードが $D_m(t+1)$ に現れない自身の子ノードをアクティブにすることに失敗する確率を表す．M 個の情報拡散系列は独立であるため，この尤度関数の値を掛け合わせることにより，全観測系列に対する尤度を求めることができる．実際には，次の対数尤度関数 $\mathcal{J}(\Theta)$ を最大化するような Θ を求める．

$$\mathcal{J}(\Theta)=\sum_{m=1}^{M}\log\mathcal{L}(\Theta;D_m) \quad (7)$$

この最大化問題は，EM アルゴリズムにおける目的関数の最大化と類似した逐次反復アルゴリズムによって解くことができる．詳細は文献 [21] を参照されたい．

5. 近年における情報拡散モデルの展開

IC モデルや LT モデルは，もっとも基本的な情報拡散モデルとして多用されるが，実際の情報拡散現象を再現するには必ずしも十分とは言えない．そのために，これまでにわれわれを含めいくつかの研究グループが新たな情報拡散モデルを提案している [8, 9, 22, 23]．具体的には，これらの基本的なモデルは，離散時間間隔でノードの状態変化が同期して起こることを前提としている．しかし，実際にはあるブログ記事を引用した記事は，元の記事の 1 時間後に投稿される場合もあれば，翌日に投稿されることもあり，ブロガーの状態変化は必ずしも同期して生じるとは限らない．このことから，われわれは IC モデル，および LT モデルを連続時間間隔における非同期状態変化を前提としたモデルに拡張し，そのパラメータ学習法も提案している [23]．

一方，情報拡散モデルのパラメータ学習には過去の情報拡散系列が必要となるが，IC モデルや LT モデルはそのパラメータ数がネットワークのリンク数に一致するため，大規模なネットワークでは学習すべきパラメータ数は膨大なものとなる．しかし，観測可能な情報拡散系列は必ずしも多くないため，限られた観測系列に過度に適合する過学習の問題が生じる．この問題を回避するため，個々のノードにその特徴を表す属性ベクトルを付与し，リンク (u,v) に対する拡散確率をノード u，v の属性ベクトルから導出するように IC モデルを拡張したノード属性つき IC モデルもわれわれは提案している [22]．このモデルでは，拡散確率導出時に用いる，属性ベクトルと同じ次元数の重みベクトルのみが学習対象となり，比較的少ない情報拡散系列からでも精度よくその値を学習することが可能である．また，たとえば，興味が似ている，もしくは出身地が同じであるようなユーザ間のリンクに対する拡散確率がそうでないユーザ間のリンクに対するものよりも高くなるなど，適切なノード属性を指定することにより，より現実的な情報拡散の再現も可能となる．

類似したアプローチとして，ノードの属性ではなく，拡散する情報のトピックに着目してリンク (u,v) におけるノード u のノード v に対する影響度，もしくは拡散確率を決定するモデルがいくつか提案されている [8, 9]．Barbieri et al. は，ユーザ u がトピック z に対してもつ影響度 (Authoritativeness) p_u^z と興味度 (Interest)θ_u^z，および情報 i に対するトピック分布

(Relevance)ρ_i^z ($z \in [1, K]$) という 3 種類のパラメータからリンク (u, v) でつながるユーザ u のユーザ v に対する影響度を導く AIR モデルと，文献 [21] と同様の EM アルゴリズムを基礎としたそのパラメータの学習法を提案している [8]．また，Chen et al. は，リンク (u, v) におけるトピック分布と拡散する情報に対するトピック分布により拡散確率 $p_{u,v}$ が定まる IC モデルの拡張，およびそのモデルの下で影響最大化問題を効率的に解く手法を提案している [9]．

6. まとめ

本稿では，確率に基づく情報拡散モデルを用いた社会ネットワーク分析の一つとして影響最大化問題を取り上げ，その近似解を効率的に求めるボンドパーコレーション法の概要を説明した．また，最尤推定の枠組みで IC モデルのパラメータを学習するための目的関数を示すとともに，いくつかのより現実的な情報拡散モデルを紹介した．社会ネットワークを介した情報拡散がわれわれの日常生活に与える影響は，よくも悪くも日々増加している．そのため，実際の情報拡散をより精緻に再現するモデル，およびそれに基づく分析手法への要求が今後も高まると思われる．とりわけ，情報の信頼性を考慮したモデルや，デマなどを迅速に察知し，その拡散を防ぐ技術への要求は高く，今後はそれらの発展にも貢献していきたいと思う．

参考文献

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

[2] M. Richardson and P. Domingos, “Mining knowledge-sharing sites for viral marketing,” In *Proceedings of KDD’02*, pp. 61–70, 2002.

[3] J. Leskovec, L. A. Adamic and B. A. Huberman, “The dynamics of viral marketing,” In *Proceedings of EC’06*, pp. 228–237, 2006.

[4] D. J. Watts and P. S. Dodds, “Influence, networks, and public opinion formation,” *Journal of Consumer Research*, **34**, pp. 441–458, 2007.

[5] E. Bakshy, J. M. Hofman, W. A. Mason and D. J. Watts, “Everyone’s an influencer: Quantifying influence on Twitter,” In *Proceedings of WSDM’11*, pp. 65–74, 2011.

[6] D. Romero, B. Meeder and J. Kleinberg, “Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter,” In *Proceedings of WWW’11*, pp. 695–704, 2011.

[7] M. Kimura, K. Saito, R. Nakano and H. Motoda, “Extracting influential nodes on a social network for information diffusion,” *Data Mining and Knowledge Discovery*, **20**, pp. 70–97, 2010.

[8] N. Barbieri, F. Bonchi and G. Manco, “Topic-aware social influence propagation models,” *Knowledge and Information Systems*, **37**, pp. 555–584, 2013.

[9] S. Chen, J. Fan, G. Li, J. Feng, L. K. Tan and J. Tang, “Online topic-aware influence maximization,” In *Proceedings of the VLDB Endowment*, **8**, pp. 666–677, 2015.

[10] D. Kempe, J. Kleinberg and E. Tardos, “Maximizing the spread of influence through a social network,” In *Proceedings of KDD’03*, pp. 137–146, 2003.

[11] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen and N. Glance, “Cost-effective outbreak detection in networks,” In *Proceedings of KDD’07*, pp. 420–429, 2007.

[12] W. Chen, Y. Wang and S. Yang, “Efficient influence maximization in social networks,” In *Proceedings of KDD’09*, pp. 199–208, 2009.

[13] W. Chen, C. Wang and Y. Wang, “Scalable influence maximization for prevalent viral marketing in large-scale social networks,” In *Proceedings of KDD’10*, pp. 1029–1038, 2010.

[14] A. Goyal, F. Bonchi and L. Lakshmanan, “A data-based approach to social influence maximization,” In *Proceedings of the VLDB Endowment*, **5**, pp. 73–84, 2011.

[15] H. Nguyen and R. Zheng, “Influence spread in large-scale social networks – A belief propagation approach,” In *Proceedings of ECML-PKDD’12*, LNAI 7524, pp. 515–530, 2012.

[16] Y. Yang, E. Chen, Q. Liu, B. Xiang, T. Xu and S. Shad, “On approximation of real-world influence spread,” In *Proceedings of ECML-PKDD’12*, LNAI 7524, pp. 548–564, 2012.

[17] M. Kimura, K. Saito and R. Nakano, “Extracting influential nodes for information diffusion on a social network,” In *Proceedings of AAAI’07*, pp. 1371–1376, 2007.

[18] M. Kimura, K. Saito and H. Motoda, “Efficient estimation of influence functions fot SIS model on social networks,” In *Proceedings of IJCAI’09*, pp. 2046–2051, 2009.

[19] K. Saito, M. Kimura and H. Motoda, “Discovering influential nodes for SIS models in social networks,” In *Proceedings of DS’09*, LNAI 5808, pp. 302–316, 2009.

[20] M. Kimura, K. Saito, K. Ohara and H. Motoda, “Efficient analysis of node influence based on SIR model over huge complex networks,” In *Proceedings of DSAA’14*, pp. 216–222, 2014.

[21] 木村昌弘，斉藤和巳，中野良平，元田浩，“社会ネットワークにおける有力ノード抽出のための情報拡散モデルの学習，”人工知能学会論文誌，**25**, pp. 215–223, 2010.

[22] K. Saito, K. Ohara, Y. Yamagishi, M. Kimura and H. Motoda, “Learning diffusion probability based on node attributes in social networks,” In *Proceedings of ISMIS’11*, pp. 153–162, 2011.

[23] 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*, **1**, pp. 30–57, 2013.