Vol.2013-ICS-170 No.6
2013/3/10

情報拡散モデルを利用した超媒介者検出法

小出 明弘[1,a)]　斉藤 和巳[1,b)]　風間 一洋[2,c)]　鳥海 不二夫[3,d)]

概要：本稿では，ネットワーク上の情報拡散現象において，情報を多くのノードへ伝搬するための橋渡しとなるノードである超媒介者を，ネットワーク構造と情報拡散モデルを利用して検出する手法を提案する．まず，任意のノードに対して情報拡散シミュレーションを行うことにより得られる影響度を，ネットワーク内の全ノードに対して算出してネットワーク全体の影響度の期待値を得る．次に，ネットワーク内の任意のノードを削除した部分ネットワークにおいて同様に影響度の期待値を算出し，この差が大きくなるノードを超媒介者として検出する．提案手法を 2 つの実ネットワークに適用し，提案手法によって検出されたノードが，実際にネットワークの情報拡散に大きく寄与していることを示す．

1. はじめに

Facebook や Twitter などのソーシャルメディアの急速な普及により，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，情報を拡散させる重要な媒体となっている．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数多く報告されている [1]， [2]， [3].

既存研究で用いられている基本的な情報拡散モデルとしては，独立カスケード（IC:Independent Cascade）モデル [5] と線形閾値（LT: Linear Threshold）モデル [6] という 2 つの対照的なモデルがある．IC モデルは情報送信者主導のモデルであり，ネットワーク上の情報送信ノードがある確率で隣接ノードに独立に情報を伝える．これに対して LT モデルは情報受信者主導のモデルであり，情報受信ノードは対象となる情報を受け入れた隣接ノードの数がある閾値以上となったときに限り，その情報を受け取る．

情報拡散の観点で最も多く研究されている問題の 1 つに影響最大化問題がある．これは，情報を出来るだけ多くのノードに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見つけ出す問題であり，これまでに先述の確率に基づく情報拡散モデルを用いてこの問題を解くアルゴリズムが多数提案されている [7]， [8].

[1] 静岡県立大学
University of Shizuoka, Suruga, Shizuoka 422–0886, Japan
[2] 和歌山大学
Wakayama Uniersity
[3] 東京大学
The Uniersity of Tokyo
[a)] j11103@u-shizuoka-ken.ac.jp
[b)] k-saito@u-shizuoka-ken.ac.jp
[c)] kazama@sys.wakayama-u.ac.jp
[d)] tori@sys.t.u-tokyo.ac.jp

一方，影響最大化問題における高影響力ノードとは異なった性質を持つ高影響力ノードとして，超媒介者と呼ばれる概念が提唱されている [4]．超媒介者とは，ネットワーク中から発信される様々な情報を受け取って，さらに多くのノードに伝える，効率的な情報伝搬の重要な役割を果たすノードである．

本稿では，ネットワーク構造と情報拡散モデルを利用して超媒介者を検出する手法を提案する．まず，ネットワーク内の任意ノードに対して情報拡散シミュレーションを行うことにより得られる影響度を，ネットワーク内の全ノードに対して算出してネットワーク全体の影響度の期待値を求める．次に，ネットワークの任意のノードを削除した部分ネットワークにおいて同様に影響度の期待値を算出し，この差を大きくするノードを超媒介者として検出する．提案手法を 2 つの実ネットワークに適用し，提案手法によって検出されたノードが実際にネットワークの情報拡散に大きく寄与していることを示す．

以下第 2 章では，情報拡散における超媒介者について述べ，第 3 章では本稿で利用する情報拡散モデルと，期待影響度について詳細に述べる．第 4 章で情報拡散に大きく寄与するノードの抽出について述べ，第 5 章で評価実験と実験結果について述べる．最後に第 6 章で本稿を結ぶ．

2. 情報拡散における超媒介者

本章では，超媒介者の定義と，検出のためのアプローチについて述べる．

2.1 超媒介者の定義

第 1 章でも述べたように，超媒介者とは，ネットワーク

中から発信される様々な情報を受け取って，さらに多くのノードに伝える，効率的な情報伝搬の重要な役割を果たすノードである．

超媒介者を検出し，このノードに情報を与えることでより多くのノードに情報を伝搬することが可能になると考えられる．また，デマのような拡散を押さえたい情報に対し，超媒介者に注意を促すことで，それ以上の拡散を抑制することが期待できる．

以下，超媒介者を検出するアプローチとして，情報拡散系列集合を利用する先行研究と，ネットワーク構造と情報拡散モデルを利用した提案手法について述べる．

2.2 情報拡散系列集合を用いた超媒介者検出

先行研究 [4] では，ネットワーク内で得られた情報拡散系列集合を利用して，超媒介者を検出する手法を提案している．情報拡散系列とは，ネットワーク内のノードを要素とした集合であり，任意のノードを情報源として発信されたある情報が，どのノードに伝わったかを表す．

情報拡散系列集合が与えられた時，この集合を，ネットワーク内の多くのノードに伝わったと考えられる高拡散系列集合と，その他の拡散系列集合に分割する．この時，超媒介者を，高拡散系列集合に頻繁に出現し，且つその他の拡散系列集合にはほとんど現れないノードであると仮定し，情報検索の分野で用いられる F 値の値が高かったノードを超媒介者として検出する．

この手法は，情報拡散系列集合が与えられれば，ネットワーク構造や拡散経路が得られなくても超媒介者を検出できることが長所である．一方，情報拡散系列集合はシステム全体の実行履歴であり，実システムの場合には運営者でないと必ずしも入手できるとは限らない．

2.3 情報拡散シミュレーションによる超媒介者検出

提案手法では，ネットワーク構造と情報拡散モデルを利用して，超媒介者を検出する．

まず, あるネットワーク構造と情報拡散モデルが与えられたとき，そのネットワーク構造に対して情報拡散シミュレーションを行い，各ノードの影響度を求めることで，ネットワーク全体での影響度の期待値を得る．次に，求めたネットワーク全体での影響度の期待値に大きく寄与したノードを超媒介者として検出する．

提案手法は，ネットワークと情報拡散モデルが与えられれば，シミュレーションによって超媒介者を求めることができるため，情報拡散集合を得る前段階や，情報拡散集合が手に入らない場合でも超媒介者を検出できる．また，ネットワーク構造を変更することで，模索的な分析が可能である．

3. 情報拡散モデルと期待影響度

本節では, 情報拡散の基本的なモデルである IC モデル [5] と LT モデル [6] について述べる．ここで，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”が”非アクティブ”のどちらかの状態しかとらない．
- 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変化するが，その逆は起こらない．
- ネットワーク上での情報の広がりは，アクティブノードの広がりとして表現される．
- 初期アクティブノード v が与えられた時，ノード v は時刻 0 で初めてアクティブになったとし，そのほかのノードは非アクティブであるとする．そして，情報拡散は離散時間 $t \geq 0$ で展開する．

また，今後の共通の設定として，有向ネットワーク $G = (V, E)$ を定義する．また，ノード集合 V の要素数を N とする．任意のノード $u, v \in V$ に対し，ノード u からノード v への有向リンク (u, v) が存在する時，ノード u をノード v の親ノードと呼び，任意のノード v に対する親ノード集合を $B(v)$ とする．また，$u, v \in V$ に対し，ノード v からノード u への有向リンク (v, u) が存在する時，ノード u をノード v の子ノードと呼び，任意のノード v に対する子ノード集合を $F(v)$ とする．

3.1 IC モデル

IC モデルは感染病の広がり方などを表すとされる基本的な確率モデルである．このモデルでは，各リンク (u, w) に対して前もって実数値 $p_{u,w}(0 < p_{u,w} < 1)$ を割り当てる．ここで，$p_{u,w}$ をリンク (u, w) 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 $t \geq 0$ で展開され，初期アクティブノード v 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 w 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 $p_{u,w}$ で成功する．成功したら，ノード w は時刻 $t+1$ でアクティブになる．もし，ノード w の複数の親ノードが時刻 t で同時に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．このプロセスが反復して行われ，次の時刻でアクティブになるノードが無くなったとき，情報拡散は終了する．

3.2 LT モデル

LT モデルは，すべてのノード $v \in V$ に対して，$\sum_{u \in B(v)} \omega_{u,v} \leq 1$ となるように前もって重み $\omega_{u,v}(> 0)$

を割り当てる．LT モデルでの拡散過程は，初期アクティブノード v 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 θ_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 u は時刻 t でアクティブな各親ノードから $\omega_{w,u}$ に従って影響を受ける．$\Gamma_t(u)$ を時刻 t でアクティブである u の親ノード集合とする．もし，アクティブな親ノードから受けた重みの合計が θ_u 以上になった場合，すなわち，

$$\sum_{w \in \Gamma_t(u)} \omega_{w,u} \geq \theta_u. \tag{1}$$

であれば，ノード u は時刻 $t+1$ でアクティブになる．このプロセスが反復して行われ，次の時刻でアクティブになるノードが無くなったとき，情報拡散は終了する．

3.3 期待影響度

ネットワーク G 上のあるノード v を初期アクティブノードとした情報拡散を考える．IC もしくは LT モデルを用いてシミュレーションした際の最終的なアクティブノード数を $a(v;G)$ と定義する．なお，これらのシミュレーションはランダム過程であるため，シミュレーションごとに $a(v;G)$ の値が異なることに注意する．次に，Kempe らの手法を利用し，ノード v を初期アクティブノードにした際の最終アクティブノード数の期待値 $\sigma(v;G)$ を推定する [8]．十分に大きな整数 M を設定し，ノード v に対して M 回の拡散シミュレーションを行い，その経験平均をとることにより，$\sigma(v;G)$ を得る．これをノード v の影響度と呼ぶ．以下に $\sigma(v;G)$ を求めるアルゴリズムを示す．

for $m = 1$ to M **do**
  Compute $a(v;G)$.
  Set $x_m \leftarrow a(v;G)$.
end for
Set $\sigma(v;G) \leftarrow (1/M)\sum_{m=1}^{M} x_m$.

さらに，$E(G)$ を求め，これをネットワーク G の期待影響度と定義する．

$$E(G) = \sum_{v \in V} \sigma(v;G)p(v). \tag{2}$$

なお，本稿では問題を簡潔にするため，ノード $v \in V$ が初期アクティブノードとなる確率 $p(v)$ は一様に $1/|V|$ とする．

以下に，$E(G)$ を推定するアルゴリズムを示す．

for $v \in V$ **do**
  for $m = 1$ to M **do**
    Compute $a(v;G)$.
    Set $x_m \leftarrow a(v;G)$.
  end for
  Set $\sigma(v;G) \leftarrow (1/M)\sum_{m=1}^{M} x_m$.
end for
Set $E(G) \leftarrow (1/|V|)\sum_{v \in V} \sigma(v;G)$.

4. 期待影響度に大きく寄与するノードの抽出

本章では，期待影響度に大きく寄与するノードの抽出について述べる．

ネットワーク G にて，任意のノード $w \in V$ を削除した部分グラフ $G \setminus \{w\}$ を考える．この時，部分グラフ $G \setminus \{w\}$ の期待影響度 $E(G \setminus \{w\})$ は以下のように定義できる．

$$E(G \setminus \{w\}) = \frac{1}{|V|-1} \sum_{v \in V, v \neq w} \sigma(v; G \setminus \{w\}). \tag{3}$$

もし，$E(G \setminus \{w\})$ の値がネットワーク G に対する期待影響度 $E(G)$ とほとんど変わらなければ，ノード w の有無にかかわらずネットワーク上のほぼ同程度のノードに情報が拡散していることから，ノード w はネットワーク上の情報拡散にほとんど関与していないといえる．一方，$E(G \setminus \{w\})$ と $E(G)$ の差が大きくなった場合，ノード w がネットワーク G から削除されたことにより，ネットワーク上で情報が拡散されなくなることを表していることから，ノード w がネットワーク上の情報拡散に大きく寄与しているといえる．

従って，本稿で対象とする超媒介者は，以下の式を満たすノード $\hat{w}$ である．

$$\hat{w} = \arg\max_{w \in V}(E(G) - E(G \setminus \{w\})). \tag{4}$$

なお，式 4 の $E(G)$ はノード w の値に関わらず定数であることから，最終的には，以下の式を満たすノード $\hat{w}$ を求める．

$$\hat{w} = \arg\min_{w \in V} E(G \setminus \{w\}). \tag{5}$$

$\hat{w}$ を求めるためのアルゴリズムを以下に示す．

for $w \in V$ **do**
  for $v \in V, v \neq w$ **do**
    for $m = 1$ to M **do**
      Compute $a(v; G \setminus \{w\})$.
      Set $x_m \leftarrow a(v; G \setminus \{w\})$.
    end for
    Set $\sigma(v; G \setminus \{w\}) \leftarrow (1/M)\sum_{m=1}^{M} x_m$.
  end for
  Set $E(G \setminus \{w\}) \leftarrow (1/(|V|-1))\sum_{v \in V, v \neq w} \sigma(v; G \setminus \{w\})$.
  if $E(G \setminus \{\hat{w}\}) > E(G \setminus \{w\})$ **then**
    Set $\hat{w} \leftarrow w$
  end if
end for

提案手法のアルゴリズムでは，先述の $E(G)$ を求めるアルゴリズムの計算量 $O(M \cdot |V| \cdot E(G))$ と比較して，およそ $|V|$ 倍の計算量 $O(M \cdot |V| \cdot (|V-1|) \cdot E(G \setminus \{w\}))$ を必要とするため，ネットワークのノード数の増加に伴い計算負

荷が問題となる．この問題を緩和するため，本稿では，ボンドパーコレーション法 [7] を利用する．$E(G)$ を推定する際，Kempe らの手法では，全ての $v \in V$ に対して初期アクティブノード v から M 回のシミュレーションをする必要があるため，計算量は $O(M \cdot |V| \cdot E(G))$ となる．一方，ボンドパーコレーション法では，ネットワーク G 上の IC 及び LT モデルが G 上のボンドパーコレーションモデルと同値であることを利用することで，計算量は $O(M \cdot |E|)$ となる．従って，期待影響度 $E(G)$ の値が大きくなるようなネットワークを対象にした場合に，特に計算速度の向上が見込まれる．実際に，先行研究では数万ノードの規模の社会ネットワークによる実験により，Kempe らの手法の数十倍の速度で $E(G)$ を計算できることを明らかにしている．

5. 評価実験

提案手法の性能を評価するため，実際に情報をやり取りした情報伝搬経路のネットワーク構造である 2 つの実ネットワークに提案手法を適用する．比較として，影響最大化問題における高影響力ノードや，ネットワークの中心性等によってノードをランキングし，実験評価する．

5.1 使用データと基本統計量

5.1.1 ブログのトラックバックネットワーク

一つ目は，ブログサイト”goo ブログ”[*1]から，2005 年 5 月に収集したブログのトラックバックネットワーク [10] である（以下，blogNW）．あるブログ記事 X がブログ記事 Y をトラックバックすると，ブログ記事 X からブログ記事 Y に対してリンクが張られる．このネットワークは「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 段まで辿って収集したもので，ノード数は 12,407 ノード，リンク数は 53,315 本である．

5.1.2 エンロンのメール送受信ネットワーク

二つ目は，エンロンの e-mail データセット [9] を利用したメールの送受信ネットワークである．各 e-mail アドレスをノードとみなし，ノード X からノード Y へメールを送信した場合，ノード X からノード Y へリンクが張られる．このネットワークのノード数は 19,603 ノード，リンク数は 210,950 本である．

5.2 ネットワークの基本統計量

表 1 に各ネットワークの基本統計量を示す．ここで，ネットワークの平均次数を $\bar{d}$ とする．続いて入次数と出次数の関係を表す入出次数相関 dc を以下の式で定義する [11].

$$dc = \frac{\frac{1}{|V|}\sum_{u\in V}|B(u)||F(u)| - \bar{d}^2}{\sqrt{\frac{1}{|V|}\sum_{u\in V}|B(u)|^2 - \bar{d}^2}\sqrt{\frac{1}{|V|}\sum_{u\in V}|F(u)|^2 - \bar{d}^2}}. \quad (6)$$

さらに，ネットワークのリンク密度 L は以下の式で定義

[*1] http://blog.goo.ne.jp/

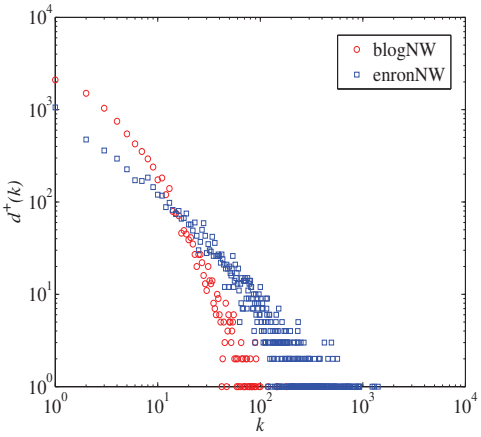

図 1 各ネットワークの出次数分布
Fig. 1 Out-degree distributions of each network.

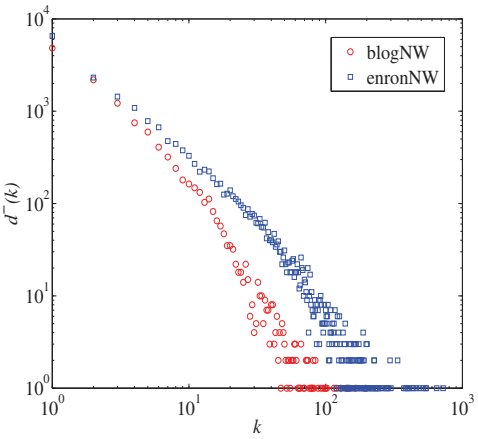

図 2 各ネットワークの入次数分布
Fig. 2 In-degree distributions of each network.

される．

$$L = \frac{|E|}{N(N-1)}. \quad (7)$$

また，ネットワーク全体に対する相互関係を表す双方向リンク率 cr は以下の式で定義される [11].

$$cr = \frac{1}{|E|}\sum_{v\in V}|F(v)\cap B(v)|. \quad (8)$$

なお，双方向リンク率は mixi ネットワーク [12] のように，リンクが全て双方向となるネットワークでは 1 となる．

平均次数は，enronNW の方が高く，リンク密度も高い．一方，双方向リンク率は blogNW が 2 倍以上高い．また，入出次数相関は両ネットワーク共に高く，入次数の大きさと出次数の大きさには強い相関がある．

続いて，2 つのネットワーク出次数分布，入次数分布を図 1，図 2 にそれぞれ示す．横軸には，親ノード集合，子ノード集合の要素数 k，縦軸には，親ノード集合，子ノード集合の要素数が k であるノード数 $d^-(k) = \{v : |B(v)| = k\}$，$d^+(k) = \{v : |F(v)| = k\}$ をそれぞれプロットしている．

表 1　各ネットワークの基本統計量
Table 1　Basic statistics of each network

	enronNW	blogNW
$\bar{d}$	10·8	4·4
dc	0·55	0·50
L	0·0011	0·00073
cr	0·21	0·50

図 1，図 2 より，入次数分布，出次数分布共にべき乗則に従っていることから，多くの社会ネットワークで見られるスケールフリー性を有している．

5.3　比較指標

5.3.1　影響最大化問題における高影響力ノード

影響最大化問題における高影響力ノードは，情報を出来るだけ多くのノードに効果的に拡散することができるノードである．

$$I(v) = \sigma(v; G). \tag{9}$$

5.3.2　次数中心性

ノード v の次数とは，ノード v とつながっているノード数である．ノード v の次数を $deg(v) = |F(v) \cup B(v)|$ を定義する．次数中心性とは、次数が高いノードほど重要ノードであるという考えに基づいた指標である [13].

$$\mathrm{dec}(v) = deg(v). \tag{10}$$

5.3.3　近接中心性

ノード v の近接度とは，ネットワーク内の他のノードへの近さを表す指標である．近接中心性とは，ほかの多くのノードへ少ないステップで到達できる，ネットワークの中心にあるノードは重要であるという考えに基づいた中心性指標である [14].

$$\mathrm{clc}(v) = (\sum_{u \in V, v \neq u} d(v,u))^{-1}. \tag{11}$$

ここで，$d(v,u)$ はノード v とノード u の最短パス長を表す．

5.3.4　媒介中心性

ノード v の媒介度とは，任意のノードペアを結ぶパスを，どの程度媒介しているかを示す指標である．媒介中心性とは，多くのノード間の橋渡しをしているノードは重要であるという考えに基づいた指標である [15].

$$\mathrm{bwc}(v) = \sum_{s \in V} \sum_{t \in V} \frac{\gamma_{s,t}(v)}{\gamma_{s,t}}. \tag{12}$$

ここで，$\gamma_{s,t}$ はノード s,t 間の最短パス数であり，$\gamma_{s,t}(v)$ はノード v を通るノード s,t 間の最短パス数を表す．

5.3.5　RageRank 値

PageRank は，Web ページの重要度を測るアルゴリズムである [17]．重要な Web ページからリンクされている Web ページは重要な Web ページであると仮定し，そのような Web ページには高い PageRank スコアを与える．以下の式の反復が収束したときの，定常ベクトル $\pi^{(k)T}$ を PageRank 値ベクトルといい，v 番目の要素 $\pi^{(k)T}(v) = \mathrm{prk}(v)$ は，ノード v の PageRank 値である．

$$\pi^{(k+1)T} = \pi^{(k)T}\mathbf{G}. \tag{13}$$

ここで，$\mathbf{G}$ は，Google 行列を表す．

5.3.6　HITS アルゴリズムの Hub 度と Authority 度

HITS アルゴリズムは，あるトピックに関するオーソリティと，関連オーソリティにリンクしているハブを求めるアルゴリズムである [16]．各ノードの Hub 度（望ましい Authority ノードにリンクを張っている度合い）と，Authority 度（望ましい Hub ノードからリンクを張られている度合い）を反復計算により算出する．

$$\mathrm{hub}^{(k)}(v) = \sum_{u \in F(v)} \mathrm{authority}^{(k-1)}(u), k = 1, 2, \cdots \tag{14}$$

$$\mathrm{authority}^{(k)}(v) = \sum_{u \in B(v)} \mathrm{hub}^{(k-1)}(u), k = 1, 2, \cdots \tag{15}$$

ここで，$\mathrm{hub}^{(k)}(v)$ は，k 反復目のノード v の Hub 度，$\mathrm{authority}^{(k)}(v)$ は，k 反復目のノード v の Authority 度をそれぞれ表す．

5.4　実験設定

4.1 節にて述べた 2 つの実ネットワークを利用して，各ネットワークの超媒介者を検出する．まず，$E(G \setminus \{w\})$ を推定する際のシミュレーション回数を $M = 1,000$ とした．

IC モデルでの拡散確率 $p_{u,v}$ は，先行研究 [4]， [8] によれば，ネットワークの平均次数の逆数として一様に設定されているため，本稿でも $p = \frac{1}{d}$ と設定し，確率 p を，任意の有向リンク (u,v) に対する拡散確率 $p_{u,v}$ として割り当てた．LT モデルに重み $w_{u,v}$ も，先行研究 [4]， [8] と同様に，任意のノード v に対して親ノード $u \in B(v)$ からの重み $w_{u,v}$ を，$w_{u,v} = \frac{1}{|B(u)|}$ で一様に与えた．

5.5　実験結果

blogNW 上で IC モデルを用いて検出した超媒介者ノードランキング上位と各指標のランキングを表 2 に，enronNW 上で IC モデルを用いたランキングを表 3 にそれぞれ示す．

1，2 列目は提案手法により検出した超媒介者ノードのランキングとそのノード ID を表し，3〜9 列目は各比較指標に基づいてランキングしたノードが，提案手法によってランキングされた場合に何位にランキングされるのかを表している．すなわち，順位の高いノードが表れるほど提案手法と関連が強い指標であることを示す．

表 2　blogNW,IC モデルの超媒介者ノードと各指標との比較
Table 2　Comparison of the ranking by Super-madiator with rankings by each index in blogNW using IC model.

Ranking by $E(G \setminus \{w\})$		Ranking by $I(v)$ /dec/prk/hub/authority/clc/bwc						
Rank	NodeID	$I(v)$	dec	prk	hub	authority	clc	bwc
1	2210	319	41	374	41	383	18	81
2	6386	83	374	82	383	41	319	90
3	4811	18	39	41	316	371	485	381
4	4230	127	383	160	371	493	215	237
5	149	503	235	39	39	60	374	1
6	7749	195	23	470	60	93	28	507
7	658	411	371	173	93	39	83	185
8	3628	485	319	59	164	346	504	258
9	10559	28	215	206	346	164	227	334
10	641	353	60	403	174	181	81	89

表 3　enronNW,IC モデルの超媒介者ノードと各指標との比較
Table 3　Comparison of the ranking by Super-madiator with rankings by each index in enronNW using IC model.

Ranking by $E(G \setminus \{w\})$		Ranking by $I(v)$ /dec/prk/hub/authority/clc/bwc						
Rank	NodeID	$I(v)$	dec	prk	hub	authority	clc	bwc
1	642	86	462	241	15	324	323	323
2	845	283	15	462	462	323	324	462
3	377	440	510	1	437	77	77	510
4	2627	508	241	391	391	204	462	1
5	4968	481	324	472	510	462	510	516
6	3332	187	516	61	514	52	204	324
7	6327	327	1	15	241	394	52	63
8	9084	448	63	27	324	124	124	241
9	657	271	323	129	336	516	516	15
10	9613	355	277	63	227	242	1	79

まず，表 2 の blogNW での，提案手法によって検出されたノードと既存の高影響度ノードとの結果を比較すると，提案手法で上位にランキングされたノードは既存の高影響度ノードとは一致しないことが分かる．また，そのほかの代表的なネットワーク指標のランキング結果と比較しても，提案手法と各ネットワーク指標にてランキングされたノードとはほとんど一致しない．表 3 の enronNW での分析結果をみても，提案手法とそのほかの指標で上位にランキングされたノードとはほとんど一致しない．従って，提案手法により既存の高影響度ノードやネットワーク指標では得られないようなノードが検出されている．なお，既存の高影響度ノードとネットワーク指標で上位にランキングされたノードを比較してもほとんど一致しないことから，既存の高影響度ノードもネットワーク指標では得られないようなノードが検出されていると言える．また，ネットワーク指標間で上位にランキングされたノードを比較すると，いくつかの指標間では類似したノードが上位にランキングされているものが見られる．

続いて，blogNW 上で LT モデルを用いて検出した超媒介者ノードのランキングと各指標のランキング結果を表 4 に，enronNW 上で LT モデルを用いた結果を表 5 にそれぞれ示す．

LT モデルを用いた場合でも，表 4 の blogNW，表 5 の enronNW 共に，提案手法とそのほかの指標で上位にランキングされたノードとはほとんど一致しない．また，IC モデルを用いた結果と同様の傾向として，ネットワーク指標間で上位にランキングされたノードを比較すると，いくつかの指標間は類似したノードが上位にランキングされている．なお，IC モデルと異なる特徴として，既存の高影響度ノードと次数や PageRank，HITS などの指標で上位にランキングされたノードが類似する傾向がある．このような結果が得られたのは，IC モデルでは，ノード v はアクティブとなった親ノードから一度でも情報が伝わればアクティブになるのに対して，LT モデルはノード v の閾値を超えるだけのアクティブとなった親ノードが必要であるため，IC モデルよりも情報が拡散しにくい．したがって，中心性が高いノードでなければ情報を拡散できる可能性がほとんど得られないからであると考えられる．

表 4　blogNW,LT モデルの超媒介者ノードと各指標との比較
Table 4　Comparison of the ranking by Super-madiator with rankings by each index in blogNW using LT model.

Ranking by $E(G \setminus \{w\})$		Ranking by $I(v)$ /dec/prk/hub/authority/clc/bwc						
Rank	NodeID	$I(v)$	dec	prk	hub	authority	clc	bwc
1	6644	499	130	42	130	220	277	299
2	4355	49	42	85	220	130	499	46
3	87	186	185	130	234	157	357	153
4	490	158	220	19	157	380	321	443
5	4387	130	158	185	185	208	42	242
6	4203	497	186	73	208	99	18	376
7	6580	321	157	2	99	185	450	21
8	5666	185	499	214	318	243	107	436
9	6805	277	321	4	243	318	49	346
10	668	42	208	34	59	461	299	117

表 5　enronNW,LT モデルの超媒介者ノードと各指標との比較
Table 5　Comparison of the ranking by Super-madiator with rankings by each index in enronNW using LT model.

Ranking by $E(G \setminus \{w\})$		Ranking by $I(v)$ /dec/prk/hub/authority/clc/bwc						
Rank	NodeID	$I(v)$	dec	prk	hub	authority	clc	bwc
1	573	3	37	385	242	3	272	272
2	1803	272	242	37	37	272	3	37
3	203	94	94	56	139	233	233	94
4	708	56	385	349	349	276	37	56
5	879	19	3	132	94	37	94	19
6	1101	37	19	68	341	210	276	3
7	1797	137	56	242	385	152	210	462
8	2849	233	462	376	3	17	17	385
9	11650	17	272	261	407	19	19	242
10	10166	342	137	462	194	270	56	342

5.6　超媒介者が期待影響度に与える影響

最後に，提案手法によって検出されたノードが，ネットワーク全体の期待影響度に寄与する度合いを評価する．表 6 に，ネットワーク G における期待影響度 $E(G)$ と，提案手法によって推定された $\hat{w}$ を削除した部分グラフでの期待影響度 $E(G \setminus \{\hat{w}\})$ をそれぞれ示す．なお，本節では提案手法において $E(G \setminus \{\hat{w}\})$ を最も小さくしたノードの値のみを評価する．

IC モデルを用いた場合の各期待影響度を比較すると，blogNW では部分グラフでの期待影響度 $E(G \setminus \{\hat{w}\})$ はネットワーク G における期待影響度 $E(G)$ の約 1.8%になる．さらに，enronNW では，$E(G \setminus \{\hat{w}\})$ が $E(G)$ の約 0.3%まで小さくなる．したがって，IC モデルを用いた提案手法は，情報拡散に大きく寄与しているノードを検出できていると考えられる．

一方，LT モデルを用いた場合の $E(G \setminus \{\hat{w}\})$ と $E(G)$ を比較すると，blogNW で約 69%，enronNW で約 57%となり，IC モデルと比較するとそれほど情報拡散に寄与しているノードを検出できていないように見える．このような結果が得られたのは，LT モデルを用いた場合には，ネットワーク全体での期待影響度 $E(G)$ 自体が非常に小さく，情報がほとんど拡散していないため，情報を媒介するような状況がほとんど見られなかったことが原因であると考えられる．

これらの結果から，ネットワーク全体での期待影響度が高くなる場合において，提案手法が情報を多くのノードへ伝搬するための橋渡しとなる超媒介者ノードを検出することができていることが示唆される．

実際に，2 つのネットワークで超媒介者として最上位にランキングされたノードをみると，blogNW では”愛国無罪”というタイトルのブログ，enronNW では当時の Enron 社の Senior Management，Jeffrey Skilling 氏であった．

本稿で使用したエンロンの e-mail データセットは，エンロン事件に関係したメールが含まれており，Skilling 氏はこの事件に大きく関与していた．エンロン事件に関しては様々な分析が行われており，Skilling 氏が Senior Management に就任後，Senior Management のメールの送受信が急激に増加したことが報告されており，Skilling 氏が情報の

収集，拡散に重要な役割を果たしていたと推測される [18]．このことからも，提案手法が超媒介者を正確に検出できることが示唆される．

表 6 提案手法とネットワーク全体の期待影響度の比較
Table 6 Comparison proposed method with influential degree of network

	blogNW		enronNW	
	IC	LT	IC	LT
$E(G)$	180.8	3.2	890.0	2.8
$E(G \setminus \{\hat{w}\})$	3.3	2.2	2.8	1.6

6. おわりに

本稿では，ネットワーク上の情報拡散現象において，情報を多くのノードへ伝搬するための橋渡しとなるノードである超媒介者を検出する手法を提案した．まず，ネットワーク内の任意ノードに対して情報拡散シミュレーションを行うことにより得られる影響度を，ネットワーク内の全ノードに対して算出してその期待値を取ることにより，ネットワーク全体での影響度を得た．次に，ネットワークの任意のノードを削除した部分ネットワークにおいて同様に影響度の期待値を算出し，この差を大きくするノードを超媒介者として検出した．評価実験として，提案手法を 2 つの実ネットワークに適用し，検出されたノードを影響最大化問題における高影響力ノードや，ネットワークの中心性等によって検出されたノードと比較した．その結果，ネットワーク内の情報拡散に大きく寄与していると考えられるノードを検出できることが示唆された．

今後は，提案手法をより大規模なネットワークに対して適用し，その性能を評価する．また，Twitter のリツイート機能のように，情報拡散現象データとして観測できるようなネットワークに本手法を適用することで，本手法の有用性を評価する．

謝辞 本稿は科研費 (22500133) の助成を受けた．

参考文献

[1] E.Bakshy, J.Hofman, W.Mason, and D.Watts, Everyone's an Influencer: Quantifying Influences on Twitter, In Proceedings of the 4th ACM International Conference on Web Seach and Data Mining, pp.65-74, (2011).
[2] J.Leskovec, L.A.Adamic, and B.A.Huberman, The dynamics of viral marketing. In Proceedings of the 7th ACM Conference on Electronic Commerce, pp.228-237, (2006).
[3] 白井嵩士，榊剛史，鳥海不二夫，篠田孝祐，風間一洋，野田五十樹，沼尾正行，栗原聡，Twitter におけるデマツイートの拡散モデルの構築とデマ拡散防止モデルの推定，第 26 回人工知能学会全国大会，(2012).
[4] K. Saito, M. Kimura, K. Ohara, and H. Motoda, Discovery of super-mediators of information diffusion in social networks, In Proceedings of the 13th international conference on Discovery science, pp.144-158, (2010).
[5] 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).
[6] 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).
[7] M.Kimura, K.Saito, R.Nakano, Extracting Influential Nodes for Information Diffusion on a Social Network, In Proceedins of the Advancement of Artificial Intelligence, pp.1175-1180, (2008).
[8] D.Kempe, J.Kleinberg, and E.Tardos, Maximizing the spread of influence through a social network, In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.137-146, (2003).
[9] B.Klimt, and Y.Yang, The enron corpus: A new dataset for email classification research, In Proceedings of the 2004 European Conference on Machine Learning, pp.217-226, (2004).
[10] Kimura, M., Saito, K., Motoda, H., Blocking links to minimize contamination spread in a social network, In Proceedings of the ACM Transactions on Knowledge Discovery from Data, 3(2), pp.9:1-9:23, (2009).
[11] K.Ohara, K.Saito, M.Kimura, and H.Motoda, Effect of In/Out-Degree Correlation on influence Degree of Two Contrasting Information Diffusion Models, In Proceedings of the 2012 International Conference on Social Computing, Behavioral Modeling, and Prediction, pp.131-138, (2012).
[12] 松尾豊，安田雪，SNS における関係形成原理-mixi のデータ分析-，人工知能学会論文誌，Vol. 22, No.5, pp.531-541 (2007).
[13] L. Freeman, Centrality in social networks: Conceptual clarication, In Proceedings of Social Networks, Vol1, No.3, pp.215-239, (1979).
[14] G. Sabidussi, The centrality index of a graph, In Proceedings of Psychometrika, 31 (4), pp.581-603, (1966).
[15] U. Brandes, A Faster Algorithm for Betweenness Centrality，In Proceedings of Journal of Mathematical Sociology, 25, pp.163-177, (2001).
[16] J.Kleinberg, Authoritative sources in a hyperlinked environent, In Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, pp.668-677, (1998).
[17] S.Brin and L.Page, The anatomy of a large scale hypertextual Web search engine, In Proceedings of the 19th International Conference on World Wide Web, pp.107-117, (1998).
[18] J. Diesner, and K. Carley, Exploration of communication networks from the enron email corpus ”it' s always about the people. enron is no different”, In Proceedings of Journal of Computational and Mathematical Organization Theory, 11(3), 201-228, (2005).