情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2009-MPS-76 No.28
Vol.2009-BIO-19 No.28
2009/12/17

成長ネットワークにおける情報伝搬モデルを用いたリンク予測

瀧上晋太郎[†1]　熊野雅仁[†2]
木村昌弘[†2]　斉藤和巳[†3]

成長するネットワークにおいて，新たに発生するリンクを予測する問題を考える．我々は，情報伝搬モデルを用いた新たなリンク予測法を提案し，大規模な実データを用いた実験により，提案法の有効性を示す．また，ネットワーククラスター分解を組み込むことによる提案法の性能向上についても調べる．

Link Prediction for Growing Networks Using Information Propagation Model

Shintaro Takigami ,[†1] Masahito Kumano ,[†2]
Masahiro Kimura [†2] and Kazumi Saito[†3]

We consider the problem of predicting newly generated links in a growing network, and propose a method based on an information propagation model. Using large real data in blog bookmark and guestbook networks, we demonstrate the effectiveness of the proposed method, and also investigate the performance improvement by incorporating network clustering.

†1 龍谷大学大学院理工学研究科電子情報学専攻
Divsion of Electronics and Informatics, Graduate School of Science and Technology, Ryukoku University
†2 龍谷大学理工学部電子情報学科
Department of Electronics and Infromatics, Faculty of Science and Technology, Ryukoku University
†3 静岡県立大学経営情報学部
School of Administration and Informatics, University of Shizuoka

1. はじめに

複雑ネットワークにおけるリンク予測問題の研究は，企業間における新たな協業の発見，人間関係における友好関係の発見，テロリストネットワークの監視，また，将来購入する商品を推薦するリコメンデーションシステムへの応用，遺伝子やタンパク質の相互作用ネットワークから新たな生物学的発見を予測する問題，さらに，情報検索，Web ハイパーリンクの自動生成などへの応用も期待されており，近年，注目を集めている[1)]．本論文では，このリンク予測問題に着目する．

リンク予測問題は，大きく二つに分類することができる．一つは，静的なネットワークにおいて，すでに観測されたリンク構造から未観測のリンク構造を予測する問題であり，もう一つは，成長ネットワーク（動的なネットワーク）において，現在のリンク構造から将来のリンク構造を予測する問題である．本論文では，後者の “成長ネットワークにおけるリンク予測問題” に取り組む．

成長する社会ネットワークのリンク予測問題に関しては，Liben-Nowell と Kleinberg[3)] によるノードの “proximity” に基づいたリンク予測法がよく知られている．ところで，現在はリンクで結ばれていないノードペアでも，一方のノードから他方のノードに情報が伝わりやすいならば，そのノード間は将来においてリンクで結ばれる可能性が高いと推測される．したがって，情報伝搬に基づいてリンクを予測するという手法の開発は，期待できるアプローチの一つと考えられる．本論文では，社会ネットワーク上での情報伝搬の基本モデルとして広く用いらている，Independent Cascade (IC) モデル[2),4),5)] に基づいた成長ネットワークにおけるリンク予測法を提案し，ブログゲストブックマークネットワークおよびブログブックマークネットワークにおける実データを用いた実験により，提案法の有効性を比較検証する．また，ネットワーククラスター分解を組み込むことによる提案法の性能向上についても調べる．

2. リンク予測問題

無向ネットワークとして表現される成長する社会ネットワークに対して，そのリンク予測問題を考える．$G = (V, E)$ をその現在のネットワーク構造とし，$G' = (V', E')$ をその一定期間後のネットワーク構造とする．ここに，V と V' はそれぞれ G と G' のノード全体の集合であり，$E(\subset V \times V)$ と $E'(\subset V' \times V')$ はそれぞれ G と G' のリンク全体の集合である．現在のネットワーク構造 G において，ノード $u \in V$ の隣接ノード全体の集合を

© 2009 Information Processing Society of Japan

Vol.2009-MPS-76 No.28
Vol.2009-BIO-19 No.28
2009/12/17

$$A(u) \ = \ \{w \in V;\ (u,w) \in E\} \tag{1}$$

で表す．また，現在のネットワーク構造 G において，共通の隣接ノードが存在するがリンクで結ばれていないノードペア全体の集合を，

$$S \ = \ \{(v,w) \in V \times V \setminus E;\ d(v,w) = 2\} \tag{2}$$

とする．ここに，$d(v,w)$ はノード v とノード w のグラフ G における距離である．S に属するリンクを “潜在リンク” と呼ぶ．潜在リンクのうち新たに生成されたリンク全体の集合を

$$L \ = \ S \cap (E' \setminus E) \tag{3}$$

とする．本論文では，現在のネットワーク構造 G が与えられたとき，L に属するリンクを予測する問題を考える．

3. 提 案 法

3.1 IC モデル

まず，Kempe らの研究[4] に従って，IC モデルの定義を述べる．

IC モデルでは，各リンク (v,w) に対して，$0 < p_{v,w} < 1$ なる実数 $p_{v,w}$ を前もって指定する必要がある．ここに，$p_{v,w}$ はリンク (v,w) を通じての “情報伝搬確率” と呼ばれる．IC モデルによる情報伝搬過程は離散時間 $t \geq 0$ で展開していく．IC モデルに基づく情報伝搬では，情報が伝わったノードを “アクティブ” と呼ぶ．ノードはその状態が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と仮定される．

IC モデルの情報伝搬過程は，初期アクティブノード（情報源ノード）が与えられたとき，次のように進んでいく．ノード v が時刻 t で初めてアクティブになったとき，v は，非アクティブであるその各隣接ノード w をアクティブにする試行を時刻 t で行い，その試行は確率 $p_{v,w}$ で成功する．もし，w の複数の隣接ノードが時刻 t で初めてアクティブになった場合は，それら隣接ノードが w 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，w 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，w は時刻 $t+1$ においてアクティブとなる．ところで，v が時刻 t で w 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，v はもはや w をアクティブにする試行を行うことはできない．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，本情報伝搬過程は終了する．

3.2 リンク生成モデル

IC モデルに基づいた次のようなリンク生成モデルにより，現在のネットワーク構造 G から一定期間後のネットワーク構造 G' を構築する．

(1) G における任意のノード $u \in V$ に対して，$v \in A(u)$ と $w \in V \setminus A(u)$ を任意に選ぶ．
(2) ノード v を初期アクティブノードとして，G 上で IC モデルに基づいた情報伝搬過程を実行する．
(3) ノード w がアクティブになった（ノード w へ情報が伝搬した）ならば，G' においてノード u とノード w の間にリンク (u,w) を新たに生成する．

我々のリンク生成モデルでは，IC モデルにおける情報伝搬確率 $\{p_{v,w}; (v,w) \in E\}$ を，あらかじめ指定する必要があることに注意しておく．

3.3 情報伝搬確率の推定

リンク生成モデルのパラメータである情報伝搬確率 $\{p_{v,w}; (v,w) \in E\}$ を，現在のネットワーク構造 G から推定する．

任意のノード $u \in V$ に対して，ノード集合 $G^+(u)$ と $G^-(u)$ を次のように定義する．

$$G^+(u) = \{(v,w) \in E;\ v,w \in A(u)\} \tag{4}$$

$$G^-(u) = \{(v,w) \in E;\ v \in A(u), w \notin A(u)\} \tag{5}$$

現在のネットワーク構造 G が観測されるのは，$(v,w) \in G^+(u)$ に対しては v から w への情報伝搬が成功し，$(v,w) \in G^-(u)$ に対しては v から w への情報伝搬が失敗したためと推測して，次の目的関数 J の最大化問題として，パラメータ $\{p_{v,w}; (v,w) \in E\}$ を最尤推定法に基づいて推定することを考える．

$$J \ = \ \log \prod_{u \in V} \left\{ \prod_{(v,w) \in G^+(u)} p_{v,w} \prod_{(v,w) \in G^-(u)} (1 - p_{v,w}) \right\} \tag{6}$$

さて，$(v,w) \in E$ に対して，ノード集合 $H^+(v,w)$ と $H^-(v,w)$ を次のように定義する．

$$H^+(v,w) \ = \ \{u \in V;\ (v,w) \in G^+(u)\} \tag{7}$$

$$H^-(v,w) \ = \ \{u \in V;\ (v,w) \in G^-(u)\} \tag{8}$$

このとき，目的関数 J はを以下のように書き表すことができる（式 (6), (7), (8) 参照）．

$$J \ = \ \sum_{(v,w) \in E} \left\{ |H^+(v,w)| \log p_{v,w} \ + \ |H^-(v,w)| \log(1 - p_{v,w}) \right\} \tag{9}$$

ここに，$(v,w) \in E$ に対して次の式が成り立つ．

$$|H^+(v,w)| \ = \ |A(v) \cap A(u)|, \tag{10}$$

$$|H^-(v,w)| \ = \ |A(v) \cup A(u)| \ - \ |A(v) \cap A(u)| \ - \ 2 \tag{11}$$

よって，式 (9), (10), (11) より，パラメータ $\{p_{v,w}; (v,w) \in E\}$ の最尤推定値 $\{\hat{p}_{v,w}; (v,w) \in E\}$ は，

© 2009 Information Processing Society of Japan

Vol.2009-MPS-76 No.28
Vol.2009-BIO-19 No.28
2009/12/17

$$\hat{p}_{v,w} = \frac{|A(v) \cap A(u)|}{|A(v) \cup A(u)| - 2} \qquad (v,w) \in E \tag{12}$$

となる．我々は，Laplace smoothing を適用して，

$$\hat{p}_{v,w} = \frac{|A(v) \cap A(u)| + 1}{|A(v) \cup A(u)|} \qquad (v,w) \in E \tag{13}$$

と推定する．

3.4 リンク予測法

我々は，潜在リンク $(v,w) \in S$ が実リンクに変化する確率（変換確率）$q_{v,w}$ を，背後にある IC モデルに基づいてノード v からノード w へ情報が伝搬する確率，

$$q_{v,w} = 1 - \prod_{u \in A(v) \cap A(w)} (1 - p_{u,v})(1 - p_{u,w}) \tag{14}$$

として推定する．

現在のネットワーク構造 G における潜在リンクのうち，一定期間後のネットワーク構造 G' において新たにリンクとなるものを k 本予測する．ここに，k は与えられた正の整数である，k 個の要素からなる予測リンク集合を，$B(k)(\subset S)$ とする．提案法では，$B(k)$ を次のように抽出する．

Step 1. 任意のリンク $(v,w) \in E$ に関して，式 (13) を用い情報伝搬確率 $\hat{p}_{v,w}$ を推定する．
Step 2. 任意の潜在リンク $(v,w) \in S$ に関して，式 (14) を用い潜在リンクが実リンクに変化する確率 $q_{v,w}$ を計算する．
Step 3. 変換確率 $q_{v,w}$ の値に関して潜在リンク $(v,w) \in S$ をランキングすることにより，予測リンク集合 $B(k)$ を抽出する．

3.5 スペクトラルクラスタリング

多くの社会ネットワークは，コミュニティー構造をもっている[6]．ここに，コミュニティー構造とは，ネットワークのクラスター分解であり，異なるコミュニティー（クラスター）間には比較的少数のリンクしか存在しないが，コミュニティー（クラスター）内ではリンクが密集しているというものである．多くの成長する社会ネットワークにおいては,「時間が経過するとともに，コミュニティー内でのリンク結合はますます増大するが，コミュニティー間のリンク結合は疎なままである.」ということが予想される．そこで，現在のネットワーク G をスペクトラルクラスタリングによりクラスター分解し，各クラスター（G より小さいネットワーク）に提案法を適用する手法を調べた．スペクトラルクラスタリングには，Ng ら[7] のアルゴリズムを用いた．

表 1 GB-10 ネットワークと GB-11 ネットワークの成長
Table 1 Growth of the GB-10 and the GB-11 networks

GB-10 ネットワーク	5 日後	15 日後
$\|E' \setminus E\|$	22,838	64,334
$\|L\|$	3,464	8,573
$\|L\|/\|E' \setminus E\|$	15.1%	13.3%

GB-11 ネットワーク	5 日後	15 日後
$\|E' \setminus E\|$	23,778	69,040
$\|L\|$	3,237	8,170
$\|L\|/\|E' \setminus E\|$	13.6%	11.8%

4. 実験評価

4.1 実験データ

Yahoo!ブログのゲストブックネットワークとアメーバブログのブックマークネットワークの実データを用いて，提案法の性能を評価した．

ブログゲストブックネットワークに関しては，二つのネットワークデータを構築した．一つ目のネットワークデータは，2008 年 10 月 16 日に $G = (V,E)$ を収集し，それぞれ 21 日と 31 日に再び，同じ部分のゲストブックネットワーク $G' = (V',E')$ を収集したものである．以降，本ネットワークデータを “GB-10 ネットワーク” と呼ぶ．また，21 日に収集された G' を “5 日後の G'” と呼び，31 日に収集された G' を “15 日後の G'” と呼ぶ．もう一つのネットワークデータは，2008 年 11 月 15 日に $G = (V,E)$ を収集し，それぞれ 20 日と 30 日に再び，同じ部分のゲストブックネットワーク $G' = (V',E')$ を収集したものである．以降，本ネットワークデータを “GB-11 ネットワーク” と呼ぶ．また，GB-10 ネットワークと同様に，20 日に収集された G' を “5 日後の G'” と呼び，30 日に収集された G' を “15 日後の G'” と呼ぶ．ここに，GB-10 ネットワークに関しては，$|V| = 44,742$，$|E| = 83,507$，$|S| = 2,363,781$ であり，GB-11 ネットワークに関しては，$|V| = 43,079$，$|E| = 79,265$，$|S| = 3,087,224$ であった．また，GB-10 ネットワークと GB-11 ネットワークにおける 5 日後および 15 日後の成長量，$|E' \setminus E|$，$|L|$ および $|L|/|E' \setminus E|$ を，表 1 に示す．

ブログブックマークネットワークに関しては，2006 年 5 月に $G = (V,E)$ を収集し，1ヶ月後に再び同じ部分のブックマークネットワーク $G' = (V',E')$ を収集した．ここに，$|V| = 56,894$，$|E| = 535,734$，$|S| = 156,874,190$ であり，$|E' \setminus E| = 41,220$，$|L| = 30,849$ であった．すなわち，新たに生成されたリンクのうち潜在リンクであったものの割合 $|L|/|E' \setminus E|$ は 75% であった．以降，本ネットワークデータを “BM ネットワーク” と呼ぶ．

© 2009 Information Processing Society of Japan

Vol.2009-MPS-76 No.28
Vol.2009-BIO-19 No.28
2009/12/17

4.2 比 較 法

提案法の性能を評価するために，Liben-Nowell と Kleinberg[3] によるノードの proximity に基づいたリンク予測法と比較した．また，ベースラインとして，予測リンク集合 $B(k)$ を抽出するために，潜在リンク集合 S から一様ランダムに k 本のリンクを選択する Random 法も調べた．

まず，Liben-Nowell と Kleinberg[3] による従来法と比較した．我々のリンク予測問題において，彼らの手法は，任意の $(v,w) \in S$ の “proximity”$px(v,w)$ を定義し，その値に従って潜在リンクをランキングすることにより，予測リンク集合 $B(k)$ を抽出することになる．彼らは，共著ネットワークを用いた実験において，以下の Adamic/Adar-proximity が最も高性能であることを示したが，我々は，Adamic/Adar-proximity を含む以下の三つの proximity と比較した．

- Common Neighors (CN):
$$px(v,w) \ = \ |A(v) \cap A(w)| \tag{15}$$
- Adamic/Adar (A/A):
$$px(v,w) \ = \ \sum_{z \in A(v) \cap A(w)} \frac{1}{\log |A(z)|} \tag{16}$$
- Preferential Attachment (PA):
$$px(v,w) \ = \ |A(v)|\,|A(w)| \tag{17}$$

以降，我々はこれらリンク予測法をまとめて “proximity 法” と呼ぶ．

4.3 評 価 尺 度

提案法および比較法では，与えられた正の整数 k に対して，ランキングにより予測リンク集合 $B(k)$ を抽出している．我々は，これらリンク予測法の性能を，ランク k における予測リンク集合 $B(k)$ の精度，

$$P(k) \ = \ \frac{|L \cap B(k)|}{|B(k)|} \tag{18}$$

で評価する．

4.4 実 験 結 果

4.4.1 GB-10 と GB-11 ネットワークにおける結果

図 1(a) は，GB-10 ネットワークにおける 10 月 21 日の予測性能を，また，**図 1(b)** は，10 月 31 日における予測性能 $P(k)$ を示している．表 1 の GB-10 ネットワークにおいて，5 日後は 3,464 のリンク，また，15 日後は 8,573 の新たなリンクが生成されているため，予測リンク集合 $B(k)$ のランク k が 1,000〜10,000 の範囲にある予測精度が問われるが，**図 1** では，ランクの上位での予測性能を確認するため，ランク k が 100 の場合についても提示した．まず，図 1(a) と図 1(b) の比較により，提案法は，5 日後よりも，15 日後の予測性能が高いことがわかる．ところで，図 1(a) では，提案法のランク k が 1,000 の場合より 100 の場合に精度が落ちている．しかし，新たなリンクの増加数を考えれば，ランク k が 100 の場合の精度よりも，1,000〜10,000 の精度に注目すべきである．しかし，いずれの場合においても，提案法は，比較法よりも予測性能が高いことがわかる．また，ランク k が 100 の場合において，従来法より，倍程度の精度が得られていることから，提案手法は，従来法よりランキングの上位から数多くの予測すべきリンクを含んでいる点で優位性があると言える．

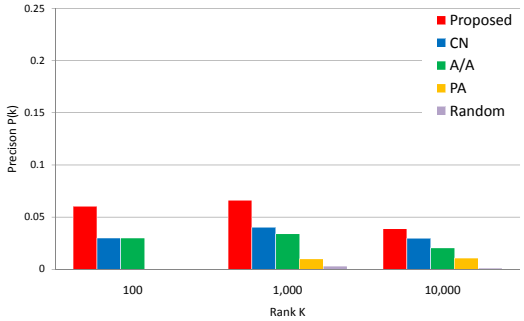
次に，GB-10 ネットワークでの予測性能が異なる期間においても成立するかを検証するため，GB-11 ネットワークでの予測性能との比較を行う．**図 2(a)**, **図 2(b)** は，それぞれ GB-11 ネットワークにおける 11 月 20 日，30 日の予測性能 $P(k)$ を示している．図 1 との比較により，いずれにおいても比較法との関係は変わっておらず，比較法に比べ，提案法の性能が明らかに高いことがわかる．

しかし，図 1，**図 2** 共に，予測精度の値を見た場合，精度が高かった 15 日後の結果においても，数%〜十数%程度の精度しかない．これは，そもそも GB-10 ネットワークの場合の潜在リンク数 $|S|$ が 2,363,781，GB-11 ネットワークの場合でも，3,087,224 の潜在リンクから数千程度の新たに生成されたリンクを予測する極めて難しい問題と言える．その点で，無作為にリンクを予測する場合と比べ，どの程度優位であるかを検証することがリンク予測問題の有効性を説明することになると思われる．そこで，無作為にリンクを予測する Random 法との相対的な比較を行う．

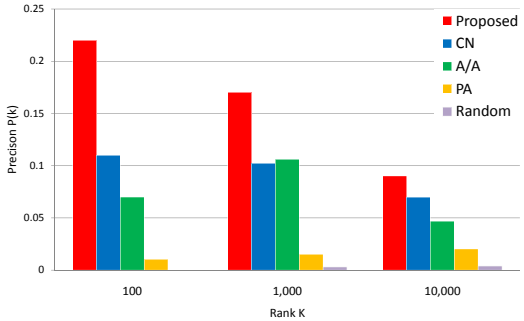
表 2 と表 3 は，Random 法の予測性能 $P_{Random}(k)$ に対する提案法の予測性能 $P(k)$ の相対値 $P(k)/P_{Random}(k)$ を表示している．ただし，Random 法との比較としては，精度の良かった 15 日後の予測結果を用いる．また，$Rank(k)$ として，100，1,000，10,000 だけでなく，15 日後までに生成されたリンク数について，10 月の場合の 8,573，また 11 月の場合の 8,170 の新規リンク増加時点における相対的倍率についても提示した．表 2 と表 3 か

© 2009 Information Processing Society of Japan

Vol.2009-MPS-76 No.28
Vol.2009-BIO-19 No.28
2009/12/17

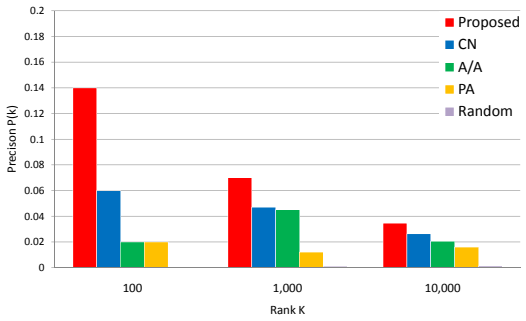


(a) GB-10 ネットワークにおける 5 日間予測

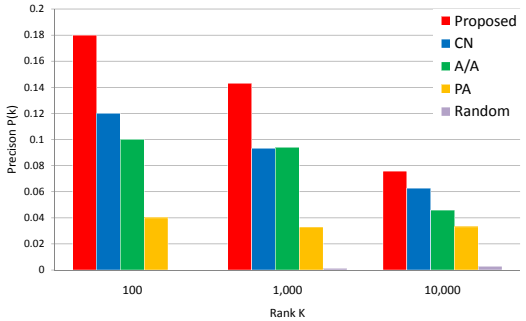


(b) GB-10 ネットワークにおける 15 日間予測

図 1 GB-10 ネットワークにおける提案法と proximity 法による予測性能
Fig. 1 Prediction performance comparison of the propose method with the proximity methods for the GB-10network



(a) GB-11 ネットワークにおける 5 日間予測



(b) GB-11 ネットワークにおける 15 日間予測

図 2 GB-11 ネットワークにおける提案法と proximity 法による予測性能
Fig. 2 Prediction performance comparison of the propose method with the proximity methods for the GB-11network

ら，Random 法に比べ，提案法は，30 倍程度の優位性があることがわかる．

次に，GB ネットワークをクラスター分解し，各クラスターに対し提案法を適用した実験の結果について述べる．まず，GB ネットワークが，ある程度のコミュニティに分解できる仮定し，その場合，コミュニティ間のリンクはそれほど増加しないと考えられることから，ク

© 2009 Information Processing Society of Japan

Vol.2009-MPS-76 No.28
Vol.2009-BIO-19 No.28
2009/12/17

表 2 GB-10 ネットワークにおける 15 日間予測結果
Table 2 Results for the fifteen-days-prediction in the GB-10network

$Rank(k)$	100	1,000	8,573	10,000
提案法	22	170	810	903
Random 法	0	3	23	39
Random 法に対する提案法の相対的倍率	-	56.7	35.2	23.2

表 3 GB-11 ネットワークにおける 15 日間予測結果
Table 3 Results for the fifteen-days-prediction in the GB-11network

$Rank(k)$	100	1,000	8,170	10,000
提案法	18	143	674	758
Random 法	0	2	20	27
Random 法に対する提案法の相対的倍率	-	71.5	33.7	28.1

ラスター分解した結果，クラスター間のリンクがどの程度減るかを検証するため，GB ネットワークを 1〜6 まで分解した場合のリンク減少数を調査した．

表 4, 表 5 は，GB ネットワークにおける 10 月 15 日と 11 月 15 日のネットワークをクラスタリングした結果である．表 4, 表 5 における，A はクラスター数 n における $|E|$ を表している．B はクラスター数 n における $|E|$ と $n=1$ における $|E|$ との割合を表している．また，C はクラスター数 n における $|S|$ を表しており，D はクラスター数 n における $|S|$ と $n=1$ における $|S|$ との割合を表している．表 4 より，$n=1$〜5 ではリンク数が減少せず，$n=6$ で急激に減少することから，五つ程度のコミュニティーが存在する可能性があると言える．また，表 5 では，$n=5$ でリンク数が減少することから，四つ程度のコミュニティーが存在する可能性がある．

このようなクラスター分解の状況下において，提案法を適用し，予測精度の変化を検証するための実験を行った．**図 3** は，10 月 16 日の GB ネットワークを $n=1$〜6 までクラスター分解した場合の 15 日後の予測性能の結果である．クラスタリングを行わない $n=1$ と比較して，$n=5$ までのクラスター分解では，提案法の予測性能に変化がないことがわかる．従って，提案法は，ある程度のクラスター分解を行っても優位性を保つことがわかる．

また，**図 4** は，11 月 15 日の GB ネットワークをクラスター分解した場合の 15 日後の予測性能の結果である．クラスタリングを行わない $n=1$ と比較して，$n=4$ までは，予測性能が下がらないことがわかる．しかし，$n=5$ では，予測性能が高くなる場合もあることがわかった．この結果から，リンク予測問題を，より小規模のクラスターに分解して適用できるという可能性とともに，より予測精度を高める手法を開拓できる可能性も期待される．

表 4 10 月 16 日のクラスタリング結果
Table 4 Clustering results of October 16

n	1	2	3	4	5	6
A	83,507	83,324	83,321	83,313	83,313	48,442
B	100.0%	99.8%	99.8%	99.7%	99.7%	58.0%
C	2,363,781	2,363,779	2,363,772	2,363,760	2,363,643	723,103
D	100.0%	99.9%	99.9%	99.9%	99.9%	30.6%

表 5 11 月 15 日のクラスタリング結果
Table 5 Clustering results of November 15

n	1	2	3	4	5	6
A	79,265	79,079	79,079	79,073	69,346	40,868
B	100.0%	99.8%	99.8%	99.8%	87.5%	51.6%
C	3,087,224	3,087,102	3,087,102	3,087,097	2,610,712	353,526
D	100.0%	99.8%	99.8%	99.8%	84.6%	11.5%

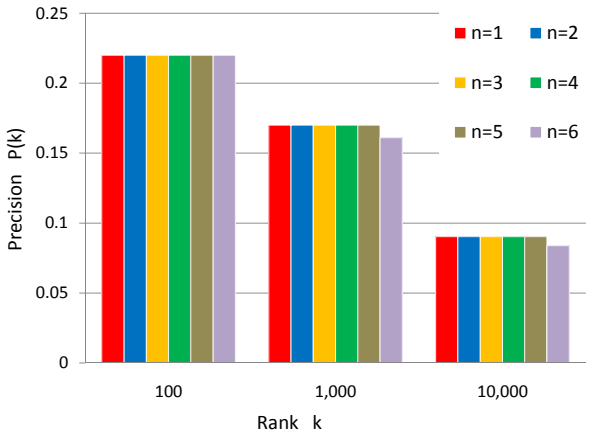


図 3 GB-10 ネットワークにおけるネットワーククラスタリングを組み込んだ手法の性能
Fig. 3 Performance improvement by incorporating network clustering for the GB-10 network

4.4.2 BM ネットワークにおける結果

図 5 は，BM ネットワークを対象としてリンク予測実験を行った結果である．BM ネット

© 2009 Information Processing Society of Japan

Vol.2009-MPS-76 No.28
Vol.2009-BIO-19 No.28
2009/12/17

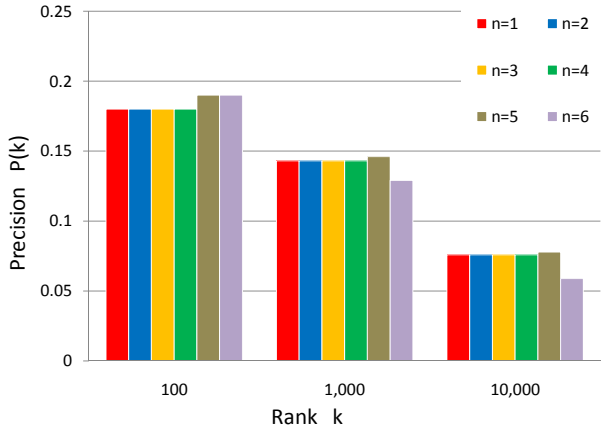


図 4 GB-11 ネットワークにおけるネットワーククラスタリングを組み込んだ手法の性能
Fig. 4 Performance improvement by incorporating network clustering for the GB-11 network

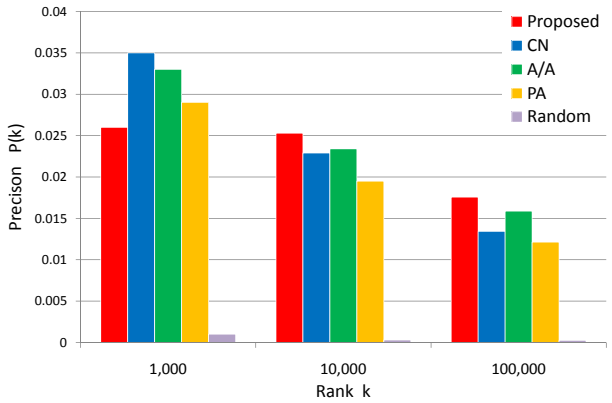


図 5 BM ネットワークにおける提案法と proximity 法による予測性能
Fig. 5 Prediction performance comparison of the propose method with the proximity methods for the BM network

表 6 BM ネットワークにおける予測性能
Table 6 Prediction performance for the BM network

$Rank(k)$	1,000	10,000	30,849	100,000
提案法	30	253	674	1,757
Random 法	1	3	9	17
Random 法に対する提案法の相対的倍率	30.0	84.3	74.9	103.4

表 7 BM ネットワークに関するクラスタリング
Table 7 Clustering results for the BM network

n	1	2	3	4	5	6
A	535,734	535,649	535,648	535,646	336,952	303,548
B	100.0%	99.98%	99.98%	99.98%	62.90%	56.66%
C	156,874,190	156,874,134	156,874,134	156,874,134	35,682,314	23,947,240
D	100.0%	99.98%	99.98%	99.98%	22.75%	15.27%

ワークを対象とした実験では，ランク k が 1,000 の場合，比較法よりも予測精度が低い結果となり，GB ネットワークを対象とした実験との違いが現れた．しかし，BM ネットワークのノード数 $|V|$ が，GB ネットワークの 44,742 に比べ，56,894 と規模が大きいだけでなく，GB ネットワークが 15 日後までの予測を行うのに対し，BM ネットワークは，1 ヵ月後の予測を行う点で，増加するリンク数の桁が異なり，1 ヵ月後に増加するリンクは 30,849 本にも及ぶ．予測精度を検証する意味では，ランク k の 10,000〜100,000 の領域で評価することが望ましく，その観点においては，比較法より提案法に優位性があることがわかる．

また，GB ネットワークの場合と同様に，Random 法との相対的倍率を比較した．表 6 より，Random 法と比較して，$Rank(k)$ のいずれの場合においても，高い優位性があり，1ヵ月後のリンク増加数と一致する 30,849 の場合においては，74.9 倍に及ぶことがわかる．また，BM ネットワークについても，クラスター分解を行った場合の予測性能の変化を検証する実験を行った．

表 7 は，BM ネットワークをクラスター分解した結果である．A はクラスター数 n における $|E|$ を表している．B はクラスター数 n における $|E|$ と $n=1$ における $|E|$ との割合を表している．また，C はクラスター数 n における $|S|$ を表しており，D はクラスター数 n における $|S|$ と $n=1$ における $|S|$ との割合を表している．

表 7 より，BM ネットワークにおいても，$n=4$ までは予測性能が落ちないことから，GB ネットワークと同様に，より小規模のクラスターに分解してリンク予測問題を適用できるという可能性とともに，より予測精度を高める手法を開拓できる可能性も期待される．

© 2009 Information Processing Society of Japan

Vol.2009-MPS-76 No.28
Vol.2009-BIO-19 No.28
2009/12/17

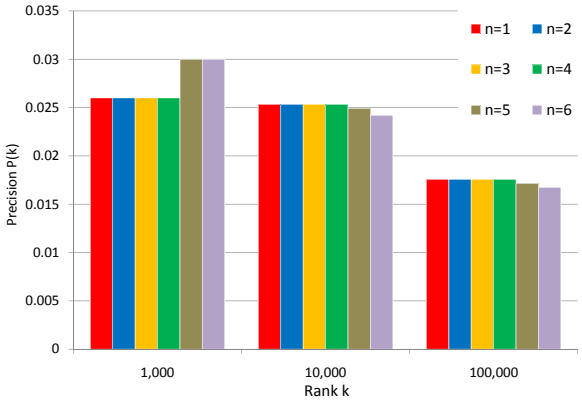


図 6 BM ネットワークにおけるネットワーククラスタリングを組み込んだ手法の性能
Fig. 6 Performance improvement by incorporating network clustering for the BM network

また，**図 6** は，BM ネットワークについて分解されたクラスターごとに提案法を適用し，予測性能を検証した結果である．図 4 と同様に，クラスター分解によって，予測性能が上がる場合があり，GB ネットワークと BM ネットワークで同様の結果が得られたことから，コミュニティー構造を持つ他のネットワークにも同様に適用し，有効性が得られる可能性を示唆しているものと思われる．

5. ま と め

情報伝搬モデルを用いた，成長する社会ネットワークのリンク予測法を提案した．大規模な実データとして GB ネットワークと BM ネットワークを用いた実験により，提案法は，Liben-Nowell と Kleinberg[3] による従来法よりも，予測性能の点で有効性を示した．

成長する社会ネットワークの多くがコミュニティー構造を内包していると仮定すれば，よいクラスター分解をすることにより，ネットワークの予測問題を，より小さな問題に置き換えることが可能となり，提案法の予測性能も向上する可能性が考えられる．そこで，ネットワークをクラスター分解した場合の提案法の予測性能の変化を調べた．

本研究において，クラスター分解を施した実験では，$n = 4$ までクラスター分解を行ったいずれの場合においても予測性能が変化しない結果が得られたことから，これら実データには，コミュニティー構造があることが示唆された．

また，$n = 5$，$n = 6$ では，予測性能が低下する場合も見受けられたが，向上する場合もあった．このことから，クラスター分解の手法を工夫することにより，提案法によって，予測性能が向上する可能性が示唆されたものと思われる．

今後，クラスター分解の手法やどの程度までのクラスター分解を行うことが望ましい結果を得られるかなど，相互の関係を検証していく予定である．

参 考 文 献

1) Getoor, L. and Diehl, C.P.: Link minig: A survey, *SIGKDD Explorations*, Vol.7, No.2, pp.84–89, 2005.
2) Goldenberg, J., Libai, B., and Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, Vol.12, No.3, pp.211–223, 2001.
3) Liben-Nowell, D. and Kleinberg, J.: The link prediction problem for social networks, *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM'03)*, pp.556–559, 2003.
4) Kempe, D., Kleinberg, J., and Tardos, E.: 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, 2003.
5) Kimura, M., Saito, K., and Motoda, H.: Blocking links to minimize contamination spread in a social network, *ACM Transactions on Knowledge Discovery from Data*, Vol.3, No.2, Article 9, 2009.
6) Newman, M.E.J. and Park, J.: Why social networks are different from other types of networks. *Physical Review E*, Vol.68, 036122, 2003.
7) Ng, A.Y., Jordan, M.I., and Weiss, Y.: On spectral clustering: Analysis and an algorithm, *Advances in Neural Information Processing Systems (NIPS'01)*, Vol.14, pp.849–856, 2001.

© 2009 Information Processing Society of Japan