

リンク張替に対するモチーフ解析の頑健性評価

伏見 卓恭[†] 大久保誠也[†] 武藤 伸明[†] 斉藤 和巳[†]

[†] 静岡県立大学経営情報学研究科 〒422-8526 静岡県静岡市駿河区谷田 52-1

E-mail: †{j09118,s-okubo,muto,k-saito}@u-shizuoka-ken.ac.jp

あらまし 複雑な有向ネットワークを特徴付ける指標として、ネットワークモチーフがある。現実問題に適用する場合、ネットワークの全体構造を完全に把握することは困難であり、ノイズが混入したネットワークデータとなる。ノイズの含まれたネットワークに対するモチーフ解析の結果が、真のネットワークに対する結果と異なることは不可避である。本研究では、大規模な複雑ネットワークに対して、リンク張替のノイズを混入することで、ネットワークモチーフ解析の頑健性を評価する。

キーワード 複雑ネットワーク, モチーフ, リンク張替, 頑健性

The Robustness Evaluation of Motif Analysis for the Link Rewiring

Takayasu FUSHIMI[†], Seiya OKUBO[†], Nobuaki MUTO[†], and Kazumi SAITO[†]

[†] School of Management and Information, University of Shizuoka

52-1 Yada, Suruga-ku Shizuoka, Shizuoka, 422-8526 Japan

E-mail: †{j09118,s-okubo,muto,k-saito}@u-shizuoka-ken.ac.jp

Abstract In complex network field, network motifs are the basic characteristics of interconnections. To apply the real network problems, it is hard to catch the whole network structures completely. Thus networks include the noise and it is natural that the motif frequency of the true network differ from that of noisy network. In this paper, we evaluate the robustness of motif analysis for the link rewiring by using the large scale networks.

Key words complex network, motif, link rewiring, robustness

1. はじめに

近年様々な分野において、大規模・複雑な事象をネットワークとしてとらえ、ノード間の相互関係やネットワーク構造、ネットワーク上での現象を分析する研究が盛んに行われている。

ネットワークを評価するための指標としてネットワークモチーフ [1] がある。ネットワークモチーフは、特徴的なノード同士の繋がり方のパターンの中で、遺伝子ネットワークの分析などで用いられている。特徴的なモチーフパターンを調べることにより、ネットワーク内で頻出するノード間の相互関係パターンやネットワーク上での現象への影響について、知見を得ることができる [4] ~ [6]。

社会科学の分野では、友人・知人などの人間関係のネットワークを対象としたソーシャルキャピタル(人間関係の豊かさ)についての研究が行われている。人間関係の豊かさは、人間同士の繋がり方が1つの要因だと考えられ、地域や文化によって異なることが知られている。これらの人間関係は、人間をノード、関係をリンクとすることで、ネットワークとしてモデル化できる。

しかしながら、このような現実の事象に適用する場合、プラ

イバシーなどの問題から、ネットワークを構成するノードが具体的に他のどのノードと関係を有するかを正確に把握することは困難である。したがって、社会調査などから得られるエゴセントリック情報(自身の次数や年齢, 所属ほか)などの部分的な情報から、ネットワークの全体構造を推定することが必要となる。構造推定問題においては、正確なネットワーク構造を推定することは不可能であり、推定結果には、リンク張替などある種のノイズが含まれることは不可避である。いくらノイズの含まれるネットワークにおいても、真のネットワークと同様にネットワークモチーフを得ることができれば、推定したネットワークを用いて研究することが可能である。

ネットワークにノイズを混入したネットワークモチーフの既存研究として Kaluza らの [2] がある。文献 [2] では、ノード削除やリンク削除などのノイズに対して、頑健なネットワークの構造やモチーフパターンの分析を行っている。

本研究では、文献 [2] とは異なるノイズを採用する。具体的には、出次数・入次数を固定したリンク張替というノイズに対するモチーフ解析の頑健性を評価する。これは、以下のような現実問題をモデル化したノイズ混入法となっている。

ある人(ノード A)を中心に考えたときに、ノード A から別

の人に相談すればノード A からの出リンク、ノード A が別の人から相談されればノード A への入リンクとなるような、相談関係に基づく社会ネットワークを考える。アンケートなどより、ある期間内に相談した人数と相談された人数やそれらの人々の属性（性別、年齢、所属ほか）などのエゴセントリック情報を取得し、その情報に基づき推定したネットワークを対象とする。アンケート結果からは具体的なリンク先を全て把握することは難しいため、推定したネットワークでのリンク集合 \hat{E} は、真のネットワークでのリンク集合 E^* とはある程度の違いが生じる。すなわちリンク張替のノイズが含まれることになる。推定したネットワークでのモチーフと真のネットワークでのモチーフが変わらない、すなわちモチーフ解析が頑健ならば、真のネットワークを完全に推定できなくても、真のネットワークにおけるモチーフを検出できる。このことを探究するために、リンク張替というノイズの下でのモチーフ解析の頑健性を評価する。

本稿の構成は次の通りである。2 章では頑健性評価の方法としてネットワークの張替およびモチーフ解析法について述べる。3 章では、評価実験に用いるネットワークデータについて述べ、モチーフ解析の頑健性について評価する。最後に、本研究のまとめと今後の課題について述べる。

2. 頑健性評価法

分析対象とするネットワーク（真のネットワークと呼ぶ）に対して、次数を固定し、任意の張替確率でリンクを張り替える。張替確率を変化させたネットワークに対して、モチーフ頻度の $Zscore$ を算出、プロットすることにより、張替確率（ノイズ混入率）と $Zscore$ の変化の関係を分析する。

2.1 リンク張替

真のネットワークのリンク張替法について説明する。張替ネットワークは真のネットワークのリンクを、各ノードの出次数・入次数を変化させずに、張替確率 p に従ってリンクをランダムに張り替えることで構築する [3]。詳細は、下記に示す。

(i) ネットワークの有向リンクリスト $L_E = (e_1, \dots, e_{|E|})$ を用意する

e_1 はリンクの両端のノードからなる順序対 $(L_{from}(1), L_{to}(1))$ であり、第 1 要素から第 2 要素へのリンクを示す。

(ii) 有向リンクリスト L_E から 2 つのノードリスト L_{from} と L_{to} を生成する

(iii) ノードリスト L_{to} の 2 つの要素を張替確率 p でスワップし、部分的に順序が変わったノードリスト L'_{to} を得る

(iv) ノードリスト L_{from} と ノードリスト L'_{to} とを結合し L'_E を得る

得られたリンクリスト $L'_E = ((L_{from}(1), L'_{to}(1)), \dots, (L_{from}(|E|), L'_{to}(|E|))) = (e'_1, \dots, e'_{|E|})$ は張替ネットワークの有向リンクを表す。但し、リンク張替の際には、自己リンクと多重リンクにならないようにする。張替モデルでのリンクのスワップを図示したものが図 1 である。図 1 では、赤と青の実線リンクがスワップ対象のリンクで、それぞれ赤と青の点線リンクに張り替えている。スワップをした結果、真のネットワークである図 1 (1) と張替ネットワークである図 1 (3)

での、各ノードの出次数・入次数は変化していない。図 2 は、多重リンクのためリンクを張り替えられない例である。

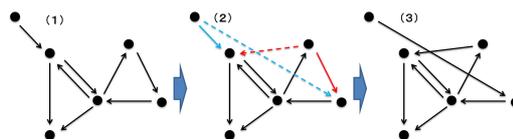


図 1 リンク張替例

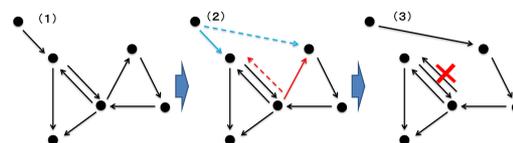


図 2 リンク張替できない例

2.2 モチーフ解析

ネットワークモチーフは、ネットワーク内のリンクパターン（モチーフパターン）の出現頻度を数え上げることで、そのネットワークの特徴的なモチーフパターンを抽出する解析法である。図 3 に [1] で示された 3 ノードのモチーフパターンを示す。

真のネットワークに対して、これら 13 種のモチーフパターンがそれぞれどの程度出現するかを数え上げる。すなわち、真のネットワークの任意の 3 ノードを抽出し、その 3 ノードがなす連結成分に関して 13 パターンのうちのどのパターンであるかを調べ数え上げる。モチーフパターン i の出現頻度を x_i とする。これをすべての 3 ノード連結成分に対し実行し、そのネットワークにおける 13 パターンの頻度ベクトル x が作られる。

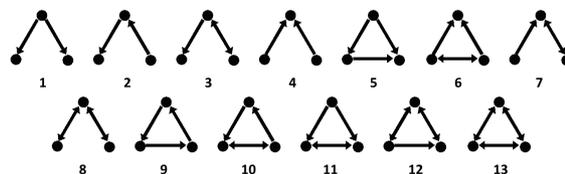


図 3 3 ノードのモチーフパターン

計算された各モチーフパターンの頻度は、同規模のネットワークにおいても同様な頻度で出現するものか、それともそのネットワークに特有なモチーフパターンなのかを検定する必要がある。真のネットワークとランダムネットワークの出現頻度の間に有意な差があるかを検定する尺度として $Zscore$ が知られている。以下に $Zscore$ の計算法について説明する。

$Zscore$ を計算するために、ランダムネットワークを生成する。ランダムネットワークの生成法は、2.1 節で説明したリンク張替法と同様である。すなわち、ノードの入次数・出次数を固定し、全リンクをランダムに張り替える（張替確率 1）。但し、自己リンクと多重リンクにならないように注意する。1000 個程度のランダムネットワークを生成し、各々に対してモチーフパターン頻度を計算する。1000 個のうち、 j 番目に生成したランダムネットワークのモチーフパターン i の出現頻度を $y_i^{(j)}$

とする．モチーフパターン i の Zscore は次式で定義される．

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \quad (1)$$

ここで， $\mu_i = \frac{1}{1000} \sum_j^{1000} y_i^{(j)}$ ， $\sigma_i = \frac{1}{1000} \sum_j^{1000} (y_i^{(j)} - \mu_i)^2$ である．

Zscore が正で大きいモチーフパターンほど，そのネットワークにおいて統計的に有意に頻出するといえる．すなわちそのネットワークの特徴的なモチーフとなる．逆に Zscore が負で絶対値の大きいモチーフパターンほど，有意に出現しないといえる．

3. 評価実験

本研究では，上述した相関関係のネットワークなどの人間関係のネットワークを対象とするため，その近似となる 4 つの大規模なネットワークを採用し，モチーフ解析の頑健性を評価する．またウェブ上の情報は，比較的完全に近い状態で収集できサンプルとして妥当性があるため，ウェブ上での社会ネットワークを採用する．

3.1 ネットワークデータ

1 つ目のネットワークは，ブログサービスサイト“ Ameba ”^(注1)の読者ネットワークである．ブログはユーザが簡単に使用できるソフトウェアにより管理されたオンライン日記で，World Wide Web で急速に普及したサービスである．“ Ameba ”の各ブログには読者のブログへのハイパーリンクを張る機能がある．ここで，ブログ X からブログ Y へのリンクは，ブログ Y がブログ X をお気に入りブログとして登録した時に生成される．従って読者ネットワークは社会ネットワークとみなすことができる．我々は 2006 年 6 月に Ameba ブログサービスサイトの 117,374 ブログの読者リストをクロールし，連結成分を抽出した．このネットワークは 56,604 ノードと 734,737 有向リンクを有している．コリンク率は 0.54 である．以下このネットワークを Ameblo ネットワークと呼ぶ．

2 つ目のネットワークは，化粧品品の口コミサイト“ @cosme ”^(注2)にお気に入り登録したユーザのファンネットワークである．サイトに登録しているあるユーザ X がユーザ Y をお気に入り登録をすると，ユーザ X からユーザ Y に対してリンクが張られる．このようにして生成されたネットワークは社会ネットワークとみなすことができる．このネットワークは，2009 年 12 月にランダムに選択したユーザから 10 段辿って収集したもので，ノード数は 45,024 でリンク数は 351,299 である．コリンク率は 0.44 である．以下このネットワークを Cosme ネットワークと呼ぶ．

3 つ目のネットワークは，エンロンの Email データセットである．各 email 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 X から Y への有向リンクを張り社会ネットワークを構築する．最大の強連結成分を抽出して生成されたネットワークは 19,603 ノードと 210,950 有向リンクで構成される．コリンク率は 0.21 である．以下このネット

ワークを Enron ネットワークと呼ぶ．

最後のネットワークは，映画情報サイト yahoo 映画^(注3)のユーザファンネットワークである．yahoo 映画では，ユーザが映画に対するレビューを記述する機能や，ユーザ同士を繋ぐファン機能が存在する．その機能に着目し，ファン機能からネットワークを構築する．ユーザをノードとし，あるユーザ X が他のユーザ Y をファン登録するとノード X からノード Y へリンクが張られる．このファンネットワークは社会ネットワークとみなすことができる．2008 年 6 月から 2009 年 6 月までの期間を取得し，ノード数 5,524 でリンク数は 76,314 となった．コリンク率は 0.56 である．以下このネットワークを Movie ネットワークと呼ぶ．

3.2 結果と考察

3.1 節で説明した 4 つのネットワークに対して，張替確率 $p = 2^{-k}$ ($k = 0, 1, \dots, 10$) で異なる張替ネットワークを生成し，それぞれ Zscore を計算した．図 5～図 8 はそれぞれ Ameblo，Cosme，Enron，Movie ネットワークに対する結果を示す．図 4 は各図の凡例である．横軸にモチーフパターンをとり，縦軸は Zscore の絶対値に対数とっている．通常モチーフ解析では，Zscore の正負および絶対値により，ネットワーク特有のモチーフであるかに焦点を当てるが，本研究では，張替確率と Zscore の変化の関係および頑健性に注目するため，負の Zscore の絶対値をとっても差支えない．異なるマーカーは異なる張替確率による張替ネットワークの Zscore を意味する．Zscore の定義より，張替確率 $p = 1$ の張替ネットワークは Zscore ベクトルが 0 になることが自明なので，プロットしていない．図 5～図 8 から，どのネットワークでもモチーフパターン 8,12,13 が特徴的であることが見てとれる．この 3 つのモチーフはどれもコリンクを含むリンクパターンである．Enron ネットワークでは，他のネットワークと比較して 8,12,13 のパターンがあまり顕著でない結果となっている．3.1 節で触れたように，Enron ネットワークは他のネットワークと比べるとコリンク率が低く，このことがモチーフ解析の結果にも反映されている．

どのネットワークでも，張替確率を大きくするに伴って，Zscore がほぼ線形に減衰している．さらに， $p = 2^{-3}$ から $p = 2^{-2}$ 程度まで，真のネットワーク（凡例：Original）での Zscore の順序を保っている．すなわち，ネットワークのモチーフ分布を保存しながら減衰している．また， $p = 2^{-2}$ 程度のノイズが含まれていても，特徴的なモチーフパターンを判別することができる．

これらのネットワークを用いた実験の結果，75%～87%程度推定できているネットワークに対するモチーフをもって，真のネットワークのモチーフを推定することができることを示唆している．すなわち，25%程度のノイズが含まれていても，8,12,13 が特徴的なモチーフパターンであることが評価できる．

4. おわりに

本研究では，大規模なネットワークデータを複数用いて，

(注1): <http://www.ameba.jp/>

(注2): <http://www.cosme.net/>

(注3): <http://movies.yahoo.co.jp/>

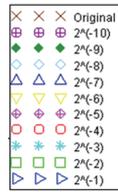


図 4 凡 例

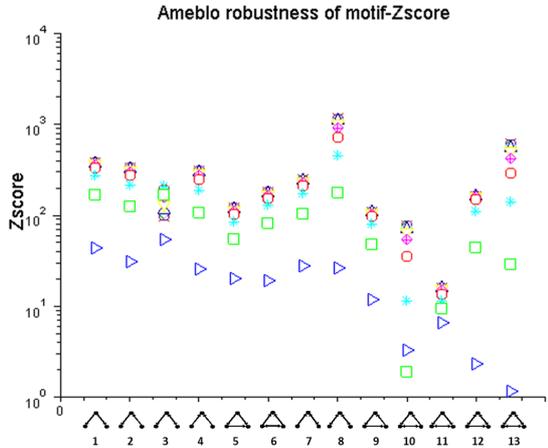


図 5 Ameblo NW Zscore

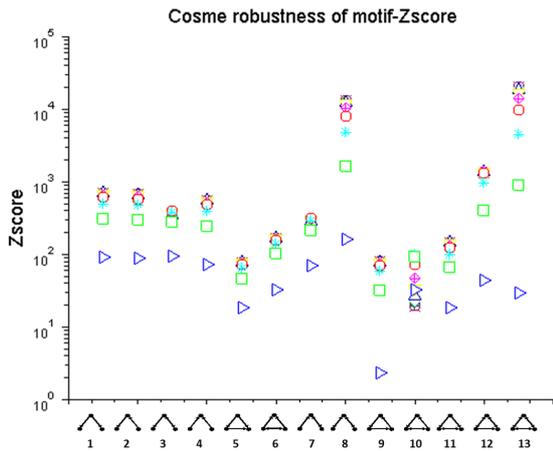


図 6 Cosme NW Zscore

ネットワークのモチーフ解析の頑健性を評価した。その結果、 $p = 2^{-3}$ から $p = 2^{-2}$ 程度のリンク張替に対して、真のネットワークでの Zscore 間の順序関係を保ち、かつ特徴的なモチーフを識別できる頑健な結果を示した。すなわち 75%~87%程度推定できているネットワークに対するモチーフをもって、真のネットワークのモチーフを推定することができることが示唆された。この実験結果より、推定した人間関係のネットワークのモチーフが、真のネットワークのモチーフ、人間関係の豊かさを示すソーシャルキャピタルの推定などに活かせることが期待できる。

今後の課題として、現実問題への適用を視野に入れ、次数固定のリンク張替以外のノイズに対するモチーフ解析の頑健性を探究していきたい。さらに、異なる特徴やネットワーク構造を

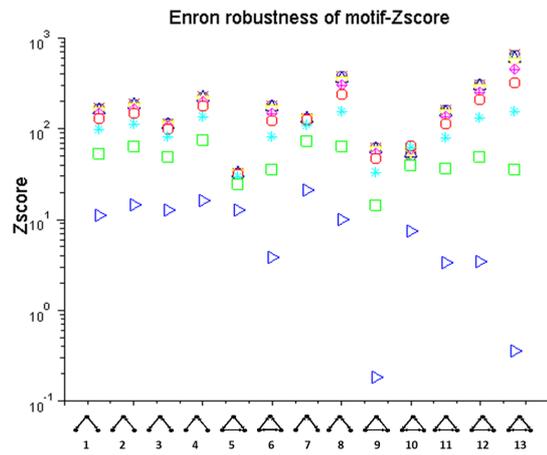


図 7 Enron NW Zscore

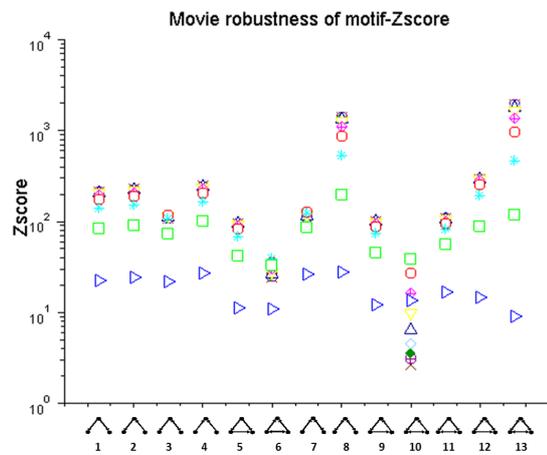


図 8 Movie NW Zscore

持つネットワークに対しての評価も進めていきたい。

謝辞 本研究は、科学研究費補助金基盤研究 (C) (No. 22500133) の補助を受けた。

文 献

- [1] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, "Network motifs: simple building blocks of complex networks." Proc. of Science 298, pp. 824-827, 2002.
- [2] P. Kaluza, M. Vingron and AS. Mikhailov, "Self-correcting networks: function, robustness, and motif distributions in biological signal processing." Proc. of Chaos 18, 026113, pp. 1-18, 2008.
- [3] M. E. J. Newman, "The structure and function of complex networks." Proc. of SIAM Review, 45, pp. 167-256, 2003.
- [4] 高田寛喜, 斉藤和巳, 上田修功, "時系列情報を考慮したモチーフパターン変化の分析", 電子情報通信学会第 7 回 Web インテリジェンスとインタラクション研究会, 2006.
- [5] 島田諭, 小出明弘, 斉藤和巳, 佐藤哲司, "QA2 部グラフにおけるモチーフを用いたコミュニティの経時的変化に関する分析", 知識共有コミュニティワークショップ (KSCWS), 2010.
- [6] 小出明弘, 斉藤和巳, 佐藤哲司, "モチーフによる QA 二部グラフの構造分析", Web とデータベースに関するフォーラム (WebDB Forum), 2010.