

Twitter の @-message で構成される成長ネットワークの分析*

小出 明弘† 齊藤 和巳† 大久保 誠也† 鳥海 不二夫‡ 風間 一洋§
 静岡県立大学† 名古屋大学‡ NTT未来ねっと研究所§

1. はじめに

ソーシャルネットワーク上の社会ネットワーク構造は、ユーザ同士の新たな交流が行われることで絶えず成長している。ネットワークの成長をモデル化することは、人間関係における友好関係の発見や、バイラルマーケティングなどへの応用が期待され、盛んに研究がおこなわれている[1, 2].

本分析では、成長モデル構築の第一歩として、Twitter 上の @-message ネットワークのリンク数増加に伴う連結成分構造において、ノード数、連結成分数、Gini係数などの指標が、どのように変化をするのかを解析する。

分析により、Twitter の @-message ネットワークは、緻密化べき乗則が成立すること、及び、ランダムに生成されたネットワークに比べ、初期の段階から各連結成分に属するノード数に偏りがあることを示す。

2. 分析手法

2.1 基本統計量分析

成長ネットワークの、時刻 t でのノード数を $n(t)$ 、リンク数を $e(t)$ 、連結成分数を $c(t)$ と定義する。多くの成長ネットワークでは、時間の経過によるノード数の増加に対して、リンク数が指数関数に比例して増加する傾向が見られることが報告されている。この特徴は、緻密化べき乗則 [3] と呼ばれ、以下の式で定義される。

$$e(t) \propto n(t)^\alpha, (\alpha > 1) \quad (1)$$

この式は、 α の増加によりネットワークが密構造になることを示しており、Twitter の @-message ネットワークでも、このような傾向がみられるのか分析した。

2.2 Gini 係数

リンク数の増加に対する連結成分数の増減だ

* Analysis of growing network which was composed @-message of Twitter

† Akihiro Koide, Kazumi Saito, Seiya Okubo · University of Shizuoka

‡ Fujio Toriumi · Nagoya University

§ Kazuhiro Kazama · Nippon Telegraph and Telephone Corporation

けでは、各連結成分のノード数の分布がわからない。これを定量的に評価するため、本分析では、主に所得配分の不平等さを測る指標である Gini 係数[4]を用いる。

時刻 t の連結成分 i ($1 \leq i \leq c(t)$)のユーザ数を $n_i(t)$ で表すと、時刻 t の Gini 係数 $G(t)$ は、次式で定義される。

$$G(t) = \frac{\sum_{i=1}^{c(t)-1} \sum_{j=i+1}^{c(t)} |n_i(t) - n_j(t)|}{(c(t) - 1) \sum_{k=1}^{c(t)} n_k(t)} \quad (2)$$

本式は、ネットワークが1つの連結成分になれば1となり、複数の同規模の連結成分により構成されるネットワークでは、0に近い値を示す指標である。

3. 実験

3.1 分析データ

3月7日から23日までのTwitterに日本語で投稿された@userを含むTweetから、各ユーザをノードとし、@userで指定されたユーザにリンクしてネットワークを作成した。各リンクに時刻を付与することで、ネットワークの成長過程を時系列的に分析する。最終的に生成されたネットワークのノード数は4,440,126、リンク数は47,402,750である。比較として、ユーザ間のリンク関係は変えずに、各リンクに付与された時刻をランダムに振り直したデータを用いる。

3.2 分析結果

時刻 t でのリンク数 $e(t)$ と、ノード数 $n(t)$ 、連結成分数 $c(t)$ 、Gini係数 $G(t)$ の関係を、それぞれ図1、図2、図3に示す。

図1では、リンク数の増加に対し、ノード数の増加は対数プロットでほぼ直線的とみなせる。これは、Twitterの@-messageネットワークにおいて、緻密化べき乗則が成立していることを示している。また、ランダムデータと実データにほとんど差が見られないことがわかる。

図2では、リンク数の増加に対して、実データの連結成分数が、ランダムに比べ増加傾向が小さいことがわかる。さらに、どちらのデータについても、連結成分数が一定の大きさまで達

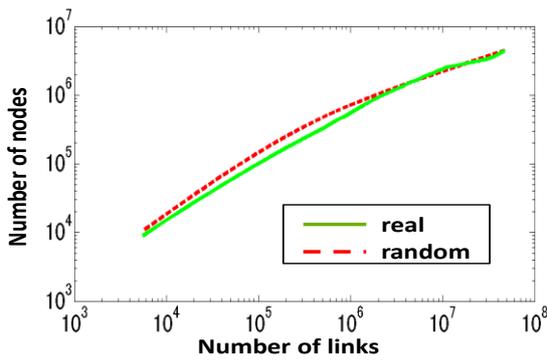


図 1: リンク数に対するノード数

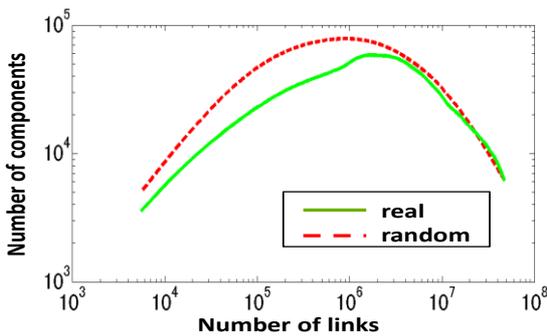


図 2: リンク数に対する連結成分数

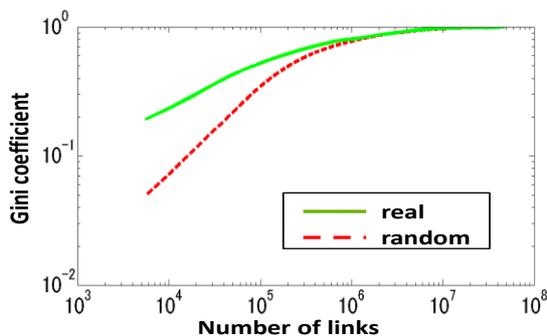


図 3: リンク数に対する Gini 係数

すると、連結成分数が減少し始めているが、実データは連結成分数がピークとなるタイミングが遅れている。また、実データはランダムに比べ、連結成分数の最大値が小さいことがわかる。

図 3 では、実データの Gini 係数が、ランダムデータに比べ、初期の段階から高い値を示していることがわかる。これは、実データでは、各連結成分に属するノード数に偏りが見られることを示している。その後、リンク数が 10^6 前後に達すると、ランダムデータとほぼ同じ値になる。

これらの分析結果により、@-message によるネットワーク構造は、初期の段階から連結成分に偏りがあり、所属ノード数の多い連結成分を中心としてネットワークが成長すると考えられる。

また、リンク数が 10^6 に達するまで、Gini 係数に増加傾向があることから、ネットワーク内の各連結成分に属するノード数の偏りは顕著になる。その一方で、連結成分は増加し続けていることを踏まえると、ほとんどのリンクは巨大化した連結成分に併合され、小規模の連結成分がネットワーク内に増加していると考えられる。

実データにランダムデータには見られない特徴が得られた理由として、各ユーザの利用形態があげられる。ユーザの活動時間は限られており、その時間内で様々なユーザへ向けて Tweet しようとする。その結果、あるユーザから複数のユーザに向けてリンクが短時間で張られる（局所的に連鎖的なツイートが発生する）ことになり、このユーザがネットワーク上でハブのような機能を有することが考えられる。このことから、各連結成分に属するノード数に偏りが見られると推測される。

4. おわりに

本研究では、Twitter 上の成長モデルを構築する第一歩として、Twitter 上における @-message ネットワークの時間変化に伴う連結成分構造を分析した。その結果、実データにはランダムデータには見られないいくつかの特徴がみられることが分かった。

今後は、本分析で得た知見を利用し、ネットワークの成長モデルを構築、評価をする。

謝辞

最後に、本研究を行なうにあたりデータをご提供いただいた、Twitter 検索を運営する @penguinana 氏に感謝する。また、本研究は NTT との共同研究および、科研費 (23500312) の助成を受けた。

参考文献

- [1] R.Albert, and A.L.Barabási,"Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, pp. 47-97,2002.
- [2] M.E.J.Newman,"The structure and function of complex networks," *SIAM Review*, 45:167-256,2003.
- [3] J.Leskovec, J. Kleinberg, and C.Faloutsos, "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations," *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, ACM, pp. 177-187, New York, NY, USA, 2005.
- [4] M.J.Salganik, P.S.Dodds, and D.J.Watts. "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," *Science* 10, pp.854-856, February 2006