

ソーシャルネットワークのコミュニティ構造が ソーシャルメディア上の投稿の拡散規模に与える影響の分析

Effects of Community Structure of a Social Network
on the Cascade Sizes of Tweets on Social Media

津川 翔
Sho Tsugawa

筑波大学 システム情報系

Faculty of Engineering, Information and Systems, University of Tsukuba

We investigate how the community structure of a social network affects information diffusion on social media. Empirical analyses using large-scale dataset of Twitter show that the community structure significantly affects the sizes of information cascades on social media. More specifically, the results of our regression analysis show that the number of retweets of a tweet spreading among different communities is significantly higher than that of a tweet spreading within a single community. We also tackle a problem of predicting the future popularity of a tweet in an early stage of its diffusion. We constructed prediction models of future popularity of a tweet using features obtained from community structures of a social network as well as the tweet's content. Through experiments, we show that the features extracted from community structure of a social network are effective for predicting the future popularity of a tweet.

1. はじめに

筆者はこれまで、ソーシャルメディアユーザ間の関係を表現したソーシャルネットワークのコミュニティ構造が、ソーシャルメディア上の投稿の拡散に与える影響を分析してきた [津川 16]。人と人との関係を表現したソーシャルネットワークの多くは、互いに密に接続されたノードのコミュニティと、それらのコミュニティ間を接続する少数のリンクで構成されるというコミュニティ構造を有している [Newman 04, Girvan 02]。このようなソーシャルネットワークのコミュニティ構造は、人と人との交流のパターンに影響を与える [Weng 13, Onnela 07, Tsugawa 15, 津川 16]。[津川 16] では、代表的なソーシャルメディアである Twitter において、ユーザのコミュニティをまたがって拡散されたツイートは、コミュニティ内で拡散されたツイートよりも、将来のリツイート数の期待値が高いこと (図 1) などを示している。

[津川 16] では、ソーシャルメディア上の投稿の拡散に影響を与える要因として、コミュニティ構造のみに着目している。一方で、コミュニティ構造以外にも様々な要因がソーシャルメディア上の投稿の拡散に影響を与えることが知られている [Suh 10, Naveed 11, Stieglitz 13]。例えば、情報を投稿したユーザのフォロワー数や投稿に URL やハッシュタグが含まれるかどうかは、投稿の拡散規模に影響を与える主要因であることが示されている [Suh 10, Naveed 11]。これらの要因と比較した時に、コミュニティ構造が投稿の拡散規模に与える影響がどの程度強いのかはこれまで明らかにされていない。

本稿では、ソーシャルネットワークにおけるコミュニティをまたがる投稿の拡散が、最終的な投稿の拡散規模に与える影響の強さを、投稿の拡散規模に影響を与える他の要因と比較することで分析する。代表的なソーシャルメディアである Twitter のデータを用いる。ツイート本文の URL の有無やツイートを投稿したユーザのフォロワー数などコミュニティ構造以外にツイートの拡散規模に影響を与える要因を制御変数とする重回帰分析により、コミュニティをまたがるツイートの拡散が、ツ

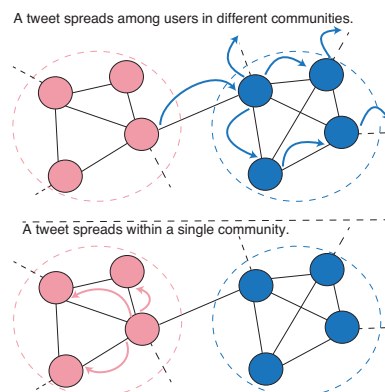


図 1: コミュニティ構造と投稿の拡散の関係：コミュニティをまたがって拡散されたツイートは、コミュニティ内で拡散されたツイートよりも、将来のリツイート数の期待値が高い [津川 16]

ィートの最終的な拡散規模にどの程度の影響を与えるかを明らかにする。さらに、ツイートの拡散規模を予測する実験により、コミュニティに関する特徴量が拡散規模の予測にどの程度有用であるかを明らかにする。

2. 分析

2.1 用いるデータ

分析には、[津川 16] と同じデータセットを用いる。本データセットには、日本語のツイートを投稿する 356,453 ユーザのフォロー関係を表現したソーシャルネットワークと、それらのユーザの投稿したリツイートの情報が含まれている。ユーザ間のフォロー関係は、Twitter API を用いて、2014 年 1 月 1 日～11 日の期間に収集した。同様に、対象ユーザ 356,453 人の投稿したリツイートは、2014 年 1 月 1 日～31 日の期間に Twitter API を用いて収集した。これにより、対象ユーザの投稿した 14,001,533 件のツイート (オリジナルツイート) と、

表 1: 構築したソーシャルネットワークと抽出したコミュニティに関する統計量

ノード数	356,453
リンク数	29,535,582
最大弱連結成分に属するノードの数	351,870
弱連結成分の数	4,558
コミュニティ数	4,584
コミュニティの規模の最大値	77,695

それらに対する 63,462,387 件のリツイートを得た。これらのツイートから、ランダムに選択した 200 万件のオリジナルツイートと、それらに対するリツイートを以降の分析に用いた。

データセット中のソーシャルネットワークに対して、代表的なコミュニティ抽出アルゴリズムである Louvain 法 [Blondel 08] を適用することにより、各ユーザの属するコミュニティを得た。本データセットのソーシャルネットワークと抽出したコミュニティに関する統計量を表 1 に示す。

2.2 重回帰分析

重回帰分析により、Twitter ユーザのソーシャルネットワークにおけるコミュニティ構造がツイートの拡散規模に与える影響を分析する。オリジナルツイート t に対する k 回目のリツイートを $r_t(k)$ 、ツイート t に対する総リツイート数を N_t と表記する。また、ツイート t を投稿したユーザを $u(t)$ 、リツイート $r_t(k)$ を投稿したユーザを $u(r_t(k))$ と表記する。各リツイート $r_t(k)$ に対して、そのリツイートが投稿された後のリツイート数、すなわち $N_t - k$ を目的変数とする。説明変数として、オリジナルツイート t を投稿したユーザ $u(t)$ とリツイート $r_t(k)$ を投稿したユーザのコミュニティが異なるか否かをを用いる。また、他の説明変数として、ツイート t の本文の URL の有無、ハッシュタグの有無、単語数、ユーザ $u(t)$ のフォロワー数、リツイートを投稿したユーザ $u(r_t(k))$ のフォロワー数、ツイート t を投稿したユーザ $u(t)$ の属するコミュニティの規模、リツイート $r_t(k)$ を投稿したユーザのコミュニティの規模、リツイートの回数 k も用いる。重回帰分析に用いる変数を表 2 に示す。

$RTnum$ (リツイート $r_t(k)$ が投稿された後のリツイート数) を目的変数とし、表 2 に示す他の変数を説明変数とする重回帰分析を行った。リツイート数の分布は裾の長い分布であるため、ここでは負の二項分布回帰 (negative binomial regression) を用いた。回帰分析により求めた偏回帰係数を表 3 に示す。

表 3 より、他の要因の影響を考慮しても、コミュニティをまたがるツイートの拡散は、ツイートの拡散規模を増加させる影響を与えていることがわかる。構築した回帰モデルにおける各変数の与える影響の大きさは、偏回帰係数を β とすると e^β から推定される。表 3 の e^β の値より、コミュニティをまたがるツイート拡散は、コミュニティ内での拡散と比べて、その後のリツイート数を約 1.4 倍にする影響を与えると推定される。一方、変数 URL および hash の e^β の値はそれぞれ、約 4.2、約 1.05 である。これらのことは、コミュニティをまたがる拡散がツイートの拡散規模に与える影響の強さは、URL の有無やハッシュタグの有無が拡散規模に与える影響の強さと比較可能であることを示している。このことは、コミュニティをまたがるツイート拡散は、ツイートの拡散規模に影響を与える主要因の一つであることを示唆している。

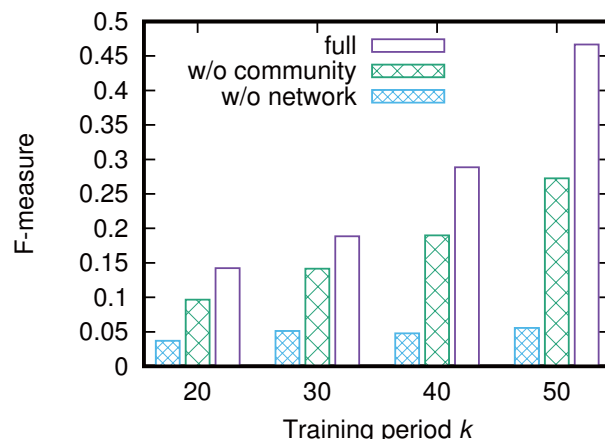


図 2: 用いる特徴量ごとのツイートの拡散規模予測精度の比較: ツイートが k 回リツイートされた時点でそのツイートの総リツイート数が 100 を超えるかどうかを予測した場合の F 値

2.3 拡散規模の予測

次に、コミュニティ構造に関する特徴量を用いて最終的なツイートの拡散規模を予測する実験を行う。これにより、コミュニティ構造に関する特徴量がツイートの拡散規模を予測するのにどの程度有用であるかを明らかにする。リツイート数そのものよりも、ツイートが大規模に拡散されるかどうかを予測することが実際には有用であると考え、ここでは、ツイートが k 回リツイートされた時点において、そのツイートの総リツイート数が 100 を超えるかどうかを予測するタスクを設定する。以降、この k を学習期間と呼ぶ。なお、リツイート数が 100 以上のツイートは、データセット中の全ツイートのうち 0.5% である。

前節で用いたデータを学習用データとし、ランダムフォレストによって予測モデルを構築する。表 2 に示す変数に加えて、リツイートを投稿した k 人のユーザのうち、ツイートの投稿者と同じコミュニティに属するユーザの割合 ($comrate$)、リツイートを投稿した k 人のユーザのフォロワー数の平均 ($avedeg$) も特徴量として用いる。これら全ての特徴量を用いた予測モデル (full)、コミュニティに関する特徴量 (com , $comsize_{orig}$, $comsize_{rt}$, $comrate$) を用いないモデル (w/o community)、ネットワークに関する特徴量 (deg_{orig} , deg_{rt} , $avedeg$ ならびにコミュニティに関する特徴量) を用いないモデル (w/o network) を構築し、用いる特徴量によって予測の精度がどの程度異なるかを調査する。ここでは、学習期間 k が 20、30、40、50 のそれぞれの場合の予測モデルを用いる特徴量ごとに構築した。

学習データに含まれないツイートをデータセットからランダムに 100 万件抽出し、それらのツイートとリツイートをテストデータとした。これらのテストデータに対して、ツイートが $k = 20, 30, 40, 50$ 回リツイートされた時点において、そのツイートの総リツイート数が 100 を超えるかどうかを予測した。図 2 に各モデルの予測精度を測る F 値を示す。

図 2 より、コミュニティに関する特徴量を用いたモデルの F 値は用いないモデルに比べて高いことがわかる。このことは、コミュニティの特徴量は投稿の拡散規模を予測する上で有用な特徴であることを示している。一方、モデルの予測精度自

表 2: 回帰分析に用いた変数

変数	説明
$RTnum$	リツイート $r_t(k)$ が投稿された後のツイート t のリツイート数
k	リツイート $r_t(k)$ が投稿されるまでのリツイート数 k
com	ツイート t とリツイート $r_t(k)$ を投稿したユーザのコミュニティが異なるかどうかを表すカテゴリ変数
$comsize_{orig}$	ツイート t を投稿したユーザの属するコミュニティに属するノード数
$comsize_{rt}$	リツイート $r_t(k)$ を投稿したユーザの属するコミュニティに属するノード数
deg_{orig}	ツイート t を投稿したユーザのフォロワー数
deg_{rt}	リツイート $r_t(k)$ を投稿したユーザのフォロワー数
URL	ツイート t が URL を含むかどうかを表すカテゴリ変数
$hash$	ツイート t がハッシュタグを含むかどうかを表すカテゴリ変数
$word$	ツイート t に含まれる単語数

表 3: 負の二項回帰分析により求めた各変数の偏回帰係数 (**: significant at 1% level)

目的変数: $RTnum$		
説明変数	Coeff. β	e^β
k^{**}	3.388e-03	1.003
com^{**}	3.528e-01	1.423
$comsize_{orig}^{**}$	2.301e-06	1.000
$comsize_{rt}^{**}$	8.180e-06	1.000
$\log(deg_{orig})^{**}$	6.473e-02	1.067
$\log(deg_{rt})^{**}$	-4.064e-02	0.960
URL^{**}	1.427e+00	4.166
$hash^{**}$	5.348e-02	1.055
$word^{**}$	2.163e-02	1.022
constant***	1.571e+00	

体はそれほど高くないこともわかる。実用上有用な予測を実現するためには、既存の拡散規模を予測するモデル [Cheng 14] にコミュニティに関する特徴量を組込むなどして、さらなる予測精度の改善が必要であると考えられる。

3. まとめと今後の課題

本稿では、ソーシャルネットワークのコミュニティ構造が、Twitter におけるツイートの拡散規模に与える影響を分析した。重回帰分析により、コミュニティをまたがるツイートの拡散は、最終的なツイートの拡散規模を増加させる影響を与えること、その影響の強さは、ツイート本文の URL やハッシュタグの有無と比較可能であることを示した。さらに、ツイートの拡散の過程において、最終的なツイートの拡散規模を予測する実験を行った。その結果、コミュニティに関する特徴量はツイートの拡散規模を予測する上で有用な特徴量であることを示した。

今後の課題として、他のコミュニティ抽出手法を用いた場合の分析が挙げられる。また、ソーシャルネットワークのコミュニティに関する特徴を用いて、ツイートの拡散の初期の段階において最終的な拡散規模を高精度に予測するモデルを構築することも重要な今後の課題である。

謝辞

本研究の一部は JSPS 科研費 17H01733 の支援を受けている。

参考文献

- [Blondel 08] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E.: Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2008, No. 10, p. P10008 (2008)
- [Cheng 14] Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., and Leskovec, J.: Can cascades be predicted?, in *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*, pp. 925–936 (2014)
- [Girvan 02] Girvan, M. and Newman, M. E. J.: Community structure in social and biological networks, *Proceedings of the National Academy of Sciences*, Vol. 99, No. 12, pp. 7821–7826 (2002)
- [Naveed 11] Naveed, N., Gottron, T., Kunegis, J., and Al-hadi, A.: Bad news travel fast: A content-based analysis of interestingness on Twitter, in *Proceedings of the ACM Web Science Conference 2011 (WebSci'11)*, pp. 1–7 (2011)
- [Newman 04] Newman, M. E. J. and Girvan, M.: Finding and evaluating community structure in networks, *Physical Review E*, Vol. 69, No. 2, p. 026113 (2004)
- [Onnela 07] Onnela, J. P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A. L.: Structure and tie strengths in mobile communication networks, *Proceedings of the National Academy of Sciences*, Vol. 104, No. 18, pp. 7332–7336 (2007)
- [Stieglitz 13] Stieglitz, S. and Dang-Xuan, L.: Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior, *Journal of Management Information Systems*, Vol. 29, No. 4, pp. 217–247 (2013)

- [Suh 10] Suh, B., Hong, L., Pirolli, P., and Chi, E.: Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network, in *Proceedings of the 2nd IEEE International Conference on Social Computing (SocialCom'10)*, pp. 177–184 (2010)
- [Tsugawa 15] Tsugawa, S. and Ohsaki, H.: Community Structure and Interaction Locality in Social Networks, *Journal of Information Processing*, Vol. 23, No. 4, pp. 402–410 (2015)
- [Weng 13] Weng, L., Menczer, F., and Ahn, Y.-Y.: Virality prediction and community structure in social networks, *Scientific Reports*, Vol. 3, p. 2522 (2013)
- [津川 16] 津川 翔: Twitter における情報拡散とコミュニティ構造の関係に関する一考察, 電子情報通信学会技術研究報告 (CQ2016-54), pp. 13–18 (2016)