マイクロブログにおけるコミュニティの抽出と分析

大澤昇平† 天笠俊之‡ 北川博之‡

†筑波大学第三学群情報学類 ‡筑波大学大学院システム情報工学研究科

1 はじめに

近年，国内のインターネット社会において，マイクロブログというサービスが数十万人規模で利用されている．国内では，Twitterやアメーバなうなどがその代表例である．マイクロブログは，ソーシャル・ネットワーキング・サイト(SNS)との多くの類似点が指摘されているが，形成される社会構造についてはほとんど明らかになっていない．

本研究では，国内最大手のマイクロブログであるTwitter上の，日本語圏ユーザで構成されるネットワークを対象に分析を行う．その結果として，これに通常の社会ネットワーク上で観測されるような，スケールフリー性や高クラスタ係数性があることを示す．次に，CNMアルゴリズムを用いてコミュニティ構造の抽出を行い，国内最大手のSNSであるmixiのネットワークについて先行研究[5]との比較を行う．最後に，これらの知見を元に，Twitterのネットワークのモデル化を行い，その形成メカニズムについて明らかにする．

2 関連研究

Java[1], Krishnamurthy[2]は，Twitterのクローリングを行い，統計を出している．この統計の中には，コミュニティ抽出に関するものも含まれているが，一部のコミュニティに関する考察に留まり，俯瞰的なコミュニティ構造に関する考察は行われていない．

湯田ら[5]は，国内最大手のSNSであるmixiのデータセットを用いて，コミュニティ構造の分析を行っている．本論文では，主にこの研究との比較を行う形で，Twitterのネットワークの特徴を明らかにしていく．

3 CNMアルゴリズム

社会ネットワークからのコミュニティ構造を抽出するためのアルゴリズムとしては，過去にクリーク・パーコレーション・メソッドやマルコフ・クラスタ・アルゴリズムなどが提案されているが，本研究ではmixiとの比較を行うため，湯田ら[5]が採用しているCNM(Clauset-Newman-Moore)アルゴリズム[4]を用いる．

CNMアルゴリズムは，与えられたグラフを，その部分グラフで構成される適切なクラスタに分割するアルゴリズムである．各クラスタは，社会ネットワークにおけるコミュニティに対応する．クラスタリングの適切さは，Newman[3]によって提案されているモジュラリティQを用いて評価される．モジュラリティQは，次式で定義される．

$$Q = \sum_i (e_{ii} - a_i^2)$$

ただし，e_{ij}はクラスタiからクラスタjに接続しているリンク密度(i.e. グラフ全体のエッジ数において占める割合)で，e_{ii}はクラスタ内のエッジ密度，$a_i \equiv \sum_j e_{ij}$はグループiから出ている全てのエッジの密度である．すなわち，どのクラスタにおいても内部のエッジ数が多く，外部のクラスタとのエッジの本数が少ないものが，適切なコミュニティ構造として評価される．

CNMアルゴリズムにおけるクラスタリングは，1ノードのみからなる原始的なクラスタ群から始まり，1回のステップごとにクラスタのペアを選び，結合していく形で行われる．このとき，クラスタのペアは，結合した際にモジュラリティが最も大きくなるように選ばれる．

4 実験

4.1 データセット

我々は，Twitterの提供するAPIにアクセスすることで，日本語圏ユーザをノード，それらのフォロー関係をエッジとするネットワークの取得を行った．日本語圏かどうかの判断は，発言の10%以上がひらがなを含んでいるかどうかで行っている．ネットワークの取得は，2009/12/1〜2009/12/31の期間中のパブリック・タイムライン上に出現した日本語圏ユーザを起点ノードとし，フォロー関係を辿って得られるすべての日本語圏ユーザの集合をデータセットとした．

本研究では，CNMアルゴリズムの適用を行うため，ネットワークの対称化を行う．対称化は，片方向にしか張られていないエッジを除去することで行った．すなわち，双方向にエッジが張られているもののみをエッジとして認めることにした．以上の手順で得られたグラフは，ノード数が475 506，エッジ数が7 276 351である．

Twitterの次数分布を図1に示す．分布が直線上に分布しており，スケールフリー性が確認できる．

4.2 実験結果

コミュニティの抽出結果を以下に示す．図2は可視化結果で図3はコミュニティ・サイズの分布である．全範囲においてベキ分布にはなっておらず，30人規模より大きいコミュニティは観察されにくい傾向がある．

Twitterのネットワークにおける統計量について，mixiのものとの比較を行う．統計量には，クラスタ係数C，ノード数対コミュニティ数比R_C，モジュラリティQを用いる．

Analizying community structure of microblog
Shohei OHSAWA†(shohei.ohsawa@kde.cs.tsukuba.ac.jp)
Toshiyuki AMAGASA‡(amagasa@cs.tsukuba.ac.jp)
Hiroyuki KITAGAWA‡(kitagawa@cs.tsukuba.ac.jp)
†College of Information Sciences, University of Tsukuba
‡Graduate School of Systems and Information Engineering, University of Tsukuba
Tennoudai , Tsukuba , Ibaraki, 305–8573 Japan

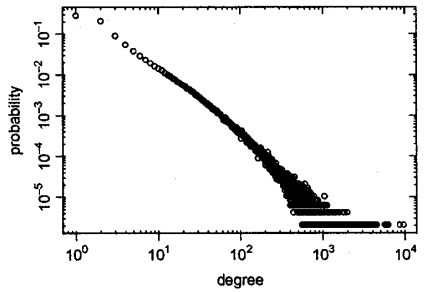

図 1: Twitter のネットワークの次数分布

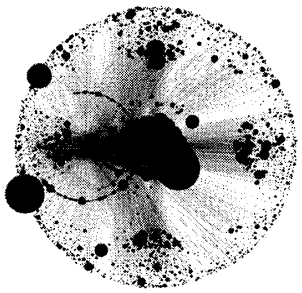

図 2: コミュニティ構造の可視化結果

4.3 CNNR モデル

既存研究では，mixi のネットワークを BA(Barabasi-Albert) モデル，WS(Watts-Strogatz) モデル，CNN(Connect Nearest Neighbor) モデルなどの単純なモデルと比較しており．スケールフリー性と高クラスタ係数性から，mixi のネットワークは CNN モデルに基づくものであると推測している (BA モデルはクラスタ係数が非常に低く，WS モデルはスケールフリー性がない)．

ただし，一方で，CNN モデルのコミュニティ分布は全範囲においてベキ分布を呈しているが，mixi のそれは逆 S 字型のグラフになっており，コミュニティが観測されにくいサイズの範囲 (ギャップ) が存在する．湯田らはこのギャップを説明可能にするため，CNN モデルの拡張である CNNR(Connect Nearest Neighbor with Random linkage) モデルを考案している．

今回の Twitter の実験結果においても，スケールフリー性と高クラスタ係数性に加えて，コミュニティ分布が完全なベキ分布になっていない特徴があり，CNNR モデルが適切なモデルであると考えることができる．[5] の図 6(e) に掲載されている CNNR モデルのコミュニティ分布と比較したところ，ランダムリンクの割合が 16% の場合とグラフの特徴がほとんど一致した．mixi の場合は 4%であるので，これは比較的高い値である．

4.4 考察

今回の実験から，Twitter のネットワークが，CNNR モデルによってモデリング可能であることを示した．これにより，Twitter のネットワークの形成メカニズムを，次のようにして説明することができると考えられる．

ノードの増加 新しいユーザ A が Twitter に登録し，知人をフォローする．

ニアレスト・ネイバーとの接続 A は，友人のプロフィール・ページやタイムライン上に出現する ReTweet の発言内容などから，自分がフォローしている友人の友人の存在を知り，フォローを行う．フォローを行われたユーザは，Twitter からの通知メールを受け取り，A のプロフィール・ページを閲覧し，フォローし返すか否かの意思決定を行う

ランダム・リンクの形成 さらに，検索機能やタイムライン上に出現するハッシュタグなどから，より幅広いユーザの存在を知り，フォローすることができる．

mixi との違いは，友人のコンテンツ上に他のユーザのプロフィール・ページが存在することで，これがランダム・リンクの割合を引き上げていると考えられる．また，紐帯の形成に合意のプロセスがなく，片方向のリンクが相互に接続することによって行われるので，これがクラスタ係数の低さに寄与していると考えられる．

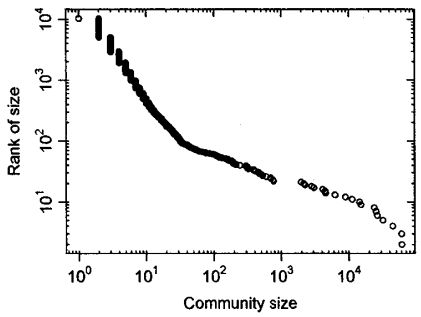

図 3: コミュニティ・サイズの分布

表 1: Twitter と mixi の特徴量の比較

	C	R_C	Q
Twitter	0.293	0.02186	0.149
mixi(湯田らによる)	0.330	0.01096	0.596

5 今後の課題

本研究では，CNM アルゴリズムを Twitter のネットワークに適用することで，コミュニティ構造の抽出を行った．その結果，mixi と同様に，コミュニティ分布が完全なベキ分布にならないことを示した．

今後は，ネットワークの経時変化に伴なうコミュニティの変化や，会話履歴を利用したネットワークを使った実験などを行う予定である．

参考文献

[1] A Java, X Song, T Finin, B Tseng: Why we twitter: understanding microblogging usage and communities, SNA-KDD 2007 workshop

[2] B Krishnamurthy, P Gill, M Arlitt: A few chirps about twitter, workshop on Online social networks

[3] M. E. J. Newman: Fast algorithm for detecting community structure in networks. Phys. Rev. E 69, 066133(2004)

[4] A. Clauset, M.E.J.Newman, and C. Moore: Finding community structure in very large networks. Phys. Rev. E 69, 066133(2004)

[5] K. Yuta, Naoki Ono, and Y. Fujiwara: A Gap in the Community-Size Distribution of a Large-Scale Social Networking Site, Arxiv preprint physics/0701168(2007)