Twitter における相互フォローネットワークと実際の友人関係の解析

理学専攻・情報科学コース　大塚 好恵

1　はじめに

近年、さまざまなソーシャルネットワーキングサービス (以下 SNS) が爆発的に普及している。SNS とは、人と人とのつながりを促進・支援する、コミュニティ型の Web サイト およびネットサービスのことを指す。

SNS を利用している人の中でも、実際の友人と SNS 上でも友人関係にある人は多い。つまり、SNS 上のネットワークを分析することにより、実際の友人関係も可視化・分析することができると考えた。本研究では、Twitter[1] に着目し、実際のユーザーのフォロー関係のネットワークを作成・ 解析することによって、コミュニティ推定を行い、実際の友人関係との比較・分析を行った。

2　コミュニティ推定とアルゴリズム

2.1　コミュニティ推定

コミュニティ推定とは、与えられたネットワーク内において、各頂点がどのコミュニティ（集団）に所属しているかを考えることを指す。コミュニティを推定することにより、似たような性質を持った頂点を分類することができる。

2.2　コミュニティのオーバーラップ

人間関係のネットワークにおいて、個人は学校、家族、部活、趣味、仕事など複数のコミュニティに所属している。このようなコミュニティの重なりをオーバーラップという。

本研究では、Ball, Karrer, Newman らによる、最尤推定法と EM アルゴリズムを組み合わせた手法 [1] を用いた。コミュニティ推定法のアルゴリズムでは、頂点を色分け（分類）していくものが多い。しかし本手法では、頂点よりも辺に着目し、辺のタイプを推定することにより、結果として頂点のコミュニティも推定することが出来るものとなっている。このアルゴリズムを実行すると、それぞれの辺について色 (タイプ) の確率が計算される。例えば、図 1(a) のようなネットワークがあるとする。実行すると、表 1 のように、各辺について色の確率が計算される。(この例ではコミュニティ数 $K = 2$ とする。) この結果に基づき、ネットワーク内のすべての辺の色を決定し、図 1(b) のようなグラフが得られる。

3　ネットワーク

3.1　データの取得

本研究では、Twitter の API を利用した。API とは、Twitter 社が提供しているサービスで、Web サイトやアプリケーションなどから Twitter の機能（データの取得、ツイートの投稿など）を呼び出すことができるものである。これを利用し、ユーザーのフォローしている/されているユーザーの一覧のデータを取得した。

表 1: 各辺の色の確率 (一部)

頂点 i	頂点 j	青の確率	ピンクの確率	結果
1	2	0	1	ピンク
1	3	0.37	0.63	ピンク
1	4	0.87	0.13	青
1	5	0.55	0.45	青

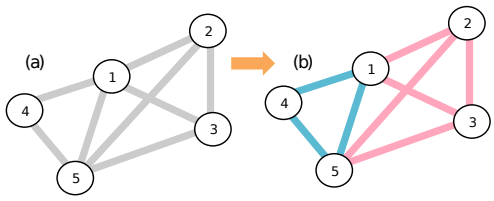


図 1: グラフの例

3.2　作成したネットワーク

まず、自分自身の Twitter のデータからネットワークを作成した。相互フォローのユーザーは 212 名であり、自分を含めて頂点数が 213 個、辺の数は 2,610 本となった。

さらに、それら 212 名のもつ相互フォローのユーザーのうち、ネットワーク内で 2 人以上と相互フォローしているユーザー 3,616 名を加えた、頂点数 3829 個、辺の数 14,490 本となる大きなネットワークも作成した。ここで、「ネットワーク内で 2 人以上と相互フォローしている」という条件をつけたのは、1 人のユーザーのみが相互フォローしている bot などの影響を避けるためである。

以下の表にネットワークの基本情報を示す。ここで、クラスタ係数とはネットワークに三角形が含まれる割合で、高いほど密につながっているネットワークといえる。距離とは、任意の点から任意の点までのステップ数のことで、平均距離はネットワーク内のすべての 2 頂点の距離の平均をとったものになる。

表 2: ネットワークの情報

	頂点数	辺の数	クラスタ係数	平均距離
小	213	2,610	0.715	1.884
大	3,829	14,490	0.607	3.541

本研究では小さい方のネットワークを対象として、以降の解析を行っていくものとする。

4　結果

本研究で使用したアルゴリズムでは、分割するコミュニティ数を指定して実行する。そこで、最適な分割数を見つけるべく、コミュニティ分割結果の良さを表す指標として知られているモジュラリティ Q という概念

[1] Twitter 社自身は、「社会的な要素を備えたコミュニケーションネットワーク」（通信網）であると規定しているが、ここでは「社会的ネットワークを構築できる」という点から、広い意味の SNS として考え、研究の対象としている。

を導入した [2]。

$$Q \equiv \frac{1}{2M} \sum_{c=1}^{N_{CM}} \left[\sum_{i,j=1; v_i, v_j \in CM_c}^{N} \left(A_{ij} - \frac{k_i k_j}{2M} \right) \right] \quad (1)$$

ここで、M はネットワーク内の全エッジ (辺) 数、N_{CM} はコミュニティ数、v_i, v_j は頂点、k_i, k_j は各頂点における次数、CM_c は c 番目のコミュニティの頂点の集合である。また、A_{ij} は隣接行列なので、頂点 v_i, v_j 間に辺があれば 1、なければ 0 をとる。

Q は 0 から 1 の間の値をとり、大きい方がいい分割である。コミュニティ数が増え、分かれすぎてしまうと Q が下がるようになっている。

コミュニティ数 2 ～ 9 でアルゴリズムをそれぞれ実行し、得られた log 最尤値は以下のグラフのようになった。コミュニティを分割すればするほど値が大きくなっていく結果になっている。

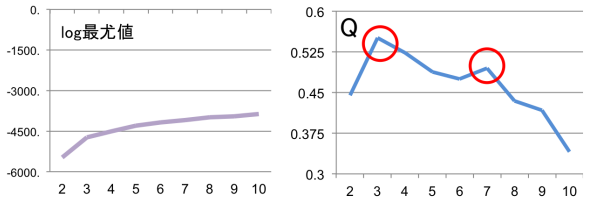
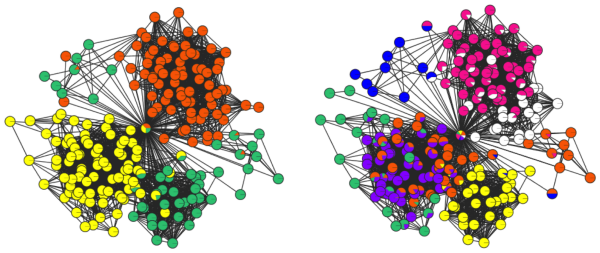


図 2: アルゴリズムを実行して得られた log 最尤値とモジュラリティ Q。

また、分割結果をもとに、それぞれのコミュニティ数について Q の計算も行った。Q に関しては 3,4,7 で高い値が得られた。この結果をもとに、コミュニティ数 3,7 の時のネットワークのグラフを描画した。頂点は、各頂点が持つ辺の色の割合を円グラフで表している。



(a) 3 で分割　　(b) 7 で分割

図 3: 分割結果

5　改良

分割結果をよく見てみると、不可解なところがいくつかあることに気がついた。このアルゴリズムでは、２人の関係性を色で示している。例えば、「○○高校の同級生」はピンク、「○○部のメンバー」はオレンジ、である。図 4(a) を見てみると、A と B のノードが青色の辺で結ばれていることが分かる。しかし、青色はこの結果では「私がしているアルバイトのメンバー」を示しており、もし A と B が同じアルバイトをしていたとしても、A と B は私と同じアルバイトをしているわけではないため、青色でつながれるのはおかしいはずである。もう 1 つの例は、図 4(b) である。ネットワーク内で次数が 2、つまり私ともう 1 人のみとつながっている頂点のもつ辺が、それぞれ異なるというものだ。つまり、私,C,D のすべての関係性が異なるという結果である。いずれの場合も、３頂点間の辺の色がすべて異なるという場合に起きているということが分かった。つまり、３人の友人の関係性がすべて異なるという状況である。実際の友人関係でもたまに存在するが、その場合共通の友人がいずれのコミュニティにも存在するはずである。また、今回の小さいネットワークでは３人の友人の関係性が異なるケースはないと考えられるため、エラーと考えられる。以上の点を踏まえ、このような場合のみ、辺の色を変更することを考えた。この例外処理を加え、さらに一部のコミュニティのみ抜き出して再度実行した結果が以下の図 5(a) である。

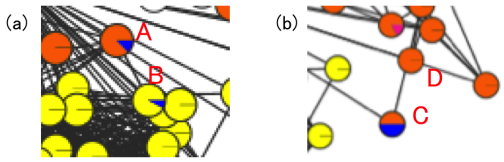
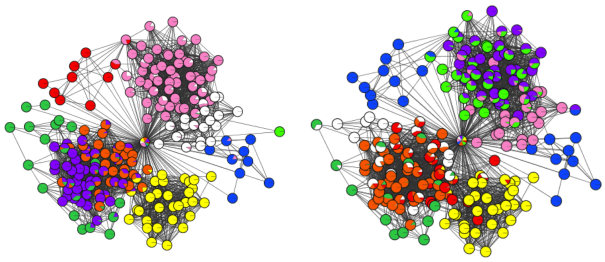


図 4: 問題のある箇所



(a) 例外処理と再分割を施した結果　　(b) アルゴリズムのみの結果

図 5: コミュニティ数 9 での分割結果の比較

このとき、最終的なコミュニティ数が 9 となったため、コミュニティ数 9 でアルゴリズムのみを実行した結果が図 5(b) である。このように、単純にアルゴリズムのみを実行しただけのものより、高精度で分割を行うことができた。

6　まとめ

Twitter におけるフォロー関係のネットワークを作成し、オーバーラップを考慮したコミュニティの分割を行った。結果に対する分析・考察より、例外的な場合の処理を追加し、単純にアルゴリズムを実行するだけよりも精度の高い分割結果を得ることができた。また、ここでは割愛するが、違うサンプルについても実行し、ネットワークの詳細に関わらず有効であることを示した。

参考文献

[1] B. Ball, B. Karrer and M.E.J.Newman, Phys. Rev. E **84**, 036103 (2011).

[2] 増田直樹・今野紀雄, 「複雑ネットワークー基礎から応用まで」, 近代科学社 (2010).