修士論文要旨 (2012 年度)

ソーシャルネットワークのノード次数に着目した局所的クラスタの抽出に関する研究

A Study for Finding Local Community Structures in Social Networks based on Degrees of Nodes

情報工学専攻　李　穎
LI　Ying

概要　ソーシャルネットワークとは，人間同士のつながりによる作られたネットワークである．それをいくつかの集合に分割すること（クラスタリング）は，ネットワークの構造を理解し，可視化する上で重要なことである．

本研究では，従来手法にとってネットワークが膨大になり計算量が大きくなるという欠点に対して，クラスタを効率よく抽出するアルゴリズムを提案する．ノード次数の大きさに基づき，局所探索アプローチを用いてクラスタを抽出するアルゴリズムを開発し，現実世界ネットワークのコミュニティ構造に近いことを示した．さらに，実装を行い，計算機実験によって既存手法と比較し，その性能を評価した．

キーワード：ソーシャルネットワーク，クラスタリング

1　研究背景

ソーシャルネットワーク (social network) とは，友人，提案，親類，嫌悪といった 1 つ以上の関係により結びつけられたノードからなる，社会的な構造である．ソーシャルネットワーク分析は，近年の社会学，人類学といった学問分野において，ポピュラーな推論・研究であると同時に，有用な方法として台頭した．この分析によって，それが家族から国家まで様々なレベルの問題に適用でき，問題解決への道を示す重要な役割を果たす．

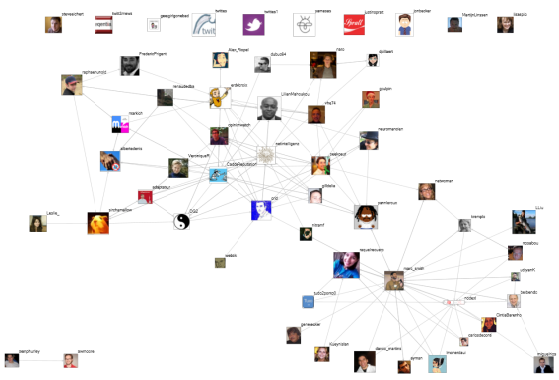

図 1　ソーシャルネットワーク

近年では，ソーシャルネットワーク分析は多くのネットワークに応用されてきた．例えば，www(world-wide web) ネットワーク [1]，疫学 [2]，科学者の共著ネットワーク [8] などであり， ネットワークという視点からデータを分析するアプローチが様々な分野で注目を浴びている．

一般的なソーシャルネットワークは,「ノード (nodes)」と「つながり (edges)」という観点から社会的隣接性を考察する．つながりとは，関係者間の結びつきをあらわすものである．その関係の密な部分 (“コミュニティ” あるいは “クラスタ”) 分析に関する研究分野では，コミュニティをより効率よくあるいは精度よく抽出するについて，多くの研究が行われてきた．

2　既存手法

2.1　GN 法

GN 法は 2004 年に Girvan と Newman によって提案されたアルゴリズムである [3，4]．GN 法では，分割型の手法を採用している．分割型の手法は過去にも研究されてきたが，従来の手法は，ネットワーク内の接続されたノードの対のうち最も類似性の低いものを見つけ出し，そのノード対を接続するエッジを取り除くという操作を繰り返し行うというものであった．これに対し，GN 法のアプローチは，最短経路に基づき最も高い **betweenness** を持つエッジに注目し，そのエッジを取り除くというものである．

betweenness とは，コミュニティ間のエッジは高い値となり，コミュニティ内のエッジは低い値となる指標値である．「もし二つのクラスタが共有するエッジが数本しか存在しないとすれば，一方のクラスタ内のノード から他方のクラスタ内のノードへ到達するためには必ずその数本のうちの一つを通らなければならない．そのため，適当な経路の集合を与えたときに通る回数の最も多いエッジは，コミュニティ間をつなぐエッジであると予想される」というものである．Girvan と Newman は計算量や実行結果から考えて，最短経路に

基づいた betweenness が最も有効であると結論づけている．

問題定義: ソーシャルネットワーク $G = (V;E)$(V:頂点集合，E:辺集合) とする．また，ネットワーク G に含まれるすべての頂点対の集合を $U(U = E * E)$，u の要素のうち最短パスが辺 $e \in E$ を通る頂点対の集合を U_e としたときの e の最短経路 *betweenness* B_e を次のように定義する．

$$B_e = \frac{|U_e|}{|U|}$$

GN 法

1. すべての辺 $e \in E$ に対し，最短経路 betweenness B_e を算出する．
2. B_e の最も高い辺を E から削除する．
3. 各連結成分をコミュニティとして出力する．
4. U，U_e を再び計算する．
5. すべての辺を削除するまで 1 に戻る．

これにより，トップダウンで樹形図が形成する．樹形図をどこで切り分ければ一番良いクラスタリング結果を得るかを考察するために，クラスタ構造を評価する指標 **modularity** を導入した．

modularity とは，クラスタのモジュール性を評価する指標である．Newman はコミュニティ分割の良さを「コミュニティ内とコミュニティ間の辺の割合」とし，この値を Modularity と呼ぶ．全ネットワークのコミュニティの集合を L，コミュニティ $i \in L$ 内部のノード同士をつなぐエッジ数の割合を e_{ii}，ネットワーク全体の辺の本数に対するコミュニティ $i \in L$ からコミュニティ $j \in L$ につながっている辺の本数の割合を e_{ij}，ネットワーク全体の辺の本数に対するコミュニティ i から出ている辺の割合を a_i としたとき，**Modularity** Q は次のように定義される．

$$Q = \sum_{i \in L}(e_{ii} - (\sum_{j \neq i} e_{ij})^2) = \sum_{i \in L}(e_{ii} - a_i^2)$$

この Q によって，クラスタ内でのつながりの強さを表すことができ，この値が大きければ良いクラスタリング結果であると言える．クラスタ数が 1 のときに $Q = 1$ で最大である．コミュニティ構造が変化する度に Q を計算し，Q が最も大きくなったときのコミュニティ構造を出力する．

頂点数を n，辺数を m としたとき，最短経路 betweenness の算出に $O(mn)$ かかる．すべての辺を削除するまでコミュニティを分割するため，GN 法の最悪計算量は $O((mn)(m)) = O(mn^2)$，ネットワークの全頂点の最大次数を d とした時，$m = O(dn)$ であるため，GN 法の最悪計算量は $O(dn^3)$ である．

2.2 Newman 法

Newman 法とは，2004 年に Newman が提案した，GN 法をより効率よく処理できるように改良したアルゴリズムである [5]．GN 法は最短経路 betweenness の計算が非常に効率が悪く，ネットワークに大きくに伴い計算量が増加することである．Newman は Q を高くすることが目的なら，Q が最も増加するようにコミュニティ同士を結合していけばよい，という考え方に従って抽出法を提案した．Modularity の増加量 ΔQ は次のように定義される．

$$\Delta Q = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j)$$

Newman 法

1. 全クラスタに対して，2 つのクラスタを結合した場合の Q の増減 (ΔQ) を計算する．
2. Q の増加を最も大きくする，2 つのクラスタを結合する．
3. 1，2 を繰り返す．Q が最大になった時点のクラスタリング結果を返す．

これにより，ボトムアップで樹形図が形成する．Q が最大になった時点に対応した部分から切り分け，クラスタリング結果を示す．Newman 法は ΔQ の算出と結合するクラスタ対の決定の際の最悪の計算量は $O(m+n)$ であるため，アルゴリズム全体で必要な最悪の計算量は $O((m+n)n) = O(dn^2)$ である．

3 提案手法

これまでの既存手法に対して，GN 法は現実的な計算量でないこと，Newman 法は GN 法より少しクラスタ精度が下回るといった問題点があった．そこで本研究ではネットワーク全体のトポロジーを必要としないという特徴から，各ノード次数の大きさにより局所探索アプローチを用いて，局所的クラスタの抽出を行う手法の提案を行う．この手法により，*GN* 法など大局的な手法より計算時間が小さくなることが利点として考えられる．

基本的な考え方：ソーシャルネットワークの性質:スケールフリー性から発想し，ソーシャルネットワーク中にごく少数のノードが膨大な次数（エッジ）を持つ一方，大多数のノードはごく少数の次数を持つという性質があるため，ソーシャルネットワークには，クラスタ内次数高い点の存在が一般である．それらの点として，周りの点に影響力を持つと仮定し，次数が小さい点を自分の所属クラスタに参加させるという考え方である．

3.1 提案手法の定義

提案手法：アルゴリズム

1. 全ノードに対して、それぞれの次数から deg 値を付く.
2. 任意隣接ノード s と t に対して,
   • s の次数は t の次数より大きい，かつ t 周りの隣接点の次数よりも大きいならば，$t \in \{s\}$
   • t の次数は s の次数より大きい，かつ s 周りの隣接点の次数よりも大きいならば，$s \in \{t\}$
3. 以上の条件を満たさなければ，それぞれ二つクラスタに所属と仮定する.
4. クラスタは隣接点を結合できなくなるまで，2, 3 を繰り返す.

これにより，1 つコミュニティ(クラスタ) を生成する．残り点から同じ方法で繰り返し，全ネットワークのクラスタリング結果を示す．提案手法は全ノード次数の算出 $O(m)$ かかり，結合するクラスタ対の決定の算出も $O(m)$ かかりため，最悪の計算量は $O(m)+O(m) = O(m)$ である．ネットワークの全頂点の最大次数を d とした時，$m = O(dn)$ であるため，$O(dn)$ である．GN 法と Newman 法の最悪計算量に対し， 提案手法はより効率よくコミュニティを探索することができる.

3.2 提案手法の改良

実際のネットワーク中に，幾つ次数が大きい点に隣接する場合は多く．提案手法によりクラスタリングを行い際に，それらの点が同じコミュニティ所属か，又別々コミュニティ所属かを判断できなくなってしまう．すべて同じコミュニティ所属を認定されてしまう．それらを解決するために，新たな制約条件を付き加ければならない.

改良のポイント：一般的に，同じコミュニティ内の点の間は密であり，同じ隣接点も多くという特徴があるため，次数が大きい点 (hub) にお互いつないでいる際に，同じ隣接ノード数の割合によって同じコミュニティ所属かどうかを判断できると推測される．提案手法の制約条件はこの割合の判断を付き加え，割合の閾値が本文で (0.4) を設定し，それ以上ならば結合する.

同様にして，クラスタ間にブリッジ点に対して，改良手法を用いて彼らの隣接ノードから所属コミュニティが区別できるようになる.

4 計算機実験

本研究で使用した社会ネットワークについて述べ，Intel(R) Core @ 2.66GHz 2.67GHz，8.00GB メモリ実験環境で，既存手法と提案手法に対する，其々クラスタリング結果のクラスタ係数，密度，及び計算量 (10 回の平均値) によって実験的評価について述べる.

4.1 Zachary karate club network

図 2 に示している，1970 年代のある米国大学における空手クラブの 34 人のメンバーの交友関係を示すソーシャルネットワークである [6].

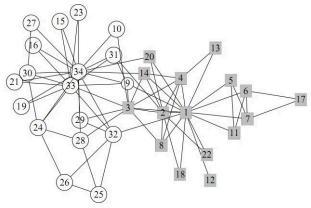

図 2　Zachary karate club network

図 2 のように，ネットワークは 2 つ大きいコミュニティが存在する．GN 法と Newman 法は 1 つのみ異なったに対して，提案手法は 2 つ異なった，提案手法の改良はすべて正し分割りができた．他のクラスタリング結果は評価以下の表 1 に示す.

表 1 Zachary karate club network についての性能評価

評価 \ method	Zachary	GN	Newman	提案手法	提案の改良
クラスタ係数	0.5706	0.6294	0.6651	0.6411	0.6856
密度	0.1390	0.2503	0.2500	0.2457	0.2519
計算量		0.2786s	0.2521s	0.2300s	0.3260s

提案手法は，クラスタ係数，密度，及び計算量についてよい結果を得られないことがわかった.

4.2 Dolphin social network

ニュージーランドで，ダウトフルとサウンドを用いて共同生活している 62 頭のイルカ社会的ネットワークである [7].

イルカネットワークも 2 つ大きいコミュニティの存在があり，各ノード間の繋がりが平均であることがわかった.

GN 法と Newman 法は 4 組を分割できた，提案手法，提案手法の改良とも 3 つ大きい組と単独ノードを分割した，他のクラスタリング結果の評価は以下の表 2 に示す．しかしながら，Zachary ネットワークと同じく，提案手法はクラスタ係数，密度，及び計算量についてよい結果を得られないことがわかった.

表 2　Dolphin social network についての性能評価

評価 \ method	Dolphin	GN	Newman	提案手法	提案の改良
クラスタ係数	0.2480	0.4364	0.2709	0.1559	0.1311
密度	0.0841	0.3896	0.4237	0.2428	0.2810
計算量		0.3274s	0.2806s	0.2453s	0.3823s

4.3　football network

2000 年秋のレギュラー大会にアメリカ大学のサッカーチーム間のスケジュールを反映したネットワークである [3]．全ネットワークは 62 チームがあり，115 回の対戦情報がある．

GN 法は 12 組を分割できた．Newman 法は 6 組を分割できた．提案手法は 1 つ大きい組みと分散のノード集合が分割したが，提案手法の改良はオリジナルのネットワークと近く 10 組を分割できた，他のクラスタリング結果の評価は以下の表 3 に示す．提案手法はラスタ係数と密度については，それほど良い結果を得ることができなかったが，提案手法の改良では GN 法と近く，かつ小さい計算量で済むことがわかった．

表 3　football network についての性能評価

評価 \ method	football	GN	Newman	提案手法	提案の改良
クラスタ係数	0.4030	**0.8531**	0.6490	0.2299	**0.8151**
密度	0.0931	**0.7804**	0.4807	0.1948	**0.7294**
計算量		3.8277s	0.6214s	0.3434s	**0.5526s**

4.4　Collaborations Between Network Scientists

2006 年 5 月に Newman によって提案されたアルゴリズムを利用した SNS データであり，実験に取り組んでいた科学者の共著ネットワークである [8]．全ネットワークは 1589 名の科学者であり，科学者達の共著論文が 2872 件である．

GN 法，Newman 法とも 274 組を分割できた．提案手法は 335 組を分割した，提案手法の改良は 318 組を分割した．他のクラスタリング結果の評価は以下の表 4 に示す．提案手法はラスタ係数と密度についてはよい結果を得ることができなかったが，提案手法の改良では GN 法と近く，かつ小さい計算量で済むことがわかった．

表 4　Scientists network についての性能評価

評価 \ method	scientists	GN	Newman	提案手法	提案の改良
クラスタ係数	0.6151	**0.4828**	**0.4795**	0.4293	**0.4722**
密度	0.0023	**0.7975**	**0.8046**	0.6921	**0.7794**
計算量		**473.244s**	55.2388s	28.9458s	**39.7718s**

5　結論

本研究ではソーシャルネットワークにおけるクラスタリング問題に対する局所探索手法を扱った．既存のアルゴリズムによって問題を定義し，既存手法の欠点を緩和するために提案手法を提案した．計算機実験を行ってそれぞれの性能を評価した．4 つの実験環境中に，スケールフリー性が強い社会ネットワークに対して，提案手法によりクラスタリングの効果的であることが分かった．しかも，全体の傾向からみると，ネットワークが膨大になり提案手法の方は計算量が小さくなることが見て取れる．

今後の課題としては，Dolphin ネットワークのようなスケールフリー性が弱いソーシャルネットワークに対しても応用できる提案手法，更に改良された提案の実験が挙げられる．

謝辞

本研究を進めるにあたり長い間，熱心なご指導，適切なご指摘を頂いた浅野孝夫教授に心から感謝いたします．また，非常に多くの助言を頂き，様々なご指導を下さった先輩方に深く感謝いたします．そして，研究室で苦楽を共にしてきた同輩，後輩に深く感謝いたします．

参考文献

[1] R.Albert, H.Jeong,and A.-L.Barabasi,Diameter of the world-wide web.Nature 401,130-131 (1999).

[2] C.Moore and M.E.J.Newman, Epidemics and percolation in small-world networks. Phys. Rev. E 61, 5678-5682 (2000).

[3] M.Girvan and M.E.J.Newman, Community structure in social and biological networks. Proc. Natl. Acad. Sci.USA 99, 7821-7826 (2002).

[4] M.Girvan and M.E.J.Newman, Finding and evaluating community structure in networks. Preprintcond-mat/0308217 (2003).

[5] M.E.J.Newman, Fast algorithm for detecting community structure in networks. Phy. Rev. E69, 066133 (2004).

[6] W.W.Zachary,An information flow model for conflict and fission in small groups.Journal of Anthropological Research 33, 452-473 (1977).

[7] Lusseau D, Schneider K, Boisseau O J, Haase P, Slooten E, Dawson SM 2003 Behav. Ecol. Sociobiol. 54 396

[8] M.E.J.Newman, Finding community structure in networks using the eigenvectors of matrices,Preprint physics/0605087 (2006).