7A-01

ネットワーク内の統計的に有意なコミュニティの抽出

水野貴之† 伊藤亮人# 新井優太‡ 秋葉拓哉† 家富洋#
国立情報学研究所，総合研究大学院大学複合科学研究科，JST さきがけ†
新潟大学大学院自然科学研究科#
リクルート住まい研究所‡

1 はじめに

2000 年以降．現実世界の様々な現象を説明する新たなパラダイムとして複雑ネットワークが注目を集めている[1]．複雑ネットワークでは，システムに所属する多数の因子をノード，それらの因子間の相互作用をエッジとして，ネットワークを用いてシステム全体を記述する．例えば，ある人間の集団をネットワークで記述するときには，ノードを個人，エッジを個人間の友人関係とする．

現実世界の様々なシステムのネットワーク内には，密な部分グラフであるコミュニティが存在する．人間の集団の例では，しばしばコミュニティは友人グループに対応している．ネットワークには，通常，複数のコミュニティが階層的に存在する．このような，コミュニティを抽出するアルゴリズムとして，Guirvan-Newman 法，Fast Greedy 法，Fast unfolding 法などがある[2-4]．これらの方法では，モジュラリティQが最大になるようにネットワークがコミュニティに分割され，全てのノードは，どれかのコミュニティに必ず属する．

$$Q = \sum_i (e_{ii} - a_i^2) \qquad (1)$$

ここで，e_{ii}はコミュニティiのエッジ密度，a_i^2はランダムグラフにおけるコミュニティiのエッジ密度である．

モジュラリティ最大化によるコミュニティの抽出は，比較的，人間の手による分類結果に近いことから，よく使われるが，いくつか改善が必要な点もある．例えば，ランダムネットワークでも，コミュニティが抽出される．本来，ランダムネットワークでは，統計的に有意なコミュニティは存在しないために，何か統計的な基準にもとづいて．各コミュニティの有意性を判断しなければいけないが，現在，統計的に明確な判断基準が存在しない．従って，現在は経験的に，ランダムネットワークでは最大モジュラリティは比較的小さくなる傾向があるために，最大モジュラリティがある閾値の以下では，ネットワーク内の全てのコミュニティに意味は無いと判断している．しかし，このような判断基準では，コミュニティのあるネットワークが，ランダム性の高いネットワークに混ざっている場合，つまり，ネットワーク内のモジュラリティに偏りがある場合に，ネットワーク全体で最大モジュラリティを算出し，その値にもとづいて，仮にある閾値より大きければ，統計的に意味のないコミュニティを拾ってしまうし，仮にある閾値より小さければ，意味のあるコミュニティを見逃してしまう．

本論文では，Guirvan-Newman 法におけるネットワーク分割の有意性を統計的に判断することにより，統計的に有意なコミュニティのみを見つける方法を提案する．

2 手法

我々は，ネットワーク内の連結成分ごとに，Girvan-Newman アルゴリズムを走らせる．Girvan-Newman アルゴリズムは，対象ネットワーク内の最大 Edge Betweenness Centrality (EBC) を持つエッジを見つけて，切断することを繰り返す．エッジeの EBC は次式で定義される．

$$g(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}} \qquad (2)$$

ここで，σ_{st}はノードsとtの間の最短経路の数を，$\sigma_{st}(e)$はエッジeを通るノードsとtの間の最短経路の数を表す．

連結成分に対して最大 EBC を持つエッジの切断を繰り返すと，ある時，２つの連結成分に別れる．ノード数n，エッジ数mの連結ランダムグラフ CRG(n,m)が，２つの連結成分に別れたときの，ノード数が少ない方の連結成分のノード数kに注目する．CRG では，エッジ数mがノード数nに比べて十分に大きいとき，別れたノードの数kの平均はわずか 1 ノードであり，その標準偏差もほとんど０である．一方で，コミュニティ構

Significant community detection in networks
† T. Mizuno, T. Akiba・National Institute of Informatics
# A. Ito, H. Iyetomi・Niigata University
‡ Y. Arai・Researcher at Recruit Sumai Co.,Ltd

造を持つ連結グラフでは，別れたノードの数kが有意に１よりも大きいことが数値的に確認できる．従って，我々は，Girvan-Newman アルゴリズムによって，ノード数n，エッジ数mの連結成分から得られた別れたノードの数kが，CRG(n,m)から統計的に得られる別れたノードの数の期待値$\overline{k_{CRG}}$＋２倍の標準偏差$2\sigma(k_{CRG})$よりも大きければ，その連結成分の分割は統計的に有意と判定する．

この統計的な有意判定を，分割して生まれる連結成分も含めて全ての連結成分に対して，有意な分割が存在しなくなるまで逐次的に繰り返す．そして，有意な分割によって作られた連結成分を有意なコミュニティと，有意ではない分割によって作られた連結成分を有意ではないコミュニティと定義する．

３　結果

図１はノード①から⑮までで形成された３つのコミュニティを持つネットワークである．これを，ノード①から㊾までの 50 ノードと 100 エッジの連結ランダムグラフに混ぜ込む．

図２は，混ぜ込んだネットワークに対して，Girvan-Newman アルゴリズムとモジュラリティ最大化により，コミュニティを抽出した結果を表す．最大モジュラリティは 0.29 であった．ノード①から⑮までで形成されたコミュニティは比較的綺麗に抽出ができてはいるが，統計的には意味のないコミュニティ，例えば，ノード㊶が属する９ノードのコミュニティも同時に抽出されてしまっている．

一方，図３は本論文で提案する手法を用いて，図２と同じ混ぜ込んだネットワークから統計的に有意なコミュニティを抽出した結果である．ノード①から⑮までで形成されたコミュニティをより綺麗に抽出できている．しかも，それら以外の有意なコミュニティは存在しないという結果を示す．

４　まとめ

本論文では，連結ラダムグラフに対する Guirvan-Newman アルゴリズムによるネットワーク分割の統計的な特徴を用いて，ネットワークから統計的に有意なコミュニティを抽出する手法を提案した．本手法は，ランダムノイズを含んだネットワークからのコミュニティ抽出に力を発揮する．

謝辞

本研究は JSPS 科研費 15KT0052 の助成を受けたものである．

参考文献

[1] Mizuno T., et. al. (2014) PLOS ONE 9, e100712.
[2] Girvan M. and Newman M. E. J. (2002) PNAS 99, 781-7826.
[3] Clauset A., et. al. (2004) Phys. Rev. E 70, 066111.
[4] Blondel V. D., et. al. (2008) J. Stat. Mech., P10008.

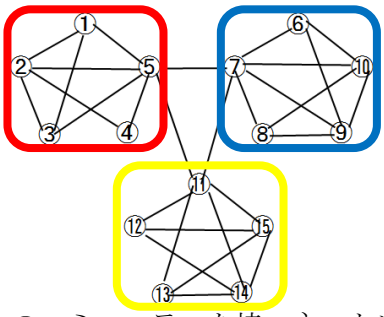

図１　３つのコミュニティを持つネットワーク

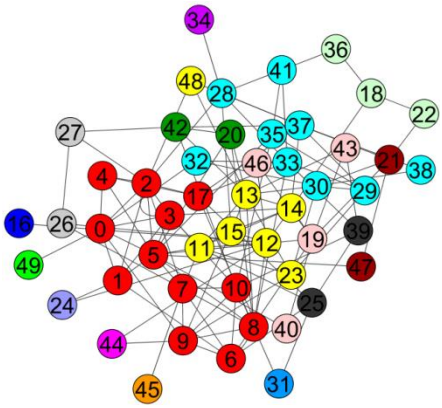

図２　図１のネットワークを連結ランダムグラフに混ぜ込んで Girvan-Newman 法によりコミュニティを抽出した結果

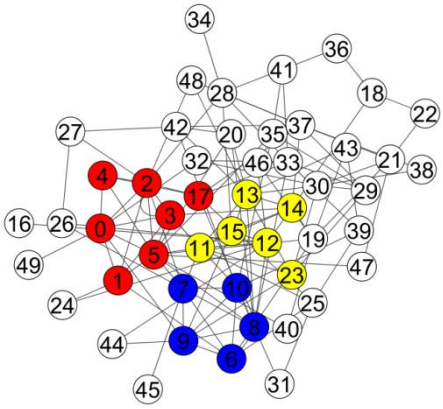

図３　図１のネットワークを連結ランダムグラフに混ぜ込んで本提案手法により統計的に有意なコミュニティを抽出した結果