A_019

混合のエントロピーを利用したコミュニティ検出アルゴリズム

Community Detection Algorithm by Using Entropy of Mixing

大久保　潤*　　田中　和之†
Jun Ohkubo　　Kazuyuki Tanaka

1　はじめに

情報処理を始めとして，人間関係や生物・化学の分野に至るさまざまな現象を一般的に理解するための理論的戦略のひとつとして“複雑ネットワーク”という概念が提案された[1, 2]．これまで，スケールフリー性・スモールワールド性の発現機構の解明や，ネットワーク上のウィルスの感染問題など，さまざまな研究が行われている．

複雑ネットワーク概念を情報処理へ応用する試みのひとつに，コミュニティ検出が挙げられる．コミュニティ検出とは，一言で言えば，データとして与えられたネットワークをいくつかのコミュニティに分割する問題である．たとえば，World Wide Web (WWW) をグラフ（ネットワーク）として表現したとする．この場合，各頂点は各ページを表し，辺はページ間のハイパーリンクを意味する．この WWW ネットワークには，ビジネスに関するページや Linux に関するページなど，さまざまなものが混在しているであろう．また，Linux の情報に関するページ同士は互いにリンクしあっている場合が多いと考えられる．そこで，頂点同士をつなぐ辺の情報（隣接行列）を元にして，Linux に関するページをコミュニティとして検出するデータマイニングの問題を考えることができる．

本稿では，熱力学的エントロピーとの類似性を利用したコミュニティ検出手法の概念を提案する．従来提案されているアルゴリズム（参考文献 [3] など）の多くは，最短経路の計算などを必要とする．したがって頂点数 N，辺の数が M であるネットワークに対するコミュニティ検出に $O(M^2N)$ の計算量を必要とする．本稿で提案する概念により，$O((M+N)N)$ の計算量をもつアルゴリズムを定式化することが可能である．さらに，提案した概念によって，コミュニティ検出手法をある種の最適化問題に置き換えることができる．コミュニティ検出の問題をある種の最適化問題として取り扱う枠組みは Newman[4] によって提案されている．Newman の手法においてはコミュニティ内の結合情報とコミュニティ間の結合情報の両方が必要とされるが，本稿で提案する手法はコミュニティ内の結合情報のみを必要とするという違いがあり，理論の枠組みとして本提案手法の方が簡潔なものとなっている．

本稿の構成は以下の通りである．2 節において基本となる概念を提示する．3 節では実際のネットワークに対して提案した概念を適用し，その検出結果に関して述べる．最後に，4 節において結論を述べた後，本稿の問題点とその改善策に関して触れる．

2　検出手法の概念

事象の乱雑さを指定する指標として熱力学におけるエントロピーがある．この熱力学的なエントロピーのアナロジーとして，ネットワークにおける“混合のエントロピー”を定義し，これを指標として採用する．このエントロピーを最小とするようにコミュニティを検出していけば，秩序だった頂点の集合，すなわちコミュニティが検出されることが期待される．

まず，熱力学的な定式化を行うために体積を定義する．ネットワークにおいて，コミュニティ i を構成している頂点の個数を n_i とし，V_i をコミュニティ i の体積とする．頂点数に対する辺の数の比から，コミュニティ i の密度 ρ_i を定義する．

$$\rho_i = \frac{e_i}{{}_{n_i}C_2} \tag{1}$$

ここで，e_i はコミュニティ i 内に存在する辺の数である．さらにこの密度を用いてコミュニティの体積を以下のように定義する．

$$V_i = \frac{n_i}{\rho_i} = n_i \times \frac{{}_{n_i}C_2}{e_i} \tag{2}$$

なお，ここではひとつの頂点から構成されるコミュニティの体積は 1 であると仮定する．

熱力学との類似性から，コミュニティ 1 と 2 が結合してコミュニティ 3 ができる際の混合のエントロピーを以下のように仮定する．

$$\Delta S = n_1 \ln \frac{V_3}{V_1} + n_2 \ln \frac{V_3}{V_2} \tag{3}$$

なお，e_{12} をコミュニティ 1 とコミュニティ 2 の間に存在する辺の数であるとすれば，式 (2) より体積 V_3 は

$$V_3 = (n_1 + n_2) \times \frac{{}_{(n_1+n_2)}C_2}{e_1 + e_2 + e_{12}} \tag{4}$$

となる．この混合のエントロピーが小さいほど体系が秩序だっていると解釈し，式 (3) の ΔS を最小にするようなコミュニティの結合操作を採用する．

* 東北大学大学院情報科学研究科，日本学術振興会特別研究員 DC
† 東北大学大学院情報科学研究科，E-mail: kazu@smapip.is.tohoku.ac.jp

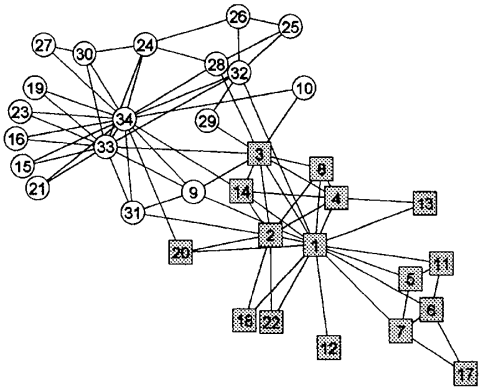

Fig. 1 karate club ネットワーク.

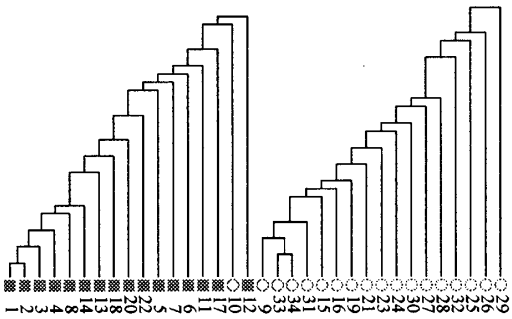

Fig. 2 karate club ネットワークに対して提案手法を適用した場合の検出結果の樹形図.

以上の議論については，熱力学的な正当性等に議論の余地がある．体積の定義により，これらの定式化は非可算的 (non-extensive) な性質を持つため，通常の熱力学とは異なる体系を持つ．完全グラフを取り扱う場合にのみ，可算的 (extensive) な性質をもち，通常の熱力学と同様の枠組みが成立することを注意しておく．

3 karate club ネットワークにおけるコミュニティ検出結果

本節では 2 節で提案された概念を実際のネットワークに対して適用し，数値実験を行う．なお，具体的なアルゴリズムは，Newman[4] によって用いられている手法とほぼ同様であり，集積的な (agglomerative) アルゴリズムである．ここでは参考文献 [5] で与えられている karate club ネットワークを例として使用する．karate club ネットワークとは，あるアメリカの空手クラブ内の人間関係を表したネットワークである．各頂点がクラブの構成員を表し，辺は人間（友人）関係を表している．図 1 にその具体的なネットワークを与える．このネットワークはもともとひとつのクラブであったが，ある時期に空手クラブの管理者と空手の先生との間で対立が起こり，クラブが二分化された．白丸で表した頂点と灰色の四角で表した頂点が，それぞれ二分化されたあとのコミュニティに属する構成員を示している．

問題は，分裂する前の人間関係（すなわち図 1 のネットワーク）からそれぞれのコミュニティを検出することができるか，ということである．事前の人間関係が既知であり，さらに分裂の仕方までわかっているため，この空手クラブのネットワークはコミュニティ検出の例題としてよく使用される．

このネットワークに対して集積的なアルゴリズムを適用した結果を樹形図として図 2 に示す．頂点 10 を除いて，正確にコミュニティを検出できていることがわかる．

4 まとめ

本稿では，ネットワークからコミュニティ構造を検出するためのアルゴリズムを熱力学的立場から提案し，karate club ネットワークに対する数値実験を通して提案アルゴリズムが良好に機能することを示した．なお，コミュニティ構造をもつネットワークを人工的に生成して数値実験を行った場合についても，提案アルゴリズムが十分良好に機能することを確認したことを補記しておく．

エントロピーの概念を用いてコミュニティ検出を行う試みは本研究が初めてであり，提案手法を拡張することにより高速かつ高性能のコミュニティ検出手法の開発へと発展することが期待される．しかしながら，本稿の提案手法における熱力学的な定式化の妥当性については議論の余地が残されており，これは今後の課題である．なお，体積の非可算性に着目することによってこの問題を避けることが可能であると考えられる．すなわち，本質的には体積の非可算性が重要であるため，熱力学とのアナロジー（混合のエントロピーの仮定等）を必要とせずに，定義された体積を最小にする最適化問題としてコミュニティ検出手法を定式化する方向性がある．これら結果については講演時に触れることにする．

謝辞

本研究の一部は文部科学省科学研究費補助金 (No.14084101, No.14084203, No.17500134, No.18·5140) の補助を得て行われたものである．

参考文献

[1] 『特集：「ネットワークが創発する知能」』，人工知能学会誌, Vol. 20, No.3 (2005).
[2] 『特集：ネットワーク生態学』，情報処理学会論文誌, Vol. 47, No. 3 (2006).
[3] M. Girvan and M.E.J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).
[4] M.E.J. Newman, Phys. Rev. E **69**, 066133 (2004).
[5] W.W. Zachary, J. Anthropol. Res. **33**, 452 (1977).