2J4-OS-04a-4

ネットワークからのコミュニティ階層構造の効果的かつ安定な検出

Effective and Robust Method for Hierarchical Community Detection

邱 シュウレ[*1]
Xule Qiu

岡本 洋[*1]
Hiroshi Okamoto

[*1] 富士ゼロックス(株)研究技術開発本部
Research & Technology Group, Fuji Xerox Co., Ltd.

現実世界の多くのネットワークでは、コミュニティ構造に階層性がある。しかしながら、コミュニティ階層構造をネットワークから効果的・効率的に検出する手法はまだ確立されていない。我々は、ランダムウォークの枠組みに基づいて、コミュニティ階層構造を効果的かつ安定に検出する機械学習アルゴリズムを構築した。さらに、この手法を用いて、実際の社会ネットワークのコミュニティ階層構造がしばしば非木型となることを発見した。

1. はじめに

全ネットワークの中の密に繋がったノード群は、ある機能あるいは役割を持つ塊だと考えられる。ネットワークを幾つかの塊に分解することにより、複雑な構造を明らかにし、ネットワーク内の役割とその相互作用を分析することができる。コミュニティ検出は、膨大なネットワークからそのような塊(「コミュニティ」と呼ぶ)を効率的かつ効果的に検出することを目指している。

現実世界の多くのネットワークのコミュニティには、重なりと階層性がある。上位のより大きなコミュニティは複数のより小さなコミュニティから構成されるということがよくある。階層構造を検出することにより、ネットワークの完全な構造を明らかにすることと共に、異なる解像度からネットワークを分析することもできる。

これまでに、階層構造を扱えるコミュニティ検出アルゴリズムがいくつか提案されている。しかしながら、いくつかの課題が残っている。

重なりと階層構造を同時に扱える方法は[3,4]ほとんどない。従来のノードにおける凝集的 (agglomerative)コミュニティ階層検出方法[2]は、コミュニティの間に重なりがないことを前提としており、各ノードをただ一つのコミュニティに割り当てるものであった。最近提案された解像度の変調を用いる方法[6,7,8]では、ある解像度において従来の凝集的方法を用いるため、重なりを扱うことができない。

従来の凝集的方法[2,3]は、ノードやリンクなどを一つずつ凝集するものであり、多くの層に単独ノードや小さな切片など、コミュニティとしての意味がないものを残してしまう。実際、文献[8]の研究により、従来の凝集的方法は、ただ一つの固定された解像度で、その解像度に対する最適なコミュニティ分解結果を見つける方法であることが示された。従って、従来方法は階層全体におけるただ一層に対する結果しか出せない。すなわち、従来の凝集的方法が与える樹状構造(dendrogram)は、正しくはコミュニティの階層構造とは見なせない。

さらに、凝集的方法では、樹状構造を登るに連れて、コミュニティの数が必ず一つずつ減ってゆく。しかしながら、実際の階層構造では、階層を登るに連れてコミュニティの数が複数減ってゆくことがしばしばある。従って、凝集的方法により得られた樹状構造は、全てのネットワークのコミュニティ階層構造を必ずしも反映しない。

連絡先:邱 シュウレ, 富士ゼロックス(株)研究技術開発本部,
〒220-8668 神奈川県横浜市みなとみらい 6 丁目 1 番.
E-mail: qiu-xule@fujixerox.co.jp

我々は、ランダムウォークの枠組みに基づいて、コミュニティの解像度を準静的変化させることにより、コミュニティ階層構造を効果的かつ安定に検出する機械学習アルゴリズムを構築した。このアルゴリズムは重なりと階層構造を同時に扱うことができる。この手法を用いて、実際の社会ネットワークのコミュニティ階層構造を分析したところ、それらがしばしば非木型となることを発見した。

2. 方法

2.1 ランダムウォークに基づくコミュニティ検出方法

本研究が提案するコミュニティ階層構造検出方法は、我々が以前に提案したコミュニティ検出方法を拡張したものである。そこで、まず、以前に提案したコミュニティ検出アルゴリズムについて簡単に振り返る。(詳細は文献[1]を参照)。

ネットワークの上を大勢の人達がリンクたどりながらノードからノードへと、ランダムに歩き回っていると考える。定常状態において、ランダムウォーカーの何人かはあるコミュニティの中を歩きまわり、他の何人かは別のコミュニティの中を歩き回っている。同じコミュニティを歩き回っている人達が同じ色の服を着ているならば、ネットワークの個々のコミュニティが色で分かれる。

ネットワークの隣接行列を $\mathbf{A}=(A_{nm})$ として、A_{nm} でノード m からノード n へのリンクの重みを表す。マルコフ性の仮定により、時刻 t にランダムウォーカーがノード n にいる確率 $p_t(n)$ は、時刻 $t-1$ の状態のみと関係があり、時間発展は、

$$p_t(n)=\sum_{m=1}^{N} T_{nm}\,p_{t-1}(m) \tag{1}$$

と表される。ただし、$T_{nm}=A_{nm}\Big/\sum_{n'=1}^{N}A_{n'm}$ は遷移確率である。充分時間が経てば、$p_t(n)$ が定常状態 $p_{t\to\infty}^{steady}(n)$ に収束する。定常状態において、ランダムウォーカーの所在位置を仮想的に観測する。観測されたデータは、ネットワークの潜在コミュニティ構造を前提として得られると考える。そこで、定常状態におけるノードの分布を

$$p^{steady}(n)=\sum_{k=1}^{K}\pi_k\,p(n\mid k) \tag{2}$$

と分解する。ただし、$\sum_{k=1}^{K}\pi_k=1$ である。π_k はコミュニティ k の事前確率、$p(n\mid k)$ はコミュニティ k におけるノードの確率分布である。左辺の式 $p^{steady}(n)$ は、服の色が区別されていない場合に観測されたランダムウォーカーの居場所の分布、右辺の式 $\sum_{k=1}^{K}\pi_k p(n\mid k)$ は、服の色でコミュニティ構造を表している場合

に観測されたランダムウォーカーの居場所の分布(コミュニティ毎のランダムウォーカーの居場所分布の総和)である。

リンクにおける各ランダムウォーカーの居場所を学習データ $\{\vec{\tau}^{(d)}\}\left(d=1,\cdots,D;\sum_{n=1}^{N}\vec{\tau}_n^{(d)}=2\right)$ ととらえ、最尤法を用いて、$\{\pi_k\}$ 及び $\{p(n|k)\}$ (このデータが得られることの背景であるネットワークのコミュニティ構造)を定める。

時刻 t のコミュニティ k における尤度関数を、多項分布と Dirichlet 分布により、次式で定義する。

$$\begin{aligned} &P\left(\{z_k^{(d)}\},\{\vec{\tau}^{(d)}\},\{p_t(n|k)\}\,|\,k\right) \\ &\sim\prod_{n=1}^{N}p_t(n|k)^{\alpha\sum_{m=1}^{N}T_{nm}p_{t-1}(m|k)}\cdot\prod_{n=1}^{N}p_t(n|k)^{\sum_{d=1}^{D}z_k^{(d)}\tau_n^{(d)}} \\ &\sim\prod_{n=1}^{N}p_t(n|k)^{\sum_{d=1}^{D}z_k^{(d)}\tau_n^{(d)}+\alpha\sum_{m=1}^{N}T_{nm}p_{t-1}(m|k)} \end{aligned} \tag{3}$$

ただし、$\vec{z}_k^{(d)}$ は観測データがコミュニティ k から生成されたかどうか表す潜在変数である。$\prod_{n=1}^{N}p_t(n|k)^{\sum_{d=1}^{D}z_k^{(d)}\tau_n^{(d)}}$ は、多項分布である。$\prod_{n=1}^{N}p_t(n|k)^{\alpha\sum_{m=1}^{N}T_{nm}p_{t-1}(m|k)}$ は、$p_t(n|k)$ の事前分布としての多項分布に共役な Dirichlet 分布である。ここで、α は Dirichlet 分布の精度を表すパラメタであり、次節で示すように、提案方法の階層構造の検出で重要な役割を果たす。全体の尤度関数は、次式で与えられる。

$$\begin{aligned} &P\left(\{z_k^{(d)}\},\{p_t(n|k)\},\{\vec{\tau}^{(d)}\}\right) \\ &\sim\prod_{k=1}^{K}\left\{\pi_k^{\sum_{d=1}^{D}\tau_n^{(d)}}\prod_{n=1}^{N}p_t(n|k)^{\sum_{d=1}^{D}z_k^{(d)}\tau_n^{(d)}+\alpha\sum_{m=1}^{N}T_{nm}p_{t-1}(m|k)}\right\} \end{aligned} \tag{4}$$

この尤度関数(を $\vec{z}_k^{(d)}$ の事後分布で平均化したもの)の最大化を、機械学習における標準手法である EM アルゴリズムに従って実行できる。各変数($\{\pi_k\}$ 及び $\{p(n|k)\}$)は次式で定められる(M-step)。

$$\begin{aligned} &p_t(n|k)=\frac{\tilde{\alpha}}{\tilde{\alpha}+\pi_k}\sum_{m=1}^{N}T_{nm}p_{t-1}(m|k)+\frac{1}{\tilde{\alpha}+\pi_k}\sum_{d=1}^{D}\gamma_{dk}\tau_n^{(d)}, \\ &\pi_k=\sum_{d=1}^{D}\gamma_{1k}\Big/D \end{aligned} \tag{5}$$

ただし $\tilde{\alpha}=\frac{\alpha}{2D}$。ベイズの定理により、$\vec{z}_k^{(d)}$ の推定が

$$\gamma_{dk}\equiv P\left(z_k^{(d)}=1\,|\,\vec{\tau}^{(d)}\right)=\frac{\pi_k\prod_{n=1}^{N}\left[p_t(n|k)\right]^{\tau_n^{(d)}}}{\sum_{k=1}^{K}\pi_k\prod_{n=1}^{N}\left[p_t(n|k)\right]^{\tau_n^{(d)}}} \tag{6}$$

で得られる(E-step)。

我々の以前の研究により、このコミュニティ検出アルゴリズムは、コミュニティの重なりを検出できることが示された。

2.2 階層構造検出

前節に述べたコミュニティ検出方法を拡張して、階層構造を検出する方法を以下に構築する。

(1) 解像度パラメタ $\tilde{\alpha}$ の準静的な変化

パラメタ $\tilde{\alpha}$ は Dirichlet 事前分布の精度に比例し、コミュニティ検出解像度を制御する。ランダムウォークを背景として考えると、$\tilde{\alpha}$ はあるコミュニティ内を歩き回っているランダムウォーカーの移動範囲、すなわち、コミュニティの広がり範囲を制御する。従って、$\tilde{\alpha}$ を変化させることにより, コミュニティの階層を導くことができると期待される。

まず $\tilde{\alpha}$ の値を十分小さい値に設定して、EMアルゴリズムを収束させる。すると、ネットワークが多数のコミュニティに分解される。これを、階層の最下層とする。次に、$\tilde{\alpha}$ の値を準静的に(十分ゆっくりと)増加させることにより、階層を下から上に導いていく。$\tilde{\alpha}$ の値を一回わずかに増やして、次に EM サイクルを一回実行する。これを繰り返す。こうすることにより、より小さなコミュニティが徐々に結合して、より大きなコミュニティができてゆく。以上の方法を Zachary`s karate club network に適用したところ、図1に示すコミュニティ事前確率 π_k の変化を得た。

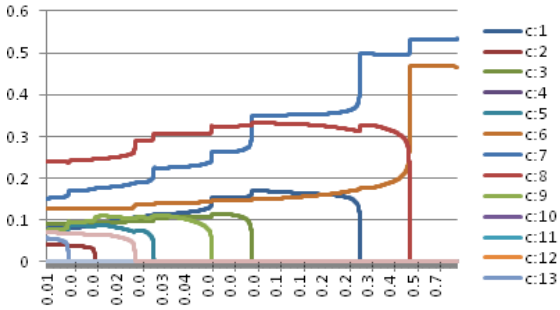

図 1: コミュニティの事前確率 π_k の変化図
(Zachary`s karate club network)

$\tilde{\alpha}$ の値が増加していくと、π_k が不連続相転移的に変化する。これは、あるところで、あるコミュニティ(あるいは複数のコミュニティ)がいきなり別のコミュニティ(あるいは複数のコミュニティ)に吸収されることを示す。これを新たな層ができた印とする。

π_k のある相転移点から、次の相転移点までの間、$\{\pi_k\}$ はほぼ静的に留まる。これが一つの層に対応する。その層のコミュニティ構造について、最も安定な結果を得るため、隣り合う二つの相転移点の中点における $\{\pi_k\}$ 、$\{p(n|k)\}$ 及び $\{\gamma_{1k}\}$ をその層のコミュニティ構造とする。

(2) 親子関係

各層のコミュニティ構造を得た後、隣接層間にコミュニティの親子関係を構成する(図 2)。

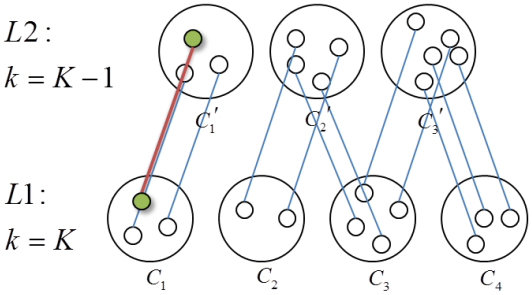

図 2: 親子関係の構成

まず、ある隣接層対において、低い方を L1、高い方を L2 とする。簡単のため、L1のコミュニティ数を K 、L2のコミュニティ数を $K-1$ と考える。また、L1 のコミュニティを $c_1,c_2,\ldots c_K$ として、L2のコミュニティを $c_1',c_2',\ldots,c_{K-1}'$ とする。次に、ネットワークの全てのノードを走査し、L1 と L2 の層にあるコミュニティの間にリンクをつけていく。

あるノードについて、まずそれぞれの層に対する帰属度を求める。そして、属しているそれぞれの層のコミュニティ(例えば c_1 と c_1' とする)の間にリンク $l_{c_1,c_1'}$ を重み

$$\omega(l_{c_1,c_1'})=\gamma_{nc_1}\cdot\gamma_{nc_1'}\Big/N \tag{7}$$

と定めると共に、コミュニティ c の重みを

$$\omega^t(c)=\omega^{t-1}(c)+\gamma_{nc}\Big/N \tag{8}$$

と更新する。ただし、$\omega(x)$ は x の重み関数である。γ_{nc} はノード n のコミュニティ c への帰属度である。コミュニティ間の関係の重みは、付けられる全てのリンクの重みの総計である。

$$\omega(L_{c_m,c_n'})=\sum_{l_{c_m,c_n'}}\omega(l_{c_m,c_n'}) \tag{9}$$

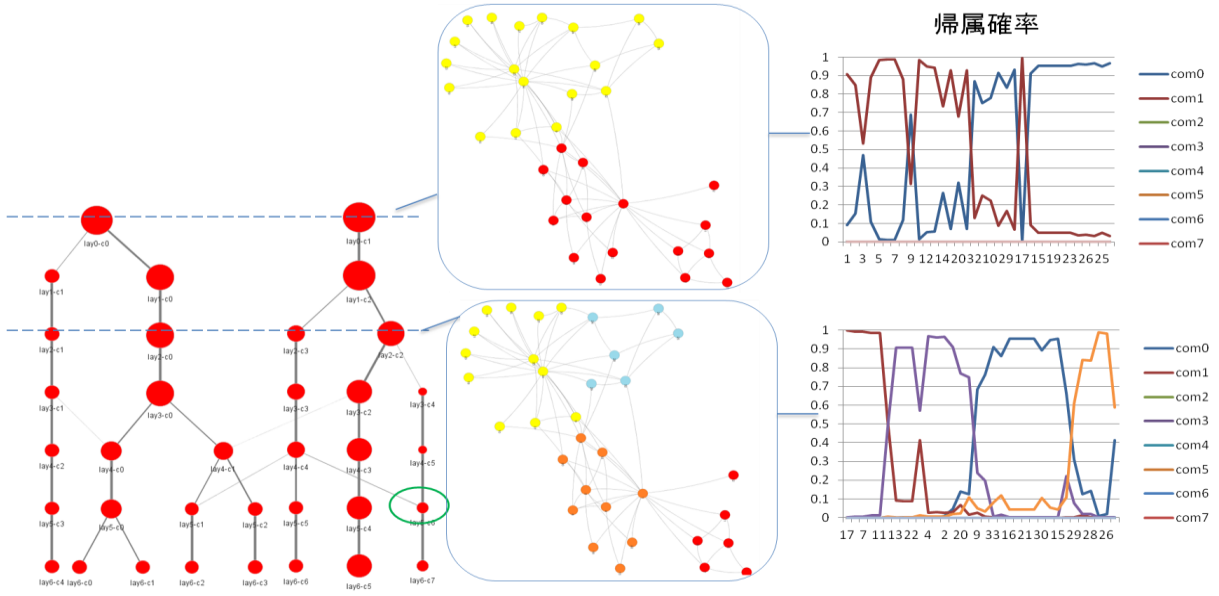

図 3:階層構造(Zachary`s karate club network)

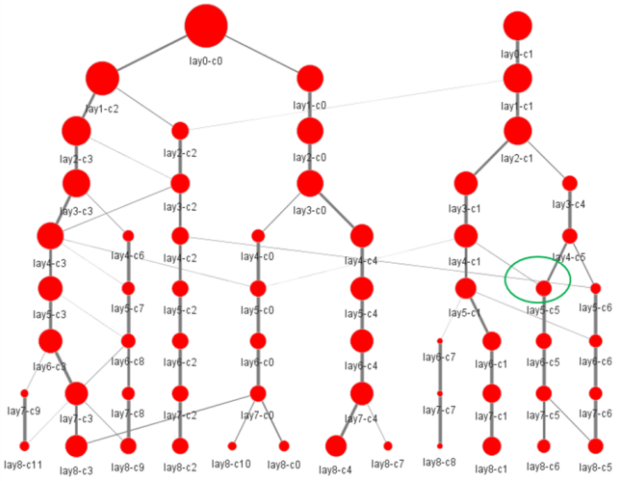

図4:階層構造(dolphin social network)

3. 結果

提案方法を karate club network に適用して、図 3 左に示すコミュニティ階層構造を得た。その中の二層に対するコミュニティ構造及びノードの帰属確率分布を、それぞれ、図 3 の中心及び右に示す。得られた帰属確率の分布から、ノードが複数のコミュニティに属している、すなわち、提案方法は重なりを扱えることが確認できる。提案方法を用いて、その他多くの社会ネットワークのコミュニティ階層構造も分析した(dolphin social network の階層構造を図4に示す)。

次に、コミュニティ階層構造が明らかに分かっているネットワーク(図5(上))に提案方法を適用した。このネットワークでは、20個のノードが強く結合して、一つのリングを形成する。五つのリングが結合して、一つのリンググループを形成する。さらに、五つのリンググループがリング状に緩く結合して、全体のネットワークが構成される。そこで、このネットワークは、二層(それぞれ、25個、5個のコミュニティからなる)の階層構造を持つ。提案方法は、この階層構造を完全に再現した(図5(下))。

一方、凝集的方法は、階層構造を登るにつれて、コミュニティの数が必ず一つずつ減ってゆくため、二層の途中で実際に存在しない層がたくさんできてしまう。すなわち、凝集的方法はこのネットワークの階層構造を正しく抽出できない。さらに、このネットワークのリング状のコミュニティのように、非クリーク(non-clique)型のコミュニティを、従来の凝集的方法では検出できないことが、文献[8]の研究により示されている。

最後に、しかしながら注目すべき順番は最初に来るべきことについて述べる。提案方法で抽出した実際の社会ネットワークのコミュニティ階層構造では、しばしば複数の親を持つ子コミュニティが存在する(図3(左)及び図4に緑円で示す)。すなわち、実際の社会ネットワークのコミュニティ階層構造が非木型になる。これは、提案方法により、はじめて見出された構造である。

安定性の証明

提案アルゴリズムでは、各変数の初期条件は乱数で決められる。従って、乱数の種を変えて、複数回の試行を行って、提案アルゴリズムから得られた階層構造の不変性(consistency)、すなわち、階層構造の安定性を調べることができる。

本研究では、階層構造の安定性を示すため、各層におけるコミュニティ分解結果の不変性を計算する方法を開発した。まず、ある層において、複数回の試行で得たコミュニティ検出結果の間で、同一コミュニティを同定する。次に、複数回の結果における各ノード n が各コミュニティ c への帰属分布を $p(n,c)$ として(ノードがランダムに帰属するとするならば、帰属分布を

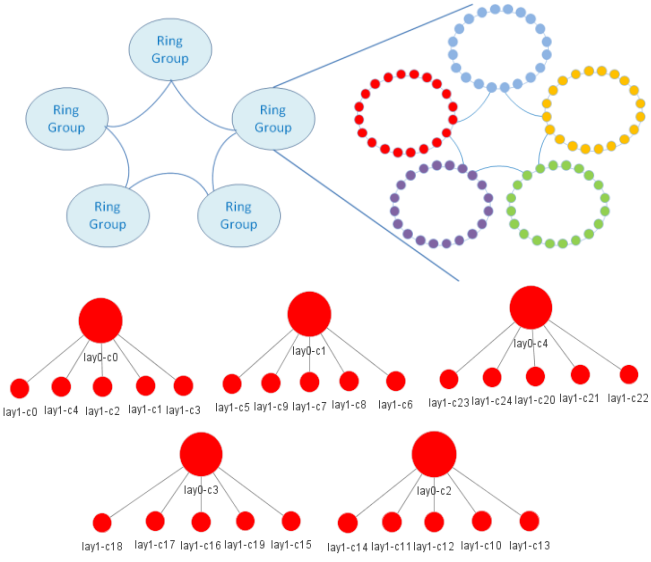

図5:階層構造(ring of ring group ネットワーク)

$p(n)\cdot p(c)$ と考える)、Kullback-Leibler divergence、KL を用い、ある層におけるコミュニティ検出安定度を定める:

$$\mathrm{KL}\left(p(n,c)\,\|\,p(n)\cdot p(c)\right)=\sum_{c=1}^{K}\sum_{n=1}^{N}p(n,c)\cdot\ln\frac{p(n)\cdot p(c\mid n)}{p(n)\cdot p(c)}$$

$$=-\sum_{n=1}^{N}p(n)\cdot\Bigg(\underbrace{-\sum_{c=1}^{K}p(c\mid n)\cdot\ln p(c\mid n)}_{\text{entropy}}\Bigg)+\Bigg(\underbrace{-\sum_{c=1}^{K}\sum_{n=1}^{N}p(n,c)}_{\text{marginal-integral}}\cdot\ln p(c)\Bigg)$$

$$=H(c)-\sum_{n=1}^{N}p(n)\cdot H(c\mid n)$$

ノード確率

(10)

ただし、$H(c)$ は全ての結果におけるコミュニティの確率分布のエントロピーである。$H(c\mid n)$ はあるノード n に関する帰属分布のエントロピーである。

式(10)を正規化したものを θ とする。θ の値の範囲は、$[0,1]$ である。もし、$\theta=1$ であれば、その層におけるコミュニティ検出安定度は100%である; $\theta=0$ であれば、コミュニティ検出結果はランダムであることを意味する。

$$\theta=\mathrm{KL}\left(p(n,c)\,\|\,p(n)\cdot p(c)\right)\Big/H(c) \tag{11}$$

この安定性指標を用いて、提案方法により安定な階層構造を検出できることをいくつかのベンチマークネットワークで確認した。ここでは、karate club network を例として述べる。50回の試行で、図6で示す安定性の結果を得た。コミュニティの数が二個、三個、四個である層において、θ は1であった。すなわち、50回の試行で得たコミュニティ分解結果は完全に一致した。それ以下の層では、安定性は徐々に下がる。コミュニティ階層構造において、より上位の部分が提案方法により正しく抽出できていることが分かる。

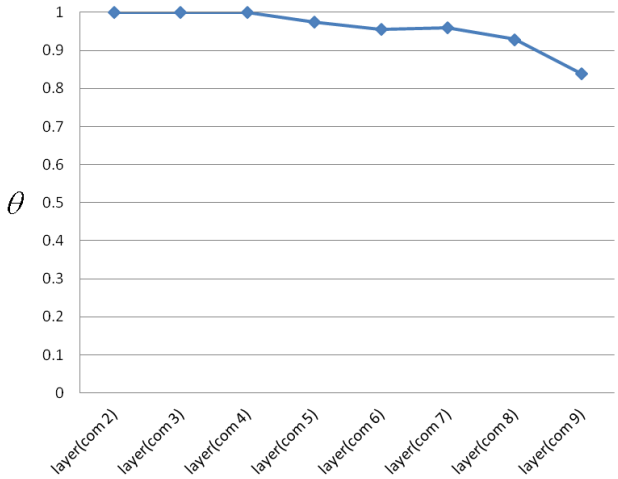

図6:階層の安定性(karate ネットワークに適用)

4. 議論

重なりと階層構造を同時に扱えるコミュニティ検出方法はこれまでほとんどなかった。我々は、マルコフ連鎖の枠組みに基づいて、コミュニティの解像度を変化させることにより重なりがある場合に階層構造を導く方法を構築した。文献[6,7,8]で提案された方法でも、我々のものと同様に解像度の変調を用いる。しかしながら、それらの方法は重なりを扱うことができず、隣接層間の親子関係も同定できない。

従来の凝集的方法により得られた樹状構造では、多くの層にコミュニティとしての意味が付かないもの(単独ノードや非常に小さな切片)が現れる。従って、このような樹状構造はネットワークの本来の階層構造ではない。さらに、各層の塊の数が層を登るたびに一つずつ減る。ネットワークの中には、層が一つ上がるとコミュニティ数が複数減る階層構造を持つものがしばしばある(例えば、図5)。提案方法はこのような階層構造を検出できることを示した。

また、提案階層検出方法の計算量は、元々のアルゴリズム[1]の計算量からたかだか加法でしか増加しない。元のアルゴリズムのコミュニティ検出する計算量が $O(M\times K\times r)$ である。ただし、M はネットワークのリンク数であり、K はコミュニティ数であり、r は EM ステップの反復回数である。今回提案したコミュニティ階層構造検出方法の計算量は $O(M\times K\times(r+T))$ である。ただし、T は、$\tilde{\alpha}$ の準静的変化に要する回数である。文献[6,7,8]で提案された方法は、各解像度において、コミュニティ検出計算を最初から実行する。従って、計算量は $O(M\times K\times r\times T)$ となり、膨大となる。

さらに、提案方法により、実際の社会ネットワークのコミュニティ階層構造は、しばしば「非木型」になることを見出した。実際、会社などにおける指揮系統のネットワークにおいて、ある組織が複数の上位組織から指揮を受ける、あるいは、同一人が複数の組織に属することがしばしばある。また、コンピュータープログラムにおいては、複数の上位モジュールが同一下位モジュールを共有することもごく普通である。我々が発見した「非木型」階層構造は、このようなことを反映でき、はじめて見出された構造である。

参考文献

1. 岡本 洋. マルコフ連鎖のモジュール分解:ネットワークからの重なりと階層構造を持つコミュニティの検出.JWEIN2014.
2. M. E. J. Newman. Communities, modules and large-scale structure in networks. Nature Phys 8, 25-31 (2012).
3. Y. Y. Ahn, J. P. Bagrow, & S. Lehmann. Link Communities reveal multiscale complexity in networks. Nature 466, 761-764 (2010).
4. A. Lancichinetti, S. Fortunato, & J. Kertesz. Detecting the Overlapping and Hierarchical Community Structure in Complex Networks. New Journal of Physics, 2009, 11: 033015 (2009).
5. M. Rosvall & C. T. Bergstrom. Multilevel Compression of Random Walks on Networks Reveals Hierarchical Organization in Large Integrated Systems. PlosOne, 6 (4): e18209 (2011).
6. P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, & J.P. Onnela. Community structure in time-dependent, multiscale, and multiplex networks. Science 328 (5980), 876-878 (2010).
7. R. Lambiotte, J. C. Delvenne, & M. Barahona. Laplacian dynamics and multiscale modular structure in networks. arXiv preprint arXiv:0812.1770, 135, (2008).
8. M.T. Schaub, J.C. Delvenne, S.N. Yaliraki, & M. Barahona. Markov dynamics as a zooming lens for multiscale community detection: non clique-like communities and the field-of-view limit. PLoS ONE 7, e32210 (2012).