Technical Report of the 2nd Workshop on Latent Dynamics (Jun 22, 2011, Tokyo, Japan)

ネットワークのコミュニティ分析とブートストラップ法

永田晴久* 下平英寿†
Haruhisa Nagata Hidetoshi Shimodaira

Abstract: 複雑ネットワーク・パラダイムにおいて、階層型クラスタリングはコミュニティ抽出の標準的な手法として用いられている。しかし、デンドログラム中のすべてのサブツリーがコミュニティ構造を持つことはなく、実際にコミュニティ構造を持つサブツリーをデンドログラム中から抽出する必要がある。本発表では、ブートストラップ法を利用して、デンドログラム中から有意なコミュニティ構造を持つサブツリーを選択する方法を提案する。

1 グラフ構造のクラスタリング

ネットワークのコミュニティ分析は、ネットワーク中から関係の強いノードを抽出する問題であり、ネットワーク科学において近年活発に研究が行われている分野である。現在広く用いられている手法は Girvan によるもの [1] や Newman によるもの [2] などがあるが、本研究では比較的最近の研究であり、単純なアルゴリズムながらクラスタ表現力の高い Ahn による方法 [3] を用いる。Ahn の方法では、ノード集合ではなく、リンク集合に対して階層型クラスタリングを適用する。

データとなるネットワークのグラフ構造を $G(V,E)$ とおく。グラフ G のノード数を $N = |V|$、リンク数を $M = |E|$、隣接集合を $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N)$ とする。2 本のリンクが共通のノード i に接するとき、その類似度 $S(e_{ij}, e_{ik})$ を次のように定義する。

$$S(e_{ij}, e_{ik}) = \frac{|n_+(j)| \cap |n_+(k)|}{|n_+(j)| \cup |n_+(k)|} \quad (1)$$

$$= \frac{\boldsymbol{a}_j \cdot \boldsymbol{a}_k}{||\boldsymbol{a}_j||^2 + ||\boldsymbol{a}_k||^2 - \boldsymbol{a}_j \cdot \boldsymbol{a}_k} \quad (2)$$

ただし、$n_+(i)$ は、ノード i とそれに隣接するノードの集合である。式 (2) は、2 ノード間に張られるリンクが複数の場合も許すような定義である。2 本のリンクが接する共通のノードを持たないときは、$S(e_{ij}, e_{kl}) = 0$ と定義する。

Ahn の方法では、類似度 S を用い、リンク集合に対して、単連結法による階層型クラスタリングを行う。結果として得られるデンドログラムでは、各サブツリーがグラフ G におけるリンクのクラスタ構造を表している。したがって、あるクラスタ中のリンクが接続するノードを列挙すれば、そのクラスタをノード集合として表現できる。

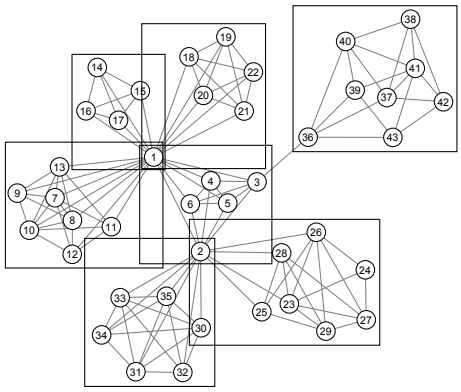

図 1: Ahn の方法に基づく階層型クラスタリングの例。実際にクラスタリングされるのはリンクであるため、コミュニティがオーバーラップしていても抽出される。

Ahn の方法には、次のような利点がある。

- クラスタはデンドログラムとして表されるので、クラスタの包含関係が一回の実行で取得できる。
- ひとつのノードを複数のクラスタに含めることができる。このような状況は多くのネットワークで出現するが、従来のノード分割を行うような方法では表現できなかった。

一方で、階層型クラスタリングを用いる手法に共通する問題点として、デンドログラム中に含まれるサブツリーが非常に多く、クラスタとしてみなせないものが多く含まれることが挙げられる。一般には、求まったクラスタのクラスタ係数の総和が最も大きくなるようにデン

*東京工業大学 大学院情報理工研究科, 152-8552 東京都目黒区大岡山 2-12-1, e-mail nagata.h.aa@m.titech.ac.jp,
Dept. of Mathematical and Computing Sciences, Tokyo Institute of Technology, 2-12-1 Ookayama Meguro-ku Tokyo 152-8552
†東京工業大学 大学院情報理工研究科, 152-8552 東京都目黒区大岡山 2-12-1, e-mail shimo@is.titech.ac.jp,
Dept. of Mathematical and Computing Sciences, Tokyo Institute of Technology, 2-12-1 Ookayama Meguro-ku Tokyo 152-8552

ドログラムを分割するが、これは上記の利点を失うことになるため、好ましい方法とは言えない。したがって、階層型クラスタリングによって求まったデンドログラムから、どのサブツリーをクラスタとみなすかの選別が問題となる。

2　ブートストラップ法の適用

この問題に対して、本研究ではブートストラップ法を用いる。このアイディアは、ブートストラップ法によってネットワークの「可能性のある変化」をシミュレートし、変化した後でも残っているクラスタを意味の強いクラスタとみなすという考えに基づく。具体的には、データであるグラフ構造をリサンプリングして新たなグラフ構造を作成し、各クラスタの生起確率をブートストラップ確率（bp）として求めることによって、クラスタの信頼性を計算する。

bp の計算方法は、バイオインフォマティクスにおける系統樹の信頼度推定で用いられている手法を流用する。元のデータに基づくデンドログラムを D、ブートストラップによって生成されたデータに基づくデンドログラムを $D_1^*, \ldots, D_B^*$ とする。また、木構造 T が T' のサブツリーであることを $T \subset T'$ と表すこととし、$l(T)$ を T の葉集合とする。このとき、D 中のサブツリー $T \subset D$ に対する bp_T は、次のように計算される。

$$\mathrm{bp}_T = \frac{\sum_{i=1}^{B} h(T, D_i^*)}{B} \tag{3}$$

$$h(T, D^*) = \begin{cases} 1 & (^{\exists}T^* \subset D^* \text{ s.t. } l(T) = l(T^*)) \\ 0 & (\text{otherwise}) \end{cases} \tag{4}$$

この bp が高いサブツリーほど、変化に強いクラスタとして抽出することにすればよい。

また、グラフ構造のリサンプリングについても、方法は自明ではない。これについては様々な方法が考えられるが、今回はエッジ集合の要素をデータと見てリサンプリングを行うこととした。すなわち、リンク集合 $E = \{e_1, \ldots, e_M\}$ に対し、重複を許してリサンプリングしたデータ $E^* = (e_1^*, \ldots, e_M^*)$ によって生成されるグラフをブートストラップサンプルとする。生成されるグラフは、次の条件を満たしている。

- 元のグラフにおいてリンクで結ばれているノード間は、リンクが複数本に増えることも、なくなることもある。
- 元のグラフにおいてリンクの存在しないノード間に、新たにリンクが張られることはない。

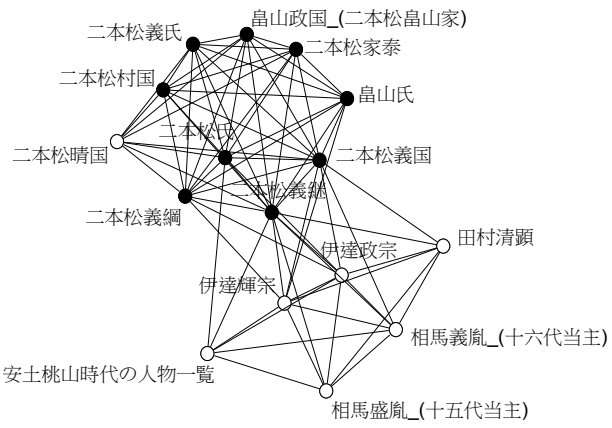

図 2: bp の高いクラスタの例。クラスタに所属するノードを黒丸で示し、その隣接ノードまでを描いている。

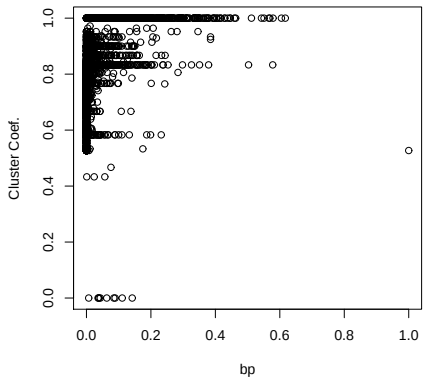

図 3: bp とクラスタ係数の関係。各点はクラスタを表す。

3　数値実験

提案した手法が実際に有効なクラスタを検出するかどうかを確かめるため、Wikipedia の記事ネットワークに対して数値実験を行った。データとして、Wikipedia の「戦国大名」カテゴリに属する記事 626 本と、その間に張られるリンク 5,341 本を用いた。この結果を、図 2、図 3、図 4 に示す。bp が高いクラスタはクラスタ係数から見ても、また実際にネットワーク図上でもコミュニティとして認められることが確かめられる。一方で、ノード数が大きいクラスタでは bp が 0 に張り付き、クラスタとして検出されなくなってしまうこともわかった。これは、大きなクラスタほどリサンプリング時の変化に影響されやすくなり、完全一致を用いるブートストラップ法では検出できなくなることに起因している。

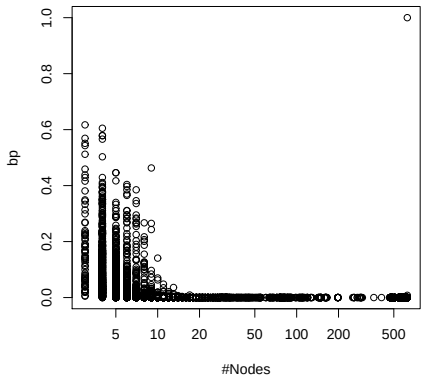

図 4: クラスタノード数と bp の関係。ノード数が大きくなると、bp は急激に小さくなる。

4　今後の課題

ブートストラップ法は、真の分布をリサンプリングデータの経験分布によって近似して、推定値を得る方法である。したがって、リサンプリング方法として、次のような方法も考えられる。

- 一部、あるいはすべての枝をランダムに張り替えてデータを生成するリサンプリング。
- BA モデルなどの生成モデルを仮定した、パラメトリックブートストラップ法によるリサンプリング。

このようなリサンプリングを考えることで、現実のネットワークに近い変化をシミュレートすることができ、より高い信頼性を持つ指標が計算できることが期待される。

また、カウント方法を工夫し、ノード数が大きいクラスタも検出できるようにしたい。

参考文献

[1] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks”, In *Proc. Natl. Acad. Sci. USA,* Vol.99, pp.7821-7826, 2002.

[2] M. E. J. Newman, “Modularity and community structure in networks”, In *Proc. Natl. Acad. Sci. USA,* Vol.103, pp.8577-8582, 2006.

[3] Y. Y. Ahn, J. P. Bagrow and S. Lehmann, “Link communities reveal multiscale complexity in networks”, In *Nature,* Vol. 466, pp.761-764, 2010.