グラフ構造のブートストラップ法とクラスター係数についての漸近評価

大阪大学 大学院基礎工学研究科　　永田 晴久
大阪大学 大学院基礎工学研究科　　下平 英寿

1　ネットワーク分析とブートストラップ法

本研究では，グラフ構造で表されるデータに対するブートストラップ法を提案する．

他の様々なデータ分析と同様に，ネットワークやグラフ構造に対する分析においても，データから求められた統計量の分布や分散を見て，統計量の信頼性を測ることは重要である．しかし，グラフ構造データの理論解析の困難さや，計算量的な観点などから，実際の分析では統計量の分布まで調べられることはこれまでほとんどなかった．そこで本研究では，グラフ構造分析における，統計量の分布を推定する簡便で扱いやすい方法の開発を目指している．

統計量の分布を推定する手法としては，ブートストラップ法を用いる．ブートストラップ法は幅広いクラスのデータで実行可能であり，また並列処理による大規模計算にも向いているため，グラフ構造の分析手法として優れている．しかし，グラフ構造に対してブートストラップ法を適用するには，リサンプリングの方法が自明でないなど，いくつかの問題がある．本研究では，これらの問題を解決した，グラフ構造に対するリサンプリング方法を提案する．

2　ポアソン分布によるリサンプリング

いま，ノード数 n の単純無向グラフ G_n が与えられているとし，G_n の隣接行列を $\boldsymbol{X}_n = (X_{ij})_n$ とする．提案法では，ブートストラップ標本となる隣接行列 $\boldsymbol{X}_n^* = (X_{ij}^*)_n$ を，次のように生成する．

$$\begin{array}{rcll} X_{ij}^* & \sim & \mathrm{Po}(1) & (i<j, X_{ij}=1) \\ X_{ij}^* & = & 0 & (i<j, X_{ij}=0) \end{array}$$

このとき，$\boldsymbol{X}_n^*$ で表されるグラフ G_n^* は，多重エッジを持つ無向グラフとなる．この方法は，G_n のエッジ集合に対するリサンプリングを改良したものと考えることができる．

3　クラスター係数の漸近評価

ネットワーク分析で用いられる統計量の一つにクラスター係数がある．クラスター係数は，ネットワーク中のあるノードに対して，その周辺のノードがどれだけ密に集まっているかを表す指標である．無向グラフ G_n において，ノード i に対するクラスター係数は次で与えられる．

$$C(i;\boldsymbol{X}_n) = \frac{\sum_{j<k} X_{ij}X_{ik}X_{jk}}{\sum_{j<k} X_{ij}X_{ik}}$$

ポアソン分布によるリサンプリングを用いると，クラスター係数の平均や分散の推定量を求めることができるようになる．たとえば，クラスター係数の平均 $\mu_n(i) = \mathbb{E}\{C(i;\boldsymbol{X}_n)\}$ に対して，ブートストラップ推定量を $\mu_n^*(i) = \mathbb{E}\{C(i;\boldsymbol{X}_n^*) \mid \boldsymbol{X}_n\}$ とすれば，$X_{ij} \sim \mathrm{Be}(p_{ij})$ $(i<j, \text{i.i.d.})$，$n \to \infty$ の条件の下で

$$\mu_n^*(i) - \mu_n(i) \longrightarrow 0 \quad \text{a.s.}$$

がいえる．したがって，n が十分に大きいとき，ブートストラップ推定量 $\mu_n^*(i)$ は $\mu_n(i)$ のよい近似となる．同様にして，分散やより高次のモーメントについても，ブートストラップ推定量の漸近評価を考えることができる．