A Study on Modularity Maximization

宮内 敦史

東京工業大学大学院社会理工学研究科経営工学専攻
 (現所属：東京工業大学大学院社会理工学研究科社会学専攻)
 指導教員：中田和秀 准教授

1. はじめに

ネットワーク解析の基本的な操作として、コミュニティ検出が知られている。コミュニティ検出とは、ネットワークをまとまりらしい部分（コミュニティ）に分割する操作であり、ネットワークの構造を理解するうえで重要な操作である。2004年、コミュニティ検出に対する評価指標として、Newman and Girvan [1] がモジュラリティを提案した。本研究では、無向グラフ $G = (V, E)$ （頂点数 n 、枝数 m ）におけるコミュニティ検出を考える。モジュラリティとは、頂点集合 V の分割 C を引数とする関数

$$Q(C) = \sum_{C \in \mathcal{C}} \left(\frac{m_C}{m} - \left(\frac{D_C}{2m} \right)^2 \right)$$

である。ここで、 m_C はコミュニティ C 内の枝数、 D_C はコミュニティ C 内の頂点の次数和である。モジュラリティが高い分割ほど、まとまりらしい分割となることが知られており、コミュニティ検出において最も支持されている評価関数である。

本論文は、モジュラリティ最大化に関する以下の二つの研究成果をまとめたものである：

(1) 2007年、二部ネットワークに適した評価指標として、Barber [2] が二部モジュラリティを提案した。二部モジュラリティとは、二部ネットワークの頂点集合 $V = R \cup B$ の分割 C を引数とする関数

$$Q_b(C) = \sum_{C \in \mathcal{C}} \left(\frac{m_C}{m} - \frac{R_C B_C}{m^2} \right)$$

である。ここで、 R_C はコミュニティ C 内の R に属する頂点の次数和であり、 B_C はコミュニティ C 内の B に属する頂点の次数和である。二部ネットワークにおけるコミュニティ検出では、二部モジュラリティが広く利用されている。通常のモジュラリティ最大化の NP-困難性が知られているのに対して、二部モジュラリティ最大化の計算複雑度は明らかにされていない。2011年、Zhan et al. [3] が通常のモジュラリティ最大化を二部モジュラリティ最大化に帰着できると述べたが、その

解析には根本的な誤りが含まれていた。本研究では、古典的な分割問題からの帰着を与え、二部モジュラリティ最大化の NP-困難性を示した。

(2) モジュラリティ最大化に対しては、これまでに数多くのヒューリスティクスが提案されてきた。しかしながら、計算機実験で用いられる大規模なネットワークでは、最適値や高精度な上界値が知られておらず、これらのヒューリスティクスが最適値にどこまで迫っているかは明らかにされていない。本研究では、モジュラリティの上界値を効率的に計算する二つの手法を提案した。どちらの手法も、モジュラリティ最大化の線形緩和問題を効率的に解く手法である。一つ目の手法は、Grötschel and Wakabayashi [4] によって提案された行生成法に簡単な工夫を施し、計算時間を大幅に短縮したものである。二つ目の手法は、行生成法と列生成法を組み合わせた手法である。この手法では、はじめに非常に小規模な部分問題を解き、逐次的に変数と制約式を追加していく。元の線形緩和問題に比べて非常に小規模な問題を解くだけで、元問題の最適解と最適値が得られることを理論的に示した。

2. 二部モジュラリティ最大化の計算複雑度

二部モジュラリティ最大化の決定問題版は、以下のように定義できる。

問題 (BIMODULARITY). 入力：二部ネットワーク $G = (V, E)$ と実数 K 、質問： $Q_b(C) \geq K$ を満たす V の分割 C は存在するか？

BIMODULARITY の強 NP-完全性を示すために、以下の決定問題を考える。

問題 (3-PARTITION). 入力： $3k$ 個の正整数の集合 $A = \{a_1, a_2, \dots, a_{3k}\}$ (ある整数 b に対して、 $\sum_{i=1}^{3k} a_i = kb$ かつ $b/4 < a_i < b/2$ ($i = 1, 2, \dots, 3k$) を満たす)、質問：集合 A の k 個の部分集合への分割で、各部分集合の要素和が b となるものは存在するか？

3-PARTITION は強 NP-完全であることが知られている。論文内では、3-PARTITION から BIMOD-

ULARITY への擬多項式時間帰着を与えることで、以下の定理を得た。

定理. BIMODULARITY は強 NP-完全である。

3. モジュラリティの上界値算出

ここでは、主な提案手法である Row and Column Generation (RCG) の解説を行う。

3.1 定式化と線形緩和問題

頂点集合を $V = \{1, 2, \dots, n\}$, 頂点 i の次数を d_i とし, $q_{ij} = A_{ij} - d_i d_j / 2m$, $C = \sum_{i=1}^n q_{ii} / 2m$ とおく。また、頂点 i と j が同一のコミュニティに属するとき 1, そうでないとき 0 をとる変数を x_{ij} とおく。すると、モジュラリティ最大化は、整数計画問題

$$\begin{aligned} \max. \quad & \frac{1}{m} \sum_{i=1}^{n-1} \sum_{j=i+1}^n q_{ij} x_{ij} + C \\ \text{s. t.} \quad & x_{ij} + x_{jk} - x_{ik} \leq 1 \quad \forall i < j < k, \\ & x_{ij} - x_{jk} + x_{ik} \leq 1 \quad \forall i < j < k, \\ & -x_{ij} + x_{jk} + x_{ik} \leq 1 \quad \forall i < j < k, \\ & x_{ij} \in \{0, 1\} \quad \forall i < j, \end{aligned}$$

として定式化できる。上記の整数計画問題において、0-1 制約を $x_{ij} \in [0, 1]$ に緩和することで、線形緩和問題 (LP) が得られる。この (LP) の最適値を求めることを目標とするが、数百頂点のネットワークでも入力すら困難である。

3.2 提案手法

提案手法 RCG は、(LP) の部分問題を扱う。はじめに、添字集合を用意する。(LP) に含まれる変数の添字全体の集合を I_{all} , 枝に関する変数の添字全体の集合を I_{init} とし, $I_{\text{out}} = I_{\text{all}} \setminus I_{\text{init}}$ とおく。ここで、 I_{out} の部分集合 I_{added} に対して、 $I_{\text{init}} \cup I_{\text{added}}$ に含まれる添字を持つ変数 (とそれらで書かれる制約式) のみを残した (LP) の部分問題を (LP(I_{added})) とおく。たとえ

ば、(LP($\emptyset$)) は枝に関する変数のみを持つ部分問題であり、(LP(I_{out})) は元問題 (LP) そのものである。

提案手法 RCG は、部分問題 (LP(I_{added})) を解きながら、以下のように変数と制約式を逐次的に加えていく手法である。

Step 1: $I_{\text{added}} = \emptyset$ とおく。

Step 2: (LP(I_{added})) を行生成法を用いて解き、暫定最適解 $\bar{x}^*$ と暫定最適値 $\bar{Q}^*$ を得る。

Step 3: 暫定最適解 $\bar{x}^*$ がある条件 (*) を満たしていれば、 $\bar{Q}^*$ を出力する。そうでなければ、適切な変数を I_{added} に加え、Step 2 に戻る。

論文内では、部分問題 (LP(I_{added})) の最適解 $\bar{x}^*$ が条件 (*) を満たすならば、(LP(I_{added})) に加えていない変数を 0 で補完することで、元問題 (LP) の最適解が得られることを示した。結果として、以下の定理を得た。

定理. RCG は (LP) の最適解と最適値を得る。

3.3 計算機実験

モジュラリティ最大化の研究で頻繁に用いられる複数のネットワークに対して、詳細な実験を行った。RCG を用いることで、5,000 頂点程度のネットワークに対しても (LP) の最適値が得られた。

参考文献

- [1] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, **69**, 026113, 2004.
- [2] M. J. Barber, "Modularity and community detection in bipartite networks," *Physical Review E*, **76**, 066102, 2007.
- [3] W. Zhan, Z. Zhang, J. Guan and S. Zhou, "Evolutionary method for finding communities in bipartite networks," *Physical Review E*, **83**, 066120, 2011.
- [4] M. Grötschel and Y. Wakabayashi, "A cutting plane algorithm for a clustering problem," *Mathematical Programming*, **45**, 59–96, 1989.