モジュラリティの上界値算出

宮内　敦史

（上智大学理工学部情報理工学科　現所属：東京工業大学大学院社会理工学研究科経営工学専攻）
指導教員：宮本裕一郎 准教授

1. はじめに

近年，ソーシャル・ネットワーキング・サービスの普及などにより，ネットワーク解析が注目されている．その中でも特に，ネットワークを密な部分ネットワークに分割する手法は，盛んに研究されている．このような密な部分ネットワークをコミュニティと呼び，ネットワークをコミュニティに分割することをコミュニティ検出という．図 1 として，その例を挙げる．

本稿では，単純無向グラフ $G=(V,E)$ におけるコミュニティ検出を考える．2004 年に，モジュラリティと呼ばれる，コミュニティ検出結果の評価指標が提案された [3]．ここで，枝数を m，隣接行列の (i,j) 成分を A_{ij}，頂点 i の次数を d_i とおく．また，頂点 i が属すコミュニティを C_i とし，δ をクロネッカーの記号とする．すると，モジュラリティは

$$Q=\frac{1}{2m}\sum_{i\in V}\sum_{j\in V}\left(A_{ij}-\frac{d_id_j}{2m}\right)\delta(C_i,C_j)$$

と表される．モジュラリティは必ず 1 以下であり，値が高いほどコミュニティ検出の結果が良いとされる．

多くの場合，コミュニティ検出問題は，モジュラリティ最大化問題として定義される．モジュラリティ最大化問題は NP–困難であるため，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．現在では，1 億頂点のグラフに対しても，高速かつ質の高いコミュニティ検出が可能である．

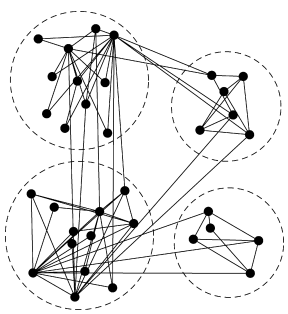

図 1　コミュニティ検出の例

しかしながら，モジュラリティ最大化問題に対する厳密解法や精度保証付き近似解法は，ほとんど提案されていない．さらに，モジュラリティの性質を利用した上界値算出法は，全く提案されていない．ヒューリスティクスの性能評価を行う際，実現されたモジュラリティが最適値あるいは上界値にどの程度迫っているのかは，検出時間と並ぶ重要な要素である．

本稿では，モジュラリティの上界値を求める手法について議論する．その中で，従来のものよりも高速かつ省メモリな手法を提案する．

2. 従来研究

2.1 自然な定式化

頂点集合を $V:=\{1,2,\ldots,n\}$ とし，$q_{ij}:=A_{ij}-d_id_j/2m$ とする．そして，頂点 i と j が同一のコミュニティに属すとき 1，そうでないとき 0 をとる変数を x_{ij} とおく．任意の頂点 i に対して $x_{ii}=1$ であるから，x_{ii} は変数として採用せず，$C:=\sum_{i=1}^{n}q_{ii}/2m$ を目的関数値に加える．さらに，任意の頂点対 i,j に対して，$x_{ij}=x_{ji}$ と $q_{ij}=q_{ji}$ が成り立つ．したがって，変数として $x_{ij}\,(i<j)$ のみを採用すると，コミュニティ検出問題は

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

として定式化できる．変数は $\Theta(n^2)$ 個，0-1 制約以外の制約式は $\Theta(n^3)$ 本ある．

定式化 (IP) において，0-1 制約を $x_{ij}\in[0,1]$ に置き換えることで，線形緩和問題 (LP) を得る．この (LP) を解くと，モジュラリティの上界値が得られる．特に，コミュニティ構造が顕著であるグラフに対しては，最適値に近い上界値を得られることがわかっている．しかし，頂点数が 200 を超えるグラフでは，実行時間とメモリ使用量の両面で，解くのが困難である．

 Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

2.2 疎な定式化

2011 年に，定式化 (IP) に比べて制約式の本数を大幅に減らしながらも，同一の最適解集合をもつ定式化が提案された [1]．ここで，頂点 i の隣接点全体の集合を $N(i)$ とし，$N(i,j) := N(i) \cup N(j) \setminus \{i,j\}$ とする．すると，提案された定式化は

$$(\mathrm{IP}_{\mathrm{sparse}}) : \max.\ \frac{1}{m}\sum_{i=1}^{n}\sum_{j>i}^{n} q_{ij}x_{ij} + C$$

$$\begin{aligned}
\text{s. t. } & x_{ki} + x_{kj} - x_{ij} \leq 1 \quad \forall k \in N(i,j) < i < j,\\
& x_{ik} + x_{kj} - x_{ij} \leq 1 \quad \forall i < k \in N(i,j) < j,\\
& x_{ik} + x_{jk} - x_{ij} \leq 1 \quad \forall i < j < k \in N(i,j),\\
& x_{ij} \in \{0,1\} \qquad\quad \forall i < j,
\end{aligned}$$

と書ける．そして，次の定理が成り立つ [1]．

定理 1. (IP) と $(\mathrm{IP}_{\mathrm{sparse}})$ は同一の最適解集合を持つ．

目的関数は (IP) のものと全く変わっていない．変わったのは，制約式の本数である．(IP) では，制約式は $\Theta(n^3)$ 本あった．これに対して，定式化 $(\mathrm{IP}_{\mathrm{sparse}})$ では，制約式の本数の上界値が

$$\sum_{i=1}^{n}\sum_{j>i}^{n}(d_i + d_j) = (n-1)\sum_{i=1}^{n} d_i = \mathrm{O}(mn)$$

で与えられる．コミュニティ検出の対象となる多くのネットワークは疎であるため，応用上のほとんどの場面では，制約式は $\mathrm{O}(n^2)$ 本であると評価できる．

$(\mathrm{IP}_{\mathrm{sparse}})$ において，0-1 制約を $x_{ij} \in [0,1]$ に置き換えることで，線形緩和問題 $(\mathrm{LP}_{\mathrm{sparse}})$ を得る．すると，先程と同様に，次の定理が成り立つ [1]．

定理 2. (LP) と $(\mathrm{LP}_{\mathrm{sparse}})$ は同一の最適解集合を持つ．

この定理により，$(\mathrm{LP}_{\mathrm{sparse}})$ で得られるモジュラリティの上界値は，(LP) で得られる値と等しいことがわかる．実験的には，より高速に求められることがわかっている．しかし，頂点数が 1,000 を超えるグラフに対しては，コミュニティ構造が自明である場合を除いて，非常に時間がかかる．さらに，1,500 頂点を超える規模のグラフでは，メモリ使用量の面でも解くのが困難である．

3. 提案手法

ここでは，モジュラリティの上界値を求める手法として，制約式制限法と変数制限法を提案する．ただし，前者に関しては単純な手法を，後者に関してはアイデアのみを示す．

3.1 制約式制限法

疎な定式化による制約式の減少は劇的であり，頂点数が 1,500 程度のグラフであれば，すべての制約式の走査が可能となった．これを利用し，制約式を順次加えて上界値を得る手法を提案する．本手法は，文献 [2] の手法に若干の工夫を施したものである．以下に，具体的な手順を示す．

Step 1. 定式化 $(\mathrm{LP}_{\mathrm{sparse}})$ について，$k \in N(i,j)$ を $k \in N(i) \cap N(j)$ とした制約式のみを採用して解き，暫定最適解 $\bar{\boldsymbol{x}}^*$ と暫定最適値 $\bar{z}^*$ を得る．

Step 2. $(\mathrm{LP}_{\mathrm{sparse}})$ のすべての制約式を走査する．$\bar{\boldsymbol{x}}^*$ がそのすべてを満たしていれば，モジュラリティの上界値として $\bar{z}^*$ を返して終了する．そうでなければ，$\bar{\boldsymbol{x}}^*$ が満たさないすべての制約式を加えて解き，$\bar{\boldsymbol{x}}^*$ と $\bar{z}^*$ を更新する．Step 2 を繰り返す．

この手法では，大抵は $(\mathrm{LP}_{\mathrm{sparse}})$ を単純に解くよりも速く上界値を求められる．しかも，多くの場合，保持する制約式は全体の 10%にも満たず，省メモリ性も高いと言える．しかし，大規模なグラフに対しては，依然としてサイズが大きい問題を扱うことになる．

3.2 変数制限法

$(\mathrm{LP}_{\mathrm{sparse}})$ を変数面でも小規模化することで，さらに省メモリ性を高めた手法である．任意の頂点対 i,j $(\{i,j\} \notin E)$ について $A_{ij} = 0$ であるから，$q_{ij} = A_{ij} - d_i d_j/2m \leq 0$ が成り立つ．したがって，$(\mathrm{LP}_{\mathrm{sparse}})$ において変数 x_{ij} $(\{i,j\} \in E)$ をすべて採用すれば，$(\mathrm{LP}_{\mathrm{sparse}})$ の緩和問題が得られる．しかし，x_{ij} $(\{i,j\} \notin E)$ を全く採用していない問題を解いても，ほぼ自明な上界値しか得られない．そこで，目的関数値を下げる方向に働くであろう変数 x_{ij} $(\{i,j\} \notin E)$ を順次採用していく．

4. 計算結果

上記の提案手法により，1,500 頂点のネットワークにおいて高精度な上界値が得られた．詳細は，紙面の都合上省略する．

参考文献

[1] Dinh, T. N. and Thai, M. T.: Finding community structure with performance guarantees in complex networks, *CoRR*, Vol. abs/1108.4034, 2011.

[2] Grötschel, M. and Wakabayashi, Y.: A cutting plane algorithm for a clustering problem, *Mathematical Programming*, **45**, 59–96, 1989.

[3] Newman, M. E. J. and Girvan, M.: Finding and evaluating community structure in networks, *Physical Review E*, **69**, 026113, 2004.

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.