グラフ分割の検出限界

川本達郎 〈産業技術総合研究所人工知能研究センター　kawamoto.tatsuro@aist.go.jp〉

グラフ分割，と言ってもあまり馴染みがないかもしれない．グラフ分割は，基本的には計算機科学・統計科学の対象であり，自然科学とは毛色が異なる面も存在するが，統計力学的なアプローチでの研究が近年も活発に進められている．

グラフ分割は，頂点と枝で構成されるグラフ（ネットワーク）から，いくつかの部分グラフに分割し，マクロなグループ構造を抽出する問題である．グラフ分割には，純粋に最適化問題としての定式化と，推論問題としての定式化の2種類が存在する．純粋な最適化問題とは，例えばグラフ上のコスト（もしくはエネルギー）最小化問題で，与えられた制約のもと，通信コストや消費電力等のコストが最小になるようにグループ分けする問題である．推論問題としてのグラフ分割とは，例えば人間関係・出版物の引用関係・遺伝子間関係・画像や地図上の要素間関係等のデータから，類似した特徴を持つグループを抽出するという問題である．

前者はコストの定義が明確で，とにかく最もコストが小さくなる解を見つけることが至上命題である．一方後者は，何をコストや制約とすべきかは明確ではなく，そもそも絶対的な正解というものが存在しない．自らコスト関数をデザインして評価する必要があり，そのためにはグラフデータがどのように生成されたものなのかという統計性も考慮する必要がある．前者に比べて後者はやたらと曖昧な問題であるが，そうは言っても社会データや画像データから特徴抽出をしたいというニーズは確固として存在する．

さらにグラフ分割は，通常，計算困難な問題になっており，最適解を求めることは技術的に困難である．そのため，使用しているアルゴリズムによってどの程度最適化がちゃんとできているのかの評価も難しい．

問題設定の曖昧性と計算困難性という，推論問題としてのグラフ分割が抱える二重の困難は，混沌とした研究の流れを生んだ．自然科学の問題と異なり，自らデザインしたコスト関数が実験結果を説明するという要請もないため，2000年代にはコミュニティ検出というテーマで膨大な数の手法提案論文が出版された．

しかし，手法だけ提案し，評価はいい加減にやっておけば良いというのでは科学として成り立たない．得られた分割結果はどの程度“正しい”のだろうか．このような推論問題に対しての真面目な取り扱いは，統計的有意性をちゃんと評価するということである．

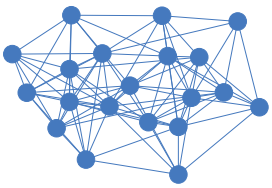
正解としてのグループ構造を持つ人工的なグラフデータを，ランダムグラフモデルによって生成してみたとき，特定のアルゴリズムが出力する分割結果がどれだけ正解を当てられるかを考えよう．正解としてのグループ構造を一様な構造に近づけていくと次第にアルゴリズムの正解率が下がっていき，あるところで，完全ランダムに分割した場合と同程度の正解率しか得られなくなってしまう．すなわち，統計的に有意な解が得られなくなる．この限界を（アルゴリズミックな）検出限界と呼び，グラフサイズが無限大の極限で，相転移として捉えることができる．

興味深いことに，グラフが十分にスパースなときは，どれだけはっきりとしたグループ構造のグラフを生成したとしても検出限界を超えてしまうことがある．この場合，アルゴリズムは完全にその機能を失ってしまうため，「どうせ正解がないのだから」という言い訳が通用しなくなり，出力結果は“正しい”とは言えなくなる．このように，理論的な後ろ盾（やそれがないこと）を増やしていくことで，より精密な議論が可能になる．

用語解説

グラフ分割：
グラフを，指定された数の部分集合（部分グラフ）に分割（分類）すること．

グラフ（ネットワーク）：
離散的なオブジェクト（人や物・文章・事象など）の（離散的な）関係性を，頂点集合とその間を結ぶ枝で表現したもの．



コミュニティ検出：
グラフ分割が与えられた数にグラフを分割するのに対し，分割数の推定までを含めた問題をコミュニティ検出と呼ぶ．また，密につながった頂点集合に分割することも定義に含まれることが多い．

統計的有意性：
得られた結果が，ランダムなイベントの結果生じた偶然ではないと判断できる場合，「統計的に有意である」と言う．