

令和元年5月31日現在

機関番号：82401

研究種目：研究活動スタート支援

研究期間：2017～2018

課題番号：17H07357

研究課題名（和文）ネットワーク上のコミュニティ検出に対する最適化基盤の確立

研究課題名（英文）Developing the Foundation of Optimization Methods for Community Detection in Networks

研究代表者

宮内 敦史（Miyuchi, Atsushi）

国立研究開発法人理化学研究所・革新知能統合研究センター・特別研究員

研究者番号：80804202

交付決定額（研究期間全体）：（直接経費） 2,300,000円

研究成果の概要（和文）：コミュニティ検出とは、ネットワーク上のまとまりらしい部分を取り出す操作であり、ネットワーク解析における基本的かつ重要な操作である。本研究では、コミュニティ検出に対して、最適化基盤の確立を行った。具体的には、コミュニティ検出の「モデル化」と「アルゴリズム設計」の両者に対して、それらを行うための枠組みを構築し、既存手法では検出し得なかった高精度なコミュニティ構造を検出できる手法を開発した。

研究成果の学術的意義や社会的意義

ネットワーク上のコミュニティ検出は、ネットワーク科学の分野で最も活発に研究されているトピックの一つである。学術的に重要なだけでなく、実社会での様々な応用が知られている。たとえば、ウェブグラフ上でコミュニティ検出を行うと、類似トピックを扱うウェブページの集合や、スパムリンクファームと呼ばれる不正リンクを多量にもつウェブページの集合を検出することができ、検索エンジンの精度向上に貢献することができる。

研究成果の概要（英文）：Community detection is a fundamental primitive in network analysis, which has a wide variety of applications in diverse domains. In this research project, we have developed the foundation of optimization methods for community detection. Specifically, we provided several frameworks for the modeling and the design of algorithms for community detection, enabling us to detect community structure with higher accuracy compared with existing methods.

研究分野：数理最適化

キーワード：複雑ネットワーク コミュニティ検出 モデル化 アルゴリズム設計 数理最適化

## 1. 研究開始当初の背景

コミュニティ検出とは、ネットワーク上のまとまりらしい部分を取り出す操作であり、ネットワーク解析の様々な場面で利用されている。コミュニティ検出法を設計するときには、多くの場合で数理最適化のアプローチが採用される。すなわち、コミュニティ検出を何らかの最適化問題としてモデル化し、それに対するアルゴリズムを設計する。これまでに様々なモデル化とアルゴリズムが提案されてきたが、より高精度なコミュニティ検出法の実現が望まれている。

## 2. 研究の目的

本研究の目的は、コミュニティ検出の「モデル化」と「アルゴリズム設計」の両者に対して、それらを行うための枠組みを構築し、コミュニティ検出に対する数理最適化によるアプローチの改善を、統一的かつ体系的に行えるようにすることである。

## 3. 研究の方法

### (1) モデル化の枠組みの構築

モデル化の枠組みの構築を行うためには、コミュニティ検出で用いられる評価関数がどのような性質を満たすべきかを考察する必要がある。ここで、評価関数とは、頂点部分集合がコミュニティらしいほど高い(あるいは低い)値をとるような関数を指す。コミュニティ検出の特殊ケースである密グラフ抽出に対しては、「密度」と呼ばれる評価関数が標準的である。そして、その評価関数を最大化する問題である「最密部分グラフ問題」は、様々なモデル化の基礎として広く利用されている。コミュニティ検出では、頂点部分集合の境界についても考慮する必要があるため、それを踏まえて様々なモデル化の基礎となるような新たなモデル化を導入する。また、モデル化の枠組みを構築する上で、現実的な状況のうち既存のモデル化では扱うことができないものを指摘し、それを克服するモデル化を導入することも重要である。特に、本研究では、グラフの枝重みに不確実性が存在する状況に着目する。

### (2) アルゴリズム設計の枠組みの構築

コミュニティ検出の特殊ケースである密グラフ抽出に対しては、アルゴリズム設計の枠組みがいくつか存在する。本研究では、これらの枠組みをコミュニティ検出に拡張する。

#### Greedy peeling の拡張による枠組みの構築

Greedy peeling は、密グラフ抽出で頻繁に利用される、アルゴリズム設計の枠組みである (Asahiro et al., *J. Algorithms*, **34**, pp. 203–221, 2000)。具体的には、ネットワークの最小次数頂点を逐次的に除去していき、そこで現れた頂点部分集合のなかで、目的関数値が最大となるものを出力するという枠組みである。Greedy peeling の特長は、その汎用性と高速性である。実際、密グラフ抽出をモデル化したほぼ全ての最適化問題に適用でき、ほぼ線形時間で走ることが知られている。さらに、出力解が理論的な精度保証をもつ場合もある。たとえば、上記の最密部分グラフ問題に対しては、 $1/2$ -近似解が得られる (Charikar, In *APPROX 2000*, pp. 84–95)。

本研究では、Greedy peeling を拡張することで、コミュニティ検出に対するアルゴリズム設計の枠組みを構築する。拡張の方法としては、たとえば以下の二つが考えられる。一つ目は、除去する頂点を決定するときに、内部への次数の小ささだけでなく、外部への次数の大きさも考慮するように変更する方法である。コミュニティは、単なる密グラフとは異なり、外部への枝の少なさも重要な要素とするため、この拡張は直感的には自然である。もう一つは、最大化したい評価関数の意味で貢献度が最小の頂点を除去するように変更する方法である。様々な最適化問題に対する理論的および実験的な解析を通じて、どのような拡張が効果的であるかを検討する。

#### 数理計画法を用いた枠組みの構築

密グラフ抽出に対しては、数理計画法を用いたアルゴリズム設計の枠組みが構築されている。Greedy peeling と比べて汎用性や高速性は劣るが、より高精度なアルゴリズムを導出できる場合が多い。たとえば、最密部分グラフ問題とその拡張に対して、線形計画法を用いた多項式時間の厳密解法が知られている (Charikar, In *APPROX 2000*, pp. 84–95; Tsourakakis, In *WWW 2015*, pp. 1122–1132)。本研究では、このような数理計画法を用いた枠組みを拡張することで、コミュニティ検出に対するアルゴリズム設計の枠組みを構築する。

#### 4. 研究成果

本研究では、主に以下の二つの成果を得た。

##### (1) コミュニティ検出に対する自然なモデル化とアルゴリズム設計に関する成果

コミュニティ検出では、頂点部分集合の内部と境界の両方を考慮する必要がある。頂点部分集合の内部（のみ）に注目する評価関数としては「密度」が標準的であり、頂点部分集合の境界（のみ）に注目する評価関数としては、「コンダクタンス」が標準的である。しかしながら、その両方に注目する自然な評価関数は知られていない。本研究では、頂点部分集合の内部と境界の両方を考慮する、新たな評価関数を導入した。具体的には、密度をもとにして、カットが大きいとペナルティを課すような評価関数となっている。そして、この評価関数を最大化する問題として、コミュニティ検出をモデル化した。

上記の最適化問題の利点は、多項式時間で最適解が得られる点である。実際、二つの多項式時間厳密解法を設計した。一つは線形計画法に基づくアルゴリズムであり、もう一つは最大流計算に基づくアルゴリズムである。どちらのアルゴリズムも最密部分グラフ問題に対するアルゴリズムを拡張することで設計した。これらのアルゴリズムは多項式時間で走るが、大規模なグラフに適用できるほど高速ではない。そこで、Greedy peeling を拡張することで、高速な精度保証付き近似解法を設計した。上記の方針の通り、除去する頂点を決定するとき、内部への次数の小ささだけでなく、外部への次数の大きさも考慮するように変更した。これにより、出力解の品質を理論的に保証することができた。

計算機実験では、提案モデル化の妥当性と提案アルゴリズムの性能を評価した。はじめに、提案モデル化の妥当性を評価するために、小中規模なグラフを用いて、最密部分グラフと提案モデル化の最適解の構造を比較した。その結果、多少の密度の犠牲は見られるものの、カットが非常に小さくなっており、モデル化の妥当性を確認することができた。つぎに、大規模なグラフを用いて、最密部分グラフ問題の近似解と提案モデル化の近似解の構造を比較した。結果の傾向は上記と変わらず、さらに、グラフの規模が大きくなるほど、提案モデル化の近似解の優位性が顕著となった。

本成果は、論文 Finding a dense subgraph with sparse cut として、国際会議 CIKM 2018 に採択された（採択率：147/862 = 17.1%）。

##### (2) 不確実性が存在する状況でのコミュニティ検出に対するモデル化とアルゴリズム設計に関する成果

コミュニティ検出のほぼ全てのモデル化は、ある強力な仮定を暗に置いている。それは、与えられるグラフの情報に不確実性が存在しない、という仮定である。実社会の応用では、真の枝重みの上下界や推定値しか得られない場合など、グラフに関して不確実な情報しか得られない状況は多い。本研究では、ロバスト最適化の理論を用いて、このような不確実な状況を扱うための新たなモデル化を導入した。コミュニティ検出の特殊ケースである密グラフ抽出に対しても同様の問題が存在するため、解決の第一歩として密グラフ抽出に取り組んだ。具体的には、「ロバスト最密部分グラフ問題」と「サンプリング・オラクル付きロバスト最密部分グラフ問題」という、二つの最適化問題を導入した。一つ目の問題は、真の枝重みの上下界しか得られない状況をモデル化しており、二つ目の問題は、真の枝重みの推定値をサンプリングすることができる状況をモデル化している。

上記の最適化問題のそれぞれに対して、理論的な精度保証をもつアルゴリズムを設計した。どちらのアルゴリズムも最密部分グラフ問題に対する既存のアルゴリズムを内部で利用しているが、不確実性を扱うための様々な工夫が施されている。一つ目の問題に対するアルゴリズムは、真の枝重みの上下界の情報を利用して枝重みを生成し、その枝重みでの最密部分グラフを出力する。これにより、決定的なアルゴリズムのなかで最良の精度を達成できることを保証した。一方、二つ目の問題に対するアルゴリズムは、真の枝重みの推定値を適当な回数サンプリングすることで枝重みを生成し、その枝重みでの最密部分グラフを出力する。これにより、擬多項式時間はかかるものの、真の枝重みでの最密部分グラフと同等の精度をもつ頂点部分集合が得られることを保証した。

計算機実験では、提案アルゴリズムの性能を評価した。その結果、単純な手法と比較して、提案アルゴリズムの両者が優れていることが確認できた。特に、二つ目の提案アルゴリズムは、実験で用いたほぼ全てのグラフに対して、真の枝重みでの最密部分グラフを検出することができた。

本成果は、論文 Robust densest subgraph discovery として、国際会議 ICDM 2018 に採択された（採択率：189/948 = 20.0%）。

上記の二つの成果については、それぞれの論文が採択された国際会議だけでなく、ネットワーク解析の世界的な研究者が所属している研究機関 ISI Foundation（イタリア，トリノ）での招待講演でも周知している。すでに数々のフィードバックが寄せられており、特に、二つ目の成果に対しては、関連する新たなモデル化の研究が行われている（Tsourakakis et al., *arXiv:1904.08178*, 2019）。

## 5 . 主な発表論文等

〔雑誌論文〕(計 4 件)

Atsushi Miyauchi and Akiko Takeda: Robust densest subgraph discovery, *Proceedings of the 18th IEEE International Conference on Data Mining (ICDM 2018)*, pp. 1188–1193, 2018. 巻なし, 査読有。

Atsushi Miyauchi and Naonori Kakimura: Finding a dense subgraph with sparse cut, *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018)*, pp. 547–556. 巻なし, 査読有。

Atsushi Miyauchi, Tomohiro Sonobe, and Noriyoshi Sukegawa: Exact clustering via integer programming and maximum satisfiability, *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*, pp. 1387–1394. 巻なし, 査読有。

Yasushi Kawase and Atsushi Miyauchi: The densest subgraph problem with a convex/concave size function, *Algorithmica*, **80**, pp. 3461–3480, 2018. 査読有。

〔学会発表〕(計 10 件)

河瀬 康志, 宮内 敦史: サイズ関数を一般化した最密部分グラフ問題, 電子情報通信学会 2019 年総合大会 COMP 学生シンポジウム, 2019 年. 招待講演。

宮内 敦史, 武田 朗子: ロバスト最密部分グラフ抽出, 日本オペレーションズ・リサーチ学会 2019 年春季研究発表会, 2019 年。

宮内 敦史, 武田 朗子: Robust densest subgraph discovery, 情報系 Winter Festa Episode 4, 2018 年。

Atsushi Miyauchi and Akiko Takeda: Robust densest subgraph discovery, The 18th IEEE International Conference on Data Mining (ICDM 2018), 2018.

Atsushi Miyauchi: Very recent developments in dense subgraph discovery, Seminar at Algorithmic Data Analytics Group, ISI Foundation, 2018. 招待講演。

Atsushi Miyauchi and Naonori Kakimura: Finding a dense subgraph with sparse cut, The 27th ACM International Conference on Information and Knowledge Management (CIKM 2018), 2018.

宮内 敦史, 藺部 知大, 鮎川矩義: Exact clustering via integer programming and maximum satisfiability, 第 17 回情報科学技術フォーラム (FIT 2018), 2018 年。

Atsushi Miyauchi: Graph clustering with theoretical performance guarantees: Mathematical programming approaches, Seminar at the Department of Computer Science, School of Engineering, Virginia Commonwealth University, 2018. 招待講演。

Atsushi Miyauchi, Tomohiro Sonobe, and Noriyoshi Sukegawa: Exact clustering via integer programming and maximum satisfiability, The 32nd AAAI Conference on Artificial Intelligence (AAAI 2018), 2018.

宮内 敦史: ネットワーク科学における数理最適化: モデル化とアルゴリズム設計, ネットワーク科学セミナー 2017, 2017 年. 招待講演。

## 6 . 研究組織

科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。