

研究終了報告書

「グラフ構造を用いた自由記述データ処理に関する研究」

研究期間：2021 年 10 月～2023 年 8 月

研究者：川本 達郎

1. 研究のねらい

近年、AI 技術や機械学習技術は急速に社会の中で活用が進んでいる。特に 2023 年に入ってから LLM と呼ばれる言語ベースの技術が大変熱を帯びている。一方で、言語処理に関連するものであれば何でも LLM によって処理が可能になるというわけではない。

例えば、言葉のニュアンスを読み解かなければならない文章や、背景知識が必要となる文章を処理する場合には、人間であっても迷ったり解釈が分かれたりすることがある。このような問題は、現実社会においては、例えば記述式アンケートの分析において直面する。このような生身の人間が(時には誤った知識に基づいて)整合しない応答をする処理については、その人間の不整合な感性こそが「正解」であり、この状況を正確に分析することは容易ではない。この問題を言語処理のみを用いて克服するのは非常に困難と言える。

そこで本課題の研究者は、言語処理の代わりにグラフ構造を用いた自由記述データの分析処理フレームワークを数年前に提案した。これは、データ収集時に、文章データのみを集めるのではなく、文章間の関係(グラフデータ)も人間から収集するフレームワークである。しかし、グラフ構造の解析に変換しても、グラフ解析としての困難が残る。本研究では、グラフ構造を用いた自由記述データ処理をする上での技術的困難性を明らかにするため、理論的な知見を広げることを狙いとした。

2. 研究成果

(1)概要

本課題では、グラフ構造を用いた自由記述データ処理に関する研究として、【課題 1】グラフの構造推論における頂点の順序制約・順序付けの問題および、【課題 2】グラフの構造推論における摂動論に注目して研究を進めてきた。また、課題1の研究を進める過程で多くの派生研究が生まれた。

【課題 1】は、提案当初は自由記述データをグラフ処理によって整理(クラスタリング)する際に起こるアルゴリズム挙動の質的な転移について研究することを目的としていた。これは事前に数値実験によって、グラフの頂点の生成順序に起因する現象だということは判明していた。これを理論的に解析するためには、問題設定をより精密化していく必要が生まれ、グラフの頂点順序とクラスタリングの間の整合性について数学的な整備をする研究(1-1)を行い、論文1本を出版した。数学的な整備を終えた後、本題の順序制約に起因するアルゴリズム挙動の質的な転移の研究を行った。しかし、非常にシンプルな解析により、ある程度一般性のある解が得られたものの、実問題に使用するようなモンテカルロ法での挙動の質的転移とは大きくずれており、実際の現象を説明する理論としての位置付けには達していないことが分かった。一方で、このグラフの頂点順序に関する研究は、大元の「自由記述データをグラフ処理によって整理する」というゴールの下で有益な派生研究を生み出し、クラスタリングアルゴリズムに関する

論文2本(1-2, 1-3)、ランダムグラフ理論に関係する論文を1本(1-4)出版するに至った。

【課題 2】は、記述データをグラフ処理する際、グラフ化したデータが持つ情報量が不足している場合にはアノテーション処理が必要となる状況を想定し、一般にグラフデータに摂動がかけられた場合のアルゴリズムの挙動の変化を研究することを目的としていた。以前に実際のアノテーション過程そのものに対応する解析をしようとした結果、あまり定性的な理解が得られない問題に直面したことから、少し問題を変えて、摂動論についての研究を行い、1本の論文を出版することができた。

その他、人工知能学会誌に【課題 1】の成果に関する解説記事を寄稿、5件の口頭発表を行った。

(2) 詳細

【課題 1】 グラフの構造推論における頂点の順序制約・順序付けの問題および派生研究

(1-1) グラフの頂点の順序づけとグループ分けについての整合性評価研究

グラフの頂点順序と、クラスタリングアルゴリズム等によって得られた推定頂点ラベルが連関を持っているのかどうかは本課題の主たる研究対象であるが、そもそもこの連関をどう定量化するかという問題がある(図1)。この論文では、Label Continuity Error (LCE) という指標を提案し、LCE に関する様々な上限値・下限値・統計値を調べた。また、これまで独立して研究が進められてきた順序推定の問題とクラスタリング問題の間の整合性を、LCE を用いて調べた。

さらに付加的な成果として、より研究が進んでいるクラスタリングアルゴリズムを参照し、対応する順序推定アルゴリズムを定式化した。興味深いことに、元のクラスタリングアルゴリズムにはハイパーパラメータの物理的意味は存在しなかったのに対し、順序推定アルゴリズムとして定式化した場合、同じハイパーパラメータには物理的意味が存在することを発見した。

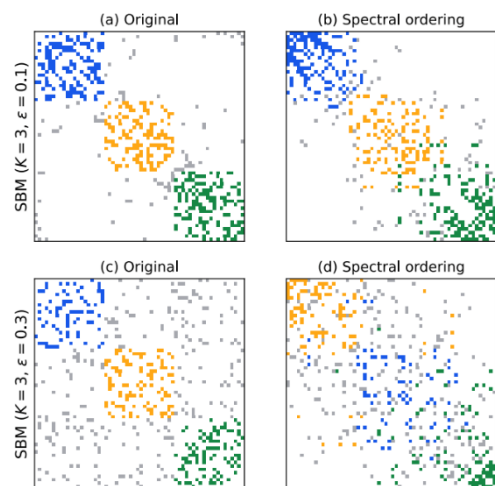


図1 並び替え前後のグラフの隣接行列。配色はグループラベルを反映している。色がはっきり分かれているほど、順序とラベルは整合しているといえる。

(1-2) グラフの頂点の順序推定アルゴリズムの提案

グラフの頂点順序推定アルゴリズム研究は古くからあるが、あまり研究者人口は多くない。近年進展が著しいブロックモデルを用いた統計推論の手法は、頂点順序推定問題においても有効である可能性があり、アルゴリズム提案を行った。頂点順序の推論は計算量的に負荷が大きく、計算量を削減する様々な近似手法を導入したが、クラスタリングアルゴリズムほどの計算量を達成するのは困難であることが明らかになった。このアプローチは昨今のグラフの統計推論アルゴリズムの型を踏襲しており、これがまだまだ発展させる余地があるという報告は、業界にとって新しい問題提起となると考えている。その意味で、本課題の当初の目的とは少し

異なるが、業界への貢献という意味で重要であると考えている。

(1-3) クラスタリングアルゴリズムの提案

グラフの頂点順序依存性の研究を進める中で、グラフのクラスタリングアルゴリズムの中でも最も有名な手法の一つである map equation の方法には、原論文で提案されているものよりも素朴な定式化が可能であることを発見し、それを single-trajectory map equation として提案した。この素朴な定式化は、これまで map equation が抱えていたオーバーフィットの問題を自然に改善する機構を持っていることを指摘し(図2)、人工データと実データを用いて確かに性能が劇的に改善されることを示した。さらに、resolution limit と呼ばれる理論限界が改善されることを示し、理論的な裏付けも明らかにした。

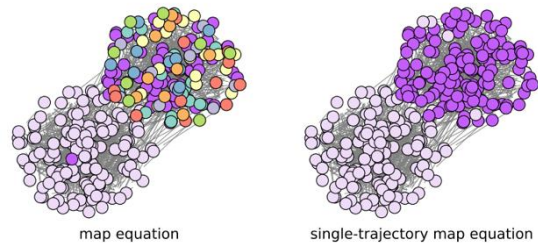


図 1 2つのグループが正解として用意されている人工グラフについて、提案手法は正しく2グループの存在を検出できている。

(1-4) ランダムグラフ理論に関する研究

グラフの頂点順序問題を研究する中で、グラフの組合せ論について、物理学的な計算手法を開拓できる可能性に気づき、この研究を行った。ランダムグラフ理論では、「ランダムグラフモデルの持つ場合の数」を求める研究が 1970 年代後半から 1990 年代前半にかけて興味が集まった。過去の一連の研究で見つかった漸近公式として有名なものに、Bender-Canfield 公式というものがあるが、これを統計力学的な計算アプローチと Hermite 多項式の展開を用いて導出できることを発見した。さらに、この手法によって得られた推定値は、非漸近領域においては Bender-Canfield 公式よりも精度が高い(実際の場合の数に近い)ことをいくつかのケースで数値計算によって示した。ここで提案した方法は、ランダムグラフのみならず、有限離散系の統計力学研究にも発展させられる可能性があると考えている。

【課題 2】グラフの構造推論における摂動論

グラフの構造推論の手法の一つとして、グラフの固有値・固有ベクトルを調べるものがある。この研究では、摂動として無向グラフの枝にランダムな向き付けを行った場合に、グラフのスペクトルがどのような影響を受けるか、統計的な解析を行った。これは、元データが有向グラフであった場合に、グラフの枝の向き情報を無視した場合にどのような問題が生じるかという逆の問題設定にも示唆を与えてくれるものとなっている。グラフのランダム有向化についての解析は、グラフに大域的な制約がない場合には比較的簡単に計算できることが明らかになった。ここでの成果を足がかりに、枝の追加による摂動がかかった場合についても研究を拡張したが、課題1の方面での進展が好調であったため、そちらの研究に注力するために、課題2の研究はここで保留にする方針をとった。

3. 今後の展開

本課題の当初の課題の一つである「順序制約がクラスタリング性能に与える影響」に対応

するには、クラスタリングアルゴリズムのほうに順序制約の影響を加味するように変更を加えるというのがより直接的な進展かもしれない。研究としてのみこの問題に取り組むのであればそれが一番かもしれないが、アルゴリズムの複雑化を行うと一般に挙動が不安定になる傾向があり、実務においては“絵に描いた餅”になる恐れがある。研究を超えて現実社会への貢献を考えると、やはりデータから順序制約の影響そのものを理解する方向性を追求するのが好ましい。ただ、現状のシンプルな理論解析を洗練していくという展開は、その末にどれほど有益な知見が得られるか不明瞭であるため、理論解析の道のりは依然として険しい。今後も紆余曲折を繰り返し、着地するまでには 10 年単位の時間がかかる可能性がある。

グラフの頂点順序依存性の研究から派生した研究については、今後多くの展開が期待される。特にランダムグラフ理論に関する研究は、有限離散系という、あまりこれまで研究されてこなかったと思われる領域に、今後少しずつ、様々な理論的展開を見出せると期待している。

最後に、当初のもう一つの課題である「グラフの摂動論」に関しては、エフォート配分の都合上あまり進めることはできなかったが、研究を進めていくことでアノテーションによる摂動論についての研究も進展させることができると考えているため、ぜひ今後も取り組んでいきたい。

4. 自己評価

研究目的の達成状況: グラフ構造のデータ処理という広い視点においては、比較的多くの成果を出せたと言える。課題提案当初の計画自体については、ある程度目的を達成できたと考えているが、完全解決というところまでは至っていない。その代わりに、研究を進める上で得た気づきを基に、様々な方向へ研究の枝葉が伸びていった。その枝葉の実りが多く、結果として多数の成果を論文出版という形で出すことができた。グラフ構造を用いて自由記述データを処理するという課題は、決して特定の課題を一つずつ順番に解決していく性質のものではなく、あらゆるアプローチを検討しながら、そこで生じる各々の問題を解決していくものだと考えている。その意味で、本課題の一連の研究によってグラフの構造推論の視野を広げることができたことは、大きな進展であったと捉えている。

研究の進め方: 研究は、神戸大の小林氏、東大博士課程の越智氏と共同研究する形で進めることができた。本課題の研究者が主導となって進めるテーマも多かったため、柔軟に研究計画を立てながら進めることができた。研究期間の初期は、まだコロナ禍の影響があり、あまり出張等ができなかったため、予算執行ができない状況であったが、最後の半年間は、いくつか出張をし、講演やヒアリングをすることができた。特に 2023 年にタイで行われた小規模ワークショップでは、ランダムグラフ理論の研究について海外の研究者と多くの交流が生まれ、まだ研究成果には結びついていないが、今後の研究体制の新しい基礎を築くことができた。また、ACT-X 内部でも、領域会議をきっかけに若手の研究者とオンラインや対面で議論が生まれた。

研究成果の科学技術及び社会・経済への波及効果: 本課題の研究成果は理論研究がメインであったため、主にグラフ解析の研究業界への貢献が大きいと考えられる。一方で、ゆくゆくは記述データ処理の基礎として機能すると期待できる。社会調査やマーケティング調査の中で、これまでにはできなかったような分析が可能となれば、様々な社会・経済への波及効果が生まれるだろう。

5. 主な研究成果リスト

(1) 代表的な論文(原著論文)発表

研究期間累積件数: 5件

1. Tatsuro Kawamoto, Masaki Ochi, and Teruyoshi Kobayashi, "Consistency between ordering and clustering methods for graphs", Phys. Rev. Research 5, 023006 (2023).
グラフの頂点順序と頂点(クラスタリング)ラベルの間の整合性について Label Continuity Error (LCE) という指標を提案し、その性質を調べたり、実際にアルゴリズム評価に活用したりした。
2. Tatsuro Kawamoto, "Single-trajectory map equation", Sci. Rep. 13, 6597 (2023).
グラフのクラスタリングアルゴリズムとして有名な map equation の手法の変種を提案。Map equation が抱えていたオーバフィット問題を軽減する手法であることを示した。
3. Tatsuro Kawamoto, "Entropy of microcanonical finite-graph ensembles", J. Phys. Complex. 4 035005 (2023).
ランダムグラフ理論で知られる Bender-Canfield 式を、統計力学的なアプローチによって導出できることを示した。

(2) 特許出願

研究期間全出願件数: 0 件(特許公開前のは件数にのみ含む)

(3) その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

【主要な共同研究による原著論文】

- Masaki Ochi and Tatsuro Kawamoto, "Contribution of directedness in graph spectra", Phys. Rev. Research 4, 033129 (2022)
- Masaki Ochi and Tatsuro Kawamoto, "Finding community structure using the ordered random graph model", Phys. Rev. E 108, 014303 (2023)

【主要な学会発表】

- 川本達郎, 越智昌毅, 小林照義, "スペクトル法によるグラフの頂点順序最適化とクラスタリングとの整合性", 日本物理学会 2022 年秋季大会
- 川本達郎, "ランダムグラフのミクロカノニカルエントロピーの統計力学的導出", 日本物理学会 2023 年春季大会
- Tatsuro Kawamoto, Teruyoshi Kobayashi, "Hypothesis testing of adjacency matrices", 9th International Discussion Meeting on Relaxations in Complex Systems (9 IDMRCS)

【著作物】

- 川本 達郎, 小林照義, 「グラフにおける頂点順序並び替え問題」, 人工知能学会誌 Vol.38 No.2 (2023 年 3 月)