氏　　名　安井 雄一郎

学位(専攻分野)　博士(統計科学)

学 位 記 番 号　総研大甲第 2403 号

学位授与の日付　2023 年 3 月 24 日

学位授与の要件　複合科学研究科 統計科学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　科学技術文献における引用ネットワークに対する確率生成モデル

論 文 審 査 委 員　主　査　南 和宏
統計科学専攻 教授
栗木 哲
統計科学専攻 教授
金藤 浩司
統計科学専攻 教授
脇田 建
東京工業大学 情報理工学院 准教授

（様式３）

博士論文の要旨

氏名　安井　雄一郎

論文題目　科学技術文献における引用ネットワークに対する確率生成モデル

本論文では学術論文や特許文献の引用ネットワーク構造に対する確率的生成モデルを提案する。学術論文や特許文献の引用構造は、文献をノードに、文献間の引用構造を有向エッジに対応させた有向グラフで表現することができる。またノードには文献の発表時刻に対応する離散時刻が付与される。本研究では対象の引用ネットワークに対して簡素な生成プロセスを定義し、シミュレーションを用いた適合を検証することで、実データがもつ性質を明らかにすることを目的とする。このように実問題に対応するネットワーク構造を理解することで、対象の実問題の理解を深めることは、様々な先行研究で用いられる典型的なネットワーク科学の枠組みである。

本論文で提案するモデルは文献間の引用が、引用する文献の種類、引用された文献の重要度、両者の発表時刻の差にもとづく確率で発生することを想定する。ロジスティック関数に従う離散時間における文献数、逆ガウス確率分布の確率密度関数に従う発表時刻差ごとの引用のされやすさ、一般化パレート分布（または指数分布）に従う各文献における参考文献の件数などにもとづき、ネットワーク構造を生成する。その際にエッジの生成は、重要度に応じてノードを選択する優先的選択（Preferential attachment; PA）機構と、選択済みノードの隣接ノードを選択し三角形型の引用構造を構築する三角形形成（Triad formation; TF）機構を組み合わせる。なお文献の重要度と種類は、ノードの入次数と出次数で近似されるものとする。

まず学術論文の書誌データベース Web of Science における確率統計分野の文献から構築された引用ネットワークを対象とする。提案モデルは離散時刻ごとの文献数にもとづきノードを生成し、一定期間さかのぼった引用構造を対象にモデリングした時刻に調整された文献の年齢分布や出次数分布にもとづきエッジの生成を行う。シミュレーションにより、引用ネットワークで重要な特徴量となる、引用数の件数を表す入次数分布、被引用数の件数を表す出次数分布、三角形型の引用構造の件数を表す三角形数の分布に関して、モデルの適合を示した。さらに arXiv の書誌データベースにおける高エネルギー物理分野の文献から構築された二種類の引用ネットワークを対象に、モデル適合を検証した。これは既存モデルでも検証に用いられる機会が多いネットワークである。また提案モデルは既存モデルと比較して、各ノードがもつ離散時刻を明示的に扱えること、実行時にデータを必要としない完全なシミュレーションを実現できることが利点である。

つづいて特許文献の引用構造から構築された引用ネットワークを対象とする。特許文献の引用ネットワークは各ノードがカテゴリやサブカテゴリにより分類され、階層的なクラスタ構造を形成している。カテゴリ内やサブカテゴリ内の引用は多く、カテゴリ間やサブカテゴリ間の引用は少ない。これは提案モデルの想定とは異なるため、あるサブカテゴリに着目して生成モデルの検討を進めたものの、三角形数の分布に対する適合に課題があることが明らかになった。データに対して提案モデルは多くの三角形を生成するノードが少ない傾向となる。

そこで提案モデルの TF 機構に対する二種類の拡張を提案する。一点目としては TF 機構が選択するノードの候補の拡張である。提案モデルでは直前の PA 機構で選択されたノードの隣接ノードを候補に設定していた。これを基準となるノードから PA 機構により選択されたノード全ての隣接ノードへ候補を拡大した。二点目としては TF 機構を実行する割合に関する拡張である。提案モデルの実行割合は定数パラメータで指定していたが、ノードごとに柔軟に設定できるように拡張を行った。まず各ノードにおける TF 機構の実行割合を予測するためのアルゴリズムを構成する。アルゴリズムは対象のネットワーク構造における各エッジが PA 機構もしくは TF 機構のどちらかで生成されたと仮定する。予測された TF 機構の実行割合の分布を観測することで、0 や 1 が過剰に存在すること、実行割合は実際によりも小さめに予測していること、ノードの出次数が大きいときに実行割合が大きくなることなどの特徴を有していることが明らかになった。そこで TF 機構の実行回数をゼロワン過剰ベータ分布でモデル化を行い、ある程度大きな出次数をもつノードでは実行割合を 1 に固定する。これらの拡張を適用した提案モデルは、特許文献の引用ネットワークや学術文献

の引用ネットワークのどちらに対しても十分に適合することを確認できた。

本研究の貢献は主に引用ネットワークの各文献がもつ離散時間の考慮、シミュレーションからデータの直接的な使用の排除、ノードごとに異なる三角形形成機構の実行割合の制御である。改善モデルは 2 ノード間に関する特徴である次数、3 ノード間に関する特徴である三角形の性質はモデリングできたものの、ネットワークに関する広域な特徴量である Scree plot へのあてはまりは依然として十分でない。また引用ネットワーク全体を対象としたときに必要な階層構造のモデリングには至っていない。これらが残された課題として挙げられる。

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis defense

博士論文審査結果

Name in Full
氏 名　　安井　雄一郎

Title
論文題目　　科学技術文献における引用ネットワークに対する確率生成モデル

安井雄一郎氏の博士論文審査を,2023 年 1 月 16 日 13 時 30 分から約２時間にわたって,本人および 4 名の委員全員の出席のもとに行った．論文発表および審査の結果，審査委員会では、本論文が学位の授与に値すると判断した．

［論文の概要］

論文は５章86ページからなり，日本語で書かれている．本論文の目的は学術論文や特許文献などの科学技術文献から構築された引用ネットワークに対する確率生成モデルを提案し，科学技術文献における引用構造の生成プロセスを明らかにすることである．

第１章では，科学技術文献の引用構造は文献をノードに文献間の引用構造を有向エッジに対応させた有向グラフで表現することができること，各ノードには文献の発表時刻に対応する離散時刻が付与されること，有向エッジが表現する引用構造には引用元に対応するノードは引用先に対応するノードに比べて同じかそれよりも新しいということなどの性質を述べている．

第２章では，これまでに提案されている一般的なネットワーク生成モデルと引用ネットワーク構造を対象としたネットワーク生成モデル（Barabási and Albert (1999), Wu and Holme (2009)等）を紹介している．また，ネットワーク間の直接的な比較は難しいため，主として解釈の容易な引用ネットワーク構造の３つの特徴量（入次数分布，出次数分布，三角形数分布）を用いてネットワークの比較を行うことを述べている．

第３章では，Wu and Holme (2009)のモデルを拡張し，ロジスティック分布に従うと仮定した時刻ごとの文献数，逆ガウス分布に従うと仮定した発表時刻差ごとの引用のされやすさ，指数分布に従うと仮定した各文献の参考文献の件数の分布に基づきネットワーク構造を生成し,その際のエッジの生成は,重要度に応じてノードを選択する優先的選択構造と,選択済みノードの隣接のノードを選択して三角形型引用構造を構築する三角形形成機構を組み合わせて確率生成モデルを構築している．本論文では，Web of Science の 1981 年から 2016 年までの 36 年分の文献データから確率統計分野に限定した 179,483 文献（1,106,622 引用）を抽出し，対象の引用ネットワークとした．また，本データセットは,データベースを所有する Clarivate Analytics 社の許可を得てデータリポジトリの Dryad にて公開している.同時に,先行研究においてモデルの適合の検証に多く用いられる arXiv の高エネルギー物理分野から生成した引用ネットワークに対しての提案モデルの有効性も検証している．

第４章では，第３章で提案した確率生成モデルをネットワーク構造が異なる National Bureau of Economic Research で公開されている米国特許データへの適用を行い，第３章

で提案したモデルの拡張をおこなった．第５章は，まとめを述べている．

［論文の評価］

科学技術文献には学術論文，特許文献，その他いろいろな関連文献がある．本論文における提案手法は，大規模な科学技術文献の引用構造を有向グラフとしてのネットーワーク構造で表現することにより，ネットワーク構造に対する生成モデルを構築し，３つの特徴量と２種類の生成機構を用いることで，そのネットワーク生成プロセスを理解するために重要な役割を果たすとともに,統計科学の博士論文として十分な意義を持つと判断される．なお第３章の内容は，査読付き英文学術誌 PLOS ONE に掲載されている．