固有値分解とテンソル分解を用いた大規模グラフデータ分析に関する研究

A Study on Large Scale Graph Analysis Using Eigen Decomposition and Tensor Decomposition

丸橋 弘治 ♥

Koji MARUHASHI

ネットワークトラフィックログやソーシャルネットワークなど、大規模グラフデータとして分析できるデータは多い。近年、グラフデータから、異常あるいは特徴的な部分構造を検知する手法がいくつか提案されているが、データ全体の構造における位置づけが考慮に入っていないという問題があった。本博士論文では、主要な構造との位置づけが特徴的な部分構造を検知することを目標とし、要素間の関係が行列あるいはテンソルで表現できるグラフデータに対し、固有値分解やテンソル分解を利用した、いくつかの新たな手法を提案する。まず、**(1)**k 一様な k 部ハイパーグラフから、主要な構造から離れ、かつ互いに密に関係しあった比較的大きなノード集合を高速に検知する手法を提案する。また、**(2)** 2部グラフから、複数の主要な構造への経路数の分布が特徴的なノード集合を効率的に発見するための手法を提案する。さらに、**(3)** 平均経路長が小さい無向1部グラフに対し、2ノード間の最短経路長を高速に概算する新たな手法を提案する。

Many data can be analyzed as large-scale graph data such as network traffic logs and social networks. The main motivation of this thesis is to detect anomalous patterns related to path capacities within whole graph structures, which have not been taken into account by any of existing works. We focus on graph data whose adjacency information can be represented as matrices or tensors. Our methods leveraged by eigen decomposition and tensor decomposition can: (1) detect *local* and *remarkable* clique-like structures in a k-partite k-uniform hypergraphs, (2) spot *patterns* of nodes related to path capacities to community structures in a bipartite grpah, and (3) estimate shortest-path distances of nodes more accurately in an undirected graph with short average distance.

♥ 株式会社富士通研究所 maruhashi.koji@jp.fujitsu.com

1. はじめに

近年の情報管理技術やセンシング技術の発達により、社会のあらゆる場所でデータが蓄積される状況になっている。その中でも、ネットワークトラフィックログやソーシャルネットワークなど、多数の要素（ノード）の関係（エッジ）の集合、すなわち大規模なグラフデータとして分析できるデータは多い。このようなグラフデータに対し、コミュニティの抽出や、リンク予測、重要ノードの抽出など、様々な分析手法が提案されてきた。特に、異常あるいは特徴的な部分構造の抽出は、ネットワークの侵入検知やクレジットカードの利用履歴からの不正検知など、重要な問題である。この問題に対する有望な手法が、比較的近年になっていくつか提案された。例えば、最小記述長原理を利用し、頻出構造では記述できない部分構造を検知する手法 [1] や、隣接行列の行列分解による近似誤差を異常として検知する手法 [2] が提案されている。さらに、各ノードの近傍グラフのノード数・エッジ数などの特徴量の多くが従う冪乗則から、大きく外れたノードを異常として検知する手法が提案されている [3]。しかしながら、これらの提案手法は、データ全体の構造における部分構造の位置付けが考慮に入っていない。

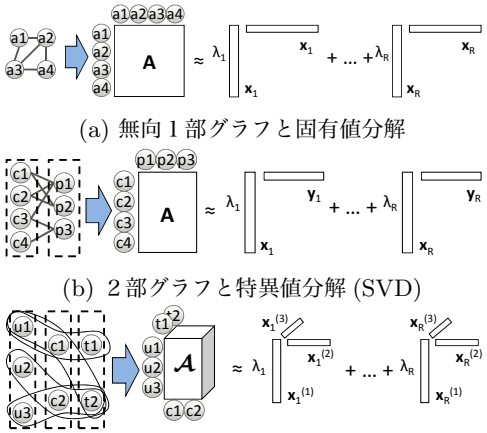
本博士論文では、ノード間の接続関係が行列あるいはテンソルで表現できるような、静的でラベル無しのグラフデータに着目し、データ全体の構造における位置づけを考慮した、特徴的な部分構造を抽出する問題に取り組む。このような行列やテンソルに対し、固有値分解や特異値分解、テンソル分解といった手法を適用することにより、各ノードに対応した複数のスコアを計算できる。これらのスコアは、データ全体で主要な構造との間の経路数に関する、いくつかの共通の性質を持つ。我々は、これらの性質を利用した、部分構造の検知と経路分析に関する、新たな手法を提案する。まず、(1)k 一様な k 部ハイパーグラフ（k,k-ハイパーグラフ）から、主要な構造から離れ、かつ互いに密に関係しあった比較的大きなノード集合を検知する手法を提案する。また、(2) 2部グラフから、複数の主要な構造への経路数の分布が特徴的なノード集合を検知する手法を提案する。さらに、(3) 無向1部グラフに対し、2ノード間の最短経路長を高速に概算する新たな手法を提案する。

なお、本稿では、ベクトルを小文字の太字 $\mathbf{x}$、行列を大文字の太字 $\mathbf{X}$、テンソルを大文字の太字カリグラフィー $\mathcal{X}$ で表す。$\mathbf{x}$ の第 i 要素は x_i、$\mathbf{X}$ の第 i,j 要素は x_{ij}、$\mathcal{X}$ の第 $(i_1, \ldots, i_k)$ 要素は $x_{i_1 \cdots i_k}$ で表す。また、ベクトルの外積を $\times$ で表す。

以降、2. で上述のスコアの共通の性質について述べる。続いて、3. で k,k-ハイパーグラフの特徴的な構造の検知、4. で2部グラフの特徴的な構造の検知、5. で無向1部グラフのノード間最短経路長推定を説明する。6. にて全体のまとめを述べる。

2. 行列・テンソル分解の共通性質

無向1部グラフ、2部グラフ、k,k-ハイパーグラフは、ノード間の接続関係を行列あるいはテンソルで表現できる（図 1）。ここで、k,k-ハイパーグラフとは、全てのノードが k 個の区画に



(a) 無向 1 部グラフと固有値分解

(b) 2 部グラフと特異値分解 (SVD)

(c) k,k-ハイパーグラフと CP 分解

図 1　グラフデータと行列・テンソル分解

分かれ、全てのエッジに互いに区画が異なる k 個のノードを含むハイパーグラフである。これらの行列・テンソルを、2 乗誤差が最小になるような、複数のベクトルの外積の和で近似することを、本稿では低ランク近似と呼ぶ。無向 1 部グラフ（ノード数 I）、2 部グラフ（ノード数 I_1, I_2）、k,k-ハイパーグラフ（ノード数 $I_1, \ldots, I_k$）の接続関係は、それぞれ、$I \times I$ の対称行列 $\mathbf{A}$、$I_1 \times I_2$ の非対称行列 $\mathbf{A}$、$I_1 \times \ldots \times I_k$ のテンソル $\mathcal{A}$ で表現できる。いずれも、エッジに対応する要素は 1（あるいは重み）、それ以外は 0 である。これらの行列・テンソルのランク R の低ランク近似は、それぞれ、2 乗誤差

$$\left\| \mathbf{A} - \sum_{r=1}^{R} \lambda_r \left(\mathbf{x}_r \times \mathbf{x}_r \right) \right\| \tag{1}$$

$$\left\| \mathbf{A} - \sum_{r=1}^{R} \lambda_r \left(\mathbf{x}_r \times \mathbf{y}_r \right) \right\| \tag{2}$$

$$\left\| \mathcal{A} - \sum_{r=1}^{R} \lambda_r \left(\mathbf{x}_r^{(1)} \times \ldots \times \mathbf{x}_r^{(k)} \right) \right\| \tag{3}$$

を最小にするような、長さ 1 のベクトル $\mathbf{x}_r$、$\mathbf{x}_r, \mathbf{y}_r$、$\mathbf{x}_r^{(1)}, \ldots, \mathbf{x}_r^{(k)}$ と、スカラー λ_r を求めることである（$r = 1, \ldots R$）（図 1）。これらは、それぞれ、固有値分解、特異値分解 (SVD)、および主要なテンソル分解である CANDECOMP / PARAFAC (CP) 分解 [4] に相当する。

本稿では、低ランク近似で得られた、ノードに対応する各ベクトルの要素の値を、ノードのスコアと呼ぶ。ノードのスコアは、接続されているノードのスコアの線形和になる。例えば、2 部グラフのノードの、特異値分解によるスコアは、

$$\begin{cases} x_{ri} = \lambda_r^{-1} \sum_{j \in C_i} a_{ij} y_{rj}, \\ y_{rj} = \lambda_r^{-1} \sum_{i \in D_j} a_{ij} x_{ri} \end{cases} \tag{4}$$

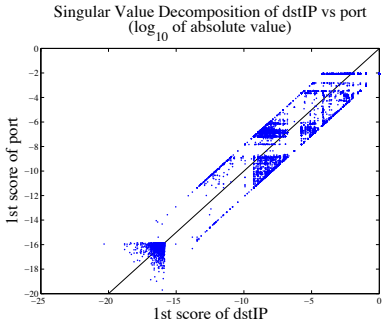


図 2　2 部グラフのエッジの、両端ノードのスコアの絶対値によるプロット（両対数）。横軸：左ノード（送信先 IP）。縦軸：右ノード（ポート番号）。

となる。ここで、C_i,D_j は、ノード i,j と接続されているノード集合である。このことから、いずれの場合でも、ノードが直接接続されている場合には、互いのスコアの絶対値の比が所定の範囲内に収まることが多いことを示すことができる（詳細は博士論文を参照）。実例として、図 2 は、通信ログにおける送信先 IP とポート番号の 2 部グラフに対し、各エッジを、両端ノードのスコアの絶対値により、両対数でプロットしたものである。図 2 は、接続されているノードのスコアの絶対値の比が、所定の範囲に収まっていることを示している。類似の性質は、無向 1 部グラフや k,k-ハイパーグラフのスコアでも観察することができる。この性質は、主要な構造のノード、すなわちスコアの絶対値が大きいノードから離れるほど、スコアの絶対値がほぼ一定の比率で減衰していくことを示している。本研究では、この性質を利用した、グラフデータの新たな分析手法を提案する。

3.　テンソル分解による特徴的な構造の検知

Web のリンク構造を利用する検索エンジンに対し、特定のコンテンツが上位に検索されやすいように、ソーシャルブックマークで特定のコンテンツ群を複数のアカウントから故意に大量に登録するといった不正が発生している。また、サーバのトラフィックを増大させてサービスを不能とする DoS 攻撃のうち、攻撃を行うマシンや時刻などを分散させることで対策を困難にする DDoS 攻撃が問題となっている。これらの不正に共通することは、比較的マイナーなコンテンツや I P アドレスといった要素の間で、相互に密な関係を形成することである。コンテンツなどの要素間の関係は、k,k-ハイパーグラフとして表現できる。そして、上述のような相互に密な関係を検知することは、主要な構造のノードと長い経路で隔てられた、クリーク状の構造を検知することに相当する。ここで、クリーク状の構造とは、k 個の区画にまたがるノードの集合で、区画の異なるほぼ全てのノードの組み合わせに対してエッジが存在するような構造を指す。テンソル分解の単純な適用により、このような構造を分離することができる。例えば、CP 分解 [4] によるスコアの絶対値が大きいノードを抽出する。しかし、この単純な方法では、主要な構造に含まれ

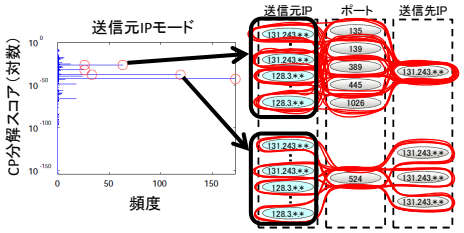


図 3　CP 分解の送信元 IP モードのスコアのヒストグラムと、スパイクから検知されたクリーク状の構造。

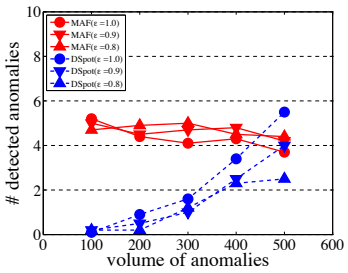


図 4　検知できた人工構造（最大 10 個）の、10 回の試行の平均値。横軸：人工構造の体積（ノードの全組み合わせ数）。縦軸：平均検知数。赤の実線：提案手法。青の点線：単純手法。形状が違うマーカーは同体積の人工構造のエッジ数の違いを表す。

る構造しか検知できない。

2. で説明したとおり、CP 分解によるスコアは、主要な構造から離れるほど、一定の比率で減衰する傾向がある。また、クリーク状の構造に含まれるノードは互いに接続されていることから、スコアの絶対値の比は所定の範囲内に収まる傾向がある。そこで我々は、k,k-ハイパーグラフの各ノードを、CP 分解によるスコアの対数が等間隔となる階級により集計し、そのヒストグラムのスパイクを検知することにより、主要な構造から離れたクリーク状の構造を高速に検知する手法を提案する [5, 6]。図 3 は、通信ログにおける送信元ＩＰとポート番号、送信先ＩＰの間の 3,3-ハイパーグラフから、CP 分解の送信元 IP モードのスコアのスパイクを抽出することにより、主要な構造から離れたクリーク状の構造を検知した例である。

また、図 4 は、通信ログに対し、ランダムに値を選んで作成した人工的なクリーク状の構造を追加し、検知できた人工構造の数を、提案手法と上述の CP 分解による単純手法の間で比較評価したものである。提案手法は、比較的小さい人工構造でも精度よく検知しているのに対し、単純手法は、小さい人工構造の検知精度が著しく悪いことがわかる。これは、人工構造が小さくなるほど主要な構造から離れた構造となり、単純手法では検知が困難になるためであると考えられる。

4.　特異値分解による特徴的な構造の検知

互いに似たノードと接続されたノード集合は、何らかの共通した性質を持った要素の集合を表すことが多い。本稿では、このようなノード集合を、コミュニティと呼ぶ。グラフデータにおいて、このようなコミュニティとの関係が異常なノードを発見することは重要である。例えば、スーパーマーケットの購買データにおいて、似た顧客に購入される商品のコミュニティの傾向、例えば各顧客の総購買数に対し乳製品の購買数が一定の割合を占める傾向が見出されたとき、総購買数に対し異常に多くの乳製品を購買する顧客は、重点的に販促活動を行う対象としたり、異常な顧客として集計から省いたりすることが考えられる。しかし、一般にコミュニティは無数に存在するため、このような異常検知に値する傾向を有するコミュニティを見出すことは困難である。また、周辺のノードとの接続関係の傾向を調べる従来手法 [3] は、コミュニティとの関係の傾向を調べるものではない。

ここで、各ノードのスコアの絶対値は、より主要な構造に近いコミュニティ内のノードとの接続数に比例する。これは、2. で述べたように、ノードのスコアは接続されているノードのスコアの線形和である（式 4）ことと、スコアの絶対値は主要な構造との距離により一定の比率で減衰すること、類似のノードと接続されているコミュニティ内のノードはスコアの絶対値の比が一定の範囲に収まる傾向があることにより、説明できる。さらに、ランクにより、スコアの絶対値が大きい主要な構造は異なる。我々は、これらの性質を利用し、２部グラフに対し、無数に存在するコミュニティの中から、上述のような傾向を持つコミュニティの存在を一度に検知するための、２つの手法を提案する [7]。１つ目の手法は、全ノードを、ノードの次数と特異値分解によるスコアによりプロットしたときの、顕著な直線状の分布を検出する。図 5(a) は、映画-俳優データセットから、＊で示した映画コミュニティへの出演数が総出演映画数に比例する傾向を持つ俳優群（上側）と、総出演映画数に関わらず一定である傾向を持つ俳優群（下側）の存在を見出した例である。２つ目の手法は、特異値分解による各ノードのスコアを次数で割った、正規化スコアを考える。そして、全ノードを、２つのランクの正規化スコアによりプロットしたときの、顕著な直線状の分布を検出する。図 5(b) は、ネットワークトラフィックログから、異なる２つのポート番号コミュニティに対し、互いに拮抗したポート番号の割合で通信される傾向のあるサーバ群の存在を見出した例である。

5.　固有値分解による経路分析

無向 1 部グラフの代表例である、人間関係を表すソーシャルネットワークにおいて、キーワードによる人物検索の結果を、自身との最短経路長によりソートするといったアプリケーションが考えられる。そのためには、特定の２つのノード間の最短経路長を、マイクロ秒単位で計算することが求められる。しかし、現在よく分析される数百万以上のノードを含むデータでは、わずか数 hop 離れたノード間の特徴量の計算でも、一般的なデスクトップＰＣ環境でミリ秒から秒オーダーの計算時間がかかってしまうという問題がある。これは、極端に多くのノードと隣接関係を持つハブノードの存在が主な原因と考えられる。

この問題に対する従来技術として、少数の Landmark ノード

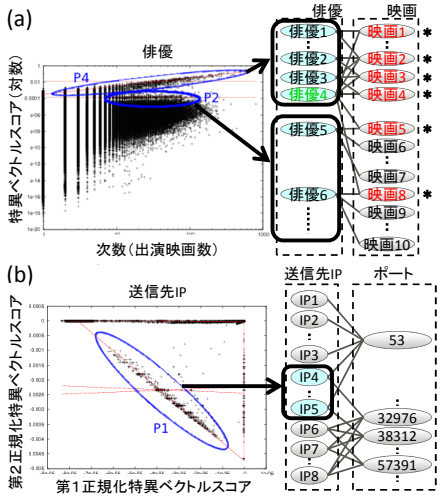


図 5 (a) 俳優ノードの、総出演映画数（横軸）とスコア（縦軸）、検出された傾向。(b) サーバノードの、第 1 正規化スコア（横軸）と第 2 正規化スコア（縦軸）、検出された傾向。

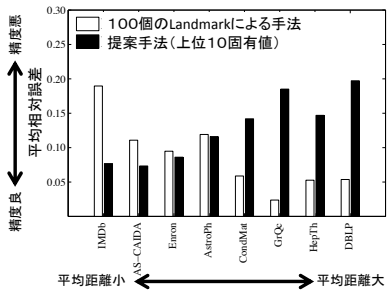


図 6 ノード間最短経路長の推定精度。横軸はデータセット。

を経由した経路長を概算値とする方法がある [8]。しかし、少数の Landmark ノードでは短い最短経路の多くをカバーできないため、近い距離において誤差が大きくなるという問題があった。我々は、固有値分解を利用して、ノード間の最短経路長を高速に概算する手法を提案する [9]。ノード間の接続関係を表す行列 $\mathbf{A}$ の k 乗 $\mathbf{A}^k$ の要素 $a_{ij}^{(k)}$ は、i 番目と j 番目のノードの間の k-hop 経路数を表す。そして、$a_{ij}^{(k-1)} = 0$ かつ $a_{ij}^{(k)} > 0$ ならば、これらのノード間の最短経路長は k と結論づけられる。我々の提案手法では、式 1 を最小にする $\mathbf{A}$ の近似行列 $\mathbf{A}' = \sum_{r=1}^{R} \lambda_r (\mathbf{x}_r \times \mathbf{x}_r)$ の k 乗 $\mathbf{A}'^k$ の要素 $a'^{(k)}_{ij}$ を、小さい k から順に計算し、閾値を超えた k を最短経路長として推定する。図 6 は、平均距離の小さいデータにおいて、Landmark 法よりも提案手法の方が精度が良いことを示している。

6. 課題と今後の展望

本研究では、静的なグラフデータに着目し、全体の構造において特徴的な部分構造を検知する手法を提案した。しかし、現実的には、さらに、時間変化のある動的なグラフデータから、これまでにない特徴を持った部分構造の発生を検知することが大きな課題となる。また、含まれる情報の粒度が異なるデータや、性質の全く異なるデータを組み合わせたときの、特徴的な部分構造を検知することも重要な課題である。今後は、今回提案した手法の改善と共に、これらの課題に取り組んでいきたい。

【謝辞】

本博士論文に多大な助言を頂いた、北川博之教授、天笠俊之准教授、櫻井鉄也教授に感謝致します。

【文献】

[1] Caleb C. Noble and Diane J. Cook. Graph-based anomaly detection. In *KDD*, pp. 631–636, 2003.

[2] Jimeng Sun, Yinglian Xie, Hui Zhang, and Christos Faloutsos. Less is more: Compact matrix decomposition for large sparse graphs. *SDM*, pp. 366–377, 2007.

[3] Leman Akoglu, Mary McGlohon, and Christos Faloutsos. Oddball: Spotting anomalies in weighted graphs. In *PAKDD*, pp. 410–421, 2010.

[4] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, Vol. 51, No. 3, pp. 455–500, September 2009.

[5] Koji Maruhashi, Fan Guo, and Christos Faloutsos. MultiAspectForensics: Pattern mining on large-scale heterogeneous networks with tensor analysis. In *ASONAM*, pp. 203–210, 2011.

[6] Koji Maruhashi, Fan Guo, and Christos Faloutsos. MultiAspectForensics: mining large heterogeneous networks using tensor. *IJWET*, Vol. 7, No. 4, pp. 302–322, 2012.

[7] Koji Maruhashi and Christos Faloutsos. EigenDiagnostics: Spotting connection patterns and outliers in large graphs. In *ICDM Workshop*, pp. 1328–1337, 2010.

[8] Michalis Potamias, Francesco Bonchi, Carlos Castillo, and Aristides Gionis. Fast shortest path distance estimation in large networks. In *CIKM*, pp. 867–876, 2009.

[9] Koji Maruhashi, Junichi Shigezumi, Nobuhiro Yugami, and Christos Faloutsos. EigenSP: A more accurate shortest path distance estimation on large-scale networks. In *ICDM Workshop*, pp. 234–241, 2012.

丸橋 弘治 Koji MARUHASHI

平成 11 年富士通（株）入社. 平成 14 年より（株）富士通研究所. 現在に至る. 平成 21 年より 1 年間、米国 Carnegie Mellon 大学にて客員研究員. 平成 26 年筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻博士後期課程修了.