産業連関ネットワーク解析のための
疎化処理と閾値の関係について

九州大学　土中哲秀
九州大学　小野廣隆

1 はじめに

近年, 産業連関ネットワークに関する分析が盛んに行われている．産業連関ネットワークは, 産業間の取引関係のデータである産業連関表を隣接行列とみなした, 点を産業, 辺を産業間の取引, 辺重みを取引量としたネットワーク（グラフ）である．産業連関ネットワーク分析の中で, 産業間の重要な部分構造を見つけるクラスタ分析はグラフ最適化問題とも親和性が高い．Kagawa らは, 正規化カット問題を応用したクラスタ分析を行い, 日本の自動車サプライチェーンにおいて, 4 つの CO_2 排出クラスタを抽出した [6]．この他にも様々な産業連関ネットワークからの各種定義に基づくクラスタ抽出・それに基づく政策提言を行う研究がなされている．

このように, 産業連関ネットワーク分析において, 重要な部分構造（組合せ構造）抽出が必要とされる場面が多く存在する．しかし, そのような部分構造抽出問題は多くの場合, 計算論的な難問であることが多く (NP 困難), 産業数が数百程度のネットワークでも素朴なアルゴリズムでは解析が不可能となる．一方, グラフアルゴリズム論の発展により, 一般には NP 困難である問題であっても, 何らかのパラメータ値が小さいグラフを対象とするならば, 高速なアルゴリズムの設計が可能となる場合がある．例えば, 木幅と呼ばれるグラフパラメータが小さなグラフでは, 多くの組合せ最適化問題を高速に解くことができることが知られている．このようなとき, 木幅に関する固定パラメータアルゴリズムが存在するという．例えば, 後述する最小 k-分割問題は, 木幅に関する $O(2^{\omega} \cdot n^3)$-時間固定パラメータ容易アルゴリズムが存在する [8]．ただし, n は頂点数, ω はグラフの木幅である．本研究ではグラフの木幅の上限値と下限値を求めることにより, 木幅に関する高性能アルゴリズムの適用可能性について考える．一般に産業連関分析ネットワークは木幅が大きく, そのままでは木幅に関する固定パラメータアルゴリズムは適用できない．そこで, 本研究ではネットワークの辺重みに閾値を導入し, 閾値未満の重みの辺を削除することにより, 対象となるグラフの木幅を小さくすることを考える（グラフ疎化）．この際の閾値の選択によっては分析対象となる構造まで壊しかねない．このようなグラフ疎化に対する影響を調べるために, モジュラリティと呼ばれるクラスタリング指標を用いて, グラフ疎化による構造の変化についても考察する．

2 産業連関ネットワーク

産業連関ネットワークは, 産業連関表から作成されたネットワーク（グラフ）である．産業連関表とは産業間の取引関係を表した正方行列であり, 各要素は産業間の取引量を表す．これを重み付き隣接行列と見立てて構成したグラフが産業連関ネットワークである．

このような生成法から, 産業連関ネットワークはいくつかの特徴的な性質を持つ．第一に, 地理的, あるいは経済圏の影響を反映した部分的に密なグラフ構造を持つ．例えば, 日本国内の県別に分けた産業を

対象とした産業連関ネットワークを考えると, (1) 同県内, あるいは近県間の産業間に重みの大きい辺が張られやすい, (2) 遠い県の関連の薄い産業間にはそもそも辺が張られない (取引量 0), 等の特徴がある. 同様の特徴は, 国際産業連関ネットワークにおける, 国別の産業の関係等でも見られる. このことは, 産業連関ネットワークの対象とする産業の分類にも依存するが, 大雑把には産業連関ネットワークは, 極めて密度が大きいいくつかの部分を全体としては比較的疎な形で結んだようなグラフとなっていることを示唆するものである. 2005 年の日本産業連関表 (397 部門) からなるグラフの例では, グラフ密度が約 0.468 (潜在的な辺のうち, 46.8%が実際に辺となっている) と密であり, また大きさ 93 のクリークを持つ.

第 2 に, 辺重みの分布がロングテール的, すなわち重みの大きな辺は少数であり, 重みの小さい辺が大部分を占めている. 2005 年の日本産業連関表 (397 部門) における辺重みの分布は図 2 のようになる. すなわち, 上述のように 2005 年の日本産業連関表の表すネットワークは密ではあるが, その辺の多くは比較的小規模の取引を表しており, 重要と考えられる大きな規模の取引を表す辺はごく少数であることを示唆している. このような辺重み分布のグラフに対して, グラフ分割問題によるクラスタリングを考えるとき, 閾値によるグラフ疎化は有効な手段であると考えられる.

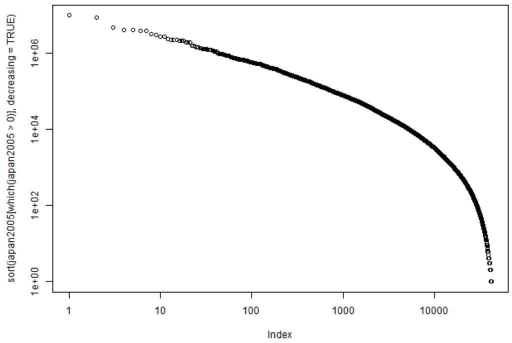

図 1: 2005 年日本産業連関表 (397 部門) の辺重み分布 (ログスケール)

3 グラフ分割問題

本節では, 産業連関ネットワークに対するクラスタ分析への応用が期待される正規化カット問題と最小 k-分割問題について述べる. はじめに, 正規化カット問題は以下のように定義される.

定義 3.1 ([9])

入力: $G = (V, E)$, 辺重み関数 $w : E \to Q^+$, 自然数 K.

出力: $\sum_{k=1}^{K} cut(C_k)/\sum_{u \in C_k, v \in V} w_{uv}$ を最小にする K 個の点部分集合 $C_1, \ldots, C_k \subseteq V$. ただし, $C_k \cap C_l = \emptyset$ $(\forall k, l \in \{1, 2, \ldots, K\}, k \neq l)$ かつ $\bigcup_{k=1}^{K} C_k = V$.

ここで w_{uv} は, $e = (u, v) \in E$ のとき, $w_{uv} = w_e$, そうでないとき, $w_{uv} = 0$ とする. また, $cut(C) = \sum_{u \in C, v \notin C} w_{uv}$ である. この問題は, 一般のカット問題と異なり, 出力されるクラスタ, すなわち K 個の点部分集合から得られる誘導部分グラフの辺の合計値でカット値を割る. これによって, クラスタの内側は密になり, 外側は疎になる. 正規化カット問題は NP-困難であるが, グラフスペクトル理論と k-平均法を用いて, 近似解を求める方法がある.

一方, 最小 k-分割問題は以下のように定義される.

定義 3.2 ([5])

入力: $G = (V, E)$, 自然数 K.

出力: $\sum_{k=1}^{K} cut(C_k)$ を最小にする K 個の点部分集合 $C_1, \ldots, C_k \subseteq V$. ただし, $||C_k| - |C_l|| \leq 1, C_k \cap C_l = \emptyset$ $(\forall k, l \in \{1, 2, \ldots, K\}, k \neq l)$ かつ $\bigcup_{k=1}^{K} C_k = V$.

すなわち, カット値が最小であり, かつそれぞれの点部分集合の大きさの差が 1 以下である K 個の点部分集合を見つける問題である. 最小 k-分割問題は NP-困難である. また $k = 2$ のとき, 特に最小 2 分割問

題と呼ばれる．最小 2 分割問題も, 同様に NP-困難である．しかし, 木幅に関する $O(2^\omega \cdot n^3)$-時間固定パラメータ容易アルゴリズムが存在する [5].

Kagawa らは正規化カット問題を応用することによりクラスタ分析を行っているが, 解が k - 平均法の初期値に影響を受けやすく, 計算結果が不安定であるという問題があった [6]. また, 応用先である産業クラスタの発見は, クラスタ間の産業の協力により CO_2 の削減や生産の強化などを行うときの効果の上昇を目指すものである．したがって, クラスタ内の産業の数がなるべく近くなるように産業を分割することが求められる．以上から, 産業連関ネットワークに対するクラスタ分析に対して, 最小 k-分割問題とそれに対するアルゴリズムの応用が期待される．次節では, 最小 k-分割問題に対する木幅に関するアルゴリズムの適用可能性を調べるために, 木幅の上限値と下限値を求める．

4 閾値と木幅

本節では, 木幅の上限値と下限値を求める．初めに木幅の性質を述べる．

木幅とは, 木分解することにより得られるグラフパラメータである．まず, 木分解の定義を与える．

定義 4.1 (木分解 [8]) *グラフ $G = (V, E)$, ノード集合 $I = \{1, 2, \ldots, N\}$, ノードのバッグ集合 $\mathcal{X} = \{X_1, X_2, \ldots, X_N \subseteq V\}$, , 木 T に対して, 以下の条件を満たす $\langle\{X_i | i \in I\}, T\rangle$ を木分解 (Tree Decomposition) という．*

1. $\bigcup_{i \in I} X_i = V$,

2. $\forall\{u, v\} \in E$ に対して, $\exists i \in I : \{u, v\} \subseteq X_i$,

3. $\forall i, j, k \in I$ に対して, T において i と k のパスの間に j が存在するとき, $X_i \cap X_k \subseteq X_j$.

このとき, 木幅は $\omega := \max_{i \in I} |X_i| - 1$ と定義される．木幅を求めることは, NP-困難であるが, $O(n^{\omega+2})$-時間アルゴリズムなどが知られている [1].

また, $G = (V, E)$ に対して, 木幅を ω, 最大クリークの大きさを s とすると以下の性質を満たす．

性質 4.1 ([3])

1. $\omega \geq s$,

2. G が弦グラフならば, $\omega = s$.

したがって, グラフの最大クリークの大きさとグラフを三角化して得られる弦グラフの最大クリークの大きさを求めることにより, 木幅の上限値と下限値を求めることができる．

実際に求めた最大クリークの大きさを図 2, 3 に示す．図 2 は 2005 年日本産業連関表 (397 部門), 図 3 は WIOD2011(40 カ国 35 部門) のデータを使用している．図の縦軸は最大クリークの大きさ, 横軸は辺重みに関する閾値を表わす．右にいくほど閾値は高くなり, グラフは疎になっている．また, 黒の点が元のグラフ, 赤の点が三角化して得られるグラフを表わす．したがって, 赤と黒の点の間に木幅が存在する．

本研究では, R のグラフパッケージ igraph と gRbase を用いて検証した [10, 4]. 図が示すように, はじめはグラフの木幅は大きいが, グラフを疎化することにより, 急激に減少していくことがわかる．

5 閾値とモジュラリティ

前節では, グラフを疎化していくときの木幅の大きさの上限値と下限値を変化を調べた．本節では, グラフ疎化によって, 重要な部分構造が失われていないかをモジュラリティと呼ばれる指標を用いて考察する．

モジュラリティとはグラフのクラスタリングの良さを表わす指標である．重み付きグラフ G において, K 個のクラスタが与えらたとき, モジュラリティは以下のように定義される．

定義 5.1 ([7])

$$Q(K) = \sum_{k=1}^{K} \left\{ \frac{\sum_{u,v \in C_k} w_{uv}}{\sum_{u,v \in V} w_{uv}} - \left(\frac{\sum_{u \in C_k, v \in V} w_{uv}}{\sum_{u,v \in V} w_{uv}} \right)^2 \right\}$$

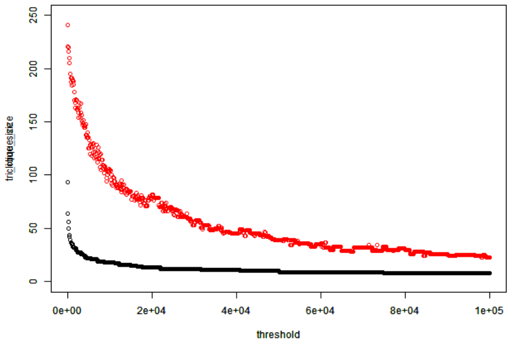

図 2: 2005 年日本産業連関表 (397 部門) の閾値と最大クリークの変化

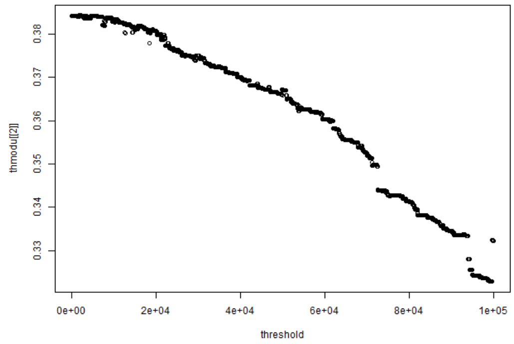

図 4: 2005 年日本産業連関表 (397 部門) の閾値とモジュラリティの変化

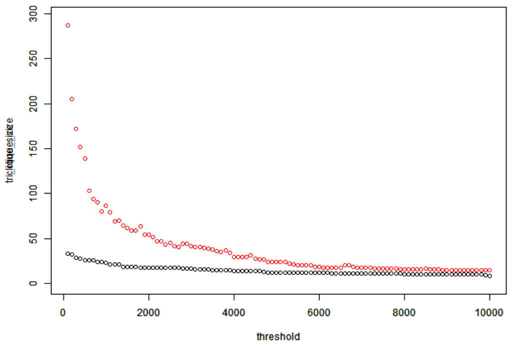

図 3: WIOD2011(40 カ国 35 部門) との閾値と最大クリークの変化

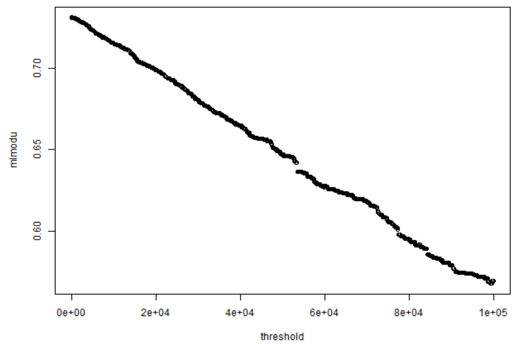

図 5: WIOD2011(40 カ国 35 部門) との閾値とモジュラリティの変化

我々は, 閾値で辺を切ったあとの疎化されたグラフにグラフクラスタリングアルゴリズムを適用し, 得られたクラスタを再び疎化前のグラフに当てはめることにより, 疎化による影響について分析した. 本研究では, グラフクラスタリングアルゴリズムとして, 計算時間が速い Blondel らのアルゴリズムを使用した [2].

2005 年日本産業連関表と WIOD2011 に対する結果をそれぞれ図 4, 5 に示す. 図の縦軸はそれぞれの閾値で切ったときのモジュラリティ, 横軸は閾値である. 最大クリークの大きさと比較すると, モジュラリティはあまり減少しない. また, 疎化したときに生じる孤立点を適切にクラスタに割り当てれば, さらに上昇することが期待される.

6　まとめと今後の課題

本研究では, 木幅の上限値と下限値を分析することにより, グラフ分割問題に対する木幅に関する固定パラメータ容易アルゴリズムの適用可能性について考察した. そのままのグラフでは木幅が大きいが, グラフを疎化することで加速度的に減少することがわかった. 一方で, グラフの構造の変化を見るためにモジュラリティを指標とした分析も行った. その結

果, モジュラリティは木幅に比べて, 減少しにくいことがわかった. すなわち, 重要な部分構造が残ったまま疎化されているといえる. 以上のことから, 産業連関ネットワークに対するグラフの疎化処理は有効であることが予想される. 今後は, 木幅に関するヒューリスティックアルゴリズムを用いてより木幅に近い近似値を求め, さらに実際にグラフ分割問題を解くことによって, グラフ疎化の解に対する影響を分析することによって, 産業連関ネットワークに対するグラフ疎化がより有効であることを示したい.

参考文献

[1] Stefan Arnborg, Derek G. Corneil, and Andrzej Proskurowski. " Complexity of finding embeddings in a k-tree", *SIAM Journal on Algebraic and Discrete Methods, Volume 8*, 1987, pp. 277-284

[2] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre. " Fast unfolding of communities in large networks", *J. Stat. Mech. (2008) P10008* , 2008

[3] Hans L. Bodlaender and Arie M.C.A," Kosterm Treewidth computations I. Upper bounds", *Information and Computation, Volume 208, Issue 3*, 2010, Pages 259-275

[4] Claus Dethlefsen and Søren Højsgaard, " A Common Platform for Graphical Models in R: The gRbase Package", *Journal of Statistical Software Vol. 14, No. 17*, 2005

[5] Klaus Jansen, Marek Karpinski, Andrzej Lingas, and Eike Seidel, "Polynomial Time Approximation Schemes for MAX-BISECTION on Planar and Geometric Graphs", *STACS 2001 Volume 2010, LNCS*, 2001, pp 365-375

[6] Shigemi Kagawa, Sangwon Suh, Yasushi Kondo and Keisuke Nansai, "Identifying environmentally important supply chain clusters in the automobile industry", *Economic Systems Research, Volume 25, Issue 3*, 2013, pp. 265-286

[7] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks", *Physical Review E 69 026113*, 2004

[8] Rolf Niedermeier, "Invitation to Fixed-Parameter Algorithms", *Oxford University Press*, (2006)

[9] Jianbo Shi and Jitendra Malik, "Normalized Cuts and Image Segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence archive Volume 22 Issue 8*, 2000 , pp. 888-905

[10] http://igraph.org/