研究ノート

地域間フローデータの階層的視覚化

Hierarchical Visualization of Inter-Regional Flow Data

中　村　有　一*
Yuichi NAKAMURA

Keywords：visualization, inter-regional flow data, maximum spanning tree, power law, scale-free

1. はじめに

通話や郵便など地域間を流れる情報の定量的なデータを分析する手法について、これまで主に統計学的な方面から研究を進めてきた。対象とするデータは地域間 OD データであり、この行列に対してクラスター分析や多次元尺度構成法（MDS）などの手法を適用することより、圏域の特定、階層構造の抽出などを試みた[1)]。これに対して、本研究ノートでは、主にグラフ理論的手法を適用し、シンプルでわかりやすい分析を試みる。また全体の基幹的な構造を強調してすっきりと表現する視覚化の効果についても考察する。

2. データと分析環境

ここで取り上げるデータは、「都道府県間通信回数（平成 25 年度）」と呼ばれるもので、電気通信事業者協会（TCA）から公開されている[2)]。このデータは加入電話と ISDN の通話回数をまとめたもので、事業者としては NTT、KDDI、ソフトバンク、その他の合計である。なお携帯電話等は含まれていない。

データの形式は、47 × 47 の OD 行列の形をとっており、自地域内の通話データもあるので、全部で 47 × 47=2209 個のデータからなる。なぜ都道府県間のデータを使うかといえば、地域数が少なすぎると階層構造をとらえにくくなるし、逆に多すぎると煩雑になって考察が難しくなると考えたからである。

今回の分析ではデータの非対称性は考慮しないこととする。元のデータを A={Aij} とすると、Aij と Aji はほぼ等しいとして、地域間のつながりを表すデータは、その合計 Dij=Aij+Aji とする。また行列の対角要素に対応する自地域内に着信するデータは対象外とする。つまり対角要素 Dii=0 とする。以上のように加工したあとにできる対称行列 D={Dij} をもとにグラフの形

* 多摩大学経営情報学部　School of Management and Information Sciences, Tama University

に変形し、グラフ理論的な手法を使って分析を行う。

主な分析環境としては、R と RStudio を利用し、グラフ理論的な手法は、ライブラリとして igraph と linkcomm を使用する[3)]。

3. べき乗則

まずデータの大まかな特徴として、べき乗則がおおむね成り立っていることを確認しておこう。「地域の順位（ランク）」と「地域の大きさ（サイズ）」を両対数グラフにプロットする。両者の関係が直線的であれば、ランクサイズルールに従うことになる。つまりべき乗則が成り立っていることを意味している。図 1 は、この点を検証するためのグラフである。ここでは都道府県ごとの発信量の合計が、地域の大きさ（サイズ）を表す指標と考える。またこのサイズの値をソートして、地域の順位（ランク）を求める。これにより 47 個の値からなる 2 個の変数が求められる。それぞれのデータの常用対数を取り、プロットしたのが図 1 である。図 1 では、プロットした点に直線をあてはめ、決定係数 R^2 を求めた。

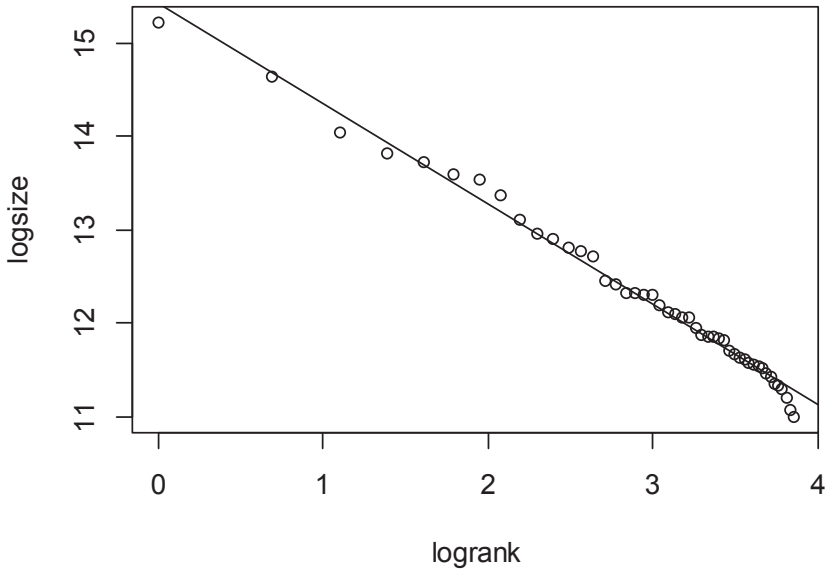

図 1　ランクとサイズの両対数プロット（直線近似の決定係数 R^2 =0.9894）

図 1 から、べき乗則はおおむね成り立っていることがわかる。同じデータについて対数を取らないでプロットするとロングテールの構造となっている。またロングテールであれば、「2 割 8 割の法則」、「パレートの法則」などと呼ばれている法則が成り立っているはずである。この法則は「上位 2 割のシェアが全体の 8 割を占める」というものである。これについても実際のデータで確認してみよう。

47 都道府県の発信量の合計をソートし、ランクで上位 9 地域までの和を求めると、全体に占める割合は、65.11% となる。上位 9 地域が全体 47 地域に占める割合は、9/47=19.15% となる。つまり上位約 20％の地域が占める割合は、全体の約 65％となる。さらに地域ごとの合計データではなく、元の対称行列 D で 2 回ずつ出てくる点を考慮して半分にした 1081 個（=（47

× 47 − 47）/2=1081）の中から上位約 20％にあたる上位 216 個のデータが全体に占める割合を求める。これは 87.70％となる。つまり上位 2 割のシェアが全体の 8 割になるという「2 割 8 割の法則」がほぼ成り立っているといえる。

べき乗則で構造が決まるようなシステムは､特徴的な大きさを持たないという意味で｢スケールフリー（scale-free）｣と呼ばれている。フラクタル的と呼ばれることもある。このようなネットワークは、ランダムに分布するネットワークに比べて、情報の集中する中心性の高い地域が支配的である。この点については、ランダムグラフとの比較を行い、その特徴を明らかにしていきたい。

4. 上位対地データによる分析

上で求めた対称行列 D で、横方向の最大値を求める。このとき何列目の要素が最大になるかを、その添え字で求める。（R では order 関数を使用する。）同様に 2 番目に大きい値、3 番目に大きい値を取るときの添え字を求める。このようにしてある地域からの発信量の中で上位 3 位（第 3 対地）までの地域を求めたものが、表 1 である。

表 1　各都道府県の第 3 対地までの地域（都道府県コードは ID 欄に表示）

ID	地域名	ID	第1対地	ID	第2対地	ID	第3対地
1	北海道	13	東京都	27	大阪府	14	神奈川県
2	青森県	13	東京都	4	宮城県	3	岩手県
3	岩手県	13	東京都	4	宮城県	2	青森県
4	宮城県	13	東京都	7	福島県	3	岩手県
5	秋田県	13	東京都	4	宮城県	3	岩手県
6	山形県	13	東京都	4	宮城県	27	大阪府
7	福島県	13	東京都	4	宮城県	14	神奈川県
8	茨城県	13	東京都	12	千葉県	11	埼玉県
9	栃木県	13	東京都	11	埼玉県	8	茨城県
10	群馬県	13	東京都	11	埼玉県	9	栃木県
11	埼玉県	13	東京都	14	神奈川県	12	千葉県
12	千葉県	13	東京都	14	神奈川県	11	埼玉県
13	東京都	14	神奈川県	11	埼玉県	12	千葉県
14	神奈川県	13	東京都	27	大阪府	11	埼玉県
15	新潟県	13	東京都	27	大阪府	20	長野県
16	富山県	27	大阪府	13	東京都	17	石川県
17	石川県	27	大阪府	13	東京都	16	富山県
18	福井県	27	大阪府	13	東京都	17	石川県
19	山梨県	13	東京都	27	大阪府	11	埼玉県
20	長野県	13	東京都	27	大阪府	15	新潟県
21	岐阜県	23	愛知県	13	東京都	27	大阪府
22	静岡県	13	東京都	23	愛知県	27	大阪府
23	愛知県	13	東京都	27	大阪府	21	岐阜県
24	三重県	23	愛知県	27	大阪府	13	東京都
25	滋賀県	27	大阪府	26	京都府	13	東京都
26	京都府	27	大阪府	13	東京都	25	滋賀県
27	大阪府	13	東京都	28	兵庫県	23	愛知県
28	兵庫県	27	大阪府	13	東京都	26	京都府
29	奈良県	27	大阪府	13	東京都	26	京都府
30	和歌山県	27	大阪府	13	東京都	26	京都府
31	鳥取県	27	大阪府	32	島根県	13	東京都
32	島根県	27	大阪府	34	広島県	13	東京都
33	岡山県	27	大阪府	34	広島県	13	東京都
34	広島県	27	大阪府	13	東京都	33	岡山県
35	山口県	34	広島県	40	福岡県	27	大阪府
36	徳島県	27	大阪府	13	東京都	37	香川県
37	香川県	27	大阪府	13	東京都	38	愛媛県
38	愛媛県	27	大阪府	13	東京都	37	香川県
39	高知県	27	大阪府	13	東京都	37	香川県
40	福岡県	13	東京都	27	大阪府	43	熊本県
41	佐賀県	40	福岡県	13	東京都	27	大阪府
42	長崎県	40	福岡県	13	東京都	27	大阪府
43	熊本県	40	福岡県	13	東京都	27	大阪府
44	大分県	40	福岡県	13	東京都	27	大阪府
45	宮崎県	40	福岡県	13	東京都	27	大阪府
46	鹿児島県	40	福岡県	13	東京都	27	大阪府
47	沖縄県	13	東京都	27	大阪府	40	福岡県

表 1 のデータをグラフとして扱い、視覚化したものが図 2 ～図 4 である。図 2 は第 1 対地のみのデータ、図 3 は第 1 対地と第 2 対地を合わせたデータ、図 4 は、第 1 から第 3 まで 3 つの対地データを合わせたデータから、グラフを描いたものである。

図 2 についてみると、以下のような知見が得られる。

（1）グラフは木（tree）の構造をとり、大部分の地域は、東京または大阪にリンクしている。

（2）福岡以外の九州の各県は福岡にリンクし、福岡は東京にリンクしている。

（3）岐阜、三重が愛知にリンクし、愛知は東京にリンクしている。

（4）山口は広島を経由して、大阪につながっている。

（5）北海道と沖縄は距離的には離れているが、東京に直接リンクしている。

以上の点はほぼ予想されるとおりである。

図３については、以下のような知見が得られる。

(1) 多くの地域で第１対地が東京で、第２対地が大阪となっている。

(2) 図２でははっきり見えてこなかった宮城、愛知、広島の圏域が現れる。

(3) 九州は大阪とのつながりが少なく、東京とのつながりが大きい。

図４では、細かい構造が現れるが、辺の交差が多くなり煩雑な印象を与える。

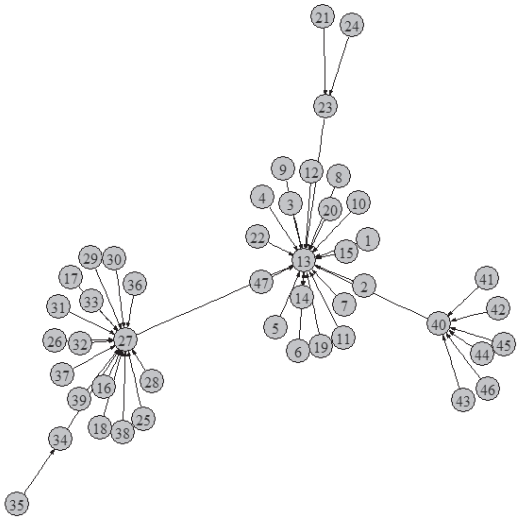

図２　第１対地のグラフ

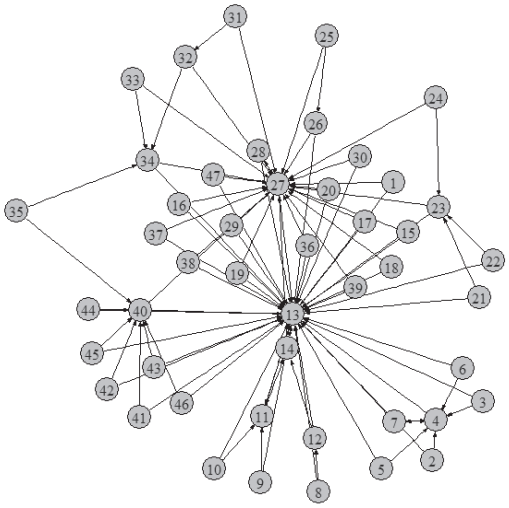

図３　第１対地～第２対地のグラフ

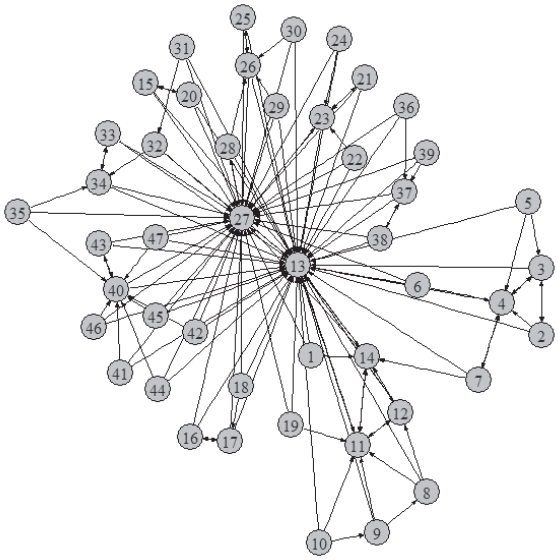

図４　第１対地～第３対地のグラフ

図5、図６は、それぞれ第２対地のみ、第３対地のみのデータをグラフ化したものである。図４では、辺の交差が多く視認性が低いが、図5、図６は、いわば図４を２か所で切り、断面

を取り出したものと考えられる。これらのグラフは、ほぼ木の構造となり、見やすくなっている。

以上のように第１対地から第３対地までのデータをグラフ化するだけでも、多くの情報が得られることがわかる。

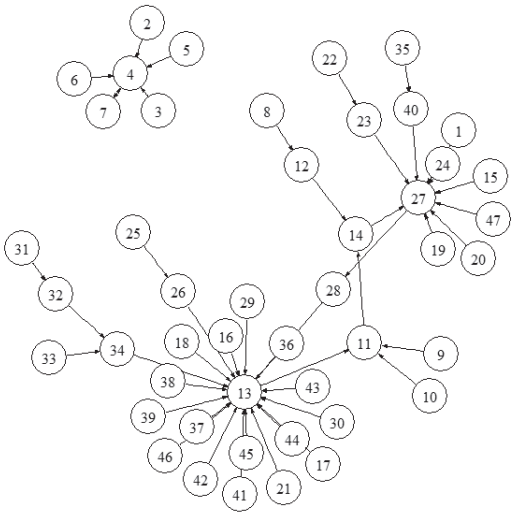

図５　第２対地のグラフ

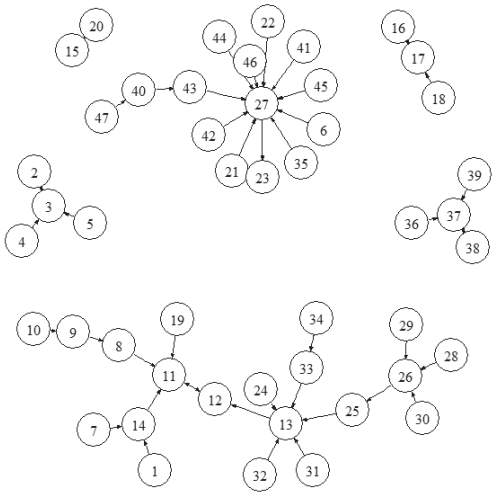

図６　第３対地のグラフ

5. 最大木による分析

次に対称行列Ｄをグラフの重みと解釈して、最大木（最大全域木、maximum spanning tree）を求めてみよう。最大木はすべての頂点を連結する木の中で、重みの和が最大となるものである。igraph ライブラリには、最大木とは逆の最小木を求める関数がある。最大木は、重みのマイナスを取って最小木を求めることにより得られる。

この結果が図７である。このグラフは、図の向きなど見た目は異なるが、図２とまったく同じ構造になっている点が興味深い。つまりデータの性格により、最大木を求めなくても第１対地をリンクするだけで同じ結果が得られている。このように基幹部分に木の構造が埋め込まれているデータでは、いろいろな方法を試しても結局、木構造が抽出される可能性が高い。

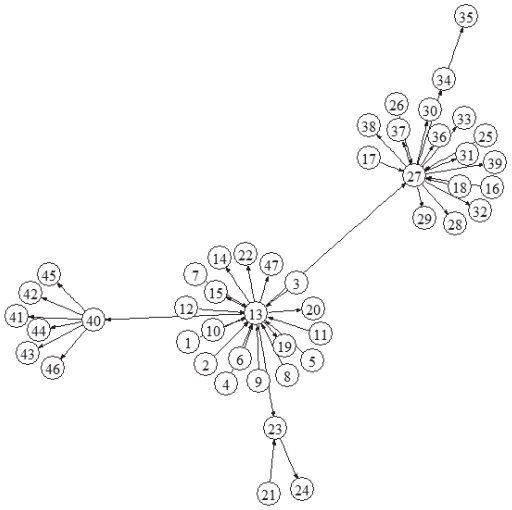

図７　最大木のグラフ

最大木分析の発展として、以下のような改良が考えられる。

(1) グラフの頂点の円の大きさを、対応する地域の規模に応じて変える。たとえば発信量の合計をあてる。これにより中核的な地域が強調され、よりメリハリのある図が得られるはずである。

(2) グラフの辺の太さを、その流量に応じて変える。太さを変化させる範囲は、ある程度調整する必要があるが、これも量的なつながりを視覚化するのに有効である。

(3) 図７の最大木グラフは、回転しても裏返しても同じ構造を表しているが、中心性の高い（たくさんの辺が集まっている）地域を上位に配置すると、階層的な構造がわかりやすく表現できる。いわば無重量状態に重力を働かせるようなイメージである。

(4) 最大木は、木の構造をもっているので、任意の２点間のルートは１つしかない。ここに木を作るときに切り捨てた流量を束ねることにより、すべての流量をグラフ上に表示できるようになる。この結果はより幹線ルートが強調される図になると思われる。

6. 今後の発展性について

グラフの辺が多くなり複雑化してくると、グラフを表示したとき、どうしても視認性が低下してしまう。特に問題となるのは、辺の交差である。これが多くなると煩雑なイメージを与える。この点を解消するために平面グラフ（planar graph）について検討してみよう。

平面グラフとは、辺が交差することなく平面上に描くことができるグラフのことである。平面グラフとなる条件については、「Kuratowski の定理」が有名である[4]。図８と図９の２種類のグラフが埋め込まれていなければ、平面グラフであり、どちらかが埋め込まれていれば平面グラフにはならないというものである。平面グラフであるかどうかを効率的に判定するためのアルゴリズムも開発されている[5),6)]。また平面グラフは辺をすべて直線分にして描くことができること（Fáry の定理）が知られている[7)]。

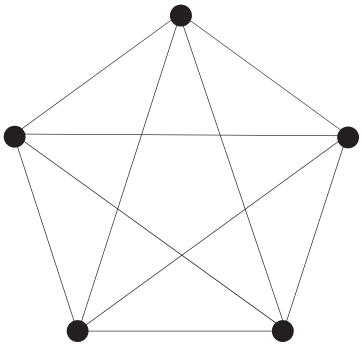

図８　完全グラフ K_5

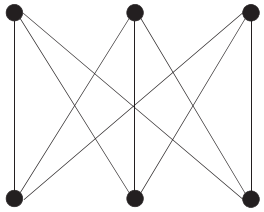

図９　完全２部グラフ $K_{3,3}$

実用上は完全に平面的にする必要はなく、ある程度の平面性があればいいと思われる。つまりグラフの平面性を判定し、平面的に描けるものはできるだけ平面的に表現することが求められる。そのためには物理的な原理を導入することが有効であろう。図 10 はグラフを見やすく描画するのに使われているモデルである。リンクのある頂点の間はバネで結ばれ、結ばれていない頂点の間ではお互いに遠ざけあう力（斥力）が働くものとしている。このようにすると図 2 のようなグラフが得られる。これと同じように、平面グラフにおいても物理的な原理を導入することにより、視認性の高いグラフを描くことが可能であると考えられる。

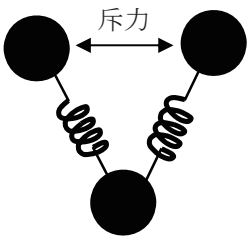

図 10　バネと斥力によるモデル

地域間のつながりを表すデータからは階層的な構造がよく現れるが、その力学的な意味をシミュレーションにより解明することはできないだろうか。初期値を設定して、地域間の関係を力学的に記述し、地域間の関係がどのように変化するかシミュレーションを行うことができれば、階層構造が生成する過程を時間的な経過を追って視覚化することができる。できるだけ単純な原理により、現実のデータがある程度再現されれば、大きな意味を持つことになるだろう。

7. おわりに

グラフ理論的なアプローチは、まだ緒についたばかりであるが、様々な手法を実際のデータに適用して、その有効性を検証していきたい。そしてその過程で作成したライブラリをまとめて公開することにより、手軽に分析できる環境を整えていきたい。

参考文献

1） 中村有一「地域間フローデータの視覚化について」経営・情報研究（多摩大学研究紀要）研究ノート 2013, No.17, pp.105-112（2013）
2） 電気通信事業者協会「テレコムデータブック 2015」第 2 章情報通信サービス利用状況　pp.21-25 http://www.tca.or.jp/databook/（2015）
3） 鈴木努「ネットワーク分析　R で学ぶデータサイエンス 8」共立出版（2009）
4） R.J. ウィルソン , J.J. ワトキンス（大石泰彦　訳）「グラフ理論へのアプローチ」日本評論社, pp.266-267（1997）
5） Hopcroft, John; Tarjan, Robert E.（1974）"Efficient planarity testing", Journal of the Association for Computing Machinery, 21（4）: 549–568
6） "Version 6.4 The LEDA User Manual", Algorithmic Solutions, 371-373（2008）
7） Fáry, István（1948）, "On straight-line representation of planar graphs", Acta Sci. Math.（Szeged）, 11: 229–233