

# 反復的なグラフサンプリングによる巨大ネットワークの局所可視化

鳥山菜海子<sup>†</sup>   吉田 光男<sup>††</sup>   伊藤 貴之<sup>†</sup>

<sup>†</sup> お茶の水女子大学 〒112-8610 東京都文京区大塚 2 丁目 1 番 1 号

<sup>††</sup> 豊橋技術科学大学 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1 番 1 号

E-mail: <sup>†</sup>{g1720529,itot}@is.ocha.ac.jp, <sup>††</sup>yoshida@cs.tut.ac.jp

**あらまし** 情報可視化の研究では、最初に全体を可視化し、徐々に局所構造に着目していく可視化が一般的とされてきた。しかし情報可視化の対象となる全てのデータにおいて、この操作手順が最良な手順であるとは限らない。例えば、その中から特徴的な部位に注目したい場合は、全体像の可視化することが必要であるとは限らない。これは大規模ネットワークの可視化においても同様である。そこで本報告では、ネットワークのサンプリングを反復して多数の部分ネットワークを作成し、それぞれのネットワークをアイコンとして表現する。その後、作成した多数のアイコンに対してクラスタリングを行い、階層型データとして可視化を行う。可視化画面では、ユーザがアイコンをクリックした際に部分ネットワークの可視化結果を確認することができる。

**キーワード** ネットワーク可視化、大規模グラフ、サンプリング、SNS

## 1 はじめに

情報社会の発展により、多様な種類の大規模なデータが容易に手に入るようになった。SNS 上での人間関係、会社の取引情報、感染症の流行に関するデータなどが大規模データの典型例である。こうした大規模なデータの可視化において、データをそのまま可視化するだけではデータの中に潜む傾向や現象を理解しづらい状況が起こりえるようになった。そこで、大規模なデータであっても一般の人々がデータに潜む重要な意味を容易に解釈できるように、あらかじめデータから本質的な部分を抽出した上で可視化する手法が活発に研究されるようになった。

情報可視化の研究では、“Overview first, then zoom and filter, details on demand” [1] とあるように、最初に全体を可視化し、徐々に局所構造に着目していく手法が一般的とされている。大規模ネットワークの可視化においても、最初にネットワーク全体を可視化したのち、局所構造に注目していくという操作手順を用いた論文が多数発表されている [2] [3]。しかし、全ての可視化でこの手法が最良であるとは限らない。これは大規模ネットワークの可視化においても同様である。典型的な状況として、ネットワークの特徴的な局所に注目したい場合においては、最初に全体像の可視化が必要ではないことがある。例えばソーシャルネットワークのデータを可視化する場合、データの所有者にとって興味深い現象は、ソーシャルネットワークデータの全体像よりも、局所部分の構造や属性であることが多い。このような背景から、大規模ネットワークの中から特徴的な局所をあらかじめ抽出し、その各局所の特徴を強調するようなネットワーク可視化手法が近年研究されている。

以上のような潮流の中で我々は、ネットワークのサンプリングを反復して多数の部分ネットワークを作成し、それぞれのネットワークを可視化する手法を研究している。具体的には、巨大ネットワークデータからグラフサンプリングを反復し、多

数の部分ネットワークを抽出したのち、ネットワークを小さなアイコンとして表現する。現時点の実装では、部分ネットワークの特徴量から円グラフを生成し、これらをアイコンとして表示している。さらに、各々の部分ネットワークから算出された特徴量に対して、次元削減とクラスタリングを適用し、適用した結果を階層型構造のデータとして可視化する。さらにこれらの手順で配置されたアイコンをクリックすると、部分ネットワークの可視化結果を表示することができる。

本報告では実行例として、各政党の Twitter 公式アカウントおよびそれらの発言をフォローしたアカウントで構成されるネットワークを可視化した事例を示す。

本報告の構成は以下の通りである。2 章では、グラフサンプリング、部分ネットワーク作成時のアイコン表現、大局的ネットワークの可視化、それぞれについて関連研究を紹介する。3 章では我々が提案する可視化手法の処理手順を述べる。4 章では本手法の実行結果と考察について、5 章ではまとめと今後の課題について述べる。

## 2 関連研究

### 2.1 グラフサンプリング

大規模ネットワークから部分ネットワークを抽出するグラフサンプリングについて、旧来から活発に研究がなされている。その中でも汎用的なグラフサンプリングの手法に関しては、いくつかのサーベイ論文 [4] [5] が発表されている。現在の我々の実装では汎用的なグラフサンプリングを利用しているが、アプリケーションごとに最適なサンプリング手法を適用することも重要である。

グラフサンプリングを活用したネットワーク可視化手法の例として、ネットワークデータをベクトル化した後にサンプリングして可視化する手法 [6] が発表されている。しかし、ネットワーク情報をベクトル化することにより、可視化表現の幅は広

がるものの、ネットワークの文脈情報が失われてしまうため、そのまま利用するには課題がある。

## 2.2 アイコン表現についての可視化

ネットワークのアイコン表現については、ノードの特徴量をアイコン化した表現、ネットワークの接続構造を可視化した表現、の2種類が提案されている。ノードの特徴量をアイコン化した表現としては、特徴量を円グラフで表現してネットワークのノードとした表現[7]が発表されている。ネットワークの接続構造を可視化した表現としては、隣接行列を描画した表現[8]や、Small multiples を利用した表現[9]が提案されている。

現状の実装では単純に特徴量を円グラフとして表現しているが、本研究には上述のような多様な表現が適用可能であると考えられるので、引き続き検討したい。

## 2.3 局所ネットワーク群の可視化

大規模ネットワークを効果的に可視化するために、重要な局所ネットワークをアイコン化する可視化手法がいくつか知られている。Yoghourdjian ら[10]は、タンパク質結合ネットワークの大局的構造をアイコン表現する試みを発表している。この手法では線形の計算時間での可視化を可能とする上に、同型グラフには同じアイコンが表示されるという利点がある。一方で、内容は異なるが構造が類似しているグラフはアイコンが似た表現になりやすい、内容は異なるが構造が違うグラフの類似性が見えにくい、という課題がある。また Chen ら[11]の手法では、ユーザ定義による局所ネットワークを基準にして、グラフの大局的構造を抽象化して可視化する手法を提案している。この手法では、ノードの構造情報をベクトル化してノードでエンコードすることで、局所ネットワークの多様な表現を可能にしている。他にもいくつかのネットワーク可視化手法が知られているが、我々の手法は「数十個・数百個単位のアイコンで表現された多数の部分ネットワーク群をその特徴によって分類表示する」ことに重点を置いたという点でこれらと異なる。

# 3 処理手順

本章では提案手法の処理手順を示す。提案手法は、1) 反復的なグラフサンプリングによる部分ネットワークの抽出、2) 部分ネットワークのアイコン表示、3) アイコン群のクラスタリングと階層的表示、4) 対話操作によって指定された部分ネットワークの画面配置、の4つの工程で構成される。

なお本報告では、伊藤らによるネットワーク可視化手法[12]と同じデータ構造のネットワークを仮定する。具体的には、ネットワーク  $G$  はノードの集合  $V = \{v_1, v_2, \dots\}$  とエッジの集合  $E = \{e_1, e_2, \dots\}$  で構成され、ノード  $v$  は特徴量となる多次元ベクタを有するものとする。

## 3.1 グラフサンプリング

本手法ではグラフサンプリングを反復することで多数の部分ネットワークを抽出する。現在の実装では Hu ら[4]が紹介する手法のうち、単純な幅優先探索にもとづくグラフサンプリ

ングを採用している。探索の初期位置となるノードには、多数のノードからリンクされている重要度の高いノードを採用している。こちらはリンク情報の入ったデータの一つずつ探索していき、リンク数が多くされているノードをまとめたデータをあらかじめ作成している。これにより、重要度の高いノードの周辺に構築される重要なネットワーク構造が高い確率で抽出される。ここでいう重要度の高いノードとは、例えば次章で後述する Twitter のネットワークの場合にはフォロー・フォロワーが多い人物を意味する。

現在のサンプリングの課題として、「重要なノードの定義」および「探索手法の性能」には、グラフの接続構造から導かれる汎用的な知見と、アプリケーション依存な知見が混在している、という点に注意が必要である。例えば次章で後述する Twitter のネットワークの場合には、ソーシャルネットワークに特化したサンプリング手法[13][14]を参考にするのが一手段であると考えられる。

## 3.2 部分ネットワークのアイコン表示

続いて、グラフサンプリングを反復して生成した多数の部分ネットワークを、アイコンとして可視化する。本手法においてアイコンが表現したい情報は、主に「部分ネットワークの接続構造」と「特徴量の大小関係」である。現状の実装では単純に、特徴量から円グラフを生成している。円グラフを採用している理由として、アイコンを確実に正方形に近い形状にしたいから、という点があげられる。ソーシャルメディアの可視化手法のサーベイ論文[15]には、円グラフ以外のアイコン表現がいくつか紹介されている。例として隣接行列によるアイコン表現[8]などは試す価値があると考えられる。

以上を踏まえて、アイコンデザインの例として、

- 円グラフ以外の形で特徴量を表現する
- 部分ネットワークの接続構造を概略的にあらわすアイコンを生成する

などの検討を今後の課題としたい。

## 3.3 アイコン集合の画面表示

上述の手法により算出された各アイコンの特徴量に対して、次元削減とクラスタリングを適用し、その結果を階層型データ可視化手法によって画面配置する。現時点での実装では、各ノードの特徴量の平均値をアイコンの特徴量とし、これに k-means 法によるクラスタリングを適用している。続いて、クラスタリング結果から得られた階層構造を、伊藤らのデータ可視化手法[16]で画面配置している。我々の実装では図3に示されるように、ウィンドウの左半分に階層構造を表示している。

## 3.4 部分ネットワークのグラフ配置

我々の実装では、アイコン集合のうち一つをクリック操作によって指定すると、そのアイコンに対応する部分ネットワークをウィンドウの右半分に表示する。部分ネットワークの表示に際して本手法では、一般的なノードリンク型のネットワーク画面配置手法を適用する。

現在の実装では、伊藤らの画面配置手法[12]を採用して、ネッ

トワークを画面配置している．この手法では3章の冒頭にて示した通り，一般的なグラフ構造に加えて，各ノードは特徴量に相当する多次元ベクトルを有するものとする．そして

- リンクで連結された共通ノードを多く有する2ノードが小さい距離を有する

- 特徴量の非類似度が小さい2ノードが小さい距離を有する

という2条件に沿って算出されたノード間距離にもとづいて，ノードのクラスタリングを実行する．ノード間距離は以下の式で算出される．

$$d = \alpha d_{vec} + (1 - \alpha) d_{adj}$$

ここで， $d_{vec}$  は隣接ノードの共通性にもとづいた2ノードの距離であり， $d_{adj}$  は特徴量の非類似度にもとづいた2ノード間の距離を示す．この2条件に沿って算出されたノード間距離にもとづいて，ノードのクラスタリングを実行する．図1は，3.1節で示したグラフサンプリングによって得られた局所ネットワークに対し，伊藤らの画面配置手法[12]を適用して可視化した例である．

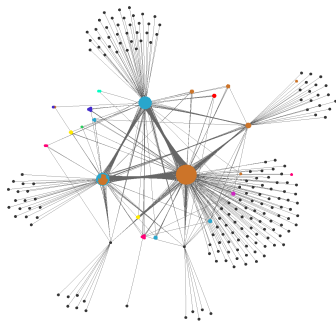


図1 グラフサンプリングによって得られた局所ネットワークに対して伊藤らの画面配置手法[12]を適用した例．

上述のクラスタリング手法は，多くのノードに連結された重要なノードをクラスタから分離する働きが強いことから，ソーシャルメディアの可視化においてインフルエンサーなどの注目度が高いノードとそれ以外のノード群の関係を表示するのに向いている．一方で，伊藤らの手法では，一般的なコミュニティ抽出型（リンクの密度が濃い部位をクラスタとする）のクラスタリングも実装しており，こちらを適用することも可能である．

以上の処理によるクラスタリングの実行後に，各クラスタをノードとみなし，クラスタを構成するノード数に応じてクラスタの半径を求める．続いて，クラスタをノードとみなしたグラフに対して力学モデルを適用して画面配置する．その後，求めた半径を用いてラプラシアンスムージングを適用し，ノードを放射状に配置することで画面配置を修正する．これらの処理を図2に示す．

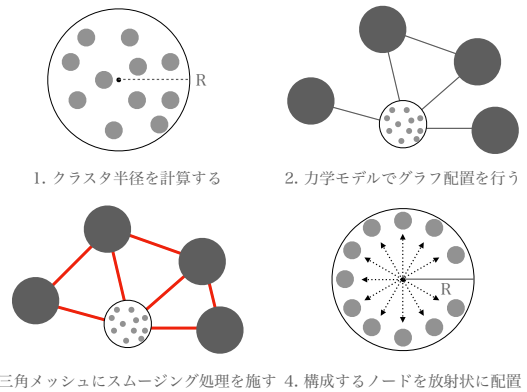


図2 画面配置手法．

なお，本手法によるクラスタリングと画面配置は数千個単位のノードを有するネットワークに対して平均10秒程度の計算時間を要する．そこで，これらの計算は前処理として済ませておいて，その計算結果を対話操作に応じて読み込む，という実装が必要となる．現状の実装では，ネットワークのクラスタリングと画面配置は別のソフトウェアとして切り離されており，計算結果はキャプチャ画像として保存されている．この実装はネットワーク可視化結果への対話操作の点で望ましくないため，改良を進めたい．

## 4 実行結果と考察

本章では実行例として，各政党のTwitter公式アカウントおよびそれらの発言をフォローしたアカウントで構成されるネットワークを可視化した事例を示す．このデータは2017年9月から10月までに収集したものであり，ノードはTwitterアカウント，エッジはフォロー関係を表している．ノードの総数は460,683個，エッジの総数は1,425,696,164本であった．また政党公式アカウントは幸福実現党・自由民主党・公明党・新党大地・自由党・社会民主党・日本共産党・日本のこころ・日本を元気にする会・維新の党・民進党・希望の党・立憲民主党の13党で，可視化時には各党には固有の色が与えられている．各アカウントは13政党の各々の公式発言を何回リツイートしたかによって得られる13次元ベクトルが特徴量として与えられる．各アイコンには対応する各アカウントの特徴量の平均値を13次元ベクトルとして算出して特徴量として用いる．

我々は上述のデータから，3.1節で示した幅優先探索によって部分ネットワークを抽出する処理を反復した．本実験では幅優先探索の開始点を，政党公式アカウントの発言を最低1回リツイートしている84,043個のアカウント[17][18]に限定した．サンプリング開始の初期位置については，フォロー・フォロワー数の多いものを注目度の高いノードから開始するようにした．具体的にはフォロー・フォロワー数が100以上のものを注目度の高いノードとし，注目度が最も高いものからサンプリ

ングを開始するようにしている。また本実験では、部分ネットワークのノード数を 5000 に固定している。今後の課題として、接続構造に応じて部分ネットワークのノード数を可変にしたい。

これらの部分ネットワークの集合をクラスタリングしてアイコンの集合で表示した結果を図 3 に示す。この可視化例ではアイコンを 6 個のクラスタに分割している。クラスタ数を動的に変更して、その都度クラスタリングを再実行することもできる。アイコンの 1 個をマウスオーバーすると、対応する部分ネットワークを画面右側に表示する。

図 3 の左側をみると、各クラスタが同様な配色の円グラフ群で構成されていることがわかる。例えば左下には、立憲民主党と共産党の発言を熱心に拡散するユーザで構成された部分ネットワークが多数あることがわかる。このような表現によって、部分ネットワーク群の特徴量（この事例では政党別の拡散回数で構成されたベクトル）の分布を把握し、拡散の構造を画面右側に表示された部分ネットワークで確認することができる。例えば 1 人の極端に強力なインフルエンサーによる「広く浅い」部分ネットワーク、少人数への拡散が何度も繰り返される「狭く深い」部分ネットワーク、といった構造の違いを観察することができる。図 4 にいくつかの例を示す。この例を見ただけでも、立憲民主党と共産党の発言の拡散には多様なパターンが存在することが示唆される。

一方で、図 3 の左側に図示されたアイコンの各クラスタを眺めても、クラスタを構成する大半のアイコンが非常に類似して見えるため、どれが興味深い部分ネットワークなのかをアイコンから推察することが難しい。このことから本報告において、特徴量だけでなく接続構造も同時に表現した複合的なアイコンのデザインとクラスタリング手法の開発が重要な課題であることがわかる。

## 5 まとめと今後の課題

本報告では、ネットワークのサンプリングを反復して多数の部分ネットワークを抽出し、これらをアイコンで表現し、階層的に可視化する手法を提案した。また、日本の各政党の Twitter 公式アカウントの発言拡散ネットワークを適用事例として、本手法による可視化の事例を紹介した。

今後の課題として、グラフサンプリングの実装について再考したい。現在の実装では、アプリケーションに依存しない汎用的なサンプリング手法を導入しており、アプリケーション固有のデータの特性や可視化の工程などを考慮したサンプリング結果となっていない。そこで今後は、例えば本報告で示したデータであれば SNS に特化したサンプリング手法の導入を検討したい。また、本手法で抽出している局所ネットワークのノード数が固定となっている点も問題である。今後はこれを見直し、適切な判断基準で探索を打ち切ることで、ネットワーク中の興味深い部位を適切なサイズで抽出できるようにしたい。

もう一つの課題として、アイコンのデザインとクラスタリングの再考があげられる。現在のアイコンは単に特徴量の平均値を算出して円グラフで表現しただけであり、ネットワークの

接続構造を表現していない。そこで今後の課題として、ネットワークの接続構造と特徴量を両方同時に表現するアイコンを表現できるようにしたい。またクラスタリングについても例えば、まず特徴量でクラスタリングした後に接続構造で再度クラスタリングする、といった拡張を試みたい。以上の観点から効果的なアイコンのデザインとクラスタリングを打ち出すことができれば、当該課題において一定の新規性を主張できると思われる。

また、部分ネットワークのグラフ配置の実装についても課題がある。現状の実装では、クラスタリングと画面配置に計算時間を要することから、あらかじめ前処理としてこれらの計算を済ませておき、画面配置結果をキャプチャ画像として保存している。そして実行時には、クリック操作に応じてキャプチャ画像を表示させている。こちらも対話操作の点で望ましくないため、改良を進めたい。

これらの改良を進めた上で、データ領域の専門家（例えば Twitter データ分析の専門家）と議論を進め、巨大ネットワークの分析ツールとしてふさわしい方向に研究を進め、また Twitter の発言拡散ネットワーク以外にも適用事例を挙げたい。

## 文 献

- [1] B. Shneiderman, The eyes have it: a task by data type taxonomy for information visualizations, IEEE Symposium on Visual Languages, 336-343, 1996.
- [2] Y. Wu et al. StreamExplorer: A Multi-Stage System for Visually Exploring Events in Social Streams, IEEE transactions on visualization and computer graphics vol. 24, No. 10, pp. 2758-2772, 2018.
- [3] J. Heer and D. Boyd, Vizster: visualizing online social networks, IEEE Symposium on Information Visualization, 2005. INFOVIS 2005., Minneapolis, MN, pp. 32-39, 2005.
- [4] P. Hu, W. C. Lau, A Survey and Taxonomy of Graph Sampling, arXiv preprint, 1308.5865, 2013.
- [5] J. Leskovec, C. Faloutsos, Sampling from Large Graphs, 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 631-636, 2006.
- [6] Zhiguang Zhou, et al., Context-aware Sampling of Large Networks via Graph Representation Learning, IEEE Trans Vis Comput Graph. 2020.
- [7] Marcum, Christopher S., and David R. Schaefer. Save Room for Pie: Adding Pie Charts to Network Visualizations in R with Statnet and Plotrix, SocArXiv, 4 Aug. 2016. Web.
- [8] F. Lekschas, et al., HiPiler: Visual Exploration of Large Genome Interaction Matrices with Interactive Small Multiples, IEEE Transactions on Visualization and Computer Graphics, Vol 24, No. 1, pp. 522-531, 2018.
- [9] X Liu, et al., CorrelatedMultiples: Spatially Coherent Small Multiples With Constrained Multi - Dimensional Scaling. Computer Graphics Forum, 37, 2018.
- [10] V. Yoghoudjian, T. Dwyer, K. Klein, K. Marriotto, M. Wybrow, Graph Thumbnails: Identifying and Comparing Multiple Graphs at a Glance, IEEE Transactions on Visualization and Computer Graphics, Vol. 24, No. 12, pp. 3081-3095, 2018.
- [11] W. Chen, et al., Structure-Based Suggestive Exploration: A New Approach for Effective Exploration of Large Networks, IEEE Transactions on Visualization and Computer Graphics, Vol. 25, No. 1, pp. 555-565, 2019.
- [12] T. Itoh, K. Klein, Key-node-Separated Graph Clustering and Layout for Human Relationship Graph Visualization, IEEE Computer Graphics and Applications, Vol. 35, No. 6,

pp. 30-40, 2015.

- [13] T. Wang, et al., Understanding Graph Sampling Algorithms for Social Network Analysis, International Conference on Distributed Computer Systems Workshop, 2011.
- [14] Z. S. Jalali, A. Rezvanian, M. R. Meybodi, Social Network Sampling Using Spanning Trees, International Journal of Modern Physics C, Vol. 27, No. 5, 1650052, 2016.
- [15] Yingcai Wu, et al., A Survey on Visual Analytics of Social Media Data, IEEE Transactions on Multimedia, Vol. 18, No. 11, pp. 2135 - 2148, 2016
- [16] T. Itoh, et al., Hierarchical Visualization of Network Intrusion Detection Data in the IP Address Space, IEEE Computer Graphics and Applications, Vol. 26, No. 2, pp. 40-47, 2006.
- [17] M. Yoshida, F. Toriumi, Analysis of Political Party Twitter Accounts' Retweeters during Japan's 2017 Election, IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 736-739, 2018.
- [18] M. Yoshida, F. Toriumi, Information Diffusion Power of Political Party Twitter Accounts During Japan ' s 2017 Election, 10th International Conference on Social Informatics (SocInfo 2018), pp. 334-342, 2018.

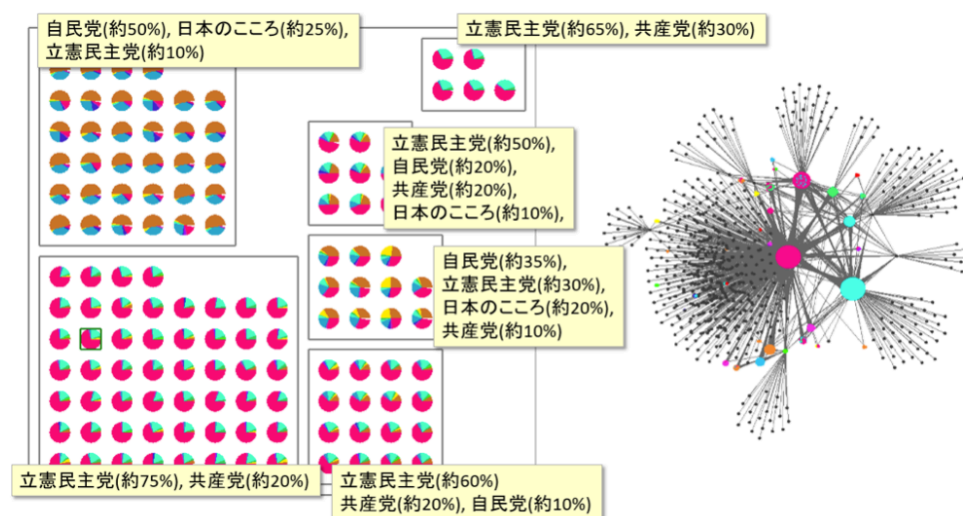


図 3 可視化例. 画面左側では部分ネットワーク集合をアイコン（現時点では円グラフ）で表示し、6 個のクラスタに分類している. 円グラフの 1 個をマウスオーバーすると、対応する部分ネットワークを画面右側に表示する. ただし右側のネットワークはリアルタイムで画面配置しているのではなく、前処理として画面配置結果を算出してキャプチャした画像を表示している.

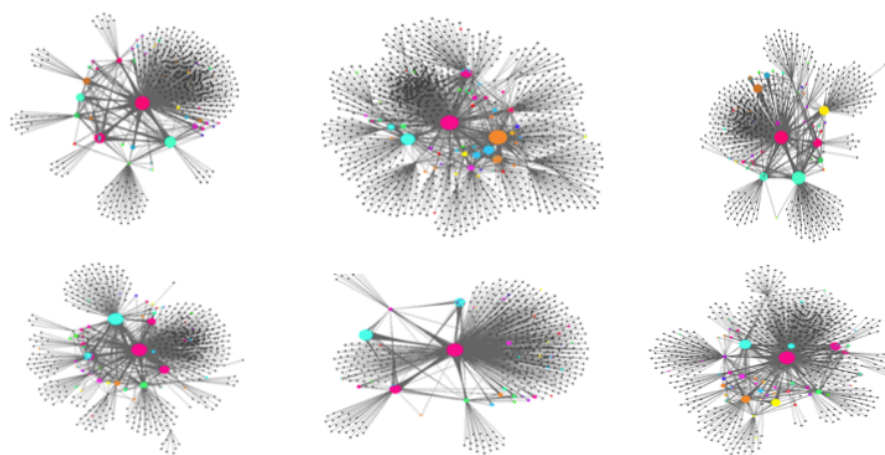


図 4 図 3 の可視化例で「立憲民主党 (約 75%), 共産党 (約 20%)」と書かれたクラスタを構成する部分ネットワークの中から 6 個を表示した例. 各々の部分ネットワークが固有の接続構造を有することがわかる.