1C5-5

3部ネットワークにおける高速なコミュニティ抽出手法

Fast method for community detection in tripartite networks

貫井 駿　　村田 剛志
Shun Nukui　Tsuyoshi Murata

東京工業大学 大学院情報理工学研究科 計算工学専攻
Department of Comuper Science, Graduate School of Information Science and Engneerings Tokyo Institute of Technology

Many social media such as Youtube, Delicious, or Hatena Bookmark, can be expressed by tripartite networks. Community detection is one of the methods for analizing these networks. Detecting community helps us understand their structures, so it is important in network analysis. Modularity-based community detection finds network partitions by optimizing modularity. In the case of tripartite networks, Fast Unfolding for Edges(FUE), proposed by Ikematsu, is the fastest among existing methods, but his method is still not fast enough for large networks. In this paper, we propose faster method which is an improved version of FUE. Our method clusters edges quickly with Louvain method and assigns community labels to nodes. In an experiment with Delicious network (around 80,000 edges), we show that our proposed method is around 2,500 times faster than FUE.

1. はじめに

ソーシャルメディアの多くは，ユーザー，リソース，タグという 3 種類のインスタンスの関係で表現できる．例えば，ソーシャルブックマークサービスのはてなブックマークや Delicious では，ユーザーがお気に入りのウェブページをタグ付けしてブックマークするという行為が行われる．このような 3 種類のインスタンスの関係は 3 部ネットワーク (3-partite 3-uniform ハイパーネットワーク) として表現できる．そのネットワークを可視化したり，構造を理解するための解析手法がコミュニティ抽出である．コミュニティ抽出とは，ネットワークから関連性の高いノードのグループ，すなわちコミュニティを発見する手法である [Fortunato 10]．コミュニティやその対応関係が発見できると，細部のデータからでは得られない情報を得ることが可能となる．一般的なネットワークにおいて最も代表的な Newman のモジュラリティ[Newman 04] では，コミュニティ内のエッジが密で，コミュニティ間のエッジが疎である分割を高く評価する．モジュラリティを用いたコミュニティ抽出では，モジュラリティを最適化するネットワーク分割を探索する．Blondel らが提案した Louvain 法 [Blondel 08] は，代表的なモジュラリティ最適化手法で，高速かつ高精度に Newman のモジュラリティを最適化可能である．

従来の 3 部ネットワークのコミュニティ抽出に関する研究では，Newman モジュラリティを 3 部ネットワークに拡張した Neubauer の 3 部モジュラリティ[Neubauer 10] を用いた手法が研究されている．既存の 3 部モジュラリティ最適化手法の中で実行速度と精度が高いとされているのが，池松が提案した Fast Unfolding for Edges(FUE) である [池松 14]．この手法は，3 部ネットワークのハイパーエッジをノードとする隣接エッジネットワークを構成する．そして，そのエッジネットワークを 3 部モジュラリティが最適化されるようにクラスタリングし，そのエッジクラスタからノードへコミュニティラベルを割り当てる．しかし，FUE は大規模なネットワークに適応するには実行速度が遅く，実ネットワークの解析は難しい．

連絡先: 貫井 駿，東京工業大学大学院情報理工学研究科計算工学専攻村田剛志研究室，東京都目黒区大岡山 2-12-1 W8-59，nukui.s@ai.cs.titech.ac.jp

そこで本稿では，従来法より高速な 3 部ネットワークのコミュニティ抽出手法の提案をする．提案手法では，3 部ネットワークをエッジネットワークに変換し，Louvain 法によりエッジクラスタを抽出し，そのエッジクラスタからノードへコミュニティラベルを割り当てる．Delicious[Delicious] の実ネットワークを用いた実験において，提案法が従来法に比べて，約 2500 倍高速であり，かつ従来法と類似したコミュニティが抽出可能であることを示す．

2. 関連研究

本節では従来のモジュラリティとその最適化手法について説明する．また 3 部ネットワークとそのコミュニティの定義についても説明する．

2.1 Newman のモジュラリティ

モジュラリティとは，ネットワークの分割の良し悪しを評価する指標である．Newman モジュラリティを用いたコミュニティ抽出において，良いコミュニティとはコミュニティ内のエッジが密で，コミュニティ間のエッジが疎であるノード集合である．Newman らは一般的なネットワーク (1 部ネットワーク) におけるモジュラリティを式 (1) のように定義した [Newman 04].

$$Q_{Newman} = \sum_{l}(e_{ll} - a_l^2) \tag{1}$$

$$e_{lm} = \frac{1}{2M}\sum_{i\in V_l}\sum_{j\in V_m} A(i,j) \tag{2}$$

$$a_l = \sum_{m} e_{lm} = \frac{1}{2M}\sum_{i\in V_l}\sum_{j\in V} A(i,j) \tag{3}$$

V_l はコミュニティl に含まれるノード集合，M は総エッジ数，$A(i,j)$ は隣接行列を表している．式 (1) において，e_{ll} はコミュニティl 内のエッジ数,a_l^2 は null モデルにおける期待値を表している．コミュニティ内エッジが密な分割のときモジュラリティ値は高い値をとり，そうでないときは低い値となる．

2.2 Louvain 法

モジュラリティを用いたコミュニティ抽出手法では，それを最適化するネットワーク分割を求める．しかし，その厳密解

を求めることは NP 困難であるため，近似手法であるモジュラリティ最適化手法を用いる．Blondel らが提案した Louvain 法は，高速かつ高精度に Newman モジュラリティを最適化可能な手法である [Blondel 08]．Louvain 法の処理手順を図 1 に示す．

1. 全てのノードをそれぞれコミュニティとして初期化し，モジュラリティQ を計算する．
2. 最適化ステップ
   (a) ノードをランダムに選択し，全てのノードに対して (b) から (d) の操作を行う．
   (b) 選択したノード v に隣接するコミュニティ集合 C_{adj} を得る
   (c) v が各 $c \in C_{adj}$ に移動したときの差分モジュラリティΔQ を計算する．その最大値を ΔQ_{max}，そのときの隣接コミュニティを c_{max} とする．
   (d) $\Delta Q_{max} > 0$ ならば v のコミュニティを c_{max}，$Q = Q + \Delta Q_{max}$ とする．
3. 融合ステップ
   コミュニティが同一なノード集合を 1 つのノードに融合する．作成した新しいネットワークに対してステップ 2 を行う．コミュニティの移動がなかった場合はアルゴリズムを終了する．

図 1: Louvain 法の処理手順

2.3 3 部ネットワークの定義

本稿で扱う 3 部ネットワークは，3-partite 3-uniform ハイパーネットワークと呼ばれるものである [Neubauer 09]．以降，3-partite 3-uniform ハイパーネットワークを単に 3 部ネットワークと呼ぶ．3 部ネットワーク G の定義は (4) で表される．

$$G = (V^X, V^Y, V^Z, E) \tag{4}$$

$$e = (i, j, k) \tag{5}$$

$$i \in V^X, j \in V^Y, k \in V^Z, e \in E \tag{6}$$

上式の V^X,V^Y,V^Z はノード集合，E はハイパーエッジの集合である．ハイパーエッジは一般的なエッジと異なり，3 つ以上のノードを接続することができる．

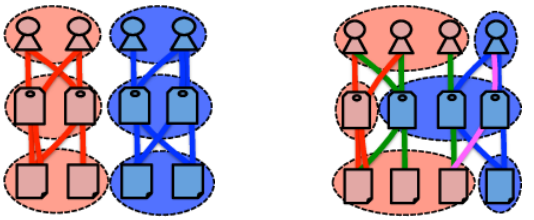
2.4 3 部ネットワークにおけるコミュニティの定義

2 部ネットワークや 3 部ネットワークなどの n 部ネットワークの定義は複数存在する．例えば Barber による n 部ネットワークのコミュニティの定義では，任意の部分のノード集合をコミュニティとしている [Barber 07]．しかし本研究では，コミュニティとして 1 種類のノードから成るものを考える．すなわち，1 つのコミュニティ内に V^X,V^Y,V^Z のノードが混ざることは許されない．

2.5 Neubauer の 3 部モジュラリティ

3 部ネットワークの分割の評価には，Neubauer が提案した式 (8) で表される 3 部モジュラリティが用いられる [Neubauer 10]．このモジュラリティはコミュニティ内のエッジの粗密ではなくコミュニティが持つ対応関係を評価する．

式 (8) 中の e_{lmn} はコミュニティl，m，n 間のハイパーエッジの本数の割合，a_l^X は X 部分におけるコミュニティl に接続



(a) モジュラリティ値：高　　(b) モジュラリティ値：低

図 2: 3 部ネットワークの分割例．

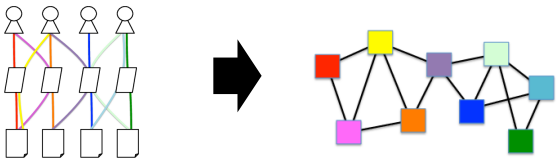


図 3: 3 部ネットワークからの隣接エッジネットワークの作成．

するハイパーエッジの本数の割合を表す．したがって，$(e_{lmn} - a_l^X a_m^Y a_n^Z)$ は，コミュニティl，m，n の対応関係の強さからその期待値を引いた値となり，それに重み α_{lmn} を掛けた値が部分モジュラリティとなる．

Neubauer の 3 部モジュラリティは，図 2(a) のように他のグループとの関係が類似しているノード集合に分割されている場合を高く評価し，反対に図 2(b) のような分割を低く評価する．

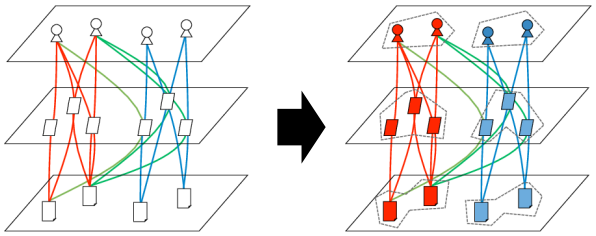
$$Q_{Neubauer} = \sum_l \sum_m \sum_n \alpha_{lmn}(e_{lmn} - a_l^X a_m^Y a_n^Z) \tag{7}$$

$$\alpha_{lmn} = \frac{1}{3}\left(\frac{e_{lmn}}{a_l^X} + \frac{e_{lmn}}{a_m^Y} + \frac{e_{lmn}}{a_n^Z}\right) \tag{8}$$

2.6 Fast Unfolding For Edges(FUE)

モジュラリティ最適化手法である Louvain 法は，Newman のモジュラリティを最適化するアルゴリズムである．そのため，Newman のモジュラリティで評価できない 3 部ネットワークに対して Louvain 法を適用することができない．

そこで池松は Louvain 法を応用して，3 部モジュラリティ最適化手法である Fast Unfolding for Edges(FUE) を提案した [池松 14]．FUE はハイパーエッジをクラスタリングすることでネットワーク分割を求める 3 部モジュラリティ最適化手法である．FUE は 3 つのタスク (i) 隣接エッジネットワークの構成，(ii) エッジクラスタを 3 部ネットワークのコミュニティへ対応付け，(iii)3 部モジュラリティ計算から成る．タスク (i) では 3 部ネットワークをハイパーエッジの隣接関係を表す隣接エッジネットワークに変換する．このネットワークは頂点が 1 つのハイパーエッジに対応しており，隣接しているハイパーエッジ同士の対応する頂点間を辺で結ぶ (図 3)．タスク (ii) では，3 部ネットワークのあるノードに接続しているエッジクラスタのうち最大のサイズのクラスタラベルをそのノードのコミュニティラベルとして割り当てる．図 4 にコミュニティラベル割り当て手法の適用例を示す．タスク (iii) では 3 部モジュラリティを用いる．これらのタスクを組み合わせた FUE のアルゴリズムを図 5 に示す．



(a) (b)

図 4: コミュニティラベル割り当て手法をすべてのノードに適用している例．異なる部分間でコミュニティは共有しない．

1. 初期化ステップタスク 1 で隣接エッジネットワーク G_{AE} を構成する．各ノードをクラスタとして初期化し，タスク 3 を適用してモジュラリティQ を計算する．
2. モジュラリティ最適化ステップ
   (a) ノード v をランダムに選択し，そのノードが隣接している全てのクラスタに移動した場合について，タスク 2，3 を適用してモジュラリティの増加値 ΔQ を求め，その最大値を ΔQ_{max}, その隣接クラスタを c_{max} とする．
   (b) Q_{max} が正値の場合 v のクラスタを c_{max} とする．
   (c) 全てのノードについて (a),(b) を適用する．
3. クラスタ融合ステップ
   クラスタを 1 つの頂点に融合し，G_{AE} を再構築する．クラスタの移動が 1 つもなければアルゴリズムを終了し，そうでない場合はステップ 2 に戻る．

図 5: FUE の処理手順.

3. 高速なエッジクラスタリング手法

本節では高速なエッジクラスタリング手法である Louvain Method for Edges(LME) を提案する．

3.1 基本的なアイデア

まず実行速度に関する FUE の問題点を指摘する．FUE におけるモジュラリティ最適化ステップにおいて，ノードを隣接クラスタに移動したときの 3 部モジュラリティの差分を計算する．しかしその度にコミュニティ割り当て手法を用いてエッジクラスタから 3 部ネットワークのコミュニティを求める必要があり，その処理が全体の実行速度の低下の原因となっている．

それを解決するために隣接エッジネットワークにおける Newman モジュラリティを最適化する手法を考える．この手法では，3 部ネットワークにおける分割を考えずに隣接エッジネットワークからエッジクラスタを求める．そうすることでコミュニティ割り当て手法を用いるのが最終的なエッジクラスタを求めた後の 1 回で済むので劇的な高速化が期待される．また，目的関数を別のモジュラリティ関数とすることによりコミュニティ抽出結果が異なることが予想される．しかし，同じコミュニティの対応に属するハイパーエッジ同士は隣接エッジネットワーク上で密に繋がっている傾向にあることを考慮すれば，Newman モジュラリティによって抽出したエッジクラスタはコミュニティ対応に相当するエッジ集合と類似したエッジ集合になると考えられる．そのため，3 部モジュラリティを最適化したときに比べて結果が大きく異ならないことが期待される．

3.2 LME のアルゴリズム

LME の処理手順を図 6 に示す．LME は，隣接エッジネットワークを構成し，エッジクラスタからノードにコミュニティラベルを割り当てるという点で FUE と共通している．しかし，FUE はエッジクラスタを求めるのに 3 部モジュラリティを最適化する．FUE において最も計算時間を要するのが，3 部ネットワークのノードにコミュニティラベルを割り当てる処理である．LME では Newman モジュラリティを最適化することにより，コミュニティラベル割り当てをする必要性をなくし，処理時間の短縮を図っている．ただし，用いている評価関数が異なるため，結果に差異が生じることに注意する必要がある．図 6 の手順 1 における隣接エッジネットワーク構成法と手順 3 におけるコミュニティラベル割り当て手法は，従来の FUE で用いた手法 (2.6 節) を適用する．

1. 3 部ネットワークから隣接エッジネットワーク G_{AE} を構成し，各々のエッジをクラスタとして初期化する
2. 初期化された G_{AE} に対して Louvain 法 (2.2 節) を適用し，エッジクラスタを得る
3. コミュニティラベル割り当て手法でエッジクラスタからノードへコミュニティを割り当てる

図 6: LME の処理手順.

3.3 実験

提案法の実行速度と抽出コミュニティを比較するため，Delicious から抽出した実ネットワークで実験を行う [Delicious]. Delicious のデータセットはそれぞれユーザー，タグ，ウェブページの 3 部で構成されている．1 つのレコードは 1 つのハイパーエッジに相当する．それに付与されているタイムスタンプ情報を用いて期間を変えて抽出し，表 1 の特徴量を持つ 5 つのデータセットを得た．各データセットに対して，提案手法と従来手法の FUE でそれぞれコミュニティ抽出を行い，実行時間を比較する．また，dataset1 において抽出されたタグコミュニティの比較を行う．実験環境として Core i7-3770 3.4GHz, RAM16GB, Python(2.7) を用いる．また，Louvain 法の実装は NetworkX を用いた Python ライブラリを用いた [Aynaud 09].

提案手法 (LME) と FUE それぞれのコミュニティ抽出に要した時間を図 7 に示す．この結果から，提案手法が FUE と比べて最大約 2500 倍高速にコミュニティ抽出可能であることがわかった．また，表 1 の dataset1 において FUE と LME それぞれで抽出したタグコミュニティ同士の類似度を図 8 に示す．類似度の尺度には Jaccard 係数を用いる．コミュニティ数が多いため，FUE はサイズが 20 以上，LME はサイズが 30 以上のものにのみ注目する．また，F_i,L_j はそれぞれ FUE と LME で抽出したコミュニティで，添字はサイズの大きさの降順となっている．LME で得られたコミュニティの多くはピークを 1 つだけ持ち，FUE で得られたコミュニティの部分コミュニティとなっていることがわかる．一方，FUE で得られた F_1 や F_2 のように，LME において細かく分割されるコミュニティが多々みられた．また，F_1 が LME によってどのようなタグ集合に分割されたかの例を表 2 に示す．この表の各タグ集合が LME によって抽出されたコミュニティに対応する．この例では LME によって F_1 のタグ集合をより細かいカテゴリに分けることができた．LME で得られたコミュニティの中には表で

表 1: Delicious データセットの特徴量

	dataset1	dataset2	dataset3	dataset4	dataset5
ユーザー数	190	258	345	453	770
タグ数	490	913	2075	4202	14930
ウェブページ数	1316	3305	3801	5923	15771
エッジ数	2252	4318	9873	19194	76776

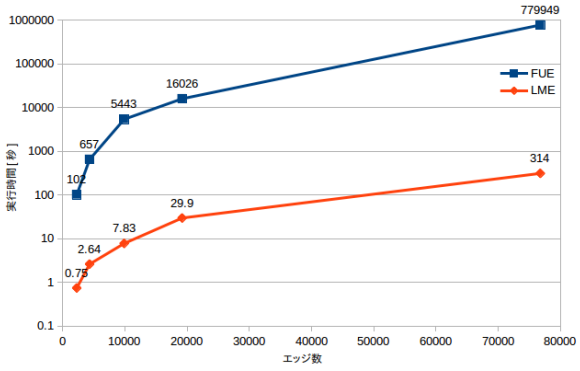


図 7: FUE(青) と LME(オレンジ) の実行時間の比較.

示したように意味の類推ができるものがいくつか見られたが，意味の類推が難しいコミュニティも見られた．

3.4 考察

FUE と LME との間に大きな実行時間の差が生じたのは，エッジクラスタリングを行うときに用いる目的関数が異なることによるものであると考えられる．目的関数を 3 部モジュラリティの代わりに Newman モジュラリティを用いることで，隣接エッジネットワークを 3 部ネットワークに変換することなくモジュラリティを高速に計算することが可能となり，計算時間が $O(NM^2)$ から $O(NM)$ に短縮された．また，最適化する目的関数が異なるために抽出されるコミュニティに差異が生じた．しかし，LME で抽出されるコミュニティが FUE で抽出されるコミュニティの部分集合になる傾向がみられ，さらにその部分集合として表 2 のように意味があると考えられるものがみられた．このことから，スケールの違いを考慮すれば LME は FUE と類似したコミュニティを抽出できると言える．

4. おわりに

本稿では Louvain 法を用いて 3 部ネットワークのハイパーエッジをクラスタリングし，高速にコミュニティ抽出する手法を提案した．Delicious の大規模な実ネットワークの実験において，提案手法は最大 2500 倍高速にコミュニティ抽出が可能となることを示した．また各手法で抽出されたタグコミュニティの結果を比較し，それらが類似していることを示した．この実験結果から，提案手法は従来では規模が大きくコミュニティ抽出が困難であった 3 部ネットワークに対しても適用可能であり，有用なツールであると言える．今後の課題としては，ネットワーク構造が異なる他のデータセットでも実験を行い，より詳細な分析を行うことが挙げられる．現在の NetworkX による実装ではメモリ効率が悪く，10 万ノードを越えるネットワークを扱うことができないため，省メモリの実装に改良することが挙げられる．

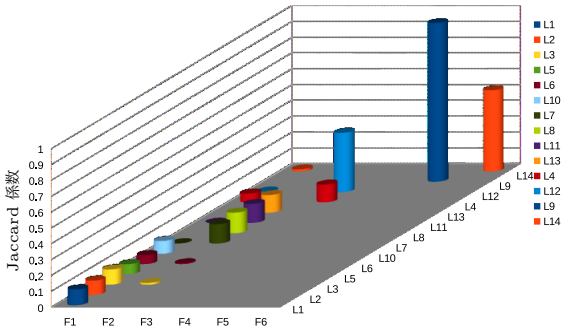


図 8: FUE で得られたコミュニティ F_i と LME で得られたコミュニティ L_j の類似度 (Jaccard 係数) の比較．ピークが 1 つの L_1 や L_2 は F_1 の部分コミュニティであると解釈できる．また，ピークが複数ある F_1 はコミュニティが分割されていることがわかる．

表 2: LME による F_1 のタグ分割例

education,math,teaching,literacy,reading...
media,television,newspaper,movie,video...
javascript,PHP,webdesign,framework,layout...
mootools,programming,web_development,html...

参考文献

[池松 14] 池松恭平, 村田剛志:3 部モジュラリティの改善とその最適化手法, 人工知能学会論文誌, Vol.29, No.2, pp.245-258, 2014.

[Fortunato 10] Santo Fortunato. “Community detection in graphs,” *Physics Reports*, Vol. 486, pp. 75174, 2010.

[Newman 04] M. E. J. Newman, “Fast algorithm for detecting community structure in networks,” *Physical Review E*, Vol.69, No. 066133, pp. 1-5, 2004.

[Blondel 08] Vincent D Blondel and Jean-Loup Guillaume and Renaud Lambiotte and Etienne Lefebvre “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, Vol. P10008, No. 066133, pp. 1-5, 2008.

[Neubauer 09] Nicolas Neubauer and Klaus Obermayer. “Towards community detection in k-partite k-uniform hypergraphs,” In *the NIPS 2009 Workshop on Analyzing Networks and Learning with Graphs*, 2009.

[Neubauer 10] Nicolas Neubauer and Klaus Obermayer. “Community detection in tagging induced hypergraphs,” In *Workshop on Information in Networks*, 2010.

[Barber 07] Michael J. Barber. “Modularity and community detection in bipartite networks,” *Physical Review E*, Vol. 76, No. 066102, pp. 1-9, 2007.

[Delicious] Delicious. http://www.delicious.com.

[Aynaud 09] Thomas Aynaud. python-louvain 0.3, http://perso.crans.org/aynaud/communities/.