3N4-OS-28b-4

ランダムウォークによる二部ネットワークからのコミュニティ検出

Random Walk based Community Detection from Bipartite Networks

邱 シュウレ[*1] 稲木 誓哉[*1] 貫井 駿[*2] 村田 剛志[*2] 岡本 洋[*1]
Xule Qiu Seiya Inagi Shun Nukui Takeshi Murata Hiroshi Okamoto

[*1] 富士ゼロックス(株)研究技術開発本部
Research & Technology Group, Fuji Xerox Co., Ltd.
[*2] 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 Computer Science, Graduate School of Information Science and Engineering Tokyo Institute of Technology

商品購入、論文共著、その他、現実世界の多くの対象は二部ネットワークで表される。我々は以前に、ランダムウォークを用いて一部ネットワークからコミュニティを検出するアルゴリズムを提案した。このアルゴリズムが二部ネットワークからも他の方法よりも高い精度でコミュニティを検出することを示す。さらに、このアルゴリズムにより二部ネットワークのコミュニティの重なりと階層構造を抽出できることを示す。

1. はじめに

二部ネットワークとは、二つの異なるタイプのノードの間にしかリンクが存在しない、すなわち、同じタイプのノード同士の間にはリンクがないネットワークである。現実世界の多くの対象は二部ネットワークで表現される。例えば、ユーザーと商品との間の購入関係、著者と論文との間の共著関係、人とイベントとの間の参加関係、俳優/女優と映画との間の共出演関係、動物と植物との間の相利共生関係、遺伝子と塩基配列との間の関係、代謝反応とそれに関与する化合物との間の関係、文書とそこに登場する単語との間の関係。

コミュニティ抽出は、一部ネットワークの場合と同様、二部ネットワークで表現される複雑系を理解するための基本的な分析作業である。コミュニティ抽出とは、ネットワークの中のノードが相対的に密に繋がったかたまり部分(これをコミュニティとよぶ)を発見するための操作である。コミュニティ抽出により、様々な情報が混じり合って表現された複雑なネットワークを、共通属性を持つノード群、すなわち、コミュニティに分解して、ネットワークの構造を整理することができる。

一部ネットワークからコミュニティを高速・高精度に抽出する方法が既に提案されている。例えば、ネットワーク分割の適切さを表す指標であるモジュラリティを高速に最適化してコミュニティ構造を探索する Louvain 法[1]、ランダムウォークのエンコードを最短化することにより高精度にコミュニティ分割を求める Infomap 法[2, 3]、ネットワークのローカル構造を利用して高速にコミュニティ分割を求める Label Propagation (LPA)法[4]、ネットワークのスケールフリーな特徴を扱えるモデル生成法 degree-corrected Stochastic Block Model (corrected-SBM)[5]、その他の方法が知られている。

しかしながら、これらの方法を二部ネットワークに直接適用することには困難がともなう。一部ネットワークで威力を発揮したこれらの方法は、ノード間のつながりが十分密であることを前提とする。しかしながら二部ネットワークでは、同じコミュニティに所属する同じタイプのノードの間に繋がりがない。そのため、一部ネットワークに比べて二部ネットワークの構造はスパースになる。

とはいうものの、すでに二部ネットワークからのコミュニティ抽出がいくつか試みられている。これらのアプローチは基本的に次の二種類に区別される:二部ネットワークを一部ノードに射影(one mode projection)し、この一部ネットワークに従来のコミュニティ抽出方法を適用する;二部ネットワークそのものからコミュニティを抽出する方法の構築を試みる。

前者のアプローチには以下の問題がともなう。

(I) 射影により二部ネットワークで表現されていた情報が一部失われる。従って、二つの異なる二部ネットワークが同一の一部ネットワークに射影されることが起こりえる [6]。

(II) 多くの場合、射影によりネットワークの規模、特にリンクの数が膨大になり、そのためにコミュニティ抽出のための計算量が大幅に増える。

(III) 二部ネットワークの中の一部のノードが極端に大きな次数を持つ、すなわち、非常に多くの他の種類のノードと繋がる場合(例えば文書-単語の二部ネットワークにおいて “the” などの stop-words がある場合)、射影により生成された一部ネットワークには、互いに強く重なる・繋がるコミュニティができてしまう[6]。このようなコミュニティ構造はモジュラリティに基づく一般的なコミュニティの定義—コミュニティ間のリンクが疎である—と相違するため、モジュラリティ最適化による方法によっては正しくコミュニティを検出できない。

(IV) 射影の方法によってコミュニティ抽出結果が変わる。重み付けして射影した場合の方が、しない場合よりも多くの情報が保存されているため、良いコミュニティ抽出結果が得られることが知られている[6,7]。

これらの問題を避けるためには、やはり二部ネットワークそのものからコミュニティを抽出する方法を検討すべきである。すでに、一部ネットワーク志向のコミュニティ抽出方法を二部ネットワークからのコミュニティ抽出に適用できるように拡張することが試みられている。代表的なコミュニティ検出方法であるモジュラリティ最適化の二部ネットワークバージョンが既に Guimera ら[7]および Barber ら [8]により提案されている。その他、LPA の二部ネットワークバージョンが Barber ら [9](LPAb)および村田ら[10](LPAb+)により、SBM の二部ネットワークバージョン(biSBM)が Daniel ら [6] により提案されている。

我々は以前に、ランダムウォークに基づくコミュニティ抽出方法を提案した。この方法によるコミュニティ抽出は、モジュラリティのような事前のコミュニティ定義を要さず、コミュニティ間に重なりがあるネットワーク、非クリーク的な構造を持つネットワーク、

連絡先:邱 シュウレ, 富士ゼロックス(株)研究技術開発本部,
〒220-8668 神奈川県横浜市みなとみらい 6 丁目 1 番.
E-mail: qiu-xule@fujixerox.co.jp

あるいは、有向ネットワークからも、安定かつ高精度にコミュニティを抽出することができる。本研究では、この方法が二部ネットワークからも高精度にコミュニティを抽出できることを示す。さらに、自動車専門用語辞書「大車林」から構成された「見出し語」－「単語」の二部ネットワークからのコミュニティ抽出を試み、我々の方法が他の代表的なクラスターリング方法よりも高い精度で見出し語をグループ分けすることを示す。

2. 方法

2.1 ランダムウォークに基づくコミュニティ検出方法

まず、以前に提案したコミュニティ検出アルゴリズムについて簡単に振り返る。(詳細は文献[11]を参照)。

提案方法はネットワーク上のランダムウォークを

$$p^{stead}(n)=\sum_{k=1}^{K}\pi_k p(n|k) \tag{1}$$

と分解することを試みる。$p^{stead}(n)$ は、ランダムウォークの定常状態における確率分布である。π_k はコミュニティ k の事前確率であり、$\sum_{k=1}^{K}\pi_k=1$ を満たす。$p(n|k)$ はコミュニティ k におけるノードの確率分布であり、Dirichlet 分布

$$P\left(p_t(1|k),\cdots,p_t(N|k)\right)\sim\prod_{n=1}^{N}\left[p_t(n|k)\right]^{\alpha\sum_{m=1}^{N}T_{nm}p_{t-1}(m|k)} \tag{2}$$

に従う。α は Dirichlet 分布の精度を表すパラメタである。$T_{nm}=A_{nm}\Big/\sum_{n'=1}^{N}A_{n'm}$ はノード m から n への遷移確率である。

このモデルを解くために、リンクにおける学習データ $\left\{\vec{\tau}^{(d)}\right\}\left(d=1,\cdots,D;\sum_{n=1}^{N}\tau_n^{(d)}=2\right)$ を用いて、尤度関数

$$P\left(\left\{z_k^{(d)}\right\},\left\{p_t(n|k)\right\},\left\{\vec{\tau}^{(d)}\right\}\right)\sim\prod_{k=1}^{K}\left\{\pi_k^{\sum_{d=1}^{D}z_k^{(d)}}\prod_{n=1}^{N}p_t(n|k)^{\sum_{d=1}^{D}z_k^{(d)}\tau_n^{(d)}+\alpha\sum_{m=1}^{N}T_{nm}p_{t-1}(m|k)}\right\} \tag{3}$$

を最大化する。ただし、学習データは $p^{stead}(n)$ に基づいて計算する(式5)。$z_k^{(d)}$ は学習データがどのコミュニティ k から生成されたかを表す潜在変数である。そこで、機械学習における標準手法である EM アルゴリズムに従ってモデルの最適化を行う。各変数($\{\pi_k\}$ 及び $\{p(n|k)\}$)は次式で定められる(M-step)。

$$\begin{aligned}p_t(n|k) &= \frac{\tilde{\alpha}}{\tilde{\alpha}+\pi_k}\sum_{m=1}^{N}T_{nm}p_{t-1}(m|k)+\frac{1}{\tilde{\alpha}+\pi_k}\frac{1}{2}\sum_{l=1}^{L}p(l)\gamma_{lk}\tau_n^{(l)},\\ \pi_k &= \sum_{d=1}^{D}\gamma_{lk}\Big/D\end{aligned} \tag{4}$$

ただし $\tilde{\alpha}=\alpha/2D$ である。l はノード n_{init} から n_{end} へのリンクである。$p(l)$ はリンク l の重要度(l の学習データにおける割合)であり

$$p(l)=p^{stead}(n_{init})\cdot T_{mn} \tag{5}$$

で求められる。g_{lk} は $z_k^{(d)}$ の推定であり、ベイズ定理により

$$\gamma_{dk}\equiv P\left(z_k^{(d)}=1|\vec{\tau}^{(d)}\right)=\frac{\pi_k\prod_{n=1}^{N}\left[p_t(n|k)\right]^{\tau_n^{(d)}}}{\sum_{k=1}^{K}\pi_k\prod_{n=1}^{N}\left[p_t(n|k)\right]^{\tau_n^{(d)}}} \tag{6}$$

で求められる(E-step)。

2.2 二部ネットワークからのコミュニティ抽出

二部ネットワークのランダムウォークでは、2種類ノードの間の「確率のスイッチング」の問題が起こる。同じ種類のノードの間にリンク存在しないため、離散マルコフ性により、t 時点に種類 A のノードにおけるランダムウォーク確率がすべて $t+1$時点に種類 B のノードに移動してしまう。そのため、永遠に A と B の間で確率の交換が行われ、ランダムウォークが定常状態に到達しない。そのため、学習データを正しく定めることができない。

しかしながら、離散マルコフ連鎖のかわりに連続時間マスター方程式を用いることにより、定常状態を求めることができる:

$$\lim_{\Delta t\to 0}\frac{p_{t+\Delta t}(n)-p_t(n)}{\Delta t}=-p_t(n)+\sum_m T_{nm}p_t(m) \tag{10}$$

$$\Rightarrow p_{t+\Delta t}(n)=\Delta t\left(-p_t(n)+\sum_m T_{nm}p_t(m)\right)+p_t(n) \tag{11}$$

Δt を1より小さな値(＜1)と設定して(11)式を繰り返し計算することにより、二部ネットワークに対しても定常状態が得られる。こうして得た定常状態を用いて、(5)式により学習データを設定する。

3. 結果

3.1 ベンチマークネットワーク「Southern Women」

Southern Women(ノード数 32、リンク数 89)は、18 人の女性の 14 個のソシャルイベントへの参加関係を表すところの、二部ネットワーク研究のためのベンチマークデータである。提案方法により Southern Women から抽出したコミュニティ構造を図 1 に示す。女性およびイベントを、それぞれ、丸および四角で表す。提案方法により検出された二つのコミュニティを青と赤で表す。ただし、異なるタイプのノードを区別するために、それぞれのコミュニティにおいて種類毎に濃い青と薄い青、濃い赤と薄い赤で表す。

女性の分解結果は、データを収集した研究者達の主張[13]、および biSBM[6] による結果と一致する。Guimera らの biModularity[7]による結果とは、女性ノード“A8”のみの分類が異なった。

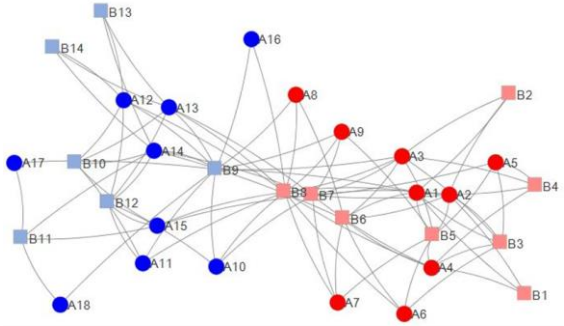

図1 「Southern Women」のコミュニティ分解

3.2 自動車専門用語辞書「大車林」

提案方法を実データ、自動車専門用語辞書「大車林」から構築された用語-説明文単語の二部ネットワーク(「用語」ノード 6,273、「単語」ノー度数 15,210、リンク数 243,870。ただし、リンクを用語と単語の共起回数で重み付けした)にも適用してみた。

人手により付与されたタグを用語の正解カテゴリ(8カテゴリ)として用いる。アルゴリズムによって得られたコミュニティ結果とその正解カテゴリの間の類似度を、アルゴリズムのコミュニティ抽出精度として F 値で計算する。

提案方法と競合として設定した他の代表的クラスタリング方法(テンソル分解モジュール Metafac、K-menas 法、混合ガウスモデル(GMM)、Ward 法)との間でコミュニティ抽出精度の比較を行った。それぞれの方法において、コミュニティ抽出を 20 回試行した。F 値の試行平均を図 2 に示す。我々の提案方法が他の方法よりも高い精度でコミュニティを抽出することがわかる。

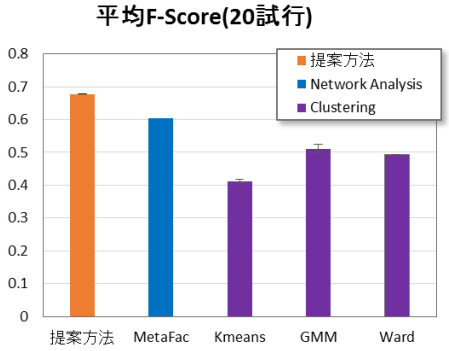

図2　大車林におけるコミュニティ抽出精度の比較結果

4. 議論

現実世界には、二部ネットワークで表現される対象が数多くある。通常一部ネットワークとして扱われているものにも、実は二部ネットワークからの射影であるものがしばしばある。例えば、研究者間の共著関係ネットワークは、研究者-論文の二部ネットワークから射影で生成されたものである。

二部ネットワークのリンク構造は一部ネットワークのそれに比べて一般にスパースである。そのため、密なリンク構造を前提とする多くのコミュニティ抽出方法を二部ネットワークからのコミュニティ検出に適用することは難しい。これまで射影により一部ネットワークに変換することが一般的に行われてきたが、このアプローチには情報損失の問題が伴う。

二部ネットワークを正しく分析するためには、「二部」構造に対応できるコミュニティ抽出方法を構築すべきである。我々は以前に、ランダムウォークに基づくコミュニティ抽出方法を提案した。この提案方法はロバストに高精度でコミュニティを抽出することができる。本研究では、この方法がさらに二部ネットワークからも高精度にコミュニティ抽出することを示した。提案方法をベンチマークネットワーク「Southern Women」および比較的巨大な実二部ネットワーク「大車林」に適用した。「Southern Women」から得られたコミュニティ構造データを収集した研究者達の主張[13]および biSBM[6]による結果と一致する。「大車林」を用いた評価実験では、人手で付与されたタグを用いて提案方法と他の方法の間のコミュニティ抽出精度を比較した。提案方法は他の方法もりも高い精度でコミュニティを抽出した。講演では、biSBM[6]、BRIM[8]、LPAb[9]、LPAb+[10]との比較評価の結果も報告する。

なお、提案方法により抽出されたコミュニティはミックスタイプ（mix-type）である。biSBM が抽出するピュアタイプ(pure-type)なコミュニティに比べ、二種類ノード間の関連性をより豊かに表現すると考えられる。

一部ネットワークについて、提案方法がコミュニティの重なりおよびコミュニティの階層構造[12]を検出できることを以前に示した。この特長は二部ネットワークからのコミュニティ抽出においても保存されると考えられる。すなわち、提案方法は二部ネットワークにおけるコミュニティの重なりと階層構造も検出することができると期待される。

参考文献

1. Blondel V.D., et al. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008.10 (2008): P10008.
2. Lancichinetti A, Fortunato S. Erratum: Community detection algorithms: A comparative analysis [Phys. Rev. E 80, 056117 (2009)][J]. Physical Review E, 2014, 89(4): 049902.
3. Rosvall M. and Bergstrom C. T.. Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. USA, 105:1118--1123, Jan. 2008.
4. Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E, 2007, 76(3): 036106.
5. Ball B, Karrer B, Newman M E J. Efficient and principled method for detecting communities in networks[J]. Physical Review E, 2011, 84(3): 036103.
6. Larremore D B, Clauset A, Jacobs A Z. Efficiently inferring community structure in bipartite networks[J]. Physical Review E, 2014, 90(1): 012805.
7. Guimerà R, Sales-Pardo M, Amaral L A N. Module identification in bipartite and directed networks[J]. Physical Review E, 2007, 76(3): 036102.
8. Barber M J. Modularity and community detection in bipartite networks[J]. Physical Review E, 2007, 76(6): 066102.
9. Barber M J, Clark J W. Detecting network communities by propagating labels under constraints[J]. Physical Review E, 2009, 80(2): 026129.
10. Liu X, Murata T. An Efficient Algorithm for Optimizing Bipartite Modularity in Bipartite Networks[J]. JACIII, 2010, 14(4): 408-415.
11. 岡本　洋. マルコフ連鎖のモジュール分解：ネットワークからの重なりと階層構造を持つコミュニティの検出.JWEIN2014.
12. 邱シュウレ, 岡本　洋. ネットワークからのコミュニティ階層構造の効果的かつ安定な検出. JWEIN2015.
13. Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell system technical journal, 1970, 49(2): 291-307.