K-024

ウェブ上で構成される二部グラフのコミュニティ対応関係

Community Correspondence in Bipartite Graph Constructed on the Web

原田　恵雨 †　鈴木　育男 †　山本　雅人 †　古川　正志 †

Keiu HARADA　Ikuo SUZUKI　Masahito YAMAMOTO　Masashi FURUKAWA

1　はじめに

現実世界に存在する関係には二部グラフとして構成できるものが数多く存在する [1]．たとえば，人間と好きな食べ物の関係，文書と単語の共起関係等である．サイバーワールドが発展した現在では，ウェブ上にも関係を表現する方法として二部グラフがよく採用されている．たとえば，Delicious[2] をはじめとするソーシャルブックマークサービス (SBM) における URL とタグの関係や，EC サイトにおける商品とレビュアーの関係などである．これらは，2 種類のオブジェクト集合をノード集合とし，異種オブジェクトをつなぐ関係をリンクとした二部グラフとして表現される．SBM は社会ネットワークを二次的なものを介したつながりとして，ウェブ上に再構築することで交流を促進させる目的がある．SBM の繋がりを解析することで，推薦やウェブマップ，ディレクトリ自動生成等の応用がなされている．これらの応用は，コミュニティと呼ばれるいくつかの類似したノード集合に分ける技術が基になっている．コミュニティの評価方法，また抽出方法が数多く提案されている [3]．これらの多くが一部グラフのみに適しており，二部グラフを適用するためには一方の部集合のみの関係に重みつき射影変換が必要となる．しかし，二部グラフの部集合はそれぞれ役割が異なるため，このような射影変換では情報が抜け落ちてしまうと考えられる．また，コミュニティ構造の理解が進むにつれ，コミュニティの重複関係が類似しているノード集合をコミュニティとする，複数のコミュニティ対応関係を持つコミュニティ構造が注目を浴びている．これはコミュニティをコミュニティ間の橋渡しであると捉え，二部グラフを全体の構造を維持したまま縮約することで得られる構造である．二部グラフを一部グラフに変換せず，かつ複数のコミュニティ対応関係を考慮したコミュニティ構造の評価・抽出方法が提案されている [4][5]．それにも関わらず，現段階では抽出精度の検証に留まり，抽出されたコミュニティ対応関係を解析する枠組みがなく，二部グラフ全体の機能理解には程遠い現状がある．

本研究では，コミュニティ対応関係と各ノードが対応するコミュニティとの関係を複数の指標を用いて調査することにより，二部グラフにおけるコミュニティ構造が果たす役割の部集合による違いを明らかにする．また，コミュニティ抽出方法により得られるコミュニティ構造が異なるため，異なる方法によって得られたコミュニティ構造による違いも比較する．

2　コミュニティ対応関係の評価方法

コミュニティ対応関係を明らかにするために，コミュニティをノードに縮約したグラフであるコミュニティグラフを導入する．コミュニティグラフの上で対応関係を測る指標を定義することで，縮約する前とは異なる統計的特徴が得られることを期待する．

2.1　コミュニティグラフ

二部グラフ $G = (V_\top, V_\perp, E)$ は上部集合 $V_\top = \{v_1, v_2, \ldots, v_{n_\top}\}$ と下部集合 $V_\perp = \{v_{n_\top+1}, v_{n_\top+2}, \ldots, v_{n_\top+n_\perp}\}$，リンク集合 $E = \{e_1, e_2, \ldots, e_m\}$ の組で表される．ただし，$n_\top$ は上部集合のノード数，$n_\perp$ は下部集合のノード数，m はリンク数，$e \in V_\top \times V_\perp$ である．G 上で，上部コミュニティ $c_\top = \{v | v \in V_\top\}$ と下部コミュニティ $c_\perp = \{v | v \in V_\perp\}$ が定義され，上部コミュニティ集合 $C_\top = \{c_{\top 1}, c_{\top 2}, \ldots, c_{\top N_\top}\}$，下部コミュニティ集合 $C_\perp = \{c_{\perp 1}, c_{\perp 2}, \ldots, c_{\perp N_\perp}\}$ が定義される．ここで，すべての i, j に対して，$c_{\top i} \cap c_{\top j} = \emptyset$ とする．このとき，コミュニティグラフ $\mathcal{G} = (\mathcal{V}_\top, \mathcal{V}_\perp, \mathcal{E}, \mathcal{W})$ は，上部集合 $\mathcal{V}_\top = C_\top$，下部上部 $\mathcal{V}_\perp = C_\perp$ とリンク集合 $\mathcal{E} = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_M\}$，リンク重み集合 $\mathcal{W} = \{\omega_1, \omega_2, \ldots, \omega_M\}$ で表される．ただし，$\epsilon \in C_\top \times C_\perp$，$M$ はリンクの本数である．$\mathrm{src}(\epsilon), \mathrm{tgt}(\epsilon)$ をそれぞれ ϵ の 1,2 番目の要素を参照する関数としたとき，ω_i は $|\mathrm{src}(\epsilon_i) \times \mathrm{tgt}(\epsilon_i) \cap E|$ として定義される．このとき，コミュニティ c 内のノード数 $|c|$ をコミュニティサイズという．また，コミュニティ c 内のノードから出ているリンク数をコミュニティの強さ $\mathrm{str}(c)$ といい，次の式で表す．

$$\mathrm{str}(c) = \sum_i \omega_i \delta(\mathrm{tgt}(\epsilon_i), c) \tag{1}$$

ここで，δ はクロネッカーのデルタとする．ただし，式 (1) は c が下部集合の要素である場合であり，c が上部集合の要素である場合，式中の src を tgt に置き換えたものとなる．

2.2　コミュニティ対応度

コミュニティ同士が互いにどれだけ関係しあうかを測る指標として，いくつかのコミュニティ対応度を提案する．まずは単純に，コミュニティ間の重みを用いた指標 α を提案する．

$$\alpha(\epsilon_i) = \omega_i \tag{2}$$

ただし，α はコミュニティの強さの分散が大きい場合には，適切な指標とはいえない．また，二部グラフの上部集合側から見た対応度とその逆とではコミュニティの強さが異なるため，コミュニティ間の重みを各部集合のコミュニティの強さで正規化した指標 $\beta_\top, \beta_\perp$ と二つの平均である β を提案する．

$$\beta_\top(\epsilon_i) = \frac{\omega_i}{\mathrm{str}(\mathrm{src}(\epsilon_i))} \tag{3}$$

$$\beta_\perp(\epsilon_i) = \frac{\omega_i}{\mathrm{str}(\mathrm{tgt}(\epsilon_i))} \tag{4}$$

$$\beta(\epsilon_i) = \frac{\beta_\top(\epsilon_i) + \beta_\perp(\epsilon_i)}{2} \tag{5}$$

† 北海道大学大学院　情報科学研究科　複合情報学専攻

表 1: コミュニティ抽出方法の違いによる，コミュニティサイズ，コミュニティの強さの違い．(A) と (B) の方法の違いは 3 章を参照．また，$E(x)$ と $SD(x)$ はそれぞれ x の平均値と標準偏差とする．

	(A)	(B)
$N_\top$ / $N_\perp$	117 / 117	13 / 13
$E(\|c\|)$ ($\top$ / $\perp$)	46.6 / 12.3	420 / 110
$SD(\|c\|)$ ($\top$ / $\perp$)	95.2 / 25.7	86.7 / 32.9
$E(\mathrm{str}(c))$	77.3	696
$SD(\mathrm{str}(c))$ ($\top$ / $\perp$)	190 / 193	112 / 118

そのリンクの片方のコミュニティに属するノードが，他方のコミュニティに属するノードに対してリンクを有する度合いが高いほど，これらの値は 1 に近づく．

部集合間の依存関係を調査するために，式 (3),(4) を用いて，指標 $\gamma_\top$，$\gamma_\perp$ を提案する．

$$\gamma_\top(\mathcal{E}) = \frac{1}{N_\top}\sum_{\epsilon\in\mathcal{E}}\left(\beta_\top(\epsilon_i)^2\right) \tag{6}$$

$$\gamma_\perp(\mathcal{E}) = \frac{1}{N_\perp}\sum_{\epsilon\in\mathcal{E}}\left(\beta_\perp(\epsilon_i)^2\right) \tag{7}$$

その部集合のコミュニティが他の部集合のコミュニティに対して，唯一つの対応関係を持つほど，これらの値は 1 に近づく．

3 実験結果

解析対象となるデータは，ウェブ上に存在しユーザの情報共有により構築された二部グラフである，CiteULike[6] のユーザ—グループ間関係 (CUL_UG) を取り上げる．CiteULike は論文専用の SBM であり，ソーシャルサイテーションサービス (SCS) とも呼ばれる．ユーザは興味のある論文を自分のポータルや自分の属するグループに投稿して他のユーザとの情報共有を図ることができる．CUL_UG はユーザ数 5456 とグループ数 1436，リンク数 9046 の二部グラフであり，今回はユーザ側を上部集合，グループ側を下部集合とした．CUL_UG のコミュニティ対応関係を調べるために，CUL_UG をコミュニティ分割し，前章で導入した提案指標を計測する．適用するコミュニティ抽出方法は，

(A) Barber[7] の指標に対して CNM 法 [8] を改良した方法

(B) 原田 [7] の指標に対して CNM 法 [8] を改良した方法

を採用する．

上記のコミュニティ抽出方法を用いた抽出結果の要約を表 1 に示す．得られたコミュニティ数は抽出方法によって開きがあり，単純に比較できないことが分かる．コミュニティサイズの平均と標準偏差に注目すると，方法 (A) では平均に対する標準偏差の値が 2 倍ほどもあり，一方，方法 (B) ではそれが 1/4 ほどに収まっている．コミュニティの強さの平均に対する標準偏差の値では，同様な現象がさらに顕著に見て取れる．ユーザ，グループの違いに注目すると，僅かにグループ側のコミュニティ集合の方が，コミュニティサイズとコミュニティの強さの両方において，平均に対する標準偏差の値の割合が大きい．

式 (6)(7) に従って，$\gamma_\top$，$\gamma_\perp$ を計測してみると，方法 (A) では $\gamma_\top = 0.844$，$\gamma_\perp = 0.889$，方法 (B) では $\gamma_\top = 0.751$，$\gamma_\perp = 0.753$ となり，方法 (A) の方が全体的に値が高い．また，どちらの方法でも僅かにグループ側の値が高い．

4 考察

依存関係を評価する指標 $\gamma_\top$ と $\gamma_\perp$ がどちらの方法でもかなり高い値を保持していることから，CUL_UG におけるコミュニティは，他の部集合との依存関係が高く，橋渡しの役割を果たすコミュニティはほとんどないことが分かる．僅かにユーザ側の $\gamma_\top$ が小さいため，グループ側よりもむしろユーザ側がコミュニティを橋渡しする役割を担っている．また，ユーザ側のコミュニティサイズ及びコミュニティの強さの分散がグループ側のそれらとは異なることも総合して考えると，CUL_UG では，よくグループに入るユーザコミュニティとその逆のパターンが多いことが予想される．

5 まとめと今後の展望

本研究では，二部グラフのコミュニティ抽出方法を用いて抽出した CiteULike ユーザ—グループのコミュニティ間対応関係を調査することによって，CiteULike のユーザ側，グループ側の役割の違いを明らかにした．その結果，各コミュニティが最も対応するコミュニティ以外ほとんど関係を持たず，閉鎖系であることが分かった．僅かに，ユーザ側がグループ間をつなぐ橋渡し役を果たしていることが予想として挙げられた．今後は，ノードが対応するコミュニティを調査することによって，ノード—コミュニティ間関係を明らかにしたい．また，CiteULike 以外の二部グラフに対しても適用することが課題として挙げられる．

参考文献

[1] M. Latapy, C. Magnien, and N. Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, Vol. 30, No. 1, pp. 31–48, January 2008.

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

[3] 池谷智行, 村田剛志. 2 部ネットワークにおけるコミュニティ検出とその評価手法. コンピュータソフトウェア, Vol. 28, No. 1, pp. 91–102, 2011.

[4] 原田恵雨, 鈴木育男, 山本雅人, 古川正志. コミュニティの対応関係を考慮した二部モジュラリティによるコミュニティ分割. コンピュータソフトウェア, Vol. 28, No. 1, pp. 127–134, 2011.

[5] Kenta Suzuki and Ken Wakita. Extracting multi-facet community structure from bipartite networks. *Computational Science and Engineering, IEEE International Conference on*, Vol. 4, pp. 312–319, 2009.

[6] Citeulike. http://www.citeulike.org/.

[7] Michael J. Barber. Modularity and community detection in bipartite networks. *arXiv:0707.1616*, Nov 2007.

[8] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, Vol. 70, No. 6, p. 66111, 2004.