F-042

ノード重要度を考慮したグラフ分割構造の学習
Learning Graph Partitioning Structures Based on Node Importance

藤巻遼平 † 山西健司 †
Ryohei Fujimaki Kenji Yamanishi

1 はじめに

近年，グラフ構造に対するマイニング技術が注目を集め，Blog/SNS ネットワーク，たんぱく質相互作用ネットワークなど，幅広い対象の分析へ応用されている．グラフ構造の分析技術は，リンクに着目した分析と，ノードに着目した分析の二つに大別される．

グラフ分割は，前者の最も代表的な話題の 1 つである．この技術は，一般的にはリンクの構造 (隣接行列) から密な部分グラフを発見し，コミュニティ構造の分析などに利用され，グラフ理論や情報理論に基づく分割アルゴリズムが提案されている [4, 6]．また，近年ではそのようなグラフ分割構造の時間変化を捉える事も重要な課題として認識されている [13]．

後者の代表的な話題としては，ノード重要度分布の推定が挙げられる．特に，スケールフリーネットワーク [1] の研究の隆盛と共に，実際のネットワークでは，リンク次数，PageRank [9] などに代表されるノード重要度が，冪分布や対数正規分布など長い裾を持った特徴的な分布に従うことが明らかとなり [1, 2]，この特徴を実応用へ生かす方策が模索されている．

本稿では，図 1 に示されるように，リンクおよびノード重要度の同時分割構造を学習するためのアルゴリズムを提案する．両者を同時に学習する事によって，グラフ中でどのスケールのノードが重要な役割を果たし，またそれらの相互作用がいかなる様相であるかを分析可能となる．より直感的には，ハブとハブの間の相互作用，ハブと非ハブの相互作用など，リンク構造だけからは知りえなかった新たな構造を学習しているといえる．これは，複雑ネットワークの背景にある階層構造 [10] を特徴付けるノード群の発見という点で重要である．さらに，本研究はこれまで議論の少なかったリンクとノード重要度の同時構造に，統計的に明確なモデルを定義したという観点からも意義深い．

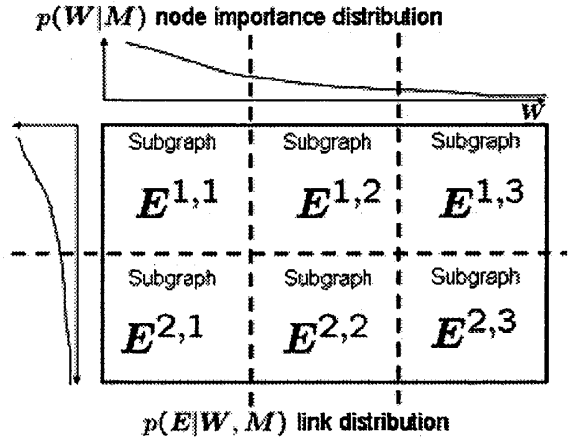

図 1: 学習の概念図

†NEC 共通基盤ソフトウェア研究所

2 同時分布モデル

本稿では，無向かつリンクが $\{0,1\}$ の 2 値をとるグラフを対象とする．まず，$\boldsymbol{E} \in \mathcal{E}^{n\times n}, \boldsymbol{W} \in \mathcal{W}^n, \boldsymbol{M}$ を，各々リンク，ノード重要度，モデルを表す確率変数とする．ただし，$\mathcal{E} = \{0,1\}$ かつ $\mathcal{W} = [0, R]$ $(R > 0)$ で，n はグラフ中のノード数とする．全体の同時分布は，

$$P(\boldsymbol{E}, \boldsymbol{W}, \boldsymbol{M}) = P(\boldsymbol{E}|\boldsymbol{W}, \boldsymbol{M}) \cdot P(\boldsymbol{W}|\boldsymbol{M}) \cdot P(\boldsymbol{M}). \quad (1)$$

となる．以下では，分割数を m $(\leq \min\{n, R\})$ として議論を進める．

まず，本稿では $\boldsymbol{E}$ は $\boldsymbol{M}$ によって m^2 の長方形の部分グラフへ分割される構造を仮定する．さらに，(p,q) 番目の部分グラフを $\boldsymbol{E}^{p,q} \in \mathcal{E}^{n_p \times n_q}$ とすると，$\boldsymbol{E}^{p,q}$ の各要素は，1 の確率が $\theta_{p,q}$ の独立なベルヌーイ分布に従うと仮定する．これは，グラフ上に拡張された確率的規則 [14] という解釈をする事が可能である．数式的には，

$$P(\boldsymbol{E}|\boldsymbol{W}, \boldsymbol{M}) = \prod_{p=1}^{m}\prod_{q=p}^{m}\prod_{i=1}^{n_p}\prod_{j=1}^{n_q}\left(\theta_{p,q}^{E_{i,j}^{p,q}}(1-\theta_{p,q})^{1-E_{i,j}^{p,q}}\right).$$

と表現される．

次に，$\boldsymbol{W}$ は $\boldsymbol{M}$ によって，m 個の領域へ分割される構造を仮定する．すなわち，値域 $[0, R]$ が $[0, a_1], (a_1, a_2], \ldots, (a_{m-1}, R]$ $(0 < a_1 < \ldots < a_{m-1} < R)$ と分割される．各領域において $\boldsymbol{W}$ の要素が一様に分布していると仮定すれば，$P(\boldsymbol{W}|\boldsymbol{M})$ はヒストグラム密度関数として表現される．すなわち，p 番目の領域に対する確率を ϕ_p とすれば，

$$P(\boldsymbol{W}|\boldsymbol{M}) = \prod_{p=1}^{m}\left(\frac{\phi_p}{R_p}\right)^{n_p}. \quad (2)$$

と表現される．

最後に，$P(\boldsymbol{M})$ に対しては，

$$P(\boldsymbol{M}) \propto 2^{-\ell(\boldsymbol{M})}, \quad \sum_{\boldsymbol{M}\in\mathcal{M}} 2^{-\ell(\boldsymbol{M})} \leq 1 \quad (3)$$

なる構造を仮定する．ここで，$\ell(\boldsymbol{M})$ に対する不等式はクラフトの不等式 [5] と呼ばれ，$\boldsymbol{M}$ に対し記述長が $\ell(\boldsymbol{M})$ である語頭符号が存在するための必要十分条件である．

3 最小記述長原理に基づく分割構造学習

一般には，分割の数や分割の位置といったモデルに関する情報は未知のため，モデル選択の問題を解く必要がある．本稿では，Rissanen [11] の最小記述長原理 (Minimum Description Length; MDL) に基づき，同時分布構造を学習する．これはリンクとノード重要度の構造を MDL によって統一的に最適化する点で，従来の MDL を利用したグラフ分割手法 [3, 4] とは一線を画すものである．

まず，$\boldsymbol{E}, \boldsymbol{W}$ に対する観測値 $\boldsymbol{e}, \boldsymbol{w}$ が与えられたとき，モデルを含めた全体の記述長を

$$\ell(\boldsymbol{e}, \boldsymbol{w}, \boldsymbol{M}) = \ell(\boldsymbol{e}|\boldsymbol{w}, \boldsymbol{M}) + \ell(\boldsymbol{w}|\boldsymbol{M}) + \ell(\boldsymbol{M}). \quad (4)$$

と展開する．ここで，$\ell(\boldsymbol{e}|\boldsymbol{w},\boldsymbol{M}) \equiv -\log P(\boldsymbol{e}|\boldsymbol{w},\boldsymbol{M})$, $\ell(\boldsymbol{w}|\boldsymbol{M}) \equiv -\log P(\boldsymbol{w}|\boldsymbol{M}), \ell(\boldsymbol{M}) \equiv -\log P(\boldsymbol{M})$ であり，log の底は 2 である．これらは，$\boldsymbol{e},\boldsymbol{w},\boldsymbol{M}$ を与えられた条件の下，語頭符号によって記述した場合の最小記述長を与える [7]．

まず，$\bar{\theta}_{p,q}$ を $\theta_{p,q}$ の最尤推定量とし，$\ell\big(vec(\boldsymbol{e}^{p,q})\big) = n_p n_q H(\bar{\theta}_{p,q}) + \frac{1}{2}\log(n_p n_q \pi/2)$ と定義すれば，リンクの記述長は

$$\ell(\boldsymbol{e}|\boldsymbol{w},\boldsymbol{M}) = \sum_{p=1}^{m}\sum_{q=p}^{m} \ell\big(vec(\boldsymbol{e}^{p,q})\big). \tag{5}$$

と計算される．ただし，n_p に関してはノード重要度の記述に含めるためリンクの記述からは除外している．また $H(\theta)$ はベルヌーイ分布のエントロピーで，$vec(\cdot)$ は行列をベクトルへ変換するオペレータとする．

次に，ノード重要度の記述は Rissanen らの提案するヒストグラムの記述方法 [12] を利用して，

$$\ell(\boldsymbol{w}|\boldsymbol{M}) = \sum_{p=1}^{m} n_p \log R_p + \log\frac{(n+m-1)!}{n!(m-1)!} + \log\frac{n!}{\prod_{p=1}^{m} n_p!}. \tag{6}$$

と記述される．ただし，$R_p = a_{p+1} - a_p$ である．

最後に，$a = dk$，ビンの最小サイズを κ，$\gamma = R/d$ とし，$\ell(m,d,\kappa,\gamma) = \log^* m + \log^* d + \log^* \kappa + \log^* \max(1, \gamma - m\kappa)$, $\ell(\boldsymbol{k}) = \frac{(\gamma - m(\kappa-1)-1)!}{(\gamma-m\kappa)!(m-1)!}$ と定義すると，モデルの記述長は

$$\ell(\boldsymbol{M}) = \ell(m,d,\kappa,\gamma) + \ell(\boldsymbol{k}). \tag{7}$$

と計算される [12]．

以上をまとめると，提案手法では以下の最適化問題を解くことによって最適な分割 $\boldsymbol{M}^*$ を計算する．

$$\boldsymbol{M}^* = \arg\min_{\boldsymbol{M}\in\mathcal{M}} \{\ell(\boldsymbol{e},\boldsymbol{w}|\boldsymbol{M}) + \ell(\boldsymbol{w}|\boldsymbol{M}) + \ell(\boldsymbol{M})\}, \tag{8}$$

計算方法の詳細に関しては紙面の都合上割愛する．

4 実験と考察

本稿では，Lesmis データ [8] に対して提案手法を適用した．このデータは，ヴィクトル・ユーゴーの代表著書「レ・ミゼラブル」中の単語の共起関係の無向グラフで，ノードは単語に対応し，リンクは対応する二つの単語が同じ章で使われた事を表している．元データ [8] は共起頻度を表す重み付無向グラフであるが，本稿では重みは無視している．グラフは 77 ノード，508 リンクから構成されている．

図 2 および図 3 に，学習されたノード重要度の分布およびリンク生成の分布を示す．まず，ノード重要度に関しては，スケールフリーネットワークの特徴とされる裾の長い分布が学習されている事が確認できる．さらに，リンクの分布に着目すると，部分グラフ $\boldsymbol{e}^{2,2}$ が $\boldsymbol{e}^{2,3}$ や $\boldsymbol{e}^{2,4}$ と比較してリンクの発生確率が高い．一般には，リンクを多く持たないノードは，リンクを多く持つノードと結びつくため，2 番目の分割領域に属するノードは非常に強い自己連結性を持ち，ノード重要度の高いハブと共にこのネットワークにおいて中心的な役割を果たしている事がわかる．このような中心的な役割を果たすノード群の発見は，マーケティングやネットワークデザインなど，実応用の観点からも重要性が高く，提案手法によって，従来の密な部分グラフを検出する手法とは異なる切り口から，ネットワークの特徴的な構造が分析可能となる．

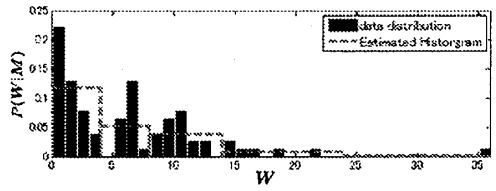

図 2: Lesmis データに関するノード重要度の観測値と推定されたヒストグラム密度関数．

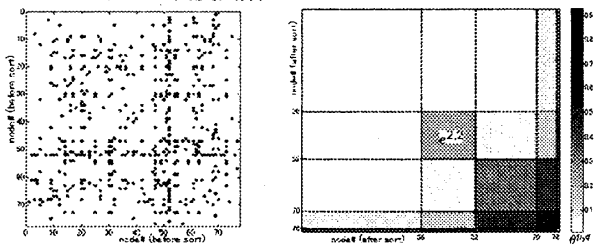

図 3: Lesmis データに関する隣接行列 (左) と推定されたリンクの分布 (右)．

5 結論

本稿では，リンクとノード重要度に関する同時分布に基づくグラフ分割方法を提案した．提案手法では，ノード重要度の情報を考慮することによって，ハブと非ハブのインタラクション構造など，リンク構造だけからは知りえなかった新たな構造を学習可能である．昨今，複雑ネットワーク分析技術の重要性が高まっており，提案手法のようにノード 2 点間のリンク以外の情報を考慮して全体の構造を推定することは，ますます重要となると考えられる．

参考文献

[1] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[2] L. Becchetti and C. Castillo. The distribution of pagerank follows a power law only for particular values of the damping factor. In *Proceedings of the 15th WWW*, 2006.

[3] D. Chakrabarti. Autopart: parameter-free graph partitioning and outlier detection. In *Proceedings of the 8th PKDD*, pages 112–124, 2004.

[4] D. Chakrabarti, S. Papadimitriou, D. S. Modha, and C. Faloutsos. Fully automatic crossassociations. In *Proceedings of the 10th KDD*, pages 79–88, 2004.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[6] G. Flake, S. Lawrence, and C. Giles. Efficient identification of web communities. In *Proceedings of the 6th KDD*, 2000.

[7] P. D. Grunwald, I. J. Myung, and M. A. Pitt. *Advances in minimum description length*. MIT Press, 2005.

[8] D. E. Knuth. *The Stanford GraphBase: A Platform for Combinatorial Computing*. Addison-Wesley, 1993.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report,Stanford Digital Library Technologies Project, 1998.

[10] E. Ravasz and A. Barabasi. Hierarchical organization in complex networks. *Phys. Rev. E*, 67:026112, 2003.

[11] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[12] J. Rissanen, T. P. Speed, and B. Yu. Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, 38(2):315–323, 1992.

[13] J. Sun, P. S. Yu, S. Papadimitriou, and C. Faloutsos. Graphscope: Parameter-free mining of large time-evolving graphs. In *Proceedings of the 13th KDD*, 2007.

[14] K. Yamanishi. A learning criterion for stochastic rules. *Machine Learning*, 9:165–203, 1992.