DEIM Forum 2023 2b-6-3

遷移確率を調整したランダムウォークに基づくハイパーグラフ埋め込み

永里　和哉†　　高邉　賢史†　　首藤　一幸††

† 東京工業大学情報理工学院　〒 152–8550 東京都目黒区大岡山 2 丁目 12–1
†† 京都大学学術情報メディアセンター　〒 606–8501 京都府京都市左京区吉田二本松町

あらまし　本論文では，ランダムウォークを用いてパイパーグラフの埋め込みを行うことを考える．ハイパーグラフ上でランダムウォークを実行し，得られたノード列を自然言語処理で用いられる skip-gram に入力する事で，グラフ構造を捉えたベクトル表現を得る事ができる．我々は，ハイパーグラフ向けに遷移確率を調整可能なランダムウォーク手法を提案する．その結果として，グラフの特徴をより適切に表現した埋め込みが可能となることを主張する．提案手法におけるパラメータを適切に調整することで，ノードのラベル分類といった機械学習タスクに対して大きなハイパーグラフであっても現実的な計算時間で高い精度の結果が得られることを実験により示した．

キーワード　グラフ処理，ハイパーグラフ，ランダムウォーク，埋め込み

1 はじめに

人と人やモノとモノの関係性を定量的に扱うにあたって，グラフは強力な数理モデルである．例えば，WWW（World Wide Web）のハイパーリンクや SNS （Social Networking Service）の友人ネットワーク，化学化合物の構造などはグラフによって数理的に表現されることが多く，近年急速に研究が進められている．中でも近年の機械学習手法の発展から，非構造データであるグラフに対してグラフ全体やグラフのノードを低次元のベクトルとして表現する埋め込み手法は注目を集め，グラフの分類やノードの分類，新しいノードが獲得するエッジの予測などに応用されている．

一方で，実ネットワークには集団として相互作用を持つ関係性も存在し，2 つのノード間を 1 つのエッジが結ぶ一般のグラフではそれらは表現しきれない．そこで，グラフの一般化であるハイパーグラフという数理モデルが研究対象となる．ハイパーグラフにおけるエッジは任意の個数のノードを含むことが可能であり，ノード間の高次の関係性を捉える事ができる．

本論文において我々は，ハイパーグラフ向けに遷移確率を調整可能なランダムウォーク手法を提案し，ノードの埋め込みに応用する．実験としてノードのラベル推定を実行し，提案手法は既存手法と同程度の推定結果を得られ，空間計算量を削減できることを示す．

2 関連研究

グラフ埋め込みに関する研究は近年多くなされている．spectral embedding [5] はグラフ信号処理に基づく手法であり，隣接するノード間のベクトルの距離が近くなるようにベクトル表現を得る．また，これを派生させたグラフニューラルネットワークの研究 [6] は非常に多くなされている．しかし，これらの手法は入力としてサイズ $n \times n$ の隣接行列を用いるため，空間計算量が大きく大規模データセットには向かない．

また，ランダムウォークを用いたグラフの埋め込み手法として DeepWalk [1] や node2vec [2] が存在する．これらの手法は，ランダムウォークで得られたノード列を skip-gram [4] に入力することで各ノードのベクトル表現を得る．ランダムウォークは並列化可能であるため，ランダムウォークを用いた手法は時間計算量の削減が期待される．

また，ハイパーグラフ上でのランダムウォークに関する研究もなされている．Carletti らの研究 [3] によると，ハイパーグラフにおけるランダムウォークの遷移において，隣接するノードを等確率で選択するのは賢明な判断とは言えない．なぜならば，ハイパーエッジに属するノード間での繋がりは連続した個人間での繋がりよりも強いことが多いからである．ゆえに，Carletti らは隣接するノードに対して，属しているハイパーエッジ内のノード数から自身を除いた数をノードの重みとし，その重みに従って遷移確率を定義した．この定義によるランダムウォークは，高次元ハイパーエッジを多く訪問することができる [3]．

また，Chitra らはエッジに依存したノードの重みが存在するハイパーグラフに対して遷移確率を定義し，ランダムウォークを定式化した [7]．Chitra らによると，ハイパーグラフにおいてノードの重みがエッジに依存しない場合，ハイパーグラフ上でのランダムウォークはクリーク表現に射影した一般の重み付きグラフ上でのランダムウォークと等価となる．

3 準　　備

3.1 記　　法

本論文では，ハイパーグラフ $\mathcal{H}(V, E)$ に対して，$V = \{1, ..., n\}$ をノードの集合，$E = \{E_1, ..., E_m\}$ をハイパーエッジの集合とする．ここで，n はノード数，m はハイパーエッジ数である．各 E_α は V の部分集合であり，特に任意の α に対して $|E_\alpha| = 2$ の時は一般のグラフと等価である．

今，ハイパーグラフの接続行列 e を

$$e_{i\alpha} = \begin{cases} 1 & \text{for } i \in E_\alpha \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

と定めると，ハイパーグラフの隣接行列 A は $A = ee^T$, ハイパーエッジ行列 C は $C = e^T e$ と表される．ここで，e はハイパーエッジとノードの包含関係を表す行列であり，A_{ij} はノード i とノード j の共有するハイパーエッジの数を，$C_{\alpha\beta}$ は $E_\alpha \cap E_\beta$ に含まれるノード数を表す．

3.2 ハイパーグラフの射影

ハイパーグラフにおいてハイパーエッジは任意個のノードを含むことができるため自由度が高く，ハイパーグラフをそのまま扱うことは難易度が高い．そこで，ハイパーエッジを全てのノード同士がエッジを持つ完全グラフ（クリーク）として表現することで，ハイパーグラフを一般の重みなし無向グラフとして扱うクリーク展開と呼ばれる手法を用いた研究が盛んになされている．しかし，クリーク展開は不可逆な変換であり，ハイパーグラフとしての情報が少なからず損なわれてしまう．そこで，本論文ではハイパーグラフをクリーク展開した重み無し無向グラフ上でのランダムウォークを比較対象の 1 つとする．

3.3 ランダムウォークと定常分布

一般に，グラフのノードを状態と定義し，ノード i からノード j への遷移確率 T_{ij} を定めると，グラフ上のノードの遷移はランダムウォークと見なすことができる．また，次の状態が現在の状態のみによって定まる確率過程をマルコフ連鎖と呼ぶ．エルゴード的なマルコフ連鎖は定常分布 $\boldsymbol{p}$ を持ち，

$$\boldsymbol{p} = \boldsymbol{p}T \tag{2}$$

を解くことで得られる．(2) 式より，ランダムウォークの定常分布は遷移確率行列 T の固有値 1 に対応する右固有ベクトルであることがわかる．

3.4 skip-gram

私たちはノードの埋め込み手法として，skip-gram を採用する [4]．skip-gram はノードの d 次元ベクトル表現を得る関数 $f : V \to \mathbb{R}^d$ を学習するニューラルネットワークであり，自然言語処理において単語の共起確率を最大化する目的で設計された．本研究ではハイパーグラフ上でランダムウォークを実行し，得られたノード列を入力とする事でノード間の関係性を捉えたノードのベクトル表現を得ることを目的とする．

4 提案手法

本章では，パラメータによって挙動を制御できるランダムウォークを提案する．データセットに合わせたパラメータを設定し，ランダムウォークを実行する事でグラフ構造を捉えたノード群を得ることができる．そして，得られたノード群を skip-gram に入力することで各ノードのベクトル表現を得ることができる．

4.1 ランダムウォーク

現在ノードが i である場合，隣接するノード j に以下で表される重み $k_{ij}^H(\beta)$ を付ける．

$$k_{ij}^H(\beta) = \sum_\alpha (C_{\alpha\alpha} - 1)^\beta e_{i\alpha} e_{j\alpha} \tag{3}$$

この定義は先行研究 [3] にパラメータ β をべき乗数として加えた重み付けであり，先行研究の拡張となっている．

そして (3) 式を用いて，ノード i からノード j への遷移確率 T_{ij} を

$$T_{ij} = \frac{k_{ij}^H(\beta)}{\sum_{l \neq i} k_{il}^H(\beta)} \tag{4}$$

と定義する．(4) 式は β の関数であり，β の値に応じて様々な挙動のランダムウォークが実現される．$\beta = 1$ の場合は先行研究 [3] と等価であり，β を 1 より大きくすると高次元ハイパーエッジをより多く訪問する挙動を示すことが予想される．また，β を小さくすると低次元ハイパーエッジを多く訪問と予想される．なお，$\beta = 0$ の時は隣接するノードを等確率に訪問するため，ハイパーエッジをクリーク展開したグラフでのランダムウォークと等価である．

以上より，提案ランダムウォークにおける私たちの仮説は以下のようにまとめられる．

- $\beta > 1$ 高次元ハイパーエッジの構造を多く抽出
- $\beta = 1$ 先行研究 [3] と等価
- $\beta < 1$ 低次元ハイパーエッジの構造を多く抽出
- $\beta = 0$　クリーク展開したハイパーグラフと等価

また，ハイパーグラフが与えられた β に対してエルゴード的であるとき，ランダムウォークは定常分布を持ち

$$p_j^\infty = \frac{d_j^H(\beta)}{\sum_j d_j^H(\beta)} \tag{5}$$

で与えられる．ここで，$d_j^H(\beta) = \sum_{l \neq j} k_{jl}^H(\beta)$ と定義した．

5 実　　験

まず簡単な人工ハイパーグラフ (図 1) を用いて，提案ランダムウォークの挙動を確認する．その後，3 つのデータセットに対してパラメータを変更したときのラベル推定タスクにおける精度を確認する．

5.1 定常状態における遷移確率

人工的なハイパーグラフ (図 1(a)) について，(5) 式で与えられた定常分布を計算し，ランダムウォークの挙動を確認した．

結果として，$\beta = -1$ の時はクリーク展開したハイパーグラフにおけるランダムウォークに比べ，高次元ハイパーエッジでの滞在確率が低く，低次元ハイパーエッジに含まれるハブでの滞在確率が高い結果が得られた．一方で，$\beta = 2$ とすると先行研究 [3] である $\beta = 1$ よりも高い確率で高次元ハイパーエッジに含まれるノードに滞在し，低次元ハイパーエッジに含まれるハブでの滞在確率は低くなった (図 5(b))．

また，各ノードの滞在確率を見ても，高い β は高次元ハイパーエッジへの高い潜在確率を誘導し，逆も同様であった (図 5(c))．

以上の結果から我々の仮説は正しく，パラメータ β を変える

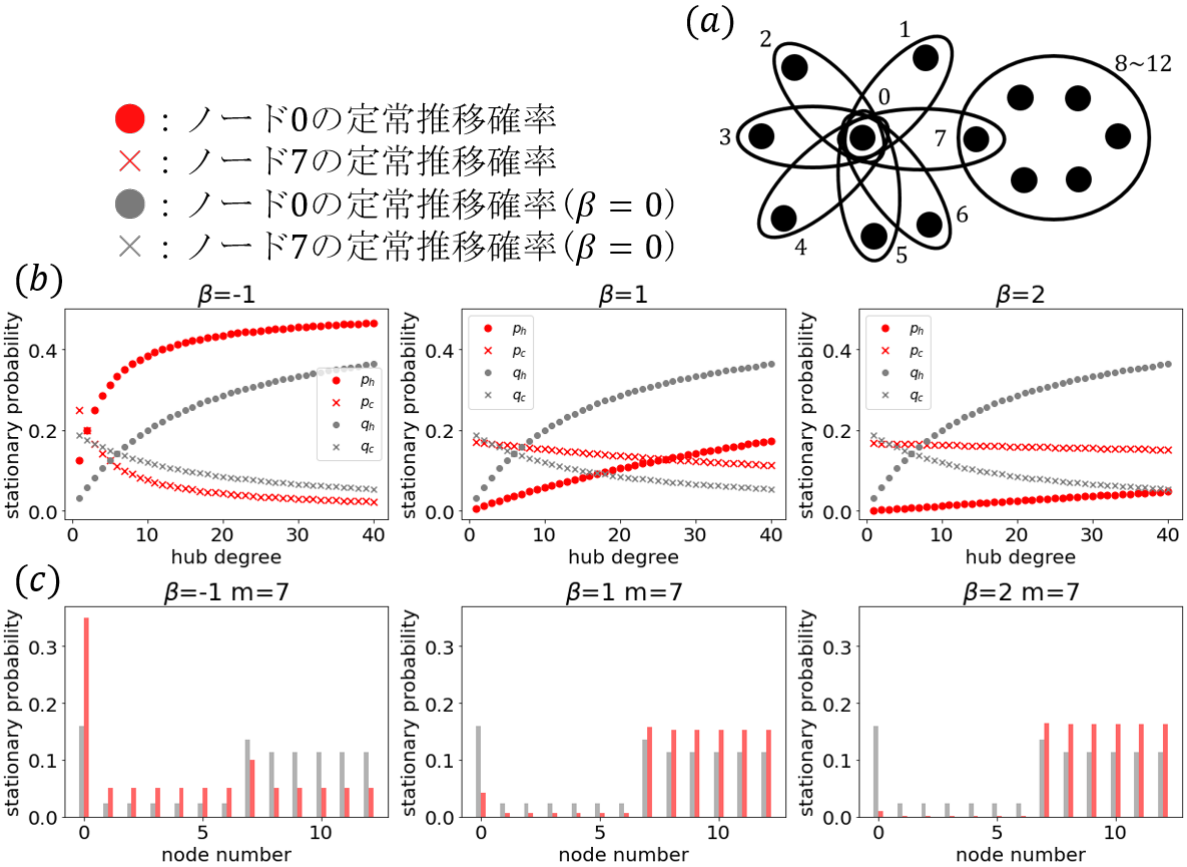

図 1　(a) は人工的なハイパーグラフ，(b) はハブの次数を変化させたときのノード 0 とノード 7 の定常推移確率，(c) はノード 0 の次数を 7 と固定したときの定常分布である．なお，(b)，(c) では提案手法（赤）とクリーク表現でのランダムウォーク（灰）を比較した．

ことでランダムウォークの挙動を制御できていることがわかる．

5.2　ラベル推定

この節では，提案ランダムウォークで得られたノード列を Skip-gram に入力し，各ノードのベクトル表現を得る事でノードのラベル推定タスクを実行する．ランダムウォークを実行する際は各状態で逐一遷移確率を計算することにより，遷移確率行列を明示的に記述することなくウォークを行う．これにより，空間計算量を大幅に削減することが可能となり大規模なデータセットに対しても実験を行うことができる．

実験の手順は以下の通りである．

（1）　ハイパーグラフ内の各ノードからそれぞれ γ 回，ウォーク長 t のノード列を提案ランダムウォークにより取得する．

（2）　取得したノード列を Skip-gram に入力し，各ノードに対して d 次元のベクトル表現を得る．

（3）　教師データの割合をノード数の 8 割とし，ロジスティック回帰にて残りのノードのラベルを推定する．

（4）　(1)~(3) を各 β に対して s 回繰り返し，F1 スコアの平均値と標準誤差を得る．

表 1　データセット

データセット	n	m	avg$\lvert E_i\rvert$	max$\lvert E_i\rvert$	ラベル数
senate-bills	294	29,157	8	99	2
contact-primary-school	242	12,704	2.4	5	11
mathoverflow-answers	73,851	5,446	24.2	1,784	1,456

データセットは表 1 [8]-[12] をを使用し，senate-bills と contact-pirmary-school データセットに対しては $\gamma \in \{8, 16, 32, 64\}$，$t = 20$，$d = 64$，$\beta \in \{-1, 0, 1\}$，$s = 20$ mathoverflow-answers に対しては $\gamma \in \{8, 16\}$，$t = 20$，$d = 128$，$\beta \in \{-1, 0, 1\}$，$s = 1$ と設定し実験を行った．ここで，$\beta = 0$ とするとクリーク展開したグラフでの deepwalk と等しく，$\beta = 1$ とした時のランダムウォークは先行研究 [3] と等しい．また，比較手法として spectral embedding [5] を用い，

表 2 にその結果を示した．なお，spectral embedding は計算過程において $n \times n$ 行列を用いるため，今回の実験において mathoverflow-answers データセットではメモリ不足のため実行することができなかった．

3 つのデータセットに対する実験結果を図 2 に示す．図 2 から，良い β を選択することにより spectral embedding とほとんど同等の精度が得られたことが分かる．特に，senate-bills データセットでは $\beta = -1$ とすることで，少ないサンプル数でも高い F1 スコアが得られ，既存手法である $\beta = 0$ を大幅に上

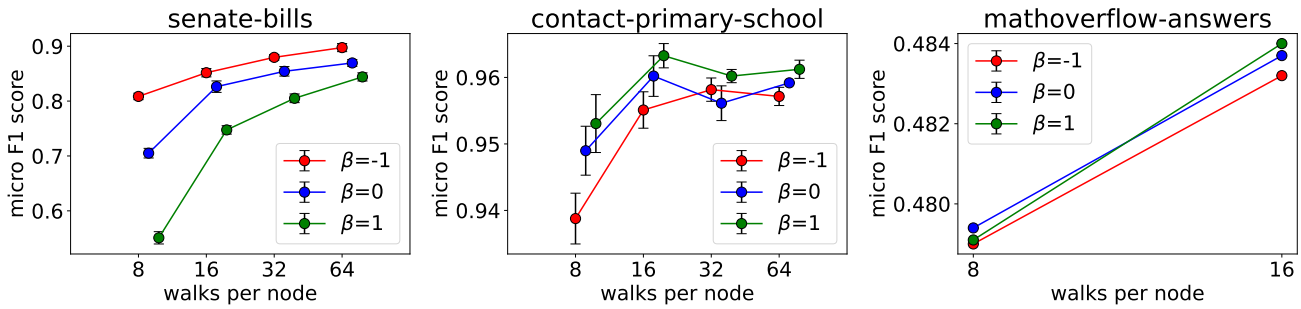

図 2　各データセットに対するラベル推定による実験結果．図のエラーバーは標準誤差を表す．

表 2　spectral embedding の実験結果

データセット	micro F1 score
senate-bills	0.949
contact-primary-school	0.980

回った．また，提案手法は空間計算量を抑える事ができるため，大規模データセットである mathoverflow-answers データセットに対しても結果を得る事ができ，更に 1,456 のラベル数において半数近いラベルを正答する事ができた．

5.3　F1 スコアのパラメータ依存性

この節では，パラメータ β による F1 スコアの変化を示す．senate-bills，contact-primary-school の 2 つのデータセットに対して，β の値をそれぞれ-3 から 3 まで，-3 から 4 まで 0.5 刻みで動かし実験を行った（図 3）．ここで，β があまりに大きいとランダムウォークは高次元ハイパーエッジから脱出できなくなり，β が小さすぎると低次元ハイパーエッジから脱出できなくなるため，これらの場合ランダムウォークによりノード間の関係性を抽出できなくなるという仮説の元，β の値を 0 付近に留めた範囲設定とした．図 3 より，senate-bills は $\beta = -0.5$ 辺りに，contact-primary-school は $\beta = 1.5$ 辺りに最大値を持つ，β に関する凸関数となっていることがわかる．よって，データセットに合わせて適切にパラメータ β を探索することにより，提案手法を用いた結果を最適化することができる可能性がある．

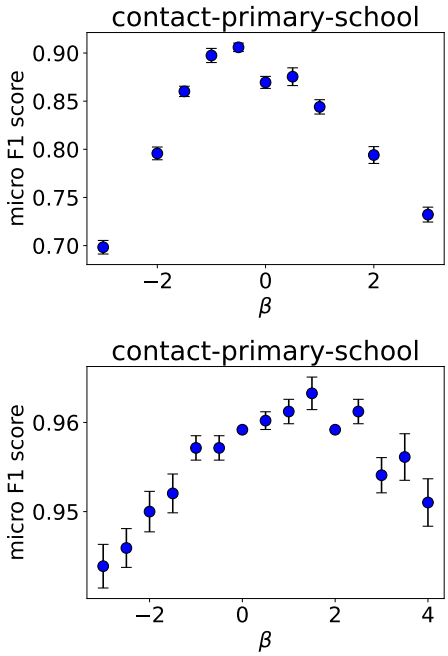

図 3　F1 スコアに対する β 依存性．図のエラーバーは標準誤差を表す．

6　結　　論

本論文において，我々はハイパーグラフ向けに遷移確率を調整可能なランダムウォークを提案し，埋め込みに応用した．

提案ランダムウォークは我々の仮説に従い，パラメータを変化させることで自在な挙動を示すことを人工的なハイパーグラフを用いる事で確認した．また，ハイパーグラフ上でランダムウォークを実行する事で大規模なデータセットであっても埋め込みを行うことが可能であり，実データを用いたラベル推定タスクにおいても既存手法と同等以上の高い精度を示した．そして，提案手法におけるパラメータはデータセットに対して最適化できる可能性があり，パラメータを最適化する事でより適切な埋め込みが可能となる．

文　　献

[1] B. Perozzi, R. AI-Rfou, and S. Skiena “DeepWalk: Online learning of social representations”, In KDD, 2014, p. 701-710.
[2] A. Grover and J. Leskovec “node2vec: Scalable feature learning for networks”, In KDD. New York: ACM; 2016. p. 855–64
[3] T. Carletti, F. Battiston, G. Cencetti and D. Fanelli “Random walks on hypergraphs”, Physical Review E, 101 ,2020 , p. 022308.
[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean. “Efficient estimation of word representations in vector space”, In ICLR, 2013.
[5] L. Tang and H. Liu. “Leveraging social media networks for classification”, Data Mining and Knowledge Discovery, 23(3):447–478, 2011.
[6] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P.S. Yu. “A Comprehensive Survey on Graph Neural Networks” , IEEE Trans. Neural Netw. Learn. Syst. 32 (1), 4–24.
[7] U. Chitra and B. J Raphael. “Random walks on hypergraphs with edge-dependent vertex weights”, In Proceedings of the 36th International Conference on Machine Learning (ICML), 2019.
[8] P S. Chodrow, N. Veldt and A R. Benson. “Generative hypergraph clustering: from blockmodels to modularity” ,Science Advances, 2021.

[9] J H. Fowler. “Connecting the Congress: A Study of Cosponsorship Networks” , James H. Fowler.

[10] J H. Fowler. “Legislative Cosponsorship Networks in the U.S. House and Senate” , Social Networks, 2006.

[11] J. Stehle, N. Voirin, A. Barrat, V. Vattuto, L. Isella, J. Pinton, M. Quaggiotto, W V. Broeck, C. Regis, B. Lina and P. Vanhems. “High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School” , PLOS ONE, 2011.

[12] N. Veldt, A R. Benson and J. Kleinberg. “Minimizing Localized Ratio Cut Objectives in Hypergraphs” , Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2020.

[13] M. Gjoka, M. Kurant, C. T. Butts, A. Markopoulou. “Walking in Facebook: A Case Study of Unbiased Sampling of OSNs” , InProc. IEEE INFOCOM, 1–9.

[14] K. Hayashi, S. G. Aksoy, C. H. Park and H. Park. “Hypergraph Random Walks, Laplacians, and Clustering” , CIKM ’20, October 19–23, 2020, Virtual Event, Ireland.

[15] D. Zhou, J. Huang, and B.Scholkopf. “Learning with Hypergraphs: Clustering, Classification, and Embedding” , In NIPS, 2007, 1601–1608.

[16] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao. “Hypergraph neural networks” , In AAAI, 2019, 3558–3565.

[17] M. Belkin and P. Niyogi. “Laplacian eigenmaps for dimensionality reduction and data representation” , Neural computation, 15(6):1373–1396, 2003.

[18] P. Goyal and E. Ferrara. “Graph embedding techniques, applications, and performance: A survey” , Knowledge-Based Systems, vol. 151, pp.78–94, 2018.