CS 292F.300 Final project proposal

Zexi Huang

TOTAL POINTS

1 / 1

QUESTION 1

1 Proposal submitted **1 / 1**

✓ **- 0 pts Correct**

This sounds great. I look forward to seeing the results.

Efficient Node Embedding by Eigendecomposition of Random-walk Similarity Matrix via Lanczos Iterations

Zexi Huang

Node embedding enables the application of classical machine learning algorithms for high-dimensional data to graph-based downstream tasks, such as link prediction, node classification, and community detection. The goal of embedding is to learn vector representations for nodes based on some notion of topological similarity (or proximity) in the graph.

In our recent work, we have shown that most existing node embedding methods implicitly or explicitly preserve two random-walk based similarity metrics, i.e., Pointwise Mutual Information (PMI) and autocovariance. The PMI similarity is the basis of several pioneering sampling-based embedding methods, such as *Deepwalk* [1] and *node2vec* [2], and its close matrix form has been found in [3]. On the other hand, the autocovariance similarity is more popular in the context of multiscale community detection [4, 5], and has recently been introduced for embedding as well [6].

In this project, we aim to find efficient ways to compute embeddings based on autocovariance. Specifically, given connected non-bipartite graph G with adjacency matrix $A \in \mathbb{R}^{n\times n}$ and degree matrix D, the τ-step random-walk transition matrix is

$$P(\tau) = \begin{cases} (D^{-1}A)^\tau & \text{for the discrete-time random-walk} \\ \exp(-\tau(I - D^{-1}A)) & \text{for the continuous-time random-walk} \end{cases}$$

Both walks have an unique stationary distribution π with its entries proportional to node degrees. Denote Π as the diagonal stationary distribution matrix. Then the autocovariance similarity matrix with Markov time τ is

$$R(\tau) = \Pi P(\tau) - \pi\pi^T$$

The embedding matrix $U^* \in \mathbb{R}^{n\times d}$ is one that minimizes the reconstruction error of $R(\tau)$:

$$U^* = \arg\min_U \|UU^T - R(\tau)\|_F^2$$

From Eckart-Young-Mirsky theorem [7], the optimal $U^* = Q\sqrt{\Lambda}$, where $R(\tau) = Q\Lambda Q^T$ is the eigendecomposition of $R(\tau)$. Note that $R(\tau)$ itself is a dense matrix, and a direct eigendecomposition would cost $O(n^3)$. However, since it is constructed from several sparse matrices, i.e., A, D, and Π, we may have more efficient ways to compute. Specifically, Lanczos iterations only require matrix-vector multiplications, and the special form of $R(\tau)$ should allow us to evaluate those efficiently.

In this project, we will explore different alternatives of efficient computation of U^* based on Lanczos iterations, and empirically compare them.

References

[1] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *SIGKDD*, 2014.

[2] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *SIGKDD*, 2016.

[3] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *WSDM*, 2018.

[4] J-C Delvenne, Sophia N Yaliraki, and Mauricio Barahona. Stability of graph communities across time scales. *PNAS*, 107(29):12755–12760, 2010.

[5] Jean-Charles Delvenne, Michael T Schaub, Sophia N Yaliraki, and Mauricio Barahona. The stability of a graph partition: A dynamics-based framework for community detection. In *Dynamics On and Of Complex Networks*, pages 221–242. Springer, 2013.

[6] Michael T Schaub, Jean-Charles Delvenne, Renaud Lambiotte, and Mauricio Barahona. Multiscale dynamical embeddings of complex networks. *PRE*, 99(6):062308, 2019.

[7] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

1 Proposal submitted 1 / 1

✓ **- 0 pts Correct**

This sounds great. I look forward to seeing the results.