Semi-Supervised Classification of Nodes in Graphs via Anonymous Walks Embeddings

Alfredo De la Fuente and Maxim Panov

1 Motivation

Graph embedding techniques arise from the need of extracting useful information from the structural nature of the data in order to be fed as input to machine learning models in different areas (natural language processing, bioinformatics [2], social network analysis [7], and recently as building blocks of reinforcement learning algorithms [5, 10, 11]). Many articles successfully surveyed the latest contributions in the field [1, 4], suggesting a growing trend of interest in this research topic due to the consistent improving results upon different models. Moreover, these surveys agree on promoting the need for a theoretical framework under which future work can be developed in this area, which is why we develop our paper following a consistent mathematical framework based on previous research [6, 12]. The main challenge of the embedding task for node classification lies on capturing the local topology as well as assigning similar embedding vectors to closely related nodes with same labels in the original graph, together with allowing its efficient extension for dynamic graphs. Therefore, we aim to develop a framework under which inductive and transductive learning for the node classification task implicitly benefits the embedding quality from the semi-supervised scenario.

2 Research Problem

We focus our analysis on a graph semi-supervised setting, under which, besides the labeled and unlabeled instances, we are provided an $(L+U) \times (L+U)$ graph adjacency matrix A, where each entry $a_{i,j}$ indicates the degree of similarity between instance i and j, regardless of whether they are labeled or not. In particular, we consider the scenario where the graph is explicitly provided to us together with additional information from the feature vectors x for each node, e.g. graph edges could reflect the friendship status in a social network between users, while feature vectors represent each user's preferences for a set of items. Thus, we aim for a classification model that predicts an output for the unlabeled instances by leveraging the nodes features, the labeled cases and the graph structure.

3 Technical Contribution

Main contribution. We propose a scalable semi-supervised node embedding algorithm that leverage the input features of each node by integrating heuristics derived from anonymous walks (see Figure 3). Our key contribution is defining an scheme that allows to efficiently learn new embeddings for semi-supervised classification task in an inductive manner, so that for any new node added to the graph we will only require a set of statistics derived from its anonymous walks, and not explicitly training the full model with the new adjacency matrix (reducing the embedding complexity time by not using the full adjacency matrix). These improvements provide an advantage over previously developed models ([3, 8, 9]) in the case of efficiently embedding nodes in static and dynamic graphs under semi-supervised setting compared to previously proposed algorithms in the literature. Our proposed model is based on Planetoid model by [12]. The framework consid-

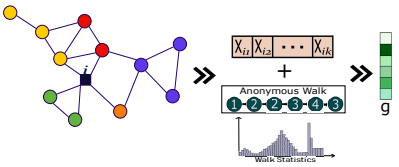

Figure 1: Proposed model scheme for semi-supervised node classification by using anonymous walks statistics and node's features as input for graph embedding.

ers a feed forward neural network where the l-th layer is a non-linear function $\mathbf{h}^l$ recursively defined as $\mathbf{h}^l(\mathbf{x}) = \text{RELU}\left(\mathbf{W}^l\mathbf{h}^{l-1}(\mathbf{x}) + \mathbf{b}^l\right)$ and $\mathbf{h}^0(\mathbf{x}) = \mathbf{x}'$; where $\mathbf{x}'$ is the $(k+w)$-dimensional vector formed by concatenating the side information $x_i \in \mathbb{R}^k$ and the statistics from anonymous walks $s_i \in \mathbb{R}^w$.

We implemented our model by modifying the code found on previous literature work references such as: Anonymous Walks repository, Planetoid implementation repository and Graph Embedding Methods (GEM) repository.The experimental results for different datasets are reported in Table 3. We are still in working progress towards running more experiments to improve our current model.

Model	Citeseer	Cora	Pubmed
GramEmb	0.582	0.579	0.698
Planetoid	0.647	0.755	0.772
GCN	0.703	0.805	0.791
Planetoid+AW	0.696	0.811	0.783

Table 1: Semi-supervised node classification accuracy for different models for each dataset.

References

[1] H. Y. Cai, V. W. Zheng, and K. Chang. "A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications". In: *IEEE Transactions on Knowledge and Data Engineering* XX.Xx (2018). ISSN: 10414347. DOI: 10.1109/TKDE.2018.2807452. arXiv: 1709.07604.

[2] N. De Cao and T. Kipf. "MolGAN: An implicit generative model for small molecular graphs". In: *ArXiv e-prints* (May 2018). arXiv: 1805.11973 [stat.ML].

[3] C. Donnat, M. Zitnik, D. Hallac, and J. Leskovec. "Learning Structural Node Embeddings Via Diffusion Wavelets". In: *ArXiv e-prints* (Oct. 2017). arXiv: 1710.10321.

[4] P. Goyal and E. Ferrara. "Graph Embedding Techniques, Applications, and Performance: A Survey". In: (2017). arXiv: 1705.02801.

[5] J. B. Hamrick, K. R. Allen, V. Bapst, T. Zhu, K. R. McKee, J. B. Tenenbaum, and P. W. Battaglia. "Relational inductive bias for physical construction in humans and machines". In: *ArXiv e-prints* (June 2018). arXiv: 1806.01203 [cs.LG].

[6] T. N. Kipf and M. Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: (2016), pages 1–14. ISSN: 0004-6361. DOI: 10.1051/0004-6361/201527329. arXiv: 1609.02907.

[7] Y. Li, C. Sha, X. Huang, and Y. Zhang. "Community Detection in Attributed Graphs: An Embedding Approach". In: *Proceedingss of AAAI* (2018), pages 338–345.

[8] M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu. "Asymmetric transitivity preserving graph embedding". In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2016, pages 1105–1114.

[9] B. Perozzi, R. Al-Rfou, and S. Skiena. "DeepWalk: Online Learning of Social Representations". In: (2014). ISSN: 9781450329569. DOI: 10.1145/2623330.2623732. arXiv: 1403.6652.

[10] M. Qu, J. Tang, and J. Han. "Curriculum Learning for Heterogeneous Star Network Embedding via Deep Reinforcement Learning". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. Marina Del Rey, CA, USA: ACM, 2018, pages 468–476. ISBN: 978-1-4503-5581-0. DOI: 10.1145/3159652.3159711.

[11] A. Sanchez-Gonzalez, N. Heess, J. T. Springenberg, J. Merel, M. Riedmiller, R. Hadsell, and P. Battaglia. "Graph networks as learnable physics engines for inference and control". In: (2018). arXiv: 1806.01242.

[12] Z. Yang, W. W. Cohen, and R. Salakhutdinov. "Revisiting Semi-Supervised Learning with Graph Embeddings". In: 48 (2016). arXiv: 1603.08861.