DEBIASING GRAPH REPRESENTATIONS VIA METADATA-ORTHOGONAL TRAINING

John Palowitch, Bryan Perozzi

SIAM Workshop on Network Science 2020
July 9–10 · Toronto

Summary

This work introduces a technique - Metadata-Orthogonal Node Embedding Training (MONET) - to control bias in graph representations from potentially sensitive metadata. We illustrate the effectiveness of MONET though our experiments on a variety of real world graphs, showing superior performance in tasks such as preventing political affiliation bias in a blog network, and preventing the gaming of embedding-based recommendation systems.

Introduction

Graph embeddings have been eminently useful in network visualization, node classification, link prediction, and many other graph learning tasks [4]. While graph embeddings can be learned directly from edge data [8], there is often accompanying node-wise metadata, like demographic or textual features. This metadata can be measurably related to a graph's structure [6], and thus metadata can enhance graph learning models [5]. However, there are applications for which it is desirable to obtain embeddings that *avoid* effects of specified *sensitive* data. For instance, the designers of a recommendation system may want to make recommendations independent of a user's demographic information or location. Simply ignoring the metadata in the model will not prevent the embeddings from learning inherent correlations between metadata and the graph structure. Thus, we propose two desiderata for controlling sensitive metadata in graph neural networks (GNNs):

D1. Metadata influence on graph topology is modeled in a partitioned subset of embedding space, providing separability to the overall graph representation.

D2. Non-metadata or "topology" embeddings are debiased from metadata embeddings with a *provable guarantee* on the level of remaining bias.

In this work we propose a novel GNN technique, MONET, that satisfies both of these desiderata. Essentially, MONET is a new GNN layer that executes training-time linear debiasing of graph embeddings, by ensuring that metadata embeddings are trained on a hyperplane orthgonal to that of the topology embeddings.

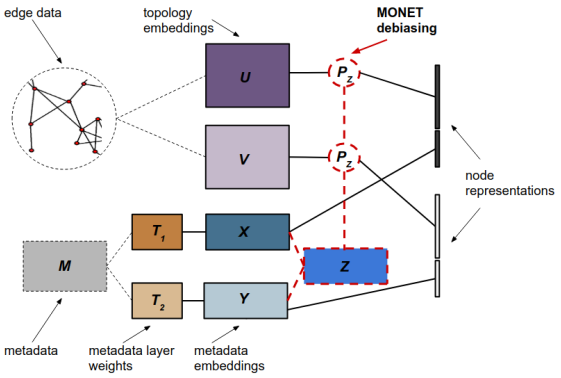

Figure 1: Illustration of the MONET unit.

Methodology

We introduce the MONET unit in an unsupervised graph embedding approach which applies the GloVe model [7] to a "corpus" of random walks [3]. Suppose we wish to embed a graph with n nodes from random-walk co-occurrence counts $C_{n\times n}$. Ignoring bias terms and loss weights [7], the GloVe loss is:

$$L_{\text{GloVe}} = \sum_{i,j\leq n} (U_i^T V_j - \log(C_{ij}))^2.$$

Above, $U, V \in \mathbb{R}^{n\times d}$ are the "center" and "context" node embeddings, which can be summed to provide an overall graph representation. Now, suppose we have node metadata $M \in \mathbb{R}^{n\times m}$. To achieve D1, we feed M through a neural network with weights T, giving metadata embeddings X and Y, as shown in Figure 1. We concatenate the metadata embeddings, yielding what we call $\text{GloVe}_{\text{meta}}$:

$$L_{\text{GloVe}_{\text{meta}}} = \sum_{i,j\leq n} (U_i^T V_j + X_i^T Y_j - \log(C_{ij}))^2.$$

The metadata embedding spaces X, Y could relieve the topology spaces U, V of the responsibility to encode metadata information, removing bias. However, we find – empirically and theoretically – that this does not fully occur. A phenomenon called "metadata leakage", which we define and prove in the full version of this work, allows topology embeddings to duplicate graph-structure metadata correlations before the metadata embeddings converge.

MONET: To prevent metadata leakage, satisfying D2, we introduce Metadata-Orthogonal Node Embedding Training (MONET), which uses the Singular Value Decomposition (SVD) of the metadata representation to directly control topology embedding bias. With $Z = X + Y$ as the metadata representation, let Q_Z be the left-singular vectors of Z. Define the projection $P_Z := I_{n\times n} - Q_Z Q_Z^T$. As illustrated in Figure 1, the MONET unit projects the embeddings U and V onto the metadata-orthogonal plane, via the operation $\mathcal{D}_Z(A) := P_Z A$:

$$L_{\text{MONET}_\text{G}} = \sum_{i,j\leq n} (\mathcal{D}_Z(U)_i^T \mathcal{D}_Z(V)_j + X_i^T Y_j - \log(C_{ij}))^2.$$

Experiments

To investigate embedding bias, we apply standard baselines, an adversarial debiasing baseline [2], $\text{GloVe}_{\text{meta}}$, and MONET_G to the Political Blogs graph [1], using blog affiliation as metadata. We measure embedding bias by the ability of a Linear SVM to predict affiliation from the embeddings on held-out test sets. As seen in Figure 2, only MONET_G's performance is consistent with a random baseline, showing exact debiasing. $\text{GloVe}_{\text{meta}}$ is still biased, showing metadata leakage. Interestingly, adversarially-debiased embeddings are still quite biased.

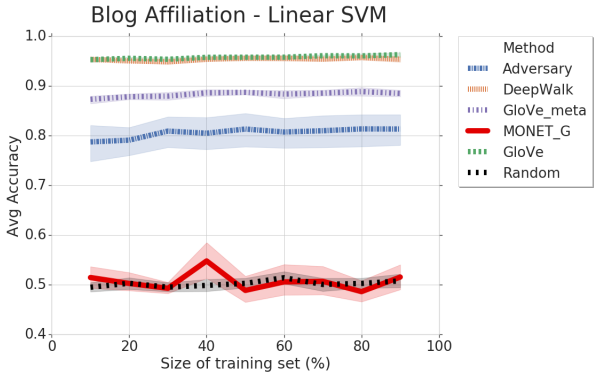

Figure 2: Prediction bias from MONET and baselines.

On another experiment, we inject an artificial ranking-inflation attack into the MovieLens graph, embedding movie (item) nodes, and using attack counts as metadata. For a given set of embeddings, we measure MRR (ranking accuracy) against the number of attacked movies with artificially inflated rankings (bias). We introduce a tuning parameter into MONET's $\mathcal{D}_Z$ to investigate bias-accuracy trade-off. As seen in Figure 3, we find that MONET_G performs 8x better than random at perfect debiasing ($\lambda = 1.0$), scaling up to baseline performance as λ decreases.

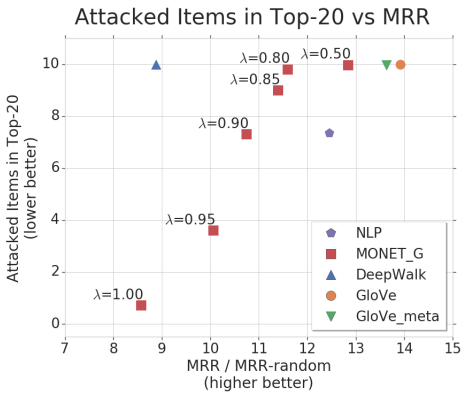

Figure 3: Item retrieval bias vs accuracy on MovieLens.

Discussion

There are two promising directions of future work. First, MONET only guarantees linear debiasing. Methods and exact guarantees for controlling non-linear associations should be investigated. Second, the performance of MONET in deeper GNNs and with high-dimensional metadata should be explored. Code for MONET and experiments is at `github.com/google-research/google-research/tree/master/graph_embedding/monet`.

References

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*. ACM, 2005.

[2] A. J. Bose and W. L. Hamilton. Compositional fairness constraints for graph embeddings. *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[3] R. Brochier, A. Guille, and J. Velcin. Global vectors for node representations. In *The World Wide Web Conference*, 2019.

[4] P. Cui, X. Wang, J. Pei, and W. Zhu. A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[5] M. E. Newman and A. Clauset. Structure and inference in annotated networks. *Nature Communications*, 2016.

[6] L. Peel, D. B. Larremore, and A. Clauset. The ground truth about metadata and community detection in networks. *Science advances*, 2017.

[7] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[8] J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. Tang. Network Embedding as Matrix Factorization: Unifying DeepWalk, LINE, PTE, and node2vec. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018.