Compositional Fairness Constraints for Graph Embeddings

Avishek Joey Bose
McGill University, Mila
joey.bose@mail.mcgill.ca

William L. Hamilton
McGill University, Mila
Facebook AI Research (FAIR), Montreal
wlh@cs.mcgill.ca

Abstract

Learning high-quality node embeddings is a key building block for machine learning models that operate on graph data, such as social networks and recommender systems. However, existing graph embedding techniques are unable to cope with fairness constraints, e.g., ensuring that the learned representations do not correlate with certain attributes, such as race or gender. Here, we introduce an adversarial framework to enforce fairness constraints on graph embeddings. Our approach is *compositional*—meaning that it can (optionally) enforce multiple different fairness constraints during inference. Experiments on standard knowledge graph and recommender system benchmarks highlight the utility of our proposed framework.

1 Introduction

Learning low-dimensional embeddings of the nodes in a graph is a fundamental technique underlying state-of-the-art approaches to link prediction and recommender systems [10]. However, in many applications—especially those involving social graphs—it is desirable to exercise control over the information contained within these learned node embeddings. For instance, we may want to ensure that recommendations are fair or balanced with respect to certain attribtues (e.g., that they do not depend on a user's race or gender) or we may want to ensure privacy by not exposing certain attributes through learned node representations. While enforcing fairness constraints on general classification models [6, 7, 12, 15, 24] and collaborative filtering algorithms [23] has received considerable attention in recent years, these techniques have yet to be considered within the context of graph embeddings—a setting that introduces particular challenges, e.g., due to the non-i.i.d. nature of graph data.

Moreover, in the case of social graphs and large-scale recommender systems, it is often the case that there are many *possible* sensitive attributes that we *may* want to enforce invariance constraints over. For instance, in a social graph we may want to empower users with the ability to specify which sensitive attributes they want to be used for different sorts of predictions—preferences that could stem from both privacy and experience preferences (e.g., users may want recommendations that do not depend on their gender). Standard fairness approaches that simply purge all information about sensitive attributes during training [15] are ill-suited to this setting, since they cannot accommodate new combinations of fairness constraints after training.

Present work. We introduce an adversarial framework to enforce *compositional* fairness constraints on graph embeddings. The key idea behind our approach is that we learn a set of *adversarial filters* that remove information about particular sensitive attributes. Crucially, each of these learned filters can be *optionally* applied after training, so the model can flexibly generate embeddings that are invariant with respect to different sets of sensitive attributes. Our work builds upon the success of recent adversarial approaches to fairness [24], disentanglement [16], and transfer learning [15]—extending these approaches to the domain of graph representation learning and introducing new algorithmic techniques to accommodate compositional constraints during inference.

Preprint. Work in progress.

2 Background and Related Work

Fair machine learning. Recent work on fairness in machine learning, including work on fairness in collaborative filtering, involves making predictions that are balanced or invariant with respect to certain sensitive variables (e.g., race or gender) [6, 7, 12, 15, 24, 23]. The usual assumption is that the sensitive variables are known *a priori*, and a standard approach is to simply purge the learned model representation space such that it is invariant with respect to the sensitive variable. Formally, in the standard "fair classification" setting we consider a data point $\mathbf{x} \in \mathbb{R}^n$, its class label $y \in \mathcal{Y}$, and a binary sensitive attribute $a \in \{0, 1\}$ (e.g., indicating gender). The high-level goal is then to train a model to predict y from $\mathbf{x}$, while making this prediction invariant or fair with respect to a [15]. There are many specific definitions of fairness, such as whether fairness refers to parity or satisfying certain preferences (see [7] for a detailed discussion).

In this paper, we consider a pragmatic measure of group fairness or demographic parity [7], which imposes the condition that the prediction algorithm should predict the same outcome across different groups with equal probability. In the social recommendation setting, this amounts to the strict requirement that a recommendation should be completely *balanced* or *invariant* with respect to a sensitive attribute. For instance, in the binary classification setting demographic parity corresponds to $P(\hat{y} = 1|a = 0) = P(\hat{y} = 1|a = 1)$. As one would expect, the utility of demographic parity drops if the true underlying rates for classification are very different, and imposing demographic parity as a constraint has been shown to significantly hinder classification performance [6]. We thus consider a relaxed form of demographic parity, which involves a tradeoff between fairness and classification [12], and we implement this tradeoff through an adversarial regularizer that criticizes the representations for encoding sensitive information [15]. This approach builds upon the recent successes of generative adversarial networks [9], adversarial fairness [15], and adversarial disentanglement [14, 16] in computer vision—adapting these techniques to the domain of graph representation learning.

Graph embeddings. A wide variety of techniques have been proposed to learn low-dimensional embeddings of nodes and edges in graphs [10], including methods based on matrix decomposition [22], random-walk based objectives [18], and graph neural networks [8]. In this work, we consider the multi-relational (i.e., knowledge graph) setting, where embeddings are learned for both entities/nodes and relations and where the primary goal is to predict (unobserved) relations between entities [17], a general setting that encompasses both link prediction and basic recommender systems [4].

We build upon the the TransD [11] model, but note that other knowledge graph embedding approaches (e.g., [5, 20, 22]) are also amenable to our framework. Let $f : \mathcal{V} \mapsto \mathbb{R}^d$, be an encoder function that maps a node/entity ID $v \in \mathcal{V}$ to a d-dimensional embedding. Learning in TransD is performed in a contrastive manner, where positive examples correspond to observed entity-relation-entity triplets (i.e., graph edges) and negative examples are formed by corrupting observed triplets by replacing true entities with randomly sampled ones. Let a positive entity-relation-entity triplet be denoted by $\xi^+ = (h^+, r^+, t^+)$, and a negative triplet be either $\xi^- = (h^-, r^+, t^+)$ or $\xi^- = (h^+, r^+, t^-)$. Given the embedding function f, the scoring rule for a triplet in TransD can then be defined as follows:

$$s(\xi, f) = \|f(h) + \mathbf{r}_p\mathbf{h}_p^\top f(h) + \mathbf{r} - f(t) + \mathbf{r}_p\mathbf{t}_p^\top f(t)\|, \tag{1}$$

where $\mathbf{r}$ is the embedding vector for r, and $\mathbf{r}_p \in \mathbb{R}^d$, $\mathbf{t}_p \in \mathbb{R}^d$, and $\mathbf{h}_p \in \mathbb{R}^d$ are projection parameters of the model. Learning can then proceed using a suitable loss function such as max-margin, where the model tries to separate the positive and negative triplets by a margin.

3 Fairness in Graph Embeddings

We now present our approach to enforce compositional fairness constraints on graph embeddings, starting with a single sensitive attribute and then generalizing to the full compositional model.

Non-compositional adversary. In this setting, we assume there is a single binary sensitive attribute, A, associated with each entity. Recall that $f : \mathcal{V} \mapsto \mathbb{R}^d$, is an encoder function that maps a node/entity to a d-dimensional embedding. For the sensitive attribute, A, we define an adversary $D : \mathbb{R}^d \mapsto [0, 1]$ that takes as input the embedding for an entity and tries to predict the binary sensitive attribute from this embedding. The objective of the adversary is thus to achieve high classification accuracy for the sensitive attribute, but in doing so it assigns a penalty to the encoder for encoding sensitive information. Training then consists of alternating updates between the encoder, whose joint objective is modified to include a fixed adversary, and then updating the adversary with a fixed encoder. The joint objective can be stated as:

$$L = L_{task}(s(\xi^{+}, f), s(\xi^{-}, f)) + \lambda_h L_{adv}(D, f(h), a) + \lambda_t L_{adv}(D, f(t), a). \tag{2}$$

The objective function is composed of two parts: the task specific loss for the encoder, i.e., $L_{task} = \max(0, \eta + s(\xi^{+}) - s(\xi^{-}))$ (with $s(\cdot)$ given by Equation 1), and the adversarial loss L_{adv}, where λ_t, λ_h are hyperparameters that trade off performance of the task with the fairness constraint. Following [15], we use

$$L_{adv}(D, \mathbf{x}, a) = 1 - \sum_{i \in \{0,1\}} \frac{1}{|(\mathbf{x}, a) \in \mathcal{B} : a = i|} \sum_{(\mathbf{x},a) \in \mathcal{B}\,:\, a=i} |D(\mathbf{x}) - a)|, \tag{3}$$

where $\mathcal{B}$ denotes a (mini)batch of data points. In the alternating steps of optimization, the encoder attempts to minimize L with the adversary fixed, while the adversary attempts to maximize a traditional cross-entropy classification loss over the sensitive attribute with the encoder fixed.

Compositional adversary. Often there are multiple sensitive attributes that we may wish to be fair with respect to. To address this challenge we introduce a compositional adversary that can (optionally) filter the embeddings to be fair with respect to a set of binary attributes. Here, we assume that there are k binary sensitive attributes $A_1, ..., A_k$, and that we may want to generate node embeddings that are fair w.r.t. any subset of these attributes. In this case, our compositional encoder, $f_c : (\mathcal{V}, \mathbb{Z}^k) \mapsto \mathbb{R}^d$ takes as input a node/entity ID, as well as a binary mask $\mathbf{m} \in \mathbb{Z}^k$, which specifies the sensitive attributes that we want the generated embedding to be invariant to. To define f_c we associate each sensitive attribute with a trainable *filter* function, $g : \mathbb{R}^d \mapsto \mathbb{R}^d$, which is trained to remove information about that specific sensitive attribute from the embedding space. We generate an embedding by summing the outputs of the filters specified by the mask, $\mathbf{m}$:

$$f_c(x, \mathbf{m}) = \frac{1}{\|\mathbf{m}\|_1} \sum_{j=1}^{k} \mathbf{m}_j g_j(f(x)), \tag{4}$$

where $\mathbf{m}_i$ is the i-th entry of the mask. To learn f_c, we additionally associate each sensitive attribute with its own unique discriminator, D_k and the compositional adversarial loss is computed as

$$\begin{aligned} L_c =& L_{task}(s(\xi^{+}, f_c), s(\xi^{-}, f_c)) + \qquad (5)\\ & \frac{1}{\|\mathbf{m}\|_1} \sum_{j=1}^{k} \lambda \mathbf{m}_j \Big(L_{adv}(D_j, g_j(f(h)), a) + L_{adv}(D_j, g_j(f(t)), a) \Big), \qquad (6) \end{aligned}$$

We term this approach compositional because the final output embedding is composed of the summation of the output of the individual filters.

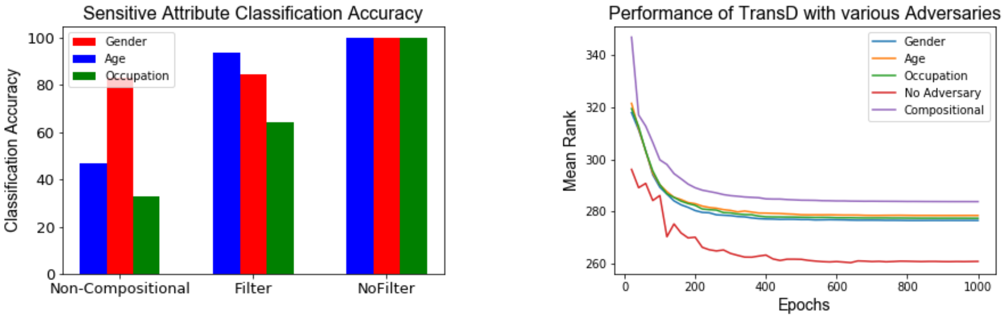
Non-binary attributes. To extend to non-binary attributes, we replace the per-sample adversarial penalty on the encoder with a multi-margin loss. In this case we have that the sensitive attribute $a \in \mathcal{A}$ is a categorical attribute and the discriminator has the form $D : \mathbb{R}^d \times \mathcal{A} \mapsto [0, 1]$, with the loss taking the form:

$$L_{adv}(D, f(x), a) = \sum_{a' \in \mathcal{A}, a' \neq a} \max(0, 1 - D(f(x), a) + D(f(x), a')). \tag{7}$$

We found this loss to be empirically superior and more stable than alternatives (e.g., cross-entropy).

4 Experiments

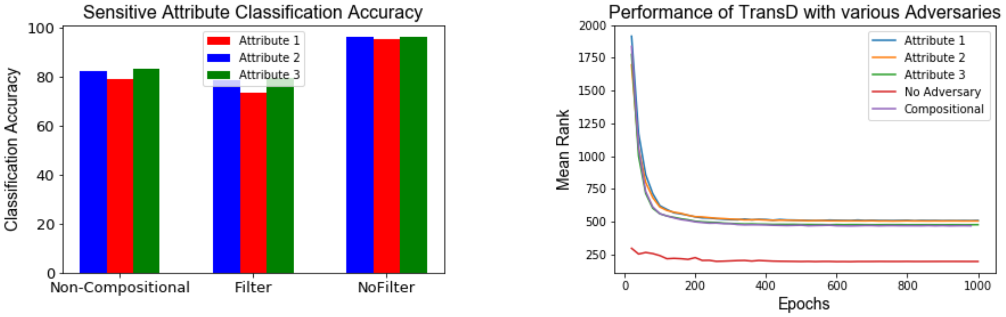
Setup and datasets. We evaluate our model on two datasets: MovieLens100k [2] and Freebase15k-237 [3]. Following [4], we view MovieLens100k as a bipartite graph between movies and users, treating the different integer rating scores as different relations, but unlike previous work, we treat the provided user features—gender, age, and occupation—as sensitive attributes (with the integer age attribute binned into 15 equal-sized bins). Freebase15k is a traditional knowledge graph completion benchmark, and we use this as a synthetic test case selecting the top-3 most common entity attributes (from [1]) as "sensitive attributes". The Appendix contains information on hyperparameter selection and other implementation details.



(a) Classification accuracy for sensitive attributes

(b) Mean rank for link prediction

Figure 1: Performance TransD with Fairness Constraints on MovieLens 100k



(a) Classification accuracy for sensitive attributes

(b) Mean rank for link prediction

Figure 2: Performance TransD with Fairness Constraints on FreeBase 15k-237

Results. Using the standard MovieLens100k and Freebase15k testsets, we evaluate both the accuracy of our model on the relation prediction (i.e., recommendation task)—evaluated using a standard relation prediction, mean-rank metric—as well as the fairness of the learned embeddings, by training an MLP classifier from scratch to predict the sensitive attributes from the generated embeddings. For both datasets, we see that the adversarial fairness objective significantly reduces the ability of a newly trained classifier to predict the sensitive attributes from the embeddings (Fig. 1.a and Fig. 2.a). Without an adversary (the “NoFilter” setting), the classification accuracy is near 100% for the sensitive attributes but drops significantly once the fairness constraints are added (both in the non-compositional and compositional filter setting). However, the performance on the relation prediction objective is also negatively impacted, especially for Freebase, highlighting the extent to which enforcing fairness can lead to worse performance on the original task. Comparing the compositional fairness approach to the non-compositional approach, where a single adversary is re-trained for each attribute, we see interesting differences across the two datasets: In the MovieLens setting, the compositional adversary performs worse than the individual adversaries, which is unsurprising given that each individual adversary needs only to focus on a single fairness objective. Interestingly, however, on Freebase15k, the compositional model is more fair than the single adversary model *on all three attributes*, indicating that jointly modeling multiple sensitive attributes can improve fairness on individual ones. Similar results have recently been observed elsewhere [19], but further theoretical analysis of this phenomena is necessary in future work.

5 Conclusion

We introduce an adversarial framework to enforce compositional fairness constraints on graph embeddings. Our approach uses a bank of adversarial filters to optionally enforce fairness constraints over a set of possible sensitive attributes. Our work sheds light on how fairness can be enforced in graph representation learning, and our compositional approach highlights how fairness could be deployed in a real-word, user-driven setting, where it is necessary to optionally enforce a large number of invariance constraints over learned graph representations.

References

[1] `https://github.com/cmoon2/knowledge_graph/tree/master/datasets/Freebase/FB15k_Entity_Types`.

[2] `https://grouplens.org/datasets/movielens/100k/`.

[3] `https://www.microsoft.com/en-us/download/details.aspx?id=52312`.

[4] Rianne van den Berg, Thomas N Kipf, and Max Welling. Graph convolutional matrix completion. *arXiv preprint arXiv:1706.02263*, 2017.

[5] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013.

[6] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[7] Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017.

[8] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *ICML*, 2017.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] W. Hamilton, R. Ying, and J. Leskovec. Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin*, 2017.

[11] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 687–696, 2015.

[12] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5967–5976, 2017.

[15] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

[16] M. Mathieu, J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun. Disentangling factors of variation in deep representation using adversarial training. In *NIPS*, 2016.

[17] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

[18] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *KDD*, 2014.

[19] Edward Raff and Jared Sylvester. Gradient reversal against discrimination. *arXiv preprint arXiv:1807.00392*, 2018.

[20] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080, 2016.

[21] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[22] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

[23] Sirui Yao and Bert Huang. New fairness metrics for recommendation that embrace differences. *arXiv preprint arXiv:1706.09838*, 2017.

[24] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

Appendix A Implementation Details

We implement each discriminator and adversarial filter as multi-layer perceptrons (MLPs) with a leaky ReLU non-linearity [21] between layers, and we use the Adam optimizer [13] with learning rate of 0.001. To generate negative triplets we randomly sample either a head or tail entity during training, with a ratio of 20 negatives for each positive triplet. Each model is trained for 1000 epochs and we use cross-validation on the validation set to select λ in Equation 2. We found that $\lambda_h = 0.1$ gives the best tradeoff between fairness and link prediction performance for MovieLens and $\lambda_h = \lambda_t = 10$ for FreeBase 15k. Note that $\lambda_t = 0$ for MovieLens as we do not use any sensitive attributes for movies. When training our compositional adversary we randomly sample the binary mask for each mini-batch. We also found that the cross entropy objective as outlined in [19] easier to train for FreeBase 15k than the max-margin objective, while the reverse is true for MovieLens 100k.