Bias in Knowledge Graph Embeddings

Styliani Bourli
Department of Computer Science & Engineering
University of Ioannina, Greece
sbourli@cs.uoi.gr

Evaggelia Pitoura
Department of Computer Science & Engineering
University of Ioannina, Greece
pitoura@cs.uoi.gr

***Abstract*—In this paper, we study bias in knowledge graph embeddings. We focus on gender bias in occupations, but our approach is applicable to other types of bias. We start by proposing measures for identifying bias in the dataset (i.e., in the KG) and then present two methods for testing whether any bias in the dataset is amplified by the embeddings. First, we look for gender-specific occupation analogies in the embeddings. Second, we test whether link prediction (i.e., occupation prediction in our case) aggregates gender bias by proposing gender-dominated occupations to people of the corresponding gender more often than expected. Then, we use a debiasing approach based on projections on the gender subspace. We present experimental results using the Wikidata dataset and pretrained TransE embeddings. Our results show that there exists gender bias in the dataset and that such bias is amplified by the embeddings. Our debiasing approach removes bias with a small penalty on accuracy.**

***Index Terms*—Knowledge graph, embeddings, bias**

I. Introduction

Knowledge graphs (KG) are multi-relational directed graphs widely used to represent knowledge in the form of (head-entity, relation, tail-entity) triplets, also known as facts. Edges represent relations between the entities (nodes) they connect. KG embeddings map the components of a knowledge graph to some low dimensional vector space (e.g., [2], [7], [12]). They have gained a lot of attention, since they are useful in a variety of tasks such as KG completion and relation extraction. Furthermore, vector operations are simpler and faster than the corresponding operations on graphs.

As machine learning algorithms are increasingly being used in various decision making processes, concerns about unfairness and bias are raised (see e.g., [4] for a survey). For example, recent work has shown that word embeddings (i.e., vector representation of words) exhibit various forms of social biases [1].

In this paper, we study bias in KG embeddings. We focus on gender bias in occupations, where we consider gender as the sensitive attribute and male and female as possible values. Our approach is readily applicable to other types of binary sensitive attributes, as well.

We start by proposing a measure for identifying bias in the dataset, that is, in the KG. Since in most cases knowledge comes from the real world, it is expected to reflect existing social biases [6], [13]. Then, we propose measures for identifying whether any bias in the dataset is encoded in, or even amplified by the KG embeddings. Our first measure is qualitative and looks for gender-specific occupation analogies in the embeddings. Our second measure uses a link prediction task to quantify bias amplification by the embeddings. Specifically, we test whether link prediction aggregates gender bias by predicting gender-dominated occupations for people of a corresponding gender more often than expected.

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

We also propose a debias approach for removing gender bias in KG embeddings of occupations. Our approach is tunable in the amount of bias it removes. We present experimental results using a Wikidata dataset and TransE embeddings. Our results show that there exists gender bias in occupations in the dataset and that bias is amplified by the embeddings. Our debias approach removes bias with a small penalty on accuracy.

The only other work on bias in KG embeddings that we are aware of is [8], that shows the existence of bias in KG embeddings. Our approach is different, since our focus is on bias amplification. More specifically, we study bias in KG embeddings in relation to the bias in the underlying KG itself, that is, we study how bias in the KG is being reflected in the embeddings. We also suggest a debias approach for KG embeddings.

The rest of this paper is structured as follows. In Section II, we introduce our methods for measuring bias. In Section III, we present our debias approach, and in Section IV, experimental results. Finally, in Section V, we present related work and in Section VI, our conclusions.

II. Measuring Bias

Knowledge graphs (KG) contain information in the form of (e_1, r, e_2) triplets, where e_1, and e_2 are entities and r their relation. A widely used KG embedding model is TransE [2], in which relationships are represented as translations in the embedding space. Let $\vec{e}$ and $\vec{r}$ be the embeddings of entity e and relation r, respectively. For a triplet (e_1, r, e_2), $\vec{e_1} + \vec{r}$ should be near to $\vec{e_2}$, if (e_1, r, e_2) holds and $\vec{e_1} + \vec{r}$ should be far away from $\vec{e_2}$, otherwise.

Next, we present our methods for measuring bias in data and in KG embeddings.

A. Bias in Data

We assume that there is gender bias in occupations, when the probability for an individual to have an occupation depends on his/her gender, i.e., the probability is higher for people of a specific gender.

*Definition 2.1: **Male, female and neutral occupations.*** For an occupation o, let $Pr(O = o|G = m)$ and $Pr(O = o|G = f)$ be respectively the probabilities that a male and a female entity have occupation o, where O stands for occupation and G for gender. Let $Pr(O = o|G = m) - Pr(O = o|G = f) = \vartheta$, and t be a positive value close to 0. We consider occupation o as a male occupation, if $\vartheta > t$, female occupation, if $\vartheta < -t$ and neutral occupation if $\vartheta \in [-t, t]$.

For a given dataset, we estimate $Pr(O = o|G = m)$ as $|M_o|/|M|$ and $Pr(O = o|G = f)$ as $|F_o|/|F|$, where $|M_o|$, $|F_o|$ are the number of male, female entities with occupation o, respectively, and $|M|$, $|F|$ the number of all male, female entities. To set threshold t, we plot the distribution of neutral occupations for different values of t and select the value of t where there is a sharp increase in the number of neutral occupations.

B. Bias in KG embeddings

Next we present methods for measuring bias and bias amplification in KG embeddings.

Analogies and projections. We consider two methods for identifying bias in embeddings. The first one is based on analogies commonly used to study bias in word embeddings [1]. Analogies are of the form *"a is to x as b is to y"*, where (a, b) is a seed pair of words such as *(man, woman)* determining a seed direction $\vec{a}-\vec{b}$, corresponding to the normalized difference between the two seed words. To define the best analogy pair (x, y) the score: $Score_{(a,b)}(x, y) = cos(\vec{a} - \vec{b}, \vec{x} - \vec{y})$ if $||\vec{x} - \vec{y}|| < \delta, \quad 0 \ otherwise$, is used. Threshold δ is a similarity threshold (often δ = 1), so as to select only semantically related x and y. The intuition is that a good analogy pair must be close to parallel to the seed direction. An example analogy pair for the $(man, woman)$ seed pair is $(computer\_programmer, homemaker)$.

For the KG embeddings and the TransE model, we use the score function:

$$Score_{(a,b)}(x, y) = cos(\vec{a} + \vec{r} - \vec{b}, \vec{x} + \vec{r} - \vec{y}) \quad (1)$$
$$if \ ||\vec{x} - \vec{y}|| \leqslant \delta, \quad 0 \ otherwise$$

where r is the relation ("*has occupation*" in our case), (a, b) is the seed pair of words (we use *(a,b) = (female, male)*), δ is a threshold for similarity (we use δ = 1) and (x, y) is a pair of occupations.

Our second method uses projections. Let $\vec{d_g} = f\vec{emale} - \vec{male}$ be the gender bias direction. Let C be a set of occupations. In the following, we use $\pi_{\vec{v}}\,\vec{u}$ for the projection of $\vec{u}$ onto $\vec{v}$. We define the projection score for each set C of occupations as:

$$proj\_score(C) = \frac{1}{|C|} \sum_{o \in C} ||\pi_{\vec{d_g}}\ \vec{o}||. \quad (2)$$

We expect the value of $proj\_score$ to be larger when C is the set of male, or female occupations than when C is the set of neutral occupations. For the latter, we expect $proj\_score$ to be close to zero.

Prediction task. To identify bias augmentation by KG embeddings, we utilize a prediction task. Given a male (resp., female) occupation, we use the embeddings to predict which individuals are more likely to have this occupation. Then, we compare the percentage of males (resp., female) individuals in the top-x predictions with the percentage of males (resp, female) individuals expected to have the specified occupation. If the predicted percentage is higher than the expected one, we conjecture that there is bias augmentation.

For TransE embeddings, to predict the most likely individuals to have a given occupation o, the score function:

$$score(e, r, o) = -||\vec{e} + \vec{r} - \vec{o}|| \quad (3)$$

is used, where e is a male/female entity, and r is the *"has occupation"* relation. We consider as Top_x, the individuals with the top x largest scores.

Let $|M_o|$ and $|F_o|$ be the number of male and female entities in the dataset having occupation o respectively and let $|E| = |M_o| + |F_o|$. The estimated number for male entities in the top-x results is $|M_o|x/|E|$, and for female entities is $|F_o|x/|E|$.

Then, we compare the predicted and the estimated percentage of male (female) individuals for different values of x and use the difference in the percentages to measure bias augmentation.

III. DEBIAS

Our methodology for removing (gender) bias from the KG embeddings follows the one in [1]. Specifically, if $\vec{e}$ is an embedding and $\vec{d_g} = f\vec{emale} - m\vec{ale}$ is the gender bias direction, then the new debiased embedding $\vec{e'}$ is obtained by subtracting from $\vec{e}$ its linear projection on $\vec{d_g}$. In our approach, we also add a tuning parameter λ to adjust how strong we want the debias to be:

$$\vec{e'} = \vec{e} - \lambda \cdot \pi_{\vec{d}}\vec{e}. \quad (4)$$

In our case, e is an occupation entity. The intuition is to extract from an occupation the bias information included in the gender bias direction. If parameter λ = 1, then the debias is "hard" and the whole gender information is extracted, whereas when λ = 0, then there is no debias.

To select an appropriate value for parameter λ for a "softer" debias, we first define the gender score function below. It expresses the mean of cosine similarities of each occupation category with the "$male$", "$female$" entity.

$$g\_score(C, g) = \frac{1}{|C|} \sum_{o \in C} cos(\vec{o}, \vec{g}), \quad (5)$$

where g is the $male$ or the $female$ entity, and C the set of [male, female, neutral] occupations. The resulting similarity ranges from -1 meaning exactly opposite, to 1 meaning exactly the same.

We choose parameter λ for "soft" debias, such as the g_score for the set of neutral occupations does not change before and after debias. Note that before debias, for the set of male occupations, we expect that its cosine similarity with $male$ entity is higher than that with $female$. The opposite holds for the set of female occupations. After the "soft" debias method is applied, these differences in cosine similarities with

$male$ and $female$ entities are expected to decrease, while they become exactly the same after "hard" debias.

IV. Experimental Evaluation

In this section, we first describe the dataset we use. Then we study bias in data and embeddings using our methods. We also analyze and evaluate the results of our debiasing approach.

A. Dataset

We use a real Wikidata dataset[1] that is pre-trained with the OpenKe library [5] using the TransE model. The dataset consists of 68,904,773 triplets, with 20,982,733 entities and 594 relations. The pre-trained embeddings have dimension 100. Since we study gender bias in occupations, we first apply a filtering procedure, to keep only triplets, that are related with gender and occupation. We also remove rare occupations (i.e., occupations possessed by less than 100 individuals). After that 1,717,427 male entities with occupation, 343,128 female entities with occupation and 730 occupations remain.

B. Measuring Bias

Bias in data. In Table I, we depict the top five male and top five female occupations using Definition 2.1 (i.e.,the ones having the largest $|\vartheta|$). Occupations such as football player, politician and lawyer are characterized as *male occupations*, while occupations such actor, singer and model are characterized as *female occupations*. The ϑ values show gender bias in data.

TABLE I: Top five male and top five female occupations.

Male Occupations (ϑ)	Female Occupations (ϑ)
football player (0.0907)	actor (-0.1532)
politician (0.0688)	singer (-0.0618)
university teacher (0.0211)	model (-0.0367)
baseball player (0.0155)	writer (-0.0204)
lawyer (0.0146)	television actor (-0.0188)

Analogies and projections. For studying bias in embeddings, we first consider analogies. We consider all (x, y) pairs, with x, y being any of the top 30 female and top 30 male occupations (i.e., the corresponding occupations with the highest $|\vartheta|$).

The top analogies for the $(female, male)$ pair are shown in Table II. Examples are (fashion model, businessperson) - "*female is to fashion model as male is to businessperson*" and (singer, conductor) - "*female is to singer as male is to conductor*". It is clear that gender bias is transferred from data to embeddings.

The $proj\_score$ (defined in Equation 2) for the male, female and neutral sets of occupations is shown on the first line of the second table in Table III. As expected the $proj\_score$ of male and female sets are greater than that of the neutral set, whose $proj\_score$ is close to zero.

TABLE II: Top five (x,y) pairs of occupations using the analogy "female is to x as male is to y".

Female Occupations	Male Occupations	Score
fashion model	businessperson	0.76
Japanese entertainer	businessperson	0.71
singer	musician	0.69
singer	conductor	0.68
model	musician	0.67

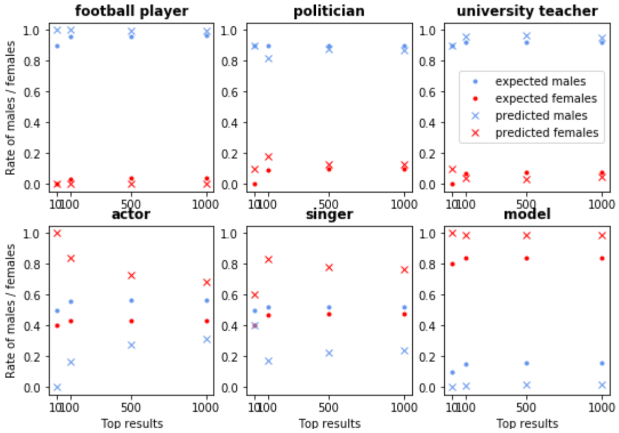


Fig. 1: Expected and predicted probabilities of top 3 male and female occupations, in the top 10, 100, 500 and 1000 results.

Prediction task. For the prediction task, we take the top-3 male and top-3 female occupations. For each of these, we compute the prediction score for all male and female entities, that is, the probability that these entities have this occupation. We then rank the results and take the Top_x entities, for x = 10, 100, 500 and 1000. We also calculate the expected percentage of males/females having each occupation. As we can see in Figure 1, in most cases the predicted value is higher than the expected value for males, if it is a male occupation and for females, if it is a female occupation. In contrast, the predicted value is lower than the expected for males, if it is a female occupation and for females, if it is a male occupation. This is a clear indication that embeddings contain and amplify bias.

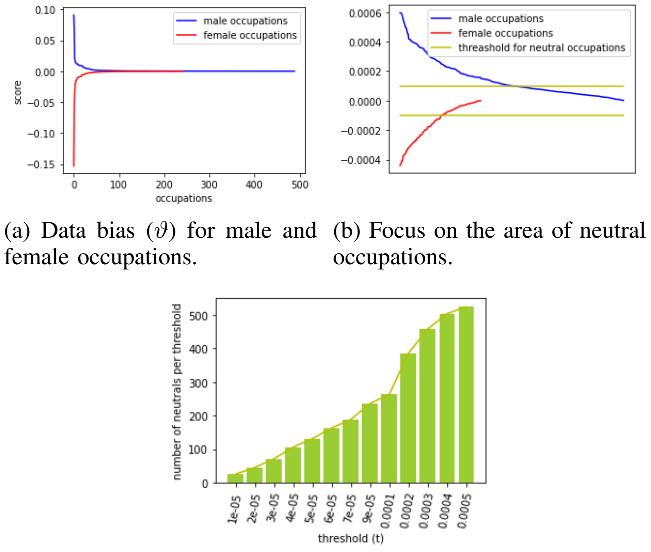
C. Debias

Firstly, for splitting occupations into male, female and neutral sets, we select $t = 0.0001$, based on Figure 2. After that there are 295, 172 and 263 occupations in the male, female and neutral set, respectively.

For debias, we used both "hard" and "soft" debias. We conclude that the appropriate value for "soft" debias is $\lambda = 0.5$, because for this value the g_score does not change for the neutral category before and after debias.

The results of $proj\_score$, defined in Equation 2, and g_score, defined in Equation 5, are shown in Table III. After "hard" debias the cosine similarity of the embeddings of male, female and neutral occupation sets are the same with the $male$ and $female$ entity. In contrast, after "soft" debias, the cosine similarity of the male occupation set is a little larger with $male$ entity, while the cosine similarity of the female occupation set

[1] Available in *http://openke.thunlp.org/*



(a) Data bias (ϑ) for male and female occupations.

(b) Focus on the area of neutral occupations.

(c) Distribution of neutral occupations using different values of t.

Fig. 2: (a) Data bias scores (ϑ) of male and female occupations using Definition 2.1, (b) focus on the area close to 0, where yellow lines are thresholds of the neutral occupation area, using t = 0.0001, (c) the distribution of neutral occupations using different threshold (t) values.

TABLE IV: HITS@10 of the top 20 male and the top 20 female occupations, for male and for female entities.

Debias	**Hits@10 (males/females)**
Before	0.94/0.94
After, λ = 0.5	0.93/0.91
After, λ = 1	0.89/0.87

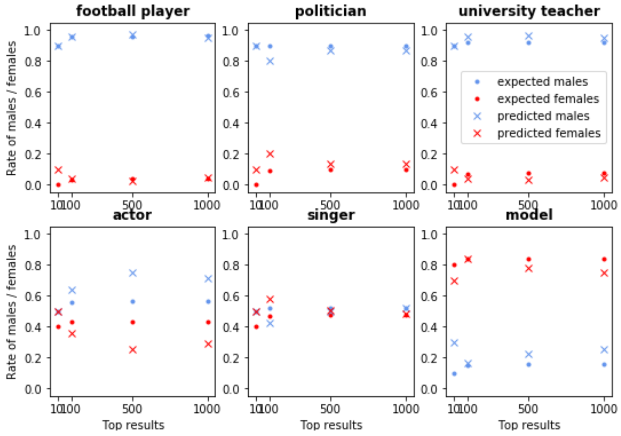


Fig. 3: Expected and predicted probabilities of top 3 male and female occupations, in the top 10, 100, 500 and 1000 results.

is a little larger with $female$ entity. This means that gender information is not totally removed. For *proj_score*, it is clear that after "hard" debias, all the embeddings are neutralize, and lose the whole gender information, while with "soft" debias this does not happen.

TABLE III: The score before and after debias for each occupation category: (Top) using *g_score* and (Bottom) using *proj_score*.

Debias	**Male (male/female)**	**Female (male/female)**	**Neutral (male/female)**
Before	0.09/-0.01	-0.08/0.18	0.04/0.04
After, λ = 0.5	0.06/0.01	-0.02/0.12	0.04/0.04
After, λ = 1	0.04/0.04	0.05/0.05	0.04/0.04

Debias	**Male**	**Female**	**Neutral**
Before	0.10	0.21	0.08
After, λ = 0.5	0.05	0.10	0.04
After, λ = 1	0.00	0.00	0.00

We now evaluate accuracy after the debias procedure using hits@10 in the top 20 male and female occupations. Specifically, we predict the occupation for males and females and we check whether the correct occupation is in the first 10 occupations with the highest scores. The results are presented in Table IV. As expected, the harder the debias is, the worst the accuracy is, since more information is removed. However, for $\lambda = 0.5$, the loss in accuracy is rather small.

We also evaluate our method using the prediction task after "soft" debias. Unlike before, as we can see in Figure 3, the predicted value is very close to the expected one, which shows that bias amplification has been decreased and embeddings contain some gender information, but without it being amplified.

V. Related Work

Algorithmic fairness and bias have been the focus of much current research (e.g., [4], [9]). For example, recent work has shown that word embeddings encode various forms of social biases [1]. In this paper, we study bias in KG embeddings. To the best of our knowledge, the only previous work in this topic is [8]. Different than [8], our focus is on whether any bias in the dataset (i.e., in the KG) is amplified by the KG embeddings. Bias amplifications has been previously studied for many applications, e.g., for recommendations [11]. We also propose the first debias method for KG embeddings. The method is based on projections on the gender subspace as in [1] and its novelty lies on tuning the amount of debias.

There has also been some recent work on achieving fairness in graph embeddings [3], [10]. The approach in [10] modifies the random walk step used in many graph embedding algorithms but not in the majority of KG embedding algorithms and thus it is not applicable to them. In [3], the learning of graph embeddings is enhanced with a set of adversarial regularization filters that remove information about sensitive attributes (e.g., gender) using projections. The approach was also tested on a TransD embedding [7] of a Freebase dataset. For this, authors report significant loss in accuracy. In our approach, the removal of any bias is done in a post-processing step and it is controlled by the bias in the input data. Thus, some gender information necessary for the prediction task accuracy remains.

VI. Conclusions

In this paper, we studied bias in KG embeddings. We proposed methods for identifying bias both in the dataset and in the KG embeddings. Our experimental results using a real Wikidata dataset and TransE embeddings have shown that gender bias in occupations exists and is amplified by the embeddings. We also proposed the first debias method for removing gender bias from occupation embeddings. Our method is tunable in the amount of bias it removes. Our experimental results have shown that the debias method is effective in removing bias with a small accuracy loss. In the future, we plan to consider additional embedding models besides TransE and more datasets and types of bias.

Acknowledgment

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EDK-00108).

References

[1] T. Bolukbasi, K-W Chang, J. Y. Zou, V. Saligrama, and A. Tauman Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*, pages 4349–4357, 2016.

[2] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795, 2013.

[3] A. Joey Bose and W. L. Hamilton. Compositional fairness constraints for graph embeddings. In *ICML*, volume 97, pages 715–724, 2019.

[4] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *FAT**, pages 329–338, 2019.

[5] Xu Han, Shulin Cao, Lv Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. Openke: An open toolkit for knowledge embedding. In *EMNLP*, 2018.

[6] K. Janowicz, B. Yan, B. Regalia, R. Zhu, and G. Mai. Debiasing knowledge graphs: Why female presidents are not like female popes. In *ISWC P&D-Industry-Blue Sky*, 2018.

[7] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao. Knowledge graph embedding via dynamic mapping matrix. In *ACL (1)*, pages 687–696, 2015.

[8] J. Fisher Joseph, D. Palfrey, C. Christodoulopoulos, and A. Mittal. Measuring social bias in knowledge graph embeddings. *arXiv preprint arXiv:1912.02761v2*, 2020.

[9] E. Pitoura, P. Tsaparas, G. Flouris, I. Fundulaki, P. Papadakos, S. Abiteboul, and G. Weikum. On measuring bias in online information. *ACM SIGMOD Record*, 46(4):16–21, 2018.

[10] T. A. Rahman, B. Surma, M. Backes, and Y. Zhang. Fairwalk: Towards fair graph embedding. In *IJCAI*, pages 3289–3295, 2019.

[11] V. Tsintzou, E. Pitoura, and P. Tsaparas. Bias disparity in recommendation systems. In *RMSE workshop at RecSyS*, 2019.

[12] Q. Wang, Z. Mao, B. Wang, and L. Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Trans. Knowl. Data Eng.*, 29(12):2724–2743, 2017.

[13] O. Zagovora, F. Flöck, and Claudia Wagner. "(weitergeleitet von journalistin)": The gendered presentation of professions on wikipedia. In *WebSci*, pages 83–92, 2017.