Understanding Gender Bias in Knowledge Base Embeddings

Yupei Du[♠] **Qi Zheng**[♡] **Yuanbin Wu**[♡] **Man Lan**[♡] **Yan Yang**[♡♢] **Meirong Ma**[♣]
[♠]Department of Information and Computing Sciences, Utrecht University, the Netherlands
[♡]Department of Computer Science and Technology, East China Normal University, China
[♢]Shanghai Key Laboratory of Multidimensional Information Processing, China
[♣]Shanghai Transsion Co., Ltd, China
y.du@uu.nl, qizheng.ecnu@outlook.com
{ybwu,mlan,yanyang}@cs.ecnu.edu.cn, meirong.ma@transsion.com

Abstract

Knowledge base (KB) embeddings have been shown to contain gender biases (Fisher et al., 2020b). In this paper, we study two questions regarding these biases: how to *quantify them*, and how to *trace their origins in KB*? Specifically, first, we develop two novel bias measures respectively for a group of person entities and an individual person entity. Evidence of their validity is observed by comparison with real-world census data. Second, we use influence function to inspect the contribution of each triple in KB to the overall group bias. To exemplify the potential applications of our study, we also present two strategies (by adding and removing KB triples) to mitigate gender biases in KB embeddings.

1 Introduction

Gender biases have been shown to have noticeable presence in a wide range of NLP models. For example, we can observe that the word embedding of "engineer" is closer to "he" than "she" (Bolukbasi et al., 2016), and co-reference systems associate "surgeon" more with masculine pronouns than with feminine ones (Rudinger et al., 2018). These biases are brought to our models from training data by our algorithms. Hence, besides revealing the existence of gender biases, it is important to quantify them and explain their origins in data.

Knowledge bases (KB, e.g. Freebase, Bollacker et al., 2007) provide accessible organizations of human knowledge by the form of triples. Each triple consists of a head entity, a relation, and a tail entity. For example, the fact that *Marie Curie is a chemist* is represented as ⟨`Marie Curie`, `people.person.profession`, `chemist`⟩. KB embeddings encode these knowledge into dense vector representations. It is important to understand gender biases in KB embeddings for two major reasons. First, KB embeddings serve as sources of prior knowledge in many downstream NLP models (e.g. pre-trained language models, Zhang et al., 2019). Clearly, if biases exist in KB embeddings, they can easily propagate into these models, and drive these models more biased. Second, Garg et al. (2018) observe that word embeddings reflect biases in the training corpora, and hence our society. Likewise, we suspect KB embeddings to reflect biases encoded in KBs, as also suggested by Radstok et al. (2021).

In this paper, we **propose two novel gender bias measures** for KB embeddings, one for a group of person entities (*group bias*) and the other for an individual person entity (*individual bias*). Furthermore, with *influence function* (Koh and Liang, 2017), we **explain the origins of group bias** at the fact triple level (i.e. how each triple in KB contribute to group bias). In practice, we use TransE (Bordes et al., 2013) to demonstrate our methods, for its popularity and simplicity. Nevertheless, most of our study can generalize to other embedding algorithms. Specifically, we make four contributions.

First, regarding **a group of person entities** with a shared relation-tail pair (e.g. of the same occupation), using **correlation analyses**, we measure their gender biases by the differences between different genders' link prediction errors.

Second, to understand the origins of the group bias, we use influence function to find its highly-influential triples in KB (i.e. triples that will change the bias most if being removed during training).

Third, regarding **a single person entity**, using **counterfactual analyses**, we develop a bias measure by measuring the change of the link prediction error when we keep everything else the same and perturb its gender. To avoid the intractable computational cost of re-training, we propose to use influence function to approximate the results.

Fourth, to further facilitate large-scale influence function based analyses, we derive a closed-form approximation of the Hessian matrix of TransE loss.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1381 - 1395

We therefore improve the time complexity of computing influence function from $\mathcal{O}(n)$ (stochastic approximation) to $\mathcal{O}(1)$.

Moreover, in further analyses, we show that both group and individual bias correlate well with real-world biases. We argue that this suggests the validity of our bias measures. We also show the accuracy of our influence function approximation by comparing with the brute-force strategy (i.e., leave-one-out re-training). Finally, to exemplify the applications of our study, we propose two simple de-biasing strategies, and demonstrate their effectiveness.

2 Preliminaries

Knowledge Base KB is a set of structural human knowledge represented by triples $\mathcal{G} = \{\langle h, r, t\rangle\}$, where h is a head entity, r is a relation type, and t is a tail entity. Moreover, these triples form a graph with entities as nodes (denoted by E, where $e \in E$ is an entity) and relations as edges. In this work, since we are particularly interested in person entities and their gender, we use $\langle s, r_g, m\rangle$ or $\langle s, r_g, f\rangle$ to represent a person s with gender male or female, where r_g is the relation of gender.[1]

TransE The entities and relations in KB can be represented with embedding vectors. These embeddings can serve in many NLP task as a source of prior knowledge. In this work, we focus on the widely used TransE (Bordes et al., 2013).

Given a triple $\langle h, r, t\rangle$, the key idea of TransE is to make vectors of h, r and t close to each other in the sense of small *link prediction* error. Concretely, TransE embeddings are learned by minimizing a margin-based ranking loss,

$$\mathcal{L} = \sum_{\langle h,r,t\rangle \in \mathcal{G}} [m + \psi(h, r, t) - \psi(h', r, t')]_+, \quad (1)$$

where m is a scalar margin and ψ is a distance measure. The lower $\psi(h, r, t)$ is, the more likely $\langle h, r, t\rangle$ forms a fact. h' and t' are two randomly sampled entities. The triple $\langle h', r, t'\rangle$ is called a *negative* sample because it is not in $\mathcal{G}$. This loss function basically says that the dissimilarity of a positive triple $\langle h, r, t\rangle$ should be smaller than a negative sample by a margin m.[2] Specifically, in this paper, we take ψ to be the L_2-norm distance $\psi(h, r, t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2^2$, where $\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^d$ are the embeddings of h, r and t, respectively.

In this paper, we use Freebase's (Bollacker et al., 2007) subset **FB5M** (Bordes et al., 2015) as the KB for training TransE embeddings and performing our analyses. See Appendix A for detailed setup.

Influence Function (Cook and Weisberg, 1982; Koh and Liang, 2017) provides an efficient way to approximate each training sample's impact on correctly predicting a test sample.

Formally, let $L(z, \theta)$ be a convex loss function on a training set $\{z_i\}_{i=1}^n$ with parameters θ. The empirical risk minimizer (ERM) is given by $\hat{\theta} = \arg\min_\theta \frac{1}{n}\sum_{i=1}^n L(z_i, \theta)$. We are interested in a training sample z's impact on $\hat{\theta}$, with a weight of ε. In this case, the new ERM is given by $\hat{\theta}_{z,\varepsilon} = \arg\min_\theta \frac{1}{n}\sum_{i=1}^n L(z_i, \theta) + \varepsilon L(z, \theta)$ (Note that if $\varepsilon = -\frac{1}{n}$, it equals to removing z).

Influence function provides an efficient method of approximating the difference between $\hat{\theta}_{z,\varepsilon}$ and $\hat{\theta}$, without retraining the model,

$$\hat{\theta}_{z,\varepsilon} - \hat{\theta} \approx \varepsilon \mathcal{I}_{\text{up,param}}(z), \quad (2)$$

where $\mathcal{I}_{\text{up,param}}(z) \overset{\text{def}}{=} -H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta})$. $H_\theta = \frac{1}{n}\sum_i \nabla^2 L(z_i, \theta)$ is the Hessian matrix of the original loss function.

Moreover, we are interested in the change of the test performance, which is a function F of the test sample z_{test} and the model parameter (LHS). By applying chain rule to F and Equation 2, we can obtain the difference of test performance. Formally,

$$F(\hat{\theta}_{z,\varepsilon}, z_{\text{test}}) - F(\hat{\theta}, z_{\text{test}}) \approx \varepsilon \mathcal{I}_{\text{up},F}(z, z_{\text{test}}), \quad (3)$$

where $\mathcal{I}_{\text{up},F}(z, z_{\text{test}}) \overset{\text{def}}{=} \nabla_\theta F(z_{\text{test}}, \hat{\theta})^\top \mathcal{I}_{\text{up,param}}(z)$.

Similarly, by splitting perturbation to first remove then add, we can also inspect the change of F when a training sample z is perturbed to z'. Denote $\theta_{-z,z',\varepsilon} = \arg\min_\theta \frac{1}{n}\sum_{i=1}^n L(z_i, \theta) - \varepsilon L(z, \theta) + \varepsilon L(z', \theta)$, and apply Equation 3 twice, we obtain

$$\begin{aligned} & F(\hat{\theta}_{-z,z',\varepsilon}, z_{\text{test}}) - F(\hat{\theta}, z_{\text{test}}) \\ \approx & \varepsilon \mathcal{I}_{\text{up},F}(z', z_{\text{test}}) - \varepsilon \mathcal{I}_{\text{up},F}(z, z_{\text{test}}) \\ \overset{\text{def}}{=} & \varepsilon \mathcal{I}_{\text{pert},F}(z, z', z_{\text{test}}). \end{aligned} \quad (4)$$

Finally, besides single sample estimation, we are also interested in inspecting the influence of removing a group of training samples. In these cases,

[1] We operate with binary gender here, because it is naturally encoded in KB.

[2] For simplicity, we consider only linear loss in the rest of this paper. This is a feasible choice both empirically and theoretically in our analyses. Empirically, in experiments we observe that the link prediction errors of all triples converge at a larger-than-margin value. Theoretically, when link prediction errors converge at values smaller than the margin, the gradients become 0. Its influence thus becomes 0, too.

Occupation	$\mathcal{B}_{\text{gr}}$	#male	#female
soldier	8.65×10^{-2}	3110	78
engineer	6.26×10^{-2}	3761	113
singer	1.46×10^{-2}	17260	13155
animator	-1.70×10^{-2}	1342	235
model	-3.11×10^{-2}	1595	5876
nurse	-9.77×10^{-2}	36	466

Table 1: A gallery of group bias ($\mathcal{B}_{\text{gr}}$) results. **#male** and **#female** are the numbers of male and female person entities in KB with this occupation respectively.

following Koh and Liang (2017), we simply add up the influence of each removed training sample. However, as noted by Koh and Liang (2017), when handling a group of samples, although influence function approximation still holds a strong correlation with the ground truth change of the parameters, the estimation can suffer from larger errors.

3 Gender Bias Measures

In this section, based on link prediction, we take two views to quantify gender biases in KB embeddings. First, using correlation analysis, we take a macro view to inspect gender biases of a group of person entities (e.g., how gender influences the overall occupation prediction accuracy of a group of engineer entities). Second, under the framework of counterfactual analysis, we take a micro view to assess gender biases of an individual person entity (e.g., how a specific engineer entity's gender influences its occupation prediction accuracy). Afterwards, we build connections between them.

In this following, we adopt occupation prediction as our running example. The reason is two fold. First, among all of the relations connected with person entities, occupation relation has the highest coverage rate (i.e. connect with the most person entities). Second, most previous relevant studies also focus on occupation. Our choice makes it easier to perform comparative studies (Garg et al., 2018; Fisher et al., 2020b).

3.1 Gender Biases of a Group

To see whether a group of entities exhibits bias, one direct solution is to deploy methods analog to those applied to analyze bias in word embeddings (Bolukbasi et al., 2016). For example, we can compute the projection of the TransE embedding of an occupation $\mathbf{o}$ to the difference between male and female entities (Bourli and Pitoura, 2020),

$$\mathcal{B}_{\text{wo}} = \mathbf{o}^\top(\mathbf{m} - \mathbf{f}),$$

where $\mathbf{m}$ and $\mathbf{f}$ are the embeddings of `male` and `female` entity respectively. However, we argue that because TransE follows a different type of learning objective (link prediction style objective instead of the vector-similarity-based ones in word embedding algorithms), directly adopt existing settings may not fully explore the semantics of TransE embeddings.

Therefore, we propose to detect group bias based on the correlation between genders and link prediction errors. Intuitively, given an occupation o, person entities of o's privileged gender will link to o with lower errors than those of unprivileged gender. Formally, we define the group bias of o as

$$\mathcal{B}_{\text{gr}} = \frac{1}{|F|}\sum_{s\in F}\psi(s, r_p, o) - \frac{1}{|M|}\sum_{s\in M}\psi(s, r_p, o),$$

where M and F are the sets of all male and female person entities with o respectively, and r_p is the relation `people.person.profession`. The higher $\mathcal{B}_{\text{gr}}$ is, the more o's embedding is biased towards male. Table 1 lists $\mathcal{B}_{\text{gr}}$ of some occupations, as well as the gender frequency of this occupation in KB. We make two observations.

First, we observe the existence of gender biases in KB embeddings, and note their consistency with real-world biases. For example, *engineer* and *nurse* have more extreme bias scores respectively towards male and female, while *singer* and *animator* have more moderate ones (quantitative analyses in §4).

Second, although the gender ratio of person entities has a great impact on $\mathcal{B}_{\text{gr}}$, it is not the only decisive factor. For example, *animator* has a gender ratio of 5.7:1, but its $\mathcal{B}_{\text{gr}}$ is biased towards female.

Inspecting the Origins of Biases The second observation motivates us to trace the origins of biases. More concretely, in the context of KB: *how do different triples contribute to $\mathcal{B}_{gr}$?* To answer this question, we apply influence function (Equation 3) with $F = \mathcal{B}_{\text{gr}}$ and observe how removing a training triple changes the overall group bias score.

One appealing property of TransE is that we are able to derive a closed-form Hessian matrix. Moreover, by further analyses, we can directly obtain a *diagonal approximation* of the Hessian matrix, and thus the Hessian inverse $\mathcal{I}_{\text{up, param}}$. Taking advantage of this, we can reduce the computation of $\mathcal{I}_{\text{up},\mathcal{B}_{\text{gr}}}$ to constant time complexity (w.r.t. training set size), which is much faster than the LiSSA algorithm (Agarwal et al., 2017) applied in (Koh and Liang, 2017), which requires $\mathcal{O}(n)$ time complexity to obtain a Hessian inverse approximation.

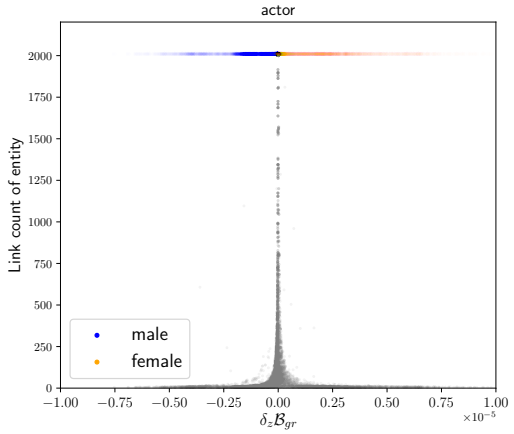


Figure 1: The relationship between different triples' $\delta_z\mathcal{B}_{\text{gr}}$ (i.e. the change of group bias $\mathcal{B}_{\text{gr}}$ if we re-train the KB embedding model without this triple) w.r.t. the occupation of *actor* and their person entities' attributes (i.e. the number of links of the entities and the gender information). Gender information is shown by blue and orange points, and node degrees is exhibited by grey points. We can observe that triples with positive (negative) $\delta_z\mathcal{B}_{\text{gr}}$ mostly contain person entities of female (male) gender. Moreover, triples contain high degree person entities are likely to have close-to-zero $\delta_z\mathcal{B}_{\text{gr}}$.

Concretely, using basic calculus, we have the following lemma and remarks. We include their detailed proof and derivations in Appendix B.

Lemma 1. *Suppose we generate the corresponding negative sample of a positive sample $\langle h, r, t\rangle$ by randomly choosing h or t and corrupting it to a random entity in E, we can derive the closed-form Hessian matrix of TransE with entries*

$$\mathbb{E}H_{\hat{\theta}} = \begin{array}{c} \\ e \\ e' \\ r \end{array} \begin{array}{c} \begin{array}{cccc} \quad & e & e' & r \end{array} \\ \begin{bmatrix} \ddots & \vdots & \vdots & \vdots \\ \cdots & \alpha_{ee} I_d & \alpha_{ee'} I_d & \alpha_{er} I_d \\ \cdots & \alpha_{ee'} I_d & \alpha_{ee} I_d & \alpha_{er} I_d \\ \cdots & \alpha_{er} I_d & \alpha_{er} I_d & 0 \\ \cdots & \vdots & \vdots & \vdots \end{bmatrix} \end{array} \begin{array}{c} \\ \\ \\ \\ \begin{matrix} r' \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ \ddots \end{matrix} \end{array},$$

where e, e' and r, r' are different entities and relations, α_{ee}, α_{er}, $\alpha_{ee'}$ are three different coefficients dependent on the frequencies of the corresponding entities and relations, and I_d is the identity matrix of $\mathbb{R}^{d\times d}$.

Remark 2. *In practice, we approximate the closed-form Hessian from Lemma 1 with its diagonal elements,*

$$\mathbb{E}H_{\hat{\theta}} \approx \text{diag}\{\underbrace{\cdots, \alpha_{ee} I_d, \cdots}_{\text{entities}}, \underbrace{\cdots, 0, \cdots}_{\text{relations}}\}.$$

Remark 3. *α_e could be zero or negative, which breaks the positive definiteness of $H_{\hat{\theta}}$. Following Koh and Liang (2017), we add λI ($\lambda > 0$) to $H_{\hat{\theta}}$ (i.e., $\alpha_{ee} \leftarrow \alpha_{ee} + \lambda$), which equals adding an L_2 regularization on parameters.*

Following Equation 3 ($\varepsilon = -1/|\mathcal{G}|$), we can compute the change of group bias (denoted by $\delta_z\mathcal{B}_{\text{gr}}$) after removing a training triple $z = \langle h, r, t\rangle$,[3]

$$\delta_z\mathcal{B}_{\text{gr}} = \frac{1}{|\mathcal{G}|}\nabla_\theta\mathcal{B}_{\text{gr}}^\top \left(\mathbb{E}H_{\hat{\theta}}^{-1}\nabla_\theta L(z, \hat{\theta})\right).$$

A triple z with positive $\delta_z\mathcal{B}_{\text{gr}}$ means that re-training without it will increase $\mathcal{B}_{\text{gr}}$ (i.e., towards "masculine") and vice versa. We note that due to the diagonal Hessian, z will have a non-zero influence iff it is reachable from o in two hops (i.e., entities of z take part in the computation of $\mathcal{B}_{\text{gr}}$). In practice, we calculate $\delta_z\mathcal{B}_{\text{gr}}$ of each triple in KB regarding $\mathcal{B}_{\text{gr}}$ of each occupation, and make three observations.

First, regarding relations in KB, we find most of the highly-influential triples (i.e. triples with highest absolute $\delta_z\mathcal{B}_{\text{gr}}$ values) to be of the profession relation (i.e., r_p) and its inverse[4]. For example, regarding the occupation of *singer*, these two relations occupy 74% of the top 1% positive triples and 78% of the top 1% negative triples. It suggests that compared with indirectly (i.e. two-hop) connected triples, triples directly connect with an entity have larger impact on it, which matches our intuitions.

Second, regarding gender, we find that most person entities in triples with high positive $\delta_z\mathcal{B}_{\text{gr}}$ are of female gender, and vice versa. Figure 1 take the occupation of *actor* as an example to illustrate this.[5] This observation agrees with previous observation: triples with person entities of male gender will drive the overall biases towards masculine, and removing them will reverse this effect.

Third, regarding graph substructure, we find that if a triple contains a high degree person entity, it usually has a nearly zero $\delta_z\mathcal{B}_{\text{gr}}$ (i.e. has small impact on other triples, see Figure 1), We suspect the reason to be as follows: the more neighbors

[3] More precisely, z is a pair of triples $(\langle h, r, t\rangle, \langle h', r, t'\rangle)$. To handle randomness of negative samples, we adopt two strategies in our implementation. First, we freeze negative samples in training epochs to get consistent results. Second, we use $\mathbb{E}H_{\hat{\theta}}$ to replace random $H_{\hat{\theta}}$ in influence functions.

[4] i.e. `people.person.people_with_this_profession`

[5] Similar patterns are observed in other occupations for both this observation and the next one.

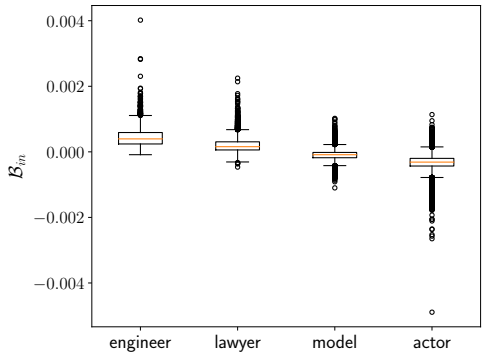


Figure 2: Distributions of $\mathcal{B}_{\text{in}}$. A positive value means the corresponding person entity's embedding is more biased towards male and vice versa. We can observe that for each occupation they are tightly distributed and consistent with real-world stereotypes.

an entity has, the more constraints its embedding needs to put on others (by link prediction). It makes the embedding less optimal to represent each constraint, and hence less influential to each triple.

3.2 Gender Biases of an Individual

Group-level correlation analyses offer us a coarse portrayal of biases. However, we are also interested in finer characterization (for each group member). Moreover, because of the complexity of KB structures, there very likely exist confounders between person entities and occupations (e.g. if two person entities of the same occupation are connected themselves, they are confounders of each other). In this case, **correlation does not imply causation**. In other words, gender differences are not guaranteed to be the only cause of $\mathcal{B}_{\text{gr}}$. Therefore, in this section, we study a further question: can we perform analyses on a specific person entity and measure its gender biases based on *how its gender change its link prediction error* (i.e. causality)?

By virtue of the structured knowledge in KB, we are able to conduct this individual-level analysis in a tractable way. The key idea is, *what if we keep everything else identical and* ***perturb*** *only the gender?* Intuitively, given an occupation o, if flipping a person entity's gender from female to male will make it easier to connect the person with o, $\mathbf{o}$ should be biased towards male. Formally, we define individual bias $\mathcal{B}_{\text{in}}$ of $\langle s, r_p, o\rangle$ as

$$\mathcal{B}_{\text{in}} = \psi(s, r_p, o)|_f - \psi(s, r_p, o)|_m,$$

where $\psi|_f$ ($\psi|_m$) is ψ computed on a version of TransE where s's gender is female (male). A high $\mathcal{B}_{\text{in}}$ means that, it is more difficult to predict s's occupation if s is female. We would thus argue that $\langle s, r_p, o\rangle$ is biased toward male. Because we keep all other attributes identical, this counterfactual definition naturally offers us causation.

The practical issue of $\mathcal{B}_{\text{in}}$ is the computation of the counterfactual: for each triple, this definition naively requires the re-training of the entire embedding model. This is intractable for large-scale analyses because of the extremely high computational cost. To avoid this issue, we apply influence function (Equation 4) for a fast evaluation of $\mathcal{B}_{\text{in}}$. Indeed, using Lemma 1 and Remark 2, we can obtain a closed-form $\mathcal{B}_{\text{in}}$ (proof in Appendix B).

Corollary 4. *Assume that for each person entity s, we have the same negative sample for $\langle s, r_p, f\rangle$ and $\langle s, r_p, m\rangle$, then*

$$\mathcal{B}_{in} \approx -\frac{4}{\alpha_s|\mathcal{G}|}\left(\mathbf{s} + \mathbf{r}_p - \mathbf{o}\right)^\top \left(\mathbf{m} - \mathbf{f}\right), \quad (5)$$

One important observation of $\mathcal{B}_{\text{in}}$ is that it is essentially a mixture of *local* graph substructure information (α_s, the degree of s in KB), and a projection of link prediction residual ($\mathbf{s}+\mathbf{r}_p-\mathbf{o}$) onto the gender difference ($\mathbf{m}-\mathbf{f}$, a reminiscence of word embedding gender subspace proposed in Bolukbasi et al., 2016). Compared with directly projecting $\mathbf{o}$ onto $\mathcal{B}_{\text{wo}}$ (a hard generalization of word embedding bias measure), the link prediction residual is more compatible with the TransE learning objective.

Figure 2 exhibits the distributions of $\mathcal{B}_{\text{in}}$ of several occupations. We make two observations from the results. First, although there are a number of outliers, most $\mathcal{B}_{in}$ are tightly distributed. It shows the consistency of the individual bias scores among different triples. Second, the bias scores correlate well with real-world gender stereotypes: *engineer* and *lawyer* lean more towards male, while *model* and *actor* lean more towards female. It suggests the validity of $\mathcal{B}_{\text{in}}$ in describing biases in KB.

Differences with Fisher et al. (2020b) A similar definition of bias is proposed in Fisher et al. (2020b) (denoted as $\mathcal{B}'_{\text{in}}$). $\mathcal{B}'_{\text{in}}$ is defined as follows: After training the embedding model to convergence, they perform one extra step of updating on the gender direction. The bias score is defined as the difference of the link prediction error before and after the update. We would like to note here the two aspects of differences between $\mathcal{B}_{\text{in}}$ and $\mathcal{B}'_{\text{in}}$.

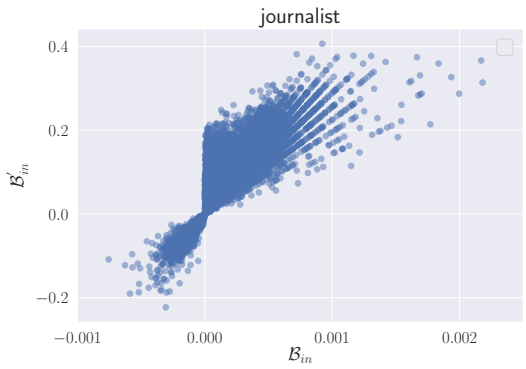


Figure 3: Relationship between $\mathcal{B}_{\text{in}}$ and $\mathcal{B}'_{\text{in}}$. We observe that these two bias measures align well. Nevertheless, there exist a substantial amount of data points with positive $\mathcal{B}_{\text{in}}$ but near zero $\mathcal{B}'_{\text{in}}$.

First, compared with $\mathcal{B}'_{\text{in}}$, $\mathcal{B}_{\text{in}}$ offers better interpretability. Concretely, in our definition, we approximate a purely counterfactual setting: flip the gender and re-train the entire model until convergence. In contrast, Fisher et al. (2020b) update the embedding after the convergence, which may not happen in real-world training.

Second, $\mathcal{B}_{\text{in}}$ takes more structural information into account. Under the case of TransE, $\mathcal{B}'_{\text{in}}$ can be expanded into the form (details in Appendix B),

$$\mathcal{B}'_{\text{in}} \propto -(\mathbf{s} + \mathbf{r}_p - \mathbf{o})^\top (\mathbf{m} - \mathbf{f}). \quad (6)$$

Compared with Equation 6, Equation 5 (approximation of $\mathcal{B}_{\text{in}}$) has an additional graph information term α_s. Intuitively, α_s serves as a normalization term: entities with more connections will be less affected by a single perturbation. In other words, the more connections an entity has, the less its link prediction error relies on one of them (i.e. gender).

Again, take the occupation of journalist as an example, we show the relationship between $\mathcal{B}_{\text{in}}$ and $\mathcal{B}'_{\text{in}}$ in Figure 3 and make two observations. First, there is a strong correlation between these two bias measures: points are approximately distributed along the diagonal. Second, we notice that there exist a substantial number of data points with positive $\mathcal{B}'_{\text{in}}$ but near zero $\mathcal{B}_{\text{in}}$. This suggests that the normalization term α_s can prevent the overestimation of biases of person entities with many connections. This also corresponds to our third observation regarding the distribution of $\delta_z \mathcal{B}_{\text{gr}}$ (§3.1).

3.3 Connections between $\mathcal{B}_{\text{gr}}$ and $\mathcal{B}_{\text{in}}$

After obtaining $\mathcal{B}_{\text{in}}$, a remaining question is: given a group of person entities, *how to use individual biases to characterize the group's overall bias?* The rationale behind is that, if we can accurately measure biases of individuals, we should be able to aggregate them to represent biases of the group.

A natural solution to this question is to directly average $\mathcal{B}_{\text{in}}$. However, in practice, we find that the averaged $\mathcal{B}_{\text{in}}$ of all occupations align poorly with $\mathcal{B}_{\text{gr}}$ ($r \approx 0.27$). We suspect this inconsistency to originate from the mismatches among different person entities' contexts in KB (i.e. different connections and local substructure). In other words, without controlling backdoor variables, simply averaging associations observed from each individual may not be suitable for representing association of the entire group (Pearl et al., 2016).[6]

In our analyses, because of the complexity of KB, it is infeasible to control all factors. Nevertheless, we can control some of them to alleviate this issue. Here, we focus on controlling gender for two reasons. First, occupations in KB are often of very imbalanced gender ratios (e.g., *nurse* connects with more female entities than male entities). At the same time, different genders usually have different distributions of $\mathcal{B}_{\text{in}}$: female entities mainly have above zero $\mathcal{B}_{\text{in}}$, while $\mathcal{B}_{\text{in}}$ of male entities distributes in a wider range.[7] Therefore, controlling gender should be able to effectively reduce the context mismatch. Second, because we treat the average link prediction error of each gender equally in group bias (§3.1), controlling gender can help us to obtain more comparable results.

We thus propose to average scores of each gender separately to calibrate this mismatch (*weighted averaging* instead of *vanilla averaging*). Formally,

$$\frac{1}{|F|} \sum_{s \in F} \mathcal{B}_{\text{in}}(\langle s, r_p, o \rangle) + \frac{1}{|M|} \sum_{s \in M} \mathcal{B}_{\text{in}}(\langle s, r_p, o \rangle).$$

We find weighted averaging align much better with $\mathcal{B}_{\text{gr}}$ ($r \approx 0.50$) and real-world biases (§4.1).

4 In-depth Analyses

4.1 Comparison with Real-world Biases

One method of inspecting the validity of a bias measure is to analyze its connection with real-world statistics (e.g. gender ratios of occupations). However, most existing datasets fail to meet our needs,

[6] Other examples of this phenomenon include *Simpson's Paradox* and *ecological fallacy*.

[7] We show $\mathcal{B}_{\text{in}}$ distribution of the occupation of *journalist* as an example in Figure 5 in Appendix C, and find similar trends in other occupations.

	vanilla		*weighted*	
	r	p	r	p
$\mathcal{B}'_{\text{in}}$	0.470	**.003**	0.590	$\mathbf{< 10^{-4}}$
$\mathcal{B}_{\text{in}}$	0.480	**.002**	0.610	$\mathbf{< 10^{-4}}$
$\mathcal{B}_{\text{gr}}$	–	–	0.668	$\mathbf{< 10^{-5}}$

Table 2: Alignment results with census data. *Vanilla* and *weighted* denotes for vanilla averaging and weighted averaging, respectively. Note that because $\mathcal{B}_{gr}$ is not applicable for averaging strategies, we simply put its score into *weighted*. Significant p values ($< .01$) are shown in bold font.

Occupation	$\mathcal{B}_{\text{gr}}$	**de-biased** $\mathcal{B}_{\text{gr}}$
architect	7.30×10^{-2}	4.32×10^{-2}
physicist	2.16×10^{-2}	0.22×10^{-2}
actor	-7.67×10^{-2}	-6.53×10^{-2}
nurse	-9.80×10^{-2}	-9.16×10^{-2}

Table 3: Examples of $\mathcal{B}_{\text{gr}}$ before and after adding dummy entities. We observe that this strategy can effectively mitigate bias, although the extent differs among different occupations.

because they have different occupation categories with FB5M (e.g. Garg et al., 2018; Du et al., 2019).

Accordingly, we take the following steps to build a new dataset. First, we collect the gender distributions of occupations in 2018 by the U.S. census data (Ruggles et al., 2020). Afterwards, we calculate their log proportions[8] and manually pair up them with occupations in KB.[9] We use it as our validation data and refer it as *census data*.

Table 2 shows the Pearson's r values and p values between census data and all five bias measures described in §3 ($\mathcal{B}_{\text{gr}}$, $\mathcal{B}_{\text{in}}$ and $\mathcal{B}'_{\text{in}}$ with both averaging strategies). Our observations are two fold.

First, both $\mathcal{B}_{\text{gr}}$ and $\mathcal{B}_{\text{in}}$ exhibit significant correlations (especially under *weighted averaging*) with census data ($p < .01$), indicating their validity of measuring gender biases in KB embeddings.

Second, individual bias measures ($\mathcal{B}_{\text{in}}$ and $\mathcal{B}'_{\text{in}}$) align better with census data under weight averaging than under vanilla averaging. This backs up our suspicion regarding contexts' mismatches, as well as our solution strategy (weighted averaging).

4.2 Accuracy of the Group Influence Approximation

Because the Hessian matrix we derived for the calculation of influence function is a diagonal approximation, and influence function of a group of training samples is only an approximation of the test performance change after re-training, one may concern the accuracy of our influence function. Therefore, in this section, we perform a validation experiment to address this concern. Specifically, for each occupation o, we first remove k triples with highest $\delta_z \mathcal{B}_{\text{gr}}$, then re-train the TransE model from scratch, and calculate their $\mathcal{B}_{\text{gr}}$ regarding o. Afterwards, we compare the sum of $\delta_z \mathcal{B}_{\text{gr}}$ with the ground truth changes in $\mathcal{B}_{\text{gr}}$. In practice, we take ks to be a arithmetic progression from 500 to 5000, with a common difference of 500.

We show a few occupations' alignment results as examples in Figure 4a-4c. We observe strong correlations between our approximation and the ground truth ($r > 0.95$ for all occupations). It suggests the accuracy of our approximation (some additional results in Appendix C).

4.3 Application: De-biasing KB Embeddings

Our study can broadly benefit relevant future research regarding societal biases and KB. As examples of such applications, based on our study in §3.1, we explore two strategies for de-biasing KB embeddings. We note that these two strategies aim to exemplify the potential impacts of our previous study, and are not necessarily the best method to de-bias KB embeddings.[10] Instead, we highly encourage future studies to build better de-biasing methods on the basis of our findings.

Strategy 1: De-biasing by Adding In Table 1, we observe that gender ratio has a substantial impact on $\mathcal{B}_{\text{gr}}$. Based on this, one natural idea of de-biasing is to balance gender proportion by adding dummy triples. The advantage of this strategy is that, because we do not remove triples, we are able to keep the information of the original KB intact.

Specifically, suppose an occupation o with M male entities and F female entities, where M is larger than F. To alleviate bias, we create $c(M - F)$[11] dummy entities and connect them with only the female gender and o. Afterwards, we re-train TransE and observe the $\mathcal{B}_{\text{gr}}$ regarding o.

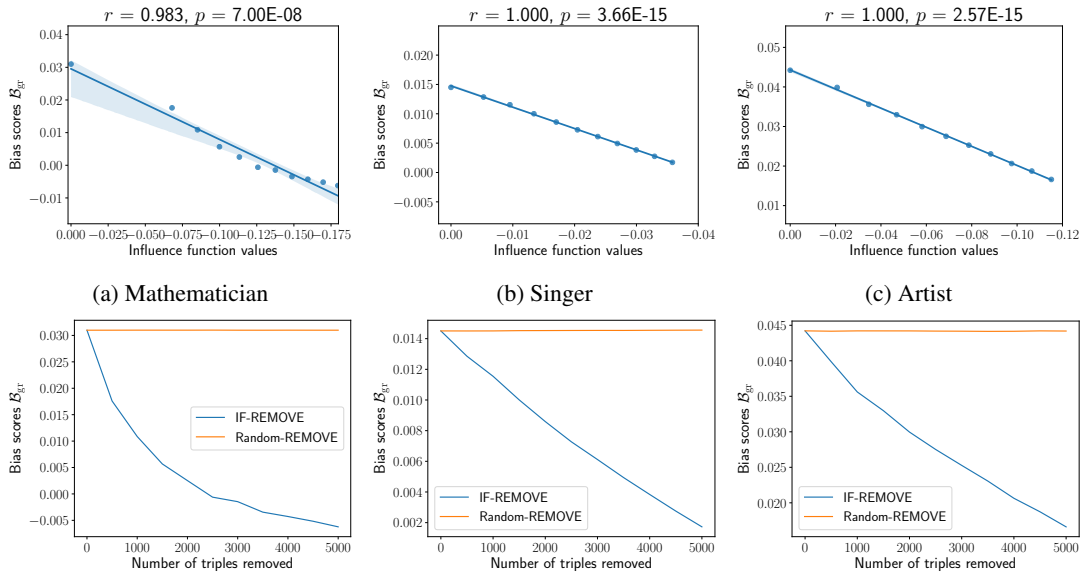
Table 3 lists a few examples of the results. We find that this de-biasing strategy overall works well.

[8] log-prop $= \frac{p}{1-p}$, where p is % of men in occupation.

[9] We apply a *many to many* pairing to match the occupation categories in census data and KB.

[10] For example, Fisher et al. (2020a) and Arduini et al. (2020) adopt adversarial loss for de-biasing KB embeddings.

[11] In practice, we set c to be 0.5, and limit the number of total added entities of each occupation to be < 10000.



(a) Mathematician (b) Singer (c) Artist

(d) Mathematician (e) Singer (f) Artist

Figure 4: 4a - 4c are the correlations between approximated influence and ground truth results (obtained by removing triples and re-training). The titles list the Pearson's r and significance p values of the alignments. 4d-4f show the $\mathcal{B}_{gr}$ of influence-function-based triples removing (IF-REMOVE) and random removing (Random-REMOVE). Compared with Random-REMOVE, IF-REMOVE reduces $\mathcal{B}_{gr}$ much more significantly.

It is worth noting that the changes of biases of some occupations (e.g. *nurse*) are smaller, which matches our previous observation: gender ratio is not the only decisive factor of $\mathcal{B}_{gr}$.

Strategy 2: De-biasing by Removing Based on our study on the origins of biases, and inspired by the validation results in §4.2, we investigate a straightforward de-biasing strategy: we simply remove the top k most influential triples based on the approximation of influence function (IF-REMOVE). Again, we take ks to be $[500, 1000, 1500, ..., 5000]$. Besides, for the purpose of controlling variable, we compare it to a naive baseline method, in which we randomly delete triples of all entities (Random-REMOVE).

Figure 4d-4f exhibit some examples of the results. We observe that comparing with the baseline, where $\mathcal{B}_{gr}$ rarely change, this de-biasing strategy is able to mitigate biases very effectively. Several additional examples are included in Appendix C.

5 Related Work

Various measures have been proposed to quantify gender biases in word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Swinger et al., 2019). Many of them are based on vector similarity (e.g. cosine similarity) between words, which matches the training objective of most word embedding algorithms (maximize the vector similarities between similar words, Mikolov et al., 2013; Pennington et al., 2014). Moreover, Garg et al. (2018) suggest that word embedding can reflect biases in the training corpora and hence our society.

Recently, a few studies have explored gender biases in KBs and their embeddings. A pioneer study by Klein et al. (2016) investigates gender gap in Wikidata across time, space, culture, occupation and language. A following study (Shaik et al., 2021) further analyzes the race and country of citizenship bias in KB regarding STEM representation. Moreover, Janowicz et al. (2018) analyze the potential bias issues in KBs from both data and schema viewpoints. Fisher et al. (2020b) propose a KB embedding bias measure based on the change of link prediction error after a one-step update towards male. Fisher et al. (2020a) and Arduini et al. (2020) propose to use adversarial training objective to mitigate biases in KB embeddings.

Influence function is a commonly used technique in robust statistics (Cook and Weisberg, 1982). Koh and Liang (2017) first use it to inspect each training point's influence on a neural network's prediction. A following study by Koh et al. (2019) investigate the accuracy of influence function on measuring the effect of removing a group of training samples, and show that its approximation has strong correla-

tions with actual effects. Afterwards, Brunet et al. (2019) apply influence function as a differential bias measure to study gender bias in word embedding. Moreover, Pezeshkpour et al. (2019) use an simplification of influence function to perform adversarial attack on link prediction.

6 Conclusion and Discussion

In this paper, we study the gender biases in KB embeddings. First, we develop two bias measures to quantify biases: one from the group level and the other from the individual level. Evidence of their validity are obtained in comparison with real-world biases. Second, to understand the origins of biases, we adopt influence functions for triple-level analysis and develop an efficient method for fast evaluation. The accuracy of this method is validated by comparing our approximation with group-truth changes after re-training. Moreover, as examples of the potential applications of our findings, we propose two de-biasing strategies for KB embeddings and obtain promising performance.

Although we focus on Freebase (FB5M) and TransE in this paper, we note that our analyses are theoretically generalizable to other commonly-used KBs and embedding training algorithms. For instance, Wikidata, another commonly-used KB, uses a different hierarchical structure to organize its data (Vrandečić and Krötzsch, 2014; Tanon et al., 2016; Wang et al., 2021). However, it still loosely follows the triple structure used in Freebase, and therefore can be pre-processed to fit in our analyses. Also, because our bias measures and bias tracing methods are built on simple and generalizable definitions (i.e., differences between link predictions errors and influence function), they can naturally be adapted to other KB embedding algorithms (Lin et al., 2015; Yang et al., 2015; Peng et al., 2021).

However, we recognize that such generalizations are not trivial efforts. Take Wikidata again for an instance, although a simple transformation is adequate for running the embedding algorithm, it is far from fully eliminating the differences between Freebase and Wikidata. For example, Wikidata does not have an inverse predicate for each relation, and has a much smaller number of overall relations (Azmy et al., 2018; Diefenbach et al., 2017). Such differences might have a large impact on the final results. Also, to perform the same analyses with other embedding algorithms, we will need to develop algorithms to facilitate the computation of their influence function (as Lemma 1), too. Therefore, we consider such generalizations to be promising future directions but out of the scope of our work.

Acknowledgement

We thank Dong Nguyen for her meticulous and valuable suggestions, as well as productive discussions. We also thank all anonymous reviewers for their constructive and helpful feedback. This research was (partially) supported by NSFC (62076097), STCSM(20511101205), and Shanghai Key Laboratory of Multidimensional Information Processing, ECNU (2020KEY001). The corresponding authors are Yuanbin Wu and Yan Yang.

Ethical Statement

Intended Usage Our work intend to provide insights of gender biases in KB and its embeddings, on how to measure these biases and how to trace the origins of them. Moreover, as discussed in §4.3, future studies can build better de-biasing methods based on our findings. In this way, our framework can contribute to the development of models that are less biased and hence potentially less harmful.

Limitations In this study, we use gender information already encoded in KB to measure and trace gender biases. However, because only binary gender is recorded in the KB that we use (Freebase), we take a narrow view of binary gender in our analyses. We hope to see more future studies on gender biases in KB embeddings that consider non-binary gender identities as well as intersectional identities.

References

Naman Agarwal, Brian Bullins, and Elad Hazan. 2017. Second-order stochastic optimization for machine learning in linear time. *Journal of Machine Learning Research*, 18(116):1–40.

Mario Arduini, Lorenzo Noci, Federico Pirovano, Ce Zhang, Raj Shrestha, and Bibek Paudel. 2020. Adversarial learning for debiasing knowledge graph embeddings. In *Proceedings of the 16th International Workshop on Mining and Learning with Graphs (MLG)*.

Michael Azmy, Peng Shi, Jimmy Lin, and Ihab Ilyas. 2018. Farewell Freebase: Migrating the SimpleQuestions dataset to DBpedia. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2093–2103, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kurt Bollacker, Patrick Tufts, Tomi Pierce, and Robert Cook. 2007. A platform for scalable, collaborative, structured information integration. In *Intl. Workshop on Information Integration on the Web (IIWeb'07)*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Styliani Bourli and Evaggelia Pitoura. 2020. Bias in knowledge graph embeddings. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 6–10.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

R Dennis Cook and Sanford Weisberg. 1982. *Residuals and influence in regression*. New York: Chapman and Hall.

Dennis Diefenbach, Thomas Tanon, Kamal Singh, and Pierre Maret. 2017. Question answering benchmarks for wikidata. In *ISWC 2017*.

Yupei Du, Yuanbin Wu, and Man Lan. 2019. Exploring human gender stereotypes with word association test. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6133–6143, Hong Kong, China. Association for Computational Linguistics.

Joseph Fisher, Arpit Mittal, Dave Palfrey, and Christos Christodoulopoulos. 2020a. Debiasing knowledge graph embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7332–7345, Online. Association for Computational Linguistics.

Josheph Fisher, Dave Palfrey, Christos Christodoulopoulos, and Arpit Mittal. 2020b. Measuring social bias in knowledge graph embeddings. *Proceedings of the Bias in Automatic Knowledge Graph Construction Workshop*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Krzysztof Janowicz, Bo Yan, Blake Regalia, Rui Zhu, and Gengchen Mai. 2018. Debiasing knowledge graphs: Why female presidents are not like female popes. In *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, USA, October 8th - to - 12th, 2018*, volume 2180 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Maximilian Klein, Harsh Gupta, Vivek Rai, Piotr Konieczny, and Haiyi Zhu. 2016. Monitoring the gender gap with wikidata human gender indicators. In *Proceedings of the 12th International Symposium on Open Collaboration*, OpenSym '16, New York, NY, USA. Association for Computing Machinery.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

Pang Wei W Koh, Kai-Siang Ang, Hubert Teo, and Percy S Liang. 2019. On the accuracy of influence functions for measuring group effects. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2181–2187. AAAI Press.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Xutan Peng, Guanyi Chen, Chenghua Lin, and Mark Stevenson. 2021. Highly efficient knowledge graph embedding learning with Orthogonal Procrustes Analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2364–2375, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating robustness and interpretability of link prediction via adversarial modifications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3336–3347, Minneapolis, Minnesota. Association for Computational Linguistics.

Wessel Radstok, Melisachew Wudage Chekol, and Mirko T. Schäfer. 2021. Are knowledge graph embedding models biased, or is it the data that they are trained on? In *Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24, 2021*, volume 2982 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2020. Ipums usa: Version 10.0.

Zaina Shaik, Filip Ilievski, and Fred Morstatter. 2021. Analyzing race and country of citizenship bias in wikidata. *arXiv preprint arXiv:2108.05412*.

Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 305–311, New York, NY, USA. Association for Computing Machinery.

Thomas Pellissier Tanon, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. From freebase to wikidata: The great migration. In *World Wide Web Conference*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Da Zheng, Xiang Song, Chao Ma, Zeyuan Tan, Zihao Ye, Jin Dong, Hao Xiong, Zheng Zhang, and George Karypis. 2020. Dgl-ke: Training knowledge graph embeddings at scale. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 739–748, New York, NY, USA. Association for Computing Machinery.

A Experimental Setup

Choices of datasets Freebase (Bollacker et al., 2007) is one of the largest publicly available KBs, with over three billion triples covering a wide range of real-world facts. Due to time and hardware constraints, in this work, we use its subset FB5M (Bordes et al., 2015) as the KB for our experiments. In practice, we find that although FB5M only holds 0.5% of the triples from Freebase, it covers a much higher percentage of human type entities and their related facts. Regarding professions, we select ones with ≥ 400 person entities and contain both male and female in FB5M.

TransE training We use DGL-KE 0.1.0 (Zheng et al., 2020) to train TransE embeddings. To get deterministic results across different training runs, we fix the random seeds and restricted the training process to run under a single thread.

Due to that TransE involves negative sampling in its training objective, we save all negative samples from the final epoch to make sure that influence function can output accurate results.[12] Regarding hyper-parameters, we use a number of dimensions of 200, a batch size of 8000, and stop training after 120000 updating steps. It takes approximately 40 minutes with a single GTX TITAN X GPU.

B Proofs and Derivations

Proof of Lemma 1. We use E and R to denote the full set of entities and relations, and N_e and N_r to denote the times of occurrence of entity e and relation r in KB. Also, we use $|\mathcal{G}|$, $|E|$, and $|R|$ to respectively denote the number of triples in KB, and the number of different entities and relations. Moreover, we define a counting function C to denote the times of occurrence of certain triples,

$$C(e_0, r_0, e_1) = \sum_{\langle h,r,t\rangle \in \mathcal{G}} \mathbb{1}(h = e_0, r = r_0, t = e_1),$$

where $\mathbb{1}$ is the indicator function. Also, we use $*$ as a wildcard element. For example,

$$C(*, r_0, e_1) = \sum_{e_0 \in E} C(e_0, r_0, e_1),$$
$$N_r = C(*, r_0, *) = \sum_{e_0, e_1 \in E} C(e_0, r_0, e_1),$$
$$C(*, *, *) = |\mathcal{G}|.$$

[12] To ensure the results from influence function are accurate, we only use the negative samples when the embeddings are close to convergence.

For the TransE loss on a single triple $\langle h, r, t\rangle$, $\varphi(h, r, t)$, it is easy to derive the second-order derivatives,

$$\begin{aligned}
&\nabla^2_{\mathbf{h},\mathbf{h}}\varphi(h,r,t) = \nabla^2_{\mathbf{r},\mathbf{r}}\varphi(h,r,t) \\
&= \nabla^2_{\mathbf{t},\mathbf{t}}\varphi(h,r,t) = \nabla^2_{\mathbf{h},\mathbf{r}}\varphi(h,r,t) \\
&= \nabla^2_{\mathbf{r},\mathbf{h}}\varphi(h,r,t) = -\nabla^2_{\mathbf{h},\mathbf{t}}\varphi(h,r,t) \\
&= -\nabla^2_{\mathbf{t},\mathbf{h}}\varphi(h,r,t) = -\nabla^2_{\mathbf{r},\mathbf{t}}\varphi(h,r,t) \\
&= -\nabla^2_{\mathbf{t},\mathbf{r}}\varphi(h,r,t) = 2I_d,
\end{aligned} \tag{7}$$

where I_d is the identity matrix of $\mathbb{R}^{d\times d}$. We observe that *the value of the second-order derivative is independent of the triple*.

The expectation of Hessian matrix of the overall loss function $\mathcal{L}$ (Equation 1) consists of five parts: $\mathbb{E}\nabla^2_{e,e}\mathcal{L}$, $\mathbb{E}\nabla^2_{e,e'}\mathcal{L}$, $\mathbb{E}\nabla^2_{r,r}\mathcal{L}$, $\mathbb{E}\nabla^2_{r,r'}\mathcal{L}$, and $\mathbb{E}\nabla^2_{e,r}\mathcal{L}$, where e, e' and r, r' denotes two different entities and relations. Because we only have a single relation in each triple, we can immediately see that $\mathbb{E}\nabla^2_{r,r'}\mathcal{L}$ is always zero. Moreover, because we train TransE embeddings with negative sampling, and the relation r is the same for positive and negative samples, we know that $\mathbb{E}\nabla^2_{r,r}\mathcal{L}$ is zero as well.

We consider two types of training samples to calculate the remaining terms: e and e' appear in a positive triple (denoted as $\nabla^2\mathcal{L}_{\text{pos}}$), and e and e' are sampled to corrupt a sample ($\nabla^2\mathcal{L}_{\text{neg}}$).

For the first case, when e appears in a positive triple $\langle h, r, t\rangle$, there will be a corresponding negative sample, with 0.5 probability to be $\langle h', r, t\rangle$, where $h' \neq h$, and 0.5 probability of $\langle h, r, t'\rangle$, where $t' \neq t$. Using Equation 7, we obtain

$$\begin{aligned}
&\mathbb{E}\nabla^2_{e,e}\mathcal{L}_{\text{pos}} \\
&= C(e,*,*)(\nabla^2_{\mathbf{h},\mathbf{h}}\varphi(h,r,t) \\
&-0.5\nabla^2_{\mathbf{h},\mathbf{h}}\varphi(h',r,t) - 0.5\nabla^2_{\mathbf{h},\mathbf{h}}\varphi(h,r,t')) \\
&+C(*,*,e)(\nabla^2_{\mathbf{t},\mathbf{t}}\varphi(h,r,t) \\
&-0.5\nabla^2_{\mathbf{t},\mathbf{t}}\varphi(h',r,t) - 0.5\nabla^2_{\mathbf{t},\mathbf{t}}\varphi(h,r,t')) \\
&= (C(e,*,*) + C(*,*,e))I_d = N_e I_d,
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\nabla^2_{e,e'}\mathcal{L}_{\text{pos}} \\
&= C(e,*,e')(\nabla^2_{\mathbf{h},\mathbf{t}}\varphi(h,r,t) \\
&-0.5\nabla^2_{\mathbf{h},\mathbf{t}}\varphi(h',r,t) - 0.5\nabla^2_{\mathbf{h},\mathbf{t}}\varphi(h,r,t')) \\
&+C(e',*,e)(\nabla^2_{\mathbf{t},\mathbf{h}}\varphi(h,r,t) \\
&-0.5\nabla^2_{\mathbf{t},\mathbf{h}}\varphi(h',r,t) - 0.5\nabla^2_{\mathbf{t},\mathbf{h}}\varphi(h,r,t')) \\
&= -2(C(e,*,e') + C(e',*,e))I_d,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\nabla^2_{e,r}\mathcal{L}_{\text{pos}} \\
&= C(e,r,*)(\nabla^2_{\mathbf{h},\mathbf{r}}\varphi(h,r,t) \\
&-0.5\nabla^2_{\mathbf{h},\mathbf{r}}\varphi(h',r,t) - 0.5\nabla^2_{\mathbf{h},\mathbf{r}}\varphi(h,r,t')) \\
&+C(*,r,e)(\nabla^2_{\mathbf{r},\mathbf{t}}\varphi(h,r,t) \\
&-0.5\nabla^2_{\mathbf{r},\mathbf{t}}\varphi(h',r,t) - 0.5\nabla^2_{\mathbf{r},\mathbf{t}}\varphi(h,r,t')) \\
&= (C(e,r,*) - C(*,r,e))I_d.
\end{aligned}$$

For the second case, since we corrupt a triple by uniformly sampling all entities, an entity is expected to be sampled as the head entity and the tail entity with the same probability of $1/2|E|$. Using Equation 7, we obtain

$$\begin{aligned}
&\mathbb{E}\nabla^2_{e,e}\mathcal{L}_{\text{neg}} \\
&= -2C(*,*,*)\nabla^2_{\mathbf{h},\mathbf{h}}\varphi(h,r,t)/2|E| \\
&= (-2|\mathcal{G}|/|E|)I_d,
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\nabla^2_{e,e'}\mathcal{L}_{\text{neg}} \\
&= -C(*,*,e')\nabla^2_{\mathbf{h},\mathbf{t}}\varphi(h,r,t)/2|E| \\
&-C(e',*,*)\nabla^2_{\mathbf{t},\mathbf{h}}\varphi(h,r,t)/2|E| \\
&-C(e,*,*)\nabla^2_{\mathbf{h},\mathbf{t}}\varphi(h,r,t)/2|E| \\
&-C(*,*,e)\nabla^2_{\mathbf{t},\mathbf{h}}\varphi(h,r,t)/2|E| \\
&= (N_e/|E| + N_{e'}/|E|)I_d,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\nabla^2_{e,r}\mathcal{L}_{\text{neg}} \\
&= -C(*,r,*)\nabla^2_{\mathbf{h},\mathbf{r}}\varphi(h,r,t)/2|E| \\
&-C(*,r,*)\nabla^2_{\mathbf{t},\mathbf{r}}\varphi(h,r,t)/2|E| \\
&= 0.
\end{aligned}$$

Putting $\mathcal{L}_{\text{pos}}$ and $\mathcal{L}_{\text{neg}}$ together, we obtain

$$\begin{aligned}
&\mathbb{E}\nabla^2_{e,e}\mathcal{L} = (N_e - 2|\mathcal{G}|/|E|)I_d = \alpha_{ee}I_d, \\
&\mathbb{E}\nabla^2_{e,r}\mathcal{L} = (C(e,r,*) - C(*,r,e))I_d = \alpha_{er}I_d, \\
&\mathbb{E}\nabla^2_{e,e'}\mathcal{L} = (N_e/|E| + N_{e'}/|E| \\
&-2C(e,*,e') - 2C(e',*,e))I_d = \alpha_{ee'}I_d, \\
&\mathbb{E}\nabla^2_{r,r}\mathcal{L} = 0, \\
&\mathbb{E}\nabla^2_{r,r'}\mathcal{L} = 0.
\end{aligned}$$

□

Derivations of Remark 2. Clearly, the magnitude of non-zero diagonal terms of the Hessian (i.e. $\nabla^2_{\mathbf{e},\mathbf{e}}\mathcal{L}$) are much larger than those of the non-zero non-diagonal terms (i.e. $\nabla^2_{\mathbf{e},\mathbf{r}}\mathcal{L}$ and $\nabla^2_{\mathbf{e},\mathbf{e}'}\mathcal{L}$), because the former only requires the occurrence of e (mostly hundreds or thousands of times for human entities), while the latter requires the co-occurrence of two terms (usually once or much smaller than the number of occurrences of the corresponding entity). We therefore propose to *approximate the Hessian matrix with its diagonal elements*. With such approximation, we estimate the expectation of the Hessian matrix as

$$\mathbb{E}H_{\hat{\theta}} \approx \text{diag}\{\underbrace{\cdots,\alpha_{ee}I_d,\cdots}_{\text{entities}},\underbrace{\cdots,0,\cdots}_{\text{relations}}\},$$

where $\alpha_{ee} = N_e - 2|\mathcal{G}|/|E|$. □

Proof of Corollary 4. We consider the case that the TransE parameter $\hat{\theta}$ is learned with $\langle s, r_p, f\rangle$ in KB and we perturb it to $\langle s, r_p, m\rangle$. The other direction is identical. Following Equation 4 by setting $F = \psi(s,r_p,o)$, $z = \langle s, r_g, f\rangle$, $z' = \langle s, r_g, m\rangle$ and $\varepsilon = -1/|\mathcal{G}|$, we have

$$\begin{aligned}
\mathcal{B}_{\text{in}} \approx \tfrac{1}{|\mathcal{G}|}&\nabla_\theta\psi(s,r_p,o)^\top H^{-1}_{\hat{\theta}} \\
&\left(\nabla_\theta L(z,\hat{\theta}) - \nabla_\theta L(z',\hat{\theta})\right).
\end{aligned}$$

Let $\mathbf{d}_1 = 2(\mathbf{s} + \mathbf{r_p} - \mathbf{o})$, $\nabla_\theta\psi(s,r_p,o)^\top$ equals

$$\begin{bmatrix} .. & \overset{s}{\mathbf{d}_1^\top} & ... & \overset{r_p}{\mathbf{d}_1^\top} & ... & \overset{o}{-\mathbf{d}_1^\top} & .. \end{bmatrix}.$$

On the other side, $\nabla_\theta L(z,\hat{\theta}) - \nabla_\theta L(z',\hat{\theta}) = \nabla_\theta\psi(s,r_g,f) - \nabla_\theta\psi(s,r_g,m)$ by cancelling negative samples. Let $\mathbf{d}_2 = 2(\mathbf{f} - \mathbf{m})$, $\mathbf{d}_3 = 2(\mathbf{h} + \mathbf{r}_g - \mathbf{f})$ and $\mathbf{d}_4 = 2(\mathbf{h} + \mathbf{r}_g - \mathbf{m})$, its transpose equals

$$\begin{bmatrix} ... & \overset{s}{\mathbf{d}_2^\top} & ... & \overset{r_g}{\mathbf{d}_2^\top} & ... & \overset{f}{\mathbf{d}_3^\top} & ... & \overset{m}{-\mathbf{d}_4^\top} & ... \end{bmatrix}.$$

Finally, by approximating $H_{\hat{\theta}}$ using $\mathbb{E}H_{\hat{\theta}}$ (Lemma 1), we see that only the product of $\mathbf{d_i}$ and $\mathbf{d}_2$ is non-zero. □

Derivations of Fisher et al. (2020b). To measure gender biases in KB embeddings, Fisher et al. (2020b) first define a function m to be the difference between link prediction error of male and female entity,

$$m(\theta) = \psi(s,r_g,m) - \psi(s,r_g,f).$$

Afterwards, the bias score of a person entity regarding an occupation o is the change of the link

prediction error after updating the entity embedding using the gradient of m (i.e., updating $\mathbf{s}$ to make m larger),

$$\begin{aligned}\mathcal{B}'_{\text{in}} &= \psi(s', r_p, o) - \psi(s, r_p, o),\\ \text{where } \mathbf{s}' &= \mathbf{s} + \eta\frac{\mathsf{d}m}{\mathsf{d}\mathbf{s}}.\end{aligned}$$

For L_2 TransE loss, the gradient equals to

$$\begin{aligned}\frac{\mathsf{d}m}{\mathsf{d}\mathbf{s}} &= 2(\mathbf{s} + \mathbf{r_g} - \mathbf{f}) - 2(\mathbf{s} + \mathbf{r_g} - \mathbf{m})\\ &= 2(\mathbf{f} - \mathbf{m}).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathcal{B}'_{\text{in}} &= \psi(s', r_p, o) - \psi(s, r_p, o)\\ &= (\mathbf{s} + 2\eta(\mathbf{f} - \mathbf{m}) + \mathbf{r}_p - \mathbf{o})^2 - (\mathbf{s} + \mathbf{r}_p - \mathbf{o})^2\\ &= (\mathbf{s} + \mathbf{r}_p - \mathbf{o})^2 + 4\eta(\mathbf{s} + \mathbf{r}_p - \mathbf{o})^\top(\mathbf{m} - \mathbf{f})\\ &+ 4\eta^2(\mathbf{m} - \mathbf{f})^2 - (\mathbf{s} + \mathbf{r}_p - \mathbf{o})^2\\ &= k + 4\eta(\mathbf{s} + \mathbf{r}_p - \mathbf{o})^\top(\mathbf{m} - \mathbf{f})\end{aligned}$$

Omitting the constant part $k = 4\eta^2(\mathbf{m} - \mathbf{f})^2$, we can find that $\mathcal{B}'_{\text{in}}$ is essentially the projection of link prediction error $\mathbf{s} + \mathbf{r}_p - \mathbf{o}$ onto gender subspace $\mathbf{m} - \mathbf{f}$, which is similar to $\mathcal{B}_{\text{in}}$. □

C Additional Results

More Figures are in the next page.

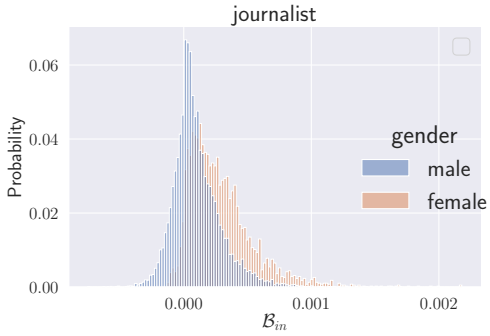
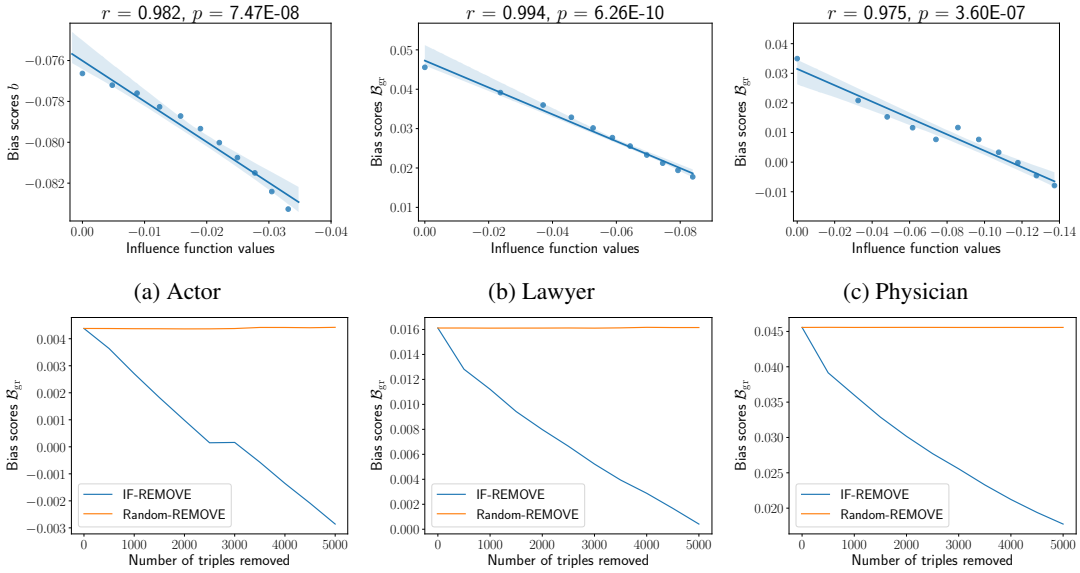


Figure 5: Different distributions of $\mathcal{B}_{\text{in}}$ between male and female entities. We can observe that the $\mathcal{B}_{\text{in}}$ distributions of different genders are clearly different.



(a) Actor (b) Lawyer (c) Physician

(d) Screenwriter (e) Musician (f) Lawyer

Figure 6: Figure 6a-6c exhibit additional results for the validation of group influence approximation. Figure 6d-6f show additional results for de-biasing KB embeddings by removing highly influential triples.