Model-Agnostic Bias Measurement in Link Prediction

Lena Schwertmann
Hasso Plattner Institute / University of Potsdam
Potsdam, Germany
lena-schwertmann@gmx.de

Manoj Prabhakar Kannan Ravi*
LexisNexis
Berlin, Germany
manoj.prabhakarkr91@gmail.com

Gerard de Melo
Hasso Plattner Institute / University of Potsdam
Potsdam, Germany
gerard.demelo@hpi.de

Abstract

Link prediction models based on factual knowledge graphs are commonly used in applications such as search and question answering. However, work investigating social bias in these models has been limited. Previous work focused on knowledge graph embeddings, so more recent classes of models achieving superior results by fine-tuning Transformers have not yet been investigated. We therefore present a model-agnostic approach for bias measurement leveraging fairness metrics to compare bias in knowledge graph embedding-based predictions (KG only) with models that use pre-trained, Transformer-based language models (KG+LM). We further create a dataset to measure gender bias in occupation predictions and assess whether the KG+LM models are more or less biased than KG only models. We find that gender bias tends to be higher for the KG+LM models and analyze potential connections to the accuracy of the models and the data bias inherent in our dataset. Finally, we discuss limitations and ethical considerations of our work. The repository containing the source code and the data set is publicly available at https://github.com/lena-schwert/comparing-bias-in-KG-models.

1 Introduction

Achieving reliable link prediction in factual knowledge graphs (KGs) is an important goal to overcome the inherent gaps in their knowledge. Such graphs are widely used by companies such as Google, LinkedIn, Amazon, and Bloomberg across a range of different real-world applications, including search, recommender systems, and voice-based question answering (Hogan et al., 2021; Weikum et al., 2021; Ji et al., 2021). Typically, information is stored in the shape of *triples* (h, r, t), consisting of a head entity h, a relation r, and a tail entity t. Entities express concepts, while relations express the connection between them, e.g., (Barack Obama, occupation, politician). Link prediction models score the plausibility of a given fact, with two distinct purposes: (i) They make the graph structure available to machine learning models, e.g., in the form of embeddings, (ii) and – if sufficiently reliable – may eventually be used to make plausible predictions of missing facts, i.e., solving the problem of automatic knowledge graph completion and refinement (Hogan et al., 2021; Paulheim, 2016).

While link prediction models are naturally evaluated for their accuracy, there have only recently been studies that assess possible biases that they may exhibit. Echoing prior position papers on bias in factual KGs (Janowicz et al., 2018; Kraft and Usbeck, 2022), we consider an analysis of bias as essential for a thorough model evaluation, especially because a) KGs contain sensitive information about humans (e.g., gender), b) historical facts naturally contain historical biases, and c) the applications of KG-based models are increasingly socially relevant due to their proliferation into widely deployed systems such as search engines and conversational agents (Kraft and Usbeck, 2022; Hogan et al., 2021). To achieve a meaningful bias analysis for the link prediction task, we argue that a *model-agnostic approach* is necessary. Only then can bias be measured comparatively across different model classes, highlighting strengths and weaknesses as well as potential causes for biased behavior.

However, existing bias measurement approaches are highly model-dependent (for a recent in-depth review, refer to Kraft and Usbeck 2022). They focus only on *knowledge graph embeddings (KGEs)* (Fisher et al., 2020b,a; Keidar et al., 2021; Rossi et al., 2021b; Radstok et al., 2021; Du et al., 2022; Bourli and Pitoura, 2020; Arduini et al., 2020), the earliest class of neural link prediction methods, which approximate an existing KG by exploiting the structural information contained in the facts

* Work conducted at Hasso-Plattner-Institute / University of Potsdam.

Findings of the Association for Computational Linguistics: EACL 2023, pages 1632–1648

of a KG (Ji et al., 2021). As KGs are incomplete and typically contain a large of number of entities that only appear in a few triples, more recent *text-based models* incorporate additional textual data sources for improved results (Ji et al., 2021). Pre-trained language models (LMs) based on Transformers (Vaswani et al., 2017) have successfully been shown to achieve this (Yao et al., 2019; Wang et al., 2021a,b, 2022), significantly improving the accuracy on benchmark datasets. However, models of this sort have not yet been investigated for bias.

We thus propose to conduct model-agnostic evaluations of link prediction models, enabling us to compare bias between representatives of KGE models and LM-based link prediction models, which we henceforth refer to as *KG only* and *KG+LM* models, respectively. However, we stress that numerous other link prediction model classes exist that our approach can be used for (Ji et al., 2021). Like previous work by Keidar et al. (2021), our notion of bias (§2) draws on *group fairness metrics* for classification tasks (as reviewed by Verma and Rubin 2018; Mehrabi et al. 2021). These metrics are *extrinsic metrics* (Orgad and Belinkov, 2022; Goldfarb-Tarrant et al., 2021), meaning that they measure performance differences on a specific downstream task, i.e., link prediction, for different social groups (e.g., gender).

Our paper makes the following **contributions**:

- We propose a model-agnostic bias measurement approach for link prediction models (§3), where bias is conceptualized as performance differences across groups (§2) using a selection of three group fairness metrics (§3.3).
- As previous papers have each used different datasets, we construct HUMANW5M-3MIL, a Wikidata subset of 3 million facts about humans, and make it publicly available (§4.1).
- We present experimental results comparing gender bias in occupation predictions between three KG only models and a KG+LM model, finding that the KG+LM model is more biased across the selected metrics (§4).
- We analyze our experimental results critically, analyzing the bias results at multiple levels of detail, and link the predictive bias to bias in the dataset (§4.4 + 4.5).

2 Bias Statement and Definitions

We follow Blodgett et al. (2020) who stress the importance of making the authors' understanding of *bias* explicit whenever it is investigated, following the taxonomy of harms (Barocas et al., 2017). Our understanding of *bias in link prediction* is based on the idea of *representational harm*, more specifically, "differences in system performance for different social groups" (Blodgett et al., 2020, p. 5456). For example, in the case of gender bias in occupation predictions, a link prediction model that predicts the occupations of women less accurately than those of men would be deemed as behaving harmfully. We consider this behavior as harmful, because deploying such a biased model in a downstream application can make it less useful for women than men. The extent of harm that is caused depends on the societal relevance of the application, e.g., it might be used for job applications or credit approval.

Beyond measuring bias in link prediction model predictions, we also investigate *data bias* in the knowledge graph (KG) datasets that we use. We consider the data to be biased if it is highly imbalanced with regard to some ideal distribution across social groups. An example of such an imbalance would be a dataset that contained significantly more facts about men than women and no facts about individuals with other gender identities.

The examples above also show that we measure bias in a specific context, defined by a *sensitive attribute* and a *target property*. A sensitive attribute is an inherent characteristic of an entity worthy of (legal or other) protection. Typical examples are gender, race, ethnicity, religion or worldview, disability, age, and sexual identity. Each sensitive attribute usually defines multiple *groups*, categorical options that a person can belong to, e.g., female gender. We view a target property as some notable property or achievement of an entity. Typical examples are their occupation, awards received, degrees, or where a person was educated. Both the target property and the sensitive attribute need to be expressed by one specific relation in the dataset. This means that the dataset-specific meaning of the respective relation matters: Here, we want to measure the influence of gender on the accuracy of occupation predictions. Following Keyes et al. (2021), we define *gender* as a multiplicitous concept expressing, e.g., identity and behaviors, going beyond bodily attributes that determine the biological *sex*

of a person. While we only discuss gender in the following, we note that for the gender identities "female" and "male" that we analyze, the distinction between gender and sex is not explicit in Wikidata, because the same entity is used to express both concepts.[1] In addition, due to data scarcity, we cannot analyze gender bias for other gender identities, such as "non-binary".[2] Due to a lack of reliable information, we do not distinguish between cisgender and transgender individuals when performing our analysis of gender bias for women and men. We refer to the HCI Gender Guidelines for further information about the terminology and concepts discussed above.[3]

3 A Model-Agnostic Approach for Measuring Bias in Link Prediction

3.1 Our Main Idea

Prior work on bias in link predictions has studied bias at the level of embeddings (Kraft and Usbeck, 2022). We instead propose to assess such bias in a model-agnostic manner by measuring bias directly on the test set predictions that each link prediction model produces. This allows us to measure bias on link prediction model classes that have not yet been evaluated with respect to bias. We focus on tail entity predictions for a given target relation, for example (Barack Obama, occupation, ?). The predictions serve as the input for group fairness metrics that measure extrinsic performance, meaning that we do not access any internal representations of the model. As group fairness metrics are usually defined for classification tasks (Verma and Rubin, 2018), we need to reframe the link prediction task accordingly (§3.2). Using these metrics allows us to investigate bias as a notion of metric-specific performance differences (§3.3). The choice of the sensitive attribute and target relation is largely subject to their prevalence in the dataset (§3.5). Due to limitations in the data (explained in Appendix B + C), our experiments focus on gender as a sensitive attribute using two gender identities and occupation as the target relation. Without loss of generality, we also use these as examples while explaining our method, but stress that our approach can be used for attributes and targets with more than two groups (§3.3).

[1] e.g. https://www.wikidata.org/wiki/Q6581072

[2] The reasons for this are further discussed in §3.5, §4.1, §6, §7, and Appendix B.

[3] https://www.morgan-klaus.com/gender-guidelines.html (Version 1.1)

3.2 Recasting Link Prediction as a Multi-Class Classification Task

Link prediction is typically defined as a ranking task, where each model produces continuous plausibility score values. For tail entity predictions – which we use for bias measurement – the model scores each possible entity in the dataset when given a combination of a head entity and a relation $(h, r, ?)$. A model that has learned the task well should thus emit high scores for the entities that are most plausible and the highest one for the entity that is the true tail entity. This enables us to reframe link prediction as a *multi-class classification* task, by defining each tail entity as a separate *class*. For example, when predicting occupation relationships, the set of all occupations in the dataset is the set of candidate tail entities, which can be viewed as class labels. A model is expected to predict true occupations, i.e., the *true label*, of a given person provided as the head entity. We define the tail entity that receives the highest plausibility score, i.e., rank 1, as the *predicted label*.

We note that this framing of the link prediction task best applies to one-to-one and many-to-one relations. For one-to-many and many-to-many relations, link prediction is technically a multi-label multi-class classification task. For a given human as head entity, there can be multiple true labels.[4] However, we cannot account for this in our approach as we want to avoid data leakage between the training/validation and test set: Our bias measurement is solely based on the test set, so the evaluation of our model predictions on the test set should be independent of the training set. We therefore only consider the occupations in the test set as true labels. We discuss a related aspect in regard to extracting information about our sensitive attribute in §3.5. This limitation applies to all sensitive attribute groups, so we assume that this does not influence the relative pattern in the bias scores between the groups, only their absolute values.

3.3 Selection of Fairness Metrics

Fairness metrics measure the performance for a given classifier, i.e., a link prediction model, using the following definitions: The predicted tail entity is denoted by $\hat{y}$, while the true one is indicated by y. The set of classes $\mathcal{Y}$ consists of candidate tail

[4] For instance, our target relation occupation is a many-to-many relationship, as a person can have multiple occupations and multiple persons can have the same occupation.

entities $t \in \{t_1, \ldots, t_{T-1}\}$ as well as the OTHER class t_T, where T is the total number of classes. We focus on sensitive attributes s with two values, $s \in \{0, 1\}$.[5] In the following, we introduce the three fairness metrics that we selected. The equations show how the *performance gap G* (Orgad and Belinkov, 2022) is measured for each target property class t, e.g., the absolute difference between the metric calculated for men versus women for a given occupation. We establish this notion of bias in our bias statement (§2).

Demographic Parity (DP) measures the *selection rate* (SelR), i.e., the probability of a given class to be predicted. For example, it answers the question "Which percentage of the persons predicted to be lawyers are men versus women?" It does not use information about the true class of the respective predictions (Verma and Rubin, 2018).

$$\begin{aligned}\text{DPG}(s,t) =& |P(\hat{y} = t \,|\, s = 1) - P(\hat{y} = t \,|\, s = 0)| \\ =& |\text{SelR}(s = 1, t) - \text{SelR}(s = 0, t)|\end{aligned} \tag{1}$$

This metric can show a potential general imbalance in the predictions. Such prediction imbalances may also reflect data bias, allowing us to analyze this connection. For using this metric in the context of binary classification and debiasing, we refer to work discussing the strengths and weaknesses of this metric (Dwork et al., 2012; Hardt et al., 2016).

Predictive Parity (PP) measures the positive predictive value (PPV), also known as *precision* (Prec), for a given class (Verma and Rubin, 2018; Chouldechova, 2017). Precision is a well-known evaluation metric that accounts for the percentage of correct predictions (true positives, TP) out of all persons predicted as belonging to that class, i.e., out of all true positives and false positives (FP). Achieving high precision for a class therefore means that if a classifier predicts a class, it is very likely that this prediction truly belongs to this class (Sokolova and Lapalme, 2009). The gap is:

$$\begin{aligned}\text{PPG}(s,t) =& |P(y = t \mid \hat{y} = t, s = 1) - \\ & P(y = t \mid \hat{y} = t, s = 0)| \\ =& |\text{Prec}(s = 1, t) - \text{Prec}(s = 0, t)|\end{aligned} \tag{2}$$

We choose this and the following metric because they are well-established in the algorithmic fairness community (Verma and Rubin, 2018; Hutchinson and Mitchell, 2019; Barocas et al., 2019), and each focuses on different capabilities of a classifier. For instance, the *precision–recall trade-off* implies a trade-off between Predictive Parity and Equality of Opportunity (Buckland and Gey, 1994). Also, an impossibility theorem from the algorithmic fairness community (Chouldechova, 2017) proves that these two notions of fairness cannot be achieved simultaneously in non-trivial scenarios.

Equality of Opportunity (EO) measures the true positive rate (TPR), also known as *recall* (Rec) (Hardt et al., 2016). For a given class, it measures the percentage of correct predictions (true positives) out of all persons that actually belong to that class. Achieving high recall for a class therefore means that the classifier identified most of the persons that truly belong to this class (Sokolova and Lapalme, 2009). The corresponding gap is:

$$\begin{aligned}\text{EOG}(s,t) =& |P(\hat{y} = t \mid y = t, s = 1) - \\ & P(\hat{y} = t \mid y = t, s = 0)| \\ =& |\text{Rec}(s = 1, t) - \text{Rec}(s = 0, t)|\end{aligned} \tag{3}$$

3.4 Analyzing Bias at Three Levels of Detail

In order to conduct a comprehensive and critical analysis, we calculate the above metrics at three levels of granularity. Each highlights a different aspect of model behavior: (i) the broadest one provides **one score per model**, (ii) a more detailed view yields **one score for each sensitive attribute group**, e.g., men vs. women, and (iii) the most detailed one provides **one score for each target property class and sensitive attribute group**, e.g., female lawyers. For (iii), we calculate the metric (selection rate, precision, or recall) for each individual target property class without averaging, e.g., $Rec(s = 1, t = 0)$. For (ii), we calculate the arithmetic mean for a specific sensitive attribute group, e.g., $s = 1$, across all T target property classes:

$$\text{EOG}(s = 1) = \frac{1}{T} \sum_{i=1}^{T} \text{Rec}(s = 1, t_i) \tag{4}$$

Calculating the average in this way means that we use *macro-averaging* assigning all classes equal importance (Sokolova and Lapalme, 2009). For (i), we invoke Equations 1–3 and average the results, again using macro-averaging:

$$\text{EOG}(s) = \frac{1}{T} \sum_{i=1}^{T} \text{EOG}(s, t_i) \tag{5}$$

[5] In theory, the approach can be extended to the non-binary case. This is shown by Keidar et al. (2021), however they do not discuss how the choice of their averaging strategy influences the interpretation of the bias score.

Table 1: Data for measuring link prediction bias on both datasets using **occupation** as the target property and **gender** as the sensitive attribute. Based on the prevalence in the test set of HUMANW5M-3MIL, we use a **minimum count threshold** of 100. This means that we consider all occupations with more than 100 occurrences as separate classes, aggregating the remaining facts in the class OTHER.

Occupation	in Test Set	with Gender	Thereof Men		Thereof Women	
other	1,534	399	336	84%	63	16%
politician	1,070	308	274	89%	34	11%
writer	262	69	58	84%	11	16%
lawyer	253	78	72	92%	6	8%
actor	158	47	34	72%	13	28%
association football player	142	32	31	97%	1	3%
poet	129	28	21	75%	7	25%
novelist	109	33	19	58%	14	42%
screenwriter	106	26	26	100%	0	0%
sum over all occupations	3,763	1,020	871	85%	149	15%

3.5 Data-Driven Choice of Target Property Classes and Sensitive Attribute

For all link prediction models, the long tail distribution typical for knowledge graph (KG) datasets (Zhang et al., 2020) presents a challenge: A small set of entities appears often, while most entities appear only a handful of times, even in large datasets. We account for this by choosing the target property classes, the sensitive attribute, and its groups based on their prevalence in the dataset. This means that each class needs to be properly represented for each sensitive attribute group, as it is also discussed in similar work in other domains (Seyyed-Kalantari et al., 2020; De-Arteaga et al., 2019). To achieve this, we reduce the number of classes significantly by **aggregating occupations below a minimum count threshold in the class OTHER**, similar to Keidar et al. (2021). The count threshold is **based on the test set of the dataset** since only this part is used for bias measurement. Using only the test set is necessary to avoid data leakage, as we directly use a model's predictions of the target property facts as input for our measurement:

Given a trained link prediction model, we extract only the facts concerned with our selected target property from the test set tail entity predictions, i.e., the (personXY, occupation, ?) facts. For each person – corresponding to the head entity – we then search the entire dataset for their sensitive attribute information, e.g., a fact stating their gender. We argue that retrieving the sensitive attribute information from the entire dataset is reasonable and does not constitute data leakage, since we only extract ground truth facts from the dataset. To be clear, **we never predict the sensitive attribute** of a person, only their target property. This means that the *data basis for the bias measurement* consists of persons with a **target property fact in the test set** and a **sensitive attribute fact somewhere in the dataset**.

While ensuring a sufficient data basis is necessary for a valid bias measurement, using a **minimum count threshold is also connected to the issues of data scarcity and data bias** (§2): (i) Facts about members of minority groups will naturally be less frequent than for those of the majority group. In addition, (ii) groups might be underrepresented due to biased selection processes in society that contributed to the creation of the data. In our case, the threshold leads to us only considering female and male as identities, while having to disregard other gender identities due to data scarcity and likely representation bias, as well. We argue that a bias analysis can still be performed under these circumstances, but that the **data basis and limitations should be clearly acknowledged**.

4 Experiments

4.1 Creating the HUMANW5M-3MIL Dataset

We created HUMANW5M-3MIL, a modified subset of Wikidata5M (Wang et al., 2021b) based on Wikidata (Vrandečić and Krötzsch, 2014), consisting of 3 million facts about humans, meaning that the head entity of each triple is always a human entity. For each entity in the dataset, a textual *description* consisting of the first section of the corresponding Wikipedia article in English is available as well as a short English *label* for each entity and relation (Wang et al., 2021b). We argue that a smaller dataset only consisting of human facts is useful to reduce the noise in the dataset and the time required to train and evaluate the models. This approach follows previous work (Bourli

Table 2: Prediction quality of all trained link prediction models on the test set measured using typical **accuracy metrics**. We report the metrics averaged over head and tail entity predictions and separately only for tail entity predictions. The best scores are highlighted in bold. The arrows express whether a high or a low value of the metric corresponds to high accuracy.

	Model	Prediction Type	MR ↓	MRR ↑	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑
KG only	TransE	averaged	188,784	19.21	16.02	20.74	24.47
		tail	11,620	38.29	32.04	41.29	48.63
	DistMult	averaged	176,300	15.62	11.14	18.42	22.44
		tail	8,405	30.76	22.00	36.34	44.09
	RotatE	averaged	221,341	14.80	11.44	17.08	19.50
		tail	19,552	29.56	22.86	34.12	38.92
KG+LM	$\text{SimKGC}_{\text{IB}}$	averaged	91,588	32.96	30.19	34.08	38.02
		tail	**255**	**64.79**	60.06	**67.04**	**73.60**
	$\text{SimKGC}_{\text{IB+SN+PB}}$	averaged	91,737	32.91	30.31	33.93	37.60
		tail	276	64.75	**60.14**	66.89	73.24

and Pitoura, 2020; Keidar et al., 2021), however the respective datasets are not publicly available. We also created this dataset due to issues we find in Wikidata5M: (i) The relation P21[6], which expresses human *sex or gender*, is not contained in the dataset, despite gender being the most frequently investigated sensitive attribute (Costa-jussà, 2019). (ii) An exploratory analysis revealed data quality issues in the entity labels such as typos or labels not matching the current English Wikidata labels. To address these issues, we merge the human facts of Wikidata5M with gender facts and English labels taken from a current Wikidata version (the truthy triples file from January 2, 2022). We ensure that each entity has a label and a description, meaning that we exclude entities that only have one or the other. For all remaining human entities, we extract the gender facts, if they exist. We limit our analysis to male and female gender, as data on non-binary gender identities and intersex people is very scarce in Wikidata (Klein et al., 2016; Zhang and Terveen, 2021)). In our case, other gender identities and intersex people are only represented by fewer than 500 occurrences combined. As the entities expressing human gender are not part of Wikidata5M and therefore lack a description, we use the first section of the Wikipedia articles for masculinity[7] and femininity[8]. The resulting dataset contains ca. 11 million triples, which we randomly sample down to 3,101,160 triples, to reduce the dataset size. The resulting dataset, HUMANW5M-3MIL, contains 1,396,220 unique entities – 1,269,907 thereof human – and 225 relations (Table 7). Table 8 shows that HUMANW5M-3MIL is representative of the larger raw dataset, when considering the manually selected candidate relations that express sensitive attributes or target properties. For instance, the *sex or gender* relation comprises ca. 13.5% of each dataset. We use comparably large evaluation sets, as our bias score calculation is only based on the test set, specifically a [0.9, 0.05, 0.05] train/validation/test random split (compared to [99.9995, 0.00025, 0.00025] for Wikidata5M), as the evaluation split size of ca. 155,000 triples is still manageable for all models we train on our dataset. Further details about the creation process of the dataset are given in Appendix B. We make the code for creating the dataset along with the data files available.[9]

4.2 Models and Training Details

We demonstrate our model-agnostic approach by comparing two model classes: knowledge graph embeddings (KGEs) that learn only from the structure contained in the knowledge graph dataset (*KG only*) and language model (LM)-based models that further also have access to the entity descriptions and relation labels (*KG+LM*).

KG only models: TransE, DistMult and RotatE. KGEs learn a dense embedding for each entity and relation in the dataset, capturing relationships between entities in a latent space (Nguyen, 2021). We choose TransE (Bordes et al., 2013) and DistMult (Yang et al., 2015) because they are common baseline models from different model families (Rossi et al., 2021a). RotatE (Sun et al., 2019) is an expressive state-of-the-art model from the same model class as TransE. We use the self-adversarial negative sampling loss (Sun et al., 2019) for all models. After hyperparameter tuning (Appendix A), we train all models for 400 epochs, using a

[6] https://www.wikidata.org/wiki/Property:P21
[7] https://en.wikipedia.org/wiki/Masculinity
[8] https://en.wikipedia.org/wiki/Femininity
[9] https://github.com/lena-schwert/comparing-bias-in-KG-models

Table 3: Bias in occupation predictions **averaged across all occupation classes**. The bias score correspond to performance gaps between predictions for men and women. The highest bias scores per fairness metric are highlighted in bold. DPG: Demographic Parity Gap, PPG: Predictive Parity Gap, EOG: Equality of Opportunity Gap. *: 3,763 occupation facts were available in total for HUMANW5M-3MIL.

Model Class	Model	DPG (Selection Rate)	PPG (Precision)	EOG (Recall)	# of Facts Used*
KG only	TransE	0.51	0.001	0.03	3,735
	DistMult	0.47	0.004	0.001	3,758
	RotatE	0.32	0.003	0.04	3,709
KG + LM	$\text{SimKGC}_{\text{IB}}$	**0.57**	**0.04**	0.08	3,726
	$\text{SimKGC}_{\text{IB+SN+PB}}$	0.54	0.02	**0.12**	3,721

Table 4: Link prediction bias results **separated for men and women** showing the absolute fairness metric scores. In some cases the absolute difference of the male and female score does not exactly match the *gap* scores in Table 3, because all results were rounded to two or three decimals. We highlight the entries with the highest difference in bold, i.e., the same entries as in Table 3. DP: Demographic Parity, PP: Predictive Parity, EO: Equality of Opportunity

Model Class	Model	DP (Selection Rate)		PP (Precision)		EO (Recall)	
		Male	Female	Male	Female	Male	Female
KG only	TransE	0.76	0.24	0.043	0.042	0.09	0.05
	DistMult	0.74	0.26	0.043	0.047	0.109	0.108
	RotatE	0.66	0.34	0.041	0.044	0.08	0.05
KG+LM	$\text{SimKGC}_{\text{IB}}$	**0.79**	**0.21**	**0.49**	**0.45**	0.32	0.41
	$\text{SimKGC}_{\text{IB+SN+PB}}$	0.77	0.23	0.51	0.49	**0.31**	**0.43**

Table 5: Deviation of predicted occupations for women from the data distribution using the KG+LM model $\text{SimKGC}_{\text{IB+SN+PB}}$. For each of the nine occupation classes, we calculate the difference between the selection rate and the distribution of the occupations in the test set of HUMANW5M-3MIL.

	Selection Rate	Data Distribution	Difference
averaged	0.23	0.15	+ 0.08
other	0.12	0.16	– 0.04
politician	0.15	0.11	+ 0.04
writer	0.25	0.16	+ 0.09
lawyer	0.12	0.08	+ 0.04
actor	0.33	0.28	+ 0.05
assoc. football player	0.08	0.03	+ 0.05
poet	0.40	0.25	+ 0.15
novelist	0.42	0.42	± 0.00
screenwriter	0.20	0.00	**+ 0.20**

batch size of 1,024, an embedding dimensionality of 512, and 32 negative samples per training triple. For TransE and DistMult a learning rate of 0.001 and for RotatE a learning rate of 0.01 is used.

KG+LM model: SimKGC LM-based models utilize pre-trained Transformers (Vaswani et al., 2017) that are fine-tuned on a knowledge graph dataset. To that end, an input sequence is created out of the entity descriptions instead of using the entity and relation IDs. We choose SimKGC (Wang et al., 2022), as it significantly outperforms earlier models with respect to accuracy and computational efficiency. It has a bi-encoder architecture using the pre-trained BERT-base (Devlin et al., 2019). One encoder learns relation-aware head entity embeddings and the other one tail entity embeddings. The plausibility scoring of triples is then simply achieved using cosine similarity. We train the $\text{SimKGC}_{\text{IB}}$ and the $\text{SimKGC}_{\text{IB+SN+PB}}$ model variants to investigate whether they exhibit different bias behavior. We do not conduct hyperparameter tuning, as the parameters for Wikidata5M used in the original paper (Wang et al., 2022) deliver strong results on our validation set. SimKGC uses the InfoNCE loss with an additive margin (Le-Khac et al., 2020). We train for 1 epoch using a batch size of 1,024, a learning rate of 3×10^{-5} and a weight decay of 0.0001. We provide further details for reproducing the experiments in Appendix A.

4.3 Evaluation Protocol

We evaluate our link prediction models for accuracy using mean rank (MR), mean reciprocal rank (MRR) as well as Hits@1, Hits@3, and Hits@10 (Rossi et al., 2021a). We calculate the ranks using the *filtered setting* (Bordes et al., 2013). Since we only use tail entity predictions for measuring bias, we compute the metrics (i) averaged across head and tail entity predictions and separately (ii) only for tail entity predictions. Following §3.5, we choose gender as a sensitive attribute and occupation as the target property for measuring bias in the trained models. We describe in Appendix C how other combinations of sensitive attributes and target property are not analyzed due to data scarcity.

4.4 Model Accuracy and Data Bias Results

Referring to Table 2, we note that the Hits@1 accuracy for tail entity predictions is the most relevant metric for bias measurement, since the tail entities with rank = 1 are used as the predicted class labels, i.e., the predicted occupation. The performance on tail entity predictions is clearly higher than the one averaged across head and tail entity predictions since there are fewer unique tail than head entities, making this prediction easier. Performance on tail entity predictions varies between 11.14 (DistMult) and 60.14 (SimKGC$_{\text{IB+SN+PB}}$). When comparing the two model classes, the KG+LM models clearly outperform the KG only models. Among the KG only models, TransE obtains the best Hits@1 result (16.02), thus outperforming the two other more recent and complex models.

Table 1 shows the absolute counts and the relative distributions of the occupation classes over the two considered gender identities (male, female). It also shows that we choose a *minimum count threshold* of 100 facts per occupation, resulting in eight distinct occupations, aggregating the remaining ones in the class OTHER. When comparing the relative distribution of facts per gender, it is evident that the data is biased: Out of the 1,020 facts that we use for bias measurement, 85% are about men and only 15% about women, while a 50%–50% distribution would be unbiased when considering these two gender identities. The occupation with the largest gender bias in the data is *screenwriter* (100% men) and the one with the smallest bias is *novelist* (58% men, 42% women). In addition, we note again that gender identities beyond women and men are severely underrepresented in the data, constituting only 0.005% of the gender facts, which is significantly lower than the 0.1–2% estimated by Goodman et al. (2019).

4.5 Results on Gender Bias in Occupation Predictions

For describing and analyzing the gender bias that our models exhibit in its occupation predictions, we consider the *three levels of detail* as introduced in §3.4. Tables 3, 4, and 5 refer to levels of detail (i), (ii), and (iii), respectively. These allow us to answer three different research questions.

Q1: Are KG+LM models more biased than KG only models?
As Table 3 shows, the bias scores are generally higher for the KG+LM models than for KG only models. Comparing the difference between the most biased models for each class shows that it is most pronounced for the demographic parity gap (DPG): $0.57 - 0.47 = 0.1$, followed by the equality of opportunity gap (EOG) $0.12 - 0.04 = 0.08$. These results suggest that the additional textual data the KG+LM models have access to leads to biased occupation predictions and that this has the most pronounced effect on DPG and EOG. The KG only models, in contrast, here manage to obtain fairly unbiased results, with scores close to zero. We note that the column "# of facts used" shows how many facts contributed to the score, since a fact can only be considered when the predicted tail entity is an occupation and not another type of entity.

Q2: Does the bias originate in higher-quality predictions for men or women?
To answer this question, we refer to Table 4, which shows the previously described results separately for men and women. For DP, we observe that the selection rate for predictions for men is generally higher. We connect this to the data distribution in Q3. For PP and EO we make two observations: First, most KG only models – which obtain essentially unbiased results – obtain dismal precision and recall scores (they only appear strong enough when evaluated using ranking metrics). Second, for the KG+LM models we observe opposing trends: With regard to precision, the prediction quality for men is slightly higher, while for recall, the prediction quality for women is noticeably higher. Especially the latter trend is surprising since the data for women is more limited. These observations show why this level of detail is important for a comprehensive

bias analysis: While KG only models exhibit far less bias, they predict occupations inaccurately despite an acceptable overall accuracy (Table 2). In addition, we conclude that predictions for men are not necessarily more accurate than those for the women, despite the significantly larger amount of data for men (85% of all occupation facts.).

Q3: Are there occupation classes that are predicted more often than expected based on their distribution in the data?
We may consider the demographic parity results for SimKGC$_{\text{IB+SN+PB}}$, our most accurate model, as an example. As explained earlier (§3.3), DP measures selection rate imbalances that we expect to mirror the data bias. Table 5 shows the per-class differences between the selection rate (predicted occupation) and the respective distribution in the data (actual occupation) when predicting the occupation of women. Whenever the difference is positive, the model predicts the given occupation for more women than expected (and vice versa). On average, the probability of women having a given occupation is overestimated by 0.08, with the occupations "poet" and "screenwriter" contributing the most to this score. Despite this, the model does predict this occupation for some women, as female screenwriters do exist in the training dataset. This might be due to the entity descriptions that this KG+LM model has access to, potentially because the person's occupation might be similar to a screenwriter or mention related words.

5 Conclusion

We present a model-agnostic approach for measuring bias in link prediction along with the first experimental study that measures bias in language model (LM)-based link prediction models (KG+LM), comparing it with bias in knowledge graph embedding (KGE) models (KG only). Using a selection of fairness metrics and analyzing our results at three levels of detail, we find that the KG+LM models are more biased. We discuss the relationship between bias, link prediction accuracy metrics and data bias. For our experiments, we create HUMANW5M-3MIL, a subset of 3 million facts about humans contained in Wikidata (Vrandečić and Krötzsch, 2014). We have made our code and the dataset available to the public to encourage further research on these topics.

6 Limitations

In the following we discuss the limitations of our work and how they might be addressed.

Our study considers a single sensitive attribute, gender, limited to two gender identities, female and male. We also note that the approach can be extended to sensitive attributes with more than two groups, requiring additional decisions on how to average the bias scores across the sensitive attribute groups in an interpretable way. This limitation is caused by data scarcity, as we describe in §3.5, §4.1 and Appendix B + C.

Using only the test set of a dataset for bias measurement has a few methodological implications: First, the bias in the test set might not be representative of the bias contained in the other splits of the data set. In our approach we used simple random splitting, where all facts are randomly distributed over the three splits, meaning that the distribution of the relations might not be the same in all splits. This approach is called the *transductive setting*, which is currently the most prevalent method of splitting knowledge graph datasets (Wang et al., 2021b). To rule out differences between the splits to some degree, a potential solution is a stratified split, conditioned on the relations in the dataset. This would enforce, for instance, that each split has the same relative amount of gender and occupation facts. This solution is however only applicable when the researcher has control over the dataset split creation process. Second, in order to have a sufficient data basis for each target property class across sensitive attribute groups (§3.5), the dataset or the test split size needs to be quite large. However, training models on large datasets requires the availability of adequate computational resources that many researchers do not have access to.

Using fairness metrics for bias measurement means that the notion of bias is closely connected to what is considered as a "misclassification". We note that we do not take into account the severity of misclassifications, e.g., that predicting a novelist to be a writer is less wrong than predicting them to be a diplomat. This would require a semantic analysis of the labels of both the true and the predicted tail entities. This might also be addressed by clustering entities with similar meanings together, e.g., predicting groups of occupations instead of single occupations.

7 Ethical Considerations

For our analysis of gender bias, we rely on factual statements contained in Wikidata[10], a crowdsourced, public knowledge graph. This means that we utilize gender information that was added to the platform by largely anonymous editors. These statements – and other statements describing demographics – might therefore not correspond to the self-identification of the respective persons or they might be incorrect, especially if human or automated data quality control mechanisms fail (Heindorf et al., 2019).

In addition, we acknowledge that knowledge graphs reflect a limited world view, because their creation process is subject to various biases, such as representation bias, popularity bias, and sampling bias (following the definitions by Mehrabi et al. 2021). In the field of knowledge graphs, these problems were first described by Janowicz et al. (2018) and recently reviewed by Kraft and Usbeck (2022). For example, facts about the non-Western world are underrepresented and persons with occupations in arts, sports, and science and technology are overrepresented (Radstok et al., 2021; Beytía et al., 2022).

One consequence of the biases mentioned above is our decision to only consider male and female gender in our analysis, as all other gender identities combined, such as non-binary, amount to fewer than 500 facts in the entire dataset. To analyze bias for these gender identities, a larger dataset or a different approach than ours would be necessary. We discuss these limitations and our understanding of gender in §2.

As described in our bias statement (§2), our notion of bias focuses on performance differences for different social groups. We note that this is a very specific, limited conceptualization of bias that could be extended by considering real-world distributions or normative connotations such as stereotypes. However, we believe that the contribution of our work is still useful for analyzing whether link prediction models work as intended, especially because it allows for comparing different model classes.

To conclude, we stress that the intended use of our approach is to identify concerning model behavior in a specific context defined by a sensitive attribute and a target property. We emphasize that the selected fairness metrics should not, e.g., be used as constraints during model training without a deeper analysis of what notions of fairness are suitable in the context of how the model will be used.

[10] https://www.wikidata.org

Acknowledgements

We thank Lisa Gotzian and Steffen Berhorst for discussions about the bias measurement approach and for feedback on the manuscript. We further thank the three anonymous reviewers for their thoughtful comments.

References

Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. 2021. PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research*, 22(82):1–6.

Mario Arduini, Lorenzo Noci, Federico Pirovano, Ce Zhang, Yash Raj Shrestha, and Bibek Paudel. 2020. Adversarial learning for debiasing knowledge graph embeddings. *MLG 2020: 16th International Workshop on Mining and Learning with Graphs - A Workshop at the KDD Conference*.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. *SIGCIS Conference*.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Pablo Beytía, Pushkal Agarwal, Miriam Redi, and Vivek K Singh. 2022. Visual gender biases in Wikipedia: A systematic evaluation across the ten most spoken languages. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:43–54.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in ai. Technical Report MSR-TR-2020-32, Microsoft.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5454–5476.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*.

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

Michael Buckland and Fredric Gey. 1994. The relationship between recall and precision. *Journal of the American Society for Information Science*, 45:12–19.

Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5:153–163.

Marta R. Costa-jussà. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1:495–496.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional Transformers for language understanding. *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)*, 1:4171–4186.

Yupei Du, Qi Zheng, Yuanbin Wu, Man Lan, Yan Yang, and Meirong Ma. 2022. Understanding gender bias in knowledge base embeddings. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:1381–1395.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, pages 214–226.

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

Joseph Fisher, Dave Palfrey, Christos Christodoulopoulos, and Arpit Mittal. 2020b. Measuring social bias in knowledge graph embeddings. *Proceedings of the Knowledge-Graph Bias Workshop*, page 7332–734.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1926–1940.

Michael Goodman, Noah Adams, Trevor Corneil, Baudewijntje Kreukels, Joz Motmans, and Eli Coleman. 2019. Size and distribution of transgender and gender nonconforming populations: A narrative review. *Endocrinology and Metabolism Clinics of North America*, 48(2):303–321. Transgender Medicine.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems (NIPS)*, pages 1–22.

Stefan Heindorf, Yan Scholten, Gregor Engels, and Martin Potthast. 2019. Debiasing vandalism detection models at Wikidata. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 2:670–680.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs. *Synthesis Lectures on Data, Semantics, and Knowledge, No. 22*.

Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un)fairness: Lessons for machine learning. *Proceedings of FAT* 2019: Conference on Fairness, Accountability and Transparency*, pages 49–58.

Krzysztof Janowicz, Bo Yan, Blake Regalia, Rui Zhu, and Gengchen Mai. 2018. Debiasing knowledge graphs: Why female presidents are not like female popes. *CEUR Workshop Proceedings*, 2180:1–5.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*.

Daphna Keidar, Mian Zhong, Ce Zhang, Yash Raj Shrestha, and Bibek Paudel. 2021. Towards automatic bias detection in knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3804–3811, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Os Keyes, Chandler May, and Annabelle Carrell. 2021. You keep using that word: Ways of thinking about gender in computing research. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

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.

Angelie Kraft and Ricardo Usbeck. 2022. The Lifecycle of "Facts": A Survey of Social Bias in Knowledge Graphs. *arXiv*.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54.

Dat Quoc Nguyen. 2021. A survey of embedding models of entities and relationships for knowledge graph completion. *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 1–14.

Hadas Orgad and Yonatan Belinkov. 2022. Choose your lenses: Flaws in gender bias evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 151–167, Seattle, Washington. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Heiko Paulheim. 2016. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8:489–508.

Wessel Radstok, Melisachew Wudage Chekol, and Mirko Tobias Schafer. 2021. Are knowledge graph embedding models biased, or is it the data that they are trained on? *CEUR Workshop Proceedings*, 2982.

Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. 2021a. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data*, 15:1–49.

Andrea Rossi, Donatella Firmani, and Paolo Merialdo. 2021b. Knowledge graph embeddings or bias graph embeddings? A study of bias in link prediction models. *DL4KG 2021: Workshop on Deep Learning for Knowledge Graphs, held as part of ISWC 2021: The 20th International Semantic Web Conference*.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew Mcdermott, Irene Y Chen, and Marzyeh Ghassemi. 2020. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *Pacific Symposium On Biocomputing*, 26:232–242.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45:427–437.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge graph embedding by relational rotation in complex space. *Proceedings of the International Conference on Learning -Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pages 6000–6010.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. *ACM/IEEE International Workshop on Software Fairness Fairness (FairWare)*, pages 1–7.

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

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, Ying Wang, and Yi Chang. 2021a. Structure-augmented text representation learning for efficient knowledge graph completion. *Proceedings of the Web Conference 2021 (WWW '21)*, pages 1737–1748.

Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022. SimKGC: Simple contrastive knowledge graph completion with pre-trained language models. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4281–4294.

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

Gerhard Weikum, Luna Xin Dong, Simon Razniewski, and Fabian Suchanek. 2021. Machine knowledge: Creation and curation of comprehensive knowledge bases. *arXiv preprint*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. *arXiv preprint*.

Charles Chuankai Zhang and Loren Terveen. 2021. Quantifying the gap: A case study of Wikidata gender disparities. In *Proceedings of the 17th International Symposium on Open Collaboration*, OpenSym '21, New York, NY, USA. Association for Computing Machinery.

Chuxu Zhang, Huaxiu Yao, Chao Huang, Meng Jiang, Zhenhui Li, and Nitesh V. Chawla. 2020. Few-shot knowledge graph completion. *Proceedings of the AAAI Conference on Artificial Intelligence - AAAI Technical Track: Knowledge Representation and Reasoning*, pages 3041–3048.

A Additional Details for Reproducing the Experiments

We list the training and evaluation time as well as the hardware used for all our models in Table 6.

KG only: TransE, DistMult, RotatE. For training the KG only models on HUMANW5M-3MIL, we largely use the parameters contained in the Graphvite configuration files from the official Wikidata5M benchmark[11] published by Wang et al. (2021b). To adapt to our dataset, we conduct minor additional hyperparameter tuning. We explore a small grid testing 6 hyperparameter combinations for each model: batch size $\in \{512, 1024\}$, learning rate $\in \{0.1, 0.01, 0.001\}$. We train TransE and DistMult for 50 epochs (training time: ca. 3.5-7 h, evaluation time: ca. 35 min) and RotatE for 20 epochs (training time: ca. 2.5-5 h, evaluation time: ca. 65 min). We choose our final parameter configuration based on the mean reciprocal rank (MRR) on the validation set, as it has been observed to be the most stable metric among the link prediction metrics (Rossi et al., 2021a).

For all models, we use a batch size of 1,024, an embedding dimensionality of 512, and 32 negative samples per training triple. For TransE and DistMult a learning rate of 0.001, for RotatE a learning rate of 0.01 is used. We train the models for 500 epochs, evaluating after each 100 epochs. Finally, the models trained for 400 epochs are used, since the MRR performance drops slightly afterwards. We use the self-adversarial negative sampling loss (Sun et al., 2019) for all models.. For TransE, we use margin γ = 12 and adversarial temperature = 0.5. For DistMult, we use margin γ = 0 and adversarial temperature = 2. Again following the Graphvite configuration files, we also apply L3 regularization with a weight of 0.002. For RotatE, we use margin γ = 6 and adversarial temperature = 0.2.

KG + LM: SimKGC. We use the pre-trained BERT-base in its "uncased" variant (Devlin et al., 2019). Since the authors trained their model on Wikidata5M, a superset of our dataset, we try using the exact same parameters as the original paper (Wang et al., 2022). We use two of their model variants to investigate whether using the self-negative (SN) and pre-batch (PB) sample types lead to different bias behavior compared to the "basic" in-batch (IB) model variant. We therefore train the $\text{SimKGC}_{\text{IB}}$ and the $\text{SimKGC}_{\text{IB+SN+PB}}$ model variants, using 2 pre-batch negatives for the latter. We train for 1 epoch using a batch size of 1,024, a learning rate of 3×10^{-5} and a weight decay of 0.0001. The remaining parameters are: 400 warmup steps for the linear learning rate scheduler, gradient clipping of 10.0, dropout 0.1, temperature τ is initialized with 0.05, additive margin γ for the InfoNCE loss is 0.02, $\alpha = 0.05$ for graph-based re-ranking is used, 2-hop neighbors are considered, and a maximal token length of 50 for the entity descriptions is used. As these parameters deliver good results on the validation set, we do not conduct hyperparameter tuning.

Implementation details. All implementations are done in Python. The code and data including files to re-create the conda environment are contained in the accompanying GitHub repository[12]. All models are based on the deep-learning framework PyTorch (Paszke et al., 2019).

- **KG only: Knowledge Graph Embeddings:** For training models on HUMANW5M-3MIL we use the model implementations and the training pipeline of the v1.8.1 PyKEEN library (Ali et al., 2021). This framework enables single-GPU training and the calculation of evaluation metrics.
- **KG + LM: SimKGC:** We use the implementation that was published alongside the paper of Wang et al. (2022). Their code[13] includes the calculation of evaluation metrics. The implementations use the Huggingface Transformers library v4.15 (Wolf et al., 2020).
- **Data Bias:** We use our own Python implementation.
- **Link Prediction Bias:** For calculating the predictive parity and equality of opportunity, we use Microsoft's fairlearn library (Bird et al., 2020), that wraps around scikit-learn's evaluation metrics. For calculating demographic parity, we modify code from the repository published by Keidar et al. (2021).[14]

B Additional Details About the Creation of the Dataset

This section describes how we create the raw version of HUMANW5M-3MIL. This raw version

[11] https://graphvite.io/docs/latest/benchmark.html

[12] https://github.com/lena-schwert/comparing-bias-in-KG-models

[13] https://github.com/intfloat/SimKGC/

[14] https://github.com/mianzg/kgbiasdetec

Table 6: Training runtime, evaluation runtime and hardware used for training all of our models. *: NVIDIA A100-SXM-80GB, $^{+}$: AMD EPYC 7502 32-Core CPU.

Model Class	Model	Train. Time	Eval. Time	GPU(s) Used	Other Hardware
KG only	TransE	27 h	35 min	1x NVIDIA A100*	10 GB RAM, 32 CPUs^{+}
	DistMult	30 h	35 min	1x NVIDIA A100*	10 GB RAM, 32 CPUs^{+}
	RotatE	55 h	66 min	1x NVIDIA A100*	10 GB RAM, 32 CPUs^{+}
KG+LM	SimKGC$_{IB}$	45 min	180 min	4x NVIDIA A100*	15 GB RAM, 50 CPUs^{+}
	SimKGC$_{IB+SN+PB}$	45 min	180 min	4x NVIDIA A100*	15 GB RAM, 50 CPUs^{+}

Table 7: Dataset statistics for the raw version of our dataset and the subsampled dataset HUMANW5M-3MIL that we use in our experiments.

	11mil. Raw Dataset	HUMANW5M-3MIL
# of entities	1,732,021	1,396,220
# of human entities	1,503,491	1,269,907
# of relations	292	225
# of train triples	-	2,791,044
# of validation triples	-	155,058
# of test triples	-	155,058
total # of triples	11,114,797	3,101,160

Table 8: Manually selected Wikidata relations of general interest for a bias analysis. This large selection can be considered as candidate relations, since they must exist in sufficient quantity to enable a robust bias analysis. We show the triple counts for each relation and the proportion of this count of the total size of each dataset. The raw dataset version contains 11,114,797 triples. Our final dataset, HUMANW5M-3MIL contains 3,101,160 triples. We ultimately only use the relations "gender" and "occupation" for our bias analysis.

Wikidata Label	Wikidata ID	Relation Expresses...	11mil. Raw Dataset		HUMANW5M-3MIL	
sex or gender	P21	gender	1,501,938	(13.51 %)	418,622	(13.50 %)
country of citizenship	P27	nationality	1,143,007	(10.28 %)	319,123	(10.29 %)
place of birth	P19	nationality	854,080	(7.68 %)	238,162	(7.68 %)
religion	P140	religion	27,805	(0.25 %)	7,827	(0.25 %)
ethnic group	P172	ethnicity	27,235	(0.25 %)	7,751	(0.25 %)
native language	P103	nationality	19,771	(0.18 %)	5,495	(0.18 %)
medical condition	P1050	disability	3,824	(0.03 %)	1,049	(0.03 %)
sexual orientation	P91	sexual orientation	484	(0.004 %)	123	(0.004 %)
occupation	P106	target property	1,095,357	(9.85 %)	305,806	(9.86 %)
educated at	P69	target property	438,207	(3.94 %)	122,195	(3.94 %)
award received	P166	target property	169,758	(1.53 %)	47,661	(1.54 %)
member of political party	P102	target property	126,285	(1.14 %)	35,233	(1.14 %)
employer	P108	target property	79,781	(0.72 %)	22,103	(0.71 %)
position held	P39	target property	75,909	(0.68 %)	21,069	(0.68 %)
field of work	P101	target property	17,757	(0.16 %)	4,956	(0.16 %)
military rank	P410	target property	16,330	(0.15 %)	4,510	(0.15 %)
nominated for	P1411	target property	12,854	(0.12 %)	3,627	(0.12 %)
academic degree	P512	target property	5,315	(0.05 %)	1,558	(0.05 %)
doctoral student	P185	target property	1,415	(0.01 %)	424	(0.01 %)

contains 11,114,797 triples, 1,732,021 entities – thereof 1,503,491 human entities – and 292 relations. To reduce dataset size, we sample it down to 3,101,160 triples, creating HUMANW5M-3MIL.

Details on extracting the labels. As an alternative to using the textual descriptions for entities, i.e., the first section of the corresponding Wikipedia article, we propose using the shorter Wikidata labels. As these contain less information for the KG+LM model to process, using labels instead of descriptions reduces the model runtime. We considered using the alias files provided with Wikidata5M[15], but found that the entity aliases have quality issues such as typos (e.g., for the 'human' entity Q5, the first alias is 'Huamn') or aliases that do not have the same meaning as the current label (e.g., for the 'universe' entity Q1, the first alias is 'Earth's universe'). After correspondence with the first author of the

[15] https://deepgraphlearning.github.io/project/wikidata5m

paper that introduced Wikidata5M (Wang et al., 2021b), we learned that they created the alias files using the "pageterms" property of the MediaWiki API[16]. The faulty aliases are thus likely a result of the use of that data source and do not represent genuine entity labels. We therefore extract entity and relation labels from the January 2, 2022 truthy triples Wikidata dump[17]. The Wikidata dump files are updated every week and contain the most recent state of Wikidata in different data formats[18]. We use the truthy triples file specifically, because it only contains non-deprecated triples, which reduces the amount of metadata contained and therefore the file size.

Details on creating the subset of human facts. We extract all facts that have a human head entity from the raw triples file provided with the original Wikidata5M files. A human head entity is identified by its "instance of human" (QXX P31 Q5) statement. This initially leads to a subset of 9,804,421 facts about humans. In order to be able to compare KG+LM models using either the shorter labels or the longer descriptions as text input, we only keep entities and relations that have both labels and descriptions. This leads to a removal of 17,528 entities and 2 relations from the dataset, due to deletions and additions that happened between the creation of Wikidata5M (based on the July 2019 Wikidata dump, Wang et al. 2021b) and the extraction of the labels (January 2022 Wikidata truthy triples dump). Removing these entities and relations means that we remove all facts that contain them, leading to 191,178 facts that are excluded in total.

Details on adding gender facts from current Wikidata. In an exploratory analysis of Wikidata5M before creating our dataset, we counted the occurrences of facts that we considered to be of general interest for a bias analysis. With respect to the gender relation (PID: P21) we found that its count is unexpectedly low (about 4,000) compared to the number of human entities in the dataset (1.5 million). Furthermore, we found that these facts express animal sex and not human gender, because the head and tail entities are non-human (tail entities: Q44148, Q43445). When filtering for gender facts in the human facts subset, we only found 384 facts overall. Through correspondence, the first author of the paper that introduced Wikidata5M (Wang et al., 2021b) informed us that they used Wikidata's "wbgetentities" API[19] to align Wikidata and Wikipedia entries. Since the Wikidata entities for male[20] and female gender[21] are linked to the same Wikipedia page describing gender[22], the API might have therefore omitted facts containing these entities. We therefore use the January 2022 truthy triples dump to extract the gender facts as well. We extract 1,243,734 facts with gender *male* and 258,204 facts with gender *female*. Persons with other gender identities, such as *non-binary*, or intersex people have fewer than 500 occurrences in the entire dataset.

We therefore consider only two gender identities within the context of this study, as the data scarcity would not allow our models to properly represent the other gender identities contained in the dataset.

Adding the gender facts for women and men entails adding two new (tail) entities to the dataset (Q6581072, Q6581097). As these entities do not have descriptions in the original dataset, we use the first section of the Wikipedia articles for masculinity[23] and femininity[24].

C Considering Additional Sensitive Attributes and Target Properties

Beyond measuring gender bias in occupation prediction, we did consider using other target properties and sensitive attributes for the analysis of the HUMANW5M-3MIL subset. However – in contrast to using "gender" and "occupation" – we found the respective data bases to be lacking.

The relation "educated at" is the target property with the second-highest counts in HumanWikidata5M. In total, the 438,207 facts have 9,330 different tail entities, i.e., educational institutions such as universities. In the test set, the 8,684 "educated at" facts still have 1,919 different tail entities, only 3 of those with more than 100 occurrences. If the minimum count threshold were set at 100, this would result in an "other" class with 8,439 facts, leading to a very imbalanced class distribution. In

[16] https://www.mediawiki.org/w/api.php?action=help&modules=query%2Bpageterms

[17] https://dumps.wikimedia.org/wikidatawiki/entities/

[18] https://www.wikidata.org/wiki/Wikidata:Database_download

[19] https://www.mediawiki.org/wiki/Wikibase/API

[20] https://www.wikidata.org/wiki/Q6581097

[21] https://www.wikidata.org/wiki/Q6581072

[22] https://en.wikipedia.org/wiki/Gender

[23] https://en.wikipedia.org/wiki/Masculinity

[24] https://en.wikipedia.org/wiki/Femininity

Table 9: Data basis for measuring link prediction bias using **occupation** as the target property and **country of citizenship** as the sensitive attribute. This shows the insufficient data basis for a bias analysis: Even the three best represented countries of citizenship (sum over all occupations ≥ 50) are not sufficiently represented across the individual occupations in the test set of HUMANW5M-3MIL.

Occupation	in Test Set	with Citizenship	USA	France	UK	Other
other	1,534	437	146	33	41	217
politician	1,070	278	104	11	9	154
writer	262	68	9	13	4	42
lawyer	253	76	50	1	1	24
actor	158	44	18	5	2	19
association football player	142	34	2	1	10	21
poet	129	38	2	8	2	26
novelist	109	29	17	3	3	6
screenwriter	106	38	9	5	2	22
Sum over all occupations	3,763	1,042	357	80	74	531

addition, the three most frequent tail entities are “Harvard University" (270 facts), “Yale University" (121 facts), and “University of Michigan" (104 facts), which represent a very limited selection of all educational institutions contained in the dataset. We therefore disregard “educated at" as a target property.

Moving on to additional potential sensitive attributes, the relation “country of citizenship" is the most promising candidate with 1,143,007 facts in HumanWikidata5M. However, when creating an overview of counts per occupation class similar to Table 1, it becomes evident that the data for each sensitive attribute group, i.e., country, is very limited (Table 9). While there are in total 18,396 country of citizenship facts in the test set, this information is only available for 1,020 of the 3,763 occupation facts. The three countries with the highest counts are all Western countries, namely USA (357 facts), France (80 facts), and the UK (74 facts). Even for these countries, the majority of the occupation classes are only represented by 0 to 5 facts. The 110 other countries represented in the test set are all aggregated in the “other" class, which is again the largest class with 531 facts. This means that the sensitive attribute groups are already quite homogeneous, while the “other" group contains the majority of diverse information about “citizenship". Compared to using two groups for “gender” as the sensitive attribute, choosing the sensitive attribute groups as above would thus result in an unrealistic and uninformative comparison. We hence decided against including “country of citizenship" as a sensitive attribute.

Similar considerations apply to the other relations of interest listed in Table 8, since these also have too few facts per target property class or a very broad distribution over sensitive attribute groups.