Are knowledge graph embedding models biased, or is it the data that they are trained on?

Wessel Radstok[1], Melisachew Wudage Chekol[1], and Mirko Tobias Schäfer[2]

[1] Data Intensive Systems Group, Utrecht University
[2] Department of Media and Culture Studies, Utrecht University

Abstract. Recent studies on bias analysis of knowledge graph (KG) embedding models focus primarily on altering the models such that sensitive features are dealt with differently from other features. The underlying implication is that the models cause bias, or that it is their task to solve it. In this paper we argue that the problem is not caused by the models but by the data, and that it is the responsibility of the expert to ensure that the data is representative for the intended goal. To support this claim, we experiment with two different knowledge graphs and show that the bias is not only present in the models, but also in the data. Next, we show that by adding new samples to balance the distribution of facts with regards to specific sensitive features, we can reduce the bias in the models.[3]

1 Introduction

For several days in early July 2018, Google and Apple's search assistants wrongfully reported that the man behind the Marvel comic books, Stan Lee, had passed away .[4] It did not take long for news articles to start popping up noting the unjustified death declaration. Although Google and Apple never officially reported on this issue, its source is likely traced back to Wikidata. On June 27th, a Wikidata user ran their own script made to parse data from Wikipedia and insert them as claims into Wikidata. This script then mistakenly pronounced Stan Lee dead. Other users soon corrected the error, which resulted in an edit war that became so bad that the page had to be temporarily locked against vandalism.

This is not the only occurrence of incorrect information in knowledge graphs causing issues in downstream search queries. In the second half of 2018, the former Guantanamo Bay detainee Omar Khadr was incorrectly reported by Google search for the query 'Canadian Soldiers'. Again the cause was the script written by the aforementioned user. Although the issue was quickly resolved after online outrage, it cropped up twice more over a period of several months. It eventually led Google to take manual action to fix the knowledge graph.[5]

[4] https://www.cinemablend.com/news/2444550/siri-is-telling-people-stan-lee-died-yesterday

[5] https://www.cbc.ca/news/technology/omar-khadr-google-search-knowledge-graph-scheer-russia-1.4999775

In addition to the presence of incorrect information in knowledge graphs due to either an error in the KG construction or intentionally supplied by content curators, KGs can also be incomplete. As an example, in Freebase [2], over 70% of person entities have no known place of birth and over 75% have no known nationality [8]. In Wikidata [14], we observe a similar behavior, for instance, over 97% of humans have no known religion and over 83% of humans have no known spoken, written or signed languages. Subsets of both Wikidata and Freebase have been widely used for testing knowledge graph completion models. However, these subsets do not take into account the incompleteness of the KGs and are prepared in a way to test solely the accuracy of models. However, if the subsets are incomplete (or unbalanced), the models can be biased. For instance, the Wikidata12K [11, 6] dataset contains 80% male and 20% female politicians. Clearly, this dataset is unbalanced and a model trained on it will likely overrepresent men in its predictions.

Indeed this is shown in our experiments using the TransE [3] model; when asked to predict people most likely to be politicians, the top 100 ranked answers contain just 12.4% while the remaining 83.6% are male. In order to mitigate such biased predictions, recently there has been a growing effort towards adapting/extending KG completion models [4, 9, 10]. These studies on bias analysis of KG embedding models focus primarily on altering the models such that sensitive features (such as gender, sexual orientation, etc) are dealt with differently from other features. The underlying implication is that the models cause bias, or that it is their task to solve it. However, we found out that the datasets on which the models are trained on are biased/unbalanced. Although algorithms for the automatic balancing of data do exist [5], these are not trivial to apply to graph datasets. Our experiments showed unsatisfactory results using these methods.

Furthermore, adapting models to remove bias means that the resulting embeddings will be bias-neutral with regards to the strength of the model used. That is, removing bias requires a bias detection model and the extent of the the bias removed depends on how much bias is detected. As a result, embeddings are not truly neutral: a more powerful model might still be able to detect biases. Therefore we argue that a domain expert must remain in the loop.

In this work, we address the problem by working directly on the data rather than altering KG embedding models. In other words, we investigate a new approach in order to balance (mitigate bias) a given dataset: we automatically extend a dataset by extracting additional facts to complete missing values of sensitive features. Moreover, so as to motivate the proposed approach, we carried out a comprehensive analysis of the distribution of sensitive features in Wikidata highlighting various skewed data distributions.

2 Related Work

We group the related work into two classes of bias analysis: (i) knowledge graphs and (ii) embedding models.

Bias analysis of knowledge graphs. [7] proposes methods to trace the provenance of crowdsourced fact checking to enable bias transparency rather than aiming at eliminating bias from a KG. Furthermore, they investigate how paid crowdsourcing can be used to understand contributors' implicit bias. Specifically, they recruit click workers to verify controversial facts and study them as they do so. I.e., they track what search engines are used and which position the URL used to validate was ranked in the result page. An example verification task is the question of whether Catalonia is a part of Spain or an independent country. The paper proposes adding both facts to the knowledge graph, with a statement testifying how much support there is for each fact.

[15] introduces ProWD, a framework and tool for profiling the completeness of Wikidata. Completeness measure is based on Class-Facet-Attribute (CFA) profiles. For example one could compare how often the attribute "educated at" or"date of birth" compare between male, German computer scientists, and female, Indonesian computer scientists.

Bias analysis of embedding models. Bourli et al. [4] present an analysis method for investigating gender bias with regards to occupation in entity embeddings. Specifically, they subtract the male embedding from the female entity embedding to get the bias vector. Projecting an occupation on this vector then gives them the bias in this occupation. Furthermore, they introduce a de-biasing approach that generates new de-biased embedding vectors from the existing one by subtracting it from the bias vector.

[10] conduct experiments on Wikidata and Freebase, and show that harmful social biases related to professions are encoded in the embeddings with respect to gender, religion, ethnicity and nationality. They first explain how traditional word embeddings metrics do not apply to KG embeddings due to the transformations applied. They then provide a method for evaluating bias. Their method operates through increasing/decreasing an entities score of a sensitive attribute (e.g., make it more male and less female) and then recording how the likelihood of a certain target triple being true changes (e.g., whether they are a nurse of a lawyer). As a followup, the authors present a novel approach to KG embedding where embeddings are trained to be neutral with respect to sensitive features using an adversarial loss function [9] . To achieve this, they add a neural-network based classifier to the scoring function: scores are penalized when this classifier can predict the value of the sensitive attribute from the existing embedding. However, this means that the embeddings are only neutral with respect to the power of the model: a more powerful model might be able to infer the sensitive values.

These (and other initiatives) indicate that there is a growing attention to bias in knowledge graphs, and efforts to make bias visible. As knowledge graphs often are collaborative repositories, it is relevant to provide users with accessible means for identifying possible bias. The examples above are helpful but limited in two ways: they either are valid for a specific knowledge graph, and/or a limited number of attributes. A general framework might provide more possibilities to map bias in knowledge graphs and enable users to become aware of the dis-

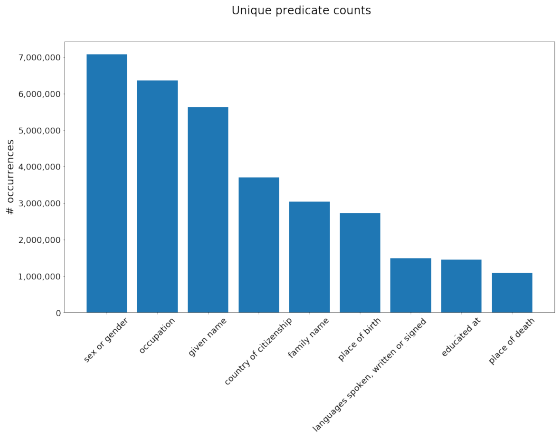


Fig. 1: How many humans have at least one occurrence of a given property?

tribution of items and attributes in a given knowledge graph. With their own subject specific expertise, these users can then decide which bias is problematic, and how to address it.

3 Wikidata Completeness Analysis

Wikidata is the large, open knowledge graph which acts as central storage for the structured data of other Wikimedia projects such as Wikipedia. Data is stored as claims or triples, containing a subject item, a property and a value. Values are entities or literals such as a quantity, a string or even a coordinate. Items are identified through URIs starting with 'Q' (e.g. Q22686 for *Donald Trump*) and properties are identified through URIs starting with 'P' (e.g., P40 for *Child*). Claims can be contextualized with additional data such as sources (for the data), ranks (in case of multiple values for a property) and qualifiers (e.g., to note that a fact was true at a specific point in time, or that a fact is disputed). A claim and its additional data are collectively referred to as a statement.

3.1 Completeness

We investigated how several properties were distributed among the class of humans in Wikidata. An item x is a human when it is an instance of (P31) human (Q5), i.e., item x must have the claim (x, P31, Q5). Using the Wikidata dump from 2021/03/31, we extracted 9,028,271 such items. We will now give a brief overview of some of our preliminary findings.

To begin we, for each item in the subset we counted whether or not a property occurs among its claims. This gives us an overview of how often a property occurs at least once. The result is displayed in Figure 1. Ignoring the predicate *instance of*, which per our definition is present on all humans, the most occurring predicates are *sex or gender* (P21), *occupation* (P106), and *given name* (P735).

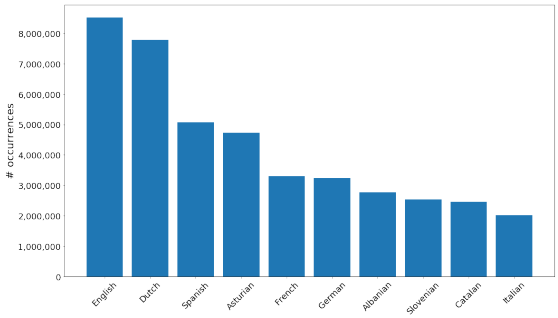


Fig. 2: How many humans have a label specified in the given language.

These occur on 7,079,543 (78%), 6,359,256 (70%) and 5,635,238 (62%) humans respectively.

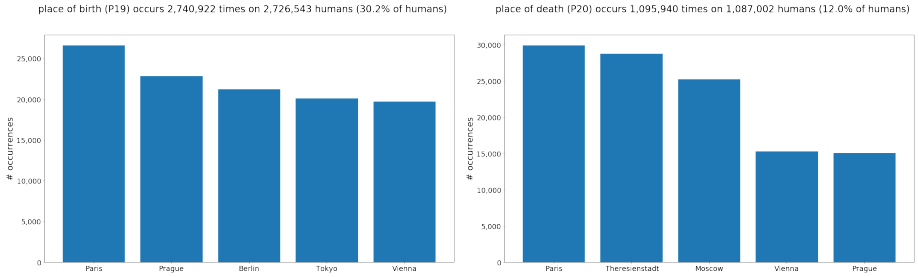
Additionally, we counted the number of languages each item had a label in. This gives us an overview of how complete Wikidata is over several languages. The result of this is displayed in Figure 2. Expectedly, the most common language is English, with 8.517.283 (94%) humans having an English label. More unexpectedly however is the fact that, in spite of being a small country with only 17 million inhabitants, the second most common label language is Dutch with 7.785.518 (86%) humans having a Dutch label.

Next, we can look at the distribution of object entities for a given predicate. I.e., given a predicate such as *place of death* (P20) we can count how many people have object values such has Moscow or Paris. From this data, we have created a bar graph for a selection of predicates in Figure 3.

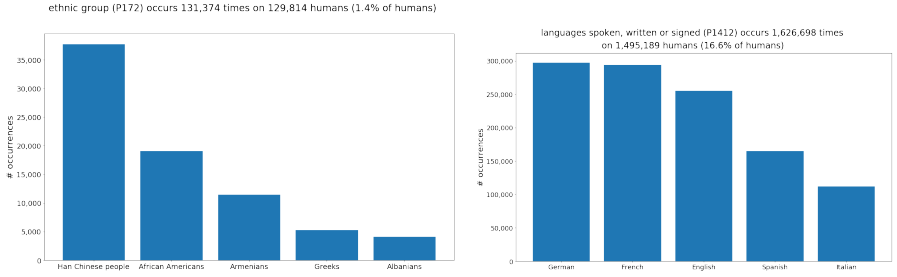
Looking at this data, it is immediately clear that it is not representative of the common population. For instance, the most common occupation by far is researcher (20%). Yet in reality, even in the USA only around 2% of the population has a PhD.[6] We of course understand that an encyclopedia covers persons of interest and not the general population. Hence it is logical that there is a bias. However, the problematic bias is not the overrepresentation of scholars but the overrepresentation of of white male scholars at western universities. If we want to inquire to what extend the population of researchers in Wikipedia is skewed we need to inquire about the presence of other occupations for persons of interest for an encyclopedia, such as athletes, activists, politicians, engineers and inventors.

We hypothesize that there are two main sources of bias present in this data. The first is availability bias, i.e., much of the data present in Wikidata is there because it could be easily imported. For instance through the use of bots. The second is interest bias, where the interests of the people who work on Wikidata end up deciding what content will dominate the dataset. Examples of this bias are the most common occupation being researcher (imported through article papers) and the second most common place of death being a concentration camp.

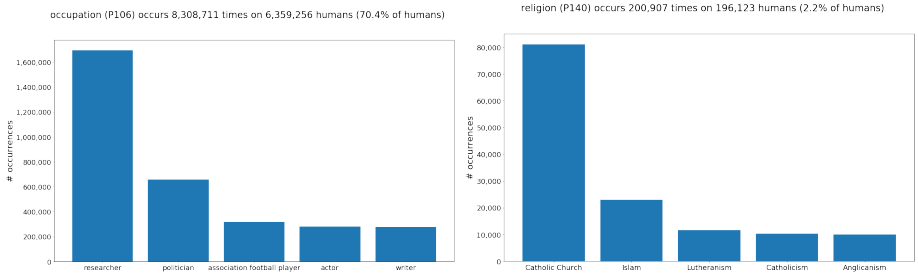
[6] `https://data.worldbank.org/indicator/SE.TER.CUAT.DO.ZS?locations=US`



(a) Most common place of birth (left) and place of death (right) entities.



(b) Most common ethnic groups (left) and languages spoken (right) entities.



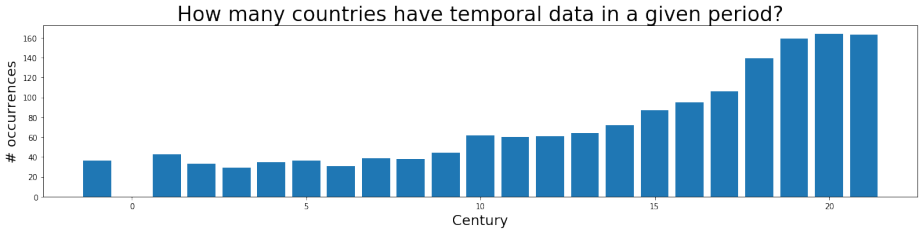
(c) Most common occupations (left) and religions (right) entities.

Fig. 3: Overview of the most common values for several predicates in Wikidata.

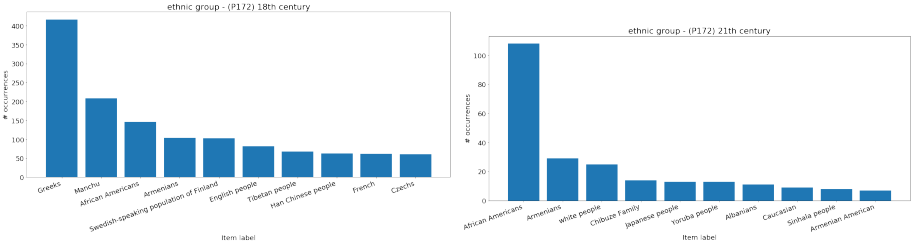
3.2 Spatiotemporal Analysis

Temporal information in Wikidata present itself in two ways. Firstly, predicates can directly have timestamps as their object value. For instance, the date of birth of a person. All predicates that can have a timestamp as object value must be *instances of* (P31) *Wikidata property with datatype 'time'* (Q18636219). There are 34 such predicates. Secondly, temporal information can be included in any other predicate through the use of qualifiers. I.e., predicates *start time* (P569) and *end time* (P570) can be applied to a triple through reification to add temporal information to that triple.

Since we are interested in how humans are represented in Wikidata, we restrict spatiotemporal analysis to the human class. Specifically, we ground data in space by looking at a persons place of birth (*P19*) and in time by looking at the place of birth (*P569*). Through this we can analyse the completeness of



(a) Number of countries with at least one fact in the given century.



(b) Comparison between the most common ethnic groups listed in Wikidata in the 18th (left) and 21st (right) century.

Fig. 4: Overview of some select metrics of Wikidata temporal metrics.

Wikidata over time. Some results are displayed in Figure 4. We observe that the further we go back in time, the less distinct countries are observed in Wikidata. I.e., facts seem to be more based on a few countries. Addtionally, we investigate the occurrences of the most common ethnic groups listed in Wikidata. Interestingly, the use of ethnic group seems to have fallen out of favour for people born more recently. In the 18th century the most common ethnic groups was Greeks with over 400 occurrences, whereas in the 21st century the most common ethnic group is African American with just over 100 occurrences.

4 Bias Analysis of Knowledge Graph Embedding Models

In this section we perform bias analysis of the knowledge graph embedding models. Specifically, we analyze the effect of balancing the data on link prediction performance. For this task we utilize two popular models, TransE [3] and DistMult [16]. We perform our experiments on two state-of-the-art knowledge graphs. The first is Wikidata12k, a subset of Wikidata extracted by [11, 6]. The second is DBP15k, a subset of DBpedia [1] originally created by [13] to test Entity-Alignment models. As we are interested in link prediction rather than entity alignment, we select a single instance of the dataset (the English version) and perform our experiments on it. All of our code is available on Github [7]

[7] `https://github.com/wradstok/KGE-bias-analyzer`

4.1 Embedding Models

For a triple (s, p, o), let (e_s, e_p, e_o) denote its embedding vectors. Taking a KG and a random initialization of the vectors as an input, a vector representation of the KG is gradually learned using a scoring function $\phi(s, p, o)$. The scoring function should reflect how well the embedding captures the semantics of the KG. The learned embeddings can be used in tasks such as classification, clustering, and link prediction. In this work, we are focussed on the last. Link prediction is the task of predicting the most likely element given a tuple where one element is missing, e.g., given a triple $(s,p,?)$, to predict the most likely object entity.

The most popular embedding model is TransE (Translating embeddings). Its scoring function is based on the intuition that the the subject and object vector should be close together after adding the predicate vector. Its scoring function is written as $\phi(s, p, o) = ||e_s + e_p - e_o||_{1,2}$. While being very powerful, it has limited expressiveness due to its is simplicty. Therefore, we also perform experiments with DistMult, a multiplicative model. Its scoring function is written as $\phi(s, p, o) = ||e_s * e_p * e_o||_{1,2,}$. In our experiments we do not use pre-trained models and instead train the embeddings from scratch.

4.2 SMOTE

One way of balancing datasets is to use Synthetic Minority Over-sampling TEchnique (SMOTE) [5]. SMOTE is an over-sampling technique that allows one to construct new examples of a given class based on existing examples in order to address imbalances in the dataset. An example would be oversampling female football players. However, SMOTE is not intended for graph datasets and as such is not trivial to apply to knowledge graphs while maintaining the underlying structure.

One way to apply SMOTE to graph data is through embedding the graph first. We use this approach to evaluate how well SMOTE is suited for our scenario. Our method is as follows. After obtaining the embeddings, we create a categorical variable with a category for each possible combination of sensitive features. In the case of 5 occupations and 2 genders, this implies 10 categories. The embeddings vectors are then combined with this categorical value associated with the sample it represents. Finally, we instruct SMOTE to generate the maximum number of samples for each possible entry using the python imbalanced-learn library [12], i.e., given that there are 1850 male association football players, we create both male and female physicists until there are 1850 of each.

However, preliminary experiments found that this method did not suffice for generating a balanced set of embeddings. Applying our evaluation method to datasets produced by above procedure did not create balanced predictions. We hypothesize that it is because SMOTE generates new examples based on existing biases. By interpolating new 'female' examples from existing female embeddings, we are only creating new examples in the same cluster. That means that locations in the embedding space which are already female become much more so. Therefore, we instead extend the datasets by sampling additional triples from the original knowledge graphs.

Dataset	# Triples	# Entities	# Pred.	# Men	# Women
Wikidata12k (original)	38,970	12,848	25	4,905	717
Wikidata12k (balanced)	51,682	15,957	25	4,905	3,610
DBpedia15k (original)	92,746	18,716	206	6,767	1,087
DBpedia15k (balanced)	95,827	27,459	206	5,916	5,917

Table 1: Dataset statistics

4.3 Sampling process

We enrich the original knowledge graphs by adding *female triples*, i.e. extra triples with female entities as subject. The data is enriched in such a way that the number of men and women associated with each of the top 5 most common occupations becomes approximately equal. The triples are obtained from the complete Wikidata and DBpedia datasets.

To ensure that the new triples have healthy connectivity with regards to the rest of the graph this is done in a three step process. Firstly, all women with the required occupations are selected from the complete knowledge graph. Secondly, from this selection the women which have the largest number of predicates which are also in the original dataset are picked. Finally, we select the women whose object values are already in the graph. The last step ensures that we do not add object entities which occur only a few times, and only with women.

4.4 Wikidata12k

Wikidata12k does not contain any information about gender or occupation. However, we can look up this data by querying the original Wikidata knowledge graph. As Wikidata12k is originally a temporal knowledge graph, we strip out the temporal information and remove and duplicate triples that may be created by this process.

The five most common occupations are association football player *Q937857* (1867), politician *Q82955* (918), actor *Q33999* (211), writer *Q36180* (184) and physicist *Q169470* (143). These occupations are not uniformly distributed with regards to gender: there are only a handful of women football players, and there is not a single woman physicist in the entire dataset.

In total, we add around 10,000 triples with female entities as subject to the Wikidata12k knowledge graph, resulting a new graph over 50,0000 triples. This increases the average number of mentions as subject (i.e., the average number of outlinks) per female entity from 3.49 in the original graph to 4.74 in the balanced graph. However, the number of outlinks still falls short of that of men, which is 5.41.

4.5 DBP15k

DBP15k is a subset of DBpedia [1] created by [13] to test Entity-Alignment models. The majority of predicates in DBP15k have very few triples associated

	Data			Prediction			
	Men	**Women**	**Women (%)**	**Men**	**Women**	**Women (%)**	**Diff (p.p.)**
Officeholder	1803	180	9.1%	85	15	15.0%	6.9
Athlete	1142	6	0.5%	100	0	0.0%	-0.5
Royalty	569	235	29.2%	60	40	40.0%	10.8
Sportsmanager	225	0	0.0%	100	0	0.0%	0.0
Scientist	216	6	2.7%	95	5	5.0%	2.3
Total	3955	427	9.7%	440	60	12.0%	-

	Data			Prediction			
	Men	**Women**	**Women (%)**	**Men**	**Women**	**Women (%)**	**Diff (p.p.)**
Officeholder	1498	1770	54.2%	16	83	83.8%	29.7
Athlete	990	1320	57.1%	55	45	45.0%	-12.1
Royalty	472	596	55.8%	25	74	74.7%	18.9
Sportsmanager	188	31	14.2%	85	15	15.0%	0.8
Scientist	195	225	53.6%	32	67	67.7%	14.1
Total	3343	3942	54.1%	213	284	57.1%	-

Table 2: Comparison between the male/female distribution and the resulting **TransE** model predictions in the original **DBPedia15k** dataset (top) and our balanced dataset (bottom). Difference column contains the difference (in percentage points) between the % of women in the predicted and the % of women predicted.

	Data			Prediction			
	Men	**Women**	**Women (%)**	**Men**	**Women**	**Women (%)**	**Diff (p.p.)**
Politician	619	96	13.4%	80	20	20.0%	6.6
Writer	100	33	24.8%	74	25	25.3%	0.5
Actor	67	89	57.1%	38	61	61.6%	4.6
Football player	1465	14	0.9%	98	2	2.0%	1.1
Physicist	108	2	1.8%	98	2	2.0%	0.2
Total	2359	234	9.0%	388	110	22.1%	-

	Data			Prediction			
	Men	**Women**	**Women (%)**	**Men**	**Women**	**Women (%)**	**Diff (p.p.)**
Politician	644	331	33.9%	53	47	47.0%	13.1
Writer	111	59	34.7%	25	74	74.7%	40.0
Actor	71	93	56.7%	33	66	66.7%	10.0
Football player	1455	1427	49.5%	8	91	91.9%	42.4
Physicist	122	44	26.5%	56	44	44.0%	17.5
Total	2403	1954	44.8%	175	322	64.8%	-

Table 3: Comparison between the male/female distribution and the resulting **TransE** model predictions in the original **Wikidata12k** dataset (top) and our balanced dataset (bottom). Difference column contains the difference (in percentage points) between the % of women in the predicted and the % of women predicted.

with them. To prevent the graph from being too sparse for an embedding model to learn we delete all predicates which occur less than 50 times.

DBpedia does not store any information about peoples sex or gender in a structured way. I.e., although a person can be of *rdf:type* of *Man* or *Woman*, manual inspection of the data did not reveal that this information was consistently present. However, most entities do contain their Wikidata identifiers. Since Wikidata does list peoples gender, we determine a persons gender by querying Wikidata for the given identifiers.

The five most common occupations are OfficeHolder (2508), Athlete (1436), Royalty (1002), SportsManager (288), and Scientist (282). Like Wikidata12k, the male/female ratio in these occupations is unbalanced, skewing heavily towards men. In addition to balancing the data by adding additional samples, we remove some male entities and their triples to create the balanced dataset.

4.6 Evaluation

To evaluate whether an embedding model contains bias with regards to gender and occupation we perform the following procedure. Firstly, we count the fraction of men and women that have a certain occupation (*P106*) x. Then, we ask the model to predict the n most likely entities for the query (*?*, *P106*, x). If the fraction of men or women returned is consistently larger than that the fraction present in the data, the model is biased. Specifically, when more men are predicted the model is biased against women and vice versa. If this bias is only present in the unbalanced dataset and not the balanced datasets, then the model reflects the data it has been trained on. However, if the bias is present in both scenarios, the models are either inherently biased or manage to pick up some form of bias in the data which is not reflected in our analysis.

4.7 Results

Our results are displayed in Tables 2 and 3 for DBpedia and Wikidata12k respectively using TransE [3], and in Tables 4 and 5 using DistMult [16]. We observe that in both original datasets, the percentage of women predicted is very low and close to the percentage of women in the dataset. The largest difference is observed on the occupation Royalty in the DBpedia15k dataset. Here the difference is just over 10 percentage points.

When we extend our view to the balanced datasets, we find that the percentage of women predicted has moved upwards with the percentage of women in the dataset. Balancing the datasets thus helps with improving the representation of minority classes in the model output. However, we do observe that the absolute differences between the number of men in the dataset, and the number of men predicted (and for women) has increased, suggesting that the model has become less accurate.

Even so, we believe a more likely explanation to be that the larger number of entities predicted induces more variance in the predictions. This explanation is strengthened by the fact that the difference is smaller when using DistMult,

	Data			Prediction			
	Men	**Women**	**Women (%)**	**Men**	**Women**	**Women (%)**	**Diff (p.p.)**
Officeholder	1803	180	9.1%	87	13	13.0%	3.9
Athlete	1142	6	0.5%	100	0	0.0%	0.5
Royalty	569	235	29.2%	68	32	32.0%	2.8
Sportsmanager	225	0	0.0%	100	0	0.0%	0.0
Scientist	216	6	2.7%	96	4	4.0%	1.3
Total	3955	427	9.7%	451	49	9.8%	-

	Data			Prediction			
	Men	**Women**	**Women (%)**	**Men**	**Women**	**Women (%)**	**Diff (p.p.)**
Officeholder	1498	1770	54.2%	47	53	53.0%	-1.2
Athlete	990	1320	57.1%	78	22	22.0%	-35.14
Royalty	472	596	55.8%	39	61	61.0%	5.2
Sportsmanager	188	31	14.2%	88	12	12.0%	-2.2
Scientist	195	225	53.6%	51	49	49.0%	-4.6
Total	3343	3942	54.1%	303	197	39.4%	-

Table 4: Comparison between the male/female distribution and the resulting **DistMult** model predictions in the original **DBpedia15k** dataset (top) and our balanced dataset (bottom). Difference column contains the difference (in percentage points) between the % of women in the predicted and the % of women predicted.

	Data			Prediction			
	Men	**Women**	**Women (%)**	**Men**	**Women**	**Women (%)**	**Diff (p.p.)**
Politician	619	96	13.4%	83	17	17.0%	3.6
Writer	100	33	24.8%	72	28	28.0%	3.2
Actor	67	89	57.1%	50	50	50.0%	-7.1
Football player	1465	14	0.9%	99	1	1.0%	0.1
Physicist	108	2	1.8%	95	5	5.0%	3.2
Total	2359	234	9.0%	399	101	20.2%	

	Data			Prediction			
	Men	**Women**	**Women (%)**	**Men**	**Women**	**Women (%)**	**Diff (p.p.)**
Politician	644	331	33.9%	51	49	49.0%	15.1
Writer	111	59	34.7%	39	61	61.0%	26.3
Actor	71	93	56.7%	37	63	63.0%	6.3
Football player	1455	1427	49.5%	50	50	50.0%	0.5
Physicist	122	44	26.5%	42	58	58.0%	31.5
Total	2403	1954	44.8%	219	281	56.2%	

Table 5: Comparison between the male/female distribution and the resulting **DistMult** model predictions in the original **Wikidata12k** dataset (top) and our balanced dataset (bottom). Difference column contains the difference (in percentage points) between the % of women in the predicted and the % of women predicted.

which is a more expressive model and can thus model the information more accurately.

Another point of note is the observation that on both datasets and almost all occupations, the sign of the difference between the percentage of women predicted and the percentage of women in the dataset is mostly positive. This means that women are actually *overrepresented* in the models predictions, indicating that the model is actually less biased than the data.

5 Conclusion

In this paper we proposed a new approach to mitigate bias in knowledge graphs embedding models by leveraging the distribution of the datasets in which the models are trained on. Specifically, rather than adapting models to mitigate bias, we instead analyze and augment the data that is fed into the model. We carried out several experiments using state of the art embedding models (namely, TransE and DistMult) and two knowledge graphs (namely DBpedia and Wikidata) and showed that balancing the data with regards to specific sensitive features (e.g. gender and occupation) improves the overall prediction capabilities of the models. Additionally, to motivate our work, we have done a completeness analysis of Wikidata using a number of sensitive features.

As a future work, we will extend the proposed approach to build a system that takes as an input a dataset and a selection of sensitive features and automatically balances the data with respect to the given features.

References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia-a crystallization point for the web of data. Journal of web semantics **7**(3), 154–165 (2009)
2. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. pp. 1247–1250 (2008)
3. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. Advances in neural information processing systems **26** (2013)
4. Bourli, S., Pitoura, E.: Bias in knowledge graph embeddings. In: 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 6–10 (2020). https://doi.org/10.1109/ASONAM49781.2020.9381459
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–3x57 (2002)
6. Dasgupta, S.S., Ray, S.N., Talukdar, P.: Hyte: Hyperplane-based temporally aware knowledge graph embedding. In: Proceedings of the 2018 conference on empirical methods in natural language processing. pp. 2001–2011 (2018)

7. Demartini, G.: Implicit bias in crowdsourced knowledge graphs. In: Companion Proceedings of The 2019 World Wide Web Conference. p. 624–630. WWW '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3308560.3317307, https://doi.org/10.1145/3308560.3317307
8. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 601–610 (2014)
9. Fisher, J., Mittal, A., Palfrey, D., Christodoulopoulos, C.: Debiasing knowledge graph embeddings. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7332–7345. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.595, https://www.aclweb.org/anthology/2020.emnlp-main.595
10. Fisher, J., Palfrey, D., Christodoulopoulos, C., Mittal, A.: Measuring social bias in knowledge graph embeddings. arXiv preprint arXiv:1912.02761 **todo** (2019)
11. Leblay, J., Chekol, M.W.: Deriving validity time in knowledge graph. In: Companion Proceedings of the The Web Conference 2018. pp. 1771–1776 (2018)
12. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research **18**(17), 1–5 (2017), http://jmlr.org/papers/v18/16-365
13. Sun, Z., Hu, W., Li, C.: Cross-lingual entity alignment via joint attribute-preserving embedding. In: d'Amato, C., Fernandez, M., Tamma, V., Lecue, F., Cudré-Mauroux, P., Sequeda, J., Lange, C., Heflin, J. (eds.) The Semantic Web – ISWC 2017. pp. 628–644. Springer International Publishing, Cham (2017)
14. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
15. Wisesa, A., Darari, F., Krisnadhi, A., Nutt, W., Razniewski, S.: Wikidata completeness profiling using prowd. In: Proceedings of the 10th International Conference on Knowledge Capture. p. 123–130. K-CAP '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3360901.3364425, https://doi.org/10.1145/3360901.3364425
16. Yang, B., Yih, W.t., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575 (2014)