Bias in Knowledge Graph Embeddings

A Thesis

submitted to the designated
by the General Assembly
of the Department of Computer Science and Engineering
Examination Committee

by

Styliani Bourli

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN DATA AND COMPUTER SYSTEMS ENGINEERING

WITH SPECIALIZATION
IN DATA SCIENCE AND ENGINEERING

University of Ioannina
February 2021

Examining Committee:

- **Evaggelia Pitoura**, Professor, Department of Computer Science and Engineering, University of Ioannina (Advisor)
- **Panos Vassiliadis**, Professor, Department of Computer Science and Engineering, University of Ioannina
- **Panayiotis Tsaparas**, Assoc. Professor, Department of Computer Science and Engineering, University of Ioannina

Acknowledgements

I would like to express my sincere thanks to my advisor, Prof. Evaggelia Pitoura, for the opportunity she gave me to perform this thesis in an interesting field such as the knowledge graph embeddings, as well as for the important help, support and guidance she has provided me.

Table of Contents

List of Figures iii

List of Tables v

Abstract vii

Εκτεταμένη Περίληψη ix

1 Introduction 1

1.1 Motivation 1

1.2 Structure 3

2 Preliminaries 4

2.1 Knowledge Graph 4

2.2 Knowledge Graph Embeddings 6

2.3 Knowledge Graph Embedding Models 7

2.3.1 The TransE model 8

2.3.2 The TransH model 8

2.3.3 Comparison of the two models 9

2.4 Notations 9

3 Bias in the dataset 10

3.1 Metrics for measuring bias in the dataset 10

3.1.1 Equal opportunity approach 10

3.1.2 Equal number approach 11

3.1.3 Measuring bias in data with the two approaches 11

4 Bias in knowledge graph embeddings 13

4.1 Methods for measuring bias in knowledge graph embeddings 13

4.1.1 Projection method . . . 13
4.1.2 Analogy puzzle method . . . 15
4.1.3 Prediction method . . . 16
4.2 Clustering fairness using knowledge graph embeddings . . . 16
4.2.1 Clustering fairness using knowledge graph embeddings . . . 17

5 Debias the knowledge graph embeddings 18
5.1 Debias approach . . . 18

6 Experiments 20
6.1 Datasets . . . 20
6.1.1 Synthetic knowledge graph . . . 21
6.1.2 Real knowledge graphs . . . 21
6.1.3 Training the knowledge graphs . . . 23
6.2 Bias in the dataset . . . 23
6.2.1 Data bias metric selection . . . 24
6.2.2 Bias results in the datasets . . . 27
6.3 Bias results in the KG embeddings . . . 31
6.3.1 Bias results in the KG embeddings using projections . . . 31
6.3.2 Detection of bias in KG embeddings using an analogy puzzle . . . 36
6.3.3 Prediction task for bias amplification detection . . . 38
6.4 Clustering results . . . 42
6.5 Debias evaluation . . . 47
6.5.1 Evaluation using projections and similarity . . . 47
6.5.2 Evaluation using prediction task . . . 49
6.5.3 Evaluation using clustering . . . 52

7 Related Work 55

8 Conclusions and Future work 57

Bibliography 58

List of Figures

2.1 Example of a knowledge graph. 5

2.2 Example of KG embeddings representation in the 2-dimensional vector space. 7

2.3 If (‘‘$Mona\ Lisa$”, ‘‘$was\ created\ by$”, ‘‘$Leonardo\ da\ Vinci$”) triple holds, then using Transe model ‘‘$\vec{Mona\ Lisa}$” + ‘‘$\vec{was\ created\ by}$” should be near to ‘‘$\vec{Leonardo\ da\ Vinci}$” and using TransH model, ‘‘$\vec{Mona\ Lisa}_{\perp}$”+ ‘‘$\vec{was\ created\ by}$” should be near to ‘‘$\vec{Leonardo\ da\ Vinci}_{\perp}$”. 9

6.1 There is a power law distribution on the number of individuals per occupation in both the original graph and the subgraph; there are few occupations that many users have and many that few have. 22

6.2 Average results of prediction task using five synthetic graphs and TransE embeddings. 27

6.3 Average results of prediction task using five synthetic graphs and TransH embeddings. 28

6.4 Plots of the distribution of neutral occupations for different values of t. The 6.4a refers to Wikidata subgraph and the gender sensitive attribute. The appopriate value for t is 0.0001. The 6.4b refers to FB13 dataset and the religion sensitive attribute. The appropriate value for t is 0.001. The 6.4c refers to FB13 dataset and the nationality sensitive attribute. The appropriate value for t is also 0.001. The 6.4d refers to FB13 dataset and the enthnicity sensitive attribute. The appropriate value for t is here 0.01. 30

6.5 Let a, b be the sensitive values. Then the upper right box of each plot refers to the occupations characterized as a both in data and in the embeddings. The bottom right box refers to the occupations marked as a in the data and b in the embeddings. The lower left box refers to the occupations marked as b both in data and in the embeddings. Finally, the upper left box refers to the occupations marked as b in the data, but as a in the embeddings. 35
6.6 The expected and the predicted numbers of individuals in top-K results using TransE embeddings. Gender concerns to Wikidata dataset, while religion, nationality and ethnicity concern to FB13 dataset. 39
6.7 The expected and the predicted numbers of individuals in top-K results using TransH embeddings. Gender concerns to Wikidata dataset, while religion, nationality and ethnicity concern to FB13 dataset. 40
6.8 Clustering results using FB13 dataset and TransE embeddings. 43
6.9 Clustering results using FB13 dataset and TransH embeddings. 44
6.10 Clustering results of occupations using FB13 dataset and TransE embeddings. 45
6.11 Clustering results of occupations using FB13 dataset and TransH embeddings. 46
6.12 Prediction task for debias evaluation using λ=0.5, λ=0.8 and K=10. . . 51
6.13 Clustering results of occupations using TransE embeddings before debias and after a debias with λ=0.8. 53
6.14 Clustering results of occupations using TransH embeddings before debias and after a debias with λ=0.8. 54

List of Tables

6.1 Information of the real datasets. 23

6.2 Information about the number of individuals and occupations of each sensitive attribute in the datasets. The gender sensitive attribute concerns the Wikidata subgraph, while the religion, nationality and ethnicity sensitive attributes concern the FB13. 29

6.3 The five most biased occupations based on gender, religion, nationality and ethnicity sensitive attributes on Wikidata subgraph and FB13 dataset. 29

6.4 The five most biased occupations based on gender, religion, nationality and ethnicity sensitive attributes on Wikidata subgraph and FB13 dataset. In Table 6.4a there are the results from TransE embeddings, while in Table 6.4b there are the results from TransH embeddings. . . 32

6.5 Pearson correlation coefficient score between data bias and embedding bias using TransE and TransH embeddings. 33

6.6 Pearson correlation coefficient score between popularity and embedding bias using our metric, let be proj, the embedding bias metric in [1] and the FB13 dataset. 34

6.7 Percentage of occupations, which are characterized as a or b both in the dataset and in the embeddings. 36

6.8 The pairs of occupations that suit in the analogy puzzle "*b is to x as a is to y*", where a and b are two sensitive values. 37

6.9 Quantification of amplification. The average results of the $predicted - expected$ score of the five most biased occupations, using TransE embeddings, K=10, 20, 50, 100 for the Wikidata subgraph (for gender), and K = 5, 7, 10 for FB13 dataset (for religion, nationality and ethnicity). 41

6.10 Quantification of amplification. The average results of the $predicted - expected$ score of the five most biased occupations, using TransH embeddings, K=10, 20, 50, 100 for the Wikidata subgraph (for gender), and K = 5, 7, 10 for FB13 dataset (for religion, nationality and ethnicity). 41
6.11 Results of the semantic relation of occupations in the clusters. 47
6.12 Let a, b be the two sensitive values. Table 6.12a gives the average projection score results of a, b and $neutral$ occupations on bias direction using TransE embeddings. Moreover Table 6.12b gives the average projection score results of a, b and $neutral$ occupations on bias direction using TransH embeddings. Both Tables present the projection score before debias, after a "soft" debias with λ=0.5 and after a "hard" debias with λ=1. 48
6.13 Let a, b be the two sensitive values. Table 6.13a gives the average cosine similarity results of a, b and $neutral$ occupations with a and b entities using TransE embeddings. Moreover Table 6.13b gives the average cosine similarity results of a, b and $neutral$ occupations with a and b entities using TransH embeddings. Both Tables present the cosine similarity score before debias, after a "soft" debias with λ=0.5 and after a "hard" debias with λ=1. 50
6.14 Hits@10 for the most 40 biased gender occupations using Wikidata dataset, and Hits@5 for the most 20 biased religion, nationality and ethnicity occupations using FB13 dataset. 52

Abstract

Styliani Bourli, M.Sc. in Data and Computer Systems Engineering, Department of Computer Science and Engineering, School of Engineering, University of Ioannina, Greece, February 2021.
Bias in Knowledge Graph Embeddings.
Advisor: Evaggelia Pitoura, Professor.

Knowledge graphs (KGs) are multi-relational directed graphs used in many tasks in recent years, including question answering, recommendation and information retrieval. They are associated with, and used by search engines such as Google, Bing, and Yahoo; and social networks such as LinkedIn and Facebook. Knowledge graph embeddings have gained a lot of attention recently, because they can map the components of a knowledge graph to a low dimensional vector space. In the era of big data, this is very important because it makes KG usage and analysis easier. But the connection of the KG embeddings production with machine learning, combined with the fact that bias learning problem using machine learning tasks receives more attention in current research, leads to concern about bias that may exists in data, transferred to the KG embeddings through learning and possibly reinforced by them. In this thesis we study the bias in KG embeddings.

We first define two approaches to quantify the bias in the dataset and after their comparison we choose the one we consider more appropriate. For measuring bias in the KG embeddings, we use a projection method and an analogy puzzle to determine quantitatively and qualitatively if the bias is transferred from the data to the KG embeddings. We also apply a prediction method to study if there is in addition a bias amplification using the KG embeddings. We further detect if the popularity of some entities, or the inequality in populations of sensitive values like male, female individuals in the dataset, affects bias in KG embeddings, and, moreover, if other tasks such as clustering affected by the bias of the KG embeddings. We then define a

debias method based on projections in the bias subspace. Its novelty lies on tuning the amount of bias it removes and in the usage of pretrained embeddings instead of the modification of the KG embedding model.

We conduct experiments using a set of real and synthetic KGs and two widely known KG embedding models. We provide a presentation and an analysis of the results. Our approaches can be easily generalized in other datasets and more KG embedding models.

Εκτεταμενη Περιληψη

Στυλιανή Μπουρλή, Δ.Μ.Σ. στη Μηχανική Δεδομένων και Υπολογιστικών Συστημάτων, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πολυτεχνική Σχολή, Πανεπιστήμιο Ιωαννίνων, Φεβρουάριος 2021.
Μεροληψία σε Ενσωματώσεις Γραφημάτων Γνώσης.
Επιβλέπων: Ευαγγελία Πιτουρά, Καθηγήτρια.

Τα γραφήματα γνώσης (knowledge graphs), είναι κατευθυνόμενα γραφήματα που περιέχουν πληροφορία διαφόρων οντοτήτων και σχέσεων του πραγματικού κόσμου. Χρησιμοποιούνται σε πολλές εφαρμογές τα τελευταία χρόνια, όπως στην ανάκτηση πληροφορίας και σε συστήματα συστάσεων, καθώς επίσης σε μηχανές αναζήτησης, όπως Google, Bing και Yahoo, αλλά επίσης και σε κοινωνικά δίκτυα, όπως το LinkedIn και το Facebook. Μερικά από τα μεγαλύτερα γραφήματα γνώσης είναι της Microsoft, του ebay, της Google και του Facebook. Υπάρχουν όμως και ανοικτά γραφήματα ελεύθερης πρόσβασης όπως το Wikidata ή παλιότερα το Freebase.

Οι ενσωματώσεις γραφημάτων γνώσεις (knowledge graph embeddings), έχουν συγκεντρώσει μεγάλο ενδιαφέρον τα τελευταία χρόνια, επειδή μπορούν και αναπαριστούν την πληροφορία των γραφημάτων γνώσης με διανύσματα σε ένα χώρο χαμηλής διάστασης. Δεδομένου του ότι ζούμε στην εποχή των μεγάλων δεδομένων, η αναπαράσταση της πληροφορίας με διανύσματα και μάλιστα σε έναν χαμηλής διάστασης διανυσματικό χώρο, βοηθάει στην ευκολότερη διαχείριση και ανάλυση των γραφημάτων. Όμως, το γεγονός ότι η παραγωγή των ενσωματώσεων είναι άμεσα συνδεδεμένη με την εφαρμογή μηχανικής μάθησης στο γράφημα, σε συνδυασμό με το πρόβλημα που έχει εντοπιστεί τα τελευταία χρόνια μεταφοράς πληροφορίας μεροληψίας μέσω της μάθησης, οδηγεί σε ανησυχία για πιθανή μετάδοση πληροφορίας που σχετίζεται με τη μεροληψία στις ενσωματώσεις, και ίσως σε ενίσχυσή της από

αυτές κατά τη χρήση τους. Σε αυτή την εργασία, μελετάμε συγκεκριμένα τη μεροληψία στις ενσωματώσεις των γραφημάτων γνώσης.

Όσον αφορά τα δεδομένα στο γράφημα αναμένουμε ότι, εφόσον προέρχονται από την πραγματική ζωή στην οποία υπάρχει συχνά ανισότητα και αδικία, η πληροφορία που έχουν περιέχει μεροληψία. Για να εξακριβώσουμε αν αυτό όντως συμβαίνει, αλλά και να μετρήσουμε την μεροληψία αυτή στα δεδομένα, ορίζουμε δύο μετρικές. Μετά από σύγκριση των δύο μετρικών επιλέγουμε αυτή που θεωρούμε καταλληλότερη. Στη συνέχεια, για να εξετάσουμε αν η μεροληψία μεταφέρεται από τα δεδομένα στις ενσωματώσεις, αλλά και για να μετρήσουμε ποσοτικά και ποιοτικά τη μεροληψία αυτή, χρησιμοποιούμε δύο μεθόδους, μία μέθοδο βασισμένη σε προβολές και ένα παζλ βασισμένο σε αναλογίες. Ωστόσο ενδιαφερόμαστε επιπλέον εκτός από το να εντοπίσουμε αν η μεροληψία μεταδίδεται στις ενσωματώσεις, αν ενδεχομένως ενισχύεται από αυτές, για αυτό και χρησιμοποιούμε μία μέθοδο βασισμένη σε πρόβλεψη. Μια ακόμα ενδιαφέρουσα μελέτη που κάνουμε είναι όσον αφορά τη σχέση της δημοφιλίας και της ανισότητας στον πληθυσμό δύο ευαίσθητων τιμών, όπως αντρών – γυναικών στα δεδομένα, με τη μεροληψία στις ενσωματώσεις, αλλά και αν άλλες εφαρμογές όπως η συσταδοποίηση επηρεάζονται από τη μερεοληψία αυτή. Επειδή τα αποτελέσματα επιβεβαιώνουν την ανησυχία μας όσον αφορά τη μεροληψία και επειδή οι ενσωματώσεις των γραφημάτων γνώσης χρησιμοποιούνται ευρέως σε πολλές σημαντικές εφαρμογές, κρίνουμε στη συνέχεια αναγκαίο τον ορισμό μίας μεθόδου αφαίρεσης της πληροφορίας αυτής από τα διανύσματα. Η καινοτομία του έγκειται στη δυνατότητα επιλογής της ποσότητας της μεροληψίας που αφαιρείται και στη χρήση προ-εκπαιδευμένων ενσωματώσεων αντί της τροποποίησης του μοντέλου παραγωγής τους.

Για να εξετάσουμε αν ισχύουν οι ισχυρισμοί μας, αλλά και για να αξιολογήσουμε τις μεθόδους μας χρησιμοποιούμε δύο πολύ γνωστά γραφήματα γνώσεων, το Wikidata και το FB13, και ένα σύνολο από συνθετικά γραφήματα. Χρησιμοποιούμε επιπλέον ενσωματώσεις που παράγουμε μέσω δύο διάσημων μοντέλων, του TransE και του TransH. Στην εργασία παρουσιάζουμε αναλυτικά όλα τα αποτελέσματα και τα συμπεράσματα από τα πειράματα μας. Είναι σημαντικό ότι οι μέθοδοι που προτείνουμε μπορούν εύκολα να επεκταθούν και να χρησιμοποιηθούν και σε άλλα γραφήματα, και σε ενσωματώσεις παραγόμενες από άλλα μοντέλα.

Chapter 1

Introduction

1.1 Motivation

1.2 Structure

1.1 Motivation

In the era of big data, knowledge graph embeddings are gaining more and more interest in the scientific world, as they can represent complex data using vectors of low dimension. Knowledge graph embeddings come from knowledge graphs (KG), that are multi-relational directed graphs widely used to represent knowledge of real-world entities in the form of triples. Each edge represents a specific relation between the two entities (nodes) it connects. Knowledge graphs are used in many applications including question answering, recommendation and information retrieval. They used also by famous search engines and social networks, such as Google and Facebook. KG embeddings map the components of a knowledge graph, such as its entities and relations, to some low dimensional vector space (e.g., [2, 3, 4]). They have gained a lot of attention, since they are useful in a variety of tasks such as KG completion and relation extraction. Furthermore, vector operations are simpler and faster than the corresponding operations on graphs.

But there is a connection of the embeddings production with machine learning tasks. This combined with the fact that as machine learning algorithms are increasingly being used, concerns about unfairness and bias in the treatment of sensitive

attributes are raised (see e.g., [5] for a survey), leads to worry about bias that may exists in data, transferred to the embeddings through machine learning and possibly reinforced by them. A recent work in word embeddings has shown that they encode various forms of social biases [6]. Bias is, generally, disproportionate weight in favor of or against an idea or thing and can have devastating effects on link prediction or recommendation systems. It could make easiest for a system to consider that a man has as occupation the occupation ”computer programmer” for example, than a woman, or depict mostly males in Google’s image search for ”CEO” [7]. So, as embeddings are used in various important applications, it is important to discover if they contain biases.

In this thesis we study bias in KG embeddings. More specifically, we start by proposing two metrics for measuring bias in the dataset. Since in most cases, knowledge comes from the real world, it is expected to reflect existing social biases [8, 9]. After the comparison of the metrics, we choose the one we consider more appropriate. Then we define a method based on projections and an analogy puzzle, to determine quantitatively and qualitatively if the bias is transferred from the data to the KG embeddings. In contrast with [1] analysis, we measure the bias in the data and compare it with the bias in the KG embeddings, we propose a method to measure the bias in KG embeddings, that is more simple and efficient, and it is also include the information of popularity of the entities we are interested in, such as occupations. We also study if the bias not only transferred to the KG embeddings, but also if it is amplified by them and if the inequality in populations of the sensitive values affects the bias. We then apply a prediction method to study if there is in addition a bias amplification using the KG embeddings. We further detect if the popularity of some entities, or the inequality in populations of sensitive values like male, female individuals in the dataset, affects bias in KG embeddings, and, moreover, if other tasks such as clustering is affected by the bias of the KG embeddings. Finally, we propose our debias approach for removing gender bias in occupation embeddings. In relation to [10], its novelty lies on tuning the amount of bias it removes using pretrained embeddings instead of modifying the KG embedding model and on removing information from the embeddings of occupations instead of individuals.

We present experimental results using a Wikidata subgraph, the FB13 dataset and also a set of synthetic graphs. For the KG embeddings, we use TransE and TransH models. In our experiments we focus on bias in occupations. Our results

using different sensitive attributes show that exists bias in occupations in the dataset and data bias is correlated with embedding bias, as well as in many cases the bias is amplified by the embeddings. They also show that popularity of occupations and inequality in the population of sensitive values, like males – females, affect bias in KG embeddings. Moreover there is influence of bias in other tasks, such clustering. Our debias approach seems to reduce the bias with a small penalty on accuracy.

1.2 Structure

This thesis contains $viii$ chapters and it is structured as follows. In section 2, we present some important preliminaries for the knowledge graphs and the knowledge graph embeddings. In section 3 we introduce our metrics for measuring bias in the dataset and in section 4 the methods for measuring bias in the KG embeddings. In section 5, we present our debias approach, and in section 6, we describe the datasets we used and give experimental results. Finally, in sections 7 and 8, we present related work and conclusions, respectively.

Chapter 2

Preliminaries

2.1 Knowledge Graph

2.2 Knowledge Graph Embeddings

2.3 Knowledge Graph Embedding Models

2.4 Notations

In this chapter we present some preliminaries that will be needed later. We start by defining what a knowledge graph is. Then, we explain the knowledge graph embeddings and their usefulness. Finally, we present two knowledge graph embedding models, used for the production of knowledge graph embeddings. These are the TransE and the TransH models, which we use in our experiments.

2.1 Knowledge Graph

A knowledge graph (KG), also known as a knowledge base, contains information about entities, such real-world objects, events, and situations, and the relations between them. The entities are the nodes of the graph, while the relations are the edges. The information is represented in the form of (e_1, r, e_2) triples, where e_1, and e_2 are entities and r the relation between them. For example in the knowledge graph of Figure 2.1, such a triple is the *("Eiffel Tower", "is located in", "Paris")* triple, where *"Eiffel Tower"* and *"Paris"* are the entities and *"is located in"* is the relation between

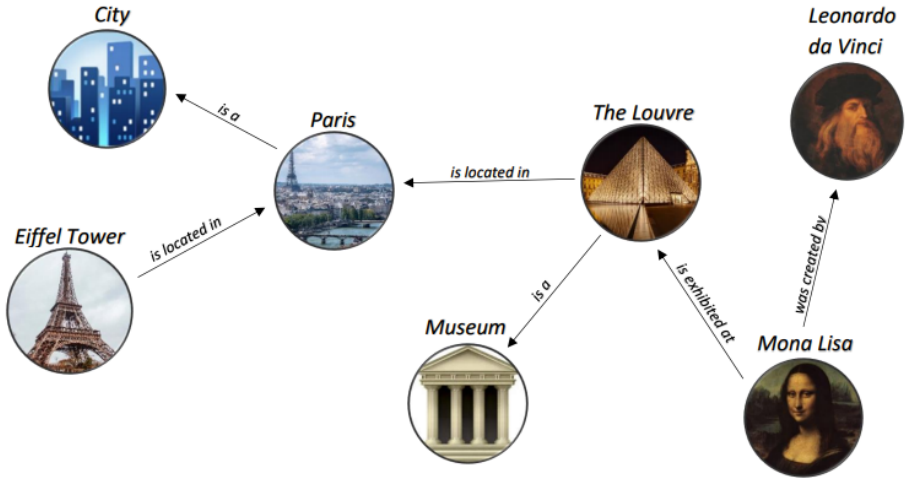


Figure 2.1: Example of a knowledge graph.

them or the *("Mona Lisa", "was created by", "Leonardo da Vinci")* triple, where *"Mona Lisa"* and *"Leonardo da Vinci"* are the entities and *"was created by"* is the relation between them.

Knowledge graphs are useful, because they combine characteristics of:

- Database, because the data can be explored via structured queries,
- Graph, because they can be analyzed as any other network data structure,
- Knowledge base, because they can be used to interpret the data and infer new facts.

Knowledge graphs are used in many tasks in recent years, including question answering, recommendation and information retrieval. They are associated with, and used by search engines such as Google, Bing, and Yahoo; knowledge-engines and question-answering services such as WolframAlpha, Apple's Siri, and Amazon Alexa; and social networks such as LinkedIn and Facebook.

Some of the biggest knowledge graphs are these of Microsoft (KG for the Bing search engine, LinkedIn data and Academics), ebay (KG that gives relationships within users and products, provided on the website), Google (KG for the categorization function across Google's devices and for the Google search engine) and Facebook (KG for making connections among people, events, ideas and news). There are also free and open knowledge graphs, like the Wikidata [11] (KG for Wikipedia and other Wikimedia sister projects) or the Freebase [1] [12] (KG that aims to create a global

[1]The Freebase was officially closed on May 2, 2016 and the data was transferred to Wikidata

resource that allow people (and machines) to access common information more effectively.).

2.2 Knowledge Graph Embeddings

The large volume of data, usually makes the traditional graph structure of the knowledge graphs hard to manipulate. For this reason, the knowledge graph embeddings are used instead.

KG embeddings encode entities and relations of a KG into vectors in some low dimensional space, while preserving the information of the KG, see Figure 2.2.

These vectors (embeddings) can be used for a wide range of tasks such as classification, clustering, link prediction, and visualization. They make it easier to do machine learning on large inputs and also vector operations are simpler and faster than comparable operations on graphs. This make's them also a useful tool for social network analysis.

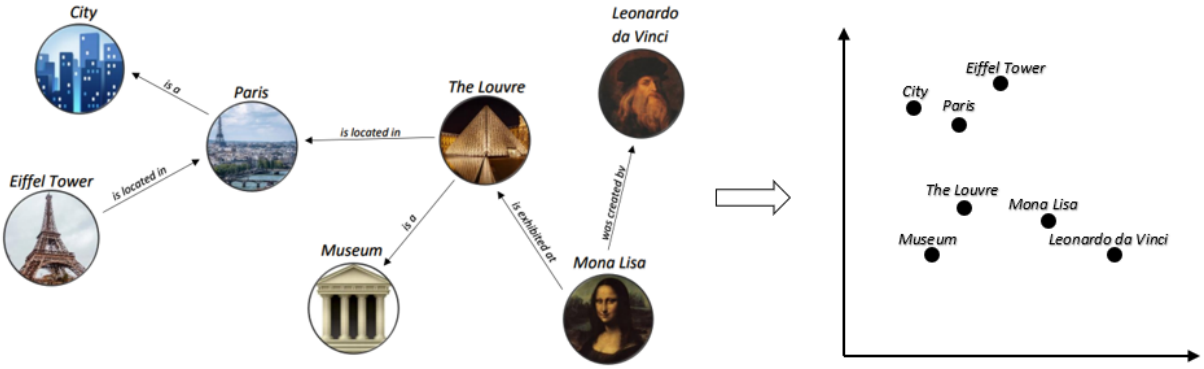


Figure 2.2: Example of KG embeddings representation in the 2-dimensional vector space.

2.3 Knowledge Graph Embedding Models

KG embeddings can be generated using a KG embedding model. A variety of KG embedding models have been proposed to embed both entities and relations in knowledge graphs into a low-dimensional vector space.

In general, if T is a set of triples, then a KG embedding model define a score function $f(e_1, r, e_2)$ for each triple $(e_1, r, e_2) \in T$. The score function $f(e_1, r, e_2)$ returns a higher score, if the triple (e_1, r, e_2) is true, than vice versa. Some KG embedding models use a margin-based loss as the training objective to learn embeddings of the entities and relations, while some others cast the training objective as a classification task. The embeddings of the entities and relations can be learned by minimizing the regularized logistic loss. The main difference among various KG embeddings models is the score function they use.

Two widely known KG embedding models, which we also use in our experiments, are the TransE [3] and the TransH [4] models. They both use a margin-based loss L, based on the score function of each model, as the training objective to learn the embeddings:

$$L = \sum_{(e_1,r,e_2)\in T} \sum_{(e_1,r,e_2)'\in T'} [\gamma + f(e_1, r, e_2) - f(e_1, r, e_2)'], \tag{2.1}$$

where $\gamma > 0$ is a margin hyperparameter and T' denotes the set of non-existing triples, which is constructed by corrupting entities and relations in the existing triples.

2.3.1 The TransE model

TransE [3] is the most representative translational distance model. The model represents the entities and the relations of the KG as translation vectors in the same embedding space. Specifically, every entity such "Paris" or "City" and every relation such "is a" is represented as a d-dimensional vector, which includes the information of this entity or relation in the graph, and is now used instead of the entity or the relation. Given a triple (e_1, r, e_2), let $\vec{e_1}$, $\vec{e_2}$ be the embeddings of entities e_1 and e_2, respectively, and $\vec{r}$ the embedding of the relation r. The idea in TransE is that $\vec{e_1} + \vec{r}$ should be near to $\vec{e_2}$, if (e_1, r, e_2) holds and $\vec{e_1} + \vec{r}$ should be far away from $\vec{e_2}$, otherwise. Figure 6.1a gives an intuition, that is if, for example, (''$Mona\ Lisa$'', ''$was\ created\ by$'', ''$Leonardo\ da\ Vinci$'') triple holds, then using Transe model, ''$\vec{Mona\ Lisa}$''+''$\vec{was\ created\ by}$'' should be near to ''$\vec{Leonardo\ da\ Vinci}$''. Specifically, the score function of the TransE model is:

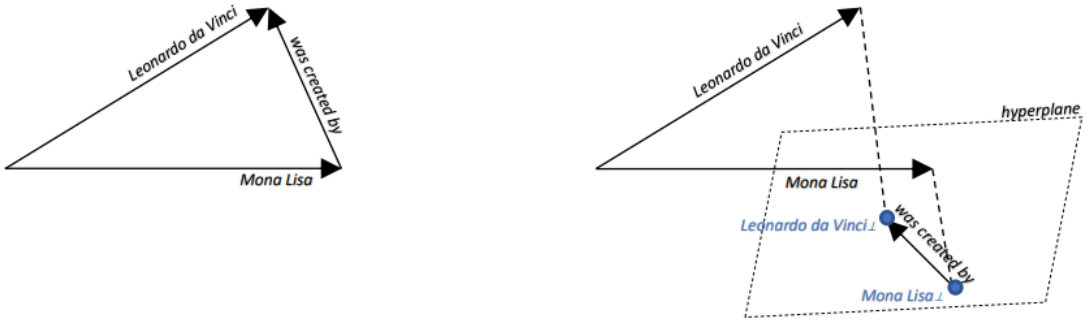
$$f(e_1, r, e_2) = -||\vec{e_1} + \vec{r} - \vec{e_2}||_{1/2} \tag{2.2}$$

and the score is high if (e_1, r, e_2) holds, while it is low otherwise.

2.3.2 The TransH model

TransH [4] is an also widely used KG embedding model. The difference of TransE is that it introduces relation-specific hyperplanes. Specifically, in TransH the representation of the entity is directly related to the relation to which it is associated. Given a triple (e_1, r, e_2), the entity representations $\vec{e_1}$ and $\vec{e_2}$ are first projected onto the r-specific hyperplane. Specifically, it is not used the d-dimensional embeddings of the entity "Paris", for example, but the projection of it on the hyperplane based on the d-dimensional embedding of the relation we are interested in, like the "is a" relation. The projections of the entities are then assumed to be connected by $\vec{r}$ on the hyperplane with low error if (e_1, r, e_2) holds, i.e., $\vec{e_{1\perp}} + \vec{r} \approx \vec{e_{2\perp}}$. $\vec{e_{1\perp}}$ and $\vec{e_{2\perp}}$ are the projection embeddings of entities e_1 and e_2 on the relation-specific hyperplane and $\vec{r}$ is the embedding of a relation r. In Figure 6.1b, for example, if (''$Mona\ Lisa$'', ''$was\ created\ by$'', ''$Leonardo\ da\ Vinci$'') triple holds, then using the TransH model, ''$\vec{Mona\ Lisa_\perp}$'' + ''$\vec{was\ created\ by}$'' should be near to ''$\vec{Leonardo\ da\ Vinci_\perp}$''. So, the score function of the TransH model:

$$f(e_1, r, e_2) = -||\vec{e_{1\perp}} + \vec{r} - \vec{e_{2\perp}}||_2, \tag{2.3}$$



(a) TransE (b) TransH

Figure 2.3: If (‘‘$Mona\ Lisa$”, ‘‘$was\ created\ by$”, ‘‘$Leonardo\ da\ Vinci$”) triple holds, then using Transe model ‘‘$\vec{Mona}\ Lisa$” + ‘‘$was\ \vec{created}\ by$” should be near to ‘‘$\vec{Leonardo}\ da\ Vinci$” and using TransH model, ‘‘$\vec{Mona}\ Lisa_{\perp}$” + ‘‘$was\ \vec{created}\ by$” should be near to ‘‘$\vec{Leonardo}\ da\ Vinci_{\perp}$”.

is high if (e_1, r, e_2) holds, while it is low otherwise.

2.3.3 Comparison of the two models

TransE is simple and efficient model, but it has flaws in dealing with 1-to-N, N-to-1, and N-to-N relations [2]. For example, if “Leonardo da Vinci has occupation painter” and also “Leonardo da Vinci has occupation engineer”; for 1-to-2 relation, then $\vec{Leonardo}\ da\ Vinci + \vec{has\ occupation}$ must be close to $\vec{painter}$, but it must also be close to $\vec{engineer}$. This means that for $has\ occupation$ relation, $engineer$ and $painter$ are similar, but this is not true. The problem gets worse as the tail entities increase. To overcome the disadvantages of TransE, TransH introduces relation-specific hyperplanes. Specifically, it enables an entity to have distributed representations when involved in different relations.

2.4 Notations

From here on we will use e_1 for the head entity of a triple, e_2 for the tail entity and r for the relation. We will also use $\vec{e}$ for the TransE embedding of an entity e and $\vec{e_{\perp}}$ for the TransH embedding of an entity e.

Chapter 3

Bias in the dataset

3.1 Metrics for measuring bias in the dataset

In this chapter we present some metrics for measuring bias in the dataset. We expect that the dataset contains bias, because the data come from real life, where bias and inequality are a constant problem. To find out if the bias exists and to quantify this bias, we consider two approaches.

3.1 Metrics for measuring bias in the dataset

Let s be a sensitive attribute (e.g. $gender$) and a, b two sensitive values (e.g. $male$, $female$). Let's also consider, that we want to discover if there is bias in occupations, because of the sensitive attribute s. To do this, we consider two approaches: we (a) look at both the expected number and the actual number of a-value entities (e.g $male\ entities$) and b-value entities (e.g. $female\ entities$) (equal opportunity approach), and (b) look only at the actual number of a-value entities and b-value entities (equal number approach).

3.1.1 Equal opportunity approach

Definition 3.1. We want the probability that an individual has a specific occupation o to be the same for entities with a value and entities with b value, or, in other words,

we ask that the probability that an individual has occupation o does not depend on the sensitive attribute s of the person.

$$P(O = o|S = a) = P(O = o|S = b) \tag{3.1}$$

3.1.2 Equal number approach

Definition 3.2. Given an occupation o, we ask that it is equally probable that the individual who has this occupation has value a or b.

$$P(S = a|O = o) = P(S = b|O = o) \tag{3.2}$$

3.1.3 Measuring bias in data with the two approaches

From the Bayes rule

$$P(S = a|O = o) = \frac{P(O = o|S = a)P(S = a)}{P(O = o)} \tag{3.3}$$

$$P(S = b|O = o) = \frac{P(O = o|S = b)P(S = b)}{P(O = o)} \tag{3.4}$$

The two definitions are equivalent if the numbers of entities with a value and entities with b value in the population are equal.

To estimate them, let,

$$P(S = a|O = o) = \frac{A_o}{A_o + B_o}, \quad P(S = b|O = o) = \frac{B_o}{A_o + B_o},$$
$$P(S = a) = \frac{A}{A + B}, \quad P(S = b) = \frac{B}{A + B}, \quad P(O = o) = \frac{A_o + B_o}{A + B},$$
$$P(O = o|S = a) = \frac{A_o}{A}, \quad P(O = o|S = b) = \frac{B_o}{B},$$

where A, B are the individuals have the values a, b, respectively. Also, A_o, B_o are the a-value individuals and the b-value individuals have an occupation o, respectively.

So, we can measure how much biased an occupation is, by taking the difference or the ratio.

Definition 3.3 (difference for **equal opportunity**).

$$\phi_o = P(O = o|S = a) - P(O = o|S = b) = \frac{A_o}{A} - \frac{B_o}{B} \tag{3.5}$$

Definition 3.4 (ratio for **equal opportunity**).

$$\phi_o = \frac{P(O = o|S = a)}{P(O = o|S = b)} = \frac{\frac{A_o}{A}}{\frac{B_o}{B}} = \frac{A_o B}{B_o A} \tag{3.6}$$

Definition 3.5 (difference for **equal number**).

$$\phi_o = P(S = a|O = o) - P(S = b|O = o) = \frac{A_o - B_o}{A_o + B_o} \tag{3.7}$$

Definition 3.6 (ratio for **equal number**).

$$\phi_o = \frac{P(S = a|O = o)}{P(S = b|O = o)} = \frac{A_o}{B_o} \tag{3.8}$$

In both *difference* definitions, $\phi_o \in [-1, 1]$. We call occupations for which $\phi_o > 0$ ***"a biased"***, occupations for which $\phi_o < 0$ ***"b biased"*** and occupations for which $\phi_o = 0$ ***"neutral"***. Also, in both *ratio* definitions, $\phi_o \in [0, \infty]$. We call occupations for which $\phi_o > 1$ ***"a biased"***, occupations for which $\phi_o < 1$ ***"b biased"*** and occupations for which $\phi_o = 1$ ***"neutral"***. To have the same range, $[-1, 1]$, in *difference* and *ratio* definitions, we can make the transformation:

- if $\phi_o < 1$, then $\phi'_o = \phi_o - 1$,
- if $\phi_o > 1$, then $\phi'_o = 1 - \frac{1}{\phi_o}$,
- if $\phi_o = 1$, then $\phi'_o = 1 - \phi_o$

In addition to occupations that have a zero score, we consider also neutrals those occupations that belong to a range close to zero. This range is defined by a constant t. To set threshold t, we plot the distribution of neutral occupations for different values of t and select the value of t for which there is a sharp increase in the number of neutral occupations in the plot.

Chapter 4

Bias in knowledge graph embeddings

4.1 Methods for measuring bias in knowledge graph embeddings

4.2 Clustering fairness using knowledge graph embeddings

In this chapter we present some methods to study first, if the bias information is transferred from the data to the KG embeddings and then to detect if there is amplification of bias using the embeddings. We also present an approach to recognize if clusters produced by the KG embeddings are fair or the bias affects the clustering procedure.

4.1 Methods for measuring bias in knowledge graph embeddings

We define here a method based on projections on a bias direction and an analogy puzzle to detect if KG embeddings contain bias. We also present a prediction method to examine bias augmentation using the embeddings.

4.1.1 Projection method

The idea in this method is that there is a bias direction, which includes the information of the bias. For example if gender is the sensitive attribute, then we consider that exists a gender bias direction with the gender information. We explain below how we choose this direction.

Given the bias direction, the projection of an occupation embedding on this direction gives the bias information, which is included in the embedding. Moreover, the norm of the projection shows how much bias information the embedding contains. So the greater the norm of this projection is, the more bias information the embedding includes.

Let $\vec{o}$ be the embedding of an occupation o, and $\vec{d}$ the bias direction. Then, the quantity of bias, which contains the occupation embedding, is:

$$||\pi_{\vec{d}}\, \vec{o}||, \tag{4.1}$$

where $\pi_{\vec{d}}\vec{o} = \vec{o} \cdot \vec{d}$ is the projection of $\vec{o}$ onto $\vec{d}$.

Bias direction

To define the best gender bias direction in [6] using word embeddings, they use the embeddings of pairs of words showing gender bias, like (male, female), (father, mother), (brother, sister), etc. Then, they use SVD to the vector resulting from the differences of the embeddings of the bias pairs. So the top-singular vector is the bias direction for the word embeddings. We use this idea, but knowledge graph embeddings are different from word embeddings. The gender information, for example, in word embeddings, learned from the sentences in the text. On the other hand in KG embeddings there are triples and specific entities that express the bias. So if we care about the gender bias, as before, the entities male and female contain all the gender bias information, because all the males are connected with the male entity with triples ("male entity", "has gender", "male") and all the females are connected with the female entity with triples ("female entity", "has gender", "female"). So we conclude that the best choice for the bias direction is the subtraction of the male, female entity embeddings ($\vec{female} - \vec{male}$). We do not need to take SVD, because we have only one pair of entities, that clearly expresses the bias. The only different for the embeddings produced by the TransH model is that they are projections on the relation-specific hyperplane, "has gender" in our case. So the direction, taking into account this information, will be ($\vec{female}_{\perp} - \vec{male}_{\perp}$). In general, for a sensitive attribute s and two sensitive values a and b the bias direction using TransE KG embeddings is:

$$\vec{d} = \vec{b} - \vec{a} \tag{4.2}$$

and the bias direction using TransH KG embeddings is:

$$\vec{d} = \vec{b_\perp} - \vec{a_\perp} \tag{4.3}$$

Comparison of our embedding bias metric with an other existing metric

The only other metric for measuring bias in KG embeddings is this in [1]. Using this metric, if $gender$ is the sensitive attribute and $male$, $female$ the two sensitive values, then the entity embeddings are first transformed to be "more male" according to the model's own encoding of masculinity. Then using all the individuals in the dataset, they analyse whether, making an entity more male increases or decreases the likelihood that they have an occupation, according to the model's score function.

So the first difference with our metric is that they use a transformation of the individuals' embeddings to measure the bias in occupations. On the other hand we use only the embeddings of the occupations and a bias direction to quantify the bias information, that is more simple but also effective. Another difference is that they use the whole number of individuals in the dataset to separate the occupations into male and female occupations. This is particularly time consuming especially in large datasets, where there are millions of individuals and occupations. Instead of this we suggest to use the similarity with the sensitive values (male, female) to separate the occupations. This is based on the idea that the information which creates the bias is this of the sensitive values that is included in the embedding. For example, doctor is considered as male occupation because the "male information" is included in the embedding of doctor. Also all the embeddings of the male individuals contain the male information. So we can use instead of the male individuals, which can be millions, the male entity and respectively, instead of the female individuals, the female entity. The last difference concerns the correlation of bias with popularity, and is analyzed in section 6 with experimental evaluation. Our metric, in contrast with the metric in [1], includes the popularity information in bias score and this is important because, as we show later, popularity affects bias.

4.1.2 Analogy puzzle method

Inspired from [6] and word embeddings, we use an analogy puzzle to detect if KG embeddings contain bias. To do this we try to find pairs of occupations, which are close to the parallel to a seed direction. The seed direction in our case is the bias

direction $\vec{d}$, as defined in 4.1.1. To define the best analogy occupation pair (x, y) the cosine similarity score:

$$cs\_score_{(a,b)}(x, y) = cos(\vec{a} - \vec{b}, \vec{x} - \vec{y}) \quad if \ \ ||\vec{x} - \vec{y}|| < \delta, \quad 0 \ \ otherwise, \tag{4.4}$$

is used. Threshold δ is a similarity threshold (we use $\delta = 2$), so as to select only semantically related x and y occupations. An example analogy pair in [6] is the *(computer programmer, homemaker)* occupation pair for the *(man, woman)* seed pair, which clearly shows bias in word embeddings.

4.1.3 Prediction method

The next method is based on a prediction task and aims to determine if there is an augmentation of bias using the KG embeddings. The idea is the following. Given an occupation, we use the embeddings to predict which individuals are more likely to have this occupation. Then, we compare the percentage of a-value (resp., b-value) individuals in the top-K predictions with the percentage of a-value (resp., b-value) individuals expected to have the specified occupation. If the predicted percentage is higher than the expected, we assume that there is an increase in bias.

For the TransE KG embeddings, to predict the individuals most likely to have a given occupation o, the score function of TransE 2.2 used. Respectively, for TransH KG embeddings the score function of TransH 2.3 used.

In both TransE and TransH KG embeddings, we consider as top-K, the individuals e with the top K scores. For two a, b sensitive values, the estimated number of a-value entities in the top-K results is $\frac{|A_o|\,K}{|A_o|+|B_o|}$, and the estimated number of b-value entities is $\frac{|B_o|\,K}{|A_o|+|B_o|}$. We compare the predicted and the estimated by the dataset percentage of a-value (resp., b-value) individuals, for different values of K.

4.2 Clustering fairness using knowledge graph embeddings

Cluster analysis or simply clustering, as defined in [13], is the process of partitioning a set of data objects (or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. Cluster analysis has been widely used in many applications such as business intelligence, image pattern recognition, Web search, biology, and security.

In knowledge graphs, clustering is often used for feature engineering in modern machine learning systems, like giving each example in the dataset with the id of the cluster it belongs to in an effort to bring expressive power to simple learning methods. Another application is for data reduction, where the cluster representations of the data can be used to replace the actual data. So it is important to study if the clusters produced using the KG embeddings are fair.

4.2.1 Clustering fairness using knowledge graph embeddings

Our approach for clustering follows the one in [14]. After training the dataset and produce the KG embeddings, we use a clustering algorithm to create clusters of the KG embeddings of the entities. In our experiments we use the K-Means clustering algorithm. K-means aims to partition n observations into K clusters in which each observation belongs to the cluster with the nearest mean.

We want the clusters produced using the KG embeddings to be fair and not reinforce biases present in data. Specifically, if a, b are two sensitive values, we want the number of a-value and b-value entities in each cluster is proportional to the data and do not exist clusters with a big increase of the number of one category. We want also this to apply for the occupations of each category in the dataset. So, to detect if there are bias clusters, we compare the number of a-value in relation to the number of b-value individuals in the dataset, with this in each cluster and also the number of b-value in relation to the number of a-value individuals in the dataset, with this in each cluster, let be $\frac{A}{B}$ and $\frac{B}{A}$, respectively. We further compare the number of a-value in relation to the b-value entities have an occupation o in the dataset, with this in each cluster, and vice versa, let be $\frac{A_o}{B_o}$ and $\frac{B_o}{A_o}$, respectively.

We also study if the clustering of the occupations is based on the semantics of the occupations or they are affected mainly by the bias. For this reason, if o_1 and o_2 are two occupations in the cluster, we calculate the $||\vec{o_1} - \vec{o_2}||$ for the TransE embeddings, and the $||\vec{o_{1\perp}} - \vec{o_{2\perp}}||$ for the TrasnH embeddings. We repeat for all the occupations in the cluster and take the average. We expect that occupations that are related semantically have very low score, close to zero, while the more unrelated the occupations are, the higher the score they have is. The idea is inspired from analogies in [6].

Chapter 5

Debias the knowledge graph embeddings

5.1 Debias approach

Machine learning is gaining more and more use in everyday life. But there are spectacular failures of learning models, such as depicting mostly males in Google's image search for "CEO" [7] or recommending doctor as a career choice for men and nurse for women. Since KG embeddings are used in a large number of applications like search engines, recommendation or link prediction systems, the need for fairness becomes imperative. For this reason, in this chapter we propose a debias approach for removing bias from the embeddings by extracting the information of the sensitive attribute, which creates the bias.

5.1 Debias approach

Our debias approach is based on [6, 15] and, in contrast with [10], we use the pretrained embeddings and do not change the model. We also remove information from the embeddings of occupations, while in [10] they remove information from the embeddings of individuals, which may have a negative effect in cases where we want to know the sensitive attribute.

Our approach uses the projection on the bias direction ($\vec{d}$), as defined in 4.1.1. More specifically, the intuition is that since the projection of a KG embedding on

the bias direction gives the bias information, we can extract this information from the original KG embedding to produce a debias one. Because many times we do not want to lose the whole information, we add a tuning parameter λ to define how strong the debias we want to be.

So if $\vec{o}$ is the embedding of an occupation o, then the new debias embedding is:

$$\vec{o'} = \vec{o} - \lambda \pi_{\vec{d}} \, \vec{o} \tag{5.1}$$

If parameter $\lambda = 1$, then the debias is "hard" and the whole gender information is extracted, while if parameter $\lambda = 0$, then there is no debias.

Chapter 6

Experiments

6.1 Datasets

6.2 Bias in the dataset

6.3 Bias results in the KG embeddings

6.4 Clustering results

6.5 Debias evaluation

In this chapter, we first describe and analyze the datasets we use. Then we compare our metrics defined in 3.1, and select the appropriate metric for measuring bias in the datasets. We present the results using this metric and we also study if the size of the population of each sensitive value or the popularity of some occupations affects bias. We also apply our methods for measuring bias in KG embeddings and analyze if the bias is transferred from the data to the KG embeddings and if it is amplified by them. Moreover, we discover if using the KG embeddings produced unfair clusters. Finally, we analyze and evaluate the results of our debiasing approach.

6.1 Datasets

In our experiments we use synthetic and real datasets. Specifically, we take the average results of a set of synthetic knowledge graphs, and we also use a subgraph of the Wikidata dataset [11] and the FB13 dataset, produced by Freebase [12], as real

datasets. We explain how we create the synthetic knowledge graphs and how we take the subgraph below.

6.1.1 Synthetic knowledge graph

Our synthetic graphs focus on two types of triples, those with a sensitive attribute, let be gender, and those with occupations. The produced triples are, for example, in form ("individual", "has occupation", "occupation") and ("individual", "has gender", "male") or ("individual", "has gender", "female"). We generate n_o occupations and n_i individuals. Let I be the set of individuals and O be the set of occupations, where $|I| = n_i$ and $|O| = n_o$. Then, we determine the number of individuals having each occupation o using a zipf distribution, with a given parameter a. Let $m(o)$ be the number of individuals having occupation o, then $\sum_{o \in O} m(o) = n_i$. After that, we separate the occupations to male occupations and female occupations and also with a parameter, related to the size of male/female entity population, we separate the individuals to males and females. Finally, we use a parameter to assign occupations to individuals, with a bias probability. Let $p_b \in [0.5, 1]$ be the probability of bias. The value 0.5 expresses no bias and 1 shows absolute bias. For each male occupation o, we assign a male individual u with probability p_b and a female individual u with probability $1 - p_b$ and for each female occupation o, we assign a female individual u with probability p_b and a male individual u with probability $1 - p_b$, until $m(o)$ individuals are assigned to o.

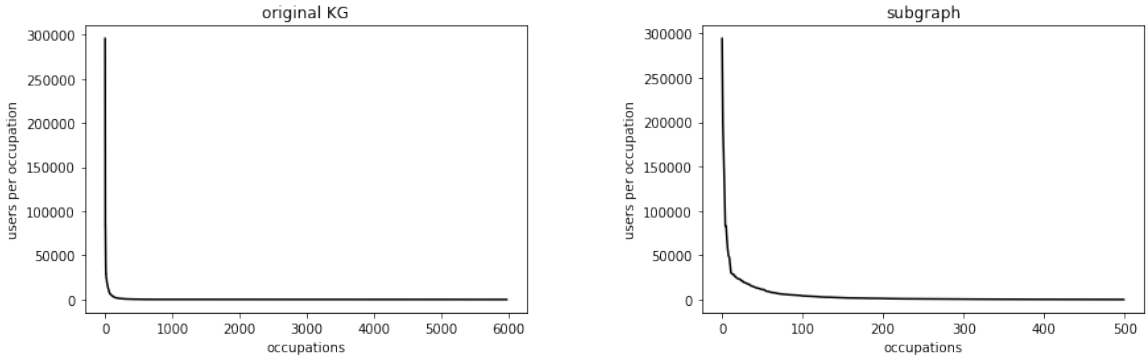
We also separate the triples of synthetic graphs in training and test sets, with 80% of the triples belong to the training set and 20% to the test set. Users and occupations of test set must exist also in training set.

6.1.2 Real knowledge graphs

Subgraph of Wikidata

Due to computational training time and memory, we do not use the real Wikidata dataset[1], which consists of 68,904,773 triples, with 20,982,733 entities and 594 relations. Instead of this we use a subgraph. To take a good subgraph, we first choose the triples which contain occupations, that more than 500 entities have them, with

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



(a) original (b) subgraph

Figure 6.1: There is a power law distribution on the number of individuals per occupation in both the original graph and the subgraph; there are few occupations that many users have and many that few have.

the constraint that must also exist the information of the gender of these entities in the triples. The idea is that we want much information for the occupations, for the training procedure, and also we want to exist the gender information of the entities, to study how this information affects the embeddings. Subsequently, we select all the triples that contain the entities of the previous selected triples, such as to add all the existing information of our entities and their relations. After that we have 16,354,458 triples, with 2,379,295 entities and 268 relations. We also calculate the percentage of the male entities and the female entities in the original graph and in our subgraph. In the original graph there are 83.56% male entities and 16.44% female entities, while in our subgraph there are 83.35% male entities and 16.65% female entities. This combined with the fact that occupations in both graphs follow a power-low distribution, see Figure 6.1, clearly shows that our subgraph is a good representative sample of the original. We also collect 10,129 triples with occupations for our experiments.

FB13 dataset

FB13 is a small subgraph of Freebase, which usually used as benchmark. It consists of 316,232 triples, with 75,043 entities and 13 relations. It also includes 23,733 test triples.

Table 6.1: Information of the real datasets.

	Wikidata subgraph	FB13
Triples	16,354,458	316,232
Entities	2,379,295	75,043
Relations	268	13

6.1.3 Training the knowledge graphs

We use the OpenKe library [16] for the training of the KG embeddings with the TransE and the TransH models, and dimension 100 for the embeddings. All the training procedures are performed on machine with 32GB RAM, GPU NVIDIA GeForce RTX 2080 with CUDA version v.10.1 and CPU Intel® Xeon® E5-2630 v3 @ 2.40GHz. During the training there was a calculation of the training error and the success of prediction in test data so as not to lead to over-training. In the synthetic graphs we used 180,000 triples for training and 20,000 triples for testing, 600 epochs for the TransE model and 400 epochs for the TransH model. The hits@10 of them was between 0.96 and 0.97, both with TransE and TransH models. In the FB13 dataset, we used also 600 epochs for the TransE model and 400 epochs for the TransH model. The hits@10 score in the test triples is 0.78 and 0.75 for the TransE and the TransH, respectively. Finally, in the subgraph of Wikidata, we used 1000 epochs for TransE, 600 epochs for TransH and the hits@(20,50,100) is (0.42, 0.67, 0.87) for the TransE and (0.55, 0.78, 0.91) for the TransH.

6.2 Bias in the dataset

In this section we compare the metrics defined in chapter 3 to select the more appropriate for measuring the bias in the dataset. We also identify if the popularity of some occupations or the inequality in the population of the sensitive values affects bias. Then we produce the data bias results using the selected metric.

6.2.1 Data bias metric selection

In this subsection we compare the metrics defined in section 3.1 to select the appropriate metric to measure the bias in our datasets. The first comparison is between the equal opportunity approach, 3.1.1, and the equal number approach, 3.1.2. The first approach takes into account the population of a sensitive value, while the second does not. Let us make an example to understand this. Let gender be the sensitive attribute, male, female the sensitive values and o an occupation. Let also be 1000 males in the dataset and 100 females, from which 90 males and 80 females have occupation o, respectively. Using the equal opportunity approach with difference, the score is -0.71< 0, so o is considered as female occupation. Respectively, if we use the equal opportunity approach with ratio, with the transformation to range $[-1, 1]$, the score is -0.89< 0, so o is also female. On the other hand, if we use the equal number approach with difference, the score is 0.06> 0, and o is considered as male occupation. Respectively, if we use the equal number approach with ratio, with the transformation to range $[-1, 1]$, the score is 0.11> 0, so o is also male. We observe that there is a significant difference in the separation of occupations into males/females whether we take into account the population or not. Because, as we show bellow, the inequality in populations affects bias, we choose the equal opportunity approach 3.1.1 as metric for measuring bias in the dataset and separate occupations into the sensitive values.

The second comparison is between the difference definition 3.5 and the ratio definition 3.6. In contrast with the ratio, the difference definition takes into account the popularity of an occupation; that is, it gives a higher bias score to the most popular occupations. For example let o_1 and o_2 be two occupations in the dataset, with 200 males and 10 females have the occupation o_1 and 800 males and 10 females have the occupation o_2. Then, for 1000 males and 100 females in the dataset, the difference definition gives 0.1 score to o_1 and 0.7 to o_2, while the ratio definition gives 0.5 score to o_1 and 0.125 to o_2. It is obvious that the most popular occupation, the o_1, has higher score with the difference definition, than the o_2, while the ratio definition gives a higher score to o_2, than to o_1.

So we need to decide if we want to use the information of the popularity in our bias metric. In general, we expect that the information of popularity is useful, because during the training, there are many triples for the popular occupations with

the bias information, and we expect that the KG embeddings of the most biased occupations to be more biased. To study if this truly happens, we follow a prediction task as described in 4.1.3. Specifically, we try to find out if the popularity affects the predictions, by amplifying bias, if:

- the prediction concerns the most biased occupations in relation to all,
- the population of males / females in the dataset is the same or different.

For this reason we create five synthetic graphs with 100,000 users, 200 occupations, bias probability 0.8, a parameter for zipf distribution 1.5 and the same number of males, females. We also create, five synthetic graphs with the same parameters, except the population, which is defined to be the 70% of the total population for males and the 30% for females. For the prediction task, we use 20,000 users from the test triples. For each occupation, we predict the most likely users to have it, and compare the number of males, females in top-K predictions with the expected one. We use TransE and TransH embeddings.

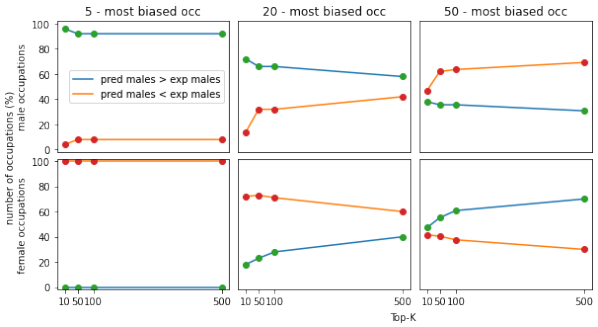
In Figure 6.2 are presented the results of the prediction task using TransE embeddings, equal opportunity approach with difference, equal opportunity approach with ratio, equal and unequal population of males/females. In Figure 6.3 there are also the results of the prediction task using TransH embeddings instead. The top-K results refer to K=10, 50, 100 and 500, while also we plot the results for the 5, 20 and 50 most biased occupations.

Let us first consider what happens when the K value increases. In equal population, using both the equal opportunity with difference and the equal opportunity with ratio approaches, we observe that as K increases, the number of male occupations with bias amplification for males and the number of female occupations with bias amplification for females reduces and respectively, the number of female occupations with bias amplification for males and the number of male occupations with bias amplification for females increases. This is expected, because the best prediction score is for the small top-K values, so we expect the results for small K values to be more representative. For the unequal population, as K increases, the same behavior as before applies to the female occupations for females, but in male occupations there is an increase in the amplification of bias for males. This is a sign that the large number of male population in the dataset maybe affects embedding bias.

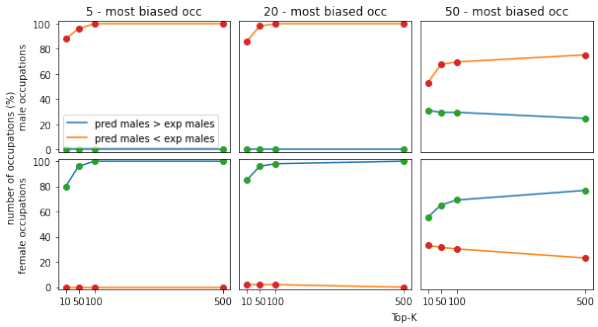
Let us now consider what is happening in the most 5, 20 and 50 biased occupations using TransE embeddings. We observe that for the difference approach, in the most cases the amplification is greater for males than for females in male biased occupations, and the amplification is greater for females than for males in the most female biased occupations. This does not seems to exist with the ratio approach, and shows that popularity is connected with bias in the KG embeddings. Moreover, in the equal population we observe that for the difference approach, as we increase the number of the biased occupations, the percentage of male occupations with male bias amplification, and the percentage of female occupations with female bias amplification is decreased. This shows that the most popular occupations are also the most biased. Given this, it is important to take popularity into account in our bias metric. An also interesting observation is that the inequality in the population affects embedding bias, since in many female occupation there is amplification of bias for males. This shows that also the choice of equal opportunity metric instead the equal number approach is correct, because it takes into account the population of a sensitive value in the dataset.

But before we make a final decision for the bias metric, let us study also the results of the TransH embeddings. We see that for the equal population, the two metrics give similar results, and there is no bias amplification of male occupations for males and of female occupations for females, so we can not draw any conclusions for popularity. But in the unequal population, using the difference approach, we observe that again it seems the most popular occupations to be also the most biased and that the inequality in the population affects bias. These observations do not exist using the ratio approach and shows that bias in TransH embeddings is affected mainly from the population and the popularity.

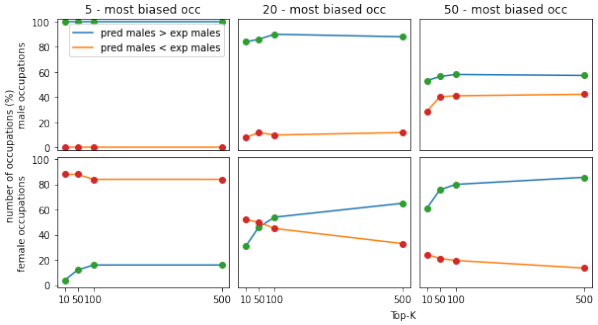
Now we can conclude that the popularity has effect on bias, since the most popular occupations in the dataset have the most biased KG embeddings, and for this reason we must take it into account, to choose our bias metric. Given this, the choice of the equal opportunity approach with difference as data bias metric is the best and for this reason we use this metric in our experiments.



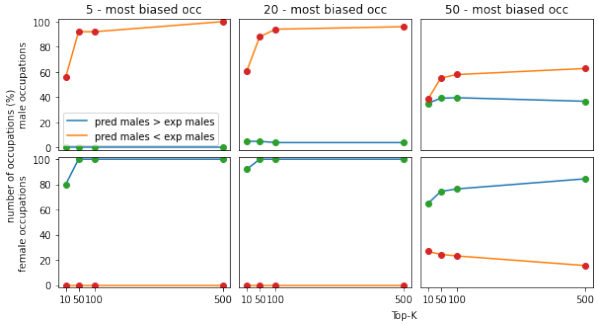
(a) Equal opportunity with difference - Equal male/female population



(b) Equal opportunity with ratio - Equal male/female population



(c) Equal opportunity with difference - Unequal male/female population

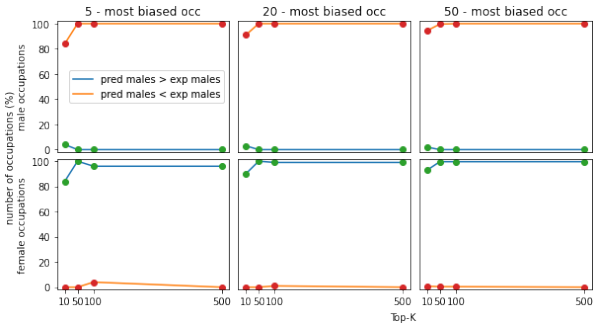


(d) Equal opportunity with ratio - Unequal male/female population

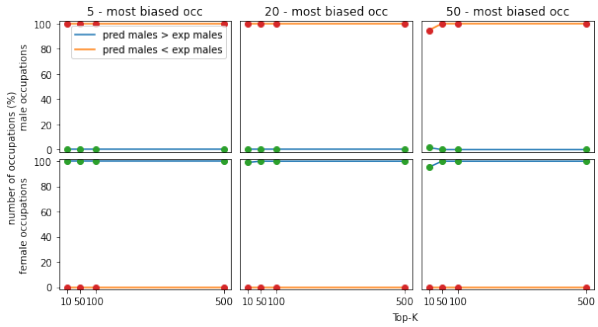
Figure 6.2: Average results of prediction task using five synthetic graphs and TransE embeddings.

6.2.2 Bias results in the datasets

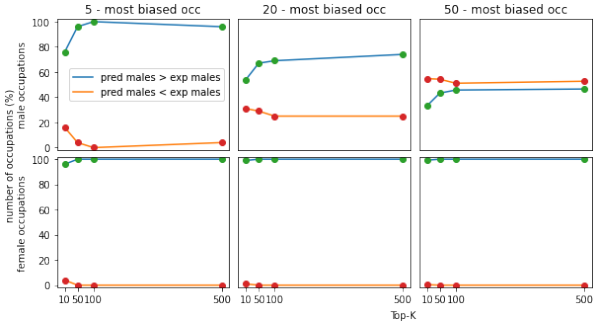
Using the equal opportunity metric as selected in previous subsection, we separate the occupations of our datasets in bias categories based on different sensitive attributes. For the Wikidata subgraph we consider as sensitive attribute the gender and as sensitive values the male and the female. For the FB13 dataset, we consider as sensitive attributes the religion, the nationality and the ethinicity and as sensitive values the Catholicism / Judaism, the United States / United Kingdom and the African American / German, respectively. Some information about the number of individuals and occupations, based on these sensitive attributes, are presented in Table 6.2. As described in chapter 3, to find the appropriate value for threshold t, we plot the distribution of neutral occupations for different values of t and select the value of t where there is a sharp increase in the number of neutral occupations. In Figure 6.4 there are the distributions of the neutral occupations of each sensitive attribute in the datasets and



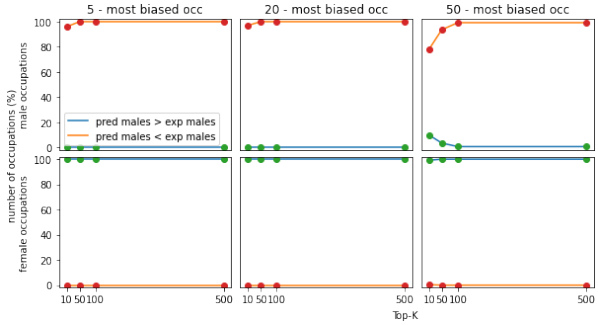
(a) Equal opportunity with difference - Equal male/female population



(b) Equal opportunity with ratio - Equal male/female population



(c) Equal opportunity with difference - Unequal male/female population



(d) Equal opportunity with ratio - Unequal male/female population

Figure 6.3: Average results of prediction task using five synthetic graphs and TransH embeddings.

also the selected value of t in each case.

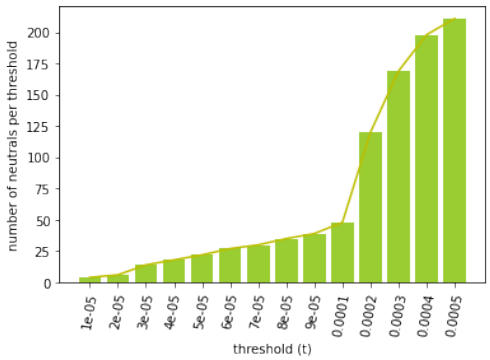
In Table 6.3, there are the five most biased occupations based on gender, religion, nationality and ethnicity sensitive attributes. The gender concerns the Wikidata subgraph, while the rest refer to the FB13 dataset. We observe that occupations like "association football player", "politician" and "lawyer" are characterized as male occupations, while occupations like "actor", "singer" and "model" are characterized as female occupations. This shows that exists gender bias in the dataset. Bias in data is also clear from the results of the FB13 dataset. For example, occupations like "politician" and "lawyer" are Catholicism, while occupations like "scientist" and "banker" are Judaism. Also, "businessperson" or "baseball player" are US occupations, while "mathematician" or "physician" are UK occupations. Moreover, occupations such as "musician", "singer" and "jazz pianist" are African American, while occupations such as "scientist", "philosopher" and "banker" are German.

Table 6.2: Information about the number of individuals and occupations of each sensitive attribute in the datasets. The gender sensitive attribute concerns the Wikidata subgraph, while the religion, nationality and ethnicity sensitive attributes concern the FB13.

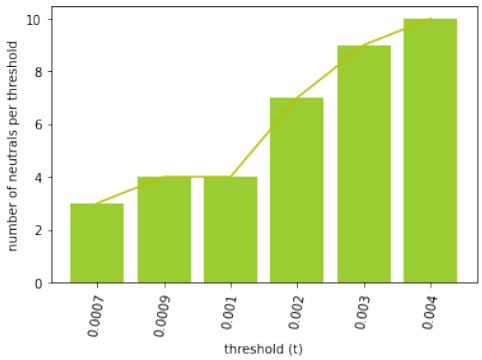
s attribute	a-value	b-value	a-ind	b-ind	a-occ	b-occ	neutrals
gender	male	female	1,681,901	336,187	275	146	79
religion	Catholicism	Judaism	753	522	19	28	4
nationality	US	UK	10,213	1,621	35	23	2
ethnicity	African Am	German	602	166	26	14	6

Table 6.3: The five most biased occupations based on gender, religion, nationality and ethnicity sensitive attributes on Wikidata subgraph and FB13 dataset.

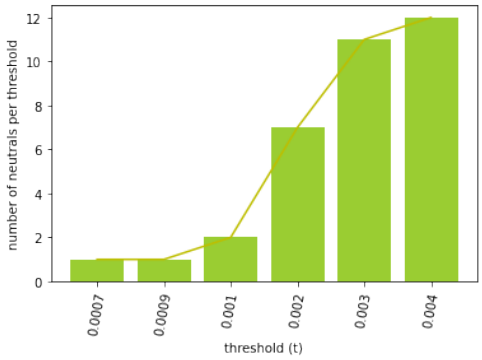
male-occupations (ϕ_o)	**female-occupations** (ϕ_o)	**Catholicism-occupations** (ϕ_o)	**Judaism-occupations** (ϕ_o)
association football player (0.0922)	actor (-0.1556)	pope (0.251)	rabbi (-0.1705)
politician (0.07)	singer (-0.0628)	politician (0.109)	scientist (-0.0828)
university teacher (0.0216)	model (-0.0374)	lawyer (0.048)	physicist (-0.0392)
baseball player (0.0157)	writer (-0.0207)	poet (0.0183)	businessperson (-0.025)
lawyer (0.0149)	television actor (-0.0192)	journalist (0.0157)	banker (-0.0236)
US-occupations (ϕ_o)	**UK-occupations** (ϕ_o)	**African American-occupations** (ϕ_o)	**German-occupations** (ϕ_o)
lawyer (0.0902)	engineer (-0.0484)	musician (0.2697)	scientist (-0.1352)
actor (0.0594)	writer (-0.0379)	singer (0.1674)	philosopher (-0.1248)
businessperson (0.0389)	mathematician (-0.0338)	actor (0.1358)	mathematician (-0.0887)
singer (0.0387)	physician (-0.028)	songwriter (0.0748)	banker (-0.0783)
baseball player (0.0267)	novelist (-0.0277)	jazz pianist (0.0365)	physicist (-0.0706)



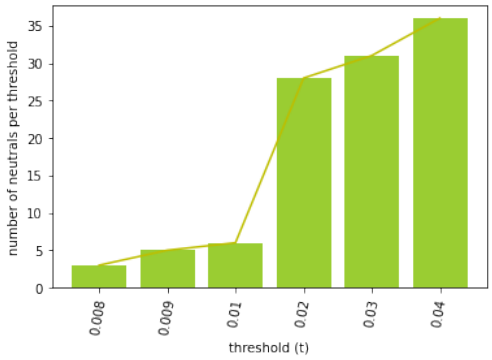
(a) gender - Wikidata subgraph - t=0.0001



(b) religion - FB13 dataset - t=0.001



(c) nationality - FB13 dataset - t=0.001



(d) ethnicity - FB13 dataset - t=0.01

Figure 6.4: Plots of the distribution of neutral occupations for different values of t. The 6.4a refers to Wikidata subgraph and the gender sensitive attribute. The appopriate value for t is 0.0001. The 6.4b refers to FB13 dataset and the religion sensitive attribute. The appropriate value for t is 0.001. The 6.4c refers to FB13 dataset and the nationality sensitive attribute. The appropriate value for t is also 0.001. The 6.4d refers to FB13 dataset and the enthnicity sensitive attribute. The appropriate value for t is here 0.01.

6.3 Bias results in the KG embeddings

As expected there is bias in the real data. In this section we use the methods defined in chapter 4 to identify if the bias is transferred from the data to the KG embeddings and also if it is amplified by them. We first give the most biased occupations using the KG embeddings. We then calculate the correlation between the embedding bias and the data bias metrics. Then, we study the existence of bias amplification using prediction. Finally we examine whether embedding bias affects other common applications, such as clustering, creating inequality and unfairness.

6.3.1 Bias results in the KG embeddings using projections

In Tables 6.4a and 6.4b are presented the most biased occupations using the score of the projection on bias direction, as described in section 4.1.1. For the bias orientation we use the cosine similarity with each sensitive value. We observe from the results that bias information is transferred from the data to the embeddings. In most cases the most biased occupations in the KG embeddings are similar to the most biased occupations in the datasets. Exception are the occupations in Wikidata subgraph using TransE embeddings, which are different from the most biased occupations in the dataset. A reason that explains this phenomenon is that TransE has flaws in dealing with 1-to-N, N-to-1, and N-to-N relations, and due to the large number of relations in the Wikidata subgraph, the information in the embeddings is not so clear.

Correlation between the data bias and the embedding bias metric

To check quantitatively, if the embedding bias metric is related to the data bias metric for all occupations, we measure the correlation between the two metric scores, which also shows if there is a correlation between embedding bias and popularity. We use Pearson correlation coefficient, which search if there is linear correlation between two variables. A value of 1 implies that a linear equation describes the relationship between the two variables perfectly. A value of -1 implies that there is inverse correlation between the two variables. A value of 0 implies that there is no linear correlation between the two variables.

Table 6.5 gives the results of the correlation. In most cases there is linear correlation between the data bias and the embedding bias. This means that the most

Table 6.4: The five most biased occupations based on gender, religion, nationality and ethnicity sensitive attributes on Wikidata subgraph and FB13 dataset. In Table 6.4a there are the results from TransE embeddings, while in Table 6.4b there are the results from TransH embeddings.

(a) Most biased occupations using TransE embeddings.

male-occupations (score)	**female-occupations** (score)	**Catholicism-occupations** (score)	**Judaism-occupations** (score)
bishop (0.3127)	softball player (0.5311)	pope (0.5137)	rabbi (0.3479)
antiquarian (0.3098)	netballer (0.524)	cardinal (0.281)	scientist (0.3411)
archbishop (0.2793)	glamour model (0.476)	soldier (0.2041)	chemist (0.2959)
flying ace (0.2685)	Playboy Playmate (0.4294)	engineer (0.1697)	physicist (0.29)
colonial administrator (0.2600)	fashion model (0.347)	politician (0.156)	psychologist (0.2781)
US-occupations (score)	**UK-occupations** (score)	**African American-occupations** (score)	**German-occupations** (score)
American football player (0.424)	philosopher (0.3058)	musician (0.4263)	philosopher (0.3685)
attorney at law (0.4012)	botanist (0.2945)	songwriter (0.4147)	mathematician (0.348)
baseball player (0.3752)	mathematician (0.2545)	singer-songwriter (0.4133)	scientist (0.3432)
businessperson (0.3722)	astronomer (0.1846)	guitarist (0.397)	rabbi (0.335)
entrepreneur (0.3649)	chemist (0.1824)	jazz pianist (0.3833)	chemist (0.3209)

(b) Most biased occupations using TransH embeddings.

male-occupations (score)	**female-occupations** (score)	**Catholicism-occupations** (score)	**Judaism-occupations** (score)
association football player (0.1766)	actor (0.1852)	pope (0.2870)	scientist (0.2244)
politician (0.1043)	singer (0.0827)	politician (0.1480)	physicist (0.1881)
American football player (0.0549)	athletics competitor (0.0679)	engineer (0.1469)	psychologist (0.1680)
military personnel (0.0533)	model (0.0470)	cardinal (0.1469)	chemist (0.1464)
Australian rules football player (0.0464)	announcer (0.0461)	lawyer (0.1300)	rabbi (0.1443)
US-occupations (score)	**UK-occupations** (score)	**African American-occupations** (score)	**German-occupations** (score)
businessperson (0.1703)	mathematician (0.1868)	musician (0.3203)	physicist (0.2731)
singer (0.1469)	philosopher (0.1860)	singer (0.2996)	mathematician (0.2389)
lawyer (0.1441)	physicist (0.1796)	songwriter (0.2330)	scientist (0.2355)
entrepreneur (0.1366)	engineer (0.1793)	American football player (0.2073)	chemist (0.2283)
musician (0.1319)	botanist (0.1703)	baseball player (0.2049)	philosopher (0.2234)

biased occupations, which are also the most popular occupations in the dataset, have the most biased KG embeddings. Exception are again the occupations in Wikidata subgraph using TransE embeddings. The reason is also here the large number of relations in the dataset. We observe that this does not happens with the TransH embeddings, which deal better with many relations in the dataset.

Table 6.5: Pearson correlation coefficient score between data bias and embedding bias using TransE and TransH embeddings.

(a) Pearson correlation coefficient score using TransE embeddings.

s attribute	a-value	b-value	score a-occ	score b-occ
gender	male	female	-0.14	-0.07
religion	Catholicism	Judaism	0.77	0.55
nationality	US	UK	0.28	0.17
ethnicity	African Am	German	0.35	0.73

(b) Pearson correlation coefficient score using TransH embeddings.

s attribute	a-value	b-value	score a-occ	score b-occ
gender	male	female	0.86	0.87
religion	Catholicism	Judaism	0.82	0.48
nationality	US	UK	0.64	0.59
ethnicity	African Am	German	0.66	0.74

So it is important that our embedding bias metric includes the information of the popularity in bias score. This is a significant difference with the embedding bias metric in [1]. From the results in Table 6.6, we see that in most cases the results using our metric have correlation with the popularity, in contrast with the other metric that in most cases has no correlation.

Bias orientation using similarity

For a sensitive attribute s and two sensitive values a and b, we consider as bias orientation the characterization of an occupation o as a-occupation or b-occupation. To do this we use the cosine similarity measure between the embedding of each occupation with the embedding of the a or the b entity, which defined as

$$cosine\ similarity = \frac{\vec{o} \cdot \vec{e}}{||\vec{o}||||\vec{e}||},$$

Table 6.6: Pearson correlation coefficient score between popularity and embedding bias using our metric, let be proj, the embedding bias metric in [1] and the FB13 dataset.

(a) Pearson correlation coefficient score results using TransE embeddings.

s attribute	a-value	b-value	score a-occ proj	score b-occ proj	score a-occ [1]	score b-occ [1]
religion	Catholicism	Judaism	0.58	0.45	0.0	-0.16
nationality	US	UK	0.11	-0.32	0.04	-0.21
ethnicity	African Am	German	0.25	0.36	-0.03	-0.17

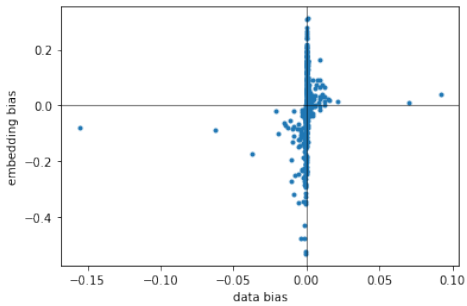
(b) Pearson correlation coefficient score results using TransH embeddings.

s attribute	a-value	b-value	score a-occ proj	score b-occ proj	score a-occ [1]	score b-occ [1]
religion	Catholicism	Judaism	0.73	0.46	-0.28	-0.01
nationality	US	UK	0.53	0.02	-0.23	0.07
ethnicity	African Am	German	0.59	0.46	-0.02	0.19

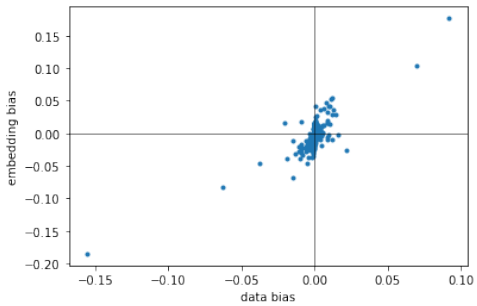
where $\vec{o}$, $\vec{e}$ let be the embeddings of an occupation o and the a or the b entity, respectively. The resulting similarity ranges from -1 meaning exactly the opposite, to 1 meaning exactly the same.

From the results in Table 6.7 we conclude that in most cases the information of the sensitive value is included to the embeddings. But it also seems that there is a small percentage of occupations that do not contain this information. So we want to see if there is any interest with these occupations.

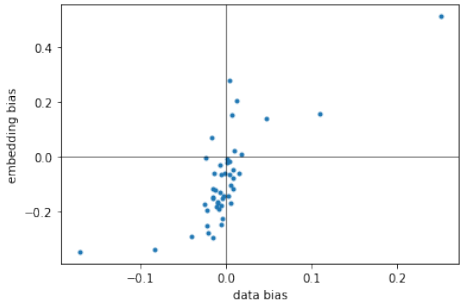
We plot the information of the data bias, embedding bias and similarity of the different sensitive attributes to find out if there is anything interesting in occupations that do not retain the sensitive information. In plots of Figure 6.5 we see that, for two sensitive values a and b, the occupations that do not retain the sensitive information and characterized as a in the dataset and b in the embeddings, or the opposite, have small data bias and small embedding bias both using TransE and TransH embeddings. This means that they are also the less popular occupations. On the other hand we observe that the occupations that have a lot of bias in the data, also have a lot of bias in the embeddings. This is another proof that our metrics express data bias and embedding bias very well.



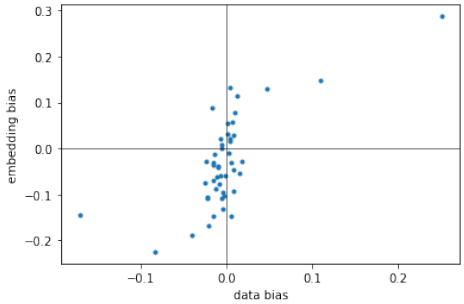
(a) gender - Wikidata subgraph - TransE



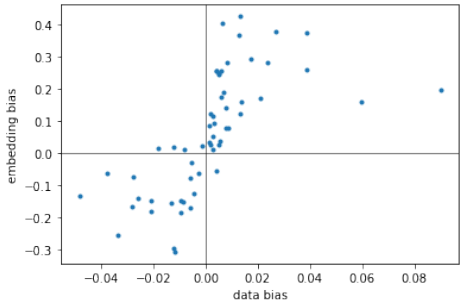
(b) gender - Wikidata subgraph - TransH



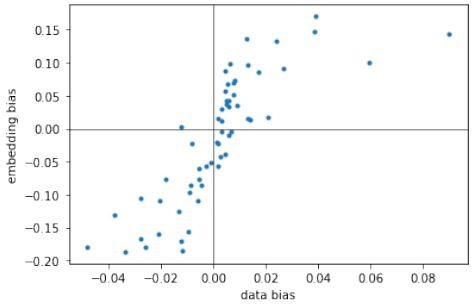
(c) relegion - FB13 dataset - TransE



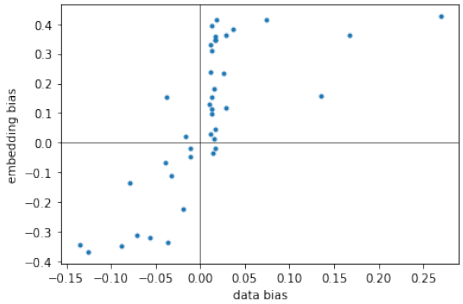
(d) relegion - FB13 dataset - TransH



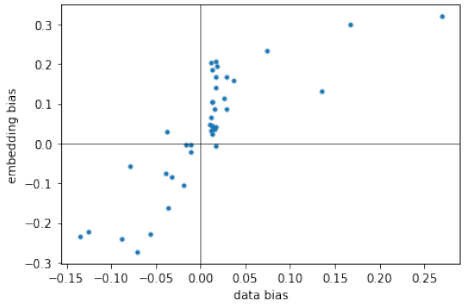
(e) nationality - FB13 dataset - TransE



(f) nationality - FB13 dataset - TransH



(g) ethnicity - FB13 dataset - TransE



(h) ethnicity - FB13 dataset - TransH

Figure 6.5: Let a, b be the sensitive values. Then the upper right box of each plot refers to the occupations characterized as a both in data and in the embeddings. The bottom right box refers to the occupations marked as a in the data and b in the embeddings. The lower left box refers to the occupations marked as b both in data and in the embeddings. Finally, the upper left box refers to the occupations marked as b in the data, but as a in the embeddings.

Table 6.7: Percentage of occupations, which are characterized as a or b both in the dataset and in the embeddings.

(a) Results using TransE embeddings.

s attribute	a-value	b-value	perc of a-occ (%)	perc of b-occ (%)
gender	male	female	74.91	97.95
religion	Catholicism	Judaism	42.11	96.43
nationality	US	UK	97.14	82.61
ethnicity	African Am	German	92.31	85.71

(b) Results using TransH embeddings.

s attribute	a-value	b-value	perc of a-occ (%)	perc of b-occ (%)
gender	male	female	72.0	88.36
religion	Catholicism	Judaism	63.16	85.17
nationality	US	UK	77.14	95.65
ethnicity	African Am	German	96.15	92.86

6.3.2 Detection of bias in KG embeddings using an analogy puzzle

In this section we try to find the occupation pairs that suit better to the sensitive values of the bias direction, to detect bias in the embeddings, as described in section 4.1.2. For the sensitive values a and b, the analogy puzzle is the "*b is to x as a is to y*", where x, y are two occupations. For example, for the $male$, $female$ sensitive values, the analogy puzzle is the "*female is to x as male is to y*".

In Table 6.8 are presented the best suited pairs for the analogy puzzles "*female is to x as male is to y*", "*Judaism is to x as Catholicism is to y*", "*UK is to x as US is to y*" and "*German is to x as African American is to y*" using TransE and TransH KG embeddings. We observe that the results for both TransE and TransH embeddings are similar. Specifically, analogies like "female is to model as male is to engineer" or "female is to model as male is to businessperson" show that there is gender bias in the KG embeddings. Also the fact that occupations about science suit better for Judaism, UK and German individuals, while respectively pope occupation for the Catholicism, businessperson for the US, and musician for the African American individuals, is also a sign that bias is transferred from the data to the KG embeddings.

Table 6.8: The pairs of occupations that suit in the analogy puzzle "*b is to x as a is to y*", where a and b are two sensitive values.

(a) "***female*** *is to* ***x*** *as* ***male*** *is to* ***y***" results using the Wikidata subgraph.

x - y (TransE score)	x - y (TransH score)
model - film director (0.91)	model - businessperson (0.71)
model - engineer (0.91)	model - engineer (0.67)
model - businessperson (0.89)	model - rugby union player (0.63)
beauty pageant contestant - engineer (0.88)	model - Australian rules football player (0.61)
model - composer (0.88)	model - priest (0.6)

(b) "***Judaism*** *is to* ***x*** *as* ***Catholicism*** *is to* ***y***" results using the FB13 dataset.

x - y (TransE score)	x - y (TransH score)
rabbi - pope (0.71)	rabbi - pope (0.71)
scientist - pope (0.67)	psychologist - pope (0.68)
physicist - pope (0.63)	scientist - pope (0.6)
conducting - pope (0.62)	conducting - pope (0.58)
mathematician - pope (0.59)	physicist - pope (0.52)

(c) "***UK*** *is to* ***x*** *as* ***US*** *is to* ***y***" results using the FB13 dataset.

x - y (TransE score)	x - y (TransH score)
mathematician - businessperson (0.55)	chemist - businessperson (0.42)
chemist - businessperson (0.51)	engineer - businessperson (0.42)
engineer - businessperson (0.51)	physicist - businessperson (0.41)
mathematician - baseball player (0.5)	mathematician - businessperson (0.4)
physician - businessperson (0.5)	surgeon - businessperson (0.39)

(d) "***German*** *is to* ***x*** *as* ***African American*** *is to* ***y***" results using the FB13 dataset.

x - y (TransE score)	x - y (TransH score)
philosopher - musician (0.6)	rabbi - musician (0.64)
rabbi - musician (0.6)	conducting - musician (0.64)
scientist - musician (0.59)	chemist - musician (0.62)
rabbi - jazz$_{pianist}$(0.59)	scientist - musician (0.6)
mathematician - musician (0.59)	physicist - musician (0.59)

6.3.3 Prediction task for bias amplification detection

As we have seen, bias exists in the data and is transferred through machine learning to the embeddings of the occupations. So it remains to be seen whether the biased occupations not only contain but also amplify biases. To do this we apply a prediction task, as described in 4.1.3. Using the gender sensitive attribute in Wikidata subgraph and the religion, nationality and ethnicity sensitive attributes in FB13 dataset, we predict for the most biased occupations of each sensitive value, who are the most likely individuals to have them. Then we compare the expected number of individuals of each sensitive value in the top-K results with the predicted one.

In Figure 6.6 are presented the results using TransE embeddings, while in Figure 6.7 are presented the results using TransH embeddings. If a, b are the two sensitive values, then we observe that in many cases the a-individuals have greater predicted score than the expected for a-occupations, and respectively, in many cases the b-individuals have greater predicted score than the expected for b-occupations. This shows bias augmentation using the KG embeddings. The only case that the expected of the a value is greater than the expected of the b value, and also there is an amplification of bias in the a value, is for the two less biased female occupations. The reason of this is, as seen from the synthetics, the big difference in the population between males and females.

Moreover, in Figures 6.9 (for TransE embeddings) and 6.10 (for TransH embeddings), we see how much the amplification was using the average of the $predicted - expected$ score subtraction, for K = 10, 20, 50 and 100 for the Wikidata subgraph, and K = 5, 7, 10 for the FB13 dataset. The results concerns the amplification of individuals of the same category with this of the sensitive value of the occupations, e.g males for male occupations. In most cases the amplification percentage shows that there is an increase of predicted individuals in relation to the expected and also that there are very view cases, where instead of amplification there is reduction. Of course, there are cases both in a and b values, where the predicted individuals could not be greater than the expected, because the expected is the 100% of the occupations, like for African American occupations. So even if in these case the expected is equal to the predicted individuals and the amplification is zero, the bias is very strong.

We conclude that the embeddings not only contain, but in many cases reinforce biases presented in data.

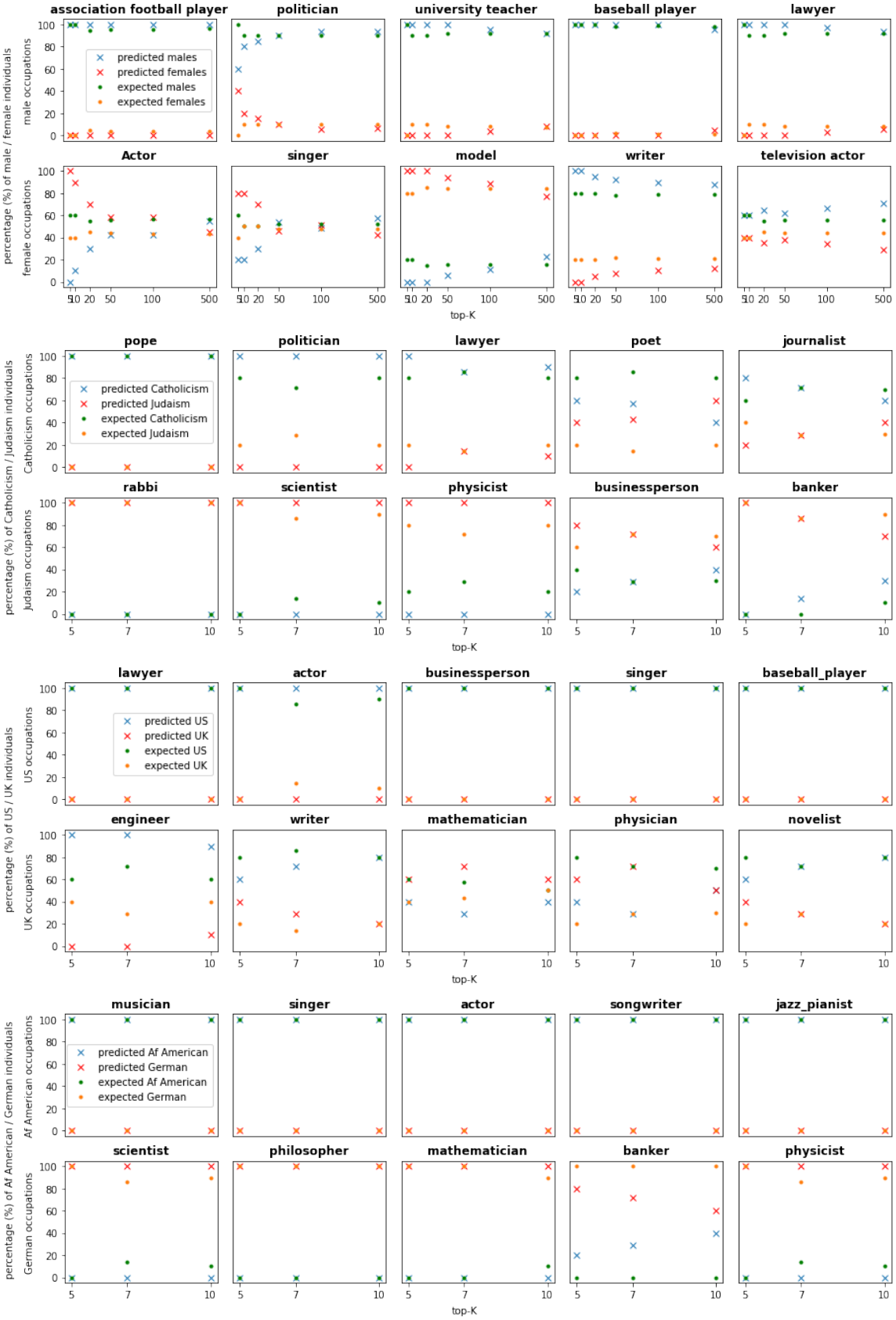


Figure 6.6: The expected and the predicted numbers of individuals in top-K results using TransE embeddings. Gender concerns to Wikidata dataset, while religion, nationality and ethnicity concern to FB13 dataset.

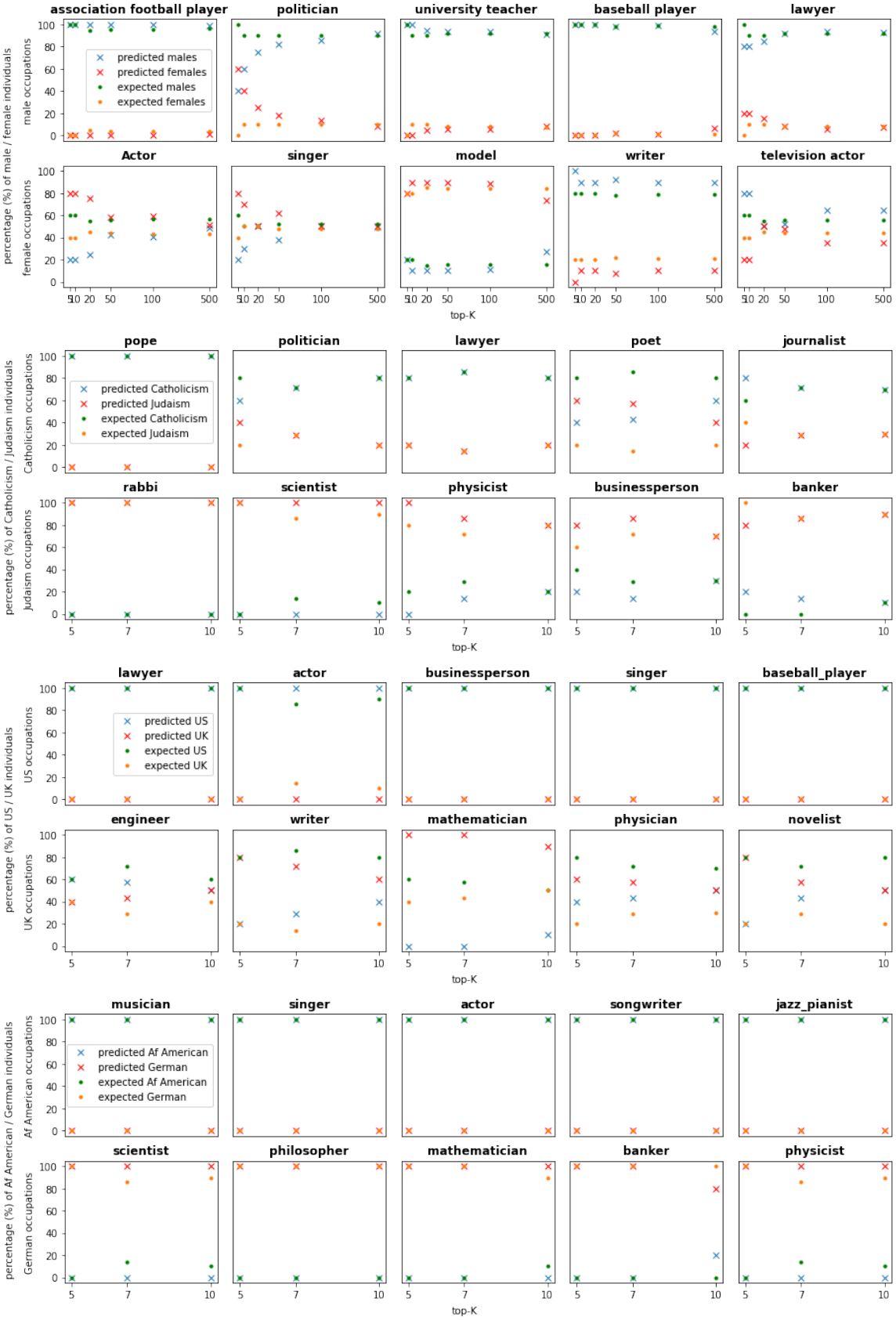


Figure 6.7: The expected and the predicted numbers of individuals in top-K results using TransH embeddings. Gender concerns to Wikidata dataset, while religion, nationality and ethnicity concern to FB13 dataset.

Table 6.9: Quantification of amplification. The average results of the $predicted - expected$ score of the five most biased occupations, using TransE embeddings, K=10, 20, 50, 100 for the Wikidata subgraph (for gender), and K = 5, 7, 10 for FB13 dataset (for religion, nationality and ethnicity).

(a) TransE embeddings - Wikidata subgraph - gender

	Top-10	Top-20	Top-50	Top-100
males	**2.0**	**4.0**	**4.4**	**3.6**
females	**16.0**	**7.0**	**0.4**	**0.4**

(b) TransE embeddings - FB13 dataset - religion / nationality / ethnicity

	Top-5	Top-7	Top-10
Catholicism	**8.0**	0.0	-4.0
Judaism	**8.0**	**8.6**	0.0
US	0.0	**2.9**	**2.0**
UK	**12.0**	**11.4**	0.0
Afr American	0.0	0.0	0.0
German	-4.0	**2.84**	-2.0

Table 6.10: Quantification of amplification. The average results of the $predicted - expected$ score of the five most biased occupations, using TransH embeddings, K=10, 20, 50, 100 for the Wikidata subgraph (for gender), and K = 5, 7, 10 for FB13 dataset (for religion, nationality and ethnicity).

(a) TransH embeddings - Wikidata subgraph - gender

	Top-10	Top-20	Top-50	Top-100
males	-6.0	-2.0	-0.4	**0.8**
females	**8.0**	**6.0**	**4.8**	**0.6**

(b) TransH embeddings - FB13 dataset - religion / nationality / ethnicity

	Top-5	Top-7	Top-10
Catholicism	-8.0	-8.6	-4.0
Judaism	**6.0**	**8.6**	**2.0**
US	0.0	**2.9**	**2.0**
UK	**44.0**	**37.1**	**28.0**
Afr American	0.0	0.0	0.0
German	0.0	**5.7**	**2.0**

6.4 Clustering results

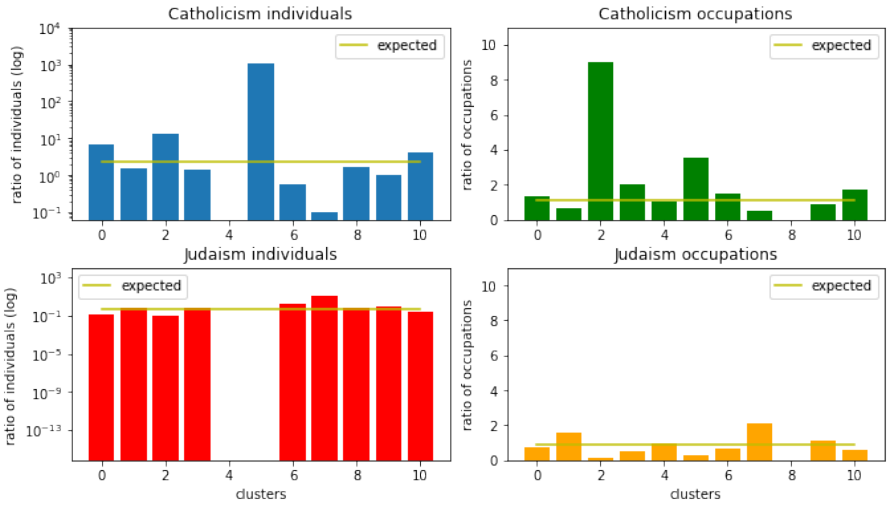
In this section we study if the clusters that produced by the KG embeddings are fair to the number of individuals and to the number of occupation of the sensitive values. We also detect if the occupations in the clusters are semantically related. To do this we use the FB13 dataset and the religion, nationality and ethnicity sensitive attributes with the Catholicism / Judaism, United States / United Kingdom and African American / German sensitive values, respectively.

For the clustering procedure we use the K-means algorithm. To select the appropriate K value for the number of the clusters we apply the "elbow method". Specifically, we produce clusters for different values of K and calculate the Sum of Squared Errors (SSE) of each value, that is the distance of points of their closest cluster center. Then we choose the K for which SSE starts to diminish, which is visible as an elbow. In all of the clustering procedures that concern the whole dataset we use K=11, while in the clustering procedures that concern the occupations we use K=6.

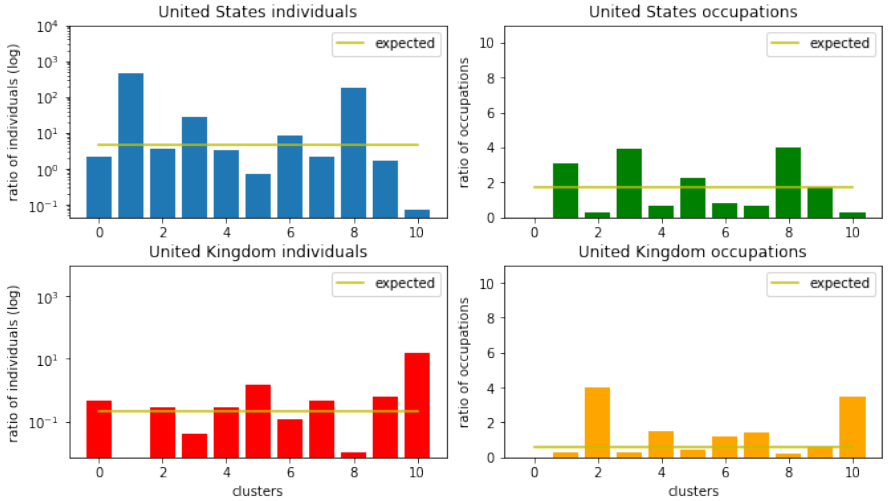
In Figures 6.8 and 6.9 there are the results of the ratio of individuals and the ratio of occupations in each cluster for the religion, nationality and ethnicity sensitive attributes using TransE and TransH embeddings. We observe that there are many clusters where there is inequality in the population of individuals. We also observe that many of the clusters, which are biased regarding to a sensitive value for individuals, they are also biased regarding to the occupations of this value. So it is obvious that the bias information in the data, is transferred to the embeddings and it is amplified by them, in the clustering procedure.

We then analyze the clusters of occupations. Using only the embeddings of the occupations, we see in Figures 6.10 and 6.11, that bias leads also to unfair clusters; clusters that amplify bias for one category. To study if occupations in the clusters are semantically related, we use the difference of occupation embeddings as defined in section 4.2. We expect that semantically related occupations have low score, while the score will be high otherwise. Actually, taking the embeddings of "writer" and "novelist" occupations, the score is 0.1 with TransE embeddings and 0.3 with TransH embeddings, while taking the embeddings of "dentist" and "carpenter" occupations, the score is 1.4 with TransE embeddings and 1.3 with TransH embeddings.

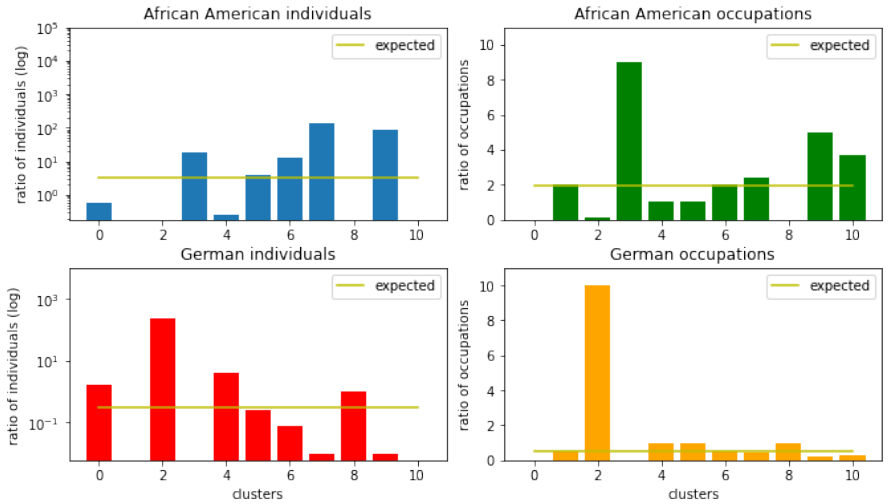
In Table 6.11 there are the semantically results of the clusters. We observe that in most cases, the occupations in the clusters are not semantically related. This means



(a) religion - TransE embeddings

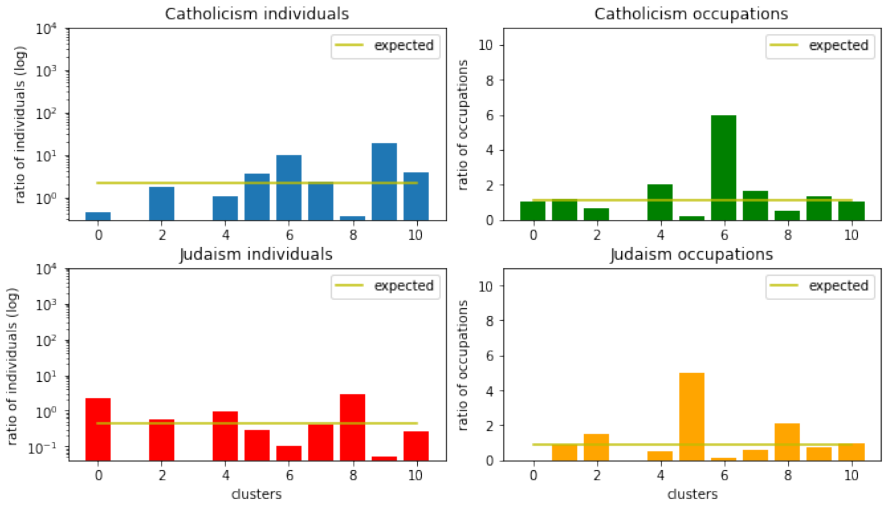


(b) nationality - TransE embeddings

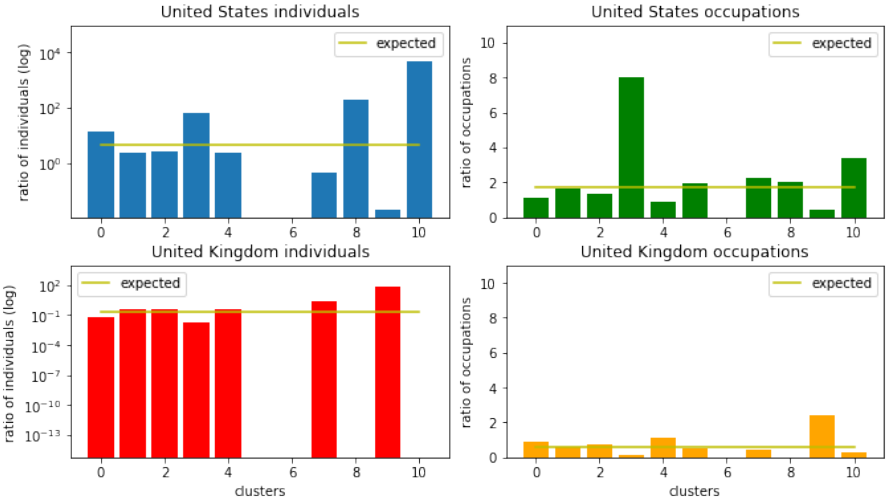


(c) ethnicity - TransE embeddings

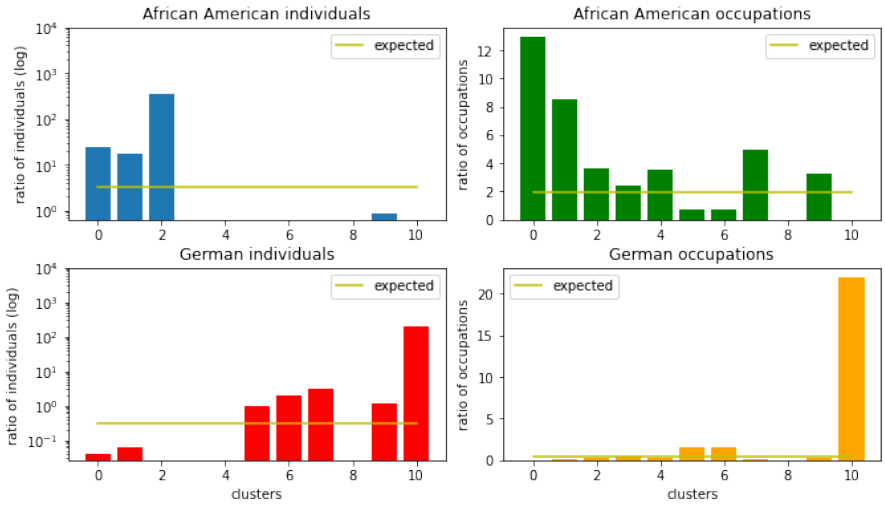
Figure 6.8: Clustering results using FB13 dataset and TransE embeddings.



(a) religion - TransH embeddings

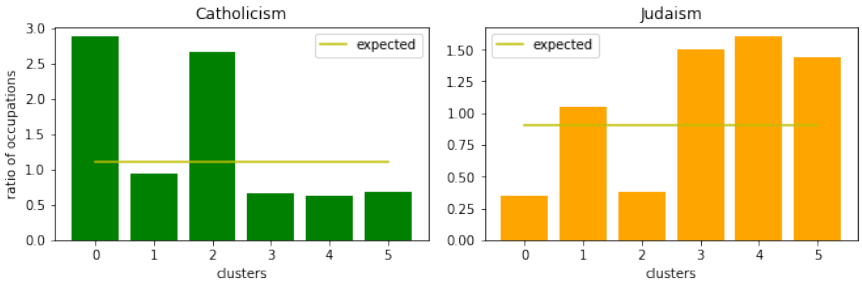


(b) nationality - TransH embeddings

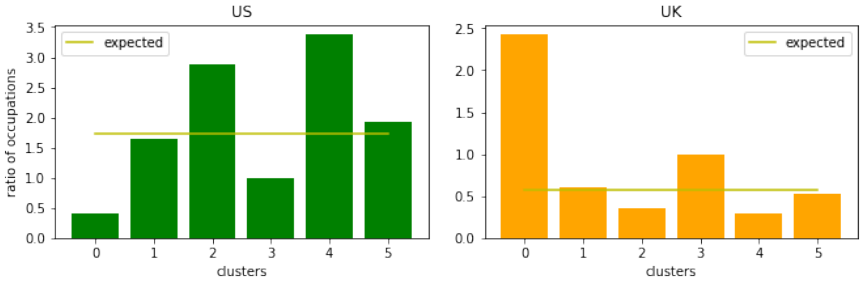


(c) ethnicity - TransH embeddings

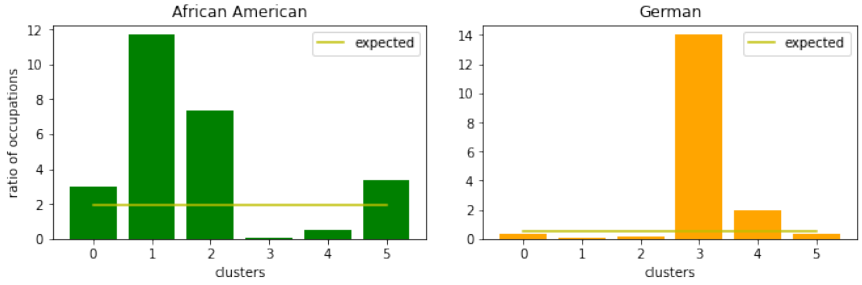
Figure 6.9: Clustering results using FB13 dataset and TransH embeddings.



(a) religion - TransE embeddings

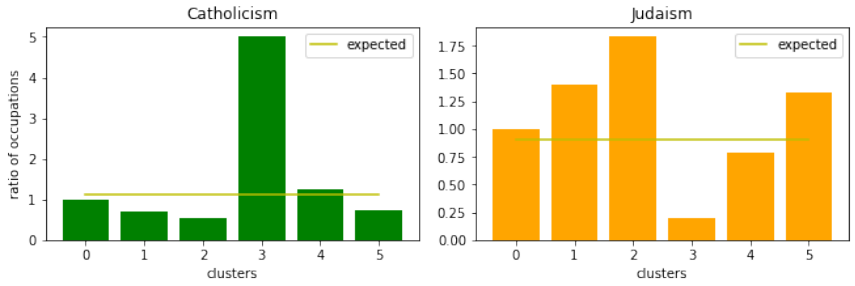


(b) nationality - TransE embeddings

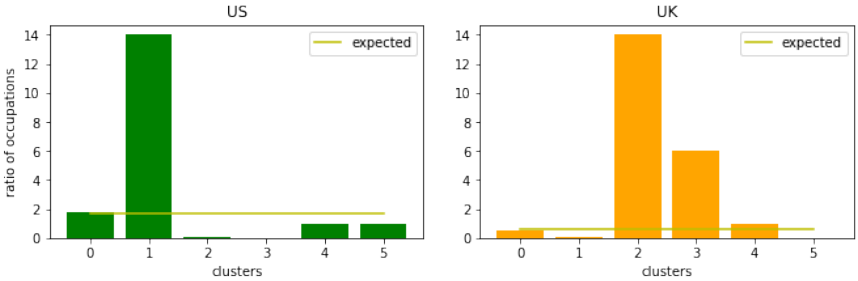


(c) ethnicity - TransE embeddings

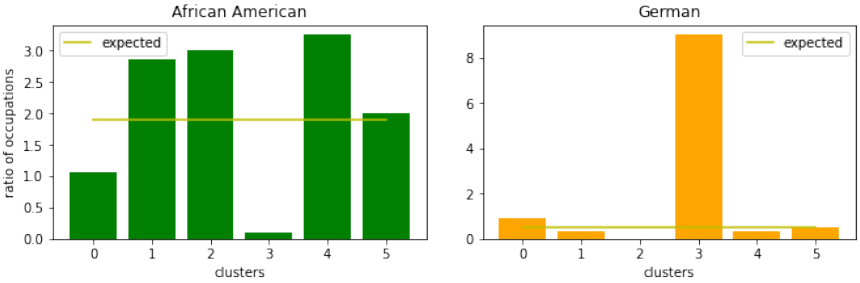
Figure 6.10: Clustering results of occupations using FB13 dataset and TransE embeddings.



(a) religion - TransH embeddings



(b) nationality - TransH embeddings



(c) ethnicity - TransH embeddings

Figure 6.11: Clustering results of occupations using FB13 dataset and TransH embeddings.

that the clustering procedure is mainly based on the sensitive information in the embeddings. This can have negative effect in the usage of clusters for feature engineering or prediction.

Table 6.11: Results of the semantic relation of occupations in the clusters.

(a) Results using TransE embeddings.

	Clust 0	Clust 1	Clust 2	Clust 3	Clust 4	Clust 5
religion	1.19	1.19	1.09	1.33	1.0	1.13
nationality	1.11	1.16	1.12	1.1	1.33	1.33
ethnicity	0.99	1.31	0.88	1.07	0.92	1.15

(b) Results using TransH embeddings.

	Clust 0	Clust 1	Clust 2	Clust 3	Clust 4	Clust 5
religion	1.12	1.09	1.34	1.15	1.22	1.29
nationality	1.33	1.06	0.96	0.62	0.66	0.0
ethnicity	1.2	1.32	0.68	0.86	0.89	0.62

6.5 Debias evaluation

As we have seen, bias exists in the data, but it is also transferred and amplified by the embeddings. It is therefore important to consider whether the debias method we proposed, is helping to reduce and perhaps eliminate bias. For this reason we first evaluate our method using projections and similarity, and then we examine the effect of the debias on the predictions.

6.5.1 Evaluation using projections and similarity

Since the projection in the bias direction expresses the bias information, we expect that the biased occupation embeddings have greater projection score than the neutrals. We also expect that the harder the debias is, the more information reduced from the embeddings, and for this reason the lower the projection score becomes. From the results in Table 6.12 we see that before debias both using TransE and TransH embeddings, the biased occupations have greater projection score than the neutrals. This means that projection truly express the bias information. We see also that after a

"soft" debias with λ=0.5 the bias information is reduced, while after a "hard" debias with λ=1, the bias information is completely lost from the embeddings.

Table 6.12: Let a, b be the two sensitive values. Table 6.12a gives the average projection score results of a, b and $neutral$ occupations on bias direction using TransE embeddings. Moreover Table 6.12b gives the average projection score results of a, b and $neutral$ occupations on bias direction using TransH embeddings. Both Tables present the projection score before debias, after a "soft" debias with λ=0.5 and after a "hard" debias with λ=1.

(a) Projection score results using TransE embeddings.

	Before debias	**Debias λ=0.5**	**Debias λ=1**
male occ	0.07	0.04	0.0
female occ	0.14	0.07	0.0
neutral occ	0.06	0.03	0.0
Catholicism occ	0.12	0.06	0.0
Judaism occ	0.17	0.09	0.0
neutral occ	0.11	0.05	0.0
US occ	0.18	0.09	0.0
UK occ	0.13	0.06	0.0
neutral occ	0.03	0.02	0.0
African American occ	0.23	0.12	0.0
German occ	0.20	0.1	0.0
neutral occ	0.11	0.05	0.0

(b) Projection score results using TransH embeddings.

	Before debias	**Debias λ=0.5**	**Debias λ=1**
male occ	0.01	0.0	0.0
female occ	0.02	0.01	0.0
neutral occ	0.0	0.0	0.0
Catholicism occ	0.08	0.04	0.0
Judaism occ	0.08	0.04	0.0
neutral occ	0.05	0.02	0.0
US occ	0.06	0.03	0.0
UK occ	0.11	0.06	0.0
neutral occ	0.06	0.03	0.0
African American occ	0.12	0.06	0.0
German occ	0.12	0.06	0.0
neutral occ	0.05	0.03	0.0

But we are also interested to see the results of the cosine similarity of the embeddings with the sensitive values after the debias. In Table 6.13 we observe that both

using TransE and TransH embeddings, as we expected, the similarity of a-occupations before the debias is greater with the a entity, while the similarity of the b-occupations is greater with the b entity. After a "soft" debias the similarity between the biased occupations and the two sensitive values begins to balance, while after a hard debias the similarity is totally balanced. This means that the bias information is eliminated, and also, as shown in Table 6.14 from the hits score the accuracy loss is small. We use the most biased occupations for the accuracy evaluation, because more information is extracted from them than in the rest, and is important to check the accuracy on them. The results are very satisfactory. It remains to be seen what happens in the predictions.

6.5.2 Evaluation using prediction task

In this subsection we study the effect of the debias in prediction. In other words, we want to find out if removing the bias information from the KG embeddings of the occupations, it reduces the occupations for which the percentage of bias value individuals is greater than or equal to the expected in the top-K results.

We select all the a-occupations and b-occupations and, using K=10, we predict which individuals are more likely to have these occupations. We compare the predicted number of the a and the b individuals in results with the expected one. We repeat this process using a "soft" debias with λ=0.5 and a "harder" debias with λ=0.8, and we study, if the percentage of occupations in which the bias value individuals is greater than or equal to the expected in the top-10 results is reduced.

In Figure 6.12 are presented the results using TransE and TransH embeddings for gender sensitive value in Wikidata subgraph, and religion, nationality and ethnicity sensitive values in FB13 dataset.

It is obvious that the reduction of bias information, e.g. of a-occupations, leads to predictions in which the b-individuals increase in the top-10 results and the a-individuals decrease. We also observe that the harder the debias gets, there are more occupations where this happens, something shows that our debias method helps the predictions to be more fair and free from bias.

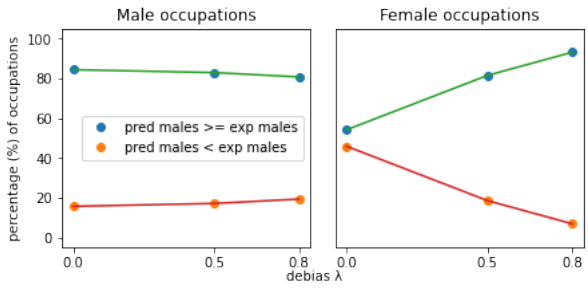
Table 6.13: Let a, b be the two sensitive values. Table 6.13a gives the average cosine similarity results of a, b and $neutral$ occupations with a and b entities using TransE embeddings. Moreover Table 6.13b gives the average cosine similarity results of a, b and $neutral$ occupations with a and b entities using TransH embeddings. Both Tables present the cosine similarity score before debias, after a "soft" debias with λ=0.5 and after a "hard" debias with λ=1.

(a) Cosine similarity results using TransE embeddings.

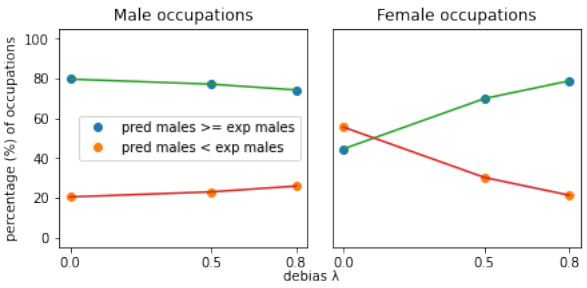
	Before debias	**Debias λ=0.5**	**Debias λ=1**
	male/female	male/female	male/female
male occ	0.11/0.10	0.11/0.11	0.11/0.11
female occ	0.11/0.14	0.12/0.13	0.12/0.12
neutral occ	0.10/0.10	0.10/0.10	0.10/0.10
	Catholicism/Judaism	Catholicism/Judaism	Catholicism/Judaism
Catholicism occ	0.08/0.05	0.08/0.06	0.07/0.07
Judaism occ	0.02/0.19	0.06/0.15	0.11/0.11
neutral occ	0.04/0.08	0.05/0.07	0.06/0.06
	US/UK	US/UK	US/UK
US occ	0.12/-0.04	0.08/0.0	0.04/0.04
UK occ	0.0/0.11	0.03/0.08	0.05/0.05
neutral occ	0.04/0.04	0.04/0.04	0.04/0.04
	African American/German	African American/German	African American/German
African American occ	0.22/-0.09	0.15/-0.01	0.07/0.07
German occ	-0.08/0.15	-0.02/0.1	0.04/0.04
neutral occ	0.07/-0.01	0.05/0.01	0.03/0.03

(b) Cosine similarity results using TransH embeddings.

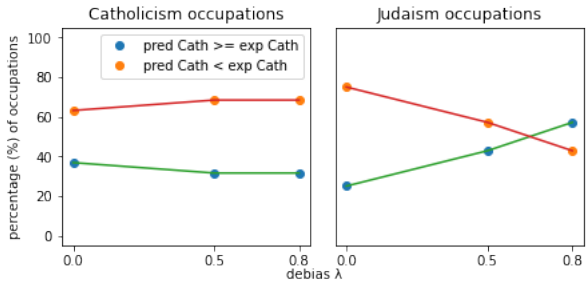
	Before debias	**Debias λ=0.5**	**Debias λ=1**
	male/female	male/female	male/female
male occ	-0.04/-0.05	-0.04/-0.05	-0.05/-0.05
female occ	-0.04/0.0	-0.03/-0.01	-0.02/-0.02
neutral occ	-0.02/-0.02	-0.02/-0.02	-0.02/-0.02
	Catholicism/Judaism	Catholicism/Judaism	Catholicism/Judaism
Catholicism occ	0.02/-0.06	0.0/-0.04	-0.02/-0.02
Judaism occ	-0.05/0.09	-0.01/0.05	0.02/0.02
neutral occ	-0.01/-0.03	-0.02/-0.03	-0.03/-0.03
	US/UK	US/UK	US/UK
US occ	0.04/0.02	0.03/0.02	0.03/0.03
UK occ	-0.17/-0.12	-0.16/-0.13	-0.15/-0.15
neutral occ	-0.03/0.0	-0.02/-0.01	-0.01/-0.01
	African American/German	African American/German	African American/German
African American occ	0.18/-0.14	0.11/-0.06	0.02/0.02
German occ	-0.11/0.14	-0.04/0.09	0.02/0.02
neutral occ	0.07/-0.03	0.04/0.0	0.02/0.02



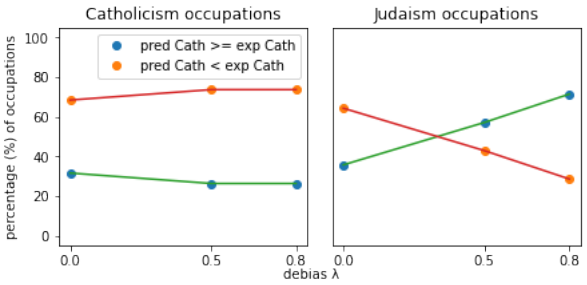
(a) Wikidata subgraph - TransE embeddings



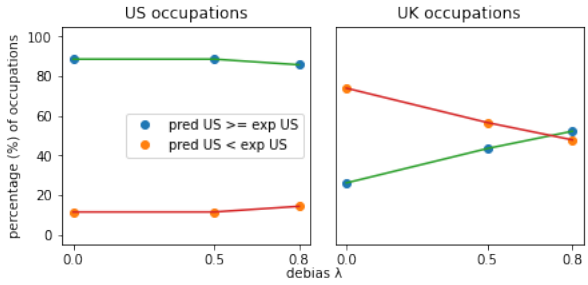
(b) Wikidata subgraph - TransH embeddings



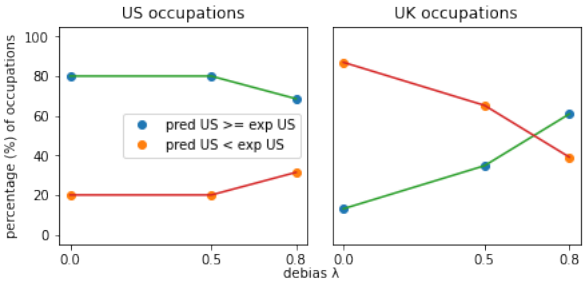
(c) FB13 dataset - TransE embeddings



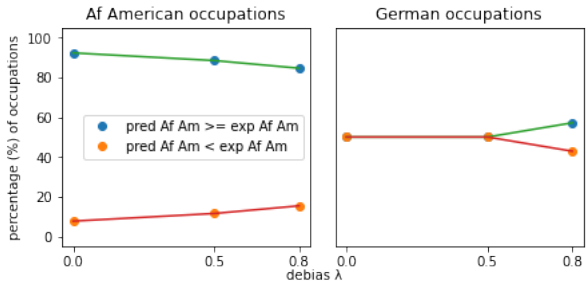
(d) FB13 dataset - TransH embeddings



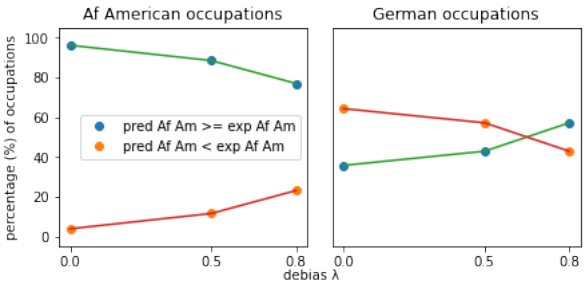
(e) FB13 dataset - TransE embeddings



(f) FB13 dataset - TransH embeddings



(g) FB13 dataset - TransE embeddings



(h) FB13 dataset - TransH embeddings

Figure 6.12: Prediction task for debias evaluation using λ=0.5, λ=0.8 and K=10.

Table 6.14: Hits@10 for the most 40 biased gender occupations using Wikidata dataset, and Hits@5 for the most 20 biased religion, nationality and ethnicity occupations using FB13 dataset.

(a) Hits@k using TransE embeddings.

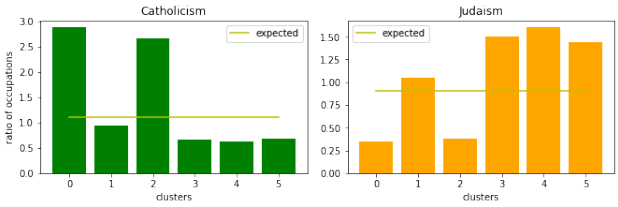
s attribute	Before debias	Debias λ=0.5	Debias λ=1
gender	0.76	0.74	0.71
religion	0.78	0.74	0.70
nationality	0.70	0.71	0.69
ethnicity	0.67	0.65	0.62

(b) Hits@k using TransH embeddings.

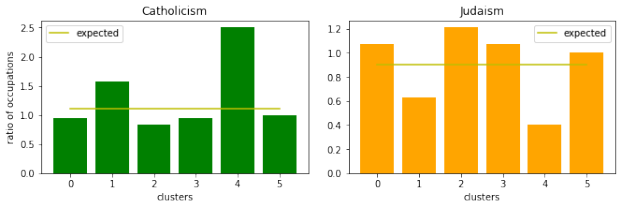
s attribute	Before debias	Debias λ=0.5	Debias λ=1
gender	0.79	0.78	0.77
religion	0.76	0.76	0.75
nationality	0.74	0.74	0.73
ethnicity	0.73	0.75	0.66

6.5.3 Evaluation using clustering

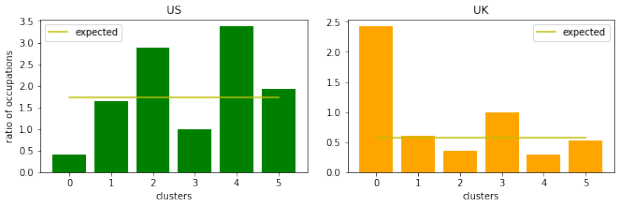
In this subsection we study if our debias approach has also effect on clustering. For the results in Figures 6.13 and 6.14, using TransE and TransH embeddings, we observe that the clusters of occupations using the debias embeddings with λ=0.8, are more fair than before. The number of individuals of the two sensitive values is closer to the expected. Of course there are still unfair clusters, and in many cases more than we would like to be. This means that maybe the clustering needs a harder debias, or a different debias approach and also that maybe the distance metric of the Kmeans algorithm (euclidean distance in our case) is not appropriate for more fair clusters production.



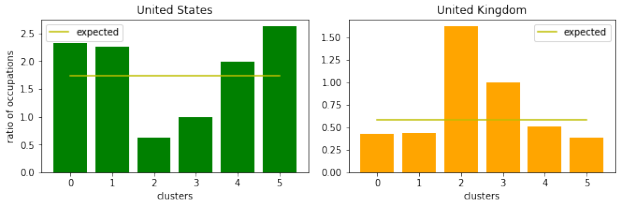
(a) religion - TransE embeddings - Before debias



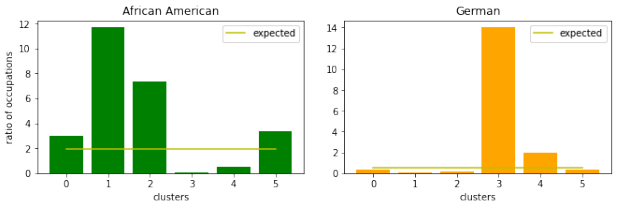
(b) religion - TransE embeddings - Debias λ=0.8



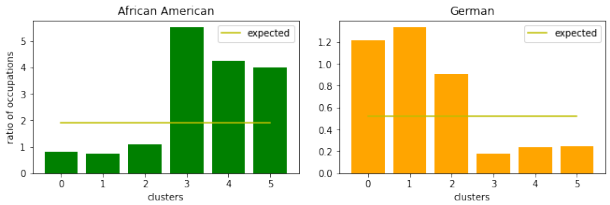
(c) nationality - TransE embeddings - Before debias



(d) nationality - TransE embeddings - Debias λ=0.8

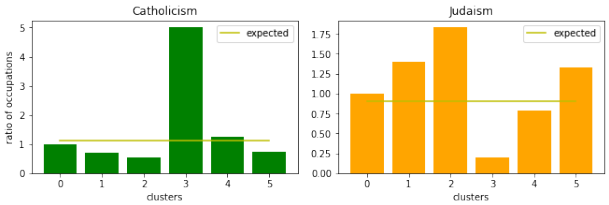


(e) ethnicity - TransE embeddings - Before debias

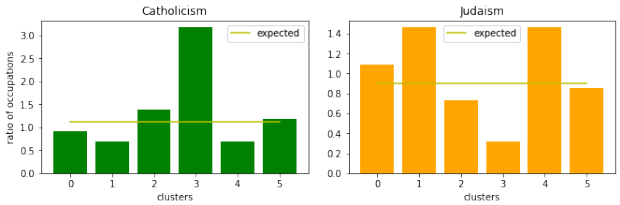


(f) ethnicity - TransE embeddings - Debias λ=0.8

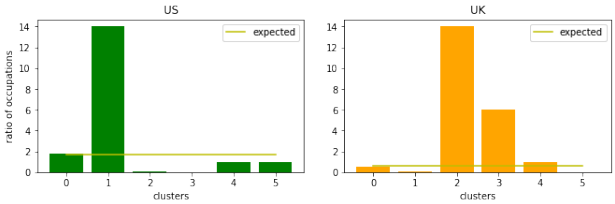
Figure 6.13: Clustering results of occupations using TransE embeddings before debias and after a debias with λ=0.8.



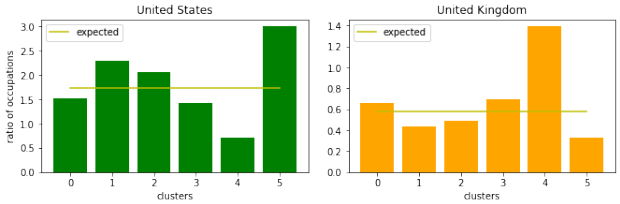
(a) religion - TransH embeddings - Before debias



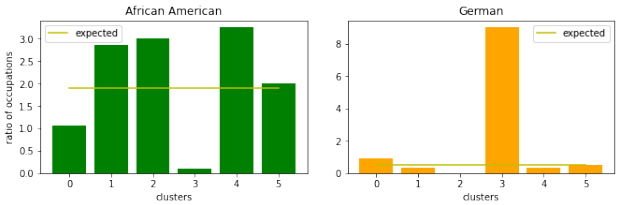
(b) religion - TransH embeddings - Debias λ=0.8



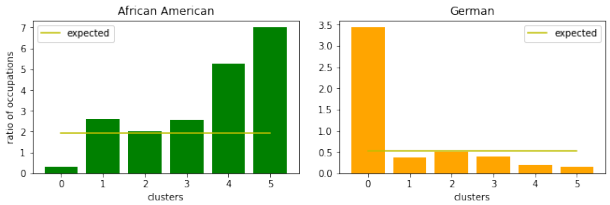
(c) nationality - TransH embeddings - Before debias



(d) nationality - TransH embeddings - Debias λ=0.8



(e) ethnicity - TransH embeddings - Before debias



(f) ethnicity - TransH embeddings - Debias λ=0.8

Figure 6.14: Clustering results of occupations using TransH embeddings before debias and after a debias with λ=0.8.

Chapter 7

Related Work

Algorithmic fairness and bias have been the focus of much current research in machine learning (see e.g., [5] for a recently survey). Because embeddings are widely used to represent complex data in a variety of machine learning tasks, and also they used in a wide range of applications, problems of bias introduced by the embeddings must receive most attention. Interest has a recent work in word embeddings, which has shown that they encode various forms of social biases [6].

In this thesis we study bias in KG embeddings [2, 3, 4]. In a recent work [1], writers apply a transformation to the entity embeddings and then using all the individuals in the dataset, they analyse whether, making an entity more male increases or decreases the likelihood that they have an occupation, according to the model's score function. In contrast with [1], we propose a more simple and efficient method, which also includes the information of popularity in bias score. We further suggest a prediction method for detection of bias augmentation using the KG embeddings. Bias amplifications has been previously studied for many applications, e.g., for recommendations [17]. We also study the bias affection in clustering, applying a clustering task as in [14]. We study the fairness in the clustering based on the population of sensitive entities as in [18], but the expected for us is the ratio of sensitive values' populations, because we have unbalanced population in the data. We also detect the semantic relation of the occupations in the clusters. Moreover, we suggest a debias approach for removing bias from the KG embeddings, which is based on projections on the gender subspace as in [6, 15]. In relation to [10], its novelty lies on tuning the amount of bias it removes using pretrained embeddings instead of the modification

of the KG embedding model and on removing information from the embeddings of occupations instead of individuals.

There has been also some other recent work on achieving fairness in graph embeddings [19, 20]. The approach in [19] modifies the random walk step, but the training procedure that is used in KGs makes this approach not applicable to them. Moreover in [20], the learning of graph embeddings is enhanced with a set of adversarial regularization filters that remove information about sensitive attributes (e.g., gender) using projections. For this dataset, the authors report significant loss in accuracy. In our approach, the removal of any bias is done in a post-processing step and it is controlled by the bias in the input data. Thus, some gender information necessary for the accuracy of the prediction task remains.

Chapter 8

Conclusions and Future work

In this thesis, we studied the bias in the KG embeddings. Especially, we first proposed two metrics for measuring bias in the dataset. We concluded that the metric, which takes into account the population of the sensitive values and gives higher score to the popular occupations was more appropriate. After that we showed that bias exists in the dataset and using a quantitative and a qualitative method, based on projections and analogies, respectively, we discovered that bias is transferred from the data to the KG embeddings. We also showed that popularity and inequality in populations of sensitive values affects bias in KG embeddings and further, that bias leads to unfair clusters when we use a clustering task. An important observation using a prediction task was, that bias not only transferred from the data to the KG embeddings, but it also amplified by them. Finally, the experimental results for our debias method has shown that it is effective in removing as much bias we want with a small accuracy loss.

In the future, we plan to consider additional embedding models besides TransE and TransH and more KGs. We plan also to measure bias in more applications and for more sensitive attributes.

Bibliography

[1] J. Fisher, D. Palfrey, C. Christodoulopoulos, and A. Mittal, "Measuring social bias in knowledge graph embeddings," *arXiv preprint arXiv:1912.02761*, 2019.

[2] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 12, pp. 2724–2743, 2017.

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

[4] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes." in *Aaai*, vol. 14, no. 2014, 2014, pp. 1112–1119.

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

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

[7] K. Janowicz, B. Yan, B. Regalia, R. Zhu, and G. Mai, "Debiasing knowledge graphs: Why female presidents are not like female popes." in *International Semantic Web Conference (P&D/Industry/BlueSky)*, 2018.

[8] O. Zagovora, F. Flöck, and C. Wagner, ""(weitergeleitet von journalistin)": The gendered presentation of professions on wikipedia," in *WebSci*, 2017, pp. 83–92.

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

[10] J. Fisher, A. Mittal, D. Palfrey, and C. Christodoulopoulos, "Debiasing knowledge graph embeddings," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7332–7345.

[11] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[12] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 1247–1250.

[13] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques third edition," *The Morgan Kaufmann Series in Data Management Systems*, pp. 83–124, 2011.

[14] M. H. Gad-Elrab, D. Stepanova, T.-K. Tran, H. Adel, and G. Weikum, "Excut: Explainable embedding-based clustering over knowledge graphs," in *International Semantic Web Conference*. Springer, 2020, pp. 218–237.

[15] S. Dev and J. Phillips, "Attenuating bias in word vectors," *arXiv preprint arXiv:1901.07656*, 2019.

[16] X. Han, S. Cao, L. Xin, Y. Lin, Z. Liu, M. Sun, and J. Li, "Openke: An open toolkit for knowledge embedding," in *EMNLP*, 2018.

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

[18] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii, "Fair clustering through fairlets," *arXiv preprint arXiv:1802.05733*, 2018.

[19] T. A. Rahman, B. Surma, M. Backes, and Y. Zhang, "Fairwalk: Towards fair graph embedding," in *IJCAI*, 2019, pp. 3289–3295.

[20] A. J. Bose and W. L. Hamilton, "Compositional fairness constraints for graph embeddings," in *ICML*, vol. 97, 2019, pp. 715–724.

Short Biography

My name is Styliani Bourli and I am a second year master student on science in data and computer systems engineering. I graduated from the Department of Computer Science and Engineering, in University of Ioannina, in 2019. I had my first publication on bias in knowledge graph embeddings in 2020. While in school, I earned a honorable diploma in 72nd Panhellenic mathematics competition "THALES". I enjoy volunteering, attending seminars and participating in teams.