# The Lifecycle of "Facts": A Survey of Social Bias in Knowledge Graphs

Link: <https://aclanthology.org/2022.aacl-main.49/>

**Motivation**

Knowledge graphs (KGs) importance:

* provide a structured and transparent form of information representation
* lie at the core of popular Semantic Web technologies
* downstream tasks

KG problem:

* manifest social biases -> propagate harmful prejudices

⇒ ethical risks must be targeted and avoided during development and application

**Approach**

**1.Bias in KG creation**

Crowd-sourcing facts

* homogeneous set of interests and knowledge
* sampling bias -> geospatial coverage of information
* Solutions
  + ask for evidence supporting their statements
  + keep track of their demographic backgrounds.

Ontologies

* hand-made and a source of bias due to the influence of opinions, motivations, and personal choices
* e.g. scientific opinions, political and religious views
* the types of information used to characterize a person entity

Extraction

* use of template sentences -> erroneous tags for female names
* names associated with different ethnicities and genders
* non-white names yielded on average lower performance scores than white names
* Solutions:
  + downsampling the training data to equalize the number of male and female instances
  + augmenting the data by artificially introducing new female instances
  + word embedding debiasing

**2. Bias in KG**

* several countries and continents are underrepresented
* disproportionate gender entities (70% M, 20% F, 10% others)
* Incorrect prominent occupation categories
* ethnic group labels
* semantic bias -> representational harms

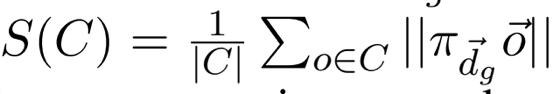
**3. Bias in KG embeddings**

Stereotypical Analogies

* whether demographics are associated with attributes in stereotypical ways
* cosine score

Projection-based measurement

Projection score metric S(C):



where:

* d\_g: one-dimensional gender direction
* pi: projection
* o: occupation vector
* C: set of occupations

Interpretation:

* high S(C) = C gender-biased
* S(C) ~ 0 = C neutral

Translational likelihood (TL) metric

* embedding of a person entity is updated for one step towards one pole of a seed dimension
* calculate TL before and after update
* absolute value averaged across all human entities high -> bias regarding the examined seed-attribute pair.
* avoids model-specificity -> generalize to any scoring function
* does not compare well between different types of embeddings
* Alternatively: flipping the entity’s gender and fully re-training the model afterward

**4. Bias in downstream tasks**

* Link prediction = prediction of relations between entities in a given KG.
* **demographic parity**: equalizes the chance of success, regardless of group => probability of predicting a label
* **predictive parity**: equalizes the chance of success, given a positive prediction, regardless of group => true positive
* <https://afraenkel.github.io/fairness-book/content/05-parity-measures.html>
* truth gender ratio not predictive of the bias metric
* validation: predicted bias values correlate to the gender distributions of occupations [QUE: so the first finding was incorrect?, or not predictive but still correlate, so there’s something else?]
* gender bias is mostly driven by triples containing gendered entities and triples of low degree. [QUE: low degree -> need more information in the triple?]

**5. Bias mitigation**

Data balancing

* added more female triples from another source (real) -> even out the binary gender distribution
* more females entities are predicted
* absolute difference between the female ratios in the data and the predictions increased,
* model less accurate and fair
* not scalable since for some domains there are no or only a limited amount of female entities

Add more synthetic triples

* inconsistent bias change
* inline with : ground truth gender ratios are not perfectly predictive of downstream task bias
* remove the triples that most strongly influenced an existing bias -> outperformed random triple removal

Adversarial Learning

* prevent prediction of a specific personal attribute from a person’s entity embedding
* adversarial loss: KL-divergence between the link prediction score distribution and an idealized target distribution
  + <https://towardsdatascience.com/understanding-kl-divergence-f3ddc8dff254>
  + -> incentives to give same probability
  + -> related information encoded in the embedding that is able to inform the same bias
* Filtering Adversarial Network (FAN)
  + filter: remove sensitive attribute information from the input
  + discriminator: predict the sensitive attribute from the output
  + separately pre-trained -> jointly trained as adversaries
  + without additional occupation classifier: high- and low-degree entities was close to random
  + with additional occupation classifier: accuracy remained unaffected

Hard debiasing

* subtract linear projection onto the previously computed gender direction from the occupation embedding
* soft version: preserve some degree of gender information by applying a weight
  + 0 < lambda < 1 to the projection value before subtraction
* Result: correlation between gender and occupation was effectively removed
* debiasing degree (lambda): trade-off with model accuracy
* Problem: only reduce gender bias, not complete removal

**Future work**

* aggregate the correlations between a set of seed dimensions and all relations in a graph
* validation: compare embedding-level bias metrics with census-aligned data to assess compatibility with real-world inequalities; obtained measurements capture all relevant aspects of the construct the instrument claims to measure
* name potential social harms as a motivator for their research on social bias in KGs
* clear working definition of social bias
* aspects of pre-existing societal biases captured in the data vs biases arising through the algorithm

**Recommendations**

Transparency and accountability

* publish bias-sensitive documentation with KG
* report demographic background of contributors
* investigate respective data dumps for potential biases and report limitations

Representativeness

* employ authors and annotators from diverse social groups and with varied cultural imprints

Algorithm bias

* combination of quantitative and qualitative measures
* evaluate content of attributions in social discourse & intended use of tech
* task/context-oriented evaluation for downstream task bias -> these bias exist independently to embedding bias -> need to measure separately