FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Social and technical biases in Knowledge Graphs

Harald Sack

Knowledge Graphs and their Role in the Knowledge Engineering of the 21st Century
Dagstuhl, 13.09.2022

What do we mean by “Bias”?

- Bias is a **disproportionate weight in favor of or against an idea or thing**, usually in a way that is closed-minded, prejudicial, or unfair. (Wikipedia)
- Bias can be thought of as “**prejudice in favor or against a person, group, or thing that is considered to be unfair**” (Jones, 2019)
- Bias is “**a particular tendency, trend, inclination, feeling, or opinion, especially one that is preconceived or unreasoned**.” (dictionary.com)

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Knowledge Graphs

- Knowledge Graphs (KGs) store human knowledge about the world in structured format, e.g., **triples of facts** or **graphs of entities and relations**, to be processed by AI systems.
- In the past decade, extensive research efforts have gone into constructing and **utilizing KGs for tasks in natural language processing, information retrieval, recommender systems**, and many more.
- **Once constructed, KGs are often considered as "gold standard" data sources that safeguard the correctness of other systems**.

Bias in Knowledge Graphs

- **Biases inherent to KGs** may become **magnified** and **spread through KG based systems (Bias Network Effect)**.
- Therefore, it is crucial that we acknowledge and address various types of bias in knowledge graph construction.

"We believe that debiasing knowledge graphs will become a pressing issue as these graphs enter everyday life rapidly."
(Janowicz et al., 2018)

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Biases in Knowledge Graphs are Different...

- **Biases in KGs**, as well as potential means to address them,
are different from those in linguistic models or image classification:
    - **KGs are sparse by nature**,
    i.e. only a small number of triples are available per entity.
    - **Linguistic models**
    learn the meaning of a term from its context within a **large corpus**.
    - **Image classification**
    learns classes from **millions of labeled images**.

Origins of Bias in Knowledge Graphs

- Biases in KGs may **originate in the very design** of the KG,
    - in the **source data** from which it is created (semi-)automatically, and
    - in the **algorithms** used to **sample**, **aggregate**, and **process that data**.

- **Source Biases**
    - typically appear in expressions, utterances, and text sources, and
    - can **carry over into downstream representations** such as **knowledge graphs** and **(knowledge graph) embeddings**.

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Human Biases

- **Reporting bias**: What people share is not a reflection of real-world frequencies
- **Selection Bias**: Selection does not reflect a random sample
- **Out-group homogeneity bias**: People tend to see outgroup members as more alike than ingroup members when comparing attitudes, values, personality traits, and other characteristics
- **Confirmation bias**: The tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses
- **Overgeneralization**: Coming to conclusion based on information that is too general and/or not specific enough
- **Correlation fallacy**: Confusing correlation with causation
- **Automation bias**: Humans often favor suggestions from automated decision-making systems over contradictory information without automation

More at https://developers.google.com/machine-learning/glossary/

Sources of Bias in Knowledge Graphs

- **Biases in KGs** can arise from multiple sources: (Janowicz et al., 2018),
  - **Data Bias:** the **data collection process** for the KG or simply **from the available data**,
  - **Schema Bias**: the **chosen ontology** or simply **embedded in ontologies**,
  - **Inferential Bias**: the result of drawing **inferences**
- Furthermore, **Biases in KG embeddings** may also arise from the **embedding method**.

Social Bias in Collaboratively Created KGs

- Knowledge Graphs are considered as **a source for "truth"**
- But what about **controversial facts**? (Demartini, 2019)
- *Is Catalunya part of Spain?*
  - *The answer might be controversial, depending whom you are asking*
- `:Calatunya :isPartOf :Spain .` or `:Catalunya a :Country .`
- As a solution, ask the crowd:
  - Provide both facts in your KG
  - Indicate for both the support from the crowd.

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Data Bias in Knowledge Graphs

- For Europe, Japan, Australia, and the US, DBpedia location density strongly correlates with population density.
- But not for large parts of Asia, Africa, and South America.
- DBpedia describes (mostly) the Western World from a Western Perspective
- This also holds for other DBpedia language versions or LOD resources, as e.g. for GeoNames (Janowicz et al., 2016)
- Similar observation in ML community (word embeddings, image search, tagging, etc.)

Schema Bias in Knowledge Graphs

- Ontologies are (mostly) developed in a **top-down** manner
  - with **application needs** in mind, or
  - certain **philosophical stances** (as for top-level ontologies).
- Ontologies are typically defined by a group of **knowledge engineers** in collaboration with **domain experts**
  - consequently (implicitly) **reflect the worldviews and biases of the development team**.
- Such ontologies will likely contain
  - most of the well-known **human biases and heuristics**, in particular **anthropocentric thinking**
- Problem:
  - A **bottom-up** strategy, as e.g. using ML to derive axioms/rules from data, will again suffer from **data biases**

Schema Bias in Knowledge Graphs

- **Encoding bias**: models depend on the selected DL fragment (and not the other way around)
- Many biases are **not directly encoded in the ontology** but **only become visible when comparing multiple ontologies together with their respective datasets**.

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Schema Bias in Knowledge Graphs

- For example, **DBpedia**, **GeoNames**, and the **Getty Thesaurus of Geographic Names (TGN)** all contain a `Theater` class.
  - data-driven perspective: **spatial statistics** for all `Theater` class members (intensity, interaction, point patterns) should yield similar results.
  - This is not the case: indicators show very distinct patterns.
  - **GeoNames** aims at containing all currently existing theaters,
  - **DBpeda** contains culturally/historically relevant theaters, and
  - **TGN** contains those that are significant for works of art.
- Differences in class extension show **implicit biases across the classes** despite their common name.

Inferential Bias in Knowledge Graphs

- **Inferential biases** in KGs arise at inferencing level, such as
  - reasoning,
  - querying, or
  - rule learning.

- Example:
  - Results of a **SPARQL** queries depend on the **entailment regimes** (e.g., simple vs. RDFS entailment).
  - In consequence, different SPARQL endpoints containing the same KG might yield different SPARQL results.

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

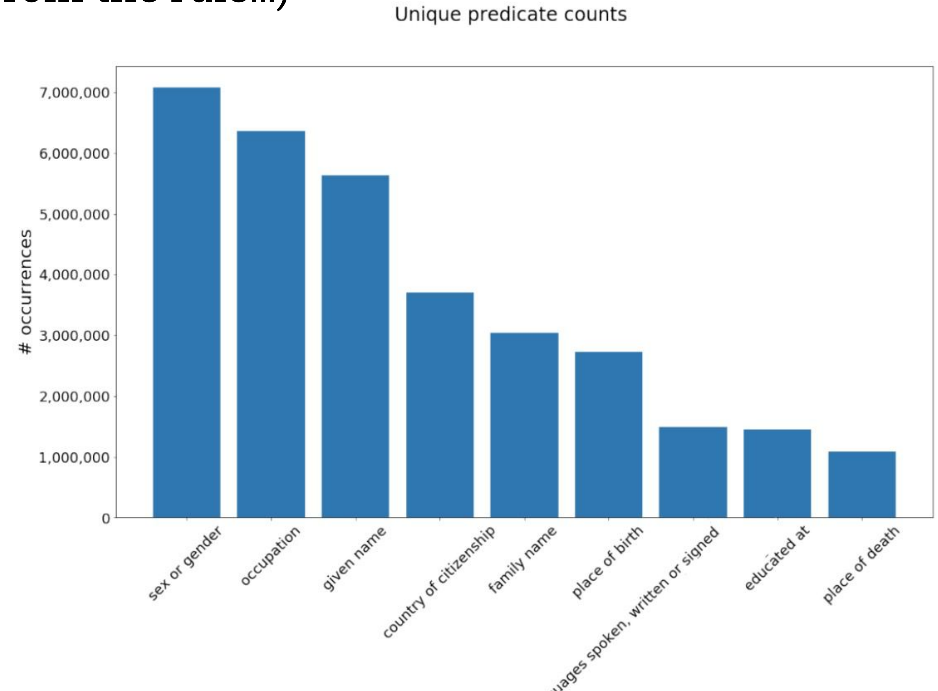
Inferential Bias in Knowledge Graphs

- Learning a (correct) model (e.g. via association rule mining) might collide with **social consensus**.
- Example:
  - Consider all **popes**, **US 5-star generals**, and **US presidents** from DBpedia.
  - These entities have one aspect in common: they are **all male**
  - Rule mining:
    - (1) if X is a pope, X is male; (*correct, by definition*)
    - (2) if X is a US 5-star general, X is male; (*correct, static enumerated class*)
    - (3) if X is a US president, X is male. (*collides with social consensus!*)
  - While these rules may be perceived as controversial, they are all correct.

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Bias in Knowledge Graph Embeddings

- Knowledge Graphs are **prone to errors**, due to
  - collaborative construction paradigm
  - automated procedures for KG construction (cannot consider all ‘exceptions’ from the rule...)
- As a result, KGs often are **incomplete**,
- which might be the cause for **bias in KG embeddings** trained on this KG.



(Radstock et al, 2021 & Vrandecic et al, 2014)

Bias in Knowledge Graph Embeddings

- If the underlying knowledge graph is biased,
  then also KG embeddings trained on this base data.
- **De-biasing KG embeddings** requires methods for
  - **Detecting bias** in KG embeddings
  - **Removing bias** from KG embeddings
- De-biasing KGEs is tricky, dependent on the underlying embedding model.

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Bias in Knowledge Graph Embeddings

- **Technical Bias** in KGE (Keidar et al., 2021):
    - Can be detected, as e.g., via **Link prediction** over the same KG with different embedding models, trained on the same KG.
- **Bias Measures**:
    - **Demographic Parity Distance** (DPD): focusses on potential bias in GT data
    - **Predictive Parity Distance** (PPD): focusses on classifier precision
        - DPD and PPD rely on classification task
        - measure the bias of sensitive relations (as e.g. “:gender”) via classification on a target relation (as e.g. on “:profession”).
    - **Translational Likelyhood**: focusses on scoring function of embedding model

Bias in Knowledge Graph Embeddings

- **Further Bias Measure**: (Fisher, 2019)
    - **Fine tuning of embeddings for bias detection**, as e.g., turning entities more "male" or "female" according to the used model and observe predictions on sensitive relations (as e.g. "occupation")

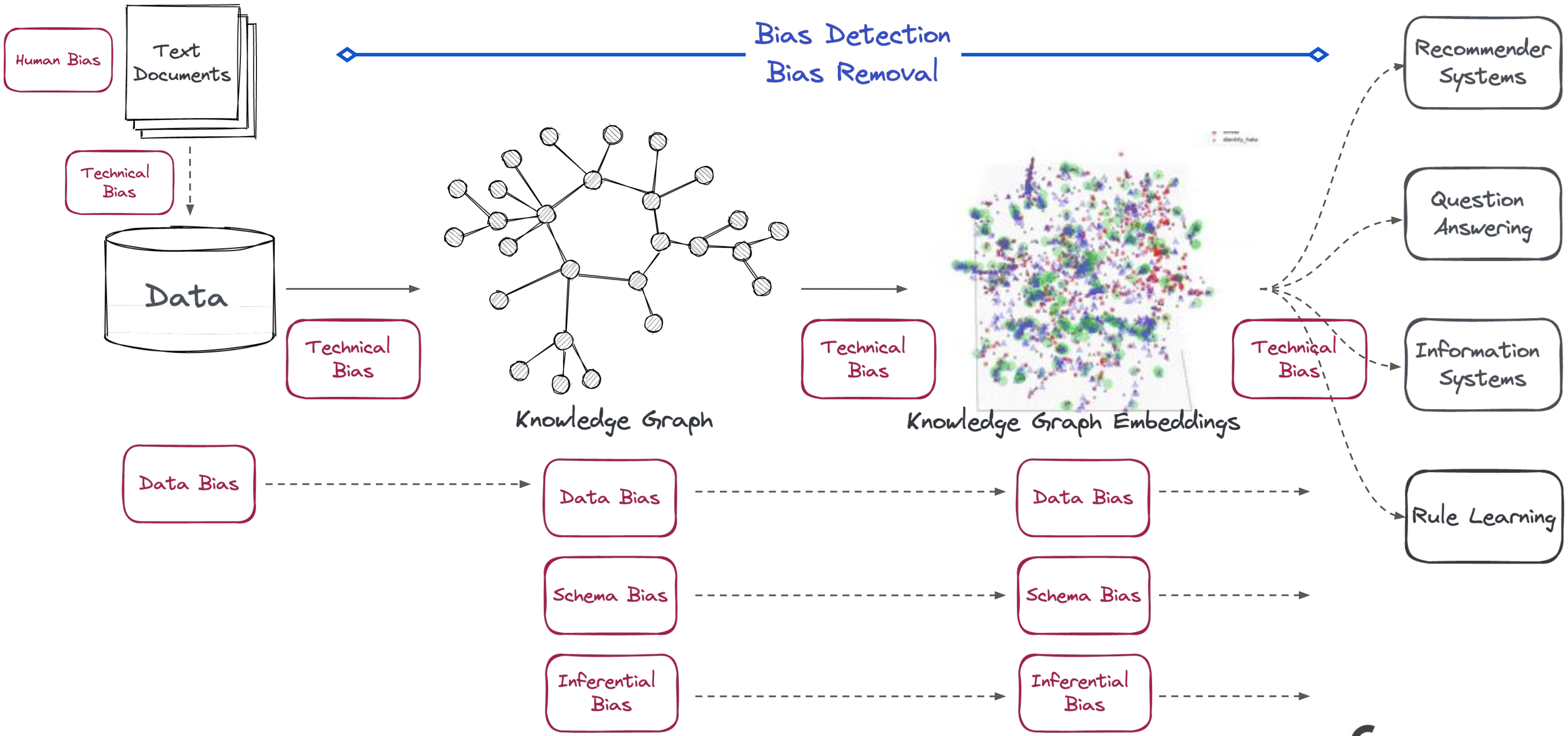
- **De-Biasing of KGE**
    - (**Bourli et al., 2020**): relies on specific detected bias on particular properties/classes (as e.g., in "occupations" and "gender") which can be balanced

    - **(Fisher et al., 2020):** trains all embeddings to be neutral with respect to sensitive relations (as e.g. "gender") by default using an adversarial loss. Sensitive information can be added back in for whitelisted cases (as e.g. "nationality" for "native language").

    - (**Arduini et al, 2020**): filtering out sensitive property information via adversarial learning (filter out, then try to predict, until acc=50%)

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

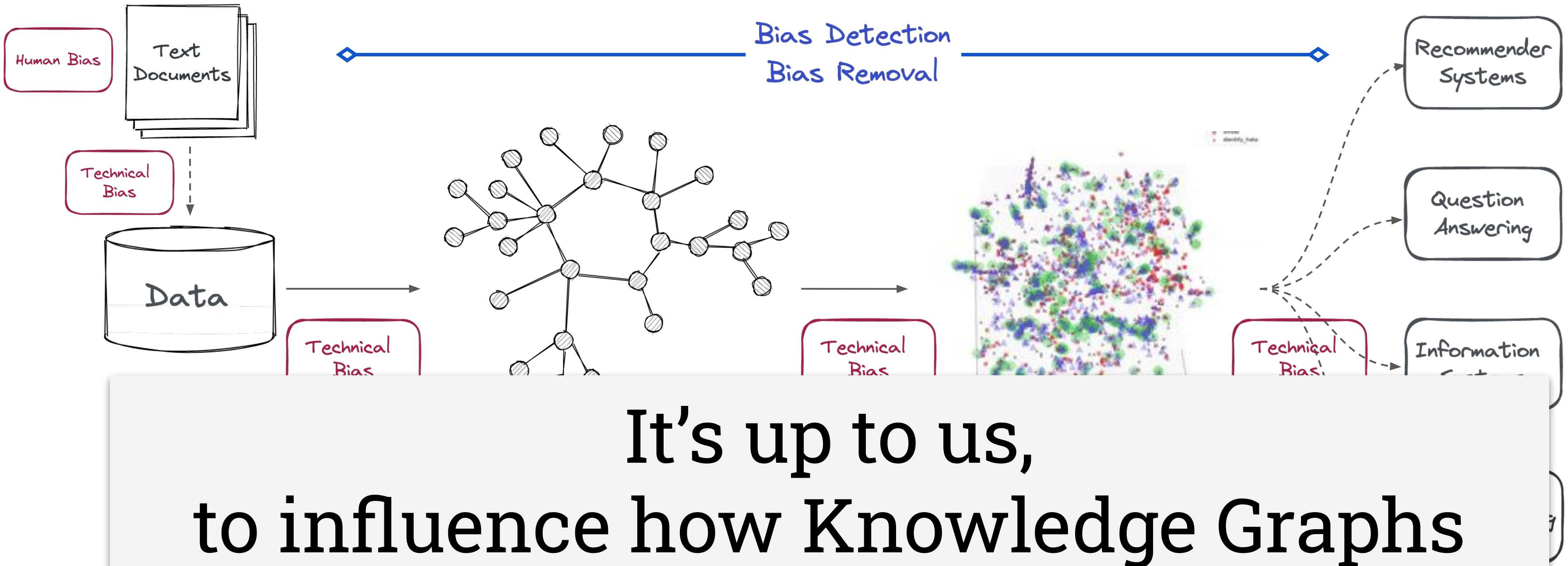
Biases in Knowledge Graphs



FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Prof. Dr. Harald Sack: “Social and technical biases in Knowledge Graphs”, Dagstuhl, 13.09.2022

Biases in Knowledge Graphs



It's up to us,
to influence how Knowledge Graphs
(and KG based systems)
evolve.

Bibliography:

[1] K. Janowicz, Bo Yan, Blake Regalia, R. Zhu, and Gengchen Mai. 2018. Debiasing knowledge graphs: Why female presidents are not like female popes. In International Semantic Web Conference.

[2] Janowicz, K., Hu, Y., McKenzie, G., Gao, S., Regalia, B., Mai, G., Zhu, R., Adams, B., Taylor, K.: Moon landing or safari? a study of systematic errors and their causes in geographic linked data. In: GIScience 2016. pp. 275–290. Springer (2016)

[3] Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57(10), 78–85 (2014)

[4] Radstok, W., Chekol, M.W., & Schäfer, M.T. (2021). Are Knowledge Graph Embedding Models Biased, or Is it the Data That They Are Trained on? Wikidata@ISWC.

[5] Demartini, G.: Implicit bias in crowdsourced knowledge graphs. In: Companion Proceedings of The 2019 World Wide Web Conference. p. 624–630. WWW '19, Association for Computing Machinery, New York, NY, USA (2019).

[6] Keidar, D., Zhong, M., Zhang, C., Shrestha, Y.R., & Paudel, B. (2021). Towards Automatic Bias Detection in Knowledge Graphs. EMNLP.

[7] Bourli, S., & Pitoura, E. (2020). Bias in Knowledge Graph Embeddings. 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 6-10.

[8] Tim Jones. 2019. Machine learning and bias. https://developer.ibm.com/articles/machine-learning-and-bias/

[9] Fisher, J. (2019). Measuring Social Bias in Knowledge Graph Embeddings. ArXiv, abs/1912.02761.

[10] Fisher, J., Mittal, A., Palfrey, D., & Christodoulopoulos, C. (2020). Debiasing Knowledge Graph Embeddings. EMNLP.

[11] Arduini, M., Noci, L., Pirovano, F., Zhang, C., Shrestha, Y.R., & Paudel, B. (2020). Adversarial Learning for Debiasing Knowledge Graph Embeddings. ArXiv, abs/2006.16309.