Debiasing Knowledge Graphs: Why Female Presidents are not like Female Popes

Krzysztof Janowicz, Bo Yan, Blake Regalia, Rui Zhu, and Gengchen Mai

STKO Lab, University of California, Santa Barbara, US
{janowicz,boyan,blake,ruizhu,gengchen_mai}@geog.ucsb.edu

Abstract. Bias, may it be in sampling or judgment, is not a new topic. However, with the increasing usage of data and models trained from them in almost all areas of everyday life, the topic rapidly gains relevance to the broad public. Even more, the opportunistic reuse of data (traces) that characterizes today's data science calls for new ways to understand and mitigate the effects of biases. Here, we discuss biases in the context of Linked Data, ontologies, and reasoning services and point to the need for both technical and social solutions. We believe that debiasing knowledge graphs will become a pressing issue as these graphs enter everyday life rapidly. This is a provocative topic, not only from a technical perspective but because it will force us as a Semantic Web community to discuss whether we want to debias in the first place and who gets a say in how to do so.

Keywords: Bias, Knowledge Graphs, Machine Learning, Ontologies

1 Introduction and Motivation

The *bias-variance* dilemma describes the struggle of trying to minimize the two main sources of error that plague machine learning algorithms in their attempt to generalize beyond their training data. While this trade-off relation is well known, it has only recently reached broad public attention due to the mainstream adoption of machine learning methods in everyday electronics and spectacular failures of their learned models such as categorizing black people as gorillas or auto-suggesting to kill all Jews in typeahead search.

While these cases generate the most public outcry, a different type of error may have subtle but more serious long-term consequences, namely representational bias such as Google's image search for "CEO" depicting mostly males. Even if this might still reflect today's reality and history, we would not want an intelligent system to *learn* this and consequently, for instance, recommend *doctor* as a career choice for men and *nurse* for women [1]. These kinds of biases are not merely a technological challenge; they are also a social issue.

There are ongoing discussions within the machine learning community about how to address representational biases (among other kinds), but these discussions have not yet reached the Knowledge Graph and Semantic Web communities despite representational issues being at their core. We believe that the heterogeneity of the Linked Data cloud, i.e., its decentralized nature and contributions

from many sources and cultures, can offer protections against biases to a certain degree, but this will come at a cost of increased *variances*.

Biases in knowledge graphs (KGs), as well as potential means to address them, differ from those in linguistic models or image classification. Instead of learning the meaning of a term by observing the context in which it arises within a large corpus or classifying buildings from millions of labeled images, KGs are sparse in the sense that only a small number of triples are available per entity. These triples are statements about the world, not usage patterns. Finally, it's important to recognize the term *bias* has different meanings across communities, implying that models can be unbiased in a machine learning sense yet simultaneously show substantial bias in a cultural context, e.g., when the training data do not reflect evolving social consensus; we address the latter kind of bias here.

In this vision paper, we describe biases from three different perspectives, namely those arising from the available data, those embedded in ontologies (i.e., the schema level), and finally those that are a result of drawing inferences. We illustrate each type with a small experiment.

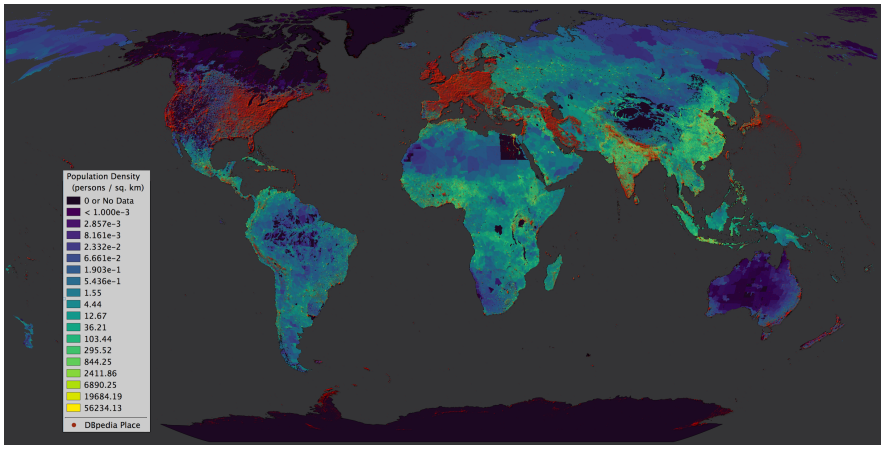

Fig. 1. Coverage of DBpedia (en) in contrast to population density.

2 Data Bias

The sheer size of the Linked Data cloud may lead to the impression that it is safe from selection biases such as sampling bias. However, the data available to both open and commercial knowledge graphs today is, in fact, highly biased. Coverage serves as an illustrative example of the underlying problem. Figure 1 plots more than one million geographically-located entities from DBpedia (en) in red. These entities, most of them being places such as Ford's Theater, link actors such as Abraham Lincoln, to events such as his assassination, to objects such as

Booth's weapon. Consequently, the global coverage of places reveals much about the underlying distribution of non-spatial data as well. The map also shows Landscan-based estimates of population density. As can be seen, coverage is not uniform. For Europe, Japan, Australia, and the US, marker density strongly correlates with population density, whereas this trend breaks for large parts of Asia, Africa, and South America. At first, one may expect that these results are reflections of using the English DBpedia version, however, the resulting pattern largely remains the same when comparing other language versions and data sources such as GeoNames, social media postings, or even government data [7].

Put more provocatively, we know a lot about the western world and we do so from a western perspective. Even more, most of what we know about other regions and cultures comes from this same western perspective. This is an artifact of history, differences in cultures of data sharing, availability of free governmental data, financial resources, and so forth. The effects are similar to what has been discussed in the machine learning community with respect to biases in word embeddings or in image search and tagging. This is not a minor issue. It translates into biased knowledge graph embeddings that increase dissimilarity to those less prototypical cases, it influences recommender system and question answering, and it learns rules from biased training data.

As a substantial part of the Linked Data technology stack is based on open standards and is available as open source implementations, we hope that a more diverse set of contributors will form around it, thereby making the Linked Data cloud more robust to biases, financial incentives, and so forth.

3 Schema Bias

In addition to data bias, knowledge graphs encounter another type of bias at the schema/ontologies level. Most ontologies are developed in a top-down manner with application needs in mind, or in case of top-level ontologies, certain philosophical stances. They are typically defined by a group of engineers in collaboration with domain experts, and thus also implicitly reflect the worldviews and biases of the development team. Such ontologies will likely contain most of the well-known human biases and heuristics such as anthropocentric thinking. The increasing use of bottom-up techniques such as machine learning to derive axioms/rules from data will not mitigate these problems as the resulting models will fall victim to the data biases discussed before. In addition, ontologies may be affected by other biases such as so-called encoding bias [4].

Anthropocentric thinking and application needs are, for instance, at play in many of the vocabularies used to describe Points Of Interest. They typically contain dozens or even hundreds of classes for various sub-classes of restaurants, bars, and music venues, but only a handful of classes for natural features such as rivers. Consider, for example, the *Place* branch of the DBpedia ontology (2015); although it contains 168 classes, the average in- and out-degree of each class is only 0.994 when the class hierarchy is visualized as a network. The average path length is 2.18, suggesting that it is relatively shallow and flat. By running

PageRank with a teleport probability of 0.85, more than 75% of the classes have a near zero PageRank score, meaning that much of its expressivity is occupied by less than 25% of classes. Simply put, the relative importance of classes is unevenly distributed in the ontology despite its general purpose characteristics.

It is worth noting that many biases are not directly encoded in the ontology but only become visible when comparing multiple ontologies together with their respective datasets. For example, DBpedia, GeoNames, and the Getty Thesaurus of Geographic Names (TGN) all contain a *Theater* class. From a data-driven perspective, one may assume that computing (spatial) statistics for all members of this class, such as intensity, interaction, and point patterns, would yield similar results across datasets [8]. However, this is not the case and those indicators will show very distinct patterns. The reason for this is that 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. Put differently, the dissimilarity in the extension of these classes is an expression of the implicit biases across the classes despite their common name.

We believe that the multitude of ontologies developed for the Semantic Web are a strength rather than a weakness as their diversity may help to mitigate some of the issues outlined above.

4 Inferential Bias

Another potential source of bias arises at the inferencing level, such as reasoning, querying, or rule learning. To start with a simple but easily overlooked example, the results of a SPARQL query depend on the entailment regimes (e.g., simple vs. RDFS entailment). However, the configuration of a particular query endpoint is outside the control of a data creator or ontology engineer and thus one may find that multiple endpoints utilizing the same ontologies and datasets, e.g., local copies of DBpedia, yield different results for the same SPARQL query.

Here we focus on another aspect, namely, the relation of learning a (correct) model that collides with social consensus. Consider for example, the use of association rule mining to infer new rules from knowledge graphs[2]. For our experiment we extracted all popes, US 5-star generals, and US presidents from DBpedia. These entities have one aspect in common: they are all male[1]. Running AMIE+ over this graph results in numerous rules and the following 3 have very high confidence scores due to a lack of negative examples: **(1)** *if X is a pope, X is male*; **(2)** *if X is a US 5-star general, X is male*; and **(3)** *if X is a US president, X is male*. While these rules may be perceived as controversial, they are all correct. The provocative point we are trying to make here is that these rules are correct for different reasons. The first case is true by definition of the concept *Pope*; hence, the learned rule is suitable for its application. The second case is an enumerated class as *US 5-star general* is a historical military rank no longer in use. Hence, while the rank is not exclusive to men, the rule still applies

[1] A triple we had to add to our graph for this example as it is not present in DBpedia

to all cases despite not being useful since it will not generate new triples. The third case is most controversial as it applies to the entire training data but does not align with social consensus. We do not want KG-based question answering systems to suggest to users that only men can become US presidents in a similar fashion to today's systems recommending women to become nurses.

5 Conclusions

In this vision paper, we highlighted the need to establish debiasing knowledge graphs as a novel research theme for the Semantic Web / Knowledge Graph community that *differs* from current mainstream research. We highlighted multiple sources of bias using small experiments. We believe that the topic is *provocative* as it walks the fine line between social responsibility and censorship. *Risk* mainly arises from the fact that debiasing itself is not a neutral task but based on social norms that may differ by countries. Will we develop methods that can be used for censorship and manipulation? As far as the *time horizon* is concerned, we believe that this will become an equally pressing issue for the SW community as it is currently in machine learning and that it should be openly addressed in workshops or panels. Finally, the topic may also be approached from a Web Science [5] perspective as well as by considering the interplay of trust and provenance [6, 3].

References

1. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Advances in Neural Information Processing Systems. pp. 4349–4357 (2016)
2. Galárraga, L.A., Teflioudi, C., Hose, K., Suchanek, F.: Amie: association rule mining under incomplete evidence in ontological knowledge bases. In: Proceedings of the 22nd international conference on World Wide Web. pp. 413–422. ACM (2013)
3. Golbeck, J., Parsia, B., Hendler, J.: Trust networks on the semantic web. In: International Workshop on Cooperative Information Agents. pp. 238–249. Springer (2003)
4. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing? International journal of human-computer studies **43**(5-6), 907–928 (1995)
5. Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., Weitzner, D.: Web science: an interdisciplinary approach to understanding the web. Communications of the ACM **51**(7), 60–69 (2008)
6. Janowicz, K.: Trust and provenance you cant have one without the other. Tech. rep., Technical Report, Institute for Geoinformatics, University of Muenster, Germany (2009)
7. 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)
8. Zhu, R., Hu, Y., Janowicz, K., McKenzie, G.: Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. Transactions in GIS **20**(3), 333–355 (2016)