A Key Information

A.1 Hosting, licensing, and maintenance plan

We used the latest Freebase data dump available at `https://developers.google.com/freebase` to generate our datasets. Freebase data dump is distributed under the Creative Commons Attribution (aka CC-BY). Our datasets along with the scripts required to generate them from Freebase datadump are available in our GitHub repository `https://github.com/idirlab/freebases`, licensed under the CC-0 license. The Innovative Data Intelligence Research Lab at UTA, where the authors are, is committed to maintaining the datasets. The lab has a track record of maintaining multiple research prototypes, demos, and datasets at `https://idir.uta.edu/`. There are several projects underway in our lab using these datasets and we will update our GitHub repository with the latest results from these projects. Any further updates to the datasets will be posted there too. Furthermore, we plan to archive the datasets at Zenodo upon paper acceptance.

A.2 Intended uses

Our datasets are intended to be used by researchers and practitioners in developing technologies based on and for knowledge graphs.

A.3 Limitations

Freebase was shut down in 2015 and its latest data dump (used by us to generate our datasets) was made available in Aug of 2015. Hence, when using the datasets one should be aware of the probability of some triples being outdated. It is also worth mentioning that Freebase domains are not equally distributed. Although Freebase covers a diverse range of domains, not all of them are well represented. For example, domain */music/* constitutes around 6.7% of the triples while domain */tennis/* accounts for only 0.001% [13].

A.4 Potential negative impacts

As it was mentioned, Freebase domains are distributed unevenly. Distribution of entities and relations is also skewed, with some entities and relations being more popular than others. If a model is trained on our datasets and is used for some real-world tasks, the bias available in data may be present in the model's output as well.

A.5 Author statement

We bear all responsibility in case of violation of rights and confirm CC-0 licenses for the included datasets.

B Details of datasets, dataset format, and dataset creation scripts

The datasets and data preprocessing scripts are made publicly available at `https://github.com/idirlab/freebases`.

The data is stored in CSV files in the form of triples which is a widely used data format for storing knowledge graph data.

As discussed in Section 6 of the paper, we provide four variants of the Freebase dataset by inclusion/exclusion of some of the Freebase's idiosyncrasies. For each of these datasets, we made three kinds of files available:

- Metadata files:
  - `object_types`: Each row maps the MID of a Freebase object to a type it belongs to.

  - `object_ids`: Each row maps the MID of a Freebase object to its user-friendly identifier.
  - `object_names`: Each row maps the MID of a Freebase object to its textual label.
  - `domains_id_label`: Each row maps the MID of a Freebase domain to its label.
  - `types_id_label`: Each row maps the MID of a Freebase type to its label.
  - `entities_id_label`: Each row maps the MID of a Freebase entity to its label.
  - `properties_id_label`: Each row maps the MID of a Freebase property to its label.
- Subject matter triples file: `fb`x, where $x \in 1, 2, 3, 4$. For each variant, depending on the nature of a task, one can choose to use one of these. For example, to exclude reverse relations but to retain CVT nodes, one can use table `fb3`. All four variants are explained in Section 6 of the paper.
- Type system file: `freebase_endtypes`. This table is built to provide the type system (Section 3 of the paper) for the dataset. Each row in this table maps an edge type to its required subject type and object type.

We also provided three types of scripts for URI simplification (parse_triples.sh), metadata separation (`FBDataDump.sh`), and processing the subject matter triples (`fb`x`.sh`, where $x \in 1, 2, 3, 4$). These scripts are used to generate the aforementioned four variants (discussed in Section 6) and their type systems.

C Basic statistics of the datasets

Statistics of our datasets can be found in Table 2 of the paper.

D Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Described in Section A.3 above.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Described in Section A.4.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Provided at the URL described in Section B.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] They are explained in Section 7 of the paper and at the URL provided in Section B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] They are explained in Section 7 of the paper and at the URL provided in Section B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We have cited all the datasets, frameworks, models, and previous work.

(b) Did you mention the license of the assets? [Yes] It is explained in Section A.1 as well as at the URL in Section B.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The datasets and data preprocessing scripts are made publicly available at the URL in Section B.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]