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.

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 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 5 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 5 of the paper.
- Type system file: `freebase_endtypes`. This table is built to provide the type system 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 5) and their type systems.

C Overview of Existing Freebase Datasets

Over the past decade, several datasets were created from Freebase. This section reviews some of these datasets and briefly discusses flaws associated with them.

FB15k [7] includes entities with at least 100 appearances in Freebase that were also available in Wikipedia based on the *wiki-links* database [11]. Each included relation has at least 100 instances. 14,951 entities and 1,345 relations satisfy these criteria, which account for 592,213 triples included into FB15k. These triples were randomly split into training, validation and test sets. This dataset suffers from data redundancy in the forms of reverse triples, duplicate and reverse-duplicate relations. Details of these issues were discussed thoroughly in [2].

FB15k-237 [36], with 14,541 entities, 237 relations and 309,696 triples, was created from FB15k in order to mitigate the aforementioned data redundancy. Only the most frequent 401 relations from FB15k are kept. Near-duplicate and reverse-duplicate relations were detected, and only one relation from each pair of such redundant relations is kept. This process further decreased the number of relations to 237. This step could incorrectly remove useful information, in two scenarios. 1) False positives. For example, hypothetically *place_of_birth* and *place_of_death* may have many overlapping subject-object pairs, but they are not semantically redundant.

Freebase86m is created from the last Freebase data dump and is employed in evaluating large-scale knowledge graph embedding frameworks [49, 25]. It includes 86,054,151 entities, 14,824 relations and 338,586,276 triples. No information is available on how this dataset was created. We carried out an extensive investigation on this dataset to assess its quality. We found that 1) 31% of the triples in this dataset are non-subject matter triples from Freebase implementation domains such as */common/* and */type/*, 2) 23% of the dataset's nodes are mediator nodes, and 3) it also has abundant data redundancy since 38% of its triples form reverse triples. As discussed in Section 3, non-subject matter triples should be removed; reverse triples, when not properly handled, lead to substantial over-estimation of link predication models' accuracy; and the existence of mediator nodes presents extra challenges to models. Mixing these different types of triples together, without clear annotation and separation, leads to foreseeably unreliable models and results. Section 6 discusses in detail the impact of these defects in Freebase86m.