1. *Reviewer 27CX*

Decent work can be polished better

Summary And Contributions:

this paper lays out a comprehensive analysis of the challenges associated with the aforementioned idiosyncrasies of Freebase, which is amongst the largest public cross-domain KGs that store common facts.

Rating: 5: Marginally below acceptance threshold

Strengths:

Comprehensive postprocessing, data organization and cleaning have been done. The work may help people better use freebase as a large existing knowledge graph database.

Confidence: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper

Weaknesses:

1-No new dataset or benchmark are proposed.

2-The organization and writing of the paper can be further polished.

3-Two minor issues make a bad impression:

This year is 2022 but the authors used the 2021 template.
The large gap in reference one looks very unprofessional.

Correctness:

Looks reasonable to me although I did not get chance to examine every details.

Clarity:

In generally clear, better writing welcomed.

Relation To Prior Work:

I am not sure whether there are similar work already done in this direction.

Documentation:

There are some useful information in the github repo the authors provide, but far from complete and comprehensive. I suggest the authors put in more work.

FEEDBACK

We sincerely thank the reviewer as their thoughtful comments helped us improve the paper significantly. The revised paper is uploaded, with more substantial changes highlighted in blue. The major changes are:

- New paragraphs in Section 1 to summarize significant contributions made by the datasets and results.

- Section 4.1 now discusses how the datasets and experimentation design can enable comparison with models built for hyper-relational facts and comparison on Wikidata.

- Section 7 now includes experiment results on FB3 and FB4 and other tasks (triple classification).

Please see below for detailed response.

1. Weakness #1 “No new dataset or benchmark are proposed.”

[Response]

We created four new variants of Freebase: FB1, FB2, FB3, and FB4. Although they are from Freebase, they are new, the differences are important, and their creation is nontrivial. Section 1 provides a summary of the work’s significant contributions. Below are a few highlights.

- “The paper fills an important gap in dataset availability. ... ours is the first-ever publicly available full-scale Freebase dataset that has gone through proper preparation.”

- “The paper also fills an important gap in our understanding of knowledge graph embedding models. … The experiments on our datasets ... inform the research community several important results that were never known before, including 1) the true performance of link prediction embedding models on the complete Freebase ...; and 2) how data idiosyncrasies ... impact model performance on the complete Freebase data.”

- “The dataset creation was nontrivial and time-consuming. It required extensive inspection and complex processing of the massive Freebase data dump, for which documents are scarce. ….we are unaware of more detailed description of these idiosyncrasies anywhere else. If researchers must learn to examine Freebase and prepare datasets from scratch, we expect many of them to have steep learning curve and the process can easily require months or even years.”

We also would like to point out that the conference website mentions “identifying significant problems with existing datasets and their use” is within its scope, at https://neurips.cc/Conferences/2022/CallForDatasetsBenchmarks. Our submission discusses several idiosyncrasies of Freebase and the associated problems of existing datasets.

2. Weakness #2 “The organization and writing of the paper can be further polished.” and Clarity: “In generally clear, better writing welcomed.”

[Response] We thoroughly revised, expanded, and polished the paper. The more substantial changes are highlighted in blue in the paper. The major changes were listed above. If there are further comments and suggestions, we will be very glad to address them.

3. Weakness #3 “Two minor issues make a bad impression”

[Response]

“This year is 2022 but the authors used the 2021 template”: when we downloaded it from the conference website, it was actually the 2021 version, although it is now the 2022 version. We did notice it, but we did not change anything since that was provided on the conference website. Nevertheless, we have changed it to reflect 2022.

Regarding “The large gap in reference one looks very unprofessional”, the gap is automatically produced by Latex. The web page title we used in the reference is exactly how it is on the website, i.e., no space before or after the column in “Wikidata:WikiProject Freebase”. Nevertheless, we have shortened the URL which makes it look slightly better.

4. “I am not sure whether there are similar work already done in this direction.”

[Response]

There is no similar prior work in this direction. The newly added paragraphs in Section 1 clarify the work's significance, including its novelty. Here are a few highlights:

"ours is the first-ever publicly available full-scale Freebase dataset that has gone through proper preparation."

"The experiments on our datasets ... inform the research community several important results that were never known before."

"In fact, we are unaware of more detailed description of these idiosyncrasies anywhere else."

"... it becomes possible to compare the performance of conventional embedding models and hyper-relational fact models on a full-scale Freebase dataset that includes multiary relationships .. Such a comparison will be the first not only for Freebase but also for any large-scale knowledge graph."

5. "There are some useful information in the GitHub repo the authors provide, but far from complete and comprehensive. I suggest the authors put in more work."

[Response] Thank you for the suggestion! We have thoroughly revised and expanded the documentation in the GitHub repository. Newly added information includes: new data files such as the URI mappings; more experiment results; detailed README.md including the statistics and examples for each dataset file, data preparation scripts usage, scripts of the experiments; and local copy of dgl-ke and LibKGE.

---

2. *Reviewer Dm86*

Review of "Creating Variants of Freebase for Robust Development of Intelligent Tasks on Knowledge Graphs"

Summary And Contributions:

This paper mainly discusses three idiosyncrasies of Freebase: a strong type system, reversed triples and mediator nodes to represent n-ary relationships. The author claims that these characteristics are powerful as wel as challenging when it comes to the advancement of KG-oriented technologies. They measure the impacts on several KG tasks such as link prediction and proprose several variants of the Freebase dataset with or without the idiosyncrasies.

Rating: 5: Marginally below acceptance threshold

Strengths:

As is stated, a strong type system could help to narrow down search in several KG and NLP areas such as query systems and could serve as an effective filter for multiple answers. They perform contrastive experienments to examine the factors.

Confidence: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper

Weaknesses:

We ca only find the analysis of the experienments on one idiocrasy, what about the other two?

Correctness:

In the table4, the link prediction result of dismult is lower than that in Toutanova[1] and we do not know whether a fitered measure is deployed in case of multiple answers.

[1] Toutanova, K., Chen, D., Pantel, P., Poon, H., Choudhury, P., & Gamon, M. (2015, September). Representing text for joint embedding of text and knowledge bases. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 1499-1509).

Clarity:

This paper is mostly well writen but contains broken parts: there idiocrasies have been listed but only one of them has been explored by experiments.

FEEDBACK

We would like to express our sincere gratitude to the reviewer for their thoughtful comments which have helped us improve the paper significantly. The revised paper is uploaded and the more substantial changes are highlighted in blue color in the paper. The major changes in the paper include:

- Several new paragraphs at the end of Section 1 to summarize the significant contributions made by the datasets and results from this work.

- Section 4.1 “Mediator Nodes” is expanded to discuss how the datasets and experimentation design can enable comparison with models built for hyper-relational facts and comparison on other datasets such as Wikidata.

- Section 7 Experiments is expanded to discuss new experiments on FB3 and FB4 and on other tasks such as triple classification.

Please see below for our detailed response to the review comments.

1. “Weaknesses: We ca only find the analysis of the experienments on one idiocrasy, what about the other two?”

[Response]

Thank you for this important comment, which inspired us to conduct more experiments. The paper now includes experiments on all three idiosyncrasies discussed.

(1.1) The experiments in the initial submission focused on how reverse triples (one of the idiosyncrasies) impact link prediction, using datasets FB1 and FB2.

(1.2) We have conducted more experiments on FB3 and FB4 as well and revised Section 7 of the paper to include the results from these new experiments, which demonstrate the impact of mediator (CVT) nodes, another idiosyncrasy discussed in the paper.

(1.3) Section 7 of the paper now also includes experiment results that show how Freebase type system we created (the third idiosyncrasy discussed in the paper) can help generate more difficult and realistic negative examples for the task of triple classification and thus avoid overestimating model performance. This part of the experiments was conducted on FB15k-237. We are doing the same experiments on our large-scale datasets FB1, FB2, FB3, and FB4, but it will require more time, given that each of these is a full-scale Freebase dataset. We are confident that we will have the results way before the final version of the paper is due for publishing, and we welcome a shepherding processing from the conference to ensure such results are included in the paper.

2. “Correctness: In the table4, the link prediction result of dismult is lower than that in Toutanova[1] and we do not know whether a fitered measure is deployed in case of multiple answers.”

[Response]

All reported results, including the original Table 4, use filtered measures. We have now clarified this in Section 7.

We have now removed the original Table 4, which was on small-scale FB15k/FB15k-237 but the paper's focus is on full-scale Freebase datasets. We believe it brings little value to keep the table while it may generate confusions such as this comment. Nevertheless, we'd like to address the reviewer's comment, as follows.

Due to different implementations and hyperparameters, results of the models might differ in different studies. For example, the hidden dimension is set as 400 in the framework we used (DGLKE), but 500 in Toutanova[1]. Generally speaking, large-scale frameworks such as DGLKE do not feature hyperparameter optimization and they usually focus on parallelizing training methods and models. This may lead to different results from frameworks and models that perform extensive hyperparameter optimization.

We would like to further point out that the objective of the experiments in the paper is not to compare different embedding models or optimize any particular model. Rather, it is to show the impacts of the three idiosyncrasies (mediator/CVT nodes, reverse triples, type system) on such models, as well as to show how they perform differently on existing small vs. new large-scale Freebase datasets. Our experiment results indeed show consistent observations across multiple representative models.

3. "Clarity: This paper is mostly well writen but contains broken parts: there idiocrasies have been listed but only one of them has been explored by experiments."

[Response]

We have thoroughly revised, expanded, and polished the paper. As mentioned above, the more substantial changes are highlighted in blue color in the paper revision. The major changes were also listed above. If there are further comments and suggestions on places that need improvement, we will be very glad to address them.

As mentioned in our response to "1. Weaknesses", with the newly included results, the paper includes experiment results on all three idiosyncrasies discussed.

---

3. *Reviewer Nfun*

Variants of Freebase dataset

Summary And Contributions:

The authors present four graph datasets based on Freebase that include or exclude various idiosyncrasies, such as freebase type system, reverses triples, and mediator nodes. They evaluate link prediction models on two of their Freebase datasets (FB1, FB2) to analyze the effects of such idiosyncrasies on tasks such as link prediction.

Rating: 4: Ok but not good enough - rejection

Strengths:

The paper is well-written and discusses the effects of these idiosyncrasies on downstream modeling tasks.

The authors evaluate several link prediction models

Confidence: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper

Weaknesses:

1-The new datasets proposed in this paper are based on Freebase, which seems to be an outdated dataset, replaced by Wikidata.

2-The authors only evaluate methods and perform experiments on FB1 and FB2 datasets, not FB3 and FB4.

3- Modifications made to the Freebase dataset are incremental data preprocessing steps.

Correctness:

The authors do not perform experiments on datasets FB3 and FB4.

Clarity:

The paper is well-written and easy to follow.

Relation To Prior Work:

The paper includes discussion of existing Freebase datasets, but does not discuss other datasets for knowledge graph modeling.

FEEDBACK

We sincerely thank the reviewer as their thoughtful comments helped us improve the paper significantly. The revised paper is uploaded, with more substantial changes highlighted in blue. The major changes are:

- New paragraphs in Section 1 to summarize significant contributions made by the datasets and results.

- Section 4.1 now discusses how the datasets and experimentation design can enable comparison with models built for hyper-relational facts and comparison on Wikidata.

- Section 7 now includes experiment results on FB3 and FB4 and other tasks (triple classification).

Please see below for detailed response.

1. Weakness #1 “The new datasets proposed in this paper are based on Freebase, which seems to be an outdated dataset, replaced by Wikidata.”

[Response]

Please refer to the several new paragraphs at the end of Section 1 which summarize our contributions. Particularly, the following point is the most relevant to this comment: “The datasets and results are highly relevant to the research community, as Freebase remains the single most commonly used dataset for link prediction, by far. We examined all full-length publications that 1) appeared in 12 top conferences during their latest years and 2) used datasets commonly used for link prediction. This amounts to 63 publications. Among them, 57 publications used datasets produced from Freebase, while 9 used datasets from Wikidata. Only 2 publications used a Freebase dataset at its full scale, specifically Freebase86m. This suggests that researchers may not be able to carry out large-scale study due to the lack of proper datasets.” (The list of the 63 publications and more details can be found in Section 1 of the paper.)

A less important feedback is that, content-wise, Wikidata cannot replace Freebase yet. Although there was a plan of transferring Freebase to Wikidata, that process was only partially accomplished and appears to be stagnant. (See references [1] and [30] in the revised paper.)

2. Weakness #2 “The authors only evaluate methods and perform experiments on FB1 and FB2 datasets, not FB3 and FB4.” and “Correctness: The authors do not perform experiments on datasets FB3 and FB4.”

[Response] Section 7 now includes the results of new experiments. One such experiment is on evaluating link prediction models on FB3 and FB4, which demonstrates the impact of mediator (CVT) nodes on link prediction. Another new experiment is triple classification to show how Freebase type system can help generate more realistic negative samples and avoid overestimating the performance of the models.

3. Weakness #3 “Modifications made to the Freebase dataset are incremental data preprocessing steps.”

[Response]

We would like to point out a few important ways in which the datasets were created differently from prior work, particularly FB15k-237.

(1) The first main difference is the way reverse relations are identified. FB15k-237 is created from FB15k by removing redundant relations. Given two relations, they calculate how much their subject-object pairs overlap and, if the overlap is greater than a threshold, they consider them as redundant. This step could incorrectly remove useful information, due to two types of mistakes. A) False positives: For example, place_of_birth and place_of_death may have many overlapping subject-object pairs, but they are not semantically redundant. B) False negatives, since FB15k-237 did not resort to the accurate reverse relation information encoded by reverse_property in Freebase. For example, we observed that FB15k-237 includes both /education/educational_institution_campus/educational_institution and /education/educational_institution/campuses but they are reverse relations according to reverse_property.

(2) Another main difference is that, in FB15k and FB15k-237, relations connected through mediator nodes are already concatenated, and it is not explained how the concatenation was made. We did extensive investigation to identify mediator (CVT) nodes and then created different variants by including CVT nodes or by excluding them through concatenation. The detection and concatenation of CVT nodes is complex and the procedure was not available from FB15k-237.

(3) Moreover, we extracted and included the Freebase type system in our datasets, which is not available in FB15k-237. Our analysis and observations about Freebase type system, which is discussed in section 3, are mostly identified by the authors of this paper and have not been discussed in any other paper.

Furthermore, section 1 now summarizes the significance of the datasets, including how nontrivial their creation is, as follows: “The dataset creation was nontrivial and time-consuming. It required extensive inspection and complex processing of the massive Freebase data dump, for which documents are scarce. None of the idiosyncrasies, as articulated in Sections 3 and 4, was defined or detailed in the data dump itself. Figuring out the details in these sections required iterative trial-and-error in examining the data. In fact, we are unaware of more detailed description of these idiosyncrasies anywhere else. If researchers must learn to examine Freebase and prepare datasets from scratch, we expect many of them to have steep learning curve and the process can easily require months or even years. Our datasets can thus speed up many researchers’ work.”

4. “Relation To Prior Work: The paper includes discussion of existing Freebase datasets, but does not discuss other datasets for knowledge graph modeling. “

[Response]

The focus of this study is on Freebase. The datasets and results are highly relevant to the research community, as Freebase remains the single most commonly used dataset for link prediction, by far, as mentioned in our response to Weakness #1 above.

Furthermore, Section 4.1 now includes a detailed discussion of hyper-relational fact, which is used for modeling multiary relationships in both Wikidata and Freebase. It further went on to explain that "given the datasets and experiment results made available in this paper, it becomes possible to compare the performance of both hyper-relational fact models and conventional models on a full-scale Freebase dataset that includes multiary relationships (specifically, FB3 in Table 1 of Section 6). Such a comparison will be the first not only for Freebase but also for any large-scale knowledge graph. For instance, to the best of our knowledge, there does not exist a full-scale Wikidata dataset with multiary relationships represented as conventional triples instead of hyper-relational facts. Therefore, conventional models have only been applied on Wikidata without multiary relationships [42], e.g., OGBL-WikiKG2 [24]. There exists no comparison of the two categories of models on Wikidata with multiary relationships."

Section 1 summarizes these discussions and further explains that "The experiment design could be similarly extended to study the impact of multiary relationships in Wikidata on embedding models and compare such models with models built for hyper-relational facts."

---

4. *Reviewer Vdxi*

Useful iteration on existing dataset with (limited) incremental value

Summary And Contributions:

The paper presents explanations of the "idiosyncrasies" of Freebase: mediator nodes, reverse triples and (administrative) metadata, and describes the challenges they introduce for data modelling tasks. The paper introduces 4 different intersections of the Freebase dataset by varying the inclusion of 2 *idiosyncrasies*: reverse triples and mediator nodes. Moreover, each intersecting dataset includes "required types" per subject or object as generated by an automated procedure based on occurrence frequencies (?) among instances. The paper concludes with a performance analysis of link prediction models on 2 out of 4 datasets in comparison with related extractions of Freebase (e.g. FB15k-237) to pinpoint differences.

Rating: 5: Marginally below acceptance threshold

Strengths:

S1: the paper provides in-depth background and discussion of the idiosyncrasies and challenges arising from them.

S2: the paper presents a substantial evaluation of various link prediction models on different datasets.

S3: the datasets address challenges that may make KG tasks like link prediction invalid, for example, due to leakage.

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

Weaknesses:

W1: The presented datasets sure have value for defeating invalid implementations of link prediction but the contribution as-is is incremental. The data cleaning steps, for example, are straightforward and overlap with prior datasets (e.g. FB15k-237). The proposed URI "simplification" seems inconvenient as it introduces a disconnect between the source dataset and the presented dataset; these simplifications could better be added as metadata. In addition, the value of the datasets are demonstrated for link prediction only.

W2: The motivation and analysis are on the weak side and at some points rely on speculation. This undermines the significance of the dataset.

Motivation: it is not clear why Freebase was chosen over e.g. DBpedia/Wikidata which are still heavily maintained and updated, and share some of the important properties as discussed in this paper. In the end, the key differentiating *idiosyncrasies* (reverse triples and mediator nodes) of Freebase are suggested to be removed from the dataset as they impose challenges for e.g. link prediction, which weakens the motivation for choosing FB.

Analysis: Some statements and conclusions seem to rely on speculations instead of analyses or facts. For example, ln. 316-318 gives an example of similar relations "place_of_birth" and "place_of_death" that *could* have been removed in the related FB15k-237 but it is unclear if this observed or implied by the procedural description of the FB15k-237 dataset as far as I can tell. Furthermore, based on a performance comparison between FB1/2 and FB15k-237 the observation of a lower accuracy on the latter is attributed to the size of the datasets but this seems ungrounded since FB15k-237 has additional differences, like the removal of trivial linkable relations, that may explain a lower performance as well. A more careful analysis is needed.

W3: The most novel enhancement of the dataset seems to be the enhancement of "required types" of objects/subjects based on some probability, but it is unclear how these probabilities are obtained, what threshold is used (and why), and an analysis of the correctness of the resulting types is missing. The added value of these types is also not supported by experiments or analysis.

Correctness:

The dataset construction seems generally correct, but as pointed out in the previous section the technical details relating to the type enhancement are missing and the decision to transform the URIs is disputable.

Clarity:

The paper is well written: the narrative is clear, and the relevant background concepts are explained carefully which makes the paper easy to follow. Section 3 and 4 present explanations and challenges of the idiosyncrasies, but these sections partly overlap and may be merged. The experiment section introduces the task of link prediction but does not formulate the problem at hand.

Relation To Prior Work:

Related datasets are discussed in a separate section and the link prediction experiment incorporates results on the most related datasets.

Documentation:

Experiments are implemented using the DGL-KE framework and therefore rely on its maintenance, but this seems OK. The authors may consider getting local copy of it. The scripts for creating the datasets are present and the data is currently hosted on Dropbox (it is claimed to be transferred to Zenodo later).

Additional Feedback:

Some minor comments:

- Table 2 shows statistics of the introduced FB intersections but while the #entities, #relations and #triples are reported for the related datasets, this table shows #entities, #properties and #triples. Is the #properties a mistake?
- Is "idiosyncrasies" the most appropriate term for the FB properties? This term may not be known to every reader and I am not sure if the meaning of it corresponds to the intended meaning of the authors.

FEEDBACK

We would like to express our sincere gratitude to the reviewer for their thoughtful comments which have helped us improve the paper significantly. The revised paper is uploaded and the more substantial changes are highlighted in blue color in the paper. The major changes in the paper include:

- New paragraphs in Section 1 to summarize significant contributions made by the datasets and results.

- Section 4.1 now discusses how the datasets and experimentation design can enable comparison with models built for hyper-relational facts and comparison on Wikidata.

- Section 7 now includes experiment results on FB3 and FB4 and other tasks (triple classification).

Please see below for detailed response.

1. "W1: ...but the contribution as-is is incremental. The data cleaning steps, for example, are straightforward and overlap with prior datasets (e.g. FB15k-237)."

We would like to point out a few important ways in which the datasets were created differently from prior work, particularly FB15k-237.

(1) The first main difference is the way reverse relations are identified. FB15k-237 is created from FB15k by removing redundant relations. Given two relations, they calculate how much their subject-object pairs overlap and, if the overlap is greater than a threshold, they consider them as redundant. This step could incorrectly remove useful information, due to two types of mistakes. A) False positives: For example, place_of_birth and place_of_death may have many overlapping subject-object pairs, but they are not semantically redundant. B) False negatives, since FB15k-237 did not resort to the accurate reverse relation information encoded by reverse_property in Freebase. For example, we observed that FB15k-237 includes both /education/educational_institution_campus/educational_institution and /education/educational_institution/campuses but they are reverse relations according to reverse_property.

(2) Another main difference is that, in FB15k and FB15k-237, relations connected through mediator nodes are already concatenated, and it is not explained how the concatenation was made. We did extensive investigation to identify mediator (CVT) nodes and then created different variants by including CVT nodes or by excluding them through concatenation. The detection and concatenation of CVT nodes is complex and the procedure was not available from FB15k-237.

(3) Moreover, we extracted and included the Freebase type system in our datasets, which is not available in FB15k-237. Our analysis and observations about Freebase type system, which is discussed in section 3, are mostly identified by the authors of this paper and have not been discussed in any other paper.

Furthermore, section 1 now summarizes the significance of the datasets, including how nontrivial their creation is, as follows: "The dataset creation was nontrivial and time-consuming. It required extensive inspection and complex processing of the massive Freebase data dump, for which documents are scarce. None of the idiosyncrasies, as articulated in Sections 3 and 4, was defined or detailed in the data dump itself. Figuring out the details in these sections required iterative trial-and-error in examining the data. In fact, we are unaware of more detailed description of these idiosyncrasies anywhere else. If researchers must learn to examine Freebase and prepare datasets from scratch, we expect many of them to have steep learning curve and the process can easily require months or even years. Our datasets can thus speed up many researchers' work."

2. "W1: … The proposed URI "simplification" seems inconvenient as it introduces a disconnect between the source dataset and the presented dataset; these simplifications could better be added as metadata."

Thank you for the suggestion regarding “URI simplification”. We now have included a mapping file from the original URIs to the simplified versions of object labels, as part of the metadata in the dataset. The inclusion of the mapping file is discussed in section 6 of the submission and is also mentioned in the GitHub repository available at https://github.com/idirlab/freebases. Furthermore, we would like to mention that many of the hosts in the original URIs (e.g., “http://rdf.freebase.com/”) are unavailable now, and thus it helps to have a dataset without these URI components. Finally, we also would like to point out that URI simplification helps improve usability for applications and tasks beyond link prediction. For instance, when we visualize a knowledge graph, it is possible only an entity’s name matters to the end users. Inclusion of the full URI will render the user interface overly cluttered.

3. “W2: … Motivation: it is not clear why Freebase was chosen over e.g. DBpedia/Wikidata which are still heavily maintained and updated, and share some of the important properties as discussed in this paper. In the end, the key differentiating idiosyncrasies (reverse triples and mediator nodes) of Freebase are suggested to be removed from the dataset as they impose challenges for e.g. link prediction, which weakens the motivation for choosing FB.”

Please refer to the several new paragraphs at the end of Section 1 which summarize our contributions. Particularly, the following point is the most relevant to this comment: “The datasets and results are highly relevant to the research community, as Freebase remains the single most commonly used dataset for link prediction, by far. We examined all full-length publications that 1) appeared in 12 top conferences during their latest years and 2) used datasets commonly used for link prediction. This amounts to 63 publications. Among them, 57 publications used datasets produced from Freebase, while 9 used datasets from Wikidata. Only 2 publications used a Freebase dataset at its full scale, specifically Freebase86m. This suggests that researchers may not be able to carry out large-scale study due to the lack of proper datasets.” (The list of the 63 publications and more details can be found in Section 1 of the paper.)

The initial version of the paper might leave the wrong impression that these idiosyncrasies are undesirable always removed. We thank you for the comment which helped us improve the paper’s clarity. The paper is now revised to show the benefits of including mediator nodes and reverse triples in data.

For example, in Section 4 we showed how datasets with mediator nodes will make it possible to evaluate conventional embedding models on multiary relationships. Conventional models cannot be applied on hyper-relational datasets, e.g., JF17K [45], because representation based on key-value pairs is alien to such models. Our datasets capture multiary relationships in Freebase through triples containing mediator nodes. Moreover, in Section 7, we showed the existence of mediator nodes will make link prediction more challenging. In short, mediator nodes are not “bad”. If a model is not designed to tackle mediator nodes, a dataset with mediator nodes could be more useful in order to focus on developing the model’s designed capacity. However, mediator nodes are necessary for modeling multiary relations, an ideal model will need to aim at tackling them too.

Regarding reverse relations, their existence leads to overestimation of model performance, and they do not provide additional information. Hence, they do not help create a more accurate model, and they do not help evaluate models more accurately either. However, including reverse relations in a dataset can allow one to compare models regarding which model better handles data leakage and thus it can help develop more robust models.

4. “W2 ….Analysis: Some statements and conclusions seem to rely on speculations instead of analyses or facts. For example, ln. 316-318 gives an example of similar relations "place_of_birth" and "place_of_death" that could have been removed in the related FB15k-237 but it is unclear if this observed or implied by the procedural description of the FB15k-237 dataset as far as I can tell.”

Thank you for the very helpful comment, which triggered us to improve the paper’s clarify a lot. This original example ("place_of_birth" and "place_of_death") is hypothetical---i.e., it is a possible scenario implied by the procedural description of FB15K-237. We have now added a new example which is observed in FB15k-237. Furthermore, we improved the writing to make it clear that there can be two types of mistakes in FB15K-237---false positives and false negatives. The original hypothetical example belongs to false positives, and the newly

added observed example belongs to false negatives. The changes are in Section 5 of the paper, and we also copy it below.

“FB15k-237 is created from FB15k by removing redundant relations. Given two relations, they calculate how much their subject-object pairs overlap and, if the overlap is greater than a threshold, they consider them as redundant. This step could incorrectly remove useful information, due to two types of mistakes. A) 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. B) False negatives, since FB15k-237 did not resort to the accurate reverse relation information encoded by reverse_property in Freebase. For example, we observed that FB15k-237 includes both /education/educational_institution_campus/educational_institution and /education/educational_institution/campuses but they are reverse relations according to reverse_property.”

5. “W2 … Furthermore, based on a performance comparison between FB1/2 and FB15k-237 the observation of a lower accuracy on the latter is attributed to the size of the datasets but this seems ungrounded since FB15k-237 has additional differences, like the removal of trivial linkable relations, that may explain a lower performance as well. A more careful analysis is needed.”

We would like to first apologize for a typo in one of our tables (Table 3 in original submission; Table 2 in the revision) in which FB1 and FB2 labels were swapped. It is corrected in the current Table 2.

Furthermore, it seems our writing led to confusion. We should compare models’ performance on FB1 (instead of FB2) with that on FB15k-237, since reverse relations are removed in both datasets. Our point is that likely the sheer size different contributed to the performance gap, which is common in machine learning in general too.

If by “trivial linkable relations” the reviewer meant the reverse relations, then “...this seems ungrounded since FB15k-237 has additional differences...” is a confusion to the authors. Because reverse relations are removed from both FB1 and FB15K-237. If “trivial linkable relations” means something else, please let us know and we will be glad to address it. Also, as mentioned in the previous comment’s response, FB15k-237 does not remove reverse relations using the truly correct method, and it could have both false positives and false negatives. However, such mistakes would be on the relatively minor side. In summary, our goal was to show that removing reverse triples from large-scale data has the same effect that we could observe on the small-scale data.

We shall mention that we have now removed the original Table 4, which shows link prediction models’ performance on small-scale FB15k/FB15k-237 datasets. But the paper’s focus is on full-scale Freebase datasets. We believe it brings little value to keep the table while it may generate confusion.

6. “W3: The most novel enhancement of the dataset seems to be the enhancement of "required types" of objects/subjects based on some probability, but it is unclear how these probabilities are obtained, what threshold is used (and why), and an analysis of the correctness of the resulting types is missing. The added value of these types is also not supported by experiments or analysis.”

Thank you for the questions! We have revised the paper accordingly to make it more clear. Specifically:

The details of how these probabilities are calculated can now be found in section 3, paragraph 6, which is highlighted in blue.

- The threshold value used is 0.95, which was mentioned in paragraph 5 of section 3. The threshold value is determined empirically, we have now provided details about how it was determined in paragraph 6.

- The correctness of the threshold is also analyzed at the end of paragraph 6. As of now, this was only done on a few cases, since it requires manual verification. We plan to carry out a larger scale verification. We are confident that we will have the results way before the final version of the paper is due for publishing, and we welcome a shepherding processing from the conference to ensure such results are included in the paper.

7. “Correctness: … the technical details relating to the type enhancement are missing and the decision to transform the URIs is disputable.”

Regarding “type enhancement”, please see above for our response #6. The paper is expanded to include this discussion in Section 3. We are grateful to your comment which led to this improvement.

Regarding “URI”, please refer to our response #2 above. Again, thank you for the suggestion which we have followed to enhance the datasets.

8. “Clarity: … Section 3 and 4 present explanations and challenges of the idiosyncrasies, but these sections partly overlap and may be merged. The experiment section introduces the task of link prediction but does not formulate the problem at hand.”

Thank you for your comments. We have gone through the paper and polished it throughout. Hopefully sections 3 and 4 now have less redundancy. e decided to still keep them separate, in order to make separate discussions of the idiosyncrasies and their associated challenges and thus to avoid confusion. If there are further comments and suggestions on places that need improvement, we will be very glad to address them.

The link prediction task is formulated in the 2nd sentence of Section 4.2: “Link prediction is the task of predicting the missing s in triple (?, p, o) or missing o in (s, p, ?).”

9. “Documentation: Experiments are implemented using the DGL-KE framework and therefore rely on its maintenance, but this seems OK. The authors may consider getting local copy of it. The scripts for creating the datasets are present and the data is currently hosted on Dropbox (it is claimed to be transferred to Zenodo later).”

Thank you for the suggestion. Following your suggestion, we made a local copy of DGLKE on our GitHub.

We will make sure to archive the dataset in Zenodo when the paper is accepted and finalized. All revisions in Zenodo are visible to the public, as different versions. Since the paper is still being reviewed and revised, we have not archived the dataset there yet in order to avoid users’ confusion due to multiple versions.

10. “Minor comment: Table 2 shows statistics of the introduced FB intersections but while the #entities, #relations and #triples are reported for the related datasets, this table shows #entities, #properties and #triples. Is the #properties a mistake? “

Properties and relations are used interchangeably in this paper, as shown in Section 2. To avoid confusion, we have replaced #properties by #relations in Table 2 (which is now Table 1) based on your suggestion.

---

5. *Reviewer Er64*

Review by reviewerEr64

Summary And Contributions:

The paper discusses several data modeling idiosyncrasies that Freebase has but are rarely found in other comparable datasets like Wikidata. The authors further measure their impacts on real-world tasks. The unique properties like mediator nodes, reverse types, type information for each node and metadata poses new benefits and challenges to the development of KGE methods. The authors further proposes four variants of Freebase dataset by the inclusion/exclusion of reverse triples and CVT nodes and evaluated existing KGE's performance on FB1 and FB2 dataset.

Rating: 3: Clear rejection

Strengths:

1. The problem, which reclaims the importance of freebase dataset is interesting and worth thinking of. The unique properties that Freebase has, like mediator nodes and no KGE methods can handle is an interesting perspective to investigate further.
2. The authors emphasize the importance of having a large-scale KG dataset in the paper, which is something worth noticing.
3. The authors perform experimental analysis across 5 KGE embedding, of different categories, including both translational and bilinear models.

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

Weaknesses:

Weakness:

Overall, this is an encouraging direction to go, but I think more work needs to be done to justify the value of this work. The weakness of this paper is summarized as follows.

Weakness 1: Inconsistency of reported performance on existing datasets with respect to the existing literature.

Weakness 2: FB1 seems to too easy and does not have much room for improvement.

Weakness 3: Some of the authors' claims seemed to be contradict with the experiment results.

Weakness 4: Some of the authors' claims are not supported by experiments.

Weakness 5: The quality of FB3 and FB4 are unsupported.

And I will explain each claim in detail here.

Weakness 1: Inconsistency of reported performance on existing datasets with respect to the existing literature. In Table4, the authors reported the performance of existing KGE methods on FB15k and FB15k-237. The results in this table is not consistent with respect to what is reported in the literature in three main aspects and I wonder if there is a specific reason for the inconsistency here.

(1) For TransR on FB15k-237, the authors reported an MRR of 0.576, which is almost twice as the number reported in [1]. In [1], they report 0.314 for TransR on FB15k-237.

(2) On FB15k-237, for methods like TransE, DistMult, Complex, RotatE, the authors reported a number that is much lower than [2].

(3) On FB15k, DGL-KE, the library they use reported an MRR of 0.726 for RotatE. However here, the authors reported an MRR of 0.685.

Weakness 2: FB1 seems to too easy and does not have much room for improvement. In Table3, the performance for the new dataset FB1 on Transe is already 0.958, and there isn't much room for improvement. So what's the meaning of proposing FB1 and using it instead of the existing KG dataset?

Weakness 3: Some of the authors' claims seemed to be contradict with the experiment results. In Table3, FB1's result is much higher than FB2 across all metrics, this should indicate that FB1 is an easier dataset than FB2. Table 2 shows that the main difference wrt. FB1 and FB2 is that FB1 removes all of the reverse pairs while FB2 retains them. According to Section 4.2, the existence of reverse triples should make the task easier because the KGE models only needs to learn the simple reverse rules. So I would expect to see FB1 has a lower performance than FB2. But it does not seem to be the case.

Weakness 4: Some of the authors' claims are not supported by experiments.In section 3 and section4, the authors claim the importance and challenges of CVT nodes and states that simply converting it to binary relationships between pairs of entities would loose information. But the authors does not really design any experiments justify their claim.

Weakness 5: The quality of FB3 and FB4 are unsupported. The authors mainly design four variants of Freebase as the new datasets. But no experiment is done for FB3 and FB4 and I cannot infer the quality of FB3 and FB4 from Table 2.

Reference

[1] Akrami et al. Realistic Re-evaluation of Knowledge Graph Completion Methods: An Experimental Study. SIGMOD 2020.

[2] Ruffinelli et al. You can teach an old dog new tricks! On training knowledge graph embeddings. ICLR 2020.

[3] Da et al. DGL-KE, https://dglke.dgl.ai/doc/benchmarks.html#fb15k, SIGIR 2020.

Correctness:

The dataset seems to be constructed in a sound way. But they don't really design extensive experiments to show the quality of the dataset. Some variants of the dataset, FB3 and FB4 is not evaluated by experiments. Also, for FB1 and FB2, there seems to some contradictions between the authors' claims and the experiment results, see weakness 2 for details about this.

Clarity:

Overall, the paper is pretty clear and well-written.

Relation To Prior Work:

The authors discussed the differences from the existing freebase variants in section 5. But the authors do not really compare their dataset with any other large-scale KG datasets like wikiKG90M. Also, previous works suggest that Freebase variants is problematic because of skewed relations and the existence of fixed-set relations [4], which makes Freebase too easy to learn. The authors do not mention anything like that in the paper.

Reference [4] Safavi et al. CODEX: A Comprehensive Knowledge Graph Completion Benchmark. EMNLP 2020.

Documentation:

The authors provide detailed procedures for data cleaning step in section 6. The authors also provide a URL to download the datasets. But there is one limitation from my point of view. Freebase is a no longer maintained database after google shut it down in 2015. If others want some additional information of Freebase, other than the ones already provided in the dataset, it would be difficult to get. And thus limits the usefulness of Freebase-based dataset.

Ethics:

No, to the best of my knowledge.

Additional Feedback:

Additional comment: If across dataset comparison is needed, AMRR is a better metric than MRR, check here https://arxiv.org/pdf/2203.07544.pdf.

FEEDBACK

We would like to express our sincere gratitude to the reviewer for their thoughtful comments which have helped us improve the paper significantly. The revised paper is uploaded and the more substantial changes are highlighted in blue color in the paper. The major changes in the paper include:

- New paragraphs in Section 1 to summarize significant contributions made by the datasets and results.

- Section 4.1 now discusses how the datasets and experimentation design can enable comparison with models built for hyper-relational facts and comparison on Wikidata.

- Section 7 now includes experiment results on FB3 and FB4 and other tasks (triple classification).

Please see below for detailed response.

1. "Weakness 1: (1) For TransR on FB15k-237, the authors reported an MRR of 0.576, which is almost twice as the number reported in [1]. In [1], they report 0.314 for TransR on FB15k-237."

[1], also our publication, used LibKGE to perform the experiments. Since this submission focuses on experiments using full-scale Freebase, we chose one of the libraries DGL-KE that is capable of it, and we used DGL-KE on both the full-scale Freebase datasets and FB15k-237, to be consistent. The performance difference mentioned in the comment could be due to different implementations and hyperparameter settings. Nevertheless, reviewer HrUH had doubt on the TransR implementation in DGL-KE. To avoid confusion, we now have taken out TransR from all results. We plan to investigate the correctness of the TransR implementation in DGL-KE, which is beyond scope here.

We shall mention that we have now removed the original Table 4, which was on small-scale FB15k/FB15k-237 datasets but the paper's focus is on full-scale Freebase datasets. We believe it brings little value to keep the table while it may generate confusion.

2. "Weakness 1: (2) On FB15k-237, for methods like TransE, DistMult, Complex, RotatE, the authors reported a number that is much lower than [2]."

In [2], the authors mentioned that the choice of training strategy and hyperparameters are very influential on model performance, often more so than the model class itself. They provide a framework called LibKGE. LibKGE aims to provide clean implementations of training, hyperparameter optimization, and evaluation strategies that can be used with any model. As they mentioned, with LibKGE's architecture and hyperparameter optimization, state-of-the-art results can be achieved.

On the other hand, large-scale frameworks typically have a narrower scope, e.g., they often do not feature hyperparameter optimization. They are usually more focused on parallelizing training models. As there are a few large-scale frameworks, we did not have many choices to conduct the experiments. The results we obtained might not be the best results possible. We should note that the scope of our work is not to compare models' results and find the best performing models. Please also see our response to the next comment regarding this.

Finally, as explained in our response to the comment above, the original table with results on FB15k/FB15k-237 is now taken out.

3. “Weakness 1: (3) On FB15k, DGL-KE, the library they use reported an MRR of 0.726 for RotatE. However here, the authors reported an MRR of 0.685.”

As demonstrated in the DGL-KE paper, several factors such as number of CPU cores or number of GPUs and other hyperparameters can affect the performance of the models. As a result, we cannot necessarily achieve the numbers they did. In our experiments we used two GPUs only. The details of the experiments’ setup can be found in section 7 of our submission and our GitHub repository at https://github.com/idirlab/freebases/tree/main/ExperimentsScripts.

As a result, these differences do not conclude that any of these results from our experiments or the ones reported by the authors of dgl-ke framework are wrong. But it shows the difference in experimental setups.

We would like to further point out that the objective of the experiments in the paper is not to compare different embedding models or optimize any particular model. Rather, it is to show the impacts of the three idiosyncrasies (mediator/CVT nodes, reverse triples, type system) on such models, as well as to show how they perform differently on existing small vs. new large-scale Freebase datasets. The drastic performance differences of the same model, trained using the same library DGL-KE, across these different settings (e.g., large-scale vs. small-scale; with CVT vs. without CVT) are the results that we focus on reporting. Compared with such drastic differences, the performance difference of a particular model in our experiment vs. an optimal result is much smaller in most cases.

4. “ Weakness 2: FB1 seems to too easy and does not have much room for improvement. In Table3, the performance for the new dataset FB1 on Transe is already 0.958, and there isn't much room for improvement. So what's the meaning of proposing FB1 and using it instead of the existing KG dataset?”

First, we’d like to point out Table index and dataset label changes, to help ease the discussion:

- Table 3 in the original submission is now expanded to Table 2 in the revision.

- FB1 in the original submission is now FB2 in the revision; and thus the original FB2 is now FB1. This correction was due to a typo in the original submission (our apologies!). See feedback to the next comment about this.

It is correct that FB1 (again, FB2 now) contains many reverse relations and thus leads to overestimation of the models. That’s what the experiment would like to verify on the full-scale Freebase datasets. Hence, for the purpose of creating a more accurate model or evaluating models more accurately, FB1 (FB2 now) is not as useful as FB2 (FB1 now). However, this dataset with reverse triples can allow one to compare models regarding which model better handles such data leakage and thus it can help develop more robust models.

Note that we provided four variants of Freebase, FB1, FB2, FB3, FB4. The differences in these four variants is the inclusion and exclusion of mediator nodes and reverse triples, as shown in Table 1 of the revised paper. They are useful for understanding the impact of mediator nodes and reverse triples, as can be seen from the results in Section 7.

Regarding “existing KG dataset”, we’d like to emphasize that existing KG datasets are either too small (e.g., FB15k-237) and/or improperly prepared for tasks such as link prediction (e.g., Freebase86m). Section 1 provides a summary of the work’s significant contributions. There is one bullet point there that is particularly pertinent to this matter, which we copy below.

“The paper fills an important gap in dataset availability. To the best of our knowledge, ours is the first-ever publicly available full-scale Freebase dataset that has gone through proper preparation. Specifically, our Freebase variants were prepared in recognition of data modeling idiosyncrasies such as mediator objects, reverse triples, and type system, as well as via thorough data cleaning. On the contrary, the Freebase data dump has all types of triples tangled together, including even data about the operation of Freebase itself; Freebase86m [52], one of the few available full-scale Freebase datasets, also mixes together everything—

operational data that are not common knowledge facts, reverse triples, and mediator objects. (Details in Section 5.)"

5. "Weakness 3: Some of the authors' claims seemed to be contradict with the experiment results. In Table3, FB1's result is much higher than FB2 across all metrics, this should indicate that FB1 is an easier dataset than FB2. Table 2 shows that the main difference wrt. FB1 and FB2 is that FB1 removes all of the reverse pairs while FB2 retains them. According to Section 4.2, the existence of reverse triples should make the task easier because the KGE models only needs to learn the simple reverse rules. So I would expect to see FB1 has a lower performance than FB2. But it does not seem to be the case."

Thank you for pointing out this. We apologize for the typo! In Table 3 of the original submission, which is now Table 2 in the revision, FB1 and FB2 labels were swapped. It is corrected in the current Table 2. Indeed FB1 has a lower performance than FB2, as you correctly pointed out, and is also now correctly shown on Table 2.

6. "Weakness 4: Some of the authors' claims are not supported by experiments. In section 3 and section4, the authors claim the importance and challenges of CVT nodes and states that simply converting it to binary relationships between pairs of entities would loose information. But the authors does not really design any experiments justify their claim."

"Star-to-Clique" (S2C) conversion is known to be irreversible and causes a loss of information. Our paper provided a reference on this (reference [45] in our revised submission).

Section 3 also has an example on this. We'd like to use that example here and further clarify. "For instance, after such a transformation for Figure 1, the new triples cannot exactly pinpoint to the work that leads to James Ivory's nomination for the BAFTA award". Imagine James Ivory was also nominated for some other award based on another work of his. Suppose the graph is transformed into binary relationships, following the procedure mentioned in that paragraph. After the transformation, it becomes impossible to figure out which work is for which nomination, thus the loss of information.

7. "Weakness 5: The quality of FB3 and FB4 are unsupported. The authors mainly design four variants of Freebase as the new datasets. But no experiment is done for FB3 and FB4 and I cannot infer the quality of FB3 and FB4 from Table 2."

Thank you for this important comment, which inspired us to conduct more experiments. Particularly, we have conducted more experiments on FB3 and FB4 as well and revised Section 7 of the paper to include the results from these new experiments.

With regard to "quality of FB3 and FB4", the four variants FB1, FB2, FB3, FB4 were created following the same procedure based on the same data preparation. Hence, they shall have equal "quality" if by that you meant the correctness of the datasets. The differences in these four variants is the inclusion and exclusion of mediator nodes and reverse triples, as shown in Table 1 of the revised paper (i.e., Table 2 in the original submission.)

8. "Correctness: The dataset seems to be constructed in a sound way. But they don't really design extensive experiments to show the quality of the dataset. Some variants of the dataset, FB3 and FB4 is not evaluated by experiments. Also, for FB1 and FB2, there seems to some contradictions between the authors' claims and the experiment results, see weakness 2 for details about this."

Thank you again for this important comment! These concerns are addressed in the feedback provided above, specifically #5 and #7.

9. "Relation To Prior Work: The authors discussed the differences from the existing freebase variants in section 5. But the authors do not really compare their dataset with any other large-scale KG datasets like wikiKG90M. Also, previous works suggest that Freebase variants is problematic because of skewed relations and the

existence of fixed-set relations [4], which makes Freebase too easy to learn. The authors do not mention anything like that in the paper."

Section 4.1 now includes a detailed discussion of hyper-relational fact, which is used for modeling multiary relationships in both Wikidata and Freebase. It further went on to explain that "given the datasets and experiment results made available in this paper, it becomes possible to compare the performance of both hyper-relational fact models and conventional models on a full-scale Freebase dataset that includes multiary relationships (specifically, FB3 in Table 1 of Section 6). Such a comparison will be the first not only for Freebase but also for any large-scale knowledge graph. For instance, to the best of our knowledge, there does not exist a full-scale Wikidata dataset with multiary relationships represented as conventional triples instead of hyper-relational facts. Therefore, conventional models have only been applied on Wikidata without multiary relationships [42], e.g., OGBL-WikiKG2 [24]. There exists no comparison of the two categories of models on Wikidata with multiary relationships." Section 1 summarizes these discussions and further explains that "The experiment design could be similarly extended to study the impact of multiary relationships in Wikidata on embedding models and compare such models with models built for hyper-relational facts." We are planning to conduct similar preprocessing and experiments on Wikidata, based on our analysis mentioned in these places.

In [4], they discuss problems of FB15k-237 but the mentioned problems do not make Freebase easy to learn as FB15k-237 is a very small subset of Freebase. Fixed set relations are discussed in one of our previous studies and we refer to them as Cartesian product relations in our publication [a]. They are mainly due to concatenation of relations. We actually did some experiments and compared the results of the models on datasets with and without Cartesian product relations. However, as the number of such relations was low, they did not significantly affect the results. Due to lack of space, we did not include such results in the paper.

[a] Realistic re-evaluation of knowledge graph completion methods: An experimental study. In SIGMOD 2020.

10. "Documentation: The authors provide detailed procedures for data cleaning step in section 6. The authors also provide a URL to download the datasets. But there is one limitation from my point of view. Freebase is a no longer maintained database after google shut it down in 2015. If others want some additional information of Freebase, other than the ones already provided in the dataset, it would be difficult to get. And thus limits the usefulness of Freebase-based dataset."

Section 1 of the paper now has a new paragraph that summarizes the significance of the work, including why Freebase. The datasets and results are highly relevant to the research community, as Freebase remains the single most commonly used dataset for link prediction, by far. We examined all full-length publications that 1) appeared in 12 top conferences during their latest years and 2) used datasets commonly used for link prediction. This amounts to 63 publications. The conferences, the papers, and the datasets used in the papers are listed in file "papers.xlsx" which can be accessed at the top directory of our GitHub repository https://github.com/idirlab/freebases. Among them, 57 publications used datasets produced from Freebase, while 9 used datasets from Wikidata. Only 2 publications used a Freebase dataset at its full scale, specifically Freebase86m. This suggests that researchers may not be able to carry out large-scale study due to the lack of proper datasets.

11. "Additional Feedback: Additional comment: If across dataset comparison is needed, AMRR is a better metric than MRR, check here https://arxiv.org/pdf/2203.07544.pdf."

Thank you for your suggestion. Currently, comparing performance across different datasets is not in the scope of the paper. But we will use AMRR when we start doing that, as you suggested.

---

6. *Reviewer HrUH*

Review

Summary And Contributions:

The paper proposes four graph datasets based on Freebase claiming to capture design principles (idiosyncrasies) better than existing datasets. The datasets have 40-60M nodes and 100-250M triples, respectively, varying the availability of Compound Value Type (CVT) nodes and inverse triples. Additionally, for each dataset, the authors provide metadata files: mappings of various ID types to labels and the type system, namely, entity types of subjects and objects for each predicate.

Rating: 3: Clear rejection

Strengths:

- The construction methodology is clear and might be interesting when executed on more up-to-date and active graphs like Wikidata or YAGO;
- The type system might be useful for some prediction tasks.

Confidence: 5: The reviewer is absolutely certain that the evaluation is correct and very familiar with the relevant literature

Weaknesses:

W1. Perhaps the first major criticism point is the source of the datasets and its relevance to the NeurIPS audience, i.e., Freebase. Freebase has been deprecated since 2015, its content is outdated, its deisgned principles have been largely improved in subsequent open knowledge graphs. The authors acknowledge this fact in the Appendix but the arguments on why the community needs a suite of datasets stuck in 2015 are not convincing. The proposed methodology might well be applied to modern, community-updated graphs like Wikidata or YAGO - and its possible benefits would have been much bigger.

W2. The defined idiosyncrasies of Freebase are mostly artifacts of obsolete modeling practices and are no longer relevant. For example, Compound Value Type nodes (CVT) were introduced to model complex relationships among several nodes, but today we have RDF*, a W3C recommendation [0] extending the RDF model to solve exactly this problem. RDF* is adopted by YAGO, and Wikidata has an analogous Wikidata Statement Model where every triple might be wrapped into a statement with additional qualifiers (entity-relation pairs). This is a much more clear and powerful model of representing complex relationships than CVT. Existing Wikidata-derived datasets like WikiPeople [1] or WD50K [2] already incorporate such complex statements with qualifiers into prediction tasks. Therefore, the claim in line 218 is incorrect:

“Current models were all built on datasets generated from DBpedia or Wikidata … none of which contains mediator nodes”

Simply because Wikidata has a more powerful Statement Model instead of CVT mediator nodes. Pertaining to Freebase, there exist JF17K [3] and m-FB15k [4] datasets with CVT nodes represented as n-ary relationships, so it’s not new from this perspective either, and it could have been reflected in the related work section.

Similarly, the efforts to build a type system stem from the lacking information in Freebase. However, modern KGs already have it by default, e.g., all properties in Wikidata have range and domain restrictions that specify subject and object types, and are also used in consistency check pipelines (for example, ShEx shapes in Wikidata).

W3. The authors proposed 4 datasets, but it seems that only one (FB1) has a practical relevance. That is, the FB2 dataset with inverse relations is already saturated and does not contain any benchmarking value since all baseline models yield 95%+ MRR and 95%-ish Hits@1 (it seems that Table 3 has a typo with swapped dataset names, the dataset with inverse edges should normally yield much higher numbers). What is the challenge of running experiments on the already saturated dataset? Then, having created FB3 and FB4, the authors do not run any experiments on them with a vague reason that “blindly applying link prediction models when CVT nodes are present may lead to foreseeable problems” (line 379). The graph with CVT nodes is essentially a hypergraph (or hyper-relational graph depending on the format), so the authors could have taken existing

hypergraph- (or hyper-relational) link prediction baselines mentioned in Section 4.1. Otherwise, it seems like the authors propose a few datasets but do not propose any meaningful tasks on them.

W4. Discussing possible tasks and applications, the authors repeatedly (in Sections 3, 4.1, 4.2) mention interactive graph query systems and data-to-text models. Visual query builders seem to be out of scope for NeurIPS while data-to-text models are mostly built on top of Wikidata-derived graphs (for obvious reasons - the task itself emerged with neural language models long after Freebase was shut down). That is, I can hardly see practical usages of data-to-text models trained on Freebase when nobody is using Freebase in real-world applications. Instead, in 2022, there are many new challenging graph reasoning tasks that might be very interesting when executed at scale, namely, inductive link prediction [5] and complex logical query answering [6] - for both of them, baseline datasets are sampled from small FB15k-237. Generally, in the presence of more challenging tasks, transductive-only link prediction on triple-only graphs (as envisioned for proposed FB1 and FB2) is not a compelling graph benchmark.

W5. TransR results of DGL-KE (and at least one other model) are incorrect due to the serious implementation error and it is known since at least September 2021 [7]. Given that the NeurIPS'22 Datasets and Benchmarking track deadline was in June 2022 (8 months as the bug is known) I wonder why the authors decided to include a clearly not-working approach into their experimental program. In fact, it might raise a question of the overall correctness of reported numbers - it would be more convincing to replicate the numbers with other frameworks like libkge [8], graphvite [9], or pykeen [10]. As of July 2022, the confirmed state of the art on FB15k-237 is Neural Bellman-Ford Networks [11] with MRR of about 0.42, and I was very surprised that a 5-year old embedding approach suddenly yielded the result suspiciously larger (0.57 MRR) than NBFNet.

References:
[0] https://w3c.github.io/rdf-star/cg-spec/editors_draft.html
[1] Guan et al. Link prediction on n-ary relational data. WWW 2019
[2] Galkin et al. Message Passing for Hyper-Relational Knowledge Graphs. EMNLP 2020
[3] Wen et al. On the Representation and Embedding of Knowledge Bases Beyong Binary Relations. IJCAI 2016.
[4] Fatemi et al. Knowledge Hypergraphs: Prediction Beyond Binary Relations. arxiv:1906.00137
[5] Teru et al. Inductive Relation Prediction by Subgraph Reasoning. ICML 2020
[6] Ren and Leskovec. Beta Embeddings for Multi-Hop Logical Reasoning in Knowledge Graphs. NeurIPS 2020
[7] https://github.com/awslabs/dgl-ke/issues/225
[8] https://github.com/uma-pi1/kge
[9] https://github.com/DeepGraphLearning/graphvite
[10] https://github.com/pykeen/pykeen
[11] Zhu et al. Neural Bellman-Ford Networks: A General Graph Neural Network Framework for Link Prediction. NeurIPS 2021

Correctness:

- Some claims are not backed up by related work (see W2 in the review)
- Reported numbers are incorrect due to a well-known implementation bug (see W5)

Clarity:

The paper is clearly written.

Relation To Prior Work:

The authors tackle two graph modalities in the proposed datasets - triple-only in FB1/FB2 and hypergraph (hyper-relational) graph in FB3/FB4. However, the related work section covers only the triple-only part. There exist popular Freebase-derived benchmarks for hypergraphs like JF17K and m-FB15K, the could have been addressed in the related work as well.

Documentation:

OK

Ethics:

No ethical concerns. The authors acknowledged in the Appendix that certain domains of intereset might be underrepresented.

Additional Feedback:

I think the paper will be much stronger if it would focus on contemporary KGs like Wikidata and propose challenging large-scale tasks over different graph modalities in addition to plain transductive link prediction on triples.

FEEDBACK

We would like to express our sincere gratitude to the reviewer for their thoughtful comments which have helped us improve the paper significantly. The revised paper is uploaded and the more substantial changes are highlighted in blue color in the paper. The major changes in the paper include:

- New paragraphs in Section 1 to summarize significant contributions made by the datasets and results.

- Section 4.1 now discusses how the datasets and experimentation design can enable comparison with models built for hyper-relational facts and comparison on Wikidata.

- Section 7 now includes experiment results on FB3 and FB4 and other tasks (triple classification).

Please see below for detailed response.

1. “W1. … the arguments on why the community needs a suite of datasets stuck in 2015 are not convincing. The proposed methodology might well be applied to modern, community-updated graphs like Wikidata or YAGO - and its possible benefits would have been much bigger.”

Please refer to the several new paragraphs at the end of Section 1 which summarize our contributions. Particularly, the following point is the most relevant to this comment: “The datasets and results are highly relevant to the research community, as Freebase remains the single most commonly used dataset for link prediction, by far. We examined all full-length publications that 1) appeared in 12 top conferences during their latest years and 2) used datasets commonly used for link prediction. This amounts to 63 publications. Among them, 57 publications used datasets produced from Freebase, while 9 used datasets from Wikidata. Only 2 publications used a Freebase dataset at its full scale, specifically Freebase86m. This suggests that researchers may not be able to carry out large-scale study due to the lack of proper datasets.” (The list of the 63 publications and more details can be found in Section 1 of the paper.)

Furthermore, Section 4.1 now includes a detailed discussion of hyper-relational fact, which is used for modeling multiary relationships in both Wikidata and Freebase. It further went on to explain that “given the datasets and experiment results made available in this paper, it becomes possible to compare the performance of both hyper-relational fact models and conventional models on a full-scale Freebase dataset that includes multiary relationships (specifically, FB3 in Table 1 of Section 6). Such a comparison will be the first not only for Freebase but also for any large-scale knowledge graph. For instance, to the best of our knowledge, there does not exist a full-scale Wikidata dataset with multiary relationships represented as

conventional triples instead of hyper-relational facts. Therefore, conventional models have only been applied on Wikidata without multiary relationships [42], e.g., OGBL-WikiKG2 [24]. There exists no comparison of the two categories of models on Wikidata with multiary relationships." Section 1 summarizes these discussions and further explains that "The experiment design could be similarly extended to study the impact of multiary relationships in Wikidata on embedding models and compare such models with models built for hyper-relational facts."

2. Regarding W2 in the review:

First, we really appreciate these comments as they motivated us to think deeper and write more clearly. Many of the changes in the paper revision were indeed triggered by comments like this. Particularly inspired by the comment, section 4.1 now discusses how the datasets and experimentation design can enable comparison with models built for hyper-relational facts and comparison on Wikidata. We are planning to conduct similar preprocessing and experiments on Wikidata, based on our analysis in Section 4.1.

We should note that, to present Wikidata facts as triples we need reification which means to add auxiliary nodes similar to CVT nodes. Wikidata RDF dumps in NTriples format presents data exactly in this way, by including such auxiliary nodes which they called statement nodes. The difference is that there is an extra direct edge between the primary subject and object. As a result, the design regarding n-ary relationships is not very different in these two datasets.

Regardless the comment about "Current models were all built on datasets generated from DBpedia or Wikidata … none of which contains mediator nodes": In that sentence we were discussing data-to-text models instead of link prediction models. To the best of our knowledge, none of the data-to-text models uses Wikipeople or WD50k. These datasets are used for hyper-relational link prediction task as discussed in Section 4.1. (We note that this sentence is now removed from the paper, since we decided to focus on embedding models and link prediction, and we removed discussions of data-to-text models)

3. "W3. The authors proposed 4 datasets, but it seems that only one (FB1) has a practical relevance. That is, the FB2 dataset with inverse relations is already saturated and does not contain any benchmarking value since all baseline models yield 95%+ MRR and 95%-ish Hits@1 (it seems that Table 3 has a typo with swapped dataset names, the dataset with inverse edges should normally yield much higher numbers). What is the challenge of running experiments on the already saturated dataset?"

Indeed, here was a typo. We thank you for the comment and we apologize for the mistake. FB1 in the original submission should be FB2, FB2 should have been FB1. This is now corrected. Also, we note that Table 3 in the original submission is now expanded to Table 2 in the revision.

It is correct that FB2 contains many reverse relations and thus leads to overestimation of the models. That's what the experiment would like to verify on the full-scale Freebase datasets. Hence, for the purpose of creating a more accurate model or evaluating models more accurately, FB2 is not as useful as FB1. However, this dataset with reverse triples can allow one to compare models regarding which model better handles such data leakage and thus it can help develop more robust models.

4. "W3.... Then, having created FB3 and FB4, the authors do not run any experiments on them with a vague reason that "blindly applying link prediction models when CVT nodes are present may lead to foreseeable problems" (line 379). The graph with CVT nodes is essentially a hypergraph (or hyper-relational graph depending on the format), so the authors could have taken existing hypergraph- (or hyper-relational) link

prediction baselines mentioned in Section 4.1. Otherwise, it seems like the authors propose a few datasets but do not propose any meaningful tasks on them."

Thank you for this important comment, which inspired us to conduct more experiments. Particularly, we have conducted more experiments on FB3 and FB4 as well and revised Section 7 of the paper to include the results from these new experiments.

Regarding using hyper-relational link prediction as baselines, we now included some discussions in Section 4.1, as summarized below. Long story short, such a comparison doesn't exist in prior work, due to lack of proper datasets from both Freebase and Wikidata. Our datasets will fill this gap.

"There is a divide between the models built for hyper-relational facts and the more conventional link prediction models such as TransE and ComplEx, in terms of both the applicable datasets and the methodologies. Conventional models cannot be applied on hyper-relational datasets, e.g., JF17K [45], because representation based on key-value pairs is alien to such models. Our work focuses on these conventional models. The datasets we create and use capture multiary relationships in Freebase through triples containing CVT nodes. In principle, the models built for hyper-relational facts could be applied on conventional datasets as well, although we are unaware of any such empirical study, not to mention such studies on datasets containing CVT nodes. In fact, as discussed in Section 1, there does not exist a full-scale Freebase dataset that is properly prepared for tasks such as link prediction. In this regard, given the datasets and experiment results made available in this paper, it becomes possible to compare the performance of both hyper-relational fact models and conventional models on a full-scale Freebase dataset that includes multiary relationships (specifically, FB3 in Table 1 of Section 6). Such a comparison will be the first not only for Freebase but also for any large-scale knowledge graph. For instance, to the best of our knowledge, there does not exist a full-scale Wikidata dataset with multiary relationships represented as conventional triples instead of hyper-relational facts. Therefore, conventional models have only been applied on Wikidata without multiary relationships [42], e.g., OGBL-WikiKG2 [24]. There exists no comparison of the two categories of models on Wikidata with multiary relationships."

5. "W4. Discussing possible tasks and applications, the authors repeatedly (in Sections 3, 4.1, 4.2) mention interactive graph query systems and data-to-text models. Visual query builders seem to be out of scope for NeurIPS while data-to-text models are mostly built on top of Wikidata-derived graphs (for obvious reasons - the task itself emerged with neural language models long after Freebase was shut down). That is, I can hardly see practical usages of data-to-text models trained on Freebase when nobody is using Freebase in real world applications. Instead, in 2022, there are many new challenging graph reasoning tasks that might be very interesting when executed at scale, namely, inductive link prediction [5] and complex logical query answering [6] - for both of them, baseline datasets are sampled from small FB15k-237. Generally, in the presence of more challenging tasks, transductive-only link prediction on triple-only graphs (as envisioned for proposed FB1 and FB2) is not a compelling graph benchmark."

Based on the comments, we removed from the paper discussions related to interactive graph query systems and data-to-text models.

Indeed, as mentioned in the comment, there are many interesting graph reasoning tasks that are usually conducted on small-scale datasets. The datasets created in our work could be used for such tasks, and we are planning for that in our feature work. Nevertheless, though, the current results are still highly relevant to the research community, as Freebase remains the single most commonly used dataset for link prediction, by far, as explained in response to W1.

6. "W5. TransR results of DGL-KE (and at least one other model) are incorrect due to the serious implementation error and it is known since at least September 2021 [7]. Given that the NeurIPS'22 Datasets and Benchmarking track deadline was in June 2022 (8 months as the bug is known) I wonder why the authors decided to include a clearly not-working approach into their experimental program. In fact, it might raise a question of the overall correctness of reported numbers - it would be more convincing to replicate the numbers with other frameworks like libkge [8], graphvite [9], or pykeen [10]. As of July 2022, the confirmed state of the art on FB15k-237 is Neural Bellman-Ford Networks [11] with MRR of about 0.42, and I was very

surprised that a 5-year old embedding approach suddenly yielded the result suspiciously larger (0.57 MRR) than NBFNet."

The TransR implementation bug is reported by a user on GitHub and has not been approved by any of the dgl-ke framework authors or any published work. Nevertheless, we now have taken out TransR from all results, and we appreciate being informed about this. We plan to investigate the correctness of the TransR implementation in DGL-KE, which is beyond scope here. Since the experiments are all on large-scale data, due to the time limitation, we have not repeated the experiments on another framework. We would like to further point out that the objective of the experiments in the paper is not to compare different embedding models or optimize any particular model. Rather, it is to show the impacts of the three idiosyncrasies (mediator/CVT nodes, reverse triples, type system) on such models, as well as to show how they perform differently on existing small vs. new large-scale Freebase datasets. The drastic performance differences of the same model, trained using the same library DGL-KE, across these different settings (e.g., large-scale vs. small-scale; with CVT vs. without CVT) are the results that we focus on reporting. Compared with such drastic differences, the performance difference of a particular model in our experiment vs. an optimal result is much smaller in most cases.

We shall mention that we have now removed the original Table 4, which was on small-scale FB15k/FB15k-237 datasets. The paper's focus is on full-scale Freebase datasets. We believe it brings little value to keep the table while it may generate confusion.

7. "Correctness: Some claims are not backed up by related work (see W2 in the review) Reported numbers are incorrect due to a well-known implementation bug (see W5)"

Thank you again for these comments. We provided the responses to W2 and W5 above.

8. "Relation to Prior Work: The authors tackle two graph modalities in the proposed datasets - triple-only in FB1/FB2 and hypergraph (hyper-relational) graph in FB3/FB4. However, the related work section covers only the triple-only part. There exist popular Freebase-derived benchmarks for hypergraphs like JF17K and m-FB15K, the could have been addressed in the related work as well."

Thank you very much for your comment. We added the discussion of hyper-relational graphs and models to Section 4.1.

9. "Additional Feedback: I think the paper will be much stronger if it would focus on contemporary KGs like Wikidata and propose challenging large-scale tasks over different graph modalities in addition to plain transductive link prediction on triples."

Thank you again for this comment. This and other similar comments have motivated us to expand Section 4.1 to discuss Wikidata and hyper-relational fact based models, we well as how our datasets and experimentation design could enable investigations and comparison of different models on Wikidata.

Nevertheless, as mentioned in response to W1, Freebase is still the single most commonly used dataset for link prediction, by far. Hence, our datasets and results are still highly relevant to the research community.

Paper Decision
Decision: Reject
Meta Review of Paper381 by Program Chairs
NeurIPS 2022 Track Datasets and Benchmarks Program Chairs

Metareview:

This submission had many issues. The authors did put a lot of effort into improving the paper since the original submission. Still, there is uncertainty about whether the experiments are sufficiently convincing, and whether this submission presents a massive improvement over previous datasets. In both cases, I'm not yet fully convinced. I'm disregarding the formatting issues raised by reviewer 27CX. Overall it still seems that this work is a bit too preliminary to warrant publication at NeurIPS.

Confidence: 3: The area chair is somewhat confident
Recommendation: Reject