Reviewer #1

Questions

- 1. Overall Rating
  - Weak Reject
- 2. Relevant for PVLDB
  - Yes
- 3. Are there specific revisions that could raise your overall rating?
  - Yes
- 5. Paper Summary. In one solid paragraph, describe what is being proposed and in what context, and briefly justify your overall recommendation.
  - In this study, the authors conduct an analysis of the challenges associated with Freebase data modeling idiosyncrasies, including CVT
    nodes, reverse properties, and type system, and provide four variants of the Freebase dataset by inclusion and exclusion
    of these idiosyncrasies. They also conducted experiments to evaluate different link prediction models on these datasets.
- 6. Three (or more) strong points about the paper. Please be precise and explicit; clearly explain the value and nature of the contribution.
  - S1. An analysis is conducted to reveal the challenges associated with Freebase data modeling idiosyncrasies, including CVT
    nodes, reverse properties, and type system.
    S2. Four variants of the Freebase dataset are provied by inclusion and exclusion of these idiosyncrasies.

    S3. Experiments are conducted to evaluate different link prediction models on these datasets.

- 7. Three (or more) weak points about the paper. Please clearly indicate whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution; write it so that the authors can understand what is seen as negative.
  - W1. From the experimental study, the comprehensive analysis is only targeted for the link prediction task. This is a kind of limitation. How about other task?

    W2. Section 4 on the challenges posed by the data modeling idiosyncrasies of Freebase. This part should clearly state the challenges.

    W3. How the four variants of the the Freebase dataset are created to address the challenges? This should be the most important part, which not much is dicussed.
- 8. Novelty. Please give a high novelty rating to papers on new topics, opening new fields, or proposing truly new ideas; give medium ratings to "delta" papers and those on well-known topics but still with some valuable contribution. (Note: For SDS and EA&B

papers, novelty does not need to be in the form of new algorithms or models. Instead, novelty for SDS can be new understanding of issues related to data science technologies in the real world. Novelty for EA&B can be new insights into the strengths and weaknesses of existing methods or new ways to evaluate existing methods.)
  - With some new ideas
- 9. Significance
  - Improvement over existing work
- 10. Technical Depth and Quality of Content
  - Syntactically complete but with limited contribution
- 11. Experiments. (Reminder: EA&B papers should have a higher bar for experiments.)
  - OK, but certain claims are not covered by the experiments
- 12. Presentation
  - Sub-standard: would require heavy rewrite
- 13. Detailed Evaluation (Contribution, Pros/Cons, Errors). Please number each point and provide as constructive feedback as possible.
  - This study is interesting. The focus of the analysis of Freebase and the creation of benchmark datasets. However, I believe a further investigation and polish of the presentation are very necessary.

    D0. Please refer to W1~W3.

    D1. If the analysis is only for the link prediction task, the title and the scope of this paper should clearly reflect `link prediction'. Otherwise, the exprimental study should contain other tasks.

    D2. How the variants of the the Freebase dataset are created is too simple - this part need to reflect the part to address the challenges.

    D3. The presentation needs a significant improvement - your contributions are not very clear from Sections 4, 5 & 6.

    D4. Experimental study: any impact on the efficiency?

Reviewer #2

Questions

- 1. Overall Rating
  - Weak Accept
- 2. Relevant for PVLDB
  - Yes

- 3. Are there specific revisions that could raise your overall rating?
    - No
- 5. Paper Summary. In one solid paragraph, describe what is being proposed and in what context, and briefly justify your overall recommendation.
    - This paper presents an analysis of the challenges associated with the idiosyncrasies of Freebase and measures their impact on knowledge graph link prediction. it provides four variants of the Freebase dataset by inclusion/exclusion of mediator objects and reverse triples. A Freebase type system is also extracted to supplement the variants. The datasets were used for experimentation using multiple link prediction algorithms and dataset variations offering useful insights on the dataset and the algorithms.

      Certainly a paper to be ready if someone would like to use Freebase for experimentation, whereas the provided datasets would promote research in the area.
- 6. Three (or more) strong points about the paper. Please be precise and explicit; clearly explain the value and nature of the contribution.
    - S1. The paper promotes understanding of embedding models for knowledge graph link prediction.

      S2. The curated dataset provided has the potential to accelerate the work of many researchers and practitioners.

      S3. The paper promotes understanding of embedding models for knowledge graph link prediction.

      S4. The curated dataset provided has the potential to accelerate the work of many researchers and practitioners.
- 7. Three (or more) weak points about the paper. Please clearly indicate whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution; write it so that the authors can understand what is seen as negative.
    - O1. Minor improvements in presentation that can be easily corrected
- 8. Novelty. Please give a high novelty rating to papers on new topics, opening new fields, or proposing truly new ideas; give medium ratings to "delta" papers and those on well-known topics but still with some valuable contribution. (Note: For SDS and EA&B papers, novelty does not need to be in the form of new algorithms or models. Instead, novelty for SDS can be new understanding of issues related to data science technologies in the real world. Novelty for EA&B can be new insights into the strengths and weaknesses of existing methods or new ways to evaluate existing methods.)
    - Novel
- 9. Significance
    - Improvement over existing work
- 10. Technical Depth and Quality of Content

  - Solid work
- 11. Experiments. (Reminder: EA&B papers should have a higher bar for experiments.)
  - Very nicely support the claims made in the paper
- 12. Presentation
  - Excellent: careful, logical, elegant, easy to understand
- 13. Detailed Evaluation (Contribution, Pros/Cons, Errors). Please number each point and provide as constructive feedback as possible.
  - Well-structured, easy to read paper.

    It promotes understanding on the Freebase intricacies, revealing also some weakness of link prediction algorithms.

    Further it offers a whole new, useful curated set of datasets to be used from now on link prediction experiments.

    It is not clear how accuracy was evaluated for extracting the Freebase Type System. How large was “the empirical evidence”?

    Section 4.1 which are the “various models” shown in Figure 3, why are those selected? This should be better explained.

Reviewer #3

Questions

- 1. Overall Rating
  - Weak Reject
- 2. Relevant for PVLDB
  - Yes
- 3. Are there specific revisions that could raise your overall rating?
  - Yes
- 5. Paper Summary. In one solid paragraph, describe what is being proposed and in what context, and briefly justify your overall recommendation.
  - The paper elaborates on the special features of Freebase, namely the CVT nodes, the reverse links and the type system. Based on the inclusion and the exclusion of these features, the authors produce four versions of the entire Freebase KG, which is publicly available. Through a series of experiments with the state-of-the-art algortihms in the field, the authors examine how each feature affects their performance. The paper is quite interesting and offers a comprehensive overview of the Freebase completion task. However, the discussion of the experimental results is rather brief, it ignores the relative performance of the considered algorithms as well as their run-time, while the paper abounds in redundant content (figures).

- 6. Three (or more) strong points about the paper. Please be precise and explicit; clearly explain the value and nature of the contribution.
  - S1. The paper provides a very nice overview of Freebase, its special characteristics and examines the way these affect the performance of KG completion methods.

    S2. The paper is well-written, well-structured and easy-to-follow.

    S1. The authors have publicly released their datasets and code.
- 7. Three (or more) weak points about the paper. Please clearly indicate whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution; write it so that the authors can understand what is seen as negative.
  - W1. I would expect a thorough discussion about the relative performance of the algorithms evaluated over the existing and the new datasts. Yet, the authors state in page 10 (end of left column) that this lies out of the scope of their work. However, one of the goals of E&A papers is to provide novel insights in to the relative performance of existing techniques so as to verify or dispute experimental analyses in the literature. Given that the authors have already performed so many experiments, it would be easy to extend their discussion into this direction.

    W2. The experimntal results ignore the time efficiency of the examined algorithms over the existing and the new datasets. How do the special characteristics of Freebase affect the run-time of the considered algorithms? In line with W1, it would be nice to rse, it would be nice to discuss the relative run-time of these algorithms.

    W3. The discussion of the performance of the same methods over the different datasets is rather superficial. See D1 for more details.

    W4. A large part of the paper is covered by redundant content, as explained in D2 in more detail.

    W5. I am a bit sceptical about the generality of the conclusions drawn from this analysis. See D3 for more details.
- 8. Novelty. Please give a high novelty rating to papers on new topics, opening new fields, or proposing truly new ideas; give medium ratings to "delta" papers and those on well-known topics but still with some valuable contribution. (Note: For SDS and EA&B papers, novelty does not need to be in the form of new algorithms or models. Instead, novelty for SDS can be new understanding of issues related to data science technologies in the real world. Novelty for EA&B can be new insights into the strengths and weaknesses of existing methods or new ways to evaluate existing methods.)
  - With some new ideas

- 9. Significance
    - Improvement over existing work
- 10. Technical Depth and Quality of Content
    - Syntactically complete but with limited contribution
- 11. Experiments. (Reminder: EA&B papers should have a higher bar for experiments.)
    - OK, but certain claims are not covered by the experiments
- 12. Presentation
    - Excellent: careful, logical, elegant, easy to understand
- 13. Detailed Evaluation (Contribution, Pros/Cons, Errors). Please number each point and provide as constructive feedback as possible.
    - The authors could improve their work in the following ways:

      D1) In Figures 3-6 and in the corresponding tables, the authors generally observe an increase or decrease in the performance of the examined algorithms, depending on the feature (e.g., CVT nodes or reverse links). This pattern is usually common across all algorithms. For example, all algorithms have higher MRR over FB+CVT+REV than over FB+CVT-REV, as shown in Figure 3. However, the degree to which this pattern applies differs from algorithm to algorithm. Yet, there is no discussion about the extent to which each feature affects each algorithm. The authors should examine in more detail how much is the MRR of each algorithm is reduced or increased in the absence or presence of CVT nodes and/or reverse links.

      D2) Figures 3, 4, 5 and 6 are rather appealing and intuitive, yet they merely visualize part of the experiments in Tables 1, 3, 5 and 8, respectively. They could be omitted or moved to an online extended version so as to save space for more interesting content, such as the relative performance of the considered algorithms, as discussed in W1 and W2 above.

      D3) The authors motivate their work by stating that Freebase is the most common dataset in the literature of KG completion. This is true to some extent, as there also other popular datasets like Wordnet (see "A comprehensive overview of knowledge graph completion. Knowl. Based Syst. 255: 109597 (2022)" for more details). The reason is every KG has some special features, like those discussed in Section 3 for Freebase, which significantly affect the performance of the proposed algorithms. Thus, the generality of the KG completion algorithms would be limited if every algorithm was examined on a single benchmark KG, especially Freebase, which was shut down in 2015. Yet, the authors do not discuss other benchmark datasets neither examine whether their methodology applies to other popular KG completion benchmarks.

      D4) In Section 3, page 4, the authors mention some experiments about the threshold \alpha, which is thus empirically set to 0.95. Please report these experiments, at least in an online extended version.

D5) Many tables, like 1, 3, 5 and 8 seem to fit in a single column with some minor modifications. This would save a lot of space for more interesting content.

D6) Please add references in the experimental setup (page 10) for the selected split ratio (90/5/5).

Minor comment:
* Please add a dot at the end of each paragraph title (e.g,. in Sections 3, 5, 6 and 7).

D4)

- 14. Revision. If revision is required, list specific required revisions you seek from the authors. Please number each point.
  - Please address W1-W4.

Reviewer #4

Questions

- 1. Overall Rating
  - Weak Reject
- 2. Relevant for PVLDB
  - Yes
- 3. Are there specific revisions that could raise your overall rating?
  - Yes
- 5. Paper Summary. In one solid paragraph, describe what is being proposed and in what context, and briefly justify your overall recommendation.
  - The paper presents an extraction and cleaning activity on freebase with the goal of creating a high quality large scale dataset for link prediction.
    Focus is put on the peculiar modelling choices and how data is represented in freebase, in particular the mixture of metadata and data, the present of inverse relationships, the the use of reification (here called mediator nodes).
    The paper present also some shallow findings when re-evaluating methods on this new set of datasets.
- 6. Three (or more) strong points about the paper. Please be precise and explicit; clearly explain the value and nature of the contribution.
  - S1 - The paper introduces a clean dataset where some important extraction tasks have been performed. The introduction of the type system is also an interesting added value

    S2 - Experiments are conducted to estimate the actual effect of various characteristics in the data that did not receive focus in previous work

S3 - The experiments identify also high variation of performances about different subsets of the graph pointing towards the need of more detailed results analysis in past and future works

- 7. Three (or more) weak points about the paper. Please clearly indicate whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution; write it so that the authors can understand what is seen as negative.
  - W1 - Freebase has not been updated anymore, this work fails to put this resource in perspective w.r.t. dataset that could be used instead, e.g., CoDex.

    W2 - The work does not provide an in-depth numerical analysis of the contents of the dataset proposed

    W3 - The experimental evaluation is overall shallow, does not critically evaluate the methods and does not report on wether the performances obtained on this dataset match existing experimental comparisons on completely different datasets.
- 8. Novelty. Please give a high novelty rating to papers on new topics, opening new fields, or proposing truly new ideas; give medium ratings to "delta" papers and those on well-known topics but still with some valuable contribution. (Note: For SDS and EA&B papers, novelty does not need to be in the form of new algorithms or models. Instead, novelty for SDS can be new understanding of issues related to data science technologies in the real world. Novelty for EA&B can be new insights into the strengths and weaknesses of existing methods or new ways to evaluate existing methods.)
  - Novelty unclear
- 9. Significance
  - Improvement over existing work
- 10. Technical Depth and Quality of Content
  - Syntactically complete but with limited contribution
- 11. Experiments. (Reminder: EA&B papers should have a higher bar for experiments.)
  - Obscure, not really sure what is going on and what the experiments show
- 12. Presentation
  - Sub-standard: would require heavy rewrite
- 13. Detailed Evaluation (Contribution, Pros/Cons, Errors). Please number each point and provide as constructive feedback as possible.
  - D1- While I think that this dataset and its availability is definitely interesting and valuable, the statement << researchers may not be able to carry out large-scale study due to lack of proper datasets.>> presented as the reason why most works focused on FB15k is far from true. Training KG embedding is extremely expensive, so the lack of scalable methods is the main reason. I suggest that statement to be removed since it is unsubstantiated.

D2- While it is true that in some works the idiosyncrasies of freebase have not been tackled explicitly in many link prediction works, leading to poor experimental evaluation, they are basically common knowledge in most data management works which (very succinctly) describe how the treated them: see [A][B] for example. They have further been explained in more detail in [C]

D3 - this work proposed to adopt the resolution of mediator nodes by concatenating edges. As the work itself explains this is a lossy transformation. I am not sure this work should add more support to this problematic practice.

D4 - I am not sure how interesting is at all the question about <<how the mixture of knowledge facts, metadata and administrative
data impact model performance.>> Compared to other more pressing questions, for instance treatment of transitive closures, effect of types and node-type prediction tasks, or performances across composite relations. Thus it seems odd that this is given precedence over studying the <<different natures of binary and multiary relations in the datasets>> and the effect on model performance

D5 - in Freebase many edges are those declaring the type of an enitity. An experiment on type prediction can be an important addition. It is instead unclear why the experiment of table 8 is not conducted on the full dataset.

D6 - TransE outperforming other methods is presented as a case of Simpson’s paradox. This actually raises the question on wether the given datasets proposed are actually worth investigation given that they can lead to misleading results. Overall, it is unclear whether the relative performance of methods here is consistent with the literature. This analysis is the core of the E&A paper and is instead quite overlooked here.

D7 - it is unclear whether the split is also shared. It is important that researchers will all reference to the same train/test/validation split

D8 - the paper needs heavy restructuring. Sections keep pointing towards sect. 6 for important details and partial comments of experimental results are mixed across all the paper which is very confusing. The paper should follow a more logical structure: present the gap in the literature, explain the methodlogy of how the datasets have been extracted, analyze their contents, and then go through a detailed experimental evaluation with clear research questions

D9 - there many sources of bias when evaluating link prediction methods, this work seems to uncritically adopt problematic metrics of quality instead see [D].

Results are even presented without confidence interval.

D10 - The 90/5/5/ split seems quite unusual with a large skew on training data. It is also unclear whether this split is stratified per relationship type

D11 - Some works explicitly test rule-based system as baselines, when it seems that many methods are often learning simple rules. With so much data available, this work could provide better inside adding such experiment to the mix.

Minor comments:

Figures like fit3 are very hard to read, plus they represent means, so something like a boxplot would be the appropriate

Sec. 41 "link predication" -> link prediction

[A] Jayaram, Nandish, et al. "Querying knowledge graphs by example entity tuples." IEEE Transactions on Knowledge and Data Engineering 27.10 (2015): 2797-2811.

[B] Mottin, D., Lissandrini, M., Velegrakis, Y., & Palpanas, T. (2016). Exemplar queries: a new way of searching. The VLDB Journal, 25, 741-765.

[C] Bast, Hannah, et al. "Easy access to the freebase dataset." Proceedings of the 23rd International Conference on World Wide Web. 2014.

[D] Mohamed, Aisha, et al. "Popularity agnostic evaluation of knowledge graph embeddings." Conference on Uncertainty in Artificial Intelligence. PMLR, 2020.

Meta Review

1. Overall Recommendation
    - Reject
2. Summary Comments
    - The paper presents an extraction and cleaning activity on freebase with the goal of creating a high quality large scale dataset for link prediction. The paper introduces a clean dataset where some important extraction tasks have been performed. The introduction of the type system is also an interesting added value. The paper is well-written, well-structured and easy-to-follow, with available code. However, there are quite a number of issues that need to be taken care before the paper is ready for publication. The most important are:

- a thorough discussion about the relative performance of the algorithms evaluated over the existing and the new datasets.
- The experimental results ignore the time efficiency of the examined algorithms over the existing and the new datasets.
- The work does not provide an in-depth numerical analysis of the contents of the dataset proposed
- The experimental evaluation does not critically evaluate the methods and does not report on whether the performances obtained on this dataset match existing experimental comparisons on completely different datasets.

As such, it is suggested that the authors revisit the paper and resubmit at a later time.