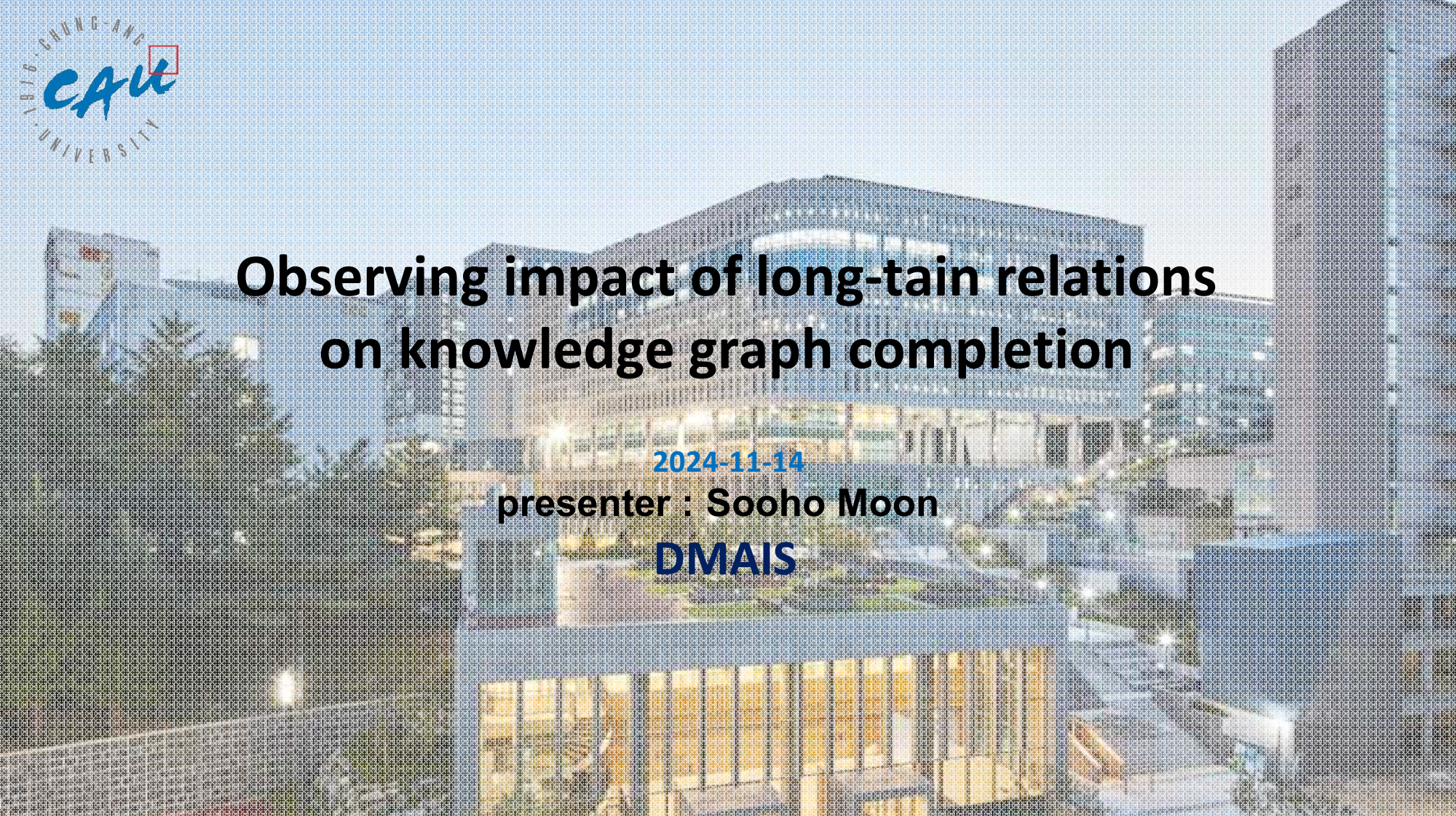


Observing impact of long-tain relations on knowledge graph completion

2024-11-14

presenter : Sooho Moon

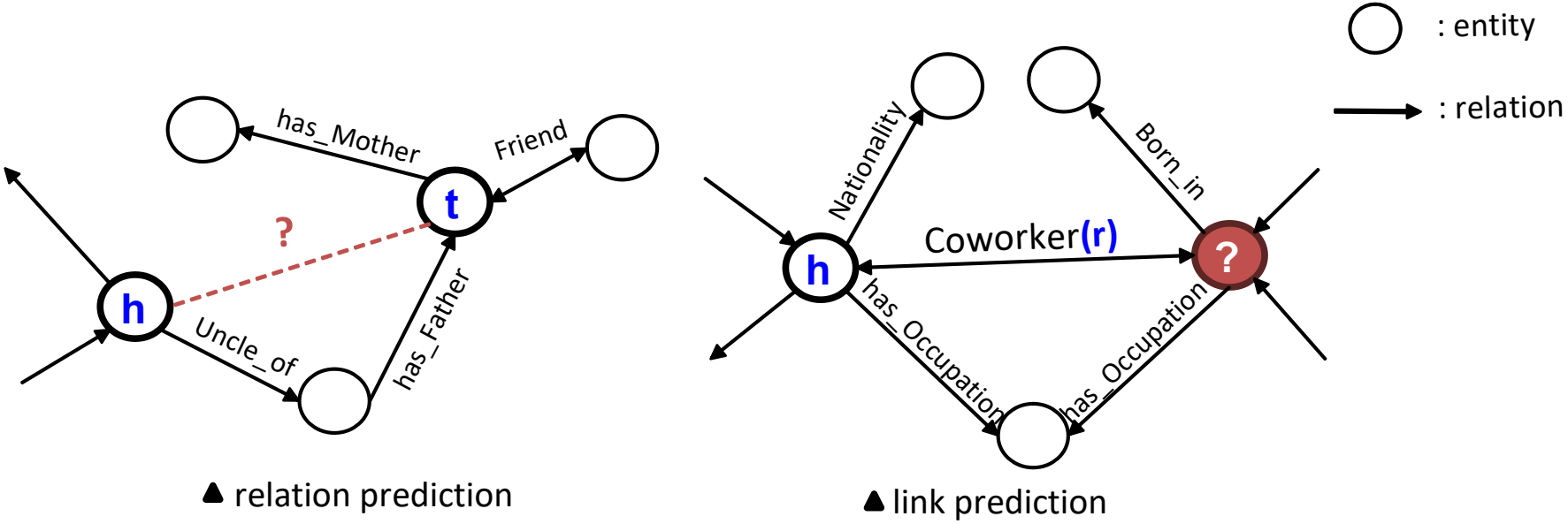
DMAIS

INDEX

- Knowledge graph completion(KGC)
- Long-tail relation symptoms of KGs
- Experiments on RotatE/Pathcon
- Result analysis
- Additional observation
- Conclusion

Knowledge graph completion(KGC)

- **Find missing relationships between entities**
  - □ Training an AI model to find **"?"** given (**?**, r, t), (h, **?**, t), (h, r, **?**)
  - □ In short, **building a model capable of reasoning over KGs**

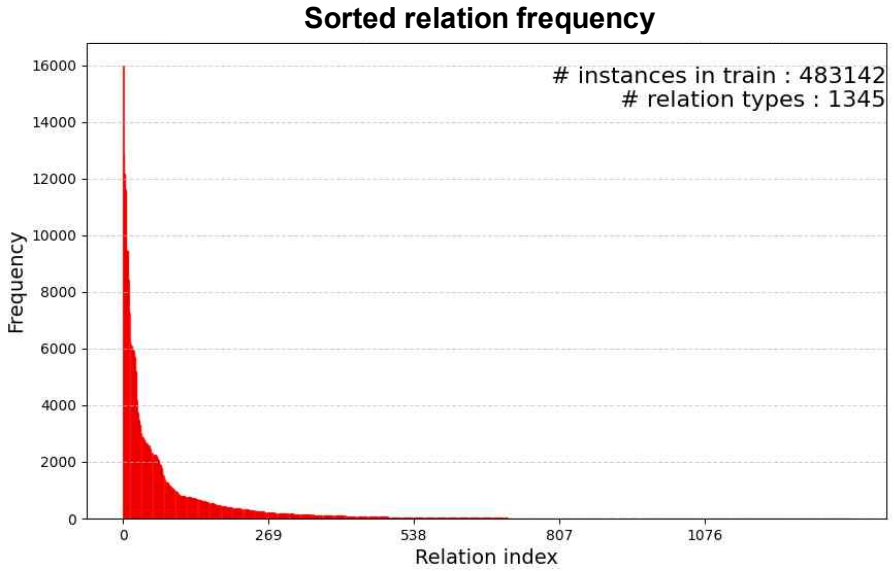


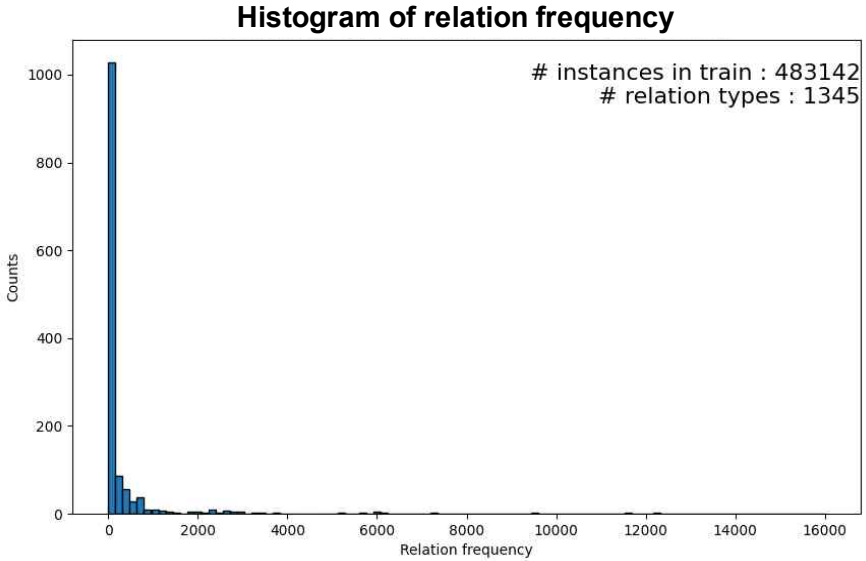
▲ relation prediction

▲ link prediction

Long-tail relation symptoms of KGs

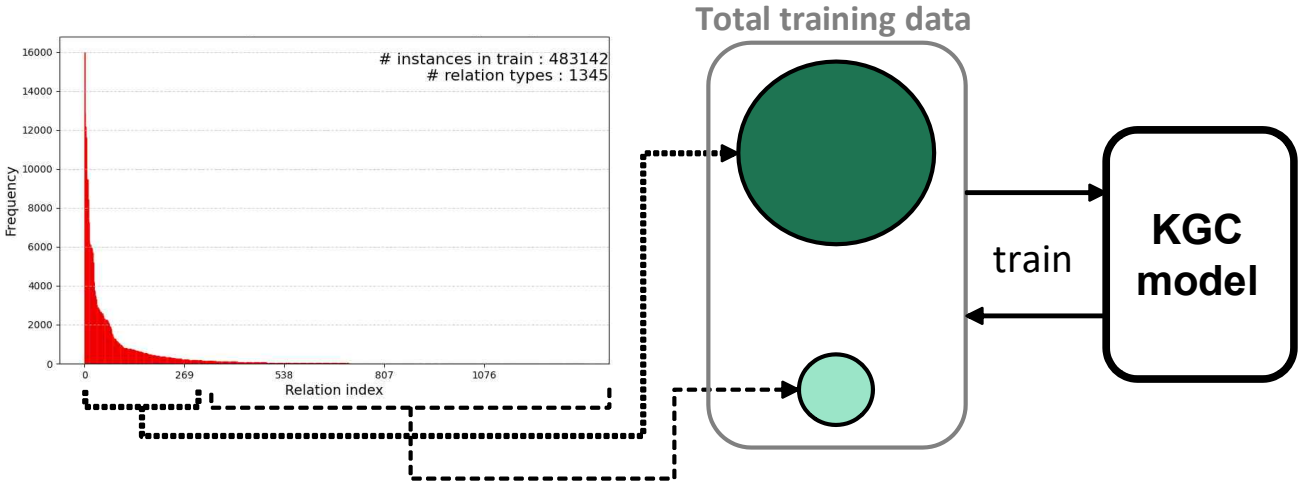
- **What is long-tail relation?**
  - □ Small portion of relations dominate the total population of the KG
  - □ Below are statistics of the FB15k dataset, a widely known and popular benchmark dataset





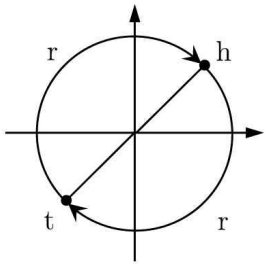
Long-tail relation symptoms of KGs

- **Why should we care?**
  - □ **Evaluation perspective :**
    High-frequency relations dominate the metrics, causing uneven evaluations
  - □ **Practical perspective :**
    Low-frequency(long-tail) relations tend to underfit, thus poor generalization in downstream tasks

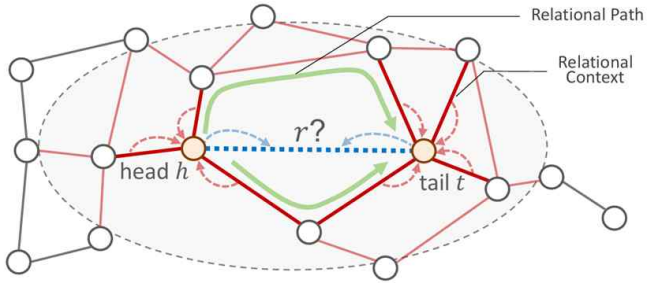


Experiments on RotatE/PathCon

- **Translation based RotatE(ICLR '19)**
  - □ **First model that achieved SOTA performance on all the benchmarks**[1]
  - □ Models individual entity, relation embeddings on a complex space
- **Path based PathCon(KDD '21)**
  - □ Combined relational context and relational paths for relation prediction



(c) RotatE: an example of modeling symmetric relations $\mathbf{r}$ with $r_i = -1$



[1] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: knowledge graph embedding by relational rotation in complex space. In ICLR.

Experiments on RotatE/PathCon

- **Can both models learn long-tail relations properly?**

  □ Are they competent enough to reason on long tail relations?

- **I want to find answers to the following questions**

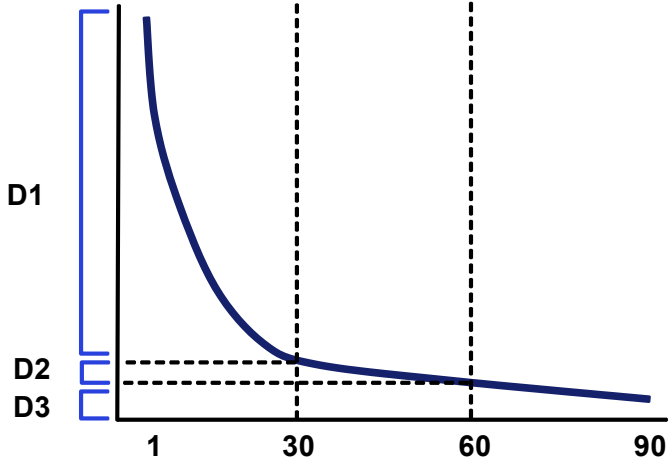
  1. Does long-tail relations have impact on the total metric?

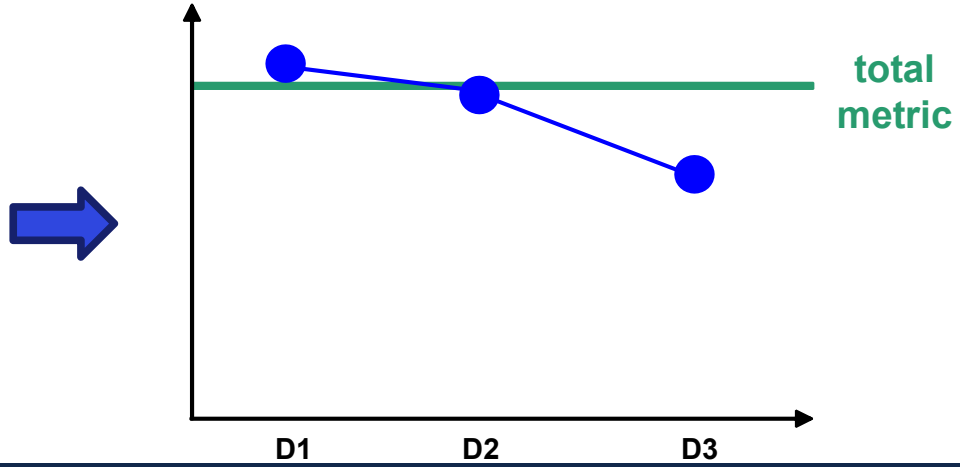
  2. If so, is the hindering factor big?

  3. Are there other things that we can think about?

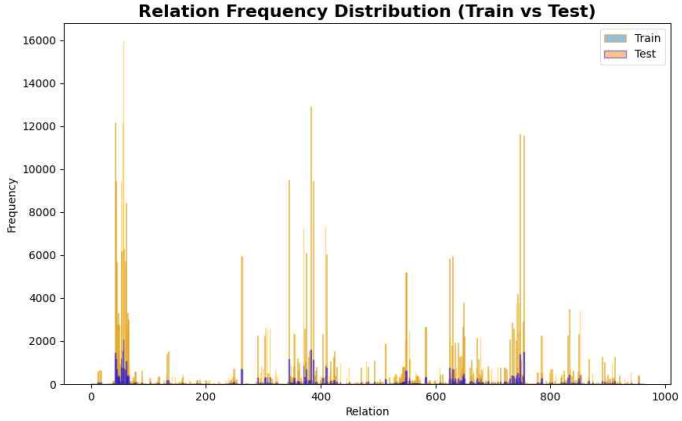
Experiments on RotatE/PathCon

- **Experiment formulation**
  - □ Count relation frequencies in the training dataset
  - □ With that, sort the relations according to frequencies in a descending order
  - □ Divide the sorted relations to k folds
  - □ Calculate the frequency of the relation which is located in the separating line, and create districts
  - □ Calculate metrics on each district





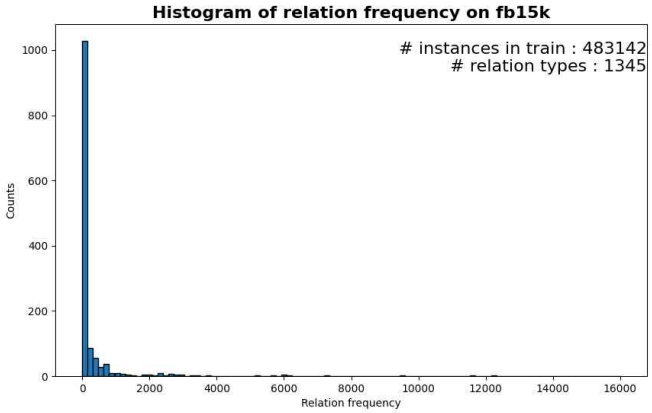
Experiments on RotatE/PathCon

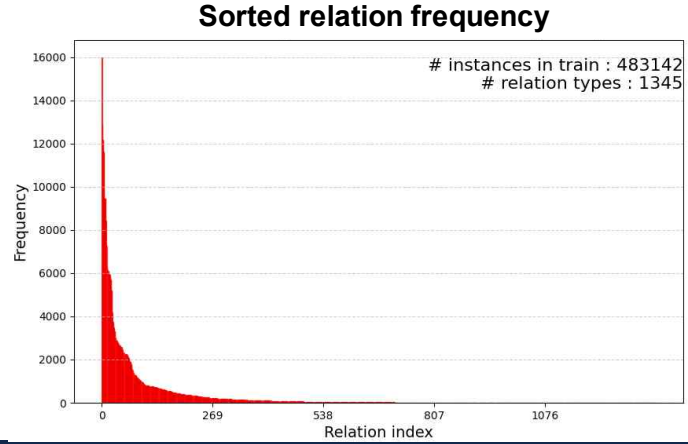


FB15k

 # train instance = 483,142
 # relation types = 1345

Top 5 least number of instance : 1, 1, 1, 1, 1



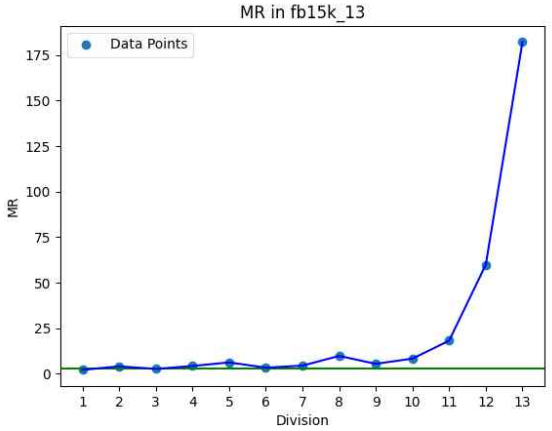


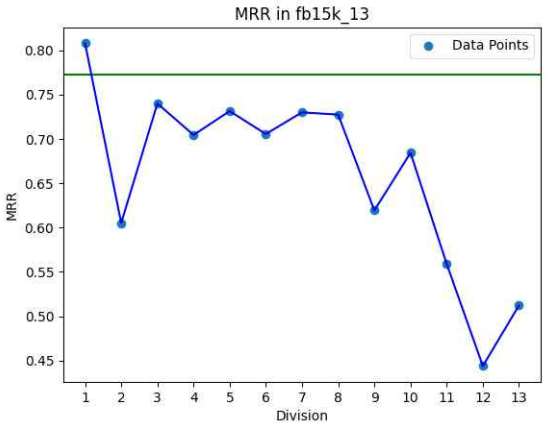
*** side note**

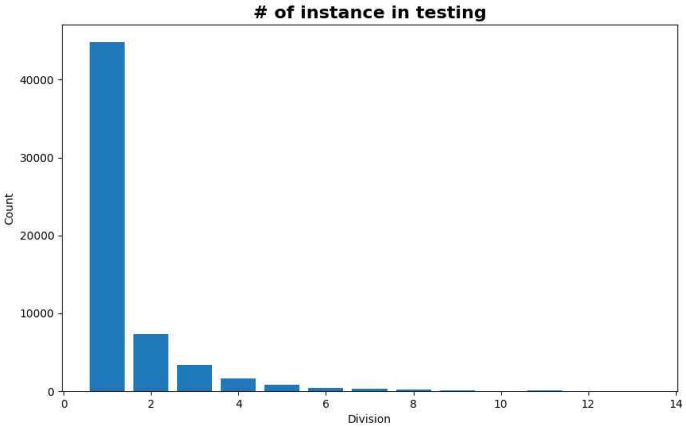
I conducted relation prediction in order to aline the test setting to our interest

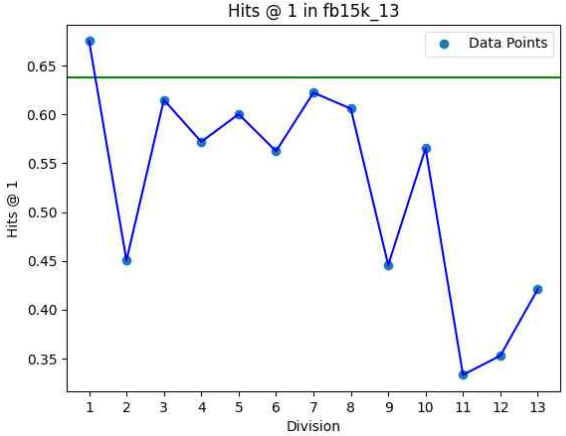
Experiments on RotatE/PathCon

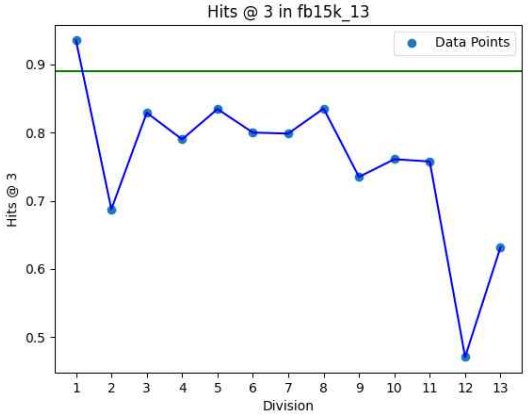
RotatE / 13

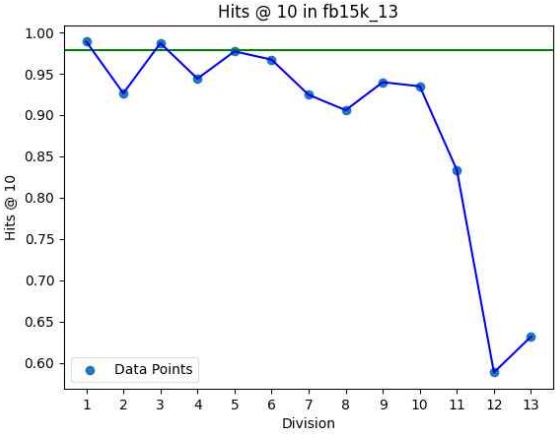




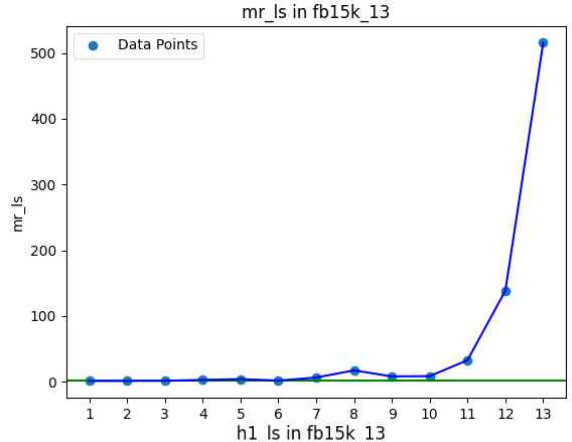


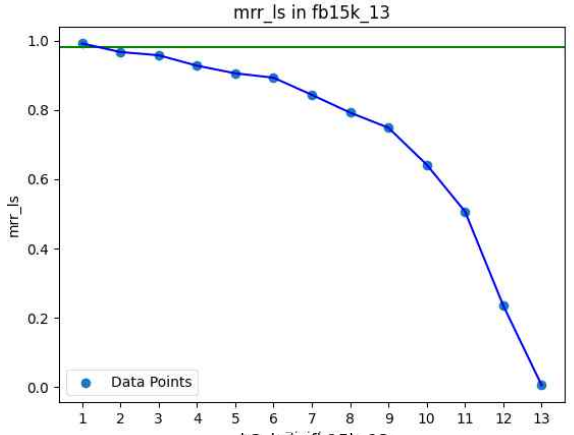




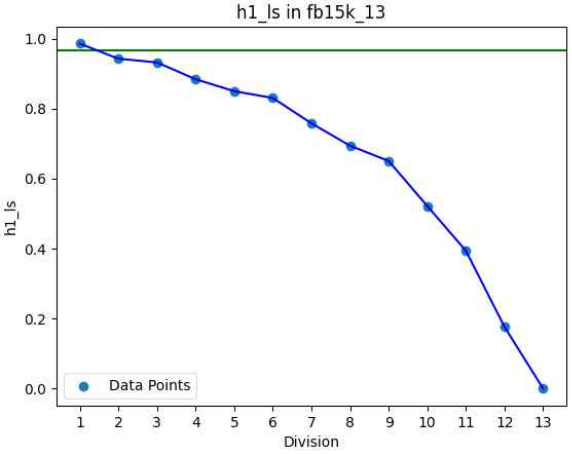


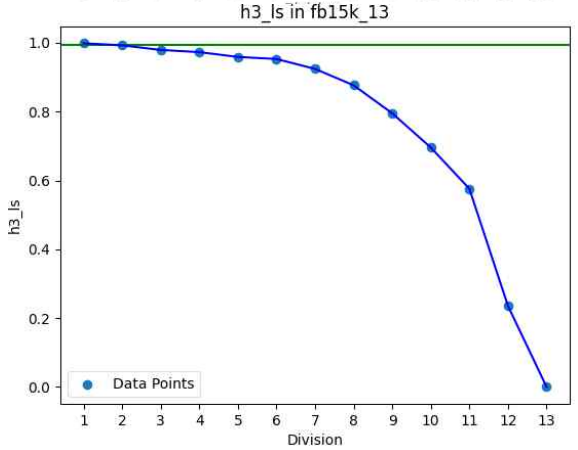
Experiments on RotatE/PathCon

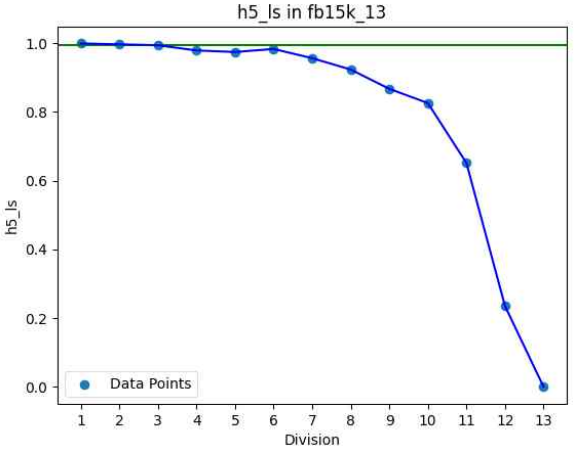




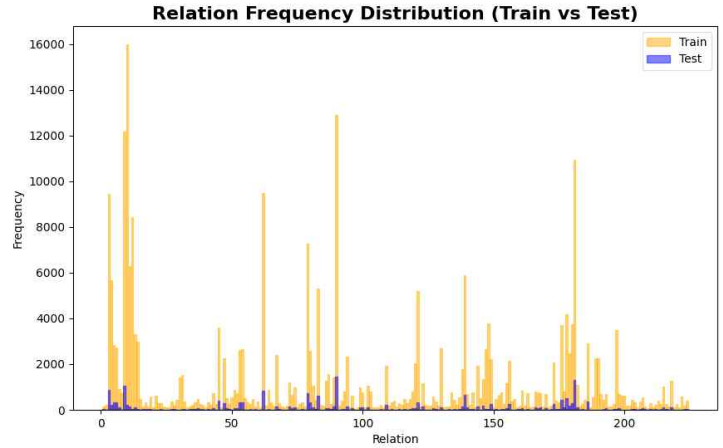
PathCon / 13







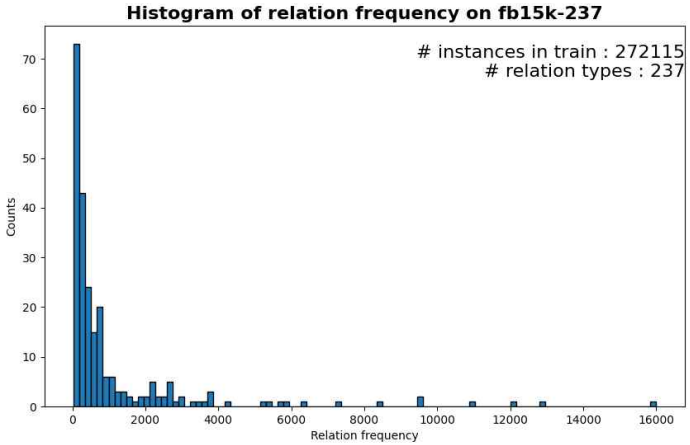
Experiments on RotatE/PathCon

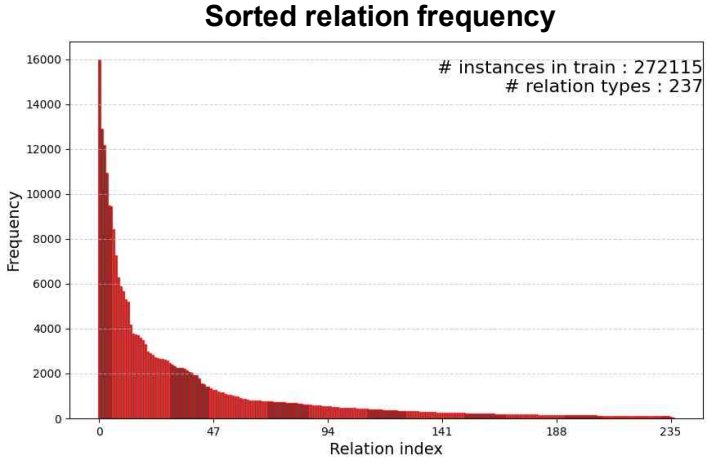


FB15k-237

 # train instance = 272,115
 # relation types = 237

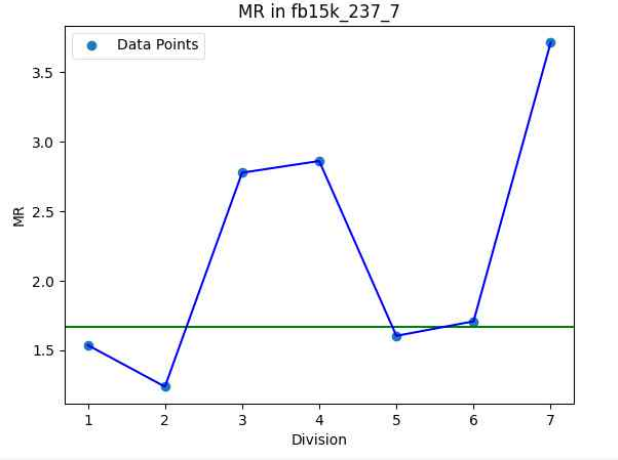
Top 5 least number of instance : 37, 90, 93, 99, 100

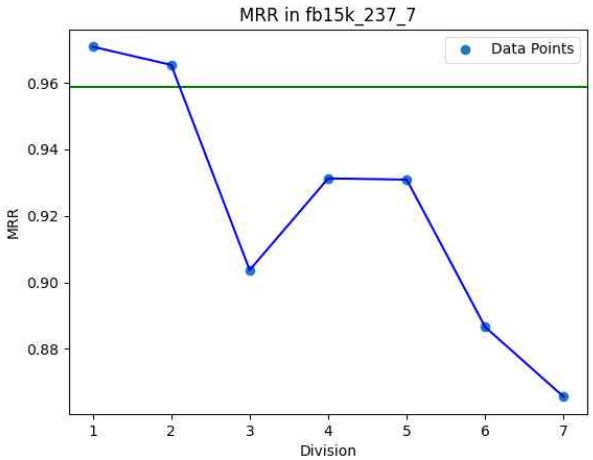


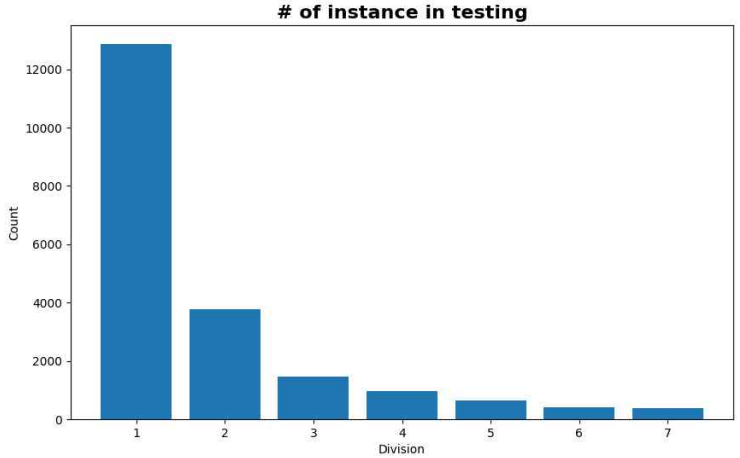


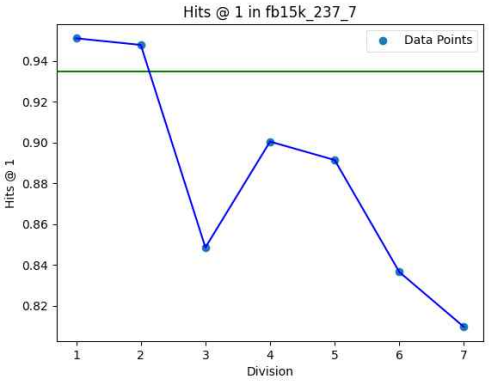
Experiments on RotatE/PathCon

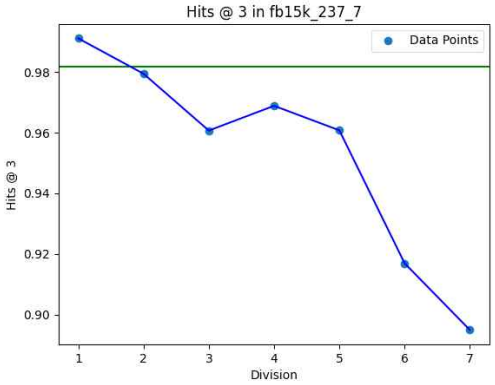
RotatE / 7

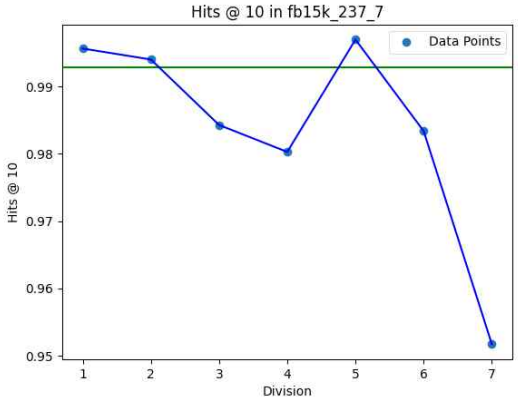




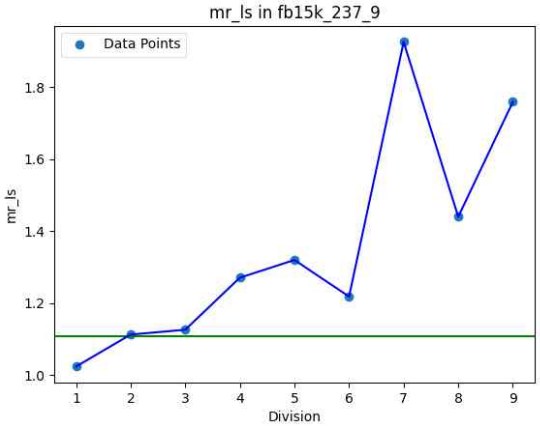


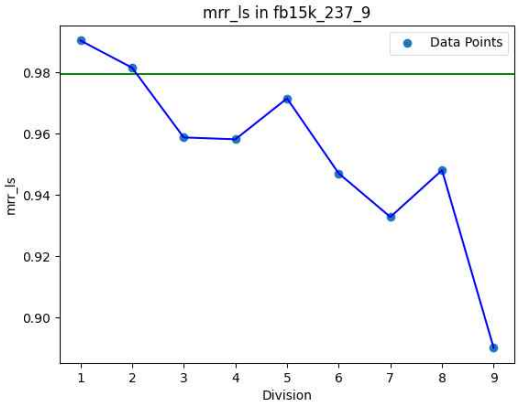




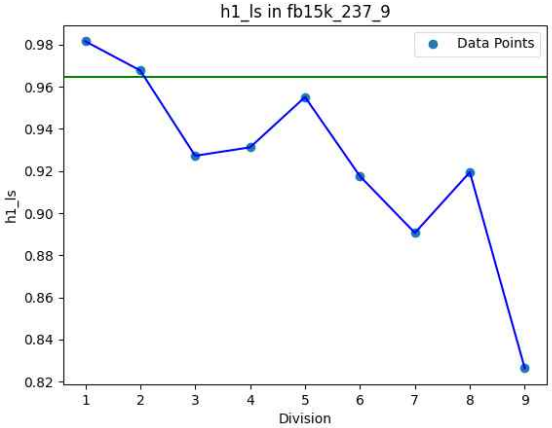


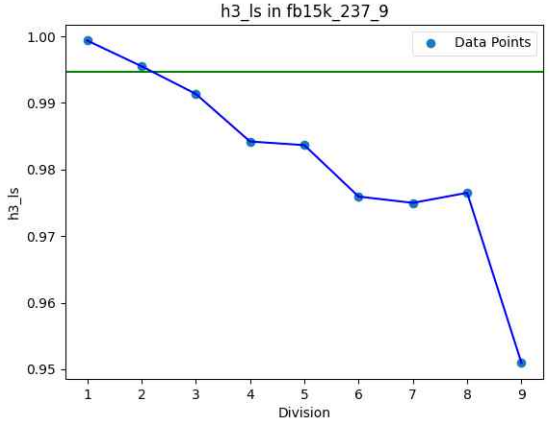
Experiments on RotatE/PathCon

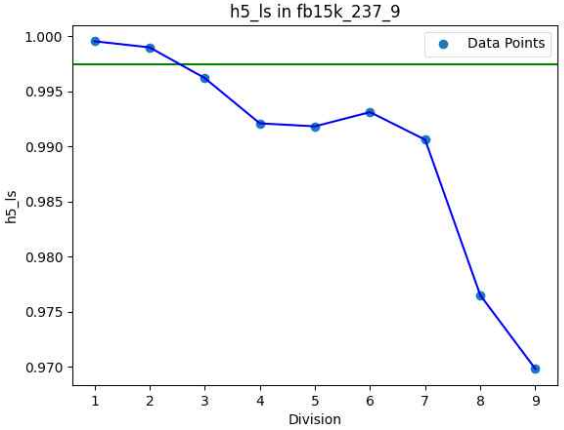




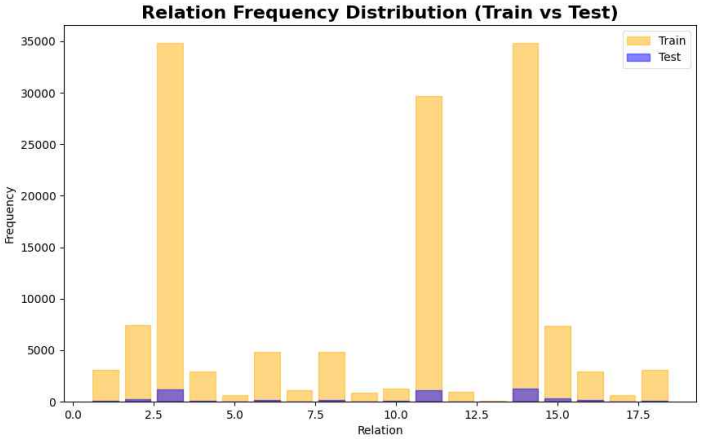
PathCon / 9

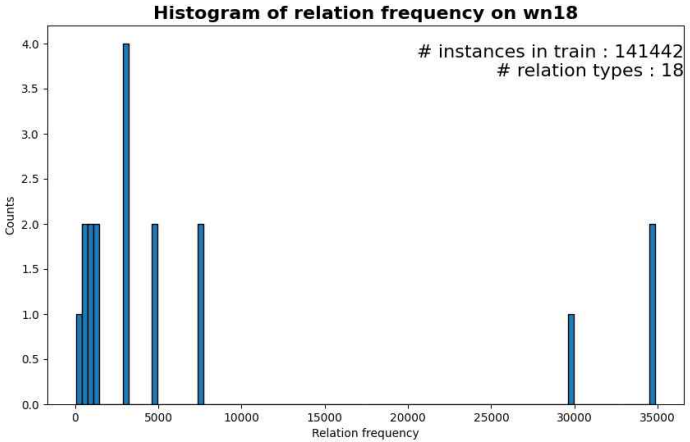






Experiments on RotatE/PathCon

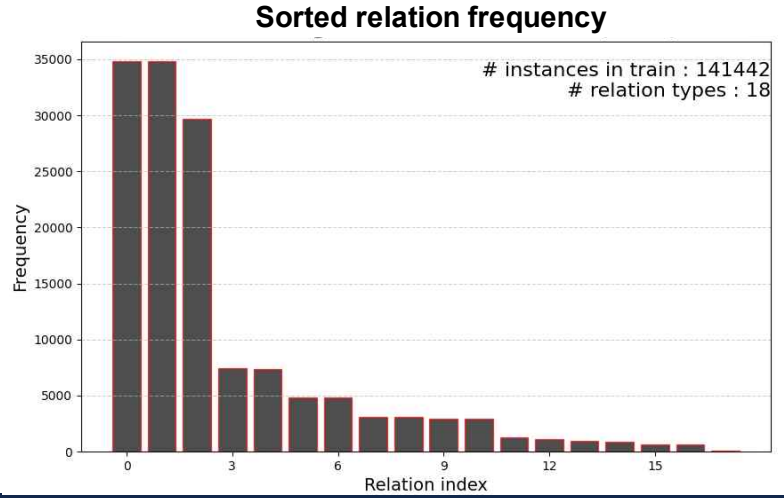




WN18

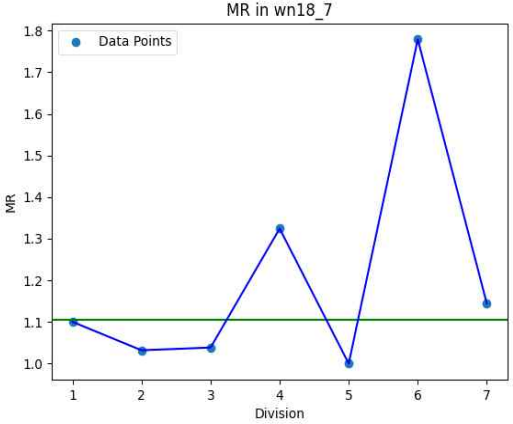
 # train instance = 141,442
 # relation types = 18

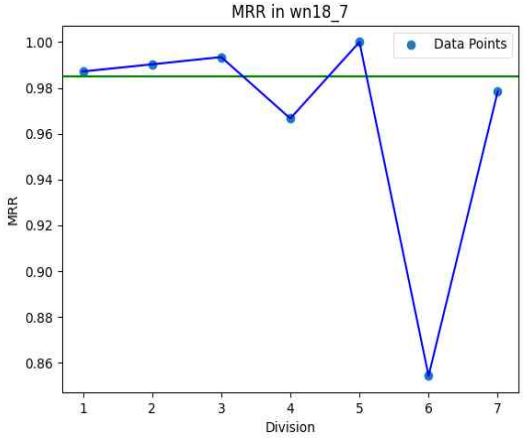
Top 5 least number of instance : 80, 629, 632, 903, 923

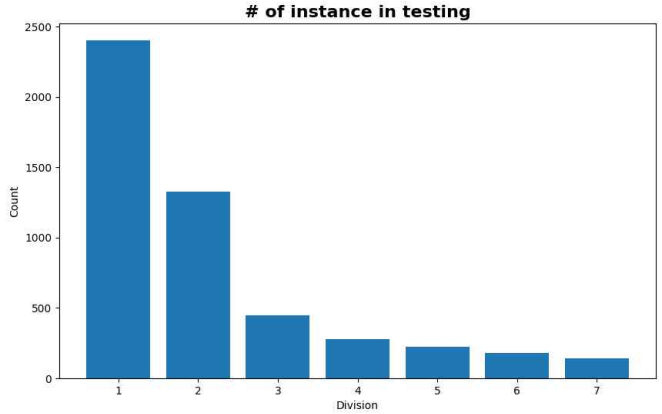


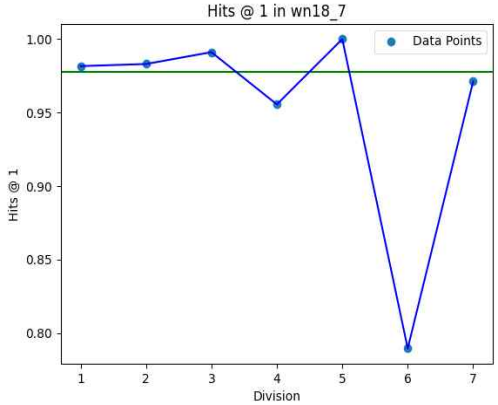
Experiments on RotatE/PathCon

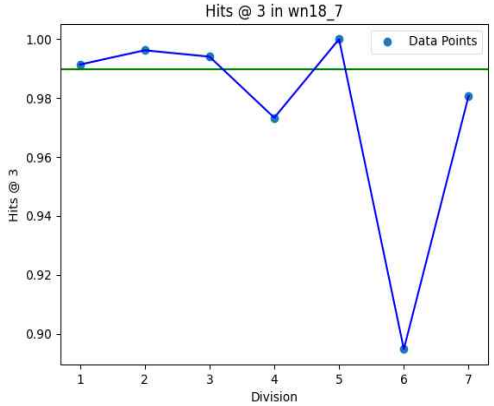
RotatE / 7

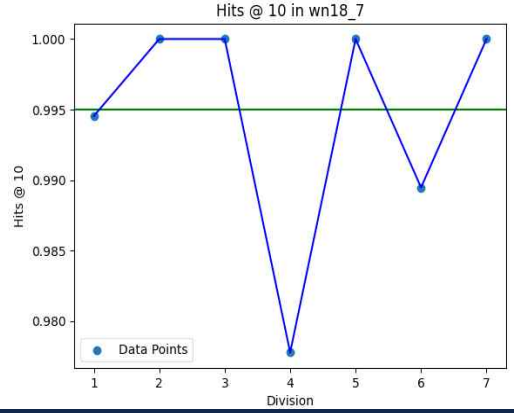




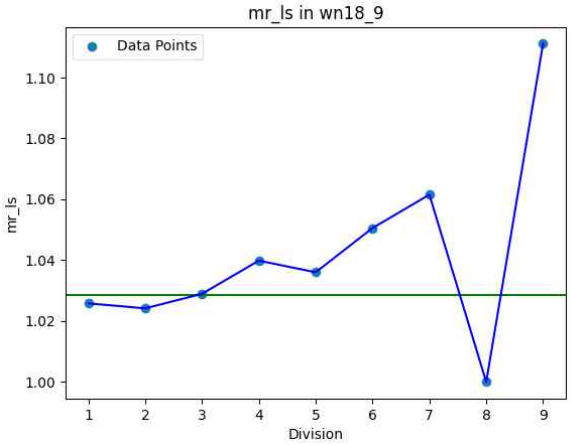


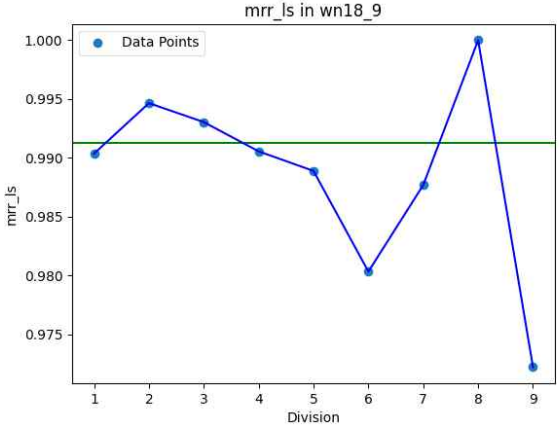




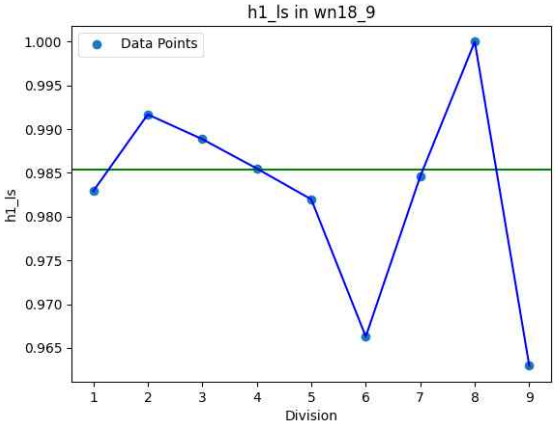


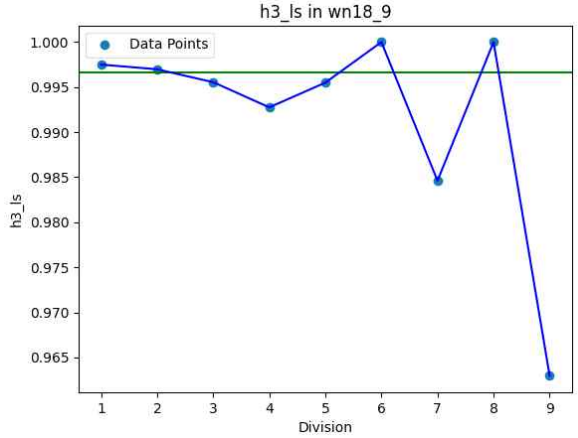
Experiments on RotatE/PathCon

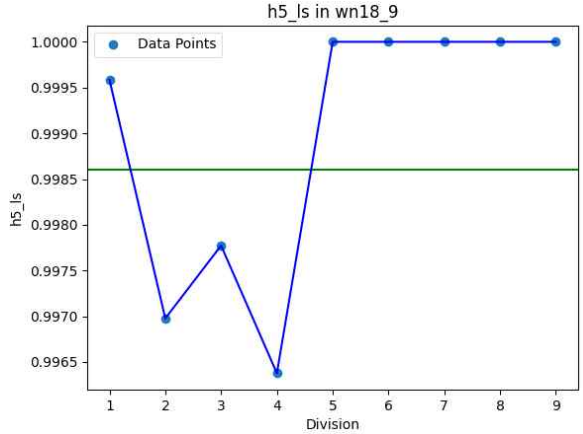




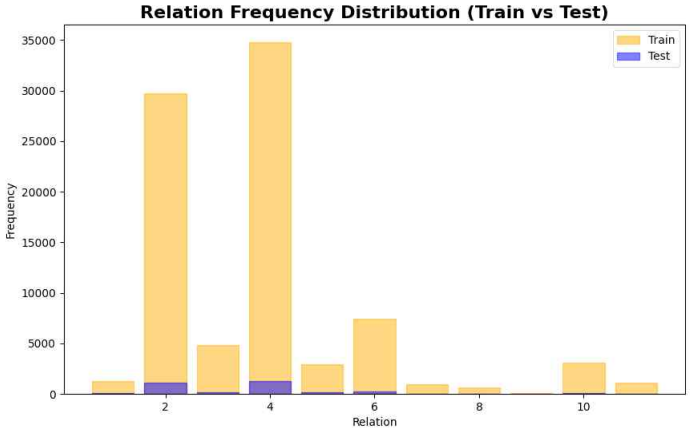
PathCon / 9







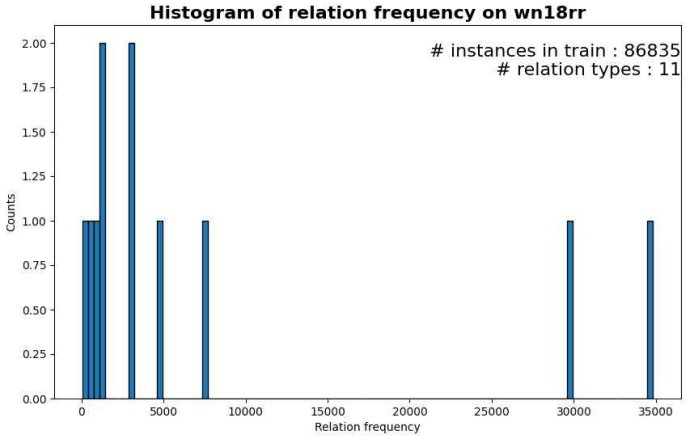
Experiments on RotatE/PathCon

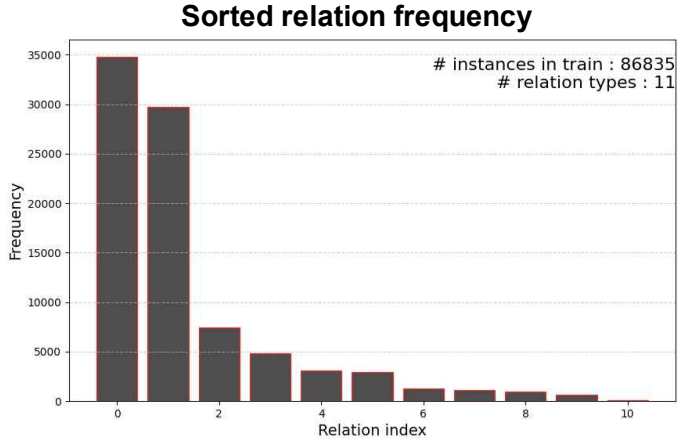


WN18RR

 # train instance = 86,835
 # relation types = 11

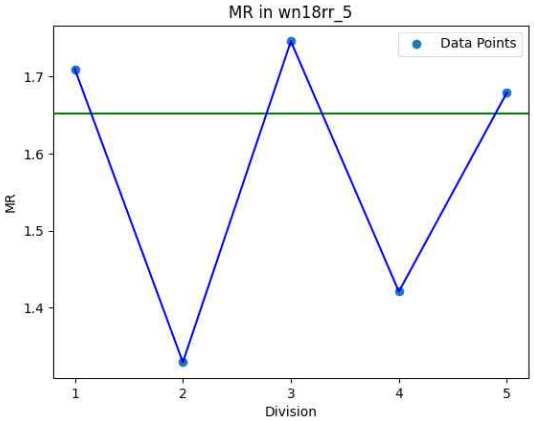
Top 5 least number of instance : 80, 629, 923, 1138, 1299

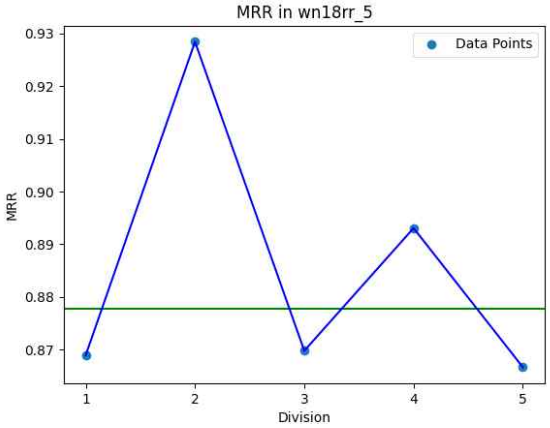


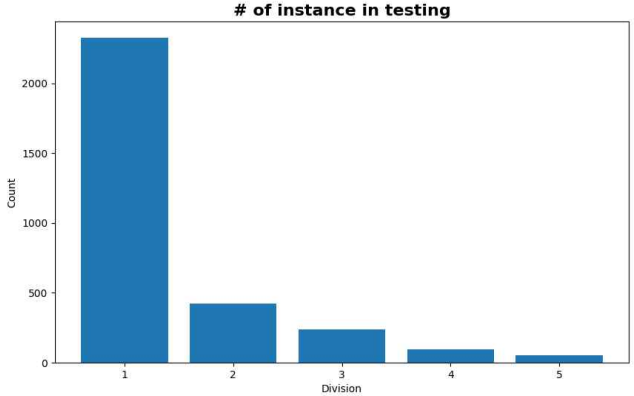


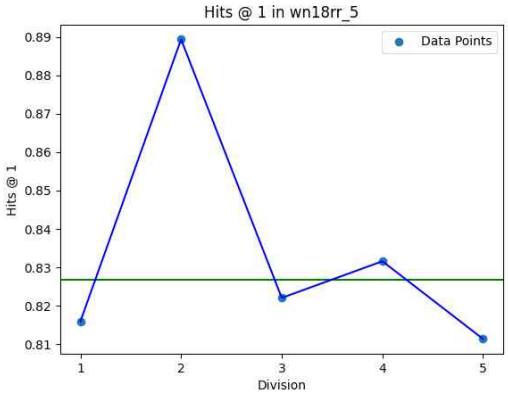
Experiments on RotatE/PathCon

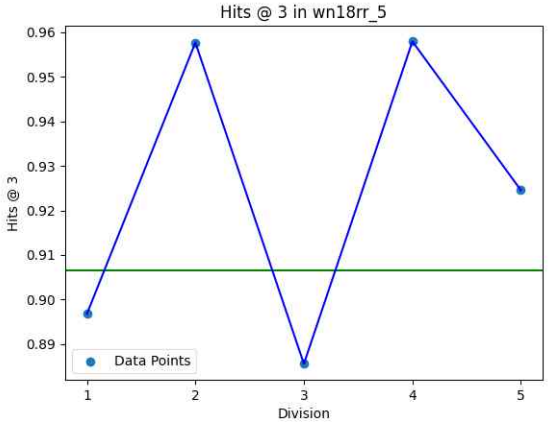
RotatE / 5

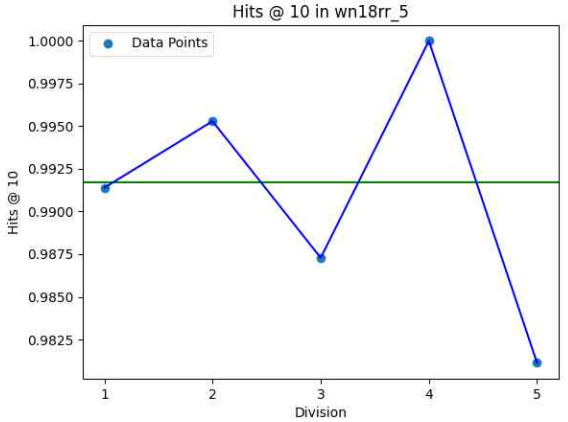




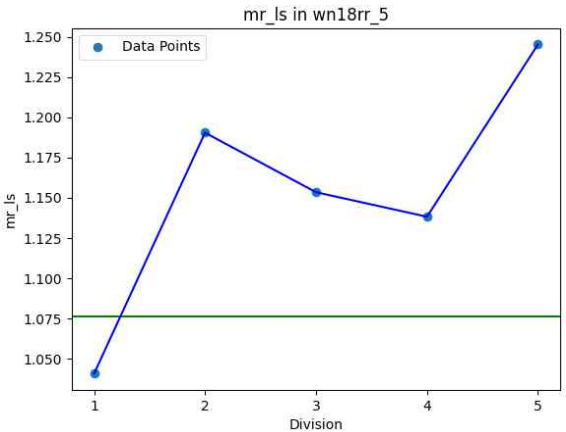


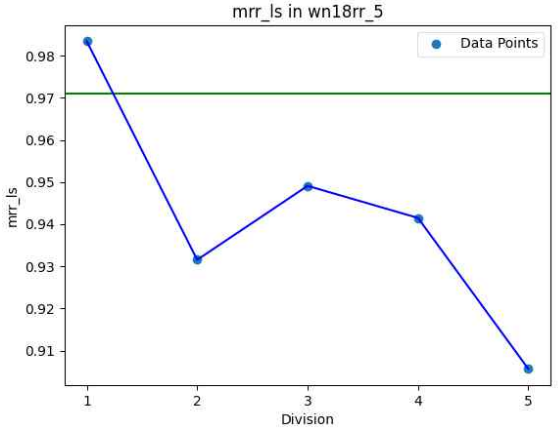




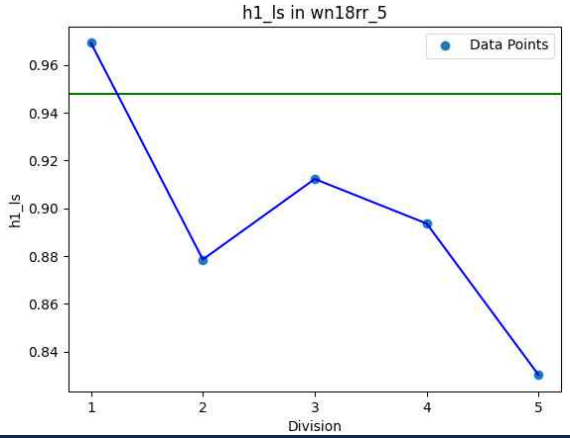


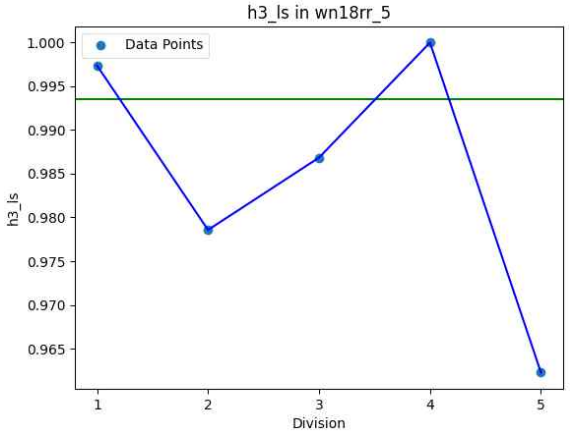
Experiments on RotatE/PathCon

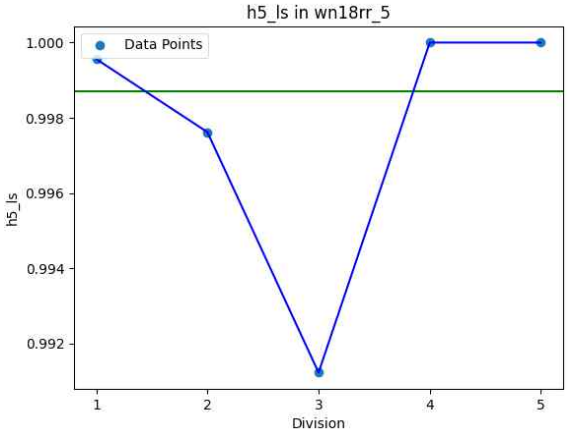




PathCon / 5



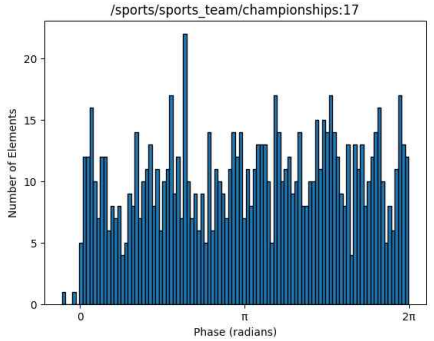




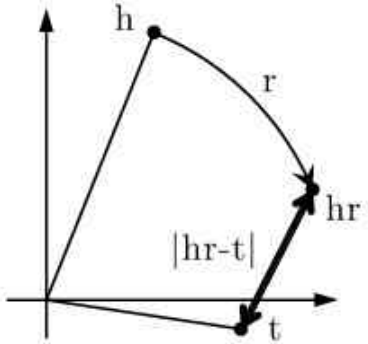
Experiments on RotatE/PathCon

analyzing RotatE relation embeddings

example of non-symmetry relation ->

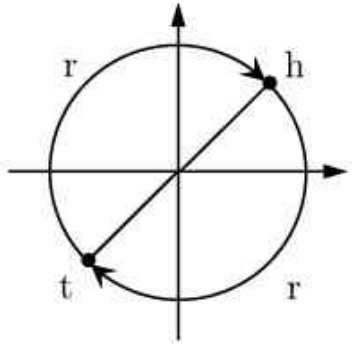


objective : $d_r(\mathbf{h}, \mathbf{t}) = \|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\|$



(b) RotatE models $\mathbf{r}$ as rotation in complex plane.

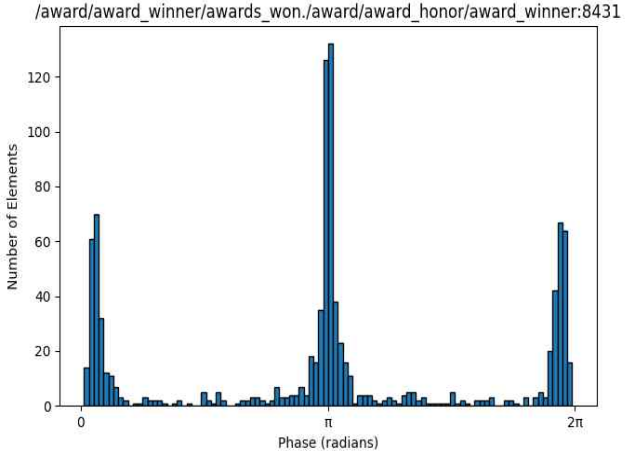
$$\mathrm{h} \circ \mathrm{r} = \mathrm{t}$$
$$\mathrm{t} \circ \mathrm{r} = \mathrm{h}$$
$$\mathrm{t} \circ \mathrm{r} \circ \mathrm{r} = \mathrm{t}$$
$$\therefore \mathrm{r} \circ \mathrm{r} = 1$$

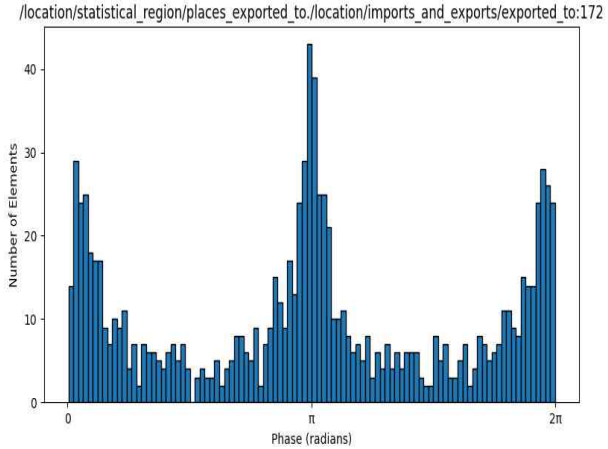


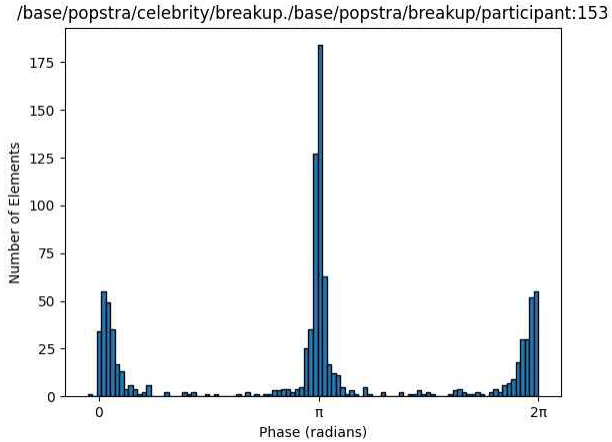
(c) RotatE: an example of modeling symmetric relations $\mathbf{r}$ with $r_i = -1$

Experiments on RotatE/PathCon

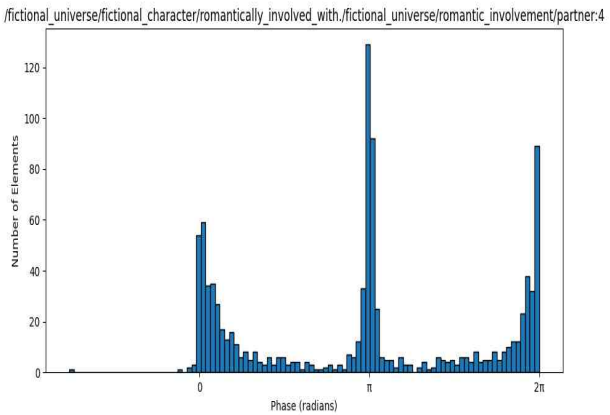
phase histograms of symmetry relations and counts

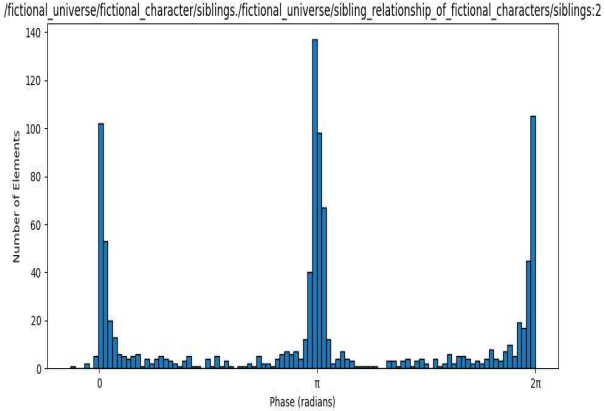


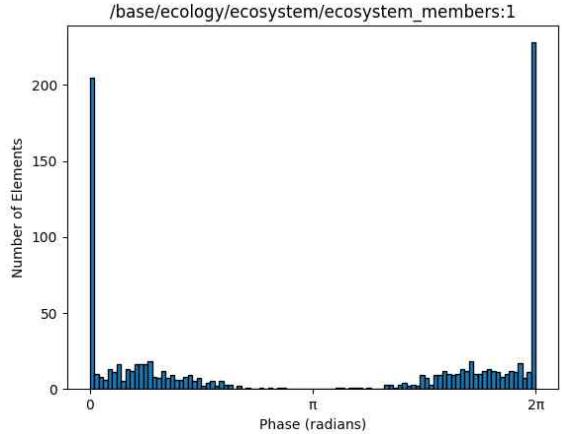


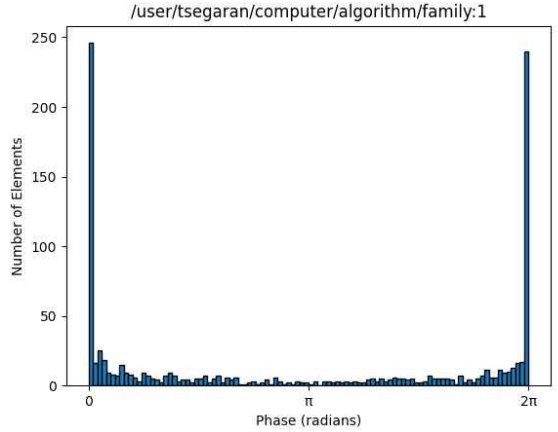


Experiments on RotatE/PathCon









Result analysis

1. Does the number of training instances indirectly affect the metric? -> Yes

Datasets like FB15k and FB15k-237 have a large number of relations, including many relations with relatively few instances, leading to metrics roughly proportional to the number of instances across intervals. However, for datasets such as WN18 and WN18RR, where the number of relations was much smaller relative to the training size, no noticeable results were observed.

2. Does the long-tail relation significantly impact the total metric? -> No

This makes sense, as the total metric is an average of individual metrics, and metrics from intervals with a disproportionately large number of instances contribute more to the total metric.

Result analysis

3. Does the number of instances significantly affect the formation of RotatE symmetry relation embeddings? -> No

Thanks to the numerous training epochs, there is no concern that the necessary characteristics of the relations are insufficiently learned.

4. Is it unnecessary to worry about long-tail relations in KGC because they have a minimal impact on the total metric? -> No

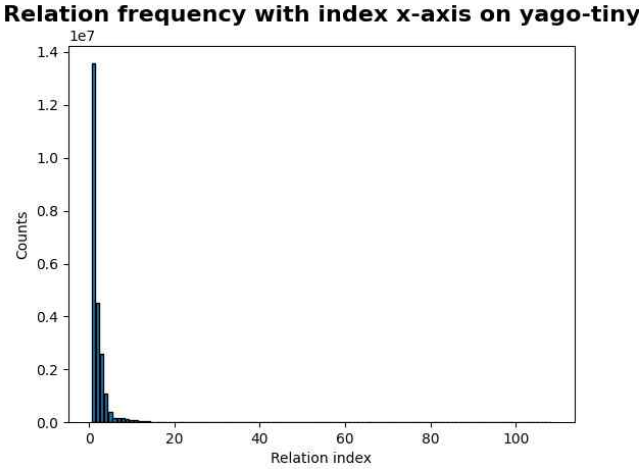
The total metric is a diluted result dominated by relations with a large number of instances. This means that despite potentially inadequate learning of long-tail relations, the environment still produces good metrics.

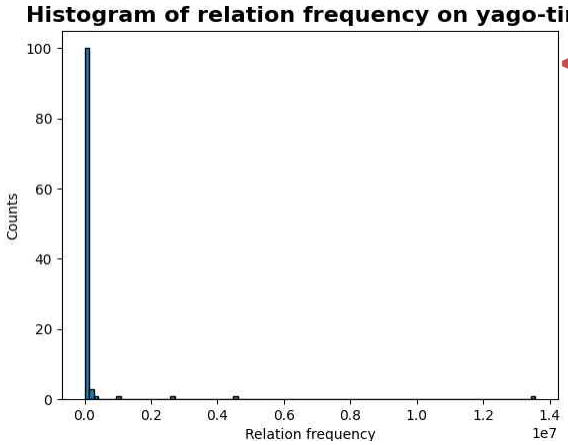
Additional observation

- **How about in real-world KGs?**
  - □ The datasets examined earlier are specifically presented for KGC (Knowledge Graph Completion). They represent subsets of large knowledge graphs. Therefore, I became curious about how real-world KGs might differ and investigated YAGO (Yet Another Great Ontology).
  - □ YAGO-4.5-0.2-tiny is a KG that consists of 23,259,536 triples, which is 1.6% of the total size of YAGO-4.5-0.2.

Additional observation

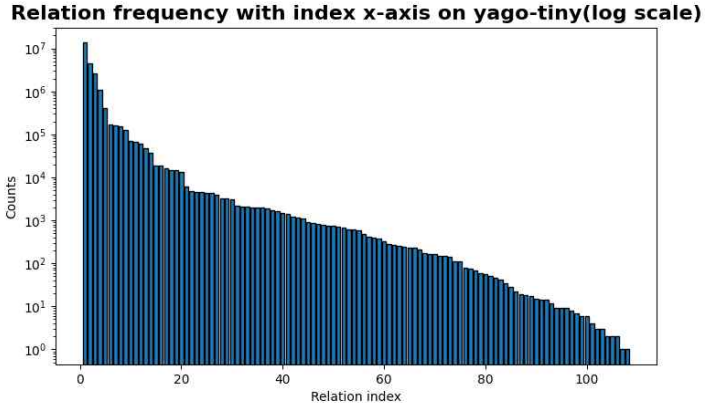
<Sorted by count>

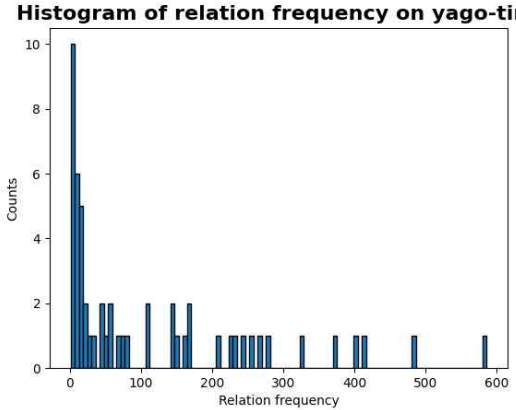




<Count histogram>

Total





First half

Conclusion

- **KGs are inherently long-tail**
  - □ If we aim to build a model, we need to acknowledge this fact
  - □ However, ordinary KGC models do not take this fact to account when modeling
  - □ Traditional metrics(MR, MRR, etc.) lack expression on evaluating reasoning models
- **More stuff to think about and possible future directions**
  - □ Experiment and compare multiple models schema wise(translation vs rule mining vs ... etc.)
  - □ Propose a new metric that can tackle the bias's of traditional metrics
  - □ Accomodate other mathodology(few-shot, etc.) for this particular nature of KG
  - □ What are the obvious examples when not accomodating long-tail symptoms?

Contact: Sooho Moon (Email: moonwalk725@cau.ac.kr)