Towards Linking Graph Topology to Model Performance for Biomedical Knowledge Graph Completion

Alberto Cattaneo[1], Thomas Martynec[2], Stephen Bonner[2], Carlo Luschi[1], Daniel Justus[1]

[1]Graphcore Research [2]Data Sciences and Quantitative Biology, AstraZeneca

Overview

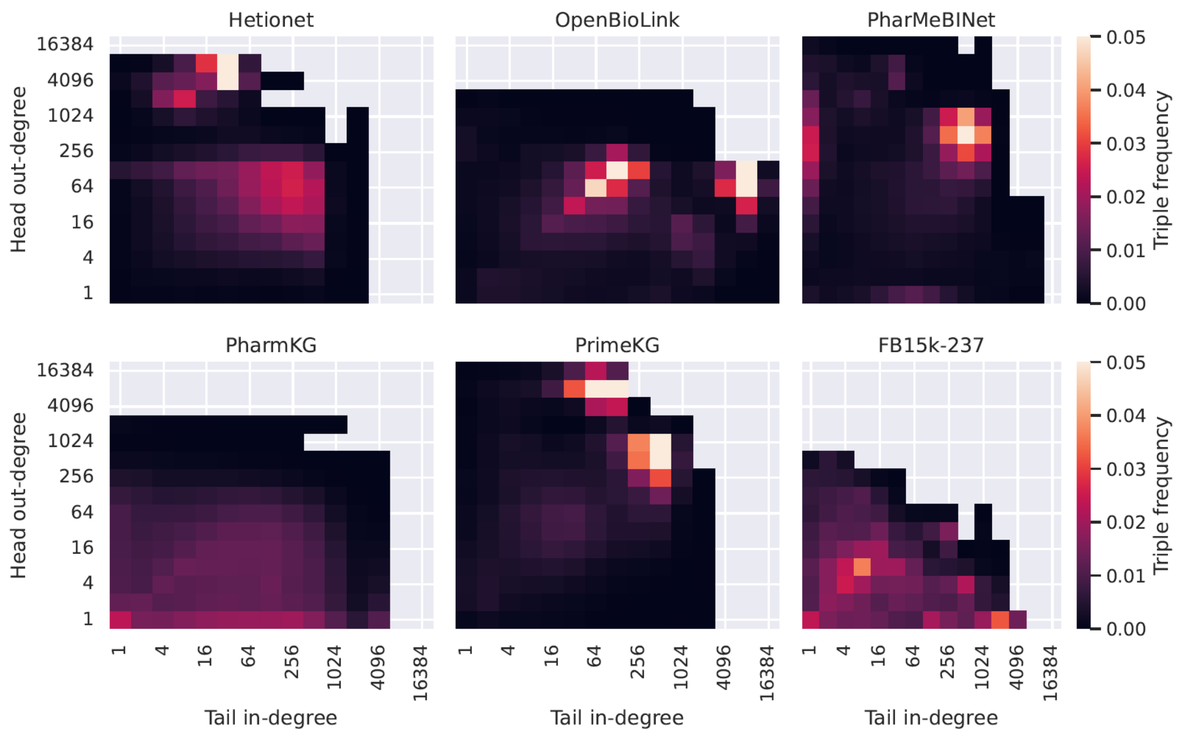
Link-prediction on Knowledge Graphs is widely used in biomedical research, from drug repurposing to biological target identification. However, little is known about the practical ability of ML models to leverage graph structure and local topology in order to make better predictions.

We conduct a comprehensive investigation into the **topological properties of public biomedical KGs** and establish **links to the accuracy of Knowledge Graph Embedding models** observed in complex real-world applications.

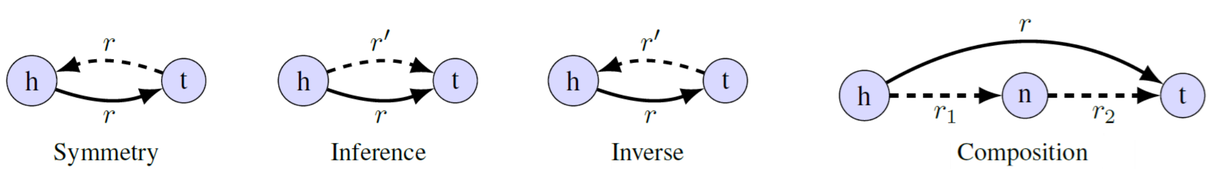
Edge Topological Properties

Edges in a KG are represented as *subject-relation-object* triples (h, r, t). We consider the following edge topological properties:

- **head out-degree** of (h, r, t): $\#\{t' : (h, r, t') \in KG\}$
- **tail in-degree** of (h, r, t): $\#\{h' : (h', r, t) \in KG\}$



- (h, r, t) is **symmetric** if $h \neq t$ and $(t, r, h) \in KG$
- (h, r, t) has **inverse** if $\exists\, r' \neq r$: $(t, r', h) \in KG$
- (h, r, t) has **inference** if $\exists\, r' \neq r$: $(h, r', t) \in KG$
- (h, r, t) has **composition** if $\exists\, r', r'', n$: $(h, r', n), (n, r'', t) \in KG$



Graph	Symmetry	Inference	Inverse	Composition
Hetionet	0.002	0.124	0.001	0.693
OpenBioLink	0.317	0.372	0.359	0.840
PharMeBINet	2.420×10^{-4}	0.052	0.002	0.598
PharmKG	0.197	0.124	0.059	0.651
PrimeKG	0	2.081×10^{-4}	0	0.807
FB15k-237	0.113	0.161	0.217	0.645

Occurrence of edge topological patterns as fraction of total triples in the datasets.

When investigating the causal effect of topological properties on the predictive accuracy of KGE models, we achieve **stronger statistical power** by adopting a different approach compared to previous studies:

- we look at the **actual head/tail degrees**, instead of using a coarser binary one/many cardinality classification;
- we consider topological properties **at the level of individual triples**, instead of averaging/pooling over relation types.

We've released **`kg-topology-toolbox`**, a Python library for computing topological metrics on any Knowledge Graph.

Read the paper

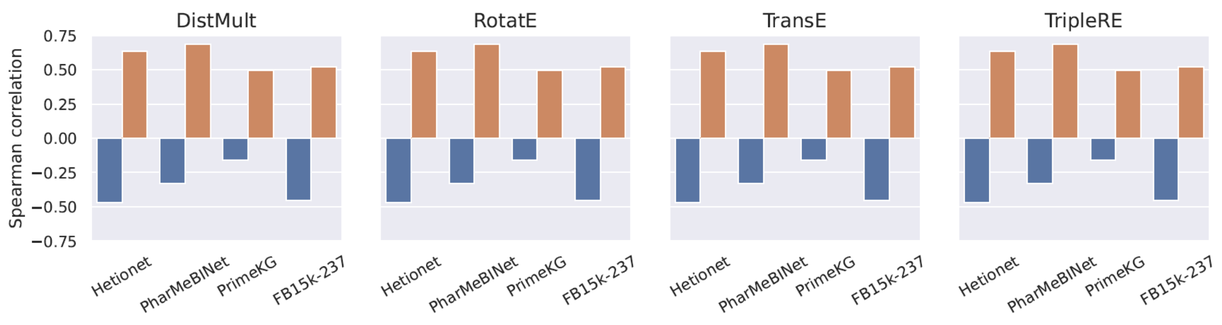
Try the library

Explore our data

Effect of Topology on Predictive Accuracy

Head and tail degrees are strongly correlated with the Mean Reciprocal Rank (MRR) of the ground truth t for $(h, r, ?)$ queries:

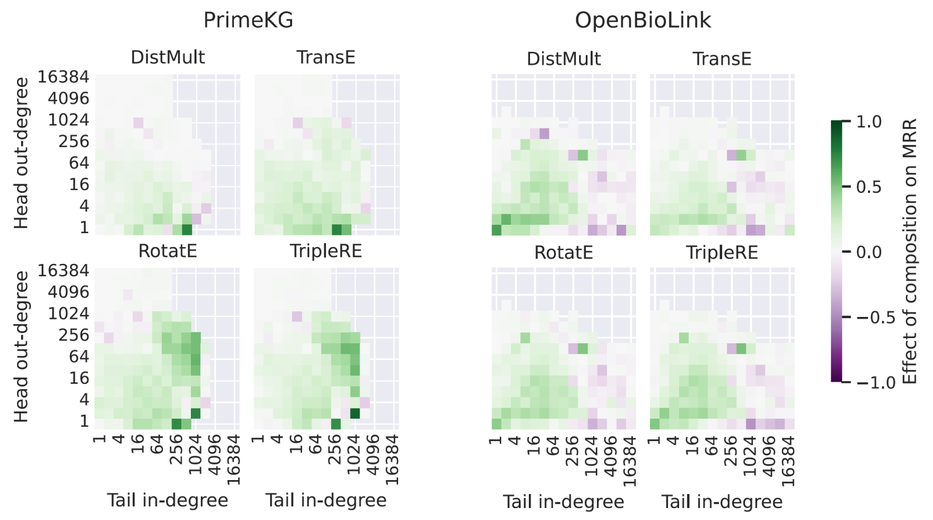
- a large in-degree of the tail node biases the model towards predicting it;
- a large out-degree of the head node implies multiple potential correct tail entities, making the task of predicting the specific one in the test set harder.



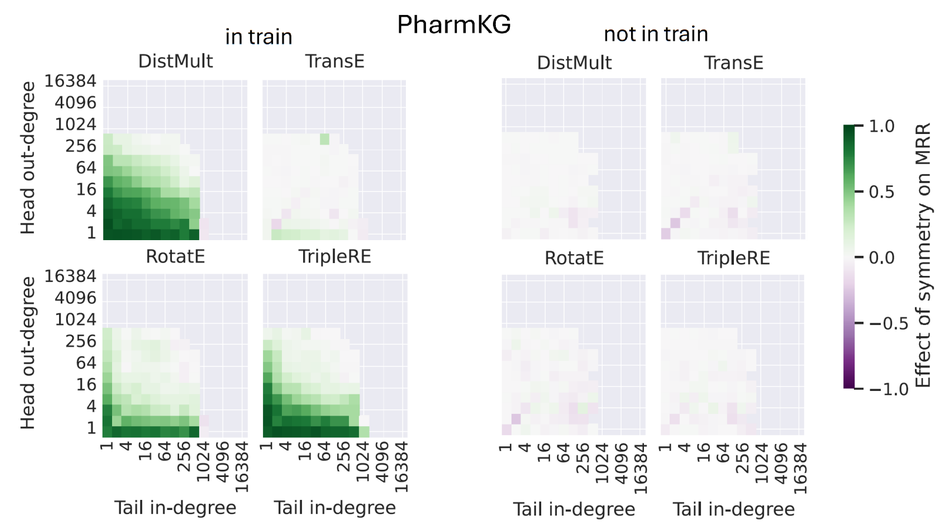
Spearman-rank correlation between head out-degree/tail in-degree and MRR of individual triples.

It is then crucial to **control for head/tail degrees** when investigating the effect of other topological patterns on MRR. Such effects are stronger when degrees are **small**.

Compositions are beneficial across all datasets and models.



For symmetry, inference and inverse, one needs to distinguish whether the **counterpart edge was seen during training**. If so, predictions become easier, assuming the KGE can model the pattern. Otherwise, we see little impact on MRR.



Test Case: Gene-Gene Interactions

We link KGE performance on gene-gene interactions across different datasets to topological properties.

- Predictions are hard due to large number of gene nodes (potential targets).
- OpenBioLink & PharmKG report better MRR thanks to high fraction of edges with inverse/inference and symmetry.
- This benefits especially DistMult (which models all edges as symmetric), while TransE is penalized by the inability to model symmetry.