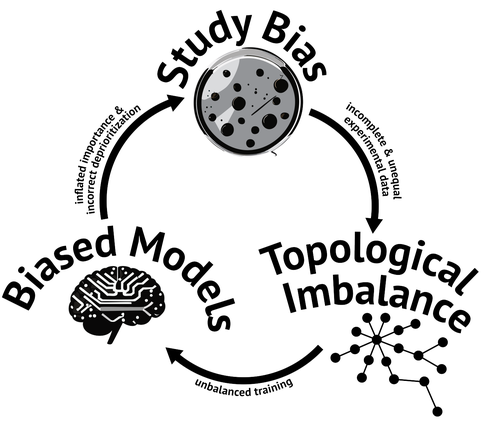
University of Colorado **Boulder**

A Vicious Cycle in Biological Knowledge Graphs

Study bias, topological imbalance, and biased predictive models are cyclicly connected

Michael Bradshaw
Ryan Layer
Department of Computer Science
& Biofrontiers Institute
michael.bradshawiii@colorado.edu
x/twitter @michael_boulder

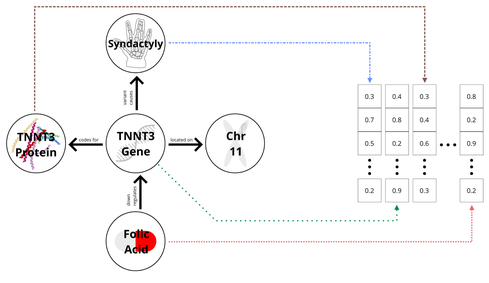
The Big Idea



Study bias, topological imbalance, and biased predictive models are cyclicly connected, creating a vicious cycle that hinders our ability to make accurate predictions about less studied genes, diseases, and drugs.

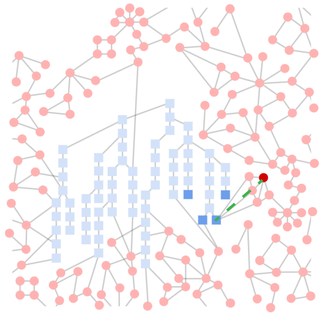
It has recently been shown how study bias in protein binding experiments creates topological imbalanced networks (Lucchetta et al. 2023). Contemporary research by Bonner et al. 2022 showed that knowledge graph embedding (KGE) link prediction (LP) models, when trained on topological imbalance networks become heavily biased towards recommending nodes with high degrees (Bonner et al. 2022). We show these overly prioritized nodes are those that have been extensively studied; when these predictions are used to generate hypotheses and direct experimental studies, creating more study bias. The cycle goes on and on, creating a system of preferential attachment, where well-studied (and subsequently connected) genes get more connections while less-studied nodes receive few to no more connections.

Background



A biological knowledge graph (KG) is a structured representation of biological information, capturing relationships and connections among various biological entities such as genes, proteins, diseases, and much more. It enables a systematic and organized way to store, integrate, and analyze complex biological data, facilitating a deeper understanding of the intricate relationships within biological systems. An increasingly popular method for using KG for computational tasks is to represent their nodes as low-dimensional knowledge graph embeddings (KGE). These KGEs describe node locations in multi-dimensional space with regard to their connection to the other nodes. Recently, KGEs have been used for many tasks including identifying disease modules, drug repurposing, and variant prioritization.

A common way these tasks are carried out is by link prediction (LP). KGE LP models predict which pairs of nodes in a KG are most likely to be connected, based on their distance to each other in the embedding space. Pairs of nodes that are not connected but that receive high scores or ranks by KGE LP models are indicative of latent connections - those that likely do exist but are yet to be observed. Output from these models is a means of hypothesis generation to direct experiment and study design to the avenues most likely to be fruitful. Shown to the right is an example of this; a patient's presenting phenotypes (dark blue squares) and variant harboring gene (dark red circle) have no direct connections to each other. An LP model may predict two of them are connected (dashed green line). Rather than exploring all possible gene-phenotype pairs related to the patient, researchers can focus on that pair most likely to have a connection - saving time and resources.



Find this project on GitHub

https://github.com/MSBradshaw/LinkPrediction

Results

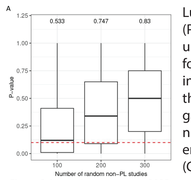
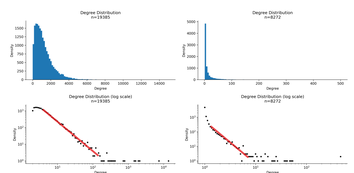
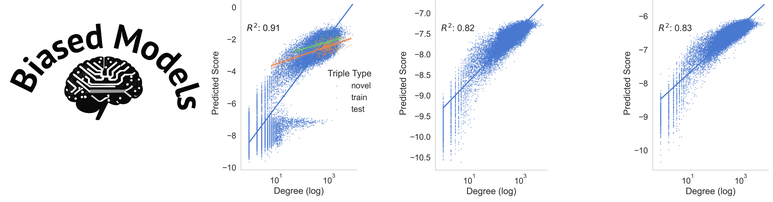


Figure 3.A from Lucchetta et al 2023

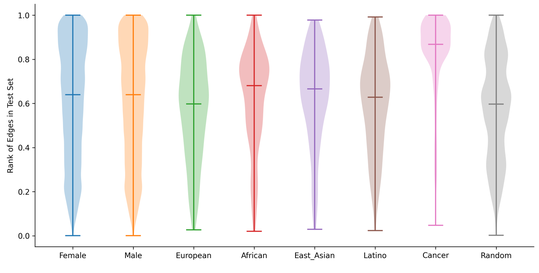
Lucchetta et al. 2023 showed that when creating protein-protein-interaction (PPI) networks from many small experimental datasets, study bias, specifically unequal use of proteins as bait or prey in binding experiments results in the formation of power-law degree distributions. In this study, they combined varying numbers of small datasets that did not exhibit power-law distributions; their results show that even with a relatively low number of experiments aggregated (100), power-law distributions form in most networks. When a large number of studies (300) is combined, it becomes almost inevitable that power-law distributions form. The test statics used here is Clauset's power-law test (Clauset et al. 2007), which uses a power-law distribution as the null hypothesis,



Topolical imbalances, like PLs do not arise only because of study aggregation. STRING is made by aggregating hundreds of small PPI experiments, databases and text-mining the literature - opening it up to the study-biases. HuRI is an all-by-all bait-capture experiment, free from study bias, but still has a PL degree distribution. A key difference in these PPI networks is their hub-nodes. In STRING the highest degree proteins are those related to cancer and common processes like ATP creation. But in HuRI, hubs proteins are not enriched for any diseases or biological processes.



Bonner et al. 2022 demonstrated how topological imbalances, like PL degree distributions, can bias knowledge graph embedding (KGE) link prediction models. Using the Hetionet knowledge graph and a variety of KGE models, Bonner showed a strong correlation between a node's degree and how well it ranked in link prediction tasks, regardless of the node type. Notably, Bonner found that higher-degree proteins rank higher in LP tasks than lower-degree proteins, even if the lower-degree protein and query node's edge were in the training set! We reproduce this same finding with a more up-to-date knowledge graph made of STRING and the Human Phenotype Ontology. You can see in this plot LP results for three phenotypes, Seizures, Premature thelarche, and Global Developmental delay. In all three there is a strong correlation between degree and predicted score. The training, and test nodes and their correlations are shown in the seizure plot. If certain nodes always ranking well because of their degree, what nodes are being deprioritized by the same system?

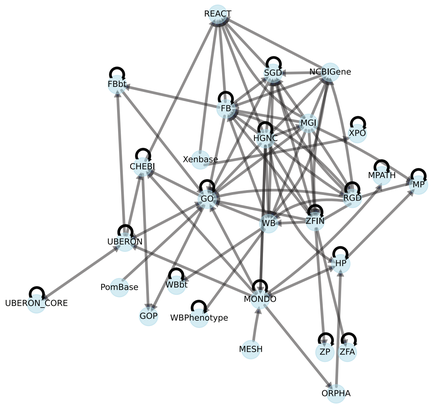


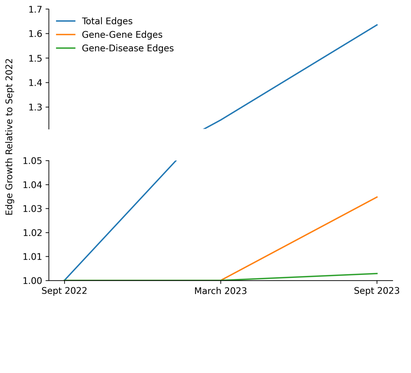
Using the Monarch Knowledge graph (Shefchek et al. 2020) and a TransE (Bordes et al.) KGE model, we perform LP experiments using genes that are differentially expressed in female and males (Oliva et al. 2020), and diseases caused by ancestry-specific variations (ASV) from European, Latino, East Asian and African populations and on gnomAD allele frequency data.

We hypothesized that since historically much of biomedical research has focused on European-male populations, we would find differences between these groups in terms of how they do in link prediction. We found no difference between these groups, except European vs East Asian, where we found the opposite of what we expected, East Asian ASV rank better than European ASVs. However, what we did find is our KGE LP model performed quite poorly on the sex and ancestry-specific groups. The median rank for edges that should be connected to proteins/diseases in these groups falls between 0.60 and 0.68 for all of them. When it comes to LP, if you are not first you are last. Typically only the top X or X% of nodes are looked at, so if most of a class's nodes are ranking < 0.90, they will never be considered. Interestingly genes related to cancer, rank quite well (median 0.87).

Link prediction is used as a form of hypothesis generation to inform experimental direction. If the results coming out of KGE LP are skewed toward high-degree nodes, it will lead research to further study high-degree nodes that have already been studied – exasperating the issue of study bias and continuing the vicious cycle of bias.

Methods





Using the Monarch Knowledge graph (Shefchek et al. 2020) and a TransE (Bordes et al.) KGE model, we perform LP experiments using genes that are differentially expressed in female and males (Oliva et al. 2020), and diseases caused by ancestry-specific variations (ASV) from European, Latino, East Asian and African populations and on gnomAD allele frequency data.

We hypothesized that since historically much of biomedical research has focused on European-male populations, we would find differences between these groups in terms of how they do in link prediction. We found no difference between these groups, except European vs East Asian, where we found the opposite of what we expected, East Asian ASV rank better than European ASVs. However, what we did find is our KGE LP model performed quite poorly on the sex and ancestry-specific groups. The median rank for edges that should be connected to proteins/diseases in these groups falls between 0.60 and 0.68 for all of them. When it comes to LP, if you are not first you are last. Typically only the top X or X% of nodes are looked at, so if most of a class's nodes are ranking < 0.90, they will never be considered. Interestingly genes related to cancer, rank quite well (median 0.87).

We train a TransE knowledge graph embedder on the Monarch KG and use it for link prediction experiments. Monarch KG, is highly heterogeneous containing 28 different nodes types with 60 types of relations connecting them. Using versions of Monarch KG from the recent past, we formulate our LP task as a forecasting problem to create a realistic test environment of how KGE LP models would have actually performed. We use versions of the KG from September 2022, March 2023 and September 2023, for training, validation and testing respectively. For our LP ranking test experiments, we find all edges relating to a certain group of genes or diseases that were added to the graph in September 2023. We then get the rank of these new connections from the model - these are the values shown in the violin plots.

Future Work

What other groups of genes/diseases/drugs are overly or under-prioritized?

If we retrain on a KGE without PPI study bias can we mitigate model bias?

Can different training techniques be used to improve outcomes?

References

Barabasi, A. L., and R. Albert. 1999. "Emergence of Scaling in Random Networks." Science 286 (5439): 509–12.
Bonner, Stephen, Ian P. Barrett, Cheng Ye, Rowan Swiers, Ola Engkvist, Charles Tapley Hoyt, and William L. Hamilton. 2022. "Understanding the Performance of Knowledge Graph Embeddings in Drug Discovery." Artificial Intelligence in the Life Sciences 2 (December): 100036.
Bordes, Antoine, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. n.d. "Translating Embeddings for Modeling Multi-Relational Data." Accessed December 4, 2023. https://papers.nips.cc/paper_files/paper/2013/file/1cec-c7a77928ca8133fa24680a88d2f9-Paper.pdf.
Clauset, Aaron, Cosma Rohilla Shalizi, and M. E. J. Newman. 2007. "Power-Law Distributions in Empirical Data." arXiv [physics.data-An]. arXiv. http://arxiv.org/abs/0706.1062.
Lucchetta, Marta, Markus List, David B. Blumenthal, and Martin H. Schaefer. 2023. "Emergence of Power-Law Distributions in Protein-Protein Interaction Networks through Study Bias." bioRxiv. https://doi.org/10.1101/2023.03.17.533165.
Oliva, Meritxell, Manuel Muñoz-Aguirre, Sarah Kim-Hellmuth, Valentin Wucher, Ariel D. H. Gewirtz, Daniel J. Cotter, Princy Parsana, et al. 2020. "The Impact of Sex on Gene Expression across Human Tissues." Science 369 (6509). https://doi.org/10.1126/science.aba3066.
Shefchek, Kent A., Nomi L. Harris, Michael Gargano, Nicolas Matentzoglu, Deepak Unni, Matthew Brush, Daniel Keith, et al. 2020. "The Monarch Initiative in 2019: An Integrative Data and Analytic Platform Connecting Phenotypes to Genotypes across Species." Nucleic Acids Research 48 (D1): D704–15.