GraphSAGE PPI Dataset Leakage

Introduction

The application of machine learning (ML) to biological datasets has become a powerful tool for understanding complex biological networks. In recent years, graph-based ML models have gained traction for analyzing biological networks, with datasets like the GraphSAGE protein-protein interaction (PPI) dataset [3, 4] serving as a popular benchmark for node classification tasks. Since its release in 2017, the GraphSAGE PPI dataset has been widely adopted for benchmarking, supporting various graph-based ML evaluations. However, recent reports of unbelievably high scores on this dataset have raised concerns about potential data leakage, suggesting that it may not accurately represent the challenges inherent in node classification [1, 6].

Despite its widespread use, the construction of the GraphSAGE PPI dataset is poorly documented, leading many researchers to use it without fully understanding its limitations. Critical details, such as how node features, labels, and tissue-specific networks were chosen, are vague [5]. The lack of information about the dataset's creation complicates the interpretation of results, limits the dataset's reliability as a generalizable benchmark, and sets back other researchers' ability to create new, comparable datasets from other gene or protein sets.

Given these concerns, the widespread use of the GraphSAGE PPI dataset as a standard benchmark has significant implications for the reliability of ML models in graph-based analysis. This project's focus is to examine the dataset's construction and uncover potential sources of data leakage. By highlighting these limitations, this investigation and project seeks to have the research community reconsider the dataset's use for evaluating graph-based models and to encourage the development of more transparent and robust benchmarks.

Approach

<u>Identifying Graphs, Features, and Labels</u>

One of the main goals is to understand and reconstruct the underlying structure and selection process used to build the GraphSAGE PPI dataset [4]. There is limited documentation on how this dataset was curated, which presents a challenge in understanding its creation. This stage of the project will involve:

- *Reconstructing Dataset Components*: I will investigate how node features and class labels were created and identify the specific nodes, tissue types, and graph structures included in the dataset. This includes tracing the dataset's creation and outlining a clear process for recreating a dataset similar to the GraphSAGE PPI dataset, ensuring transparency and reproducibility.
- *Selection Criteria*: I will analyze and document the selection criteria for the graphs and nodes, identifying potential biases toward well-studied proteins or genes that may affect model training. This includes checking for consistent representation of specific nodes across graphs and examining edge patterns to be able to assess overrepresentation of well-studied nodes, which may indicate sampling biases.

<u>Exploring Signs of Data Leakage</u>

Building on the information about the dataset's components, this part of the project will explore the possibility of data leakage and assess its impact on model evaluation through a series of experiments and tests:

- *Performance Testing with PyTorch Geometric Models*: With the implementation of GraphSage PPI dataset in PyTorch Geometric [2], I will run baseline models on the original data to serve as points of reference for model performance.
- *Alternative Dataset Testing*: Using insights gained from analyzing the dataset creation process, I will create a new dataset or use other well-known datasets, such as the OmniDataset [7], to test model performance. By applying the same models to these alternatives, I want to compare the results with the GraphSAGE PPI dataset, which may reveal dependencies on dataset-specific patterns.
- *Ablation Studies*: By leaving out parts of the GraphSage PPI dataset, I will do ablation studies to see the model sensitivity to different dataset components to find potential overfitting or data leakage.
- *Randomization and Fully Connected Graph Experiments*: To probe dataset structure, I will randomize edges within the dataset's graphs and evaluate the models' performances under these conditions. Additionally, I will generate fully connected graphs to observe how evaluation metrics change in response to different graph structures.
- *Feature-Label Correlation Analysis*: I will analyze correlations between feature vectors and class labels across graphs to identify potential data leakage. This analysis will quantify correlations to determine if feature vectors unintentionally predict labels, which would indicate data leakage.

Significance

The GraphSAGE PPI dataset is widely used in benchmarking ML algorithms, yet its lack of details of its creation and potential data leakage threaten the validity of findings and progress in graph model research. By analyzing the dataset's construction, this project will bring answers to its limitations.

Once completed, I expect to have an answer on how the datasets' biases, ambiguities, and data leakage affect graph model performance. Key questions I hope to answer by examining the dataset's creation include: How were the networks constructed and/or chosen and what criteria were used to select specific nodes? What factors in the dataset's generation process might have contributed to high performance scores? How are node features and labels defined, and do they introduce correlations that could lead to data leakage? Additionally, how does well-studied genes and proteins in the dataset influence a models' generalizability, and could this bias be responsible for high performance results? By answering these questions, this project aims to uncover the specific parts that are leading to high scores with the GraphSAGE PPI dataset.

Through this investigation and project, I anticipate gaining valuable insights into best practices for dataset curation, documentation, and transparency, which are essential for reproducibility and model validation in ML applications and model creation.

References

Note: I am in communication with Dr. Arjun Krishnan regarding the history and origins of the GraphSAGE PPI dataset to gain additional insights and context for this project.

[1] Chen, M., Wei, Z., Huang, Z., Ding, B., & Li, Y. (2020). *Simple and Deep Graph Convolutional Networks*. Proceedings of Machine Learning Research. Retrieved October 28, 2024, from https://proceedings.mlr.press/v119/chen20v.html

[2] DGL Team. (2018). Source code for dgl.data.ppi — DGL 2.2.x documentation. Retrieved October 28, 2024, from https://docs.dgl.ai/_modules/dgl/data/ppi.html#PPIDataset

[3] Hamilton, W., Ying, Z., & Leskovec, J. (2017). *Inductive Representation Learning on Large Graphs*. NIPS papers. Retrieved October 28, 2024, from http://papers.nips.cc/paper/by-source-2017-671

[4] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). GraphSAGE: Inductive Representation Learning on Large Graphs. Retrieved October 28, 2024, from https://snap.stanford.edu/graphsage/

[5] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). *williamleif/GraphSAGE: Representation learning on large graphs using stochastic graph convolutions.* GitHub. Retrieved October 28, 2024, from https://github.com/williamleif/GraphSAGE

[6] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2017, October 30). *[1710.10903] Graph Attention Networks*. arXiv. Retrieved October 28, 2024, from https://arxiv.org/abs/1710.10903

[7] Zitnik, M., & Leskovec, J. (2017). SNAP: Predicting multicellular function through multi-layer tissue networks. Retrieved October 28, 2024, from https://snap.stanford.edu/ohmnet/