Graph Neural Network (GNN) Efficiency Improvements by Using Graph Sampling Strategies

Kunxiao Gao & Yuzhe Guo, Supervisor: Luana Ruiz

Introduction

Graph Neural Networks (GNNs): Powerful tools for processing graph-structured data, widely used in social network analysis, recommendation systems, and biological networks:
Challenges in GNNs: High computational cost and memory demands when training large graphs.
Graph Sampling Motivation:
1. Reducing graph size while maintaining predictive accuracy.
2. Optimizing performance by preserving key graph structures via sampling methods better than random sampling in smaller sampled subgraphs.

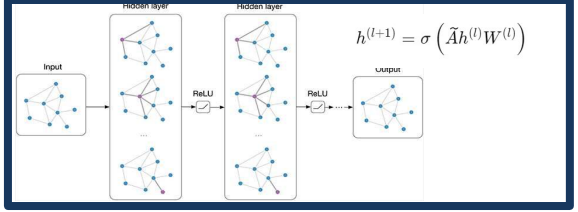
Objectives

1. Compare the performance of leverage score sampling, adapted BFS sampling, and random sampling for GNN predictions.
2. Reduce computational overhead while maintaining predictive accuracy.
3. Assess how well GNNs trained on sampled subgraphs generalize to larger or different graph structures.

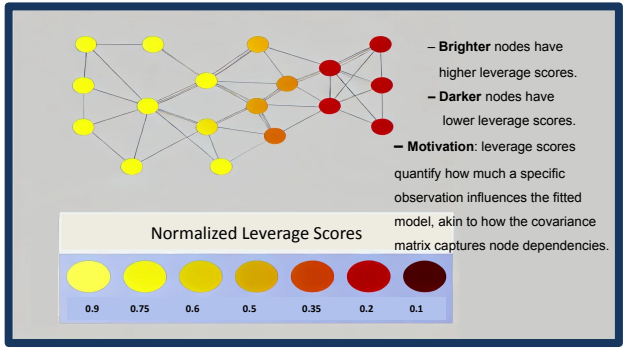
DATA

The MovieLens-100k dataset contains a total of **100,000 ratings**, provided by **943 users** for **1,682 movies**. Each rating in the dataset is an integer ranging from **1 to 5**, representing the user's level of preference or satisfaction with a given movie. To build the **movie similarity graph W**, pairwise similarities between movies are calculated using user ratings from the training subset by measuring normalized covariance for movies rated by the same users. Only the top **40** strongest similarities for each movie are kept. Finally, normalized W. **Tree distance** also will be recorded to measure distance between 2 graph.

GNN

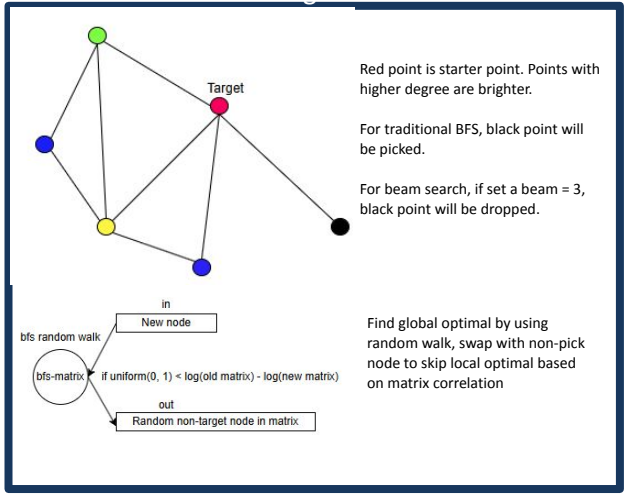


Leverage Scores Sampling



– **Brighter** nodes have higher leverage scores.
– **Darker** nodes have lower leverage scores.
– **Motivation**: leverage scores quantify how much a specific observation influences the fitted model, akin to how the covariance matrix captures node dependencies.

BFS Algorithm



Red point is starter point. Points with higher degree are brighter.

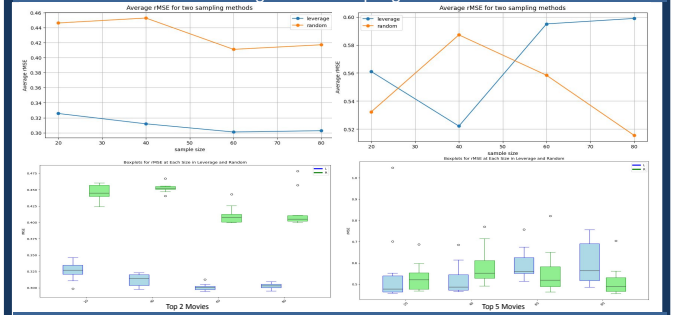
For traditional BFS, black point will be picked.

For beam search, if set a beam = 3, black point will be dropped.

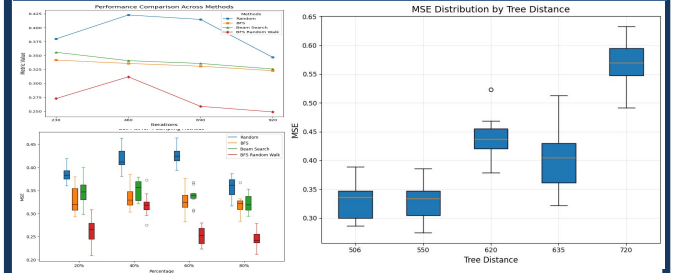
Find global optimal by using random walk, swap with non-pick node to skip local optimal based on matrix correlation

Results

Leverage Scores Sampling Methods



BFS, Beam Search, Random Walk



Conclusions

- **Leverage Scores Sampling** outperformed random sampling in predicting less movies rating at the same time. Achieved a Root Mean Square Error (RMSE) of **0.3**, which is **26.8% lower** than the RMSE of the random sampling method, at sample size 60%.
- BFS-random walk sampling outperformed random sampling in one-movie rating prediction. Achieved RMSE of **0.25** in the end. And **15%** lower than random sampling at 40% graph size.
- Future work and improvement should be focusing on applying sampling methods on predicting more movie nodes at the same time.
- The two sampling methods have the potential to lead to reduce computational overhead while maintaining predictive accuracy.