Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Community based Feature Aggregation in GNNs

David Klingenfuß

2910709

Prof. Dr. Ingo Scholtes

Chair of Machine Learning for Complex Networks
University of Würzburg

Supervisor: Jan Pichowski and Lisi Qarkaxhija

Würzburg, January 20, 2026

1. Motivation and Problem Statement

In Graph Neural Networks, nodes are typically updated by aggregating Information from their neighbors using Message Passing. That way, those models can capture structural patterns in the Data.

In this thesis, we will try an alternatice approach to node feature aggregation, based on community affiliation. The central idea is to perform clustering methods on the graph and then use the community affiliations of nodes to aggregate their node features, instead of their neighbors. The goal is to identify under which circumstances this approach performs comparably or even better than neighbor focused Message Passing. We expect this approach to change the strictly local inductive bias from normal GNNs by focusing on different patterns in the graphs.

2. Approach

The central challenge of this approach is obtaining meaningful clusterings and it is to be expected to be the main bottleneck, since we will not have a ground truth community labels for comparison. Therefore we will start by systematically trying out different clustering methods and comparing their performances. They will range from strictly graph-structure based algorithms, that ignore node features to approaches that incorporate both structural aswell as feature information. With synthetic graph data tailored to those clustering methods, we will try to find out under which conditions they work well, to reuse those insights later on actual data.

Since the primary focus is to compare standard message passing with the proposed community-based aggregation approach, the models we will use will deliberately be kept simple. A key aspect on the other hand will be to compare performance across datasets, to find out which data characteristics favor or hinder our approach.

There will be several approaches to community based node aggregation that we will try. The simplest way is to perform a single clustering on the graph and use those node affiliations for feature aggregation. Additionally, multiple clusterings can be applied all at once, either by focusing on different criteria or with different granularity, in the hopes of combining the strengths of different clustering methods and capturing both large scale and fine grained local patterns. Finally, a hybrid approach of community based and neighbor focused feature aggregation will be examined, by using both community affiliation and neighbors for feature aggregation.

3. Planned Procedure

- Building a custom data generation pipeline based on torchgeometrics SBM Dataset to be able to easily generate custom "Pre Clustered" synthetic graphs, that have features and are labelled, since all different kinds of those graphs will be needed throughout the project
- Creating an initial model using Infomap Clustering and a simple implementation of "CommunityPassing" to test the approach for a simple classification task
- Varying size and connectivity of data to compare how it affects effectivity of the approach
- Generating Data that shows feature similarity within same clusters and testing feature based clustering for the model
- Using different clustering methods to create different partitions of the graph, each contributing to aggregation, to see if this helps to extract more information
- Using the same clustering method, but in different sparsity to get both fine grain and sparse partitions of the graph, to aggregate nodes heavily with features within small block and only lightly with features from far reaching blocks

- Combining message passing and community based approach to aggregate features and testing
- In depth analysis of all synthetic tests so far and analyzing for which data properties they worked well, to project this on real data
- Using actual data like arxiv and mutag to test our approach and comparing how well the approach worked depending on the data and prediction task

4. Literature and Tools

- SBM to generate data
  https://pytorch-geometric.readthedocs.io/en/2.5.0/generated/torch_geometric.datasets.StochasticBlockModelDataset.html
- PathpyG for visualization
  https://www.pathpy.net/0.2.0-dev/
- arxiv and mutag datasets, aswell as others that will be added later
  https://www.kaggle.com/datasets/Cornell-University/arxiv, https://huggingface.co/datasets/graphs-datasets/MUTAG