Research Proposal

Previously we had been exploring various methods for improving the SeHGNN model's performance through augmenting the dataset. After doing some initial analysis of the results from the model, we noticed that correctly predicted papers tended to have more papers that cited it. This area was then further explored by masking the future citations edges during training to see the difference in results. After running this with k-fold cross validation, the results are as shown in table 1 below.

Table 1. K-Fold Results

Fold Number	Masking Test Accuracy (%)	No Masking Test Accuracy (%)
1	56.9339	57.2422
2	56.8633	56.2332
3	56.4668	56.7438
4	56.2142	56.2842
5	54.6022	56.2842
6	53.5593	55.1284
7	54.1402	56.0182
8	53.4914	55.3827
9	54.4724	55.9204
10	53.1982	55.3829
Avg:	**54.99419**	**56.06202**

Running the SeHGNN model with ten k-folds shows that papers with their future citations masked tend to have a lower accuracy. Based on these results, we predict that augmenting the graph's edges for future citations will increase the model's accuracy.

The goal for this project is to address this issue by adding a new edge in order to emphasize the connection between two papers that share a citation. In order to do this, we will add a new edge called "paper related_to paper". One thing we also noticed during the dataset analysis is that the graph is sparse, so adding this information could potentially help improve the model's accuracy. One primary benefit of this is that the new information is added according to knowledge already existing within the graph.

Potential Other Augmentation Methods

- Topology methods to check for common neighbors
    - Linking two related works: papers who cite a similar paper
- Implement sequential methods
    - If paper 1 cites paper 2, and paper 2 cites paper 3, then augment an edge for paper 1 citing paper 3
- Check to see how flipping the direction of the citation edges impacts the accuracy.
- Temporal Graph Learning Augmentation

I will meet weekly with Eric and Gang to discuss my progress and help plan for the future. I suggest bi-weekly written reports to Professor Jiang to share the progress that has been made on the project. The final goal is to have a deliverable that is a workshop paper or main conference paper at next year's KDD or ICML conferences.