Follow-up questions on assignment 5

I adopted a two-step process for this assignment. First, with the unsupervised version of the GraphSAGE algorithm I computed the node embeddings. And the second part is the supervised learning part where I am training a classifier for link prediction tasks.

I had following doubts while working on this assignment,

1. Size of the negative samples

We are given a graph in which the degree distribution of the nodes is such that the 75th percentile is two. Even if we assume that the given information of the graph is incomplete (i.e., there exists some links which we don’t know about), the 75th percentile being equal to 2 a low number. In other words, the adjacency matrix is sparsely populated. This is evident from the density of the graph which is equal to 0.04%. If we calculate ratio of non-existing links to existing link the value is 2290 i.e., for every existing link there exists 2290 possible but non-existent links.

In this case, link prediction task can also be looked at as a milder version of anomaly detection problem that correctly detects an even which occurs with frequency of 1/2290. Therefore, how to determine the correct size of the negative sample?

The negative sampling is used in two places, objective function of the unsupervised GraphSAGE algorithm and in creation of the training data for the classifier. Negative sampling sort of achieves different purposes in both of these tasks, how to determine size of the negative sample for each?

1. Where to travel in GraphSAGE?

In GraphSAGE we aggregate information from neighborhood at a distance. But as this is a directed graph, there exists two neighborhoods on either side of the node. Which neighborhood to consider in this case?

1. Drastic changes in the classification performance

Classification with text embeddings only and node embeddings only are very different from each other.