Group 34

Darren Banh 12462836  
Md Sazid Banna 12786442  
Zohaib Lodhi 12767354

# Methodology

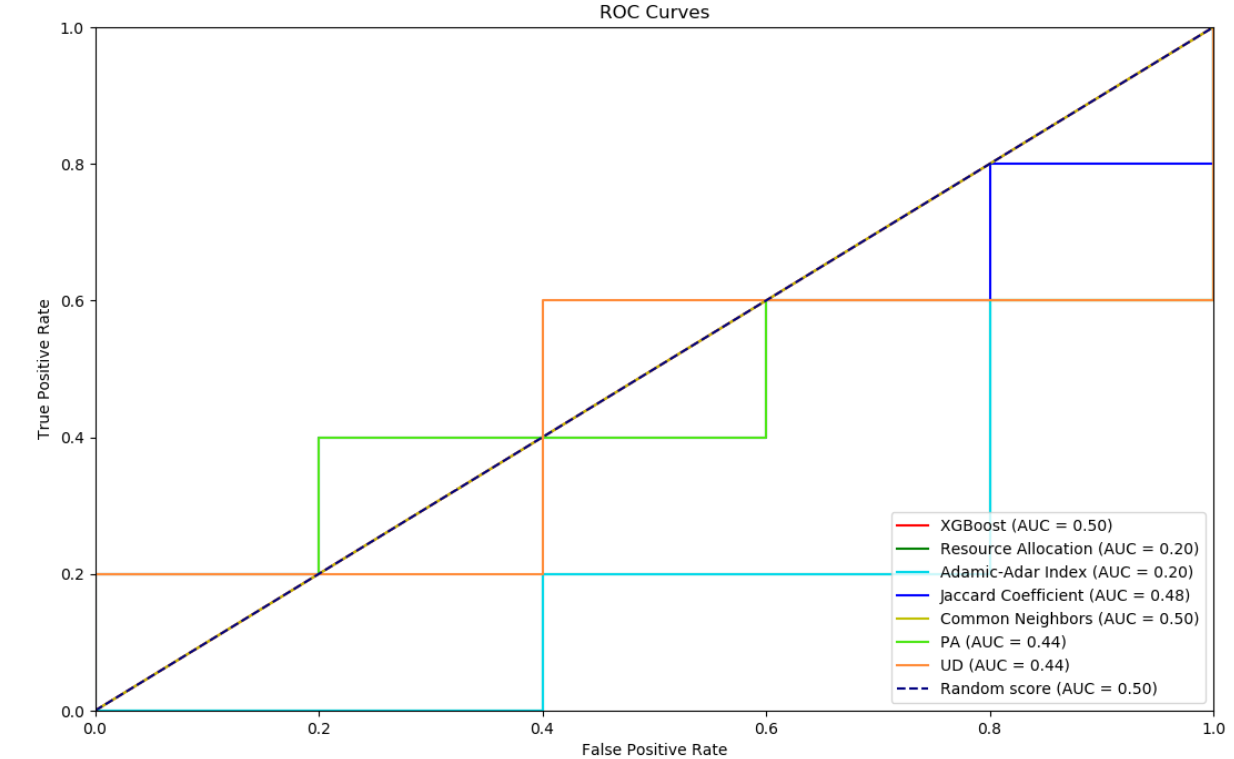
In order to work with the data, we used notepad++ to convert the .txt files into csv. For prediction, we have used networkit and networkx.

Networkit is an open-source toolkit for large-scale network analysis. Written in C++, Networkit provides faster processing compared to other available libraries. Because of networkit’s prebuilt algorithms and available examples, it was used.

We have used networkit’s implementation of following algorithms to compare the performance of link prediction.

* Resource Allocation
* Adamic-Adar
* Jaccard Coefficient
* Common Neighbors
* Adjusted Random
* Same Community
* Udegree index

The training data has 18,616 nodes and 178442 edges while the test data has 11,517 nodes and 19667 edges. After calculating, we got the ROC scores below.



We can see that XGBoost and Common Neighbor had the highest score of 0.50 which isn’t good at all. Due to segmentation fault and bug in the test environment, we had to use networkx and networkit interchangeably which might explain the low score and unusual graph.

But when comparing the features importance in XGBoost algorithm, we can see that degree of first given node (UD) and preferential attachment has the most importance despite the ROC score of the two being 0.44

