**Linear Regression:**

Average mean squared error across all Go-ids

"(0, 1) PPI" 0.001211887408350228

"(0, 1) SIMP" 0.0012126315661762492

"(2,) PPI" 0.0012127442848385173

"(2,) SIMP" 0.0012112071926545654

"(0, 2) PPI" 0.0012127048064502438

"(0, 2) SIMP" 0.001210476789533034

"(0, 1, 2) PPI" 0.0012118307398820312

"(0, 1, 2) SIMP" 0.0012110116447171836

"(1, 2) PPI" 0.0012132378607974665

"(1, 2) SIMP" 0.0012115913303623323

"(0,) PPI" 0.0012129864311300038

"(0,) SIMP" 0.0012120214216284723

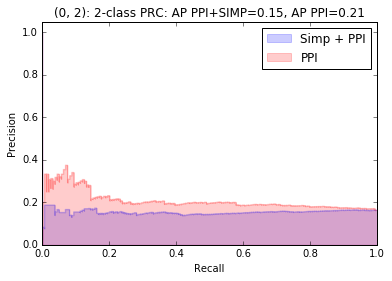
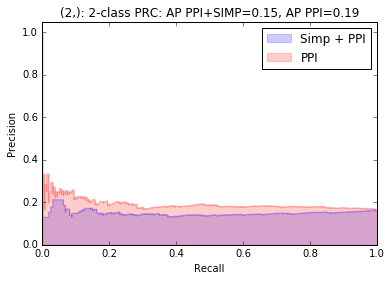
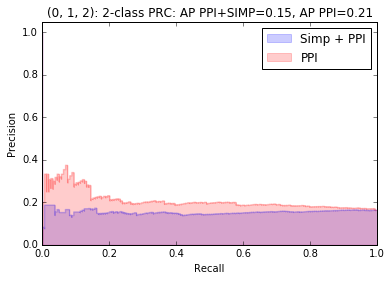
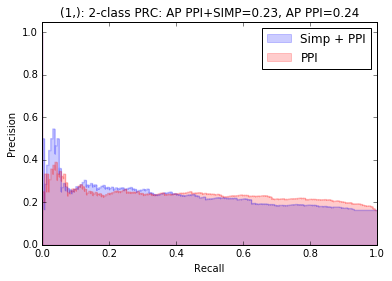
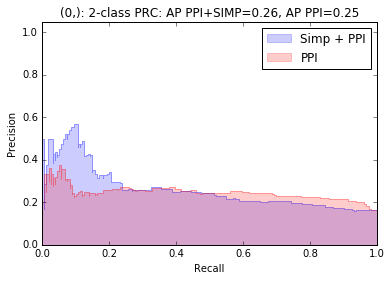
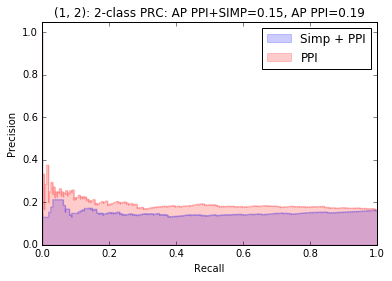
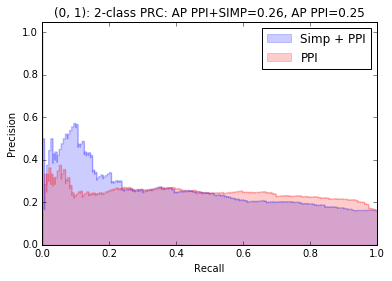
"(1,) PPI" 0.0012137853300388825

"(1,) SIMP" 0.0012131195409431218

As we can see, Regression is not ideal for this problem. All the models we generated using regression have a really low prediction capacity and really high error rate. Since the problem is a classification problem, regression models are expected to perform poorly.

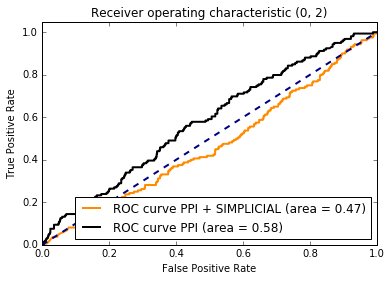
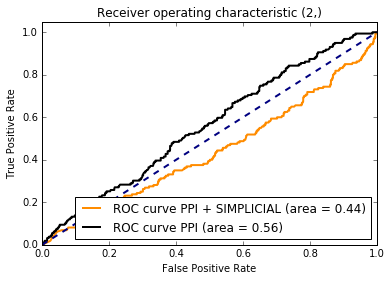
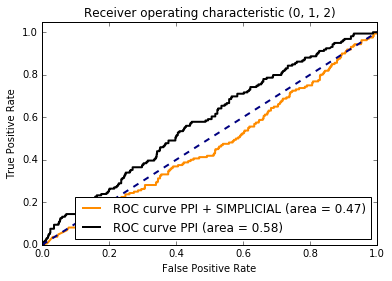
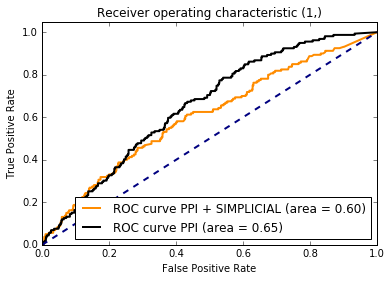
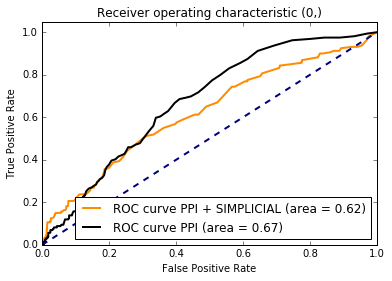
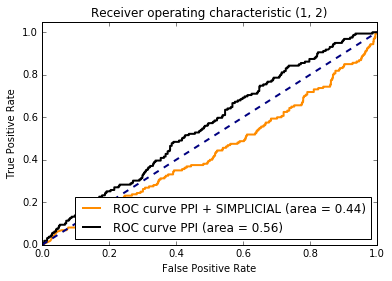
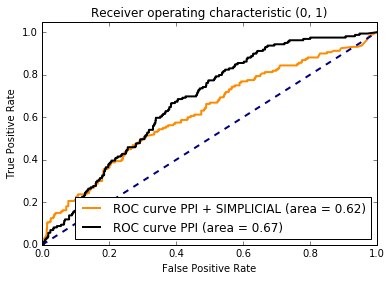
To train our model, we have taken all the combinations of centralities. In the above table, the tuples show what combinations we have used with 0 being betweenness centrality, 1 being closeness centrality and 2 being degree distribution.

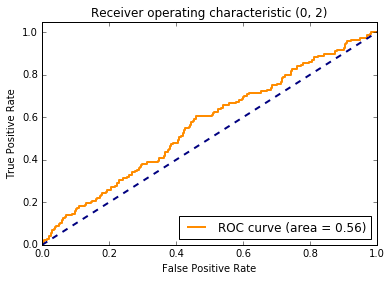
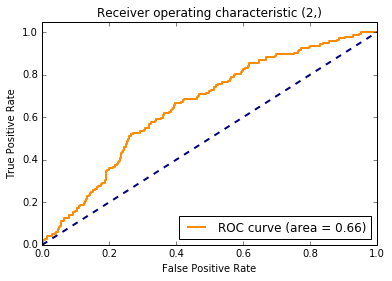
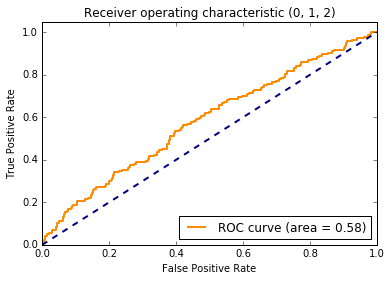
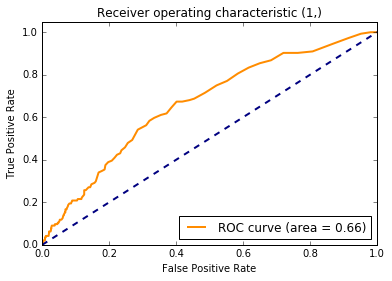
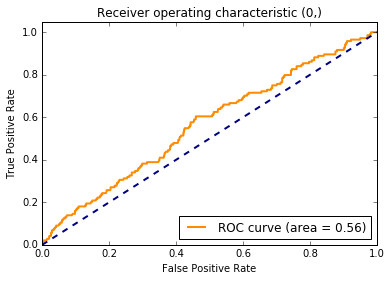
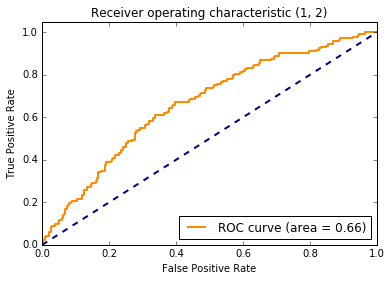
**Logistic Regression:**

****

What above figures show, is the difference Precision and Recall curves for different combination of input and how the resulted logistic regression model performed in both ppi and simplicial complexes network. As we can notice, our maximum average precision over all the combination of input is 25% in simplicial complexes and 22% in ppi networks. Something else we can observe is that our precision drops really suddenly once we increase the recall. Which means, only very few instances are right if we want to be precise. Precision drops to a constant value and stays around it as the recall increases. It is clear that prediction capacity of this model is quite bad. Some combinations seems to perform better under simplicial complexes and some combinations performs worse.

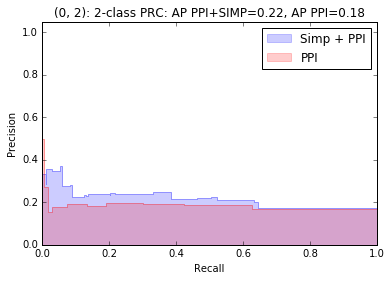
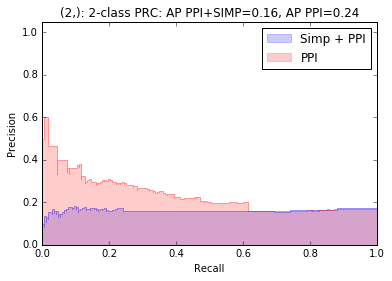
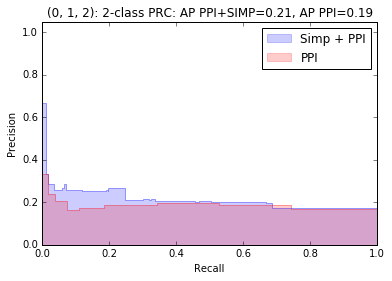
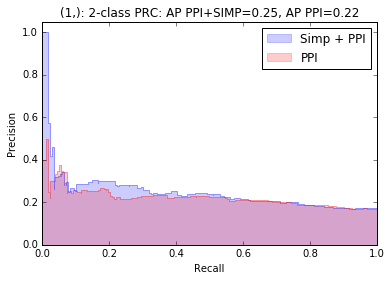
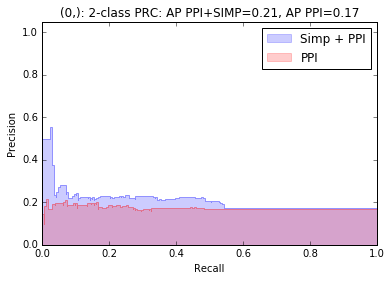
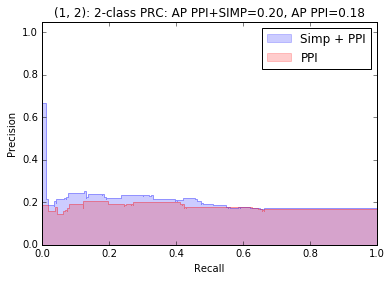
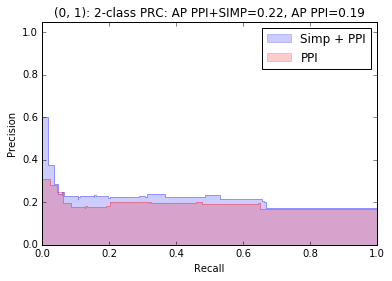
-- EXPAND ON ALL THESE POINTS--





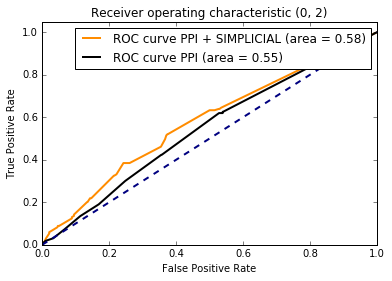
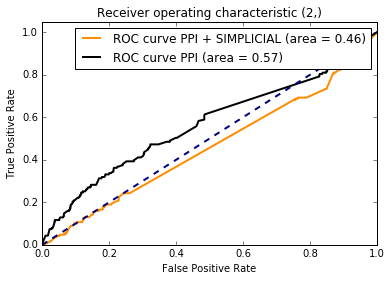
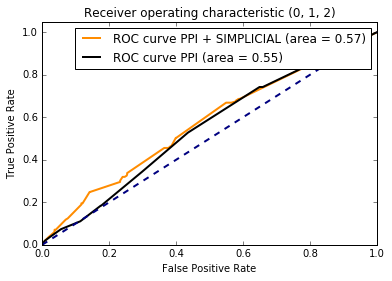
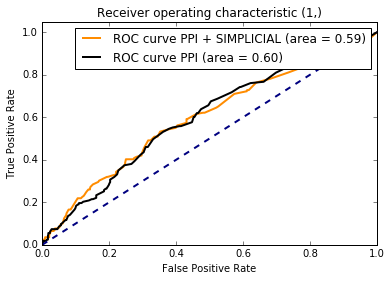
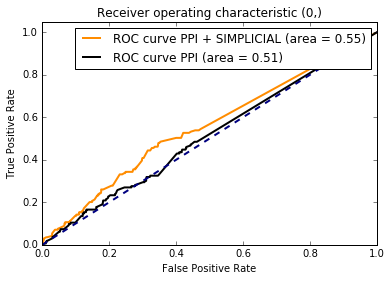
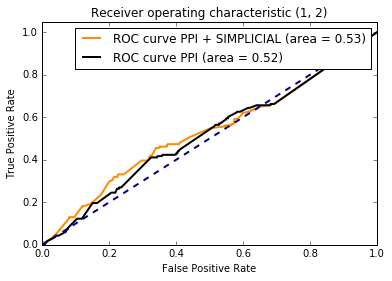
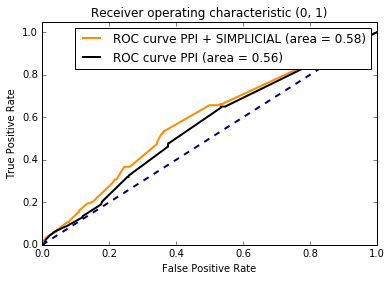
Above shown Receiver Operating Characteristic (ROC) are plotted based on the combination of models we generated from different combinations of input. The dotted line represents the True and False positive output we get from a random classifier when we move the threshold of the prediction from high to low. The Orange curve, ROC curve, represents the True and False positive rate we would get from the models as we move our threshold. Even though three out of 7 models perform bit better than a random classifier, these model’s prediction power is quite low and have a high False positive rate. As we can see, some of the combinations have better True and False positive rate considering the area under the curve which shows how good a model performs. (First, third and fifth models have the area value of 0.66 which is more than the other combinations)

**Random Forest Classification:**



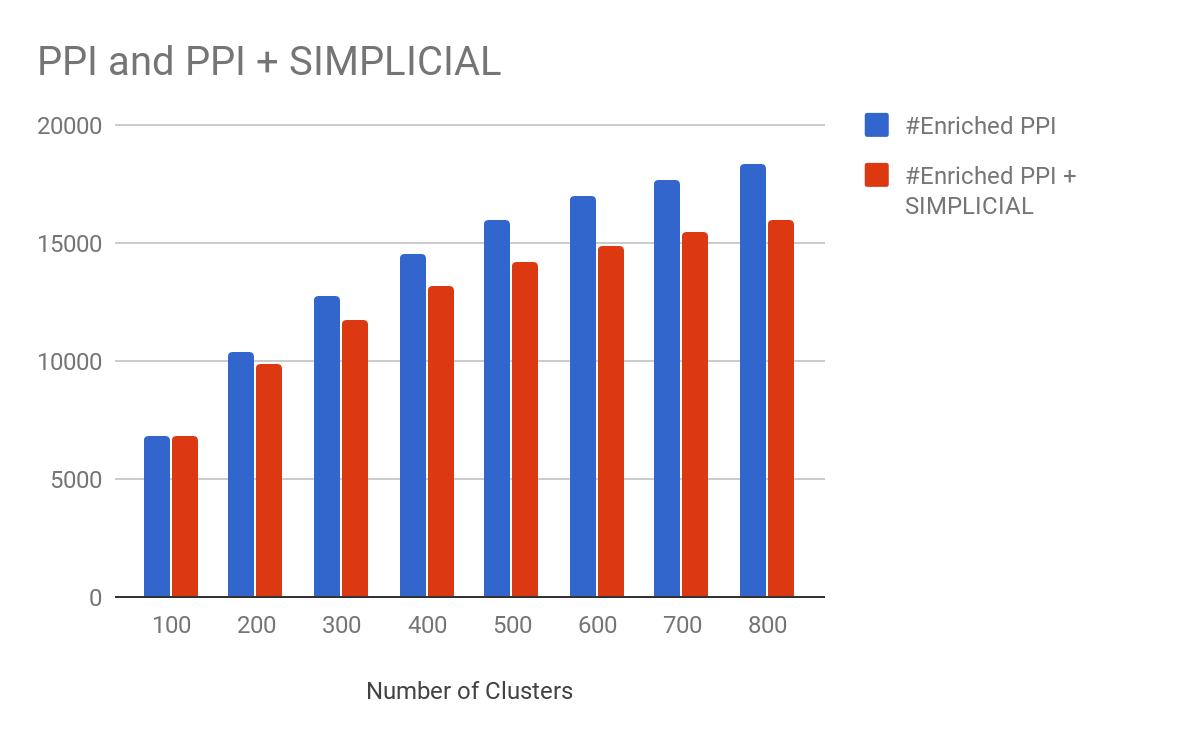
Above figures depict different Precision and Recall curves for different combination of centrality inputs and how Random Forest Classification model performed in both ppi and simplicial complexes networks. According to the figures, maximum average precision over all the combinations of centralities as inputs is less than 20% in ppi networks and simplicial complexes around 20%.

Also it is important to note that, some models seems to have a slower drop in precision in both Simplicial Complexes and ppi networks. We can see that simplicial complex models only perform better under certain combination of topological measures but the maximum average precision suggest that this enhancement in performance is not significant. For all the models, the precision barely increases as the recall increases and it’s almost constant after a certain recall in ppi networks.



ROC curves for Random Forest Classification shows the similar performance, if not worse compared to the one we got for Logistic regression. On average, random forest classification gives us worse performance and worse average area under the curve. Also we can note that simplicial complexes do not perform better in some combinations of inputs as measures.

# Agglomerative clustering:

****

We applied agglomerative clustering on the genes centrality data that we have collected on both ppi and simplicial complexes. We analysed each cluster to see if any of the GO terms are enriched in them. We calculated the p value for each GO term in each cluster for both ppi and simplicial complexes.(https://www.nature.com/articles/srep35098#supplementary-information) We repeated the experiment by changing the number of clusters from 100 - 800 incrementing by 100 every time. The above figure depicts the number of clusters we have in the experiment and the number of enriched GO terms in that experiment. As we can see clustering based on all three centralities (betweenness, closeness and degree distribution) in simplicial complexes, did not give us more enriched GO terms. The figure also shows that as we increase the number of clusters, the number of enriched GO terms increases in both PPI and simplicial complexes. For future, this clustering may be used for predicting GO terms in the clusters. Also we can use combinations of centralities for our clustering model in future to see if we get better enrichments in both simplicial complexes and ppi networks.