Assignment 2

COMP 596: Network Science

Due on October 4th 2020

1. Select 3 (or more) centrality measures and find the top 5 most important nodes in the Enron dataset. Who are the top ranked people? [30%] aggregate all emails sent at different times into a static snapshot with an edge weight showing how many emails in total have been send from one node to the other

2. Select 2 (or more) community detection or graph clustering algorithms which cluster a given graph, and compare their performance.

   (a) apply the algorithms to cluster the graphs [30%]
   Experiment on these three sets of data:
   real-classic: strike, karate, polblog, polbooks, football [10%]
   real-node-label: citeseer, cora, pubmed [10%]
   Here, the goal is to use the classification labels as clustering labels, and see how well we can find those labels without using the feature vectors.
   synthetic: LFR [10%]
   The common practice is to sample for varying values of μ which controls how well separated are the communities, i.e. generating synthetic graphs with $\mu = .1$ to $\mu = .9$, reporting average performance for 10 realizations at each difficulty level, see `https://arxiv.org/abs/0805.4770`, Fig 5 for example. N = 1000, or 5000 are common settings. For this experiments, you can use $\mu = .5$, n=1000, tau1 = 3, tau2 = 1.5, average degree=5, min community=20.

   (b) compare their performance [40%]
   report average performance, in terms of how much the results agrees with known clusters using both NMI and ARI measures, also report the modularity
   report average performance per each type of benchmarks: real-classic [10%], real-node-label[20%], synthetic[10%] and an overall average

bonus the best performing algorithm [10%]

Feel free to use any package or code off-the-shelf, or implement your own algorithm, measures. Submit the report in pdf and code as separate attachments, through Mycourses.