COMP 596 Assignment 2

Yifan Bai Shih-Chieh Fuh

Part 1

The three measures applied in Part 1 are:

- degree centrality
- eigenvalue centrality
- PageRank

The two tables below show the top 5 nodes by node number and by name, respectively.

Top5	1	2	3	4	5
Degree	63	7	117	19	54
Eigenvalue	7	63	62	60	117
Pagerank	19	117	54	63	7

Table 1: Top 5 nodes by node number

Top5	1	2	3	4	5
Degree	kenneth.lay	sally.beck	jeff.skilling	jeff.dasovich	tana.jones
Eigenvalue	sally.beck	kenneth.lay	john.lavorato	louise.kitchen	jeff.skilling
Pagerank	jeff.dasovich	jeff.skilling	tana.jones	kenneth.lay	sally.beck

Table 2: Top 5 nodes by name

Part 2

The two community detection algorithms applied can be found at:

Algorithm 1: Clauset-Newman-Moore greedy modularity maximization

https://networkx.github.io/documentation/stable/reference/algorithms/generated/networkx.algorithms.community.modularity_max.greedy_modularity_communities.html

Algorithm 2: Louvain algorithm

https://python-louvain.readthedocs.io/en/latest/api.html

The performance of the real-classic datasets and real-node-label datasets are shown in the tables below. MOD1, NMI1, ARI1 are modularity, NMI, and ARI for algorithm 1. Similarily, MOD2, NMI2, and ARI2 represent the respective measures for algorithm 2. The higher measure between the two algorithm is highlighted in yellow.

Overall, algorithm 2 performs better in terms of modularity and also NMI/ARI in real-classic datasets. However, algorithm 1 performs better with real-node-label datasets in terms of NMI/ARI, as well as certain datasets in the real-classic category. We have also noticed a low measure for NMI/ARI in real-node-label datasets. It seems that the number of communities clustered by the selected algorithms is far more than the number of communities provided in the true label, which perhaps could explain the low NMI/ARI to some degree.

	MOD1	NMI1	ARI1	MOD2	NMI2	ARI2
football	0.5785	0.5364	0.7624	0.6054	0.8069	0.8903
karate	0.3807	0.6803	0.6925	0.4188	0.4619	0.5866
polbooks	0.502	0.6379	0.5308	0.5268	0.5366	0.5219
polblogs	0.4248	0.5264	0.3771	0.4271	0.5208	0.3775
strike	0.5557	0.6647	0.7704	0.562	0.7978	0.8841
Avg	0.4883	0.6091	0.6266	0.508	0.6248	0.6521

Table 3: Statistics for real-classic datasets

	MOD1	NMI1	ARI1	MOD2	NMI2	ARI2
citeseer	0.9089	0.0805	0.351	0.9145	0.08	0.349
cora	0.8637	0.2028	0.4355	0.8678	0.192	0.4411
pubmed	0.7222	0.1718	0.2063	0.768	0.0741	0.1833
Avg	0.8316	0.1517	0.3309	0.8501	0.1153	0.3245

Table 4: Statistics for real-node-label datasets

For synthetic datasets, the average values of 10 simulations are used to plot the figures below, in a range of mu from 0.1 to 0.9. By comparing algorithm 1 and algorithm 2, it seems that algorithm 2 performs better in terms of modularity. For all three measures, we observe a decreasing trend as mu increases. Although only with a slight difference, it seems that overall, algorithm 2 performs better than algorithm 1 in synthetic datasets.

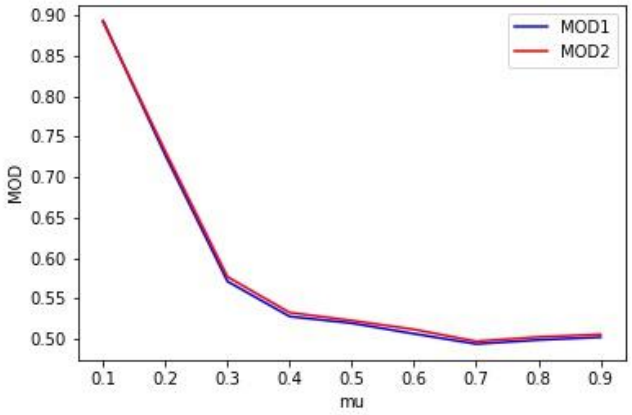

Fig. 1: Modularity in synthetic datasets

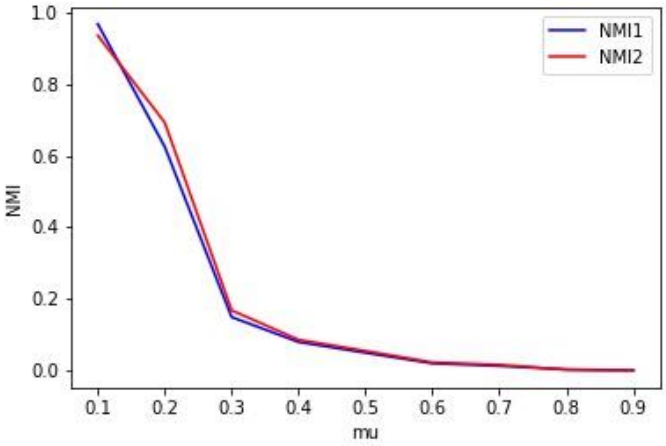

Fig. 2: NMI in synthetic datasets

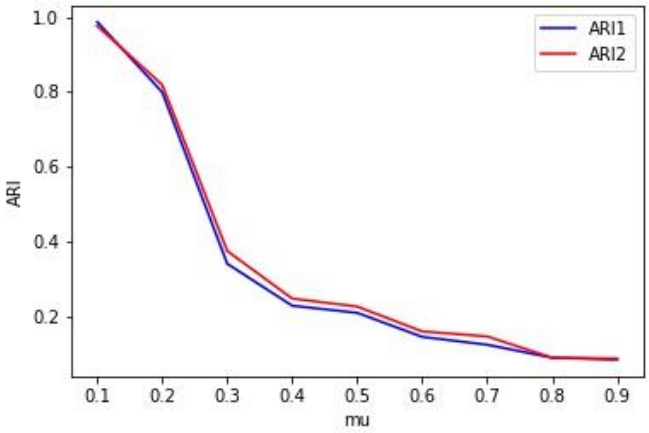

Fig. 3: ARI in synthetic datasets

We also report the average across the three distinct types of datasets. The average is defined as the average of the three average values from each type of datasets:

	MOD1	NMI1	ARI1	MOD2	NMI2	ARI2
Avg_real_classic	0.488329	0.609124	0.62664	0.508033	0.624814	0.65208
Avg_real_node_label	0.831603	0.151716	0.330914	0.850106	0.115336	0.324481
Avg_synthetic	0.582109	0.210866	0.334701	0.585704	0.218746	0.34787
Avg_total	0.634014	0.323902	0.430752	0.647948	0.319632	0.441477

Table 5: Overall performance