Assignment 2

COMP 599: Network Science

Due on October 4th 2022

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? [20%]
   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. Graph Clustering [50%]

   Select two (or more) community detection or graph clustering algorithms, apply them on the following real world datasets and evaluate their performances:
   – real-classic: strike, karate, polblog, polbooks, football
   – real-node-label: citeseer, cora, pubmed
   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.

   (a) Algorithmic Complexity: Derive and report the complexity of the chosen algorithms. [10%]

   (b) Qualitative Evaluation: Visualize the obtained clusters using Gephi tool or any other graph visualization tool of your choice. Report the visualizations and comment your observations. [20%]

   (c) Quantitative Evaluation: Evaluate the quality of the clusters using label independent (topology only) metrics for both sets of graphs and label dependent metrics for graphs with labels. [20%]
   topology based metrics: Modularity and Conductance
   label dependent metrics: NMI and ARI.

3. Evaluation using synthetic datasets: [30%]

   (a) Create a set of synthetic dataset using 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) Qualitatively evaluate the chosen algorithms in the previous questions on this synthetic datasets and report your results [10%]

   (c) Compute the Average Modularity and Conductance measures for each of the following sets of datasets (i) Real-Classic, (ii) Real-Node-Label and (iii) Synthetic datasets, compare them and report your observations

4. **[Bonus:]** Compare the results with an algorithm published/proposed in the last 4 years and report your observations [10%]

5. [**Bonus:**] Compare the results on 3 more real world datasets not included in the assignment and report your observations[10%]

6. [**Bonus:**] Compare the ranking of algorithm observed on LFR benchmarks with similar size graphs generated by FARZ generators and report your observations[10%]

Feel free to use any package or code off-the-shelf for the community detection algorithms or centrality measures, or implement the algorithm/measures on your own, you can also implement your own graph clustering algorithm.
Submit the report in pdf and code as separate attachments, through Mycourses.