Assignment 1: Understanding Complex network through different Network Models and Centrality Measures

Re-submit Assignment

Due 5 Sep by 1:00 **Points** 0 **Submitting** a file upload
Available 20 Aug at 0:00 - 5 Sep at 11:29 16 days

Submission is individual.

Theme A: Network Model (Weightage: 30 Points)

You are given following real world network data

1. IMDB Movie Dataset: **https://github.com/zyz282994112/GraphInception/tree/master/data (https://github.com/zyz282994112/GraphInception/tree/master/data)**
2. SLAP Gene Dataset: **https://github.com/zyz282994112/GraphInception/tree/master/data (https://github.com/zyz282994112/GraphInception/tree/master/data)**
3. Foursquare Restaurant Review Dataset: **https://sites.google.com/site/yangdingqi/home/foursquare-dataset (https://sites.google.com/site/yangdingqi/home/foursquare-dataset)**
4. DBLP Co-authorship network: (https://snap.stanford.edu/data/com-DBLP.html)
5. Higgs Twitter Dataset: (**https://snap.stanford.edu/data/higgs-twitter.html (https://snap.stanford.edu/data/higgs-twitter.html) (Links to an external site.) (https://snap.stanford.edu/data/higgs-twitter.html)**)

** Notes:

- Please refer to Graph Inception paper (Deep Collective Classification in Heterogeneous Information Networks, Section 7) for IMDB Movie and SLAP Gene datasets schema details.
- From Foursqaure dataset webpage, use the NYC Restaurant Rich Dataset.
- For Higgs Twitter Dataset, use the Mention Network dataset.

For all these datasets, study the following.

1. Degree distribution
2. Clustering Coefficient
3. Average path length
4. Giant Cluster component (size in terms of number of node and diameter)

Create a random graph G(n, p) corresponding to each of the network dataset, where n is the number of nodes and p is the probability of forming a link between two nodes. Keep the number of node same with that of the real world network data, and generate edges with different values of p. Study the above graph properties against different values of p and compare with that of the real world network.

Theme B: Network Centrality measures (Weightage: 30 Points)

For the above network datasets (real work networks) and estimate the following centrality.

1. Degree
2. Closeness
3. Betweenness
4. Eigenvector
5. Katz
6. PageRank
7. HITS

Plot the centrality distribution for each network and analyse distribution across the networks.

Submission: Write a detailed but brief report of all the above investigations. You are encouraged to include any information such as observations, inferences, correlation etc. with respect to the above network properties which will enhance the quality of your submission.

You will need to submit the report and program code before the due date.

Implementation: You are free to use any of the publicly available network library (such as Networkx) and visualization tools (such as Gephi) to implement and analyse above estimates. However, you are also encouraged to implement the above estimates of your own (using any language or script of your choice) and compare the output from your implementation with that of the library to check the correctness. You can earn an additional **points upto 40** by implementing the estimates of your own.

Submission instruction will be provided on the canvas.