CSC 407/607 Network Analysis
Fall 2022

Graduate Project

This is the project for graduate students to work on the latest Yelp dataset.

Access Yelp Dataset from https://www.yelp.com/dataset . Note that you will be asked to sign a dataset license form. The compressed file is about 4GB in size. You will be focusing on the <reviews.json> file in the dataset.

This project is mostly about businesses with at least R reviews in two different metropolitan areas. Use the following approach:

a. Identify the total number of reviews for each business in one metropolitan area.

b. Keep only those businesses with at least R reviews and remove all others.

c. Among the remaining businesses, when two businesses have been reviewed by the same reviewer, connect them.

d. Draw this graph and compute the degree centrality of all vertices. Compute the average degree centrality of all nodes.

e. Repeat for a second metropolitan area and compute the new average degree centrality of all nodes in the new metropolitan area.

f. Do you see any difference in these two values? Explain.

Notes:

Note 1. Your value of R should be reasonable so that the graph can be shown. And I expect everyone working independently to pick unique numbers, for example, 95 (and please do not pick this number).

Note 2. Feel free to use any programming language that you are good at. If you need a Linux server to help you with computation, try to explore ways to login to the university Linux server: linux.uncg.edu

Note 3. Your project report should include:

i. Sample list of businesses that satisfy the requirement.

ii. Sample list of businesses with common reviewers (show both businesses and the common reviewer)

iii. Graph in part d) and e).

iv. Sample degree centrality of businesses.

v. Answer to part f).

vi. Source codes, and full results in files.