CAP 6315 Project Part #1 – Social Network EDA

This term-long project will guide you through the full big data pipeline using one real-world large dataset. You will work with the same dataset both project parts.

You may work individually or in pairs. If working in pairs, you must submit one submission for every project part that includes both students' first and last names and Z-numbers on the first page.

Use of AI tools (ChatGPT, Copilot, Gemini, etc.) is not allowed for any part of this project. Projects that were generated by AI will receive 0 points, will result in potential F in the course, and will be reported for academic dishonesty.

Project Part #1 Learning Objectives:

- Students will be able to use NetworkX library to create and visualize social networks, as well as extract meaningful metrics from it.
- Students will learn to interpret the results of social network analysis by using network properties such as network diameter, edge density, and clustering coefficient.
- Perform exploratory data analysis (EDA) on node-level attributes using descriptive statistics and visualizations.
- Communicate insights from social network and attribute-based analysis using clear visualizations and written interpretations.
- Compute and interpret multiple centrality and influence measures in social networks.
- Apply and compare community detection algorithms.

Submission Instructions

- ***Your submission should include the complete Jupyter Notebook (or Google Colab) following the steps outlined below***. Be sure to include everything! It's helpful if you add the steps as comments in your code, followed by the code block or markdown cells that represent the answers.
- ***Leave the code outputs in the notebook***. Codes without visible outputs will lose 0.25 points for each missing output.
- If working in pairs, one submission is fine.
- ***Please export your code in .html format.***
- ***Note: You must submit your code in .html format. If not, 10% of the grade will be deducted.***
  Your name and Z-number must appear at the top of the code. If they are missing, 10% of the project part #1 grade will be deducted. Include your name and Z-number on top of your code, even if you are working alone.
- ***Note that everything should be done using PySpark and NetworkX. Using pandas to work with data is not allowed. 20% of the grade will be deducted If PySpark is not used.***

Dataset

Dataset: Twitch social network **https://snap.stanford.edu/data/twitch-social-networks.html** - you can use the data from any language.

Nodes: streamers

Node attributes: days, views, mature, partner

Edges: follower/friend connections

Steps

Do not forget to add your names and Z-number(s) on top of the code.

- ***Using NetworkX library and PySpark, follow the steps below:***
  1. Load the data using PySpark and build a graph using networkX library. (0.5 points)
  2. Visualize the graph. (0.5 points)
  3. Calculate the number of nodes in the graph. (0.5 points)

4. Calculate the number of edges in the graph. (0.5 points)
5. Calculate the network diameter. (0.5 points)
6. Add a markdown cell. In your own words, interpret the meaning of the obtained value for network diameter. (0.5 points)
7. Plot a histogram representing the degree distribution. (0.5 points)
8. Add a markdown cell and discuss what you observe by looking at the generated degree distribution. (0.5 points)
9. Calculate the average degree of a node. (0.5 points)
10. Add a markdown cell. In your own words, explain what the obtained average degree of a node means in the context of this dataset? (0.5 points)
11. Calculate the average shortest path length between any two nodes. (0.5 points)
12. Add a markdown cell. In your own words, explain what the obtained value of the average shortest path length means in the context of this dataset. (0.5 points)
13. Plot the distribution of the shortest path length. (0.5 points)
14. Add a markdown cell and explain what you observe in this histogram. (0.5 points)
15. Calculate the edge density of this graph. (0.5 points)
16. Add a markdown cell and answer the following question. What can we conclude about this graph considering its edge density? (0.5 points)
17. Load the node attribute file and display the first few rows of the dataset. (0.5 points)
18. Plot a histogram of the *views* (0.5 points)
19. Add a markdown cell and describe the distribution of Is it symmetric, skewed, or does it contain outliers? (0.5 points)
20. Create a bar plot showing the number of streamers who are partners vs non-partners. (0.5 points)
21. Add a markdown cell and interpret the proportion of partners in this dataset. (0.5 points)
22. Create a bar plot showing the number of streamers with mature content enabled vs not enabled. (0.5 points)
23. Add a markdown cell and interpret the results. (0.5 points)
24. Create a boxplot comparing views by partner status. (0.5 points)
25. Add a markdown cell and discuss whether partner streamers appear to differ from non-partners in terms of views. (0.5 points)
26. Compute degree, betweenness, closeness, Eigenvector, and PageRank centrality for all nodes. (2.5 points)
27. Report the top 10 node ids for each centrality (2.5 points)
28. Add a markdown cell answering the following question: What do you observe when comparing the top 10 nodes across all centrality measures and PageRank? (0.5 points). You may discuss: if some nodes appear across multiple measures, and which nodes appear only for certain measures.
29. Create boxplots comparing centralities and PageRank values between partner streamers and non-partner streamers (5 × 0.5 = 2.5 points)
30. Add a markdown cell and explain what you observe under each plot (5 × 0.5 = 2.5 points). Hint: Are partners more central/influential?
31. Apply the Louvain community detection algorithm. Report the number of obtained communities. (0.5 points)
32. Compute the average views and days per community (1 point).
33. Add a markdown cell and discuss what you observe (0.5 point).

Total points: 25