DATA620
Miraj Patel

Project 1 Proposal

Data Source Identification:

For this project, I have selected the GitHub Social Network dataset, available via the Stanford Large Network Dataset Collection (SNAP). This dataset represents a large-scale web graph of developers and their professional "follow" relationships.

- Nodes: Developers on the GitHub platform.
- Edges: Directed "follow" relationships between developers.
- Categorical Variable: The dataset includes a node attribute called “ml_target”. This is a binary categorical variable where:
  - 0 represents a Web Developer.
  - 1 represents a Machine Learning (ML) Developer.
- Web Source: The data will be accessed and downloaded directly from the SNAP Repository.

Data Loading Plan:

To ensure the analysis is computationally efficient and produces a clear visualization, I will implement a subsampling strategy.

- Data Extraction: I will use the Python requests library to programmatically fetch the node features (categorical data) and the edge list (the graph structure) from the web source.
- Subsampling: To work with a manageable sample size, I will filter the dataset to include a subset of approximately 1,500 to 2,000 (number not set) nodes.
- Graph Construction: I will use the pandas library to merge the categorical labels with node IDs and then load the resulting structure into a “networkx.Graph()” object.

Centrality Analysis & Hypothetical Outcome:

The project will focus on comparing Degree Centrality and Eigenvector Centrality across the two categorical groups (Web vs. ML).

I hypothesize that while Web Developers may have a higher average Degree Centrality (more total followers), Machine Learning Developers will exhibit significantly higher Eigenvector Centrality.