Miguel Gomez
09/20/2025
Data 620

The network data I selected is Facebook Large Page-Page Network from SNAP.

To analyze the Facebook Large Page–Page Network, I would begin by downloading the dataset from SNAP which is available in a zip folder. It contains three files that can be combined for analysis:

- **musae_facebook_edges.csv**, which contains the list of undirected edges representing mutual "likes" between Facebook pages
- **musae_facebook_features.json**, which holds descriptive attributes of the pages (such as text-based features extracted from the profile descriptions)
- **musae_facebook_target.csv**, which assigns each page to one of four categorical groups: politicians, governmental organizations, television shows, or companies.

To prepare the data, I would first load the edge list into a graph using pandas, JSON and NetworkX in Python. Then I would merge in the node labels from the target file and attach any additional descriptive attributes from the features file if needed. Once the network is set up, I would figure out the degree centrality for each node to measure the number of direct ties each page has. The results could then be grouped by the categorical label, allowing comparisons of centrality across categories.

Some of the steps I would need to write code for in Python to achieve the following is:

- Loading edge list
- Loading node Targets
- Loading node features
- Converting features as a dict with node ids as keys
- Computing degree centrality
- Adding centrality scores to graph
- Creating a dataframe for analysis
- Generating a summary by category

A hypothetical outcome might show that company pages generally have higher degree centrality, showing broad cross-category visibility, while political pages display lower average centrality but with a few highly connected outliers. This pattern would suggest that companies are positioned as hubs of connectivity in this network, while politicians rely on concentrated influence within smaller communities.