Degree Distribution of Male Professions

Overview

As a woman in data science, I understand how it feels to be a minority in a male dominated field. However, how often is the other side of the story shared? This idea mainly sparked up through conversations with my friends who are psychology majors, and both my male and female friends made a note on how their lectures mainly consist of women. And although there is a more even spread at a highly competitive and private institution like Boston University, there is a parallel where there are more female teachers than male. A career field where knowledge on numbers and coding, skills that have a difficult connotation is associated with men, where career fields focused on children and emotions are associated with women. This project studies how gender representation varies across different job categories and if gender stereotypes are statistically proven.

Data set

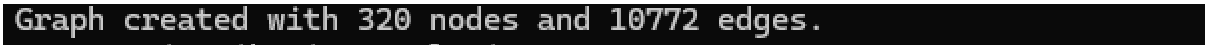
https://github.com/fivethirtyeight/data/tree/master/male-flight-attendants

Explanation

To delve into degree distribution, I used nodes and edges to represent the connections between the categories (or professions) that have similar male pecentages. Each node contains a tuple. Graphs are particularly effective for modeling complex networks, such as social networks or job categories, where relationships are key to understanding the data's structure. To create the graph, I used the ‘create_graph’ function. To compute the degree of each node and the number of nodes reachable with two hops, I used the functions ‘calculate_degrees’ and ‘calculate-two-hop-neighbors’. Analyzing degrees provides insights into how interconnected different job categories are, while two-hop neighbor analysis reveals indirect relationships that may not be immediately obvious.

The ‘analyze_distribution’ function calculates mean, standard deviation, and performs power-law fitting using parameters such as α (alpha) and x_min. It also conducts a Kolmogorov-Smirnov test to evaluate fit. To define the Kolmogoroz-Smirnov test, it is a statistical method used to compare a sample distribution to a reference distribution or to compare two sample distributions. These parameters help validate assumptions about gender representation patterns across job categories. Lastly, to visualize the data, I used the ‘plot_distribution’ function to physically show the degree distributions.

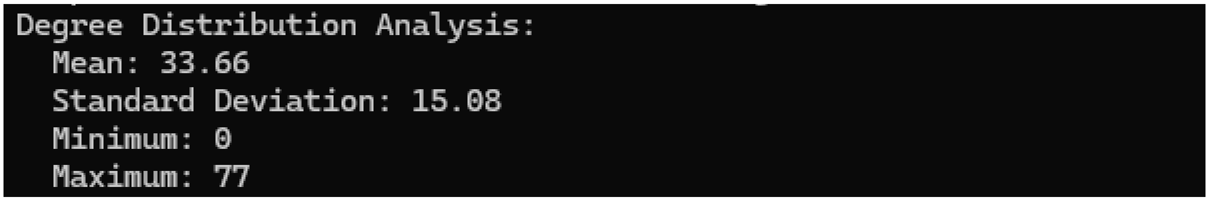
Output



```
Graph created with 320 nodes and 10772 edges.
```

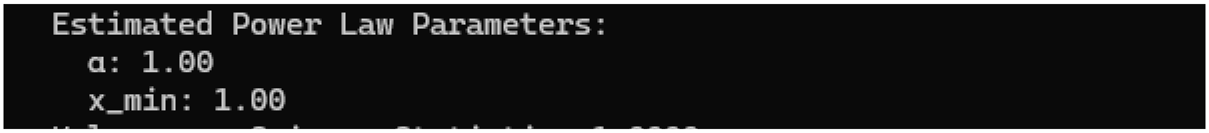
Each node represents a job category. There are 320 job categoriesin the dataset.Each edge represents a connection between two job categories based on their male percentage similarity. There are 10,772 connections formed between these categories, indicating a dense network where many job categories are closely related in terms of gender representation.

Degree Distribution Analysis:



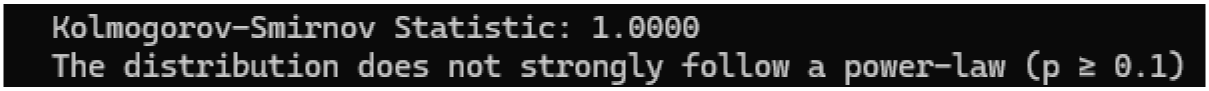
```
Degree Distribution Analysis:
  Mean: 33.66
  Standard Deviation: 15.08
  Minimum: 0
  Maximum: 77
```

The mean being 33.66 indicates the average number of connections (edges) per job category. On average, each job category is connected to about 34 other categories. A standard deviation of 15.08 suggests that while most job categories have around 34 connections, some have significantly more or fewer connections. At least one job category has no connections, which is told by the minimum of 0, indicating that it is isolated from others in terms of gender representation. The most connected job category has 77 edges, meaning it is connected to 77 other job categories.



```
  Estimated Power Law Parameters:
    α: 1.00
    x_min: 1.00
```

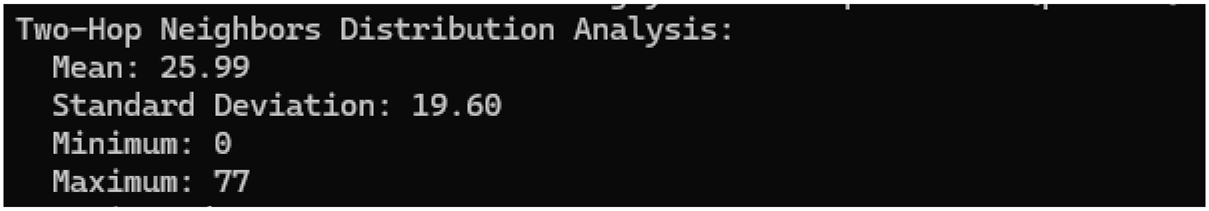
An α value of 1 suggests a relatively flat distribution, which is characteristic of many real-world networks but may not strongly imply a power-law behavior. ‘X_min’ represents the minimum value in the dataset considered for fitting the power-law model. Here, it is set to 1, which means that only degrees greater than or equal to one were used in the analysis.



```
  Kolmogorov-Smirnov Statistic: 1.0000
  The distribution does not strongly follow a power-law (p ≥ 0.1)
```

A K-S statistic of 1 indicates a poor fit; specifically, it suggests that the observed data deviates significantly from what would be expected under a power-law model. The conclusion that "the distribution does not strongly follow a power-law" indicates that there is insufficient evidence to claim that the degree distribution follows a power-law.

Two-Hop Neighbors Distribution Analysis



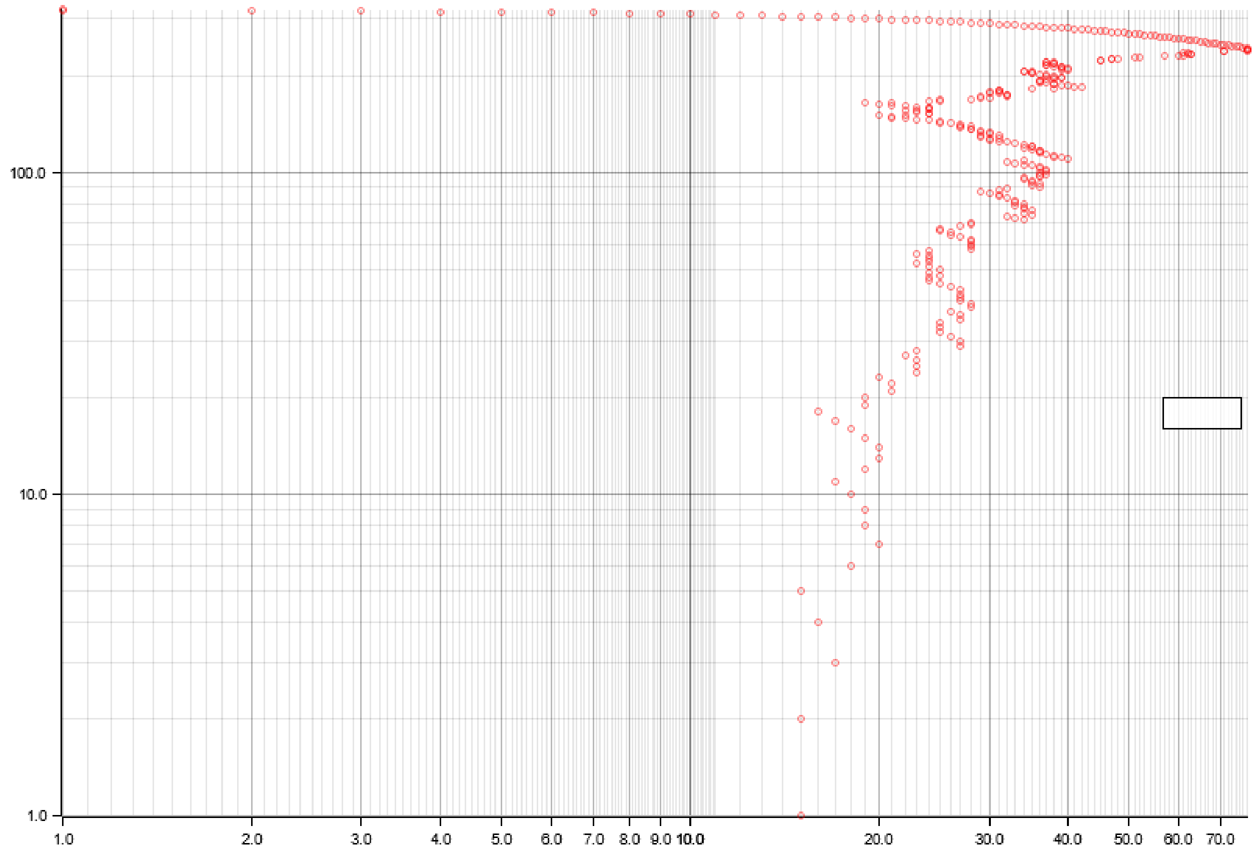
```
Two-Hop Neighbors Distribution Analysis:
  Mean: 25.99
  Standard Deviation: 19.60
  Minimum: 0
  Maximum: 77
```

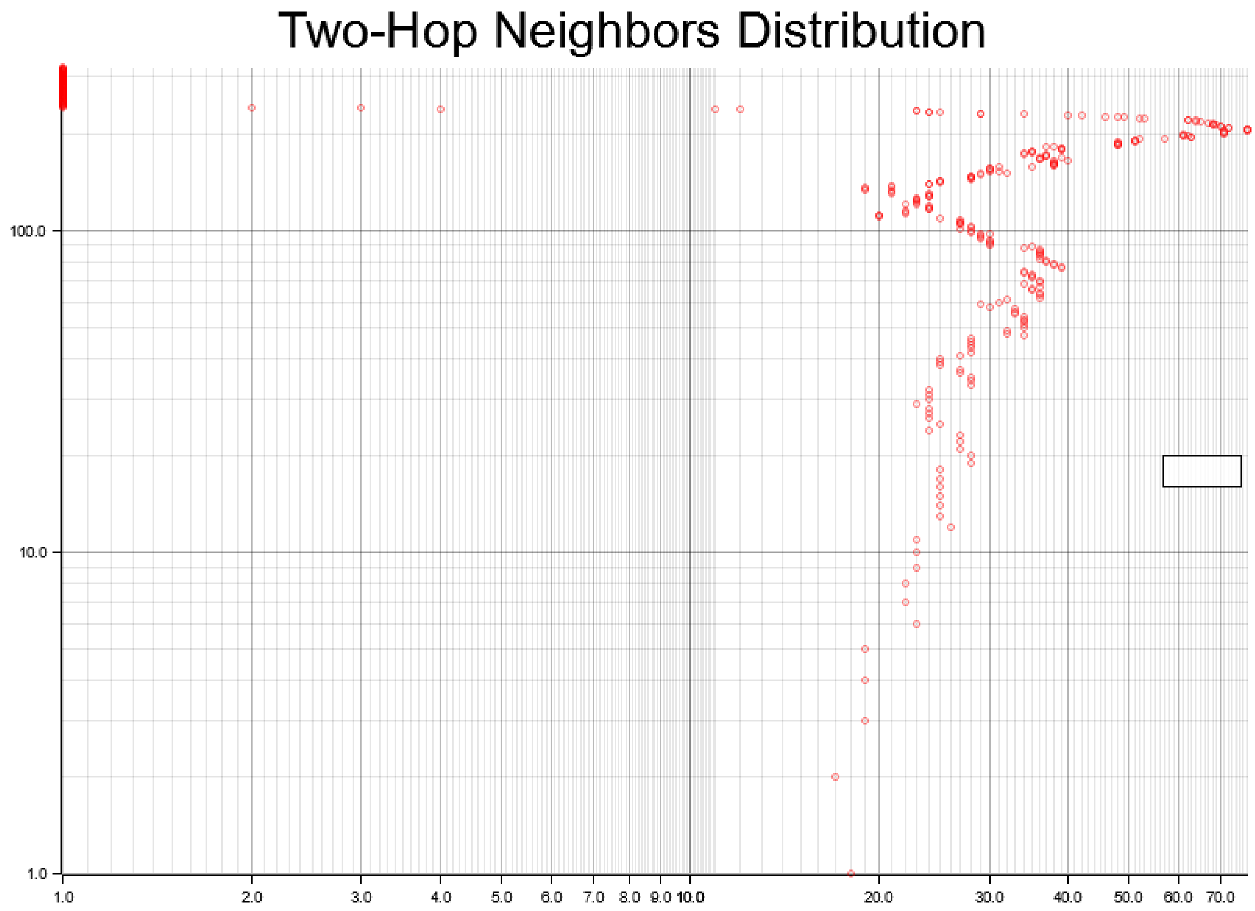
Similar to the degree distribution analysis, this section analyzes how many job categories can be reached by traversing exactly two connections (edges). A mean of 25.99 indicates that on average, each job category can reach about 26 other categories through two hops. A standard deviation of 19.60 shows variability in this reachability; some categories can reach many others while others can reach few. Again, the minimum and maximum values indicate that at least one job category cannot reach any other category within two hops, while another can reach up to 77 categories.

The Estimated Power Law Paramters and Kolmogrov-Smirnov Statistics remain the same

The code also creates graphs which can be seen below:

Degree Distribution





.

Conclusion

Drawing the data back to the categorical data, the dataset concludes that there is persistent gender segregation in many occupations, with significant underrepresentation of males in care-oriented and administrative roles, and underrepresentation of females in technical and physical labor-intensive jobs. This opens up the conversation about if we as a society will see a change in this, or if gender roles are a fixed aspect of society.