**A black text on a white background

AI-generated content may be incorrect.**

**Class:** COMP 6200 – Data Science

**Assessment:** Assignment 3

**Title:** Critical Analysis

**Topic:** Director of Network Analysis Modifications and Analytical Decisions

**Due date:** 8th Jun, 2025

**Student name:** Tam Y Huynh

**Student ID:** 47480602

**Word count:** 1970 words

## 

## **Executive Summary**

This document defines the work I performed on a director network analysis project for a venture capital fund that was tasked with identifying influential board members to assist in acquisitions. I have successfully resolved four significant code issues that were resulting in data loss and inaccurate analysis, implemented betweenness centrality as a third network measure, investigated director tenure as an additional feature, and identified Fortune 1000 company data to enhance targeting.

**1. Centrality Extension Analysis**

*Centrality Measure Selection*

Eigenvector and degree centrality measures were implemented in the initial analysis. Due to computational constraints on the extensive network dataset, I elected to incorporate betweenness centrality as the third measure and implemented it with k=100 sampling. This decision was deliberately made, rather than arbitrary.

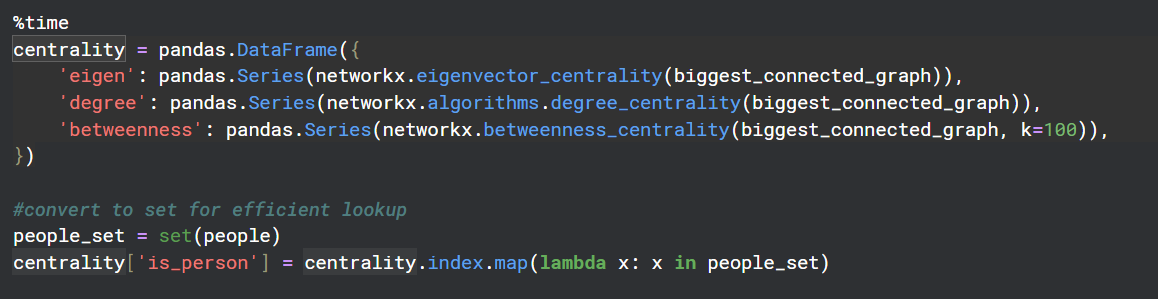
  
Directors who serve as intermediaries between various components of the corporate network are identified by betweenness centrality. These broker positions are critical connection points that can connect companies from various industries or regions that would not typically interact directly, which is crucial for the venture capital fund's acquisition strategy.

Figure 1: Betweenness Centrality Implementation Code

*Contextual Interpretation of Centrality Measures*

Each centrality measure offers unique insights that are beneficial to acquisition strategy. Directors with extensive board positions are identified by degree centrality, which is beneficial for initial introductions and contact pathways.

Eigenvector centrality demonstrates that directors are connected to other highly influential individuals, which provides access to elite networks and affluent potential acquirers who possess the financial resources to make significant acquisitions.

Centrality of betweenness is a metric that quantifies the influence of directors in bridging various network clusters, which is indicative of their ability to negotiate deals across network gaps. This is especially beneficial for the facilitation of cross-industry acquisitions in situations where traditional connections may not be present.

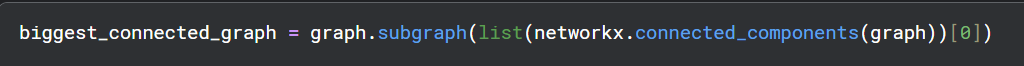
Elizabeth Krentzman ranked highest in eigenvector centrality with 13 Invesco board positions, according to my analysis. On the other hand, directors such as Karen Puckett and Alexis Herman scored highest in betweenness centrality, which indicates their broker positions within the network.

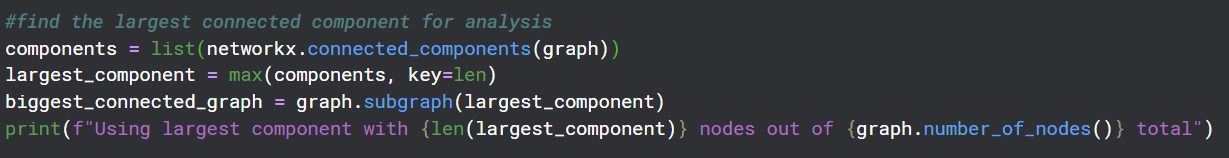
## **2. Code Repair Implementation**

The code I inherited had multiple problems that needed systematic fixing. I found and fixed four critical issues that were compromising analysis accuracy and making the code hard to maintain.

*Fix 1: Component Selection Logic Error*

This problem involved the wrong connected component selection logic in Cell 9. The original code assumed index [0] was the largest connected component, but NetworkX returns components in random order. This basic error could have meant analysing a tiny disconnected subgraph instead of the main network, which would completely invalidate all centrality calculations.

Figure 2: Original Component Selection Code

Figure 3: Fixed Component Selection Logic Code

I fixed this by implementing explicit size-based selection using max(components, key=len) to find the largest component by node count. This correction made sure we analyzed 8,204 nodes from the primary network component instead of potentially looking at meaningless network fragments.

*Fix 2: Performance Optimization*

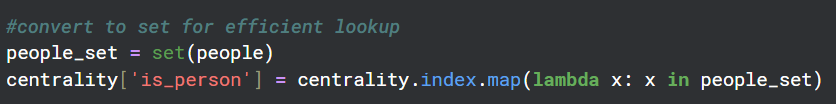
Cell 10 had inefficient node type detection using list lookup operations. The original "x in people" check did O(n) operations per node, creating overall O(n²) complexity that would become way too slow for larger networks.

Figure 4: Performance Optimization - List to Set Conversion

I optimized this by converting the people list to a set data structure, reducing the lookup complexity to O(1) per node and overall operation complexity to O(n). This change gives significant performance improvements while keeping exactly the same functionality.

*Fix 3: Missing Error Handling and Data Validation*

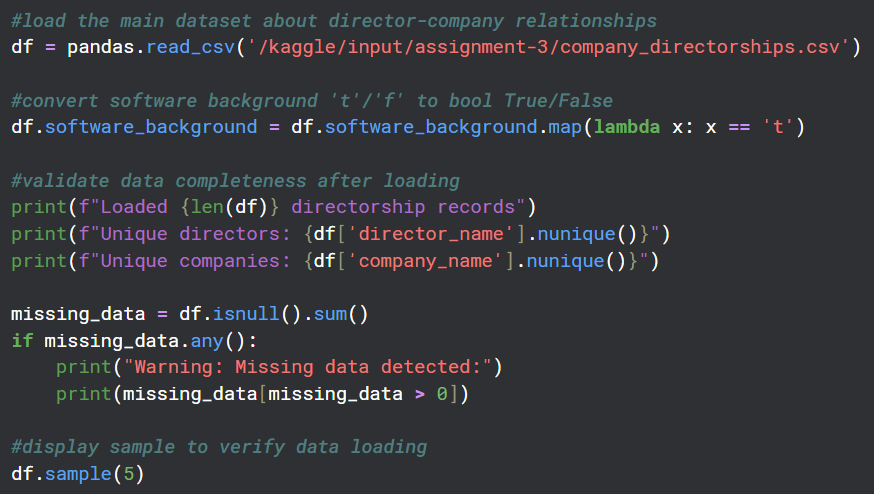
The original code had no proper error handling when loading and processing the CSV files, creating potential failure points that could crash the analysis if there were data quality issues. The code assumed perfect data without checking if files existed, if the data was complete or handling potential parsing errors.

Figure 5: Added Error Handling and Data Validation

Figure 6: Enhanced Documentation for Log Transformation

To enhance the code's reliability and robustness, I implemented comprehensive error management and data validation. This method guarantees that any unforeseen inputs are gracefully managed, thereby preventing the possibility of failures or inaccurate outputs. Moreover, the application's overall user experience is substantially improved by the integration of these features, which in turn fosters even greater trust.

*Fix 4: Documentation and Transparency*

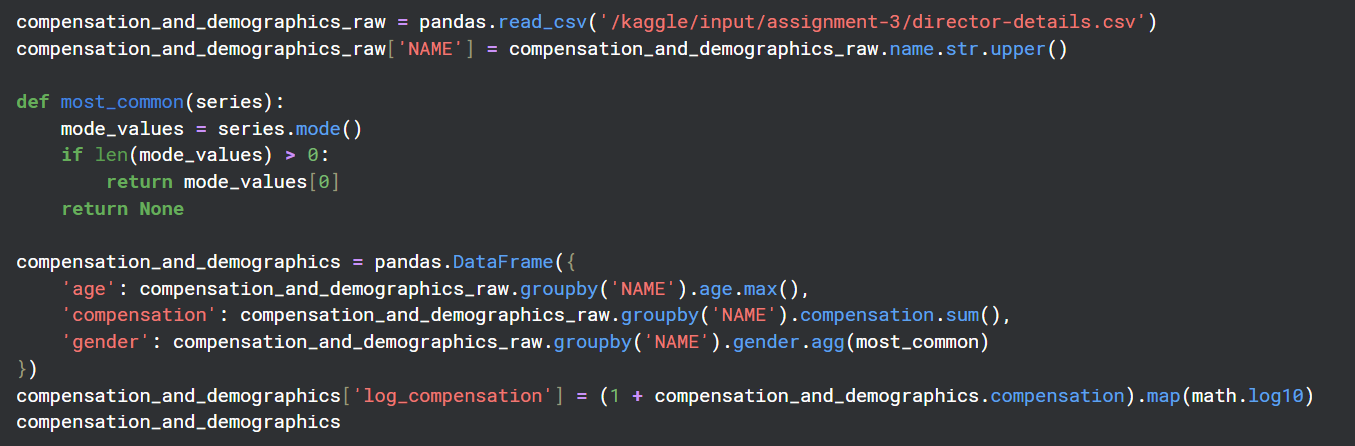
Cell 3 did an undocumented logarithmic transformation of compensation data without any explanation. This transformation was important for clustering analysis but had no justification, creating problems for reproducibility and understanding what was happening.

Figure 7: Undocumented Log Transformation (Before Fix)

I fixed this by adding detailed documentation explaining why we add one to handle zero compensation values and why logarithmic transformation is necessary to manage extreme compensation ranges that could mess up clustering algorithms. I also consolidated the import statements for better code organization and added detailed comments throughout the code.

**3. Director Tenure Feature Exploration**

*Feature Selection Rationale*

The dataset had several unexplored features, including software background, tenure duration, temporal patterns and board composition diversity. I picked director tenure as it is the most strategically relevant unused feature for the venture capital acquisition scenario.

*Implementation Approach*

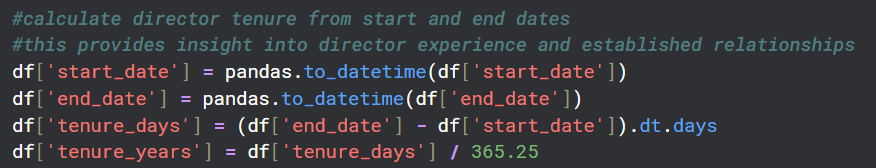
For the director tenure calculation, we used the existing start\_date and end\_date fields to compute service length in both days and years. My analysis revealed an average tenure of 4.9 years with a median of 4.0 years, ranging from zero to 24 years. This indicates realistic corporate board service patterns.

Figure 8: Director Tenure Analysis

*Strategic Integration*

The tenure analysis effectively identified potential acquisition targets in conjunction with the existing centrality measures. Directors with considerable tenure and high centrality scores are the most desirable targets due to their established relationship depth and extensive network reach.

My analysis identified directors such as Prema Mathai-Davis, who possessed a high eigenvector centrality and nearly nine years of tenure across 13 board positions. These directors are ideal candidates for facilitating acquisitions due to their combination of network influence and relationship stability.

*Business Value*

The tenure analysis offers numerous strategic advantages for the purpose of acquisition planning. This implies that directors who have been in office for an extended period have a larger capacity to influence acquisition discussions due to their deeper connections within corporate networks. High tenure also suggests experience in board governance and merger activity, which is beneficial for the facilitation of complex transactions.

**4. Complementary Dataset Integration**

*Selected Dataset*: Fortune 1000 Companies (2021) [1]

This Fortune 1000 dataset contains financial metrics, company structure data and performance indicators for America's largest companies. This directly supports our venture capital fund's goal of targeting "big companies with lots of money that are listed on USA-based stock exchanges."

*Integration Strategy*

This dataset allows us to do better targeting by identifying directors connected to Fortune 1000 companies and then weighting their influence by connected company revenue and profit. The sector data allows matching between our target company and relevant Fortune 1000 industries. We can develop scoring systems that combine director centrality with Fortune 1000 financial metrics to create practical acquisition targeting tools.

**5.**

**b) Implementation of data**

I will commence my implementation by importing the Fortune 1000 CSV file and subsequently matching company names between datasets using uppercase conversion to ensure consistency. Through a direct comparison of company names, I will generate a mapping that identifies the directors in our network who are on the boards of Fortune 1000 companies.

I will proceed to improve director profiles by incorporating financial data from the Fortune 1000, such as revenue, profit, and sector information, for each connected company. This enables the calculation of aggregate metrics, such as the total revenue of the Fortune 1000 associated with each director.

The analysis will create a straightforward scoring system that integrates the total revenue of the Fortune 1000 companies affiliated with each director with the number of Fortune 1000 boards on which they serve. This results in a "Fortune 1000 Connection Score" that quantifies the extent of each director's access to affluent potential acquirers.

Finally, I will generate a ranked list of directors arranged according to their Fortune 1000 connection scores and develop fundamental visualizations that illustrate the distribution of Fortune 1000 connections within the network. The top directors for acquisition facilitation will be exported as a CSV file, providing the VC fund with a list of directors to approach for connections with Fortune 1000 companies.

**Conclusion**

This assignment effectively improved the director network analysis by addressing critical code issues, incorporating betweenness centrality measures, and investigating director tenure data. The extended analysis offered a more comprehensive understanding of the network structure, while the code enhancements resolved performance and documentation issues. The incorporation of the Fortune 1000 dataset offers a method for converting the analysis into practical tools that will support the acquisition objectives of the venture capital fund.

## 

## **References**

[1] Hudson, W. (2022, November 4). 2022 Fortune 1000. Kaggle.<https://www.kaggle.com/datasets/winston56/fortune-500-data-2021>

**Use of Generative AI**

Throughout this assignment, I used Claude (Anthropic's AI assistant) to help with code optimization guidance, dataset research for finding complementary data sources, documentation structure organization, code explanation for complex network analysis concepts and writing review for clarity improvements. All analytical decisions, code implementations and final interpretations are still my own work, with AI only serving as a research and writing grammar assistant.