COMP6200 Critical Analysis

Analytical Rationale and Code Modifications Summary

Md Mahdi Mohtasim, ID: 48926434
mdmahdi.mohtasim@students.mq.edu.au
Macquarie University

June 8, 2025

Introduction

This report documents the modifications and analytical choices made in my final submission notebook for Assignment 3. The aim was to identify influential directors using network centrality and enrich that analysis with external data from wealthy companies.

Task 1 – Centrality Extension

To enhance the network analysis of director influence, I extended the notebook by incorporating an additional centrality measure alongside the existing ones.

Existing Centrality Measures

- **Degree Centrality**: In this context, a director with high degree centrality sits on many company boards and is widely connected. Such individuals are often valuable for maximizing access and outreach across the network.
- **Eigenvector Centrality**: This measure prioritizes directors who are not only well-connected but are also connected to other highly influential people. It is useful for identifying elite or prestigious individuals who may have significant sway in corporate decision-making.

Added Centrality Measure: Closeness Centrality

I extended the analysis by adding **closeness centrality** (implemented as `closeness_weighted` in the code). Closeness centrality measures how quickly a node can reach all other nodes in the network using the shortest weighted paths.

In this project, edge weights were constructed using director-specific attributes such as `compensation`, `role_score`, and `service_years`. This allowed the closeness measure to reflect not just structural distance but weighted strategic position in the graph.

A high closeness value suggests that a director can efficiently communicate with or influence others, making them ideal **connectors or brokers** in the board network—exactly the kind of individuals who might facilitate acquisition opportunities.

This added metric provides a richer and more balanced perspective when combined with degree and eigenvector centrality, capturing visibility, prestige, and strategic positioning in the network.

Task 2: Code Repair

Fix 1: Improved Code Documentation

The original notebook had multiple dense code blocks without inline comments or explanations. This made it difficult to follow the logic and purpose behind key transformations.

- Inline comments were added to each code block.
- Organized cells into sections and sub-sections for better document structure.
- Commented out unnecessary clustering code, which was irrelevant to the goal

Fix 2: Standardized Names Using Regex Cleanup

I improved the name-cleaning step by applying regular expressions to replace all non-alphabetic characters with spaces and to collapse multiple consecutive spaces into a single space. This significantly improved name-matching accuracy across datasets, particularly when integrating the complementary dataset.

The names `CLAIRE BABINEAUX- FONTENOT` and `CLAIRE BABINEAUX-FONTENOT` were previously treated as distinct individuals. After standardization, they are now correctly recognized as the same person, ensuring more accurate aggregation and merging across sources.

Fix 3: Grouped Duplicate Director Entries by Company

The original dataset contained multiple duplicate rows for the same director within the same company. This was likely due to directors having multiple terms, partial SEC filings, or data entry inconsistencies.

To resolve this, I grouped the data by both `director_name` and `company_name`, and then summed the `service_years` for each group. This ensures that each director-company relationship is uniquely represented in the processed dataset.

This fix improves data integrity by:

- Preventing service years from being artificially inflated.
- Avoiding the creation of **multiple nodes** for the same director in the network graph, which would distort centrality and influence metrics.

Prior to this fix, the dataset contained approximately **93,000** director–company relationships. After grouping, we now have about **40,000** unique relationships, significantly reducing redundancy and improving network accuracy.

Fix 4: Enhanced Influence Scoring (New Metric)

Previously, the influence of directors was measured using only **eigenvector centrality**, which, while powerful, may not fully capture a node's overall impact in the board network.

To improve this, I introduced a new feature: `influence_score`, a composite metric calculated using:

- **Eigenvector Centrality** – Measures prestige through connections to influential peers.
- **Closeness Centrality** – Captures how efficiently a director can reach others in the network.
- **Degree Centrality** – Reflects the breadth of a director's direct board memberships.

By combining these three dimensions, the `influence_score` allows for a more comprehensive and balanced ranking of directors. It accounts for both reach and relational quality, enhancing the accuracy and fairness of identifying top influencers in the network.

Task 3 – Feature Exploration in Existing Dataset

To enrich the network analysis, I engineered two new features from the existing dataset:

- **role_score**: A custom scoring metric that assigns weights based on a director's role in the company. For example, directors holding executive titles such as CEO or Chairperson receive higher scores than ordinary board members.
- **service_years**: Calculated as the difference between the start and end years of a director's service, this metric captures the tenure of a director on the board.

These engineered features help capture two critical aspects of a director's potential influence:

- **Authority**, represented by their role in the organization.
- **Experience**, indicated by the length of service.

By incorporating these dimensions into centrality and influence metrics, we better identify high-impact individuals who may play a key role in facilitating corporate acquisitions.

Task 4: Complementary Dataset- Fortune 1000 (2024)

To enhance the analysis, we incorporated a complementary dataset from Kaggle:
"2024 Fortune 1000 Companies" — Dataset Link

This dataset provides an updated list of Fortune 1000 companies for the year 2024. It helps us:

- Identify whether a director is currently associated with a **top-performing company**.
- Enrich centrality and influence analysis by **flagging high-profile affiliations**.
- Support visualizations or filtering of **top-tier networks**.

A complementary feature called `worked_in_top_company` was added using this dataset to tag directors with Fortune 1000 affiliations.

Task 5: Refinement

Task 5a – Presentation Visualizations

To effectively communicate my findings to a non-technical audience, I created a visual presentation that summarizes the work process, key metrics, and outcomes of the analysis. This presentation is structured to guide a venture capital (VC) team through the logic and impact of using board network analysis to identify high-potential directors. A visual summary has been provided in the accompanying .ppsx presentation file.

Task 5b – Implementation of Complementary Dataset (Fortune 1000)

To enhance the analysis, I incorporated the Fortune 1000 dataset, which lists the top U.S. companies by revenue. Since the VC fund's goal is to facilitate an acquisition by a wealthy U.S.-based firm, this dataset is highly relevant for identifying directors with valuable corporate affiliations.

Each director was flagged with a boolean indicator `worked_in_top_company`, which is set to `True` if they served on the board of any Fortune 1000 company. This flag was merged into the main dataset after cleaning and standardizing the company names using regular expressions.

The influence analysis was driven by three core centrality measures:

- **Degree Centrality**
- **Eigenvector Centrality**
- **Closeness Centrality**

These values were standardized and combined into a single composite score called `influence_score`. Directors with a high `influence_score` and a `worked_in_top_company` flag set to True were prioritized in the final ranking.

This implementation added real-world business relevance by highlighting directors who are not only structurally important in the network, but also possess experience in large, acquisition-capable firms.

Conclusion

All five tasks have been completed in the notebook and documented in this report. My analysis integrates network theory, real-world data, and visual storytelling to support the VC firm's decision-making goals.