**Report:** *PageRank for Identifying Central People in News Articles*

## Introduction

This report presents the implementation and results of a PageRank-based iterative method to identify the most important people occurring in news articles. The data used is a plain text file containing an undirected and unweighted graph of a social network of co-occurrence in news articles. The graph has been constructed from a subset of 3000 news articles from the Reuters-21578 corpus by identifying person names.

## Methodology

The PageRank algorithm is implemented from scratch using Python, with the power iteration method. The provided network was undirected, and therefore, before applying the PageRank algorithm, it was first converted to a directed network.

1. For each edge vertex1 vertex2, an edge in the opposite direction, i.e., vertex2 vertex1, was included.
2. The teleportation rate was set to 0.10.

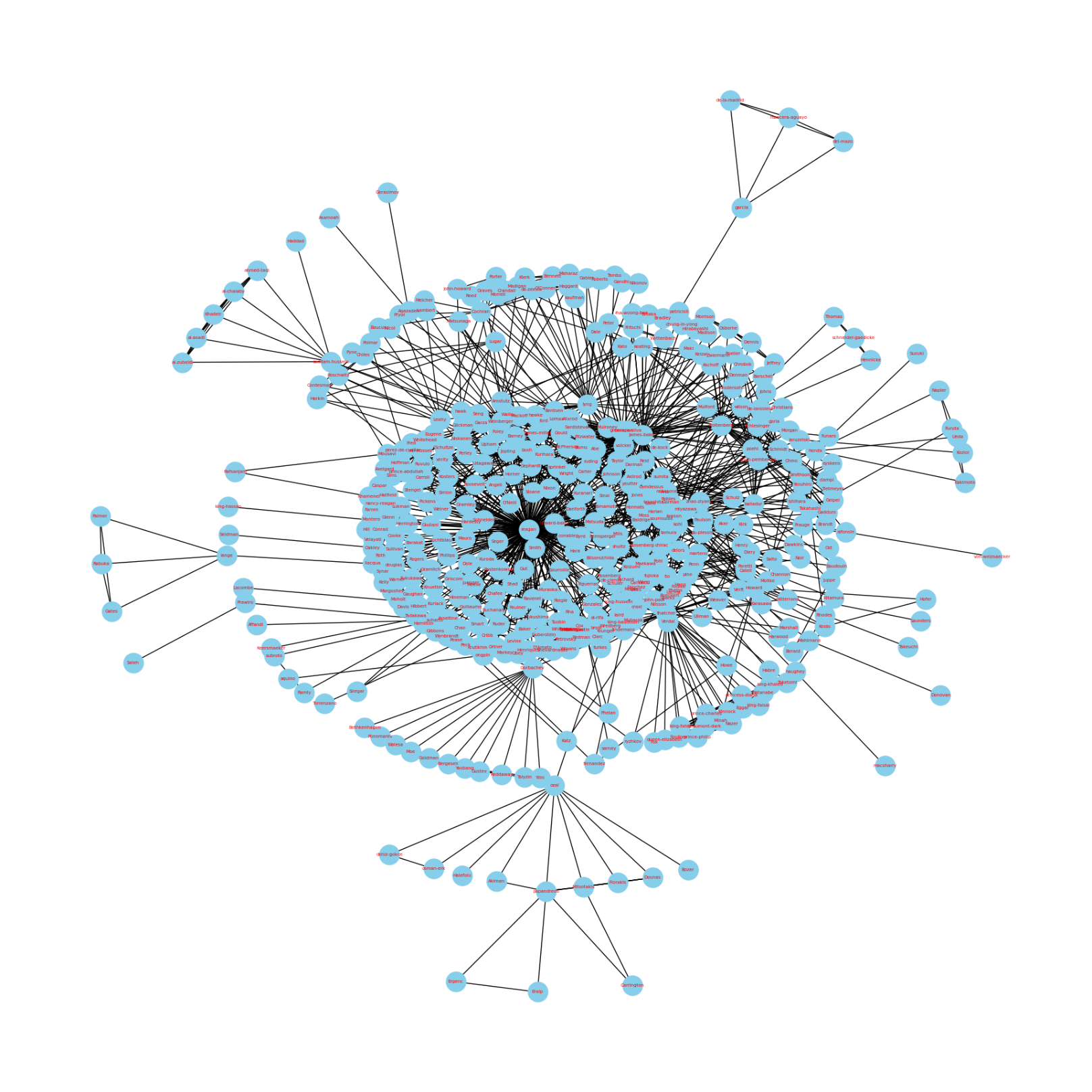
## Results

The PageRank algorithm was run on the directed network to determine the most central people in the co-occurrence graph. The names of the top 20 people, as well as their PageRank scores, are presented in the table below:

| Rank | Name | PageRank Score | Node Number |
| --- | --- | --- | --- |
| 1 | Reagan | 0.06556774409502332 | Node 7 |
| 2 | James Baker | 0.02152838254339281 | Node 2 |
| 3 | Nakasone | 0.016321790735326854 | Node 6 |
| 4 | Thatcher | 0.01405100683504674 | Node 38 |
| 5 | Stoltenberg | 0.011032520911152957 | Node 43 |
| 6 | Lyng | 0.011001261817001048 | Node 12 |
| 7 | Weinberger | 0.00969327172531046 | Node 83 |
| 8 | Volcker | 0.009640655541153413 | Node 3 |
| 9 | Howard Baker | 0.009296323591989826 | Node 8 |
| 10 | Yeutter | 0.008759694982874483 | Node 13 |
| 11 | Gorbachev | 0.008154409808426516 | Node 256 |
| 12 | Greenspan | 0.008039708487115724 | Node 57 |
| 13 | Miyazawa | 0.007676814974598526 | Node 33 |
| 14 | Gephardt | 0.007116743507893871 | Node 151 |
| 15 | Kohl | 0.007002209233901579 | Node 24 |
| 16 | Poehl | 0.00673408373018457 | Node 45 |
| 17 | Shultz | 0.006181267819023077 | Node 96 |
| 18 | Howard | 0.005893464789654408 | Node 294 |
| 19 | Ozal | 0.005544136360031637 | Node 36 |
| 20 | Chirac | 0.0051600557344136325 | Node 31 |

## Graph Visualization (Vertices and Nodes):

The graph of all the people given in the data with vertices, links and nodes are visualized using python as follow:



## Discussion

The names listed are prominent figures from the 1980s, which makes sense given that the dataset is from 1987 news articles. For example, Ronald Reagan was the 40th President of the United States during that time, and James Baker served as the Secretary of the Treasury and later as the Secretary of State under Reagan. Similarly, Margaret Thatcher was the Prime Minister of the United Kingdom, and Yasuhiro Nakasone was the Prime Minister of Japan.

This suggests that the PageRank algorithm is working as expected, identifying the most frequently mentioned and therefore presumably most important people in the news articles. However, to fully validate the results, a deep understanding of the context (i.e., the events and personalities of the time) is needed, or the results should be compared with a benchmark or ground truth if available.

This would help us confirm that our algorithm is truly capturing the essence of that fascinating decade.

## Screenshots of Running Results:

Here are the screenshots of the results of our algorithm:

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

## Conclusion

The PageRank iterative algorithm is perfectly able to Identify Central People in News Articles as we can compare the results with the graph presented in the report.