ONLINE SOCIAL NETWORK ANALYSIS - PROJECT 1 - CS 579

SHALLUM ISRAEL MARYAPANOR - A20547274
SANJANA RAYARALA - A20548132

INTRODUCTION

Social media platforms, which provide a wealth of data ready for study, have become essential components of contemporary culture. In this project, we investigate how to gather, display, and evaluate social media data in order to learn more about the dynamics and architecture of networks. Our goal is to create social networks by scraping data from a selected platform and using different network metrics to find trends and patterns.

This project has several different goals. First and foremost, we aim to obtain hands-on expertise with data gathering methods while taking into account both technological and moral factors like user privacy and data usage guidelines. Second, we want to efficiently depict the gathered data using graph analysis tools so that we can understand the network topology. Ultimately, our goal is to obtain significant understanding of the composition and dynamics of the social network under investigation through the computation of network metrics like degree distribution and clustering coefficient.

With this study, we hope to show that we can successfully negotiate the challenges of social media data analysis and derive useful insights from the findings. We contribute to the expanding body of knowledge in this subject and provide the foundation for future research endeavors by thoroughly documenting our approach, problems encountered, and the consequences of our findings.

DATA COLLECTION

Utilizing RapidAPI, a Python script that communicates with the Twitter API, data for this project was obtained from Twitter. To be more precise, the aforementioned.php URL made it easier to retrieve the accounts that a given user—in this case, Andy Murray—followed. The gathered information created a network of 120 user account nodes, or nodes, and edges that represented the directed link "User A following User B." Using Andy Murray's Twitter handle as an input, the script made HTTP GET queries to the API endpoint using the Python requests package. The Twitter handles of the accounts that were followed were then extracted by parsing the JSON-formatted data that the API had returned.

Challenges Encountered

During the course of gathering data, many difficulties surfaced:

API Rate Limit : The quantity of queries that can be made using the Twitter API in a certain amount of time is limited by rate restrictions. After much investigation, an API with a 1000 request/hour capacity was found, allowing rate constraints to be complied with.

Completeness of Data: It was not always possible to retrieve the following lists in their entirety due to rate limitations and possible privacy settings of individual Twitter accounts. This restriction might affect how comprehensive the network graph is.

Cursor control: To get significant amounts of data, careful cursor control was required due to Twitter's API pagination. By managing cursors well, pagination's drawbacks might be avoided and a larger collection of accounts could be retrieved.

User Privacy and Data Usage Policies

Throughout the project, it was crucial to follow Twitter's user privacy policy and data usage standards. With consideration for users' privacy concerns, the Twitter API only offers access to publicly available data. Only publicly accessible account information was gathered for this research, in compliance with Twitter's privacy policy.

Accessible in the footer section under "Privacy Policy" and "Terms of Service," respectively, Twitter's official website offers extensive documentation about its rules addressing user privacy and data usage. Furthermore, standards for appropriate API usage are outlined in the Twitter Developer Agreement and Policy, which place a strong emphasis on data security, user privacy, and prohibitions on using data for surveillance. Below link also contains data usage policy (Refer point 4)

Privacy policy link:
https://cdn.cms-twdigitalassets.com/content/dam/legal-twitter/site-assets/privacy-policy-2023-10-17/en/x-privacy-policy-2023-10-17.pdf

Data Visualization

For visualizing the results, we utilized both Gephi software and the NetworkX library in Python.

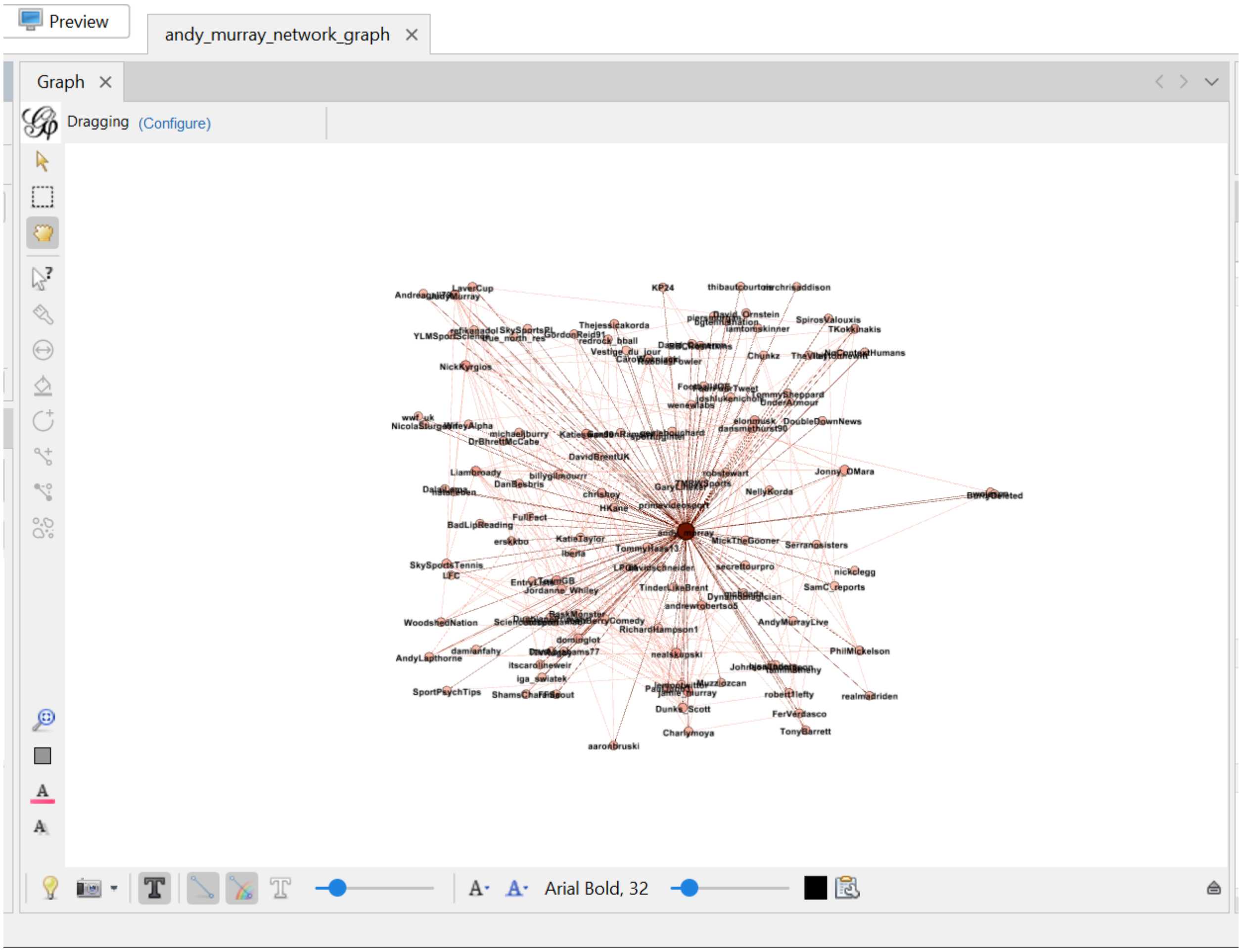
Gephi
Gephi is well known for having a strong and user-friendly graphical user interface, which makes it especially good for visualizing intricate networks. Its intuitive features make it easy to explore and understand network architecture.

NetworkX
A Python package called NetworkX was created expressly for building, modifying, and examining intricate networks. It offers all the features needed for thorough computational analysis and network behavior modeling. Users can efficiently complete complex computational jobs with NetworkX.

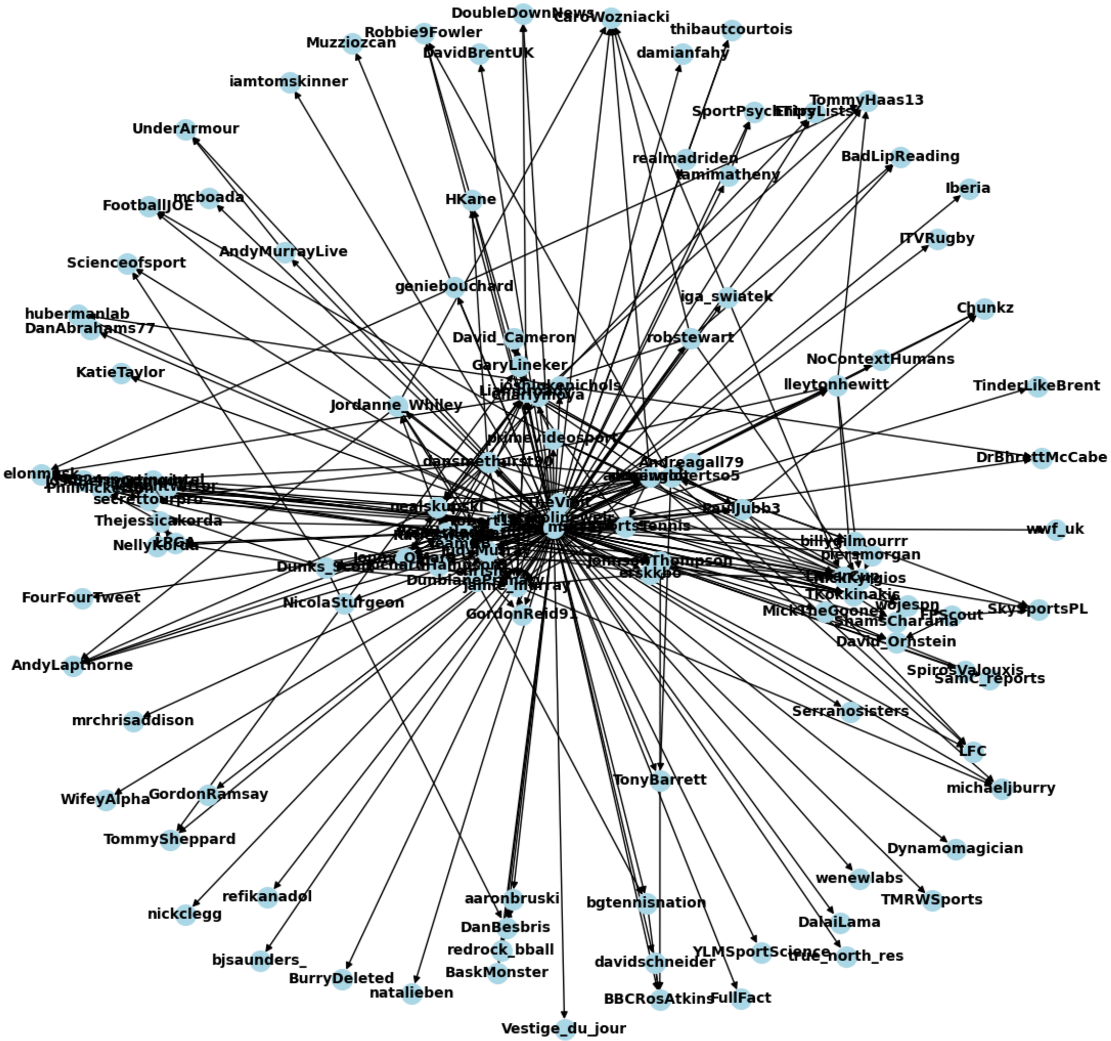
Combinatorial Method
Gephi and NetworkX were both used, which provided a synergistic approach to data visualization. Through the utilization of Gephi's exploratory features and graphical interface, we were able to obtain important knowledge about how complex networks are represented visually. In addition, the use of NetworkX

allowed for thorough computational analysis and the modeling and investigation of network dynamics and behaviors.

Below is the network graph for the user Andy murray using Gephi:



Below is the graph using NetworkX (Python):



Network Measures: Choosing the Right Ones

Average Degree Distribution**:** Offers information about the overall connection of a network, which is important to comprehend the potential for information transmission.

Clustering Coefficient: Assists in subgroup identification by assessing users' propensity to create communities.

PageRank**:** Indicates important influences in the spread of information by measuring the impact of particular nodes.

Degree Distribution:

Shows that there are few highly linked nodes in a sparsely connected network, possibly including Andy Murray.

In-Degree Distribution: The majority of users only have a small number of followers, whereas certain well-known accounts have larger numbers of followers.

Out-Degree Distribution: Users tend to follow fewer other users, indicating that they are selective about whose patterns they follow.

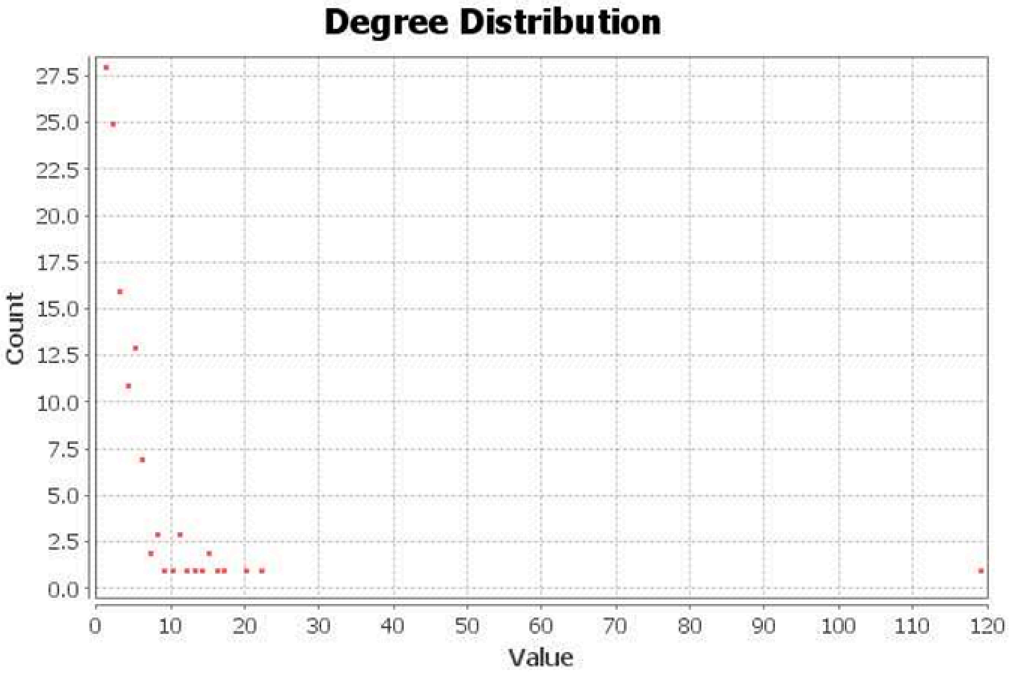
Conclusions: Information flow in the network is driven by influential nodes, and it shows a modest level of communication.

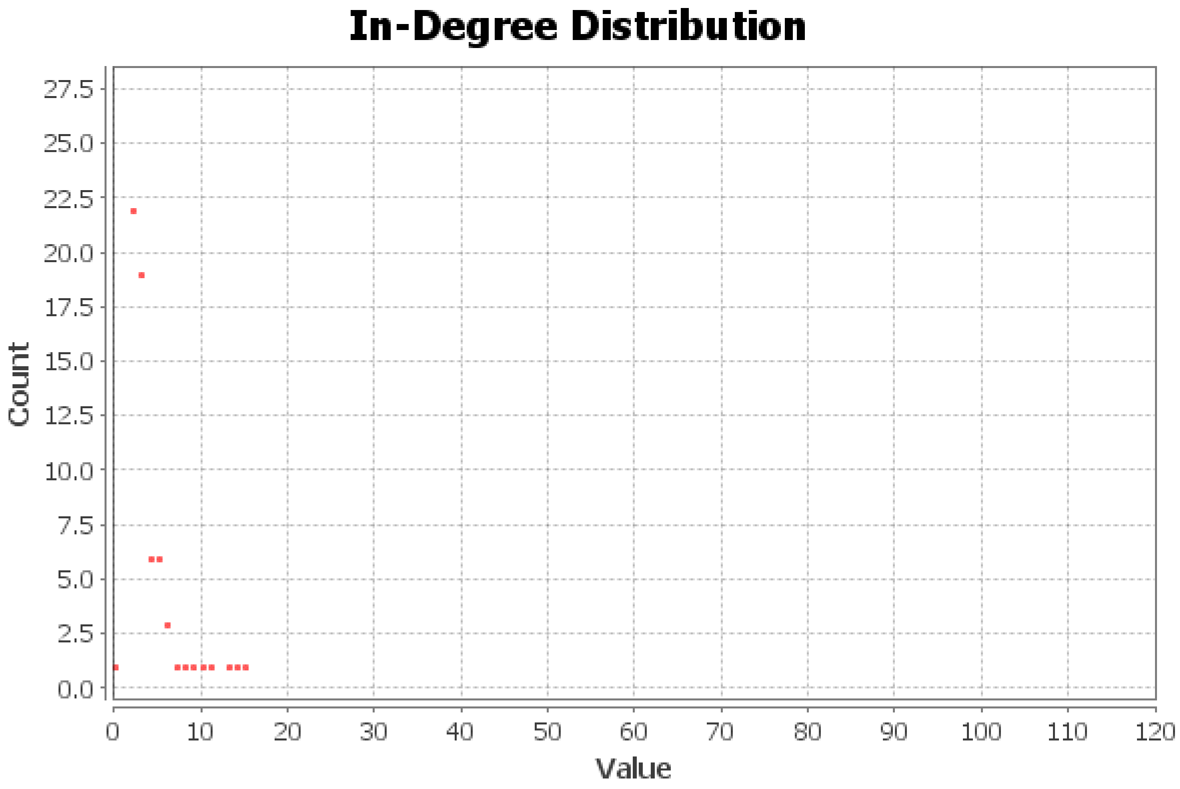
Important influencers have a big say in how information spreads throughout the network.

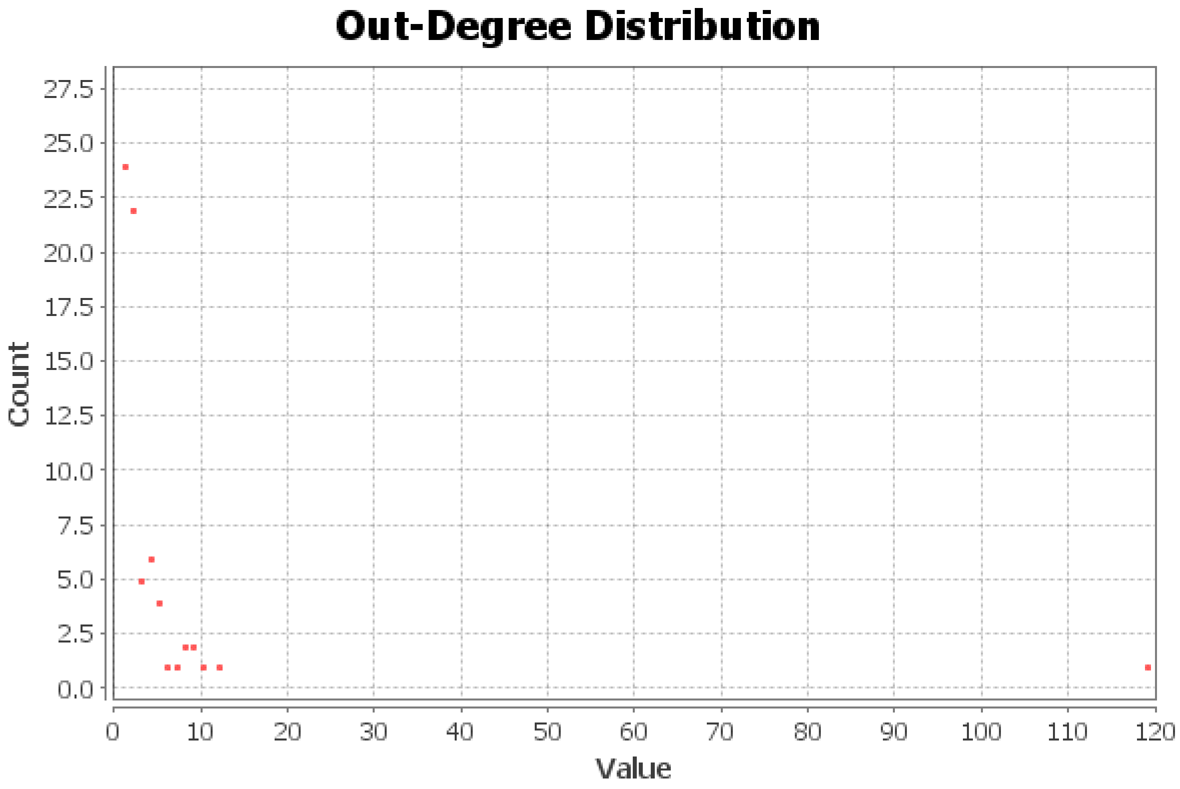
Degree Report

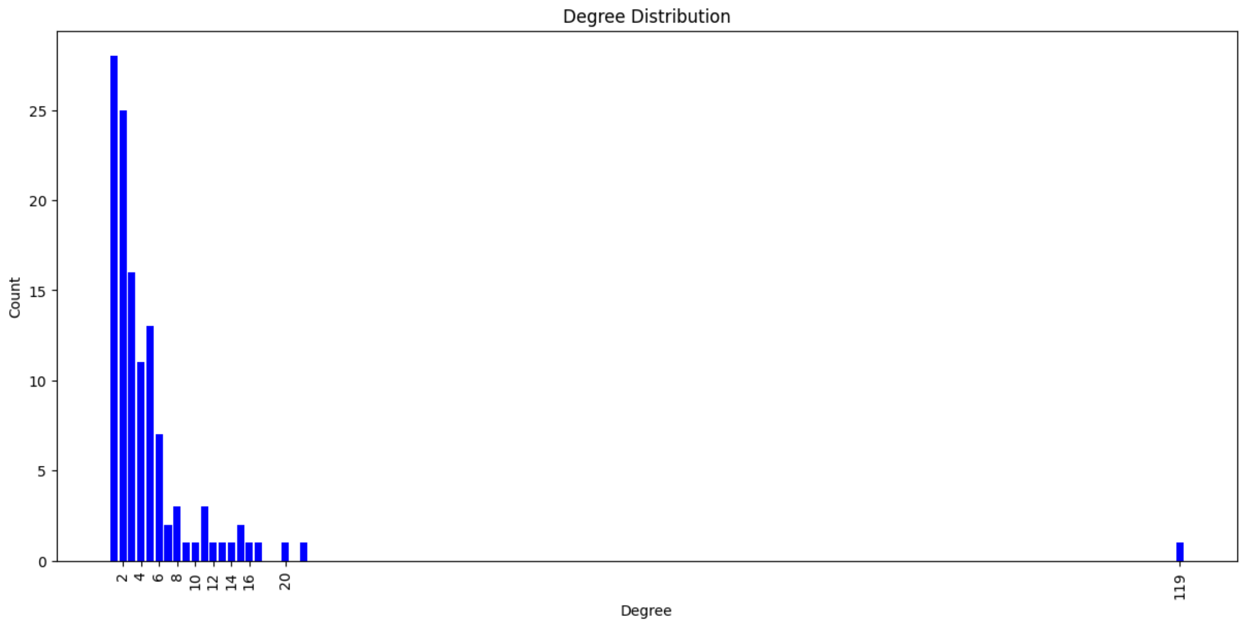
Results:

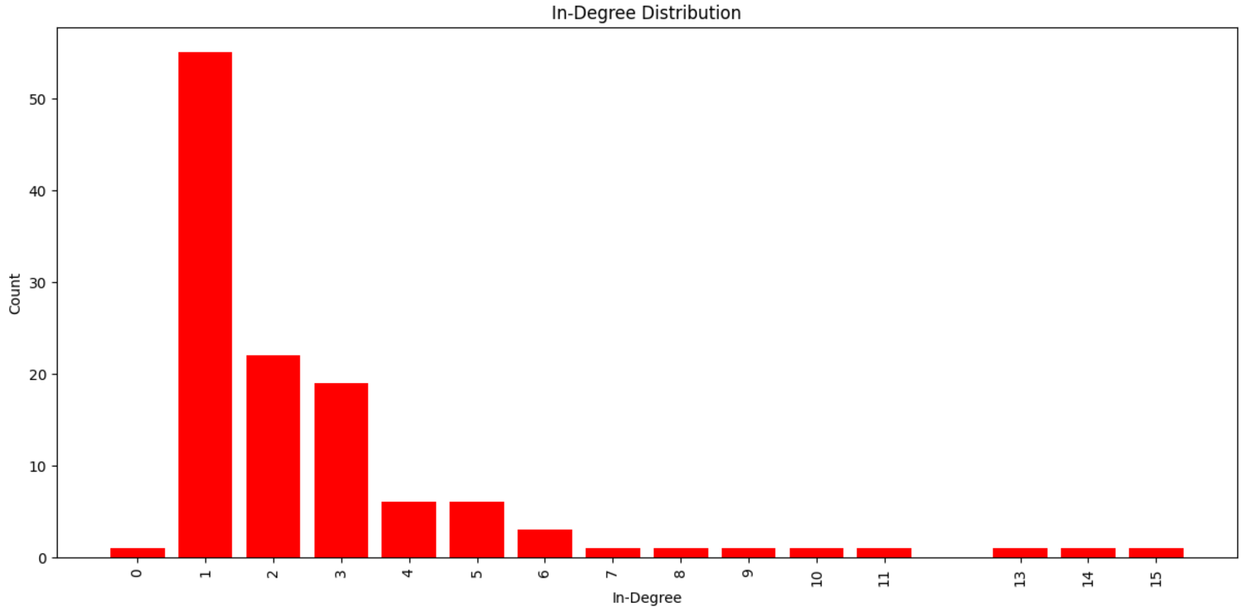
Average Degree: 5.250

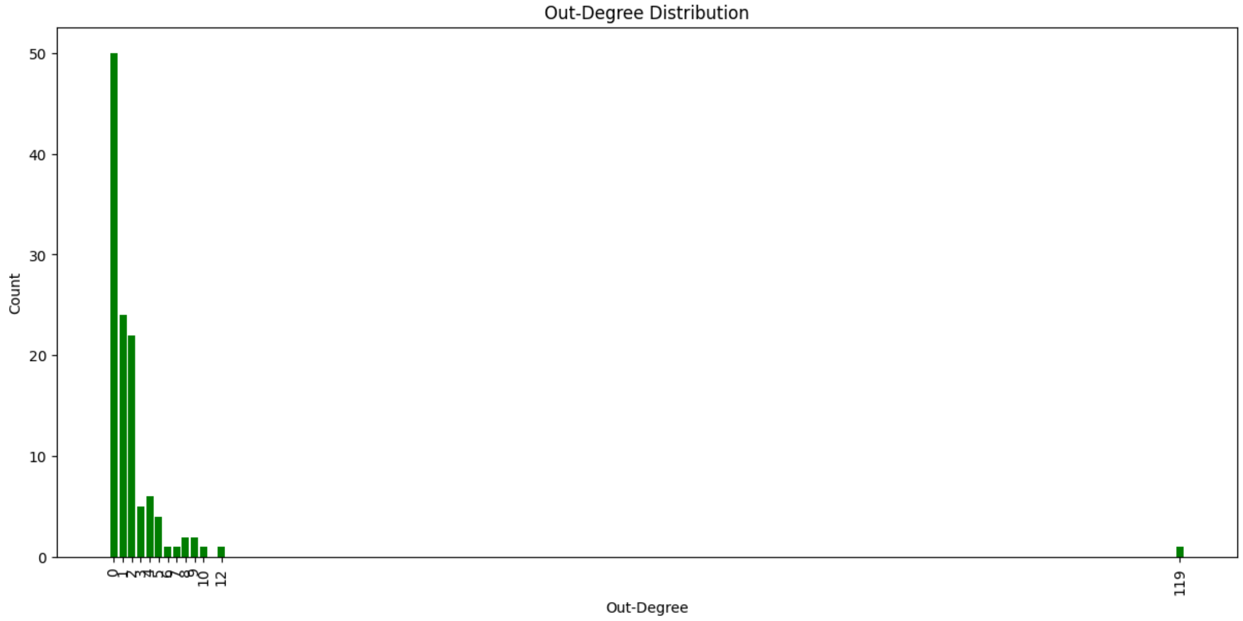



In-Degree Distribution
Count
27.5
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
10
20
30
40
50
60
70
80
90
100
110
120
Value


Out-Degree Distribution
Count
27.5
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
10
20
30
40
50
60
70
80
90
100
110
120
Value


Degree Distribution
Count
0
5
10
15
20
25
2
4
6
8
10
12
14
16
20
119
Degree


In-Degree Distribution
Count
0
10
20
30
40
50
0
1
2
3
4
5
6
7
8
9
10
11
13
14
15
In-Degree


Out-Degree Distribution
Count
0
10
20
30
40
50
0
1
2
3
4
5
6
7
8
9
10
12
119
Out-Degree

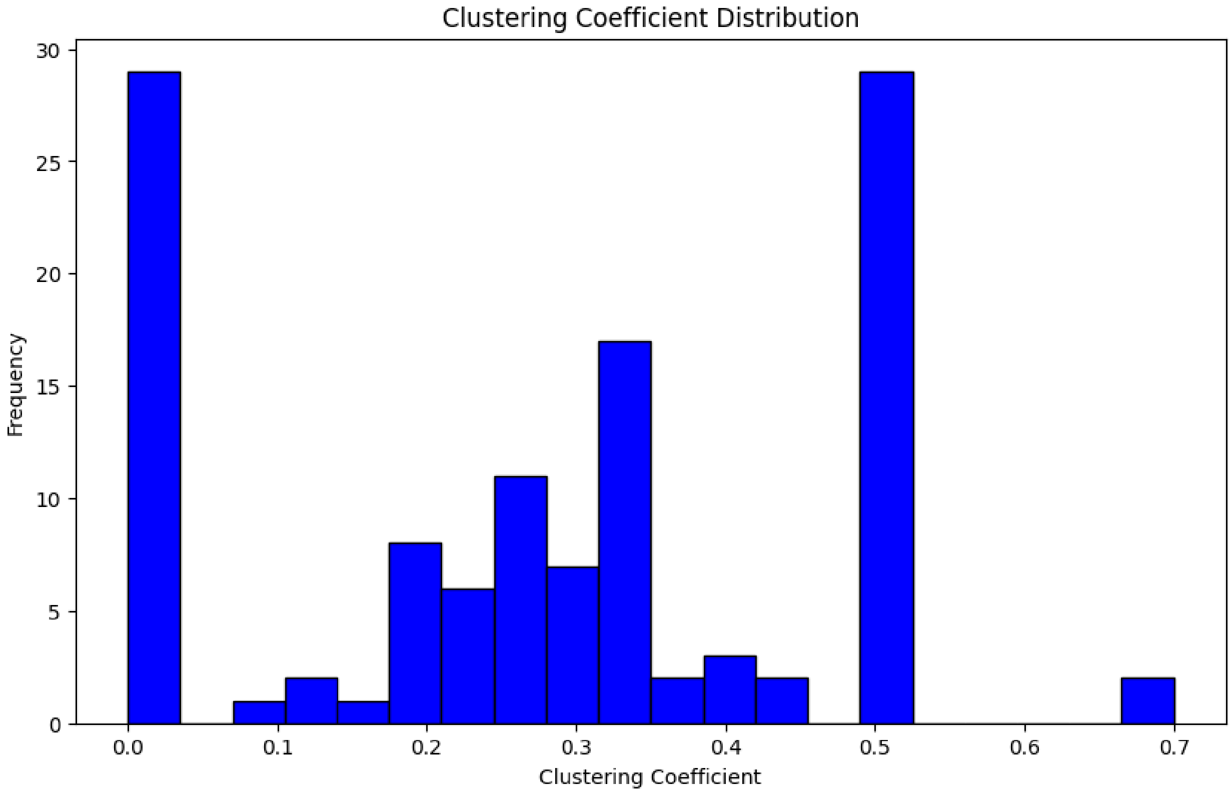
Coefficient of Clustering

Histogram: Moderate clustering is visible, indicating connectedness within local groups.

Variation: Heterogeneous clustering throughout the network is indicated by peaks in the histogram.

Findings: The network appears to have smaller communities based on moderate clustering.

Information dissemination dynamics may be influenced by community structures.



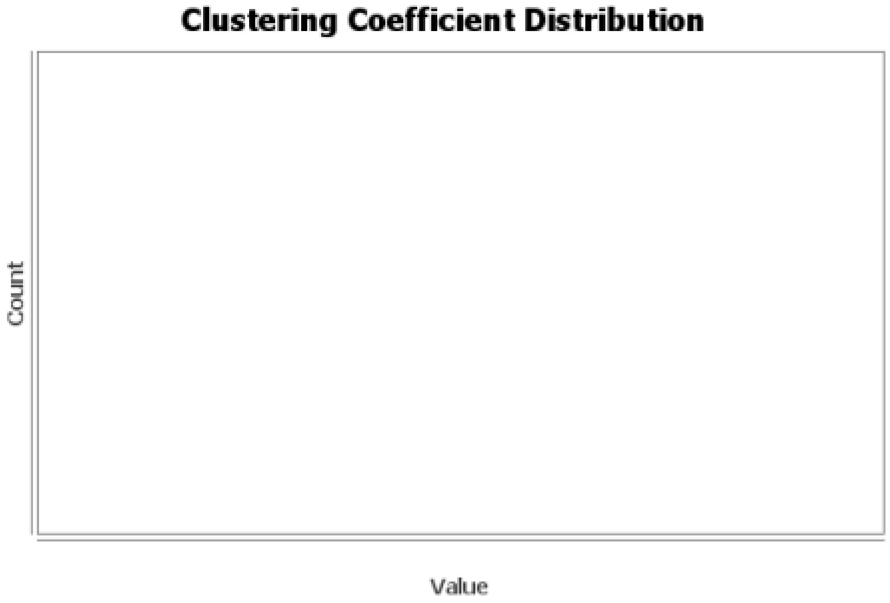
Clustering Coefficient Metric Report

Parameters:

Network Interpretation: directed

Results:

Average Clustering Coefficient: 0.273
The Average Clustering Coefficient is the mean value of individual coefficients.



Algorithm:

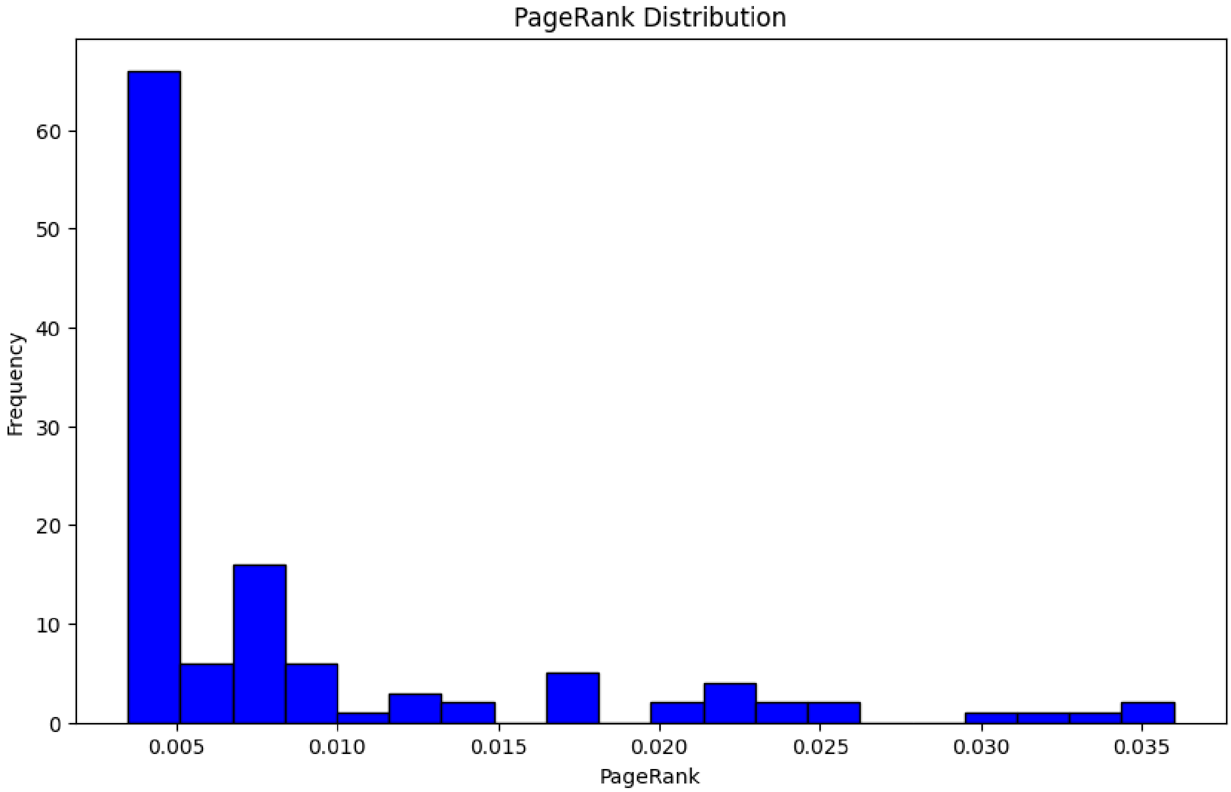
Simple and slow brute force.

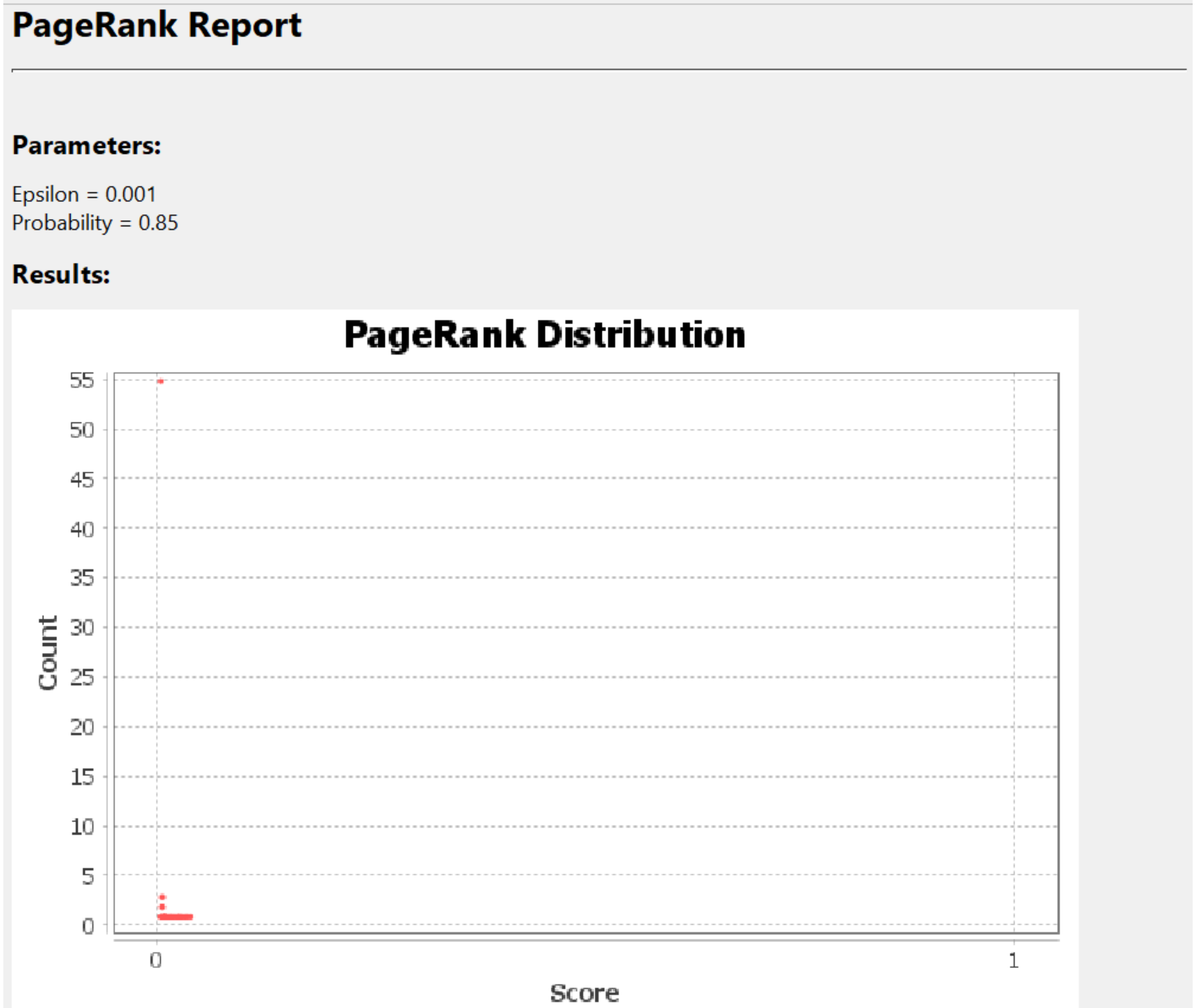
PageRank Notes:

Histogram: Shows a distribution that is right-skewed and has a small number of very important nodes.

Influential Nodes: Andy Murray and other nodes with high PageRank scores act as central centers for the spread of information.

Findings: The influence hierarchy of the network is based on a "hub-and-spoke" model, where a small number of important nodes have a substantial impact.





PageRank Report

Parameters:

Epsilon = 0.001
Probability = 0.85

Results:

PageRank Distribution

Analysis of the Findings and Issues Raised by the Findings

Influence Dynamics: Elements that give some nodes their high influence scores. Community Structures: Common interests or real-world relationships as the foundation for clustered groups?

The function of content is to affect network position through shared material. Effects of Network Evolution: Stability of the Influence Hierarchy and Gradual Changes in Network Measures.

Next Measures for Additional Research

- Temporal Analysis: Monitor the evolution of networks to comprehend their dynamics.
- Analyze content: Examine how network position relates to content.
- Finding and evaluating online communities is known as community detection.
- Influence Factors: Examine the characteristics of nodes with influence.
- Network Robustness: Use node removal simulations to evaluate the resilience of the network.
- Comparison with Other Networks: To find special aspects, compare Andy Murray's network with others.

CONCLUSION:

The examination of Andy Murray's Twitter friendship network yielded significant understanding of the network's composition, interconnection, and dynamics. Following data collecting, visualization, and network analysis, the following important conclusions were made:

Network Connectivity: Information flows through the network at a moderate rate, driven by important influencers. An examination of the degree distribution showed a network that was poorly connected but had a few strongly connected nodes, maybe including Andy Murray.

Community Structures: The network may contain smaller communities, as indicated by moderate clustering coefficients. These communities could have an impact on the dynamics of information dissemination, modifying the information flow inside the network.

Influence Dynamics: PageRank study revealed important nodes that are important for the spread of information. According to the "hub-and-spoke" model of the network, a small number of important nodes have a big impact on the network as a whole.

Future Directions: A number of queries about community structures, influence dynamics, and the effect of content on network dynamics were brought up. To learn more about the evolution and dynamics of the network, more research may delve into community detection, content analysis, and temporal analysis.

In summary, this examination of Andy Murray's Twitter friendship network sheds light on the composition and dynamics of social networks led by celebrities. Comprehending the complexities of these networks is essential to understanding the patterns of information diffusion and the influence of influential individuals on the establishment of online discourse.

References:

[1] For Twitter Developer use cases and tutorials
https://developer.twitter.com/en

[2] For RapidApi use cases and documentation
https://docs.rapidapi.com/docs/consumer-quick-start-guide

[3] For usage of methods in the networkx package
https://networkx.org/documentation/stable/tutorial.html

[4] Data visualization: Gephi
https://gephi.org/tutorials/gephi-tutorial-quick_start.pdf

[5] Matplot lib
https://matplotlib.org/

[6] Pandas
https://pandas.pydata.org/

[7] Python requests
https://pypi.org/project/requests/