CS 579: Online Social Network Analysis

Project -1: Social Media Data Analysis

Submitted by:
Aman Agrawal (A20446346)
Disha Sharma (A20443440)

Platform Chosen: Twitter

Objective:

The objective of the project is to crawl social media data for a subset of its users (100-500) and report analysis on the extracted data.

About the project:

In this project, we have collected a list of Twitter Users and developed a network, which is used to later perform the analysis on the obtained graph using Python. For the reasons mentioned later in this report, the users have been specifically selected to include only the major National Football League (NFL) players. Further, the connections of each of these players are filtered to include only those users that are part of our network.

Two different graphs have been studied from the network:

- Directed Graph: a connection starting from User A to User B is established in this graph only if user A follows User B
- Undirected Graph: a connection between User A and User B is established in this graph only if both the users follow each other.

Finally different network measures such as degree distribution, measures of centrality, etc are calculated for each of these graphs.

Project Requirements:

Other than the standard Python Libraries, following two libraries are required to run the project:

- python-twitter: to collect user information and their followers
- Networkx: to draw the network and perform graphical operations and retrieve inferences.

Project Outline:

The project is broadly divided into following major steps:

1. Data collection and filtering
2. Building a network structure
3. Calculating network measures

1. Data collection and filtering

- To start with, we selected the Twitter platform for this project because a vast number of people use twitter which in turn generates a lot of data. This amount of data will help us better understand and provide better insights which will achieve the goal of the project.
- But as per Twitter API rules, the information retrieval from a Twitter account is possible only if either the account is a 'public' account or if the API owner follows that account.
- As per general observation, celebrities usually have their account public. To narrow down, we formed a network of only the players of the National Football League.
- But, in order to avoid the manual task of listing down these players, the 'scraper.py' file (attached with the project) was used to get a list of names of these players.
- This file, using 'beautifulsoup', fetches the information from https://www.fantasypros.com/nfl/cheatsheets/top-players.php.
- Parsing the information returns the required list.
- Using this list and the Twitter API from 'python-twitter' package, we had our required list of users and their information.
- A simple check for the 'verified' users only, ensures that the resultant users are indeed the players' genuine accounts.

2. Building a network structure

- Used networkx library to build the graph of the network, using the information collected in the previous step
- As a result, a directed graph and an undirected graph are generated.

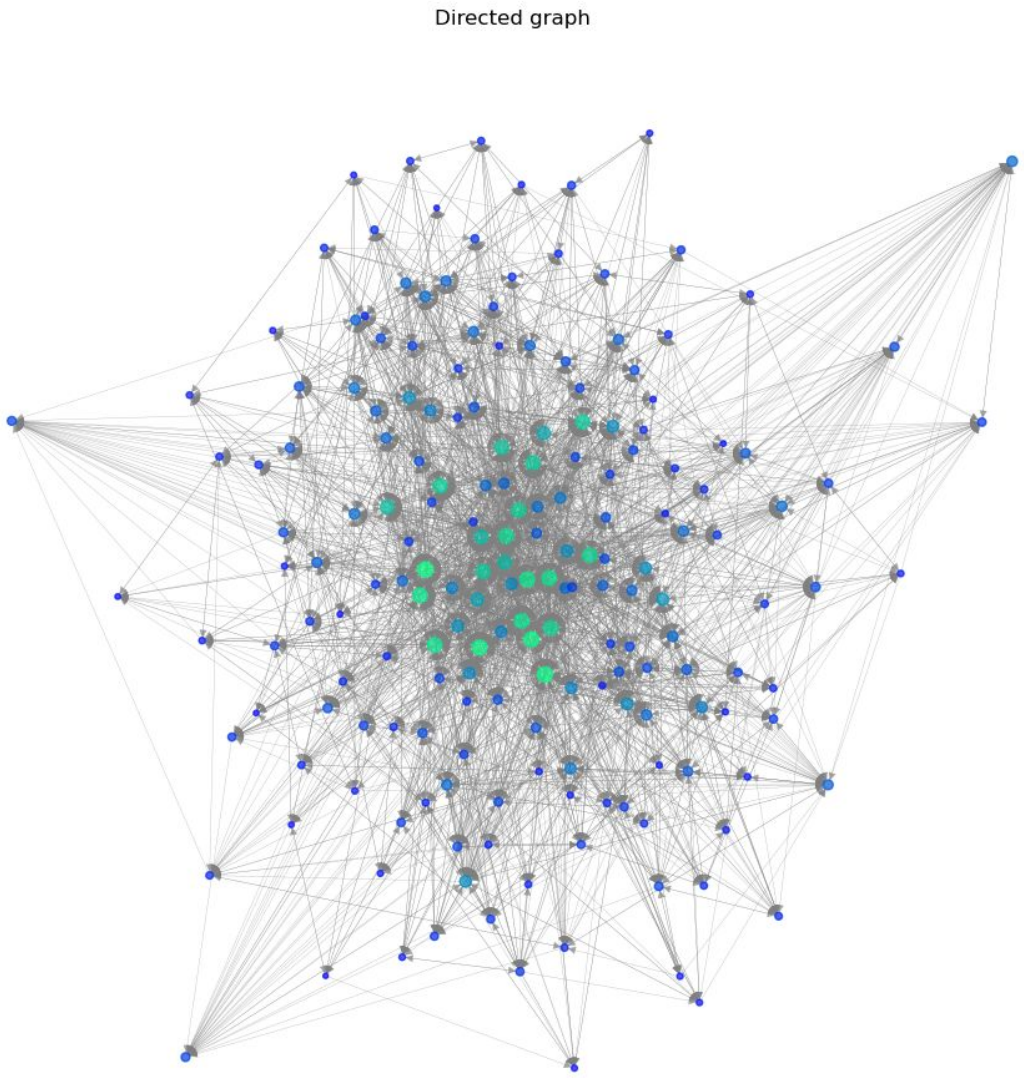
3. Calculating Network Measures

- Following network measures were calculated for each of the graph:
    - Degree distribution and plotting it on a histogram
    - Centrality measures such as: betweenness, closeness, eigenvector, pagerank.
    - Diameter of the graph
    - Reciprocity

Results

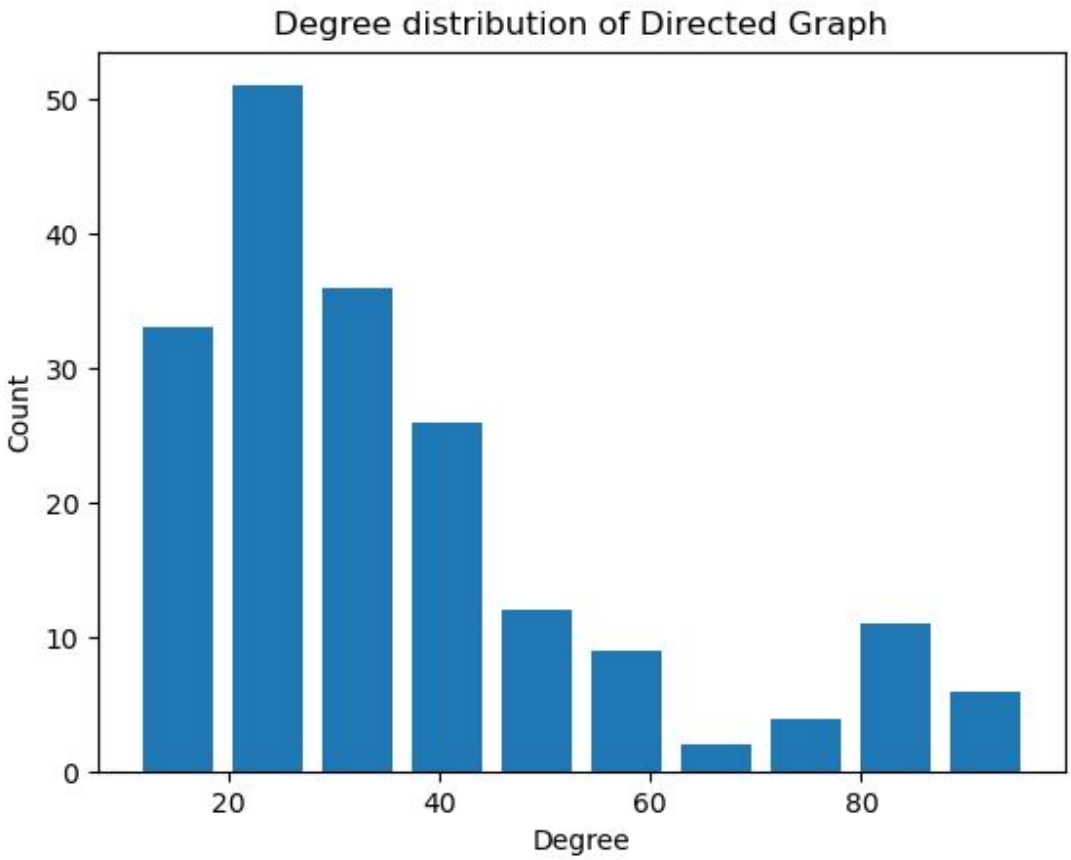
For the Directed Graph of the network:

Graph



- The above directed graph is obtained. Each edge from node A to node B implies that user A follows user B on twitter.
- For the ease of visualization, size of each node is set to be proportional to the indegree of the node and appropriately color mapped.

Degree Distribution



Network Measures:

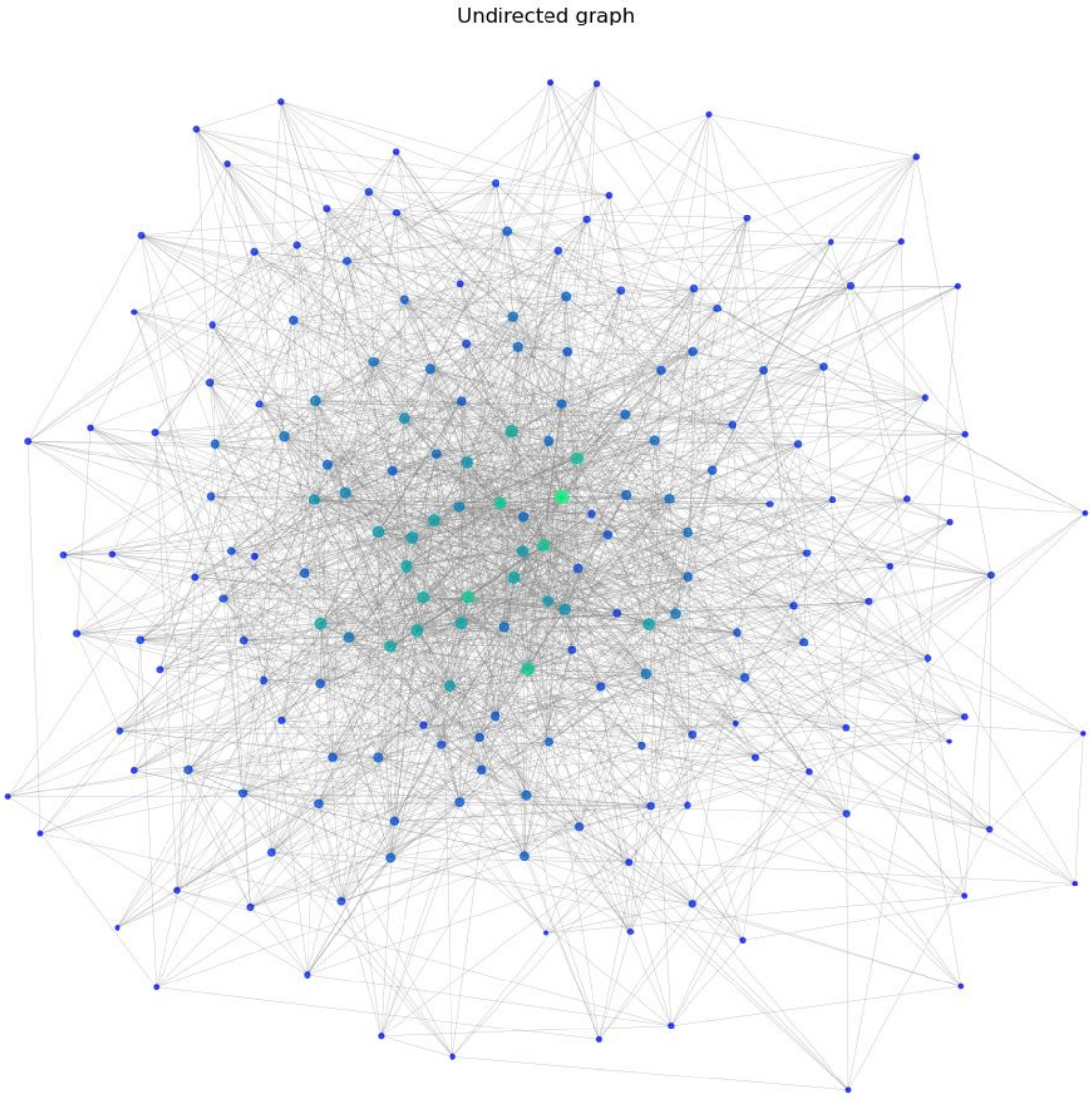
- For the directed graph, the centrality measures are:

Centrality Measure	1	2	3	4	5
Top 5 betweenness centrality	'James Conner', 0.0421	'George Kittle', 0.0402	'Ezekiel Elliott', 0.0344	'Patrick Mahomes II', 0.033	'Trey Burton', 0.0315
Top 5 closeness centrality	'Jarvis Juice Landry', 0.5727	'Ezekiel Elliott', 0.5659	'Mark Ingram II', 0.5642	'Deandre Hopkins', 0.5625	'Saquon Barkley', 0.5592
Top 5 eigenvector centrality	'Jarvis Juice Landry', 0.1943	'Lamar Jackson', 0.1816	'Ezekiel Elliott', 0.1766	'Keenan Allen', 0.1733	'Deandre Hopkins', 0.1695
Top 5 pagerank centrality	'Mark Ingram II', 0.0195	'Larry Fitzgerald', 0.0162	'Mike Evans', 0.0137	'Tom Brady', 0.0135	'Tyler Lockett', 0.0135

- Diameter of the graph: 4
- Reciprocity of the graph: 0.7527
- Correlation coefficient: 0.2155

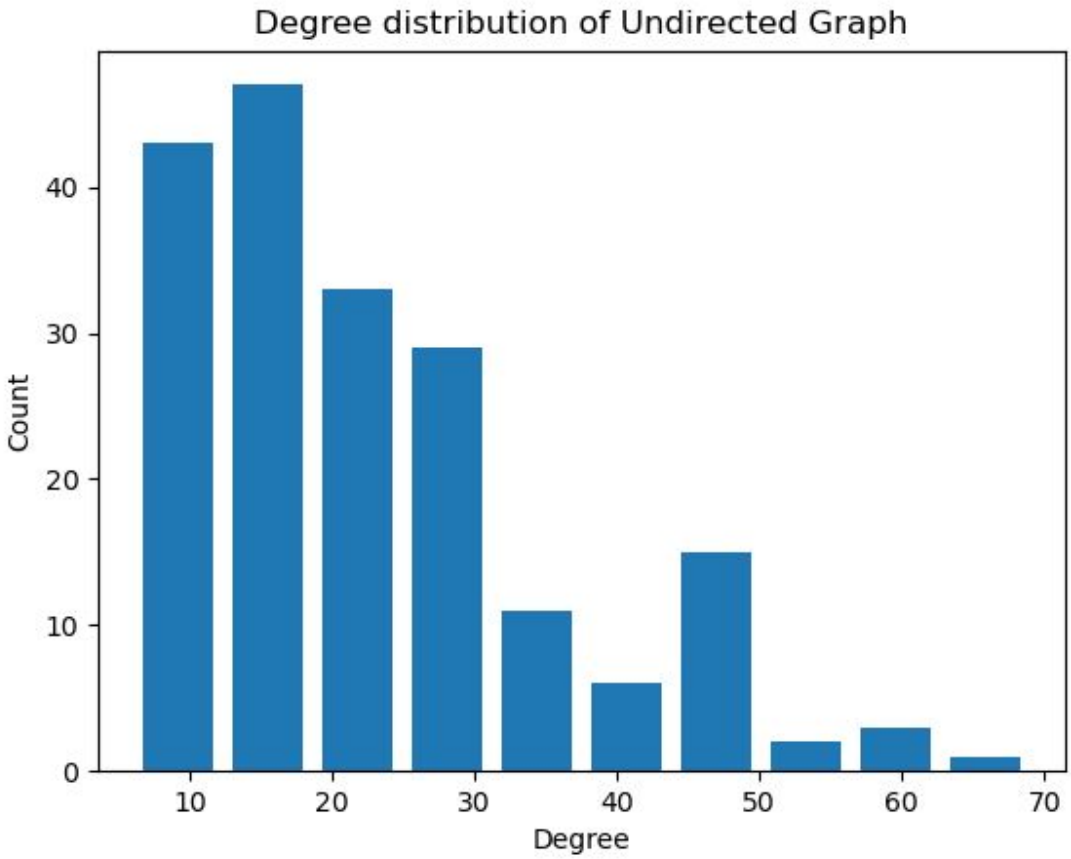
For the undirected graph of the network:

Graph



- The above undirected graph is obtained. Each edge from node A to node B implies that both the users follow each other on twitter.
- For the ease of visualization, size of each node is set to be proportional to the degree of the node and appropriately color mapped.

Degree Distribution



Network Measures:

- For the undirected graph, the centrality measures are:

Centrality Measure	1	2	3	4	5
Top 5 betweenness centrality	'Trey Burton', 0.048	'Miles Sanders', 0.0327	'George Kittle', 0.0298	'DIGGS', 0.0292	'Ezekiel Elliott', 0.0264
Top 5 closeness centrality	'Trey Burton', 0.6077	'Ezekiel Elliott', 0.587	'DIGGS', 0.5851	'Miles Sanders', 0.5833	'Jarvis Juice Landry', 0.5833
Top 5 eigenvector centrality	'Trey Burton', 0.1716	'Jarvis Juice Landry', 0.165	'Ezekiel Elliott', 0.1618	'DIGGS', 0.1549	'Lamar Jackson', 0.1513
Top 5 pagerank centrality	'Trey Burton', 0.0145	'Miles Sanders', 0.0121	'DIGGS', 0.012	'Ezekiel Elliott', 0.0119	'Jarvis Juice Landry', 0.0115

- Diameter of the graph: 3
- Reciprocity of the graph: 0.0
- Correlation coefficient: 0.2520

Inferences

- As expected, the degree distribution follows the power law curve. This can be observed from the histogram plot shown above for both the graphs
- As it can be seen from the table, different centrality measures have different top 5 users in the network. This accounts for different parameters each measure stresses upon during the calculation.
- Nodes with higher degree are clustered towards the centre of the graph

References

- https://developer.twitter.com/en
- https://networkx.github.io/documentation/stable/index.html
- https://blackboard.iit.edu/bbcswebdav/pid-915758-dt-content-rid-16872144_1/xid-16872144_1