CS 579

Online Social Network Analysis

Twitter Network Analysis

Project 1

Chandni Patel

A20455322

Omkar Pawar

A20448802

Introduction

Twitter is one of the major social media platforms and has a lot of information shared daily. People from all domains use Twitter as their place to share updates about things going on in their lives and any ideas from their mind. In this project, we attempt to perform network analysis of twitter users and implement the terminologies learned in class to better understand users from the graph. Based on 300 users of twitter from Bollywood industry, we perform a create a friendship network and see how all these celebrities are connected. This graph allows us to highlight the structure of the network's relationships and identify users whose position is particular.

Why Twitter?

A lot of people use Twitter as their primary source of sharing information. It is a good source to acquire knowledge about any domain that we wish. Most of the user profiles are public, so scraping data from twitter is easy and has fewer restrictions. All the information on twitter is publicly available and can be accessed using Twitter API.

Due to all these factors, we chose Twitter as our platform for analysis.

Twitter API

Twitter API is where we get our data. You need a developer's account to access data using this API. After setting a developer account, you get four keys for authentication. We used these keys in our python script and "twitter" library for python to find friends of specific users.

Approach

Following are the three steps that we take to perform analysis on twitter network

1. Data Collection

We used Twitter API and twitter library in python to write a jupyter script to collect the data.

Our main focus is on the Indian Bollywood industry to get the users. To do so, T-Series is a good place to start with. We can get people that TSeries follows and propagate forward in the network. TSeries follows many celebrities in the industry and it is important we define some criteria to include only some specific nodes in the graph to keep the calculations lucid and easy to interpret.

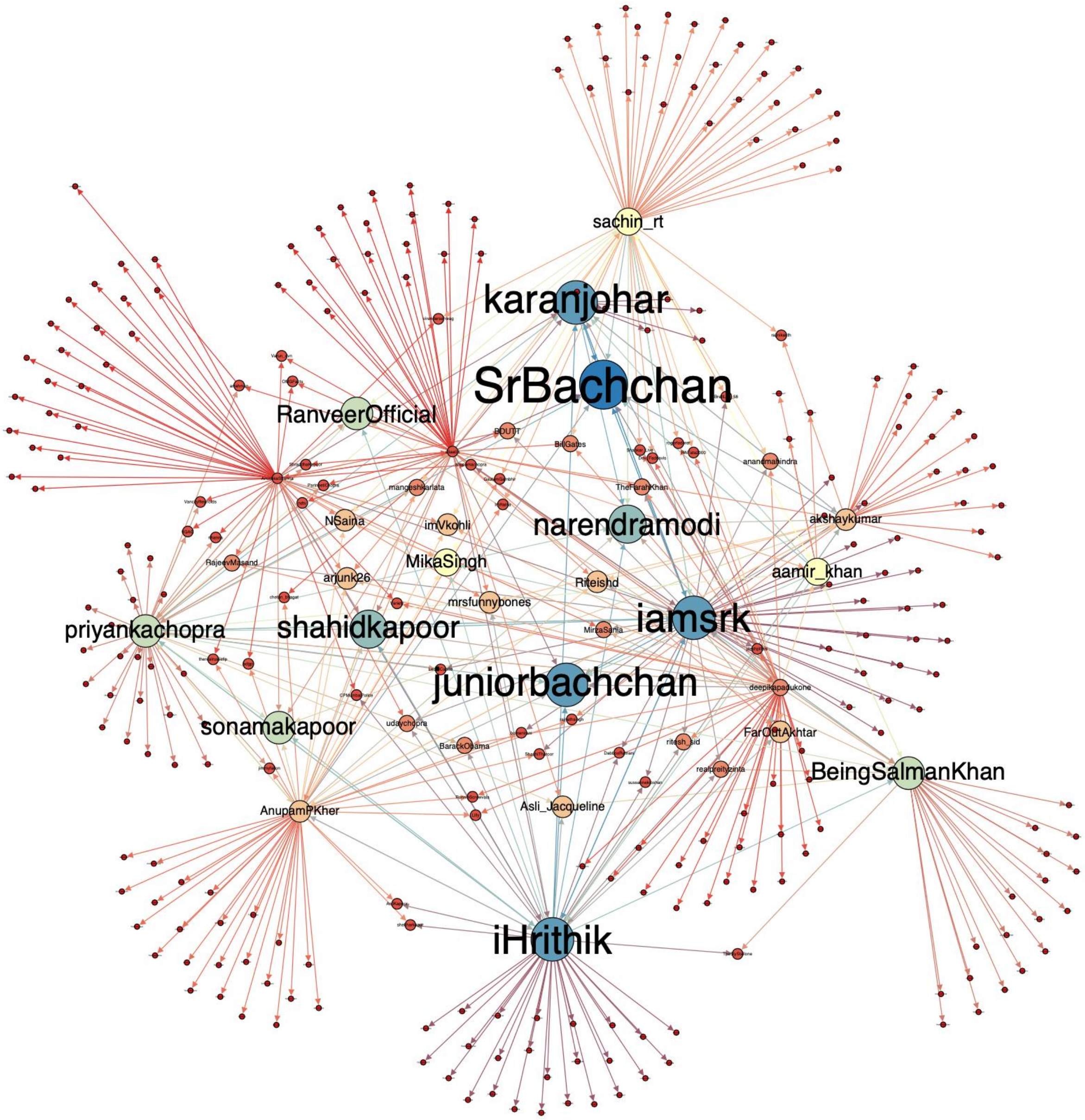
- Defined a function to filter top 50 friends (sorted by number of followers) of a user (TSeries).
- Take 13 celebrities' data to form a network. By data, we mean the top 50 friends of these celebrities using the function defined above.

Once we have all the data at hand, we save it using .csv format and then use this data to draw network diagrams and calculate network measures.

2. Data Visualization

Data collected from the above step is imported in Gephi for visualization. Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free.

Here is the outcome of our twitter network.



While we created the network, we used few network measures to interpret the graph more efficiently and find interesting nodes from the graph. The graph is directed, and the outgoing edge shows that the user node follows the person to which the arrow is pointed. Other viz features are as follows

- **Size of nodes**: In-degree

This is to make the node bigger if it has higher in-degree. In degree is the number of edges coming towards a node.

- **Color of nodes**: In-degree

Color scale that we used here is a blue-red palette. Blue has higher in- degree and turns to red as indegree reduces.

- **Color of edges**: Corresponding to the color of node

The edges that come out from a blue node are blue and red are red. This is useful to see what people users follow. The edge color corresponds to node color.

- **Label Size:** In-degree

More the in-degree of the node, bigger its label. This helps us to highlight nodes with high in-degree.

3. Network Measures

Let's take a look at the network measures for our twitter network.

- **Number of Nodes and Edges**

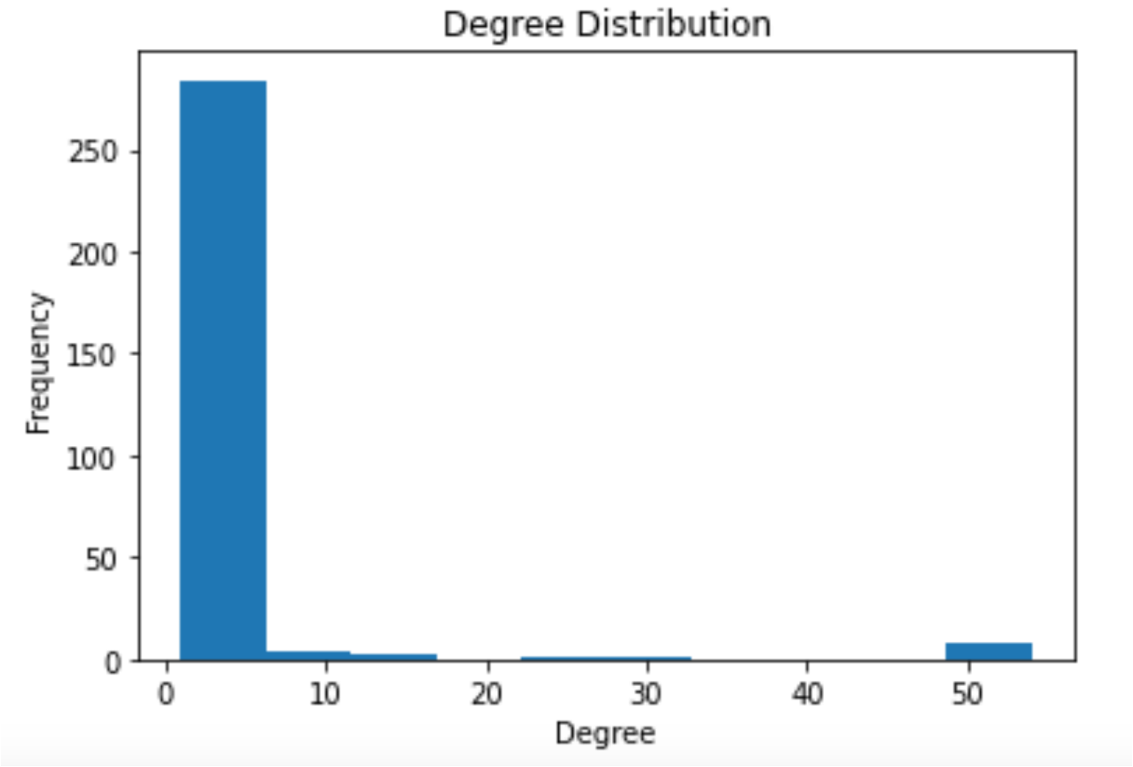
Nodes: 300

Edges: 470

- **Degree Distribution**

The degree of a node in a network is the number of connections it has to other nodes. The degree distribution of a graph is the probability distribution of these degrees over the whole network.

Average Degree: **1.567**



- **Clustering Coefficient**

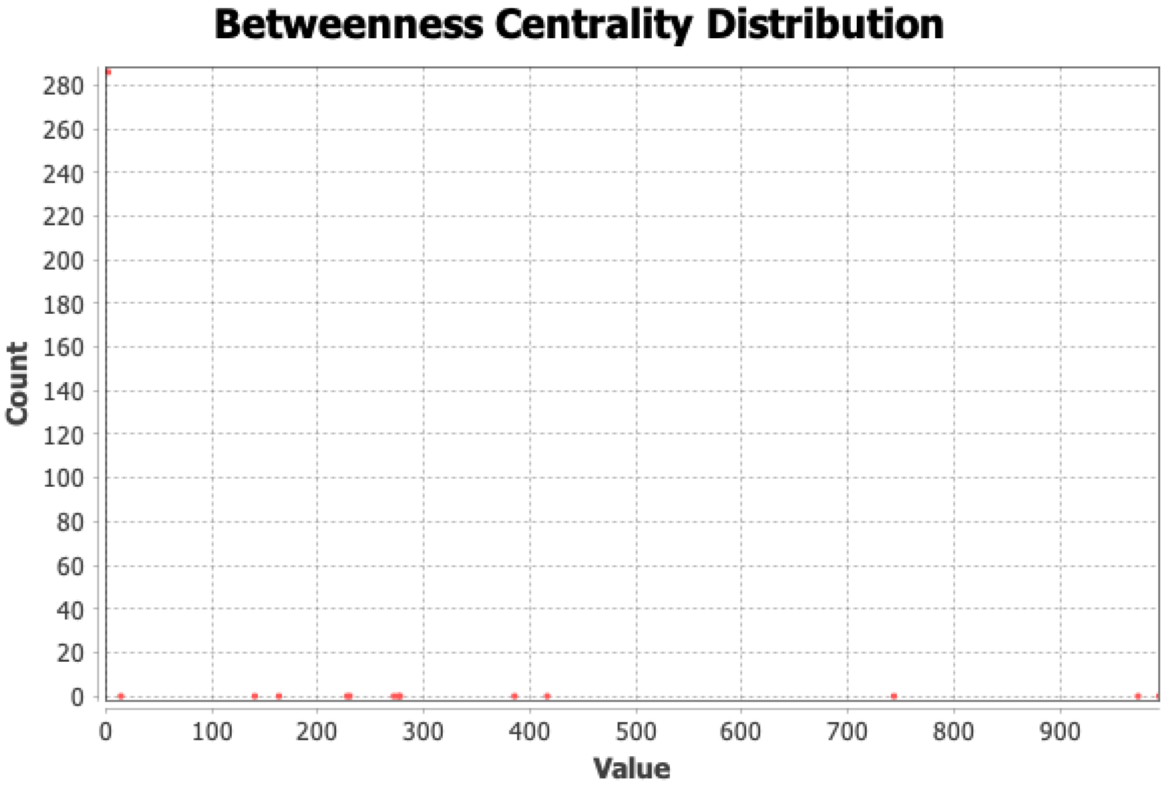
A clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together.

Average Clustering Coefficient: **0.120**

- **Betweenness centrality**

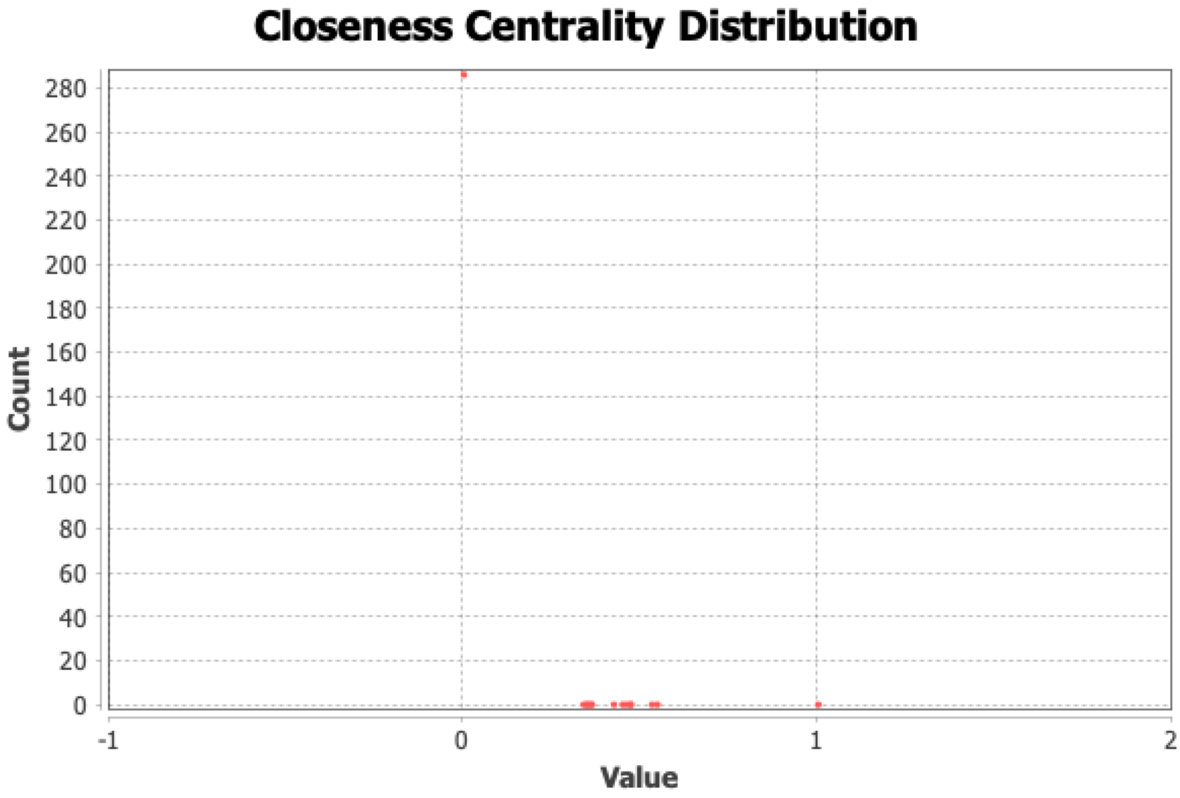
Another way of looking at centrality is by considering how important nodes are in connecting other nodes.

Average Path length: 2.41738645862357



- **Closeness centrality**

In a connected graph, closeness centrality of a node is a measure of centrality in a network. It is calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph.



Network Analysis

Following information is obtained from the data visualization and the network measures:

- Certain users are followed more by most of the Bollywood celebrities like SrBachchan, iamsrk, karanjohar, iHrithik, and juniorbachchan.
- In spite of being from the same industry, these Bollywood celebrities are following more people and pages of their individual interests.
- Each of them are following very few people in common from Bollywood itself, breaking the perception of a close-knit Bollywood community.