]

Social World Connectivity among the Indian Celebrities

Harshit Bhatt

hbhatt2014@my.fit.edu

PhD. Computer Science

Florida Institute of Technology

Research Paper on Twitter Data Mapping Project

Complex Networks (CSE 5656)

Fall 2014

Abstract

With this work I have developed a social world(twitter) relationship among the Indian celebrities and groups based on their mention in tweets. Using Twitter API I have collected tweets all around the globe concerning majority of verified Indian celebrities and groups. The data collected through Twitter API has been passed through various algorithms check to final result into a meaningful graph for various research options.

Deducting the data helps in redefining the inter network of the celebrities and gave out various facts regarding users(tweeters) affiliation to the celebrities. The networks gave some stark information about the celebrities, as we saw in different networks the major player(according betweenness centrality) were different due to the different factors on which the respective network was developed. The networks could be said biased due to various events concerning specific celebrities during tweet collection time which spiked the mention of those celebrities and their relationship.

The whole project helped me to gauge a new perspective of the Indian Celebrities in the social media which has not been thought about deeply in the past researches. This work not only concern about the celebrities but also about the mindset of tweeting crowd - about what they are more likely to tweet about and if done further research on the semantics of the tweets it could lead in defining personal relationships between specific tweeters and celebrities.

Introduction

The main purpose of devising this network was to get a gauge of the celebrities fame and their inter-network through the Social Media Data. The data here is the tweets, and only those tweets are included which mention or are related to, any of these Celebrities or groups. Only those celebrities and groups are included who are verified from Twitter.

Celebrities in the list are from the field of Indian Cinema, Cricket and Politics. Due to the popularity of the celebrities among their fan base I was able to collect a large amount of data with the help of Twitter Streaming API.

The JSON data given by Twitter API includes a wide range of information about a tweet, which helped in defining the base table for the network, through which I was able to deduce various other meaningful sub-tables and views which modeled the various graphs in my result.

Tables which mapped the various graphs were on:

- List of number of unique users for each celebrity
- List of common users between each pair of celebrities
- List of total number of each celebrity mentioned in tweets
- List of each pair of celebrities who are mentioned in the same tweet and their count
- List of every user and their mention counts to the celebrity(this list is for each celebrity)
- List containing hashtags related to celebrities
- List having common hashtags reference between the celebrities

There were three networks that were constructed with these tables and their result has been shown in the Result section of this report. The networks had a lot of edges between them which I reduced with the filters in Gephi to get a common overview of the network so only the most prominent nodes were left in the network. The nodes size was taken in accordance to their betweenness centrality which gives a more real clout of the celebrities in the network, the edge strength is similar as initially thought of - the number of relations between the celebrities more thick edge means more celebrities are related to each other. The network structure is developed through OpenOrd layout of Gephi with node color in accordance to the modularity class and community division also done with help of it.

With help of Gephi I was able to calculate degree distribution, betweenness and closeness centrality, eccentricity, density and modularity of the network whose result I have shown in the Result section for further detailed study of the networks.

Building the Twitter Network

- The first step was to set up Twitter Streaming API which was processed by getting my special token and secret key from twitter for this application.
- The python library Tweepy was used as a bridge to collect tweets from the Twitter Inc.
- The whole program is set up in the Python language, which helped to collect tweets and its other relevant information, which was further saved into MySQL database tables so as to map the networks.
- The main table for the network is named celebrity_tweets which includes the first data that is collected directly from the tweets.
- The fields are ->

  user_twitter_id -> it has the twitter_id of the person who has send that tweet.

  user_name -> user twitter handle.

  tweet -> the actual tweet which has been sent by the user.

  users_mention_id -> twitter_id of the celebrities that are mentioned in the tweet.

  hashtags -> hashtags that are in the tweet and are related to those celebrities.
- To get all (most of) the celebrities in India I first choose three most famous Indian celebrities(based on general prospective) which by their twitter handle are @SrBachchan, @iamsrk, @msdhoni.
- With the help of Twitter API I was able to collect all the Verified friends and friends of friends of these celebrities.
- Each of these friends details are saved by their respective table name having fields as :

  user_id -> celebrity twiiter id

  user_name -> celebrity twitter handle
- The same was done for the first three celebrities for their friends list and their own details.
- Now a list was made to remove the duplicates from these tables and get a unique list of celebrities which was saved in the table named celebrities.
- This list was saved to a table named celebrities which holds detail for every celebrity used in our tweet collection. It has fields as :

  celebrity_twitter_id -> celebrity twiiter id

  celebrity_twitter_name -> celebrity twitter handle

- There were total of 467 unique celebrities and verified groups which I finally got as the input for data collection.
- The program is now executed with the input of these unique celebrities.
- The data collected is saved in the celebrity_tweets table as discussed above.
- The data was collected within a range of 23 days and total tweets collected were around 352,000.
- After the data collection some relevant tables were deduced from the main table, which were respectively related to the specific networks which we have mapped to show the final results.
- Those were Six tables namely :

a. celebrity_mentions_count -> this table stores the total mentions which a celebrity received during the time period. It has fields as :

i. celebrity_twitter_id -> celebrity twitter id

ii. nodes -> celebrity twitter handle

iii. weight -> total user mentions

b. celebrity_common_mention_tweets -> this table gives the number of tweets in which the celebrities are commonly mentioned. It has fields as :

i. source -> one of the celebrities twitter handle in the pair

ii. celebrity1_id -> first celebrity twitter id

iii. target -> the other celebrity twitter handle in the pair

iv. celebrity2_id -> second celebrity twitter id

v. weight -> number of tweets having celebrities common mentioned

c. celebrity_specific_distinct_hashtags_count -> this table gives the total number of different hashtags that are connected to the celebrity. It has fields as :

i. celebrity_twitter_id -> celebrity twitter id

ii. nodes -> celebrity twitter handle

iii. weight -> total distinct hashtags connected with the celebrity

d. celebrity_hashtags_common -> this table gives the number of hashtags that are common between each pair of celebrities. It has fields as :

i. source -> one of the celebrities twitter handle in the pair

ii. celebrity1_id -> first celebrity twitter id

iii. target -> the other celebrity twitter handle in the pair

iv. celebrity2_id -> second celebrity twitter id

v. weight -> number of hashtags that are common between them

e. celebrity_number_unique_users -> this table gives the total unique users that a celebrity got mentioned by. It has fields as :

i. celebrity_twitter_id -> celebrity twitter id

ii. nodes -> celebrity twitter handle

iii. weight -> number of unique users for the celebrity

f. celebrity_users_common -> this table gives the number of users that are common between each pair of celebrities. It has fields as :

i. source -> one of the celebrities twitter handle in the pair

ii. celebrity1_id -> first celebrity twitter id

iii. target -> the other celebrity twitter handle in the pair

iv. celebrity2_id -> second celebrity twitter id

v. weight -> number of users that are common between them

- I have also saved the users details who have mentioned the celebrity in another database with tables name respective of the celebrity name containing the fields as :

  user_id -> user twitter id who has mentioned that celebrity

  user_name -> users twitter handle

  count -> number of times that celebrity has been mentioned by the user

- Three networks were constructed with the help of the above six tables. With help of igraph python library and Gephi Software we were able to map these networks. The networks are as following :

a. The first network uses celebrity_users_common and celebrity_number_unique_users tables to generate a relationship which proposes the common users between each pair of celebrities, in this network :

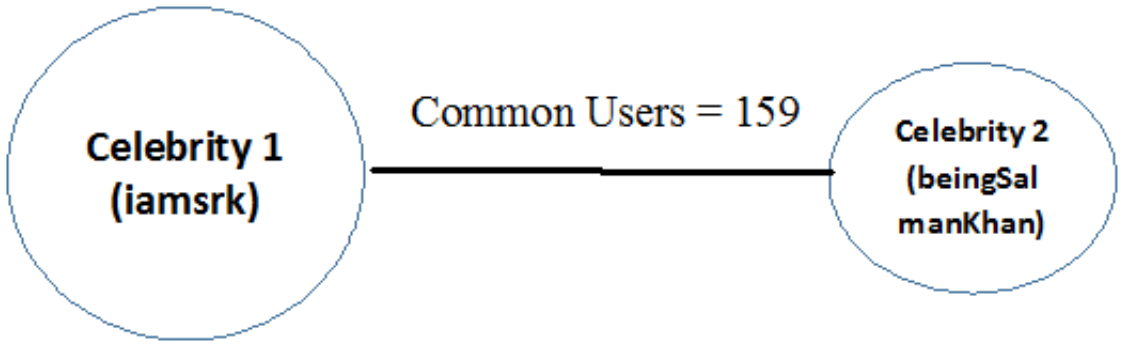
i. Nodes -> Celebrities

ii. Node Size -> Celebrities unique users count

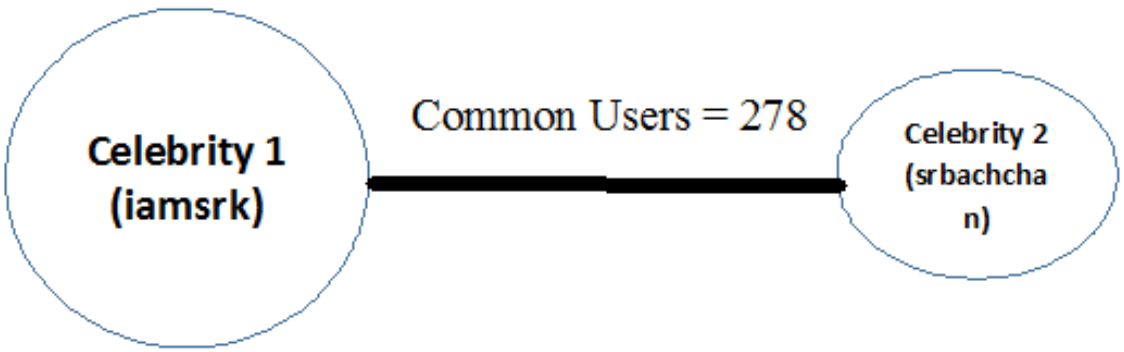
iii. Edges -> Common Users between Celebrities

iv. Edge Strength -> Common Users count

For example =>



Unique Users = 6353 **Unique Users = 852**



Unique Users = 6353 **Unique Users = 797**

b. The second network uses celebrity_mentions_count and celebrity_common_mention_tweets tables to generate a relationship which proposes how commonly two celebrities are mentioned in the same tweets, in this network :

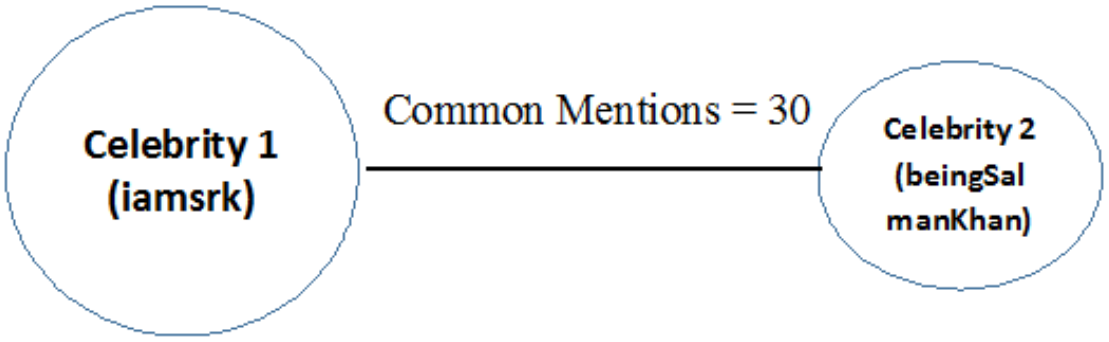
i. Nodes -> Celebrities

ii. Node Size -> Celebrities mention count

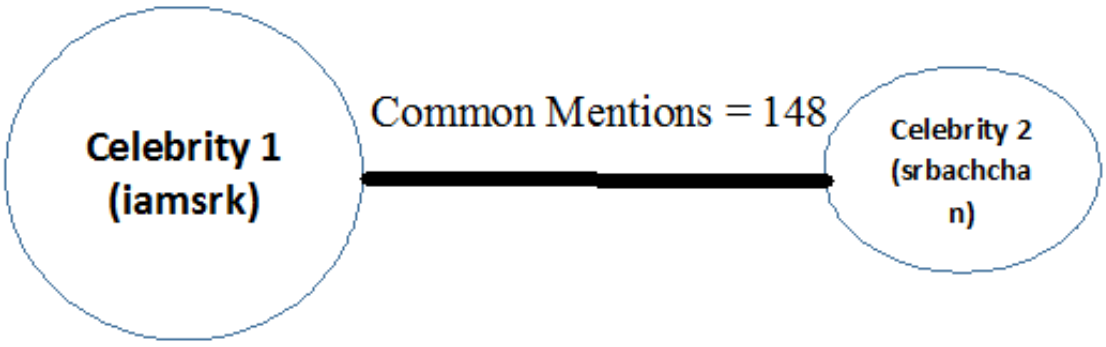
iii. Edges -> Celebrities common in a tweet

iv. Edge Strength -> count of Celebrities commonly mentioned in a Tweet

For example =>



Total Mentions = 23443 **Total Mentions = 1693**



Total Mentions = 23443 **Total Mentions = 1087**

c. The third network uses celebrity_specific_distinct_hashtags_count and celebrity_hashtags_common tables to generate a relationship which proposes how many hashtags are common between the celebrities, in this network :

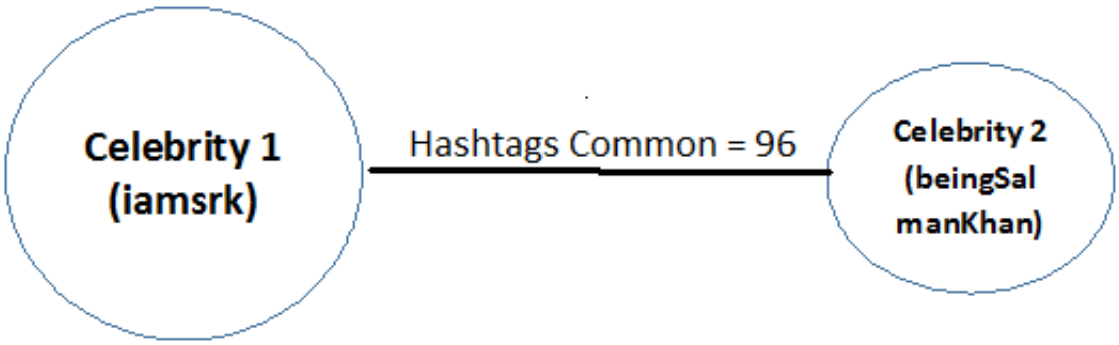
i. Nodes -> Celebrities

ii. Node Size -> Celebrities hashtags count

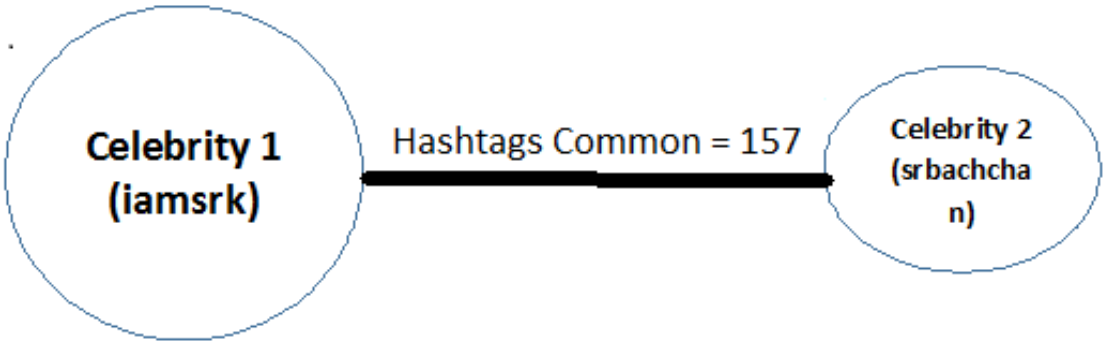
iii. Edges -> Celebrities common hashtags

iv. Edge Strength -> count of Celebrities common hashtags

For example =>



Total Mentions = 1165 **Total Mentions = 276**



Total Mentions = 1165 **Total Mentions = 376**

Related Work

Many such networks regarding the relationships among celebrities have been developed in the past by researchers. As the celebrities hold a special status in general public mind and hence their demeanor, their words affect people a lot which has made researchers to work on this field with extreme enthusiasm. We here would be talking about three main research which made me to opt for this topic.

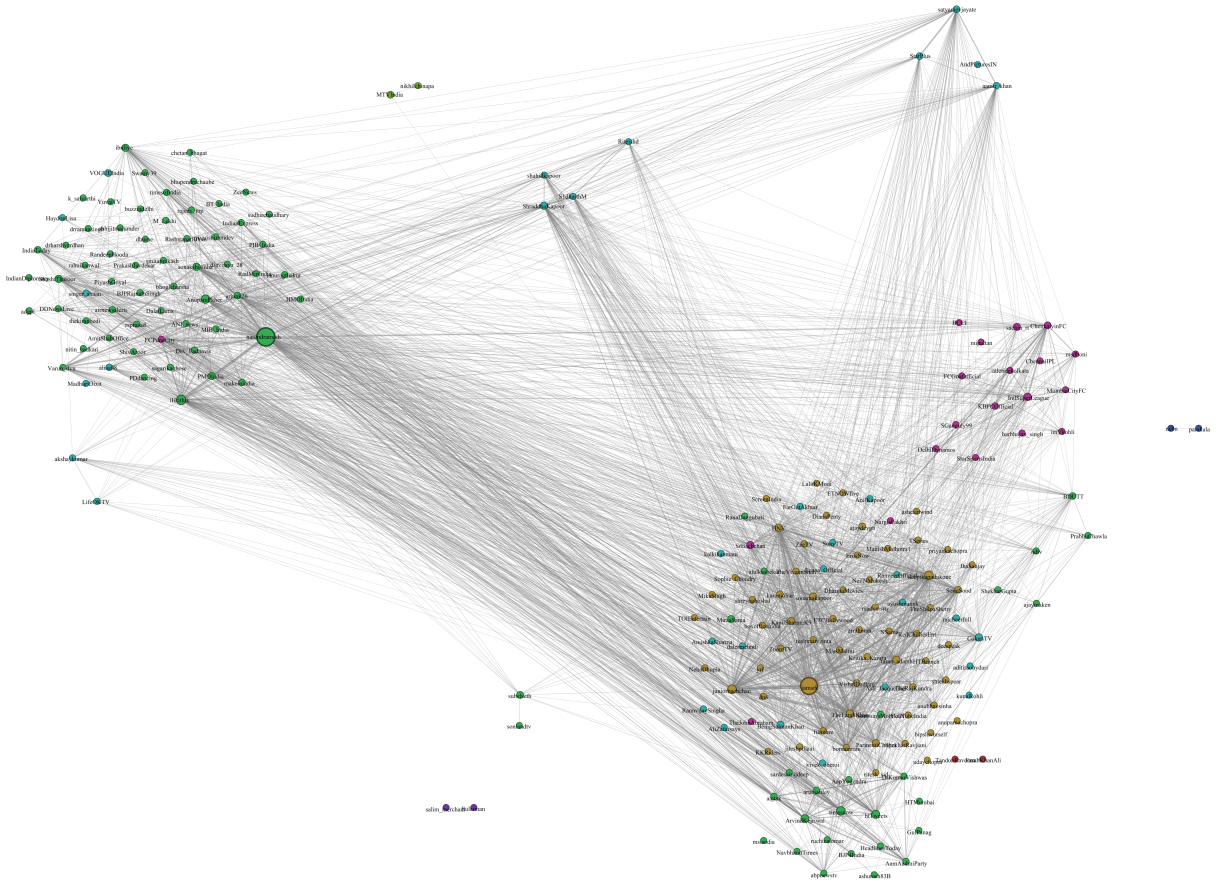
Most of the research in this field is based on the public database of IMDB which include every actor and their movies and where they have worked with each other and hence a relationship could be formed.

There has been no such research on Indian celebrities social media relationship in the past which made me to delve into this field. My main inspiration comes from a surprising source - a blog which I read while surfing on the internet of Fafadia Tech[3] where they have shown a relationship between the Indian actors on the context of those celebrities working together, the data for which they got from the IMDB databse but I tried to extend that work by taking Social Media in the context and using real world common public tweets for drawing a relation between these celebrities which could show a more genuine relationship between the Indian celebrities.

Some other work related to the affiliation between the Movies and Actors in the Hollywood was conducted by researchers[1] where they conducted the same experiment as above for the actors and their affiliation with the movies they have worked on using the IMDB database.

Another work related on the analysis and measurement of social network[2] also helped me to finalize the basic theme and initial setup of my network where researchers have shown how does the social network would be used as a genuine medium of measuring various real world entities in the near future as more people are concentrating on this medium for communicating and expressing their thoughts and ideas.

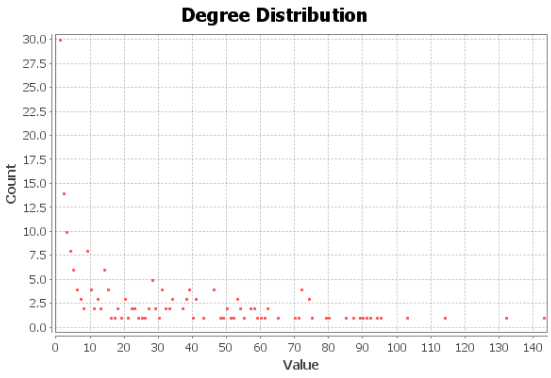
Experimental Results



Common Users between Celebrities Community Network

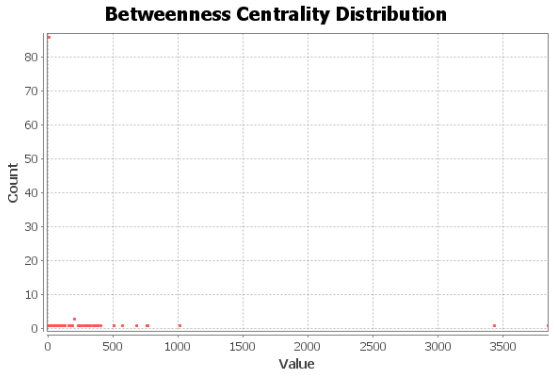
Graph for the Communities Distribution of Common Users Network

Results => Modularity: 0.351, Modularity with resolution: 0.280, Number of Communities: 8

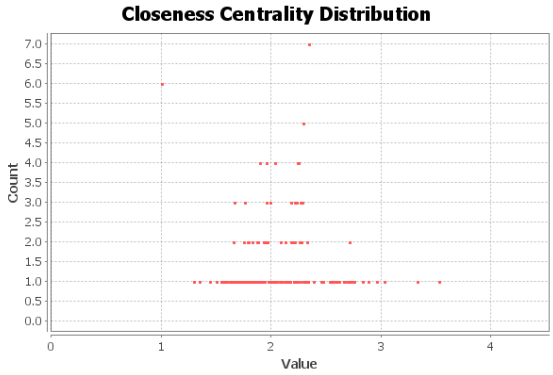


Graph for the Degree Distribution of Common Users Network

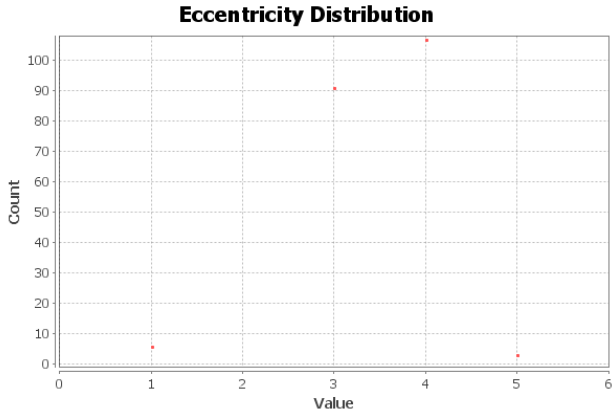
Results => Average Degree: 26.947



Graph for Betweenness Centrality Distribution of Common Users Network

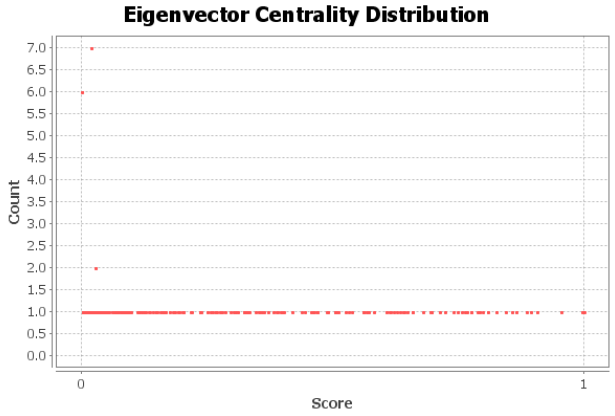


Graph for Closeness Centrality Distribution of Common Users Network



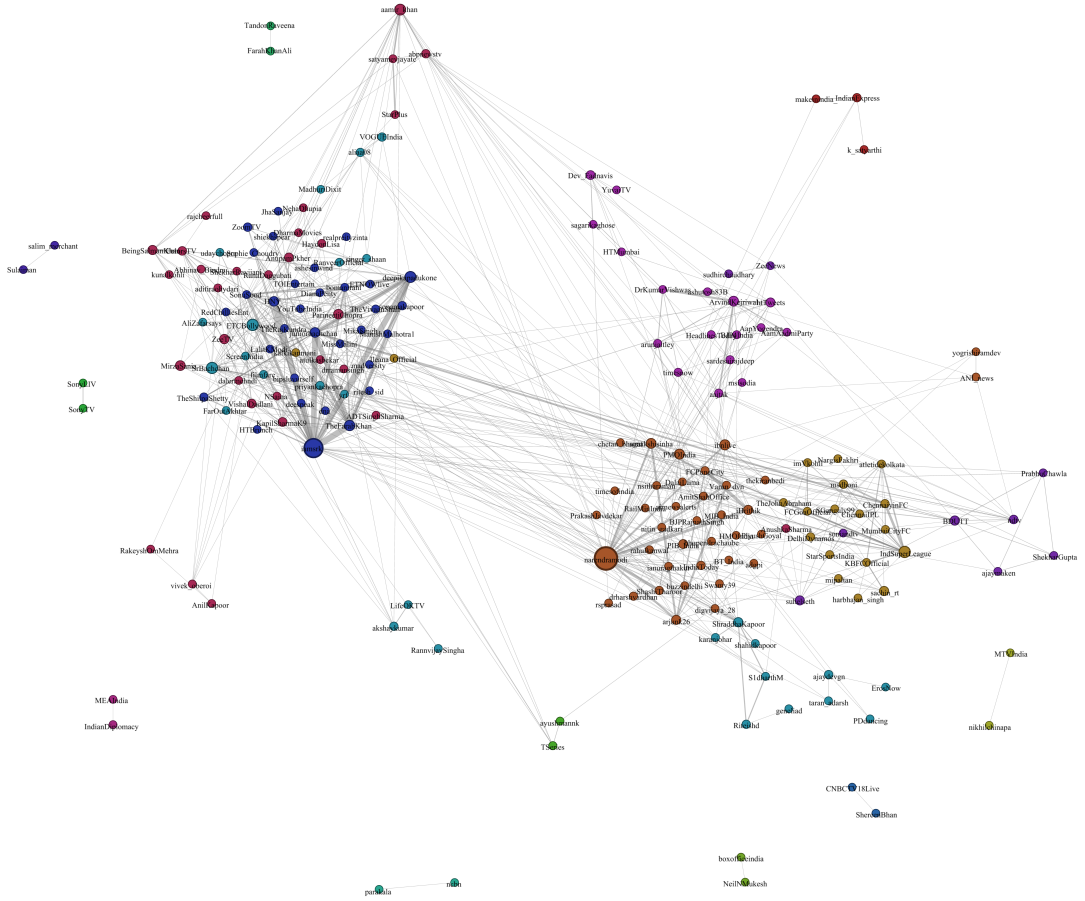
Graph for Eccentricity Distribution of Common Users Network

Results => Diameter: 5, Radius: 1, Average Path length: 2.0907, Number of shortest paths: 40206

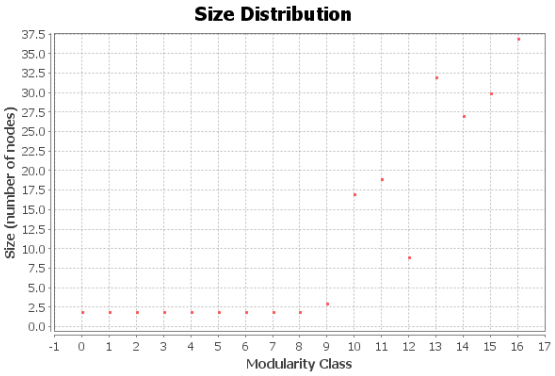


Graph for EigenVector Centrality Distribution of Common Users Network

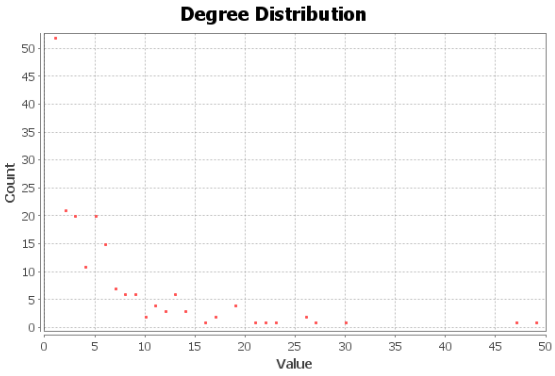
Parameters => Number of iterations: 100, Sum change: 0.006464



Celebrities mention in same tweet Community Network



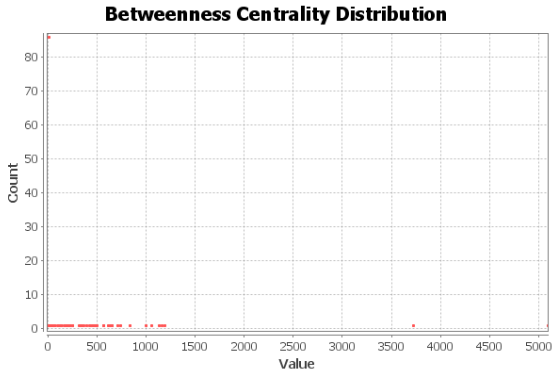
Graph for the Communities Distribution of Common Mention Network



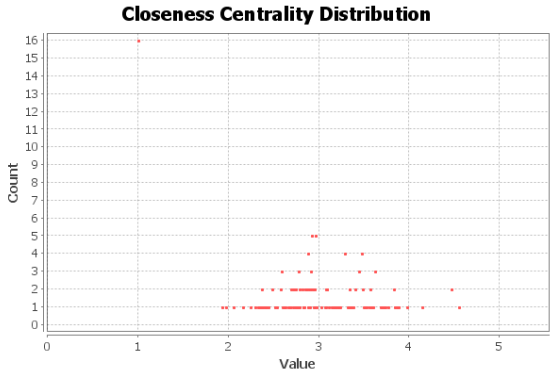
Graph for the Degree Distribution of Common Mention Network

Results => Modularity: 0.580, Modularity with resolution: 0.495, Number of Communities: 17

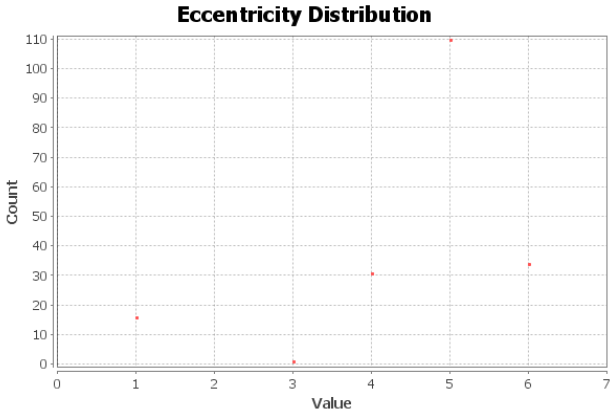
Results => Average Degree: 6.021



Graph for Betweenness Centrality Distribution of Common Mention Network

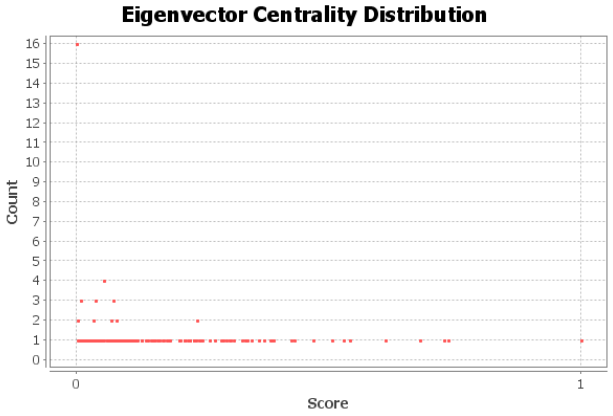


Graph for Closeness Centrality Distribution of Common Mention Network



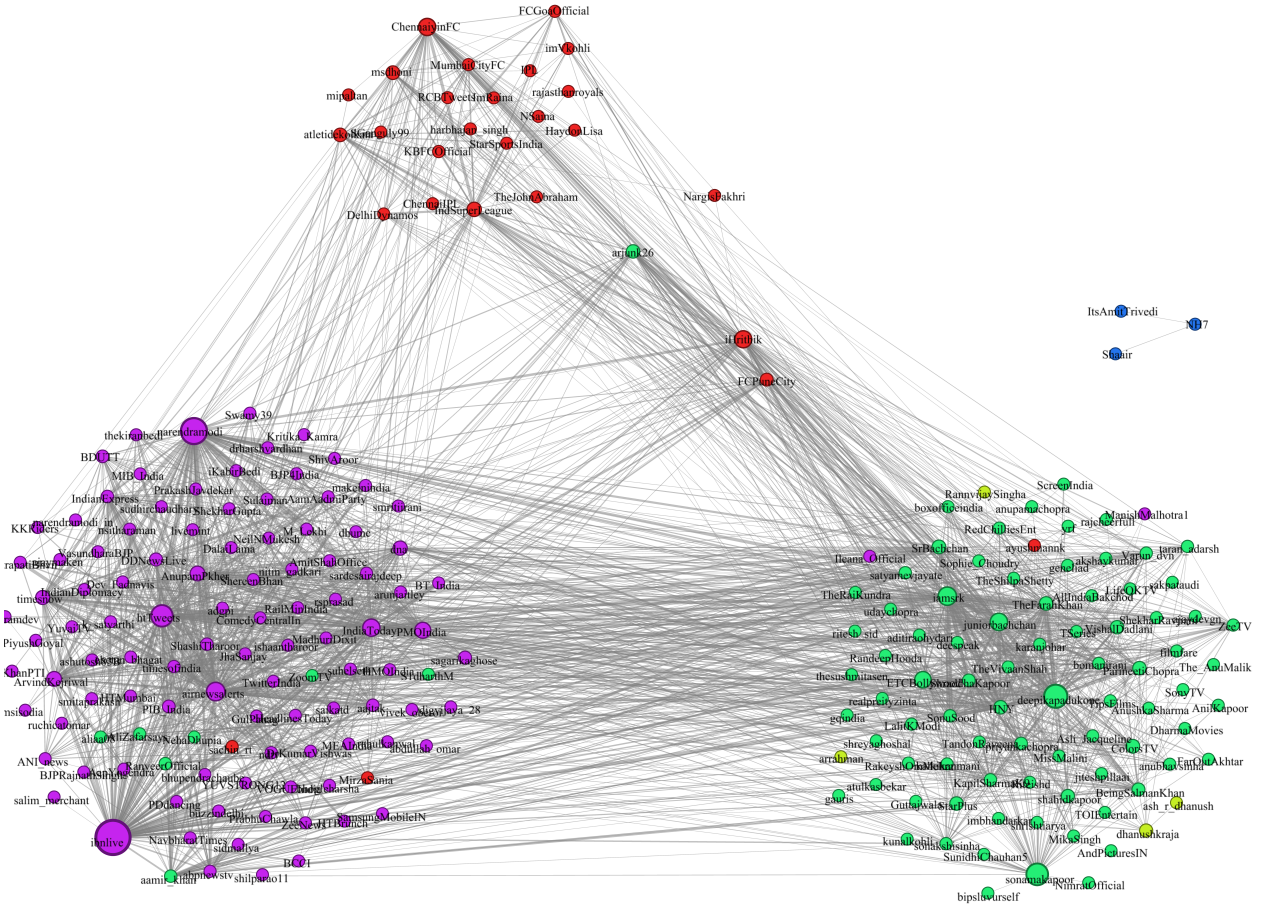
Graph for Eccentricity Distribution of Common Mention Network

Results => Diameter: 6, Radius: 1, Average Path length: 3.0195, Number of shortest paths: 30816

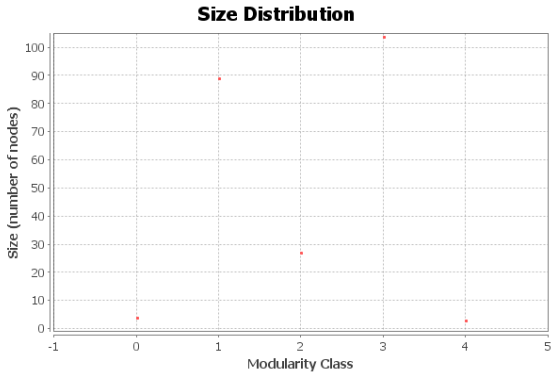


Graph for EigenVector Centrality Distribution of Common Mention Network

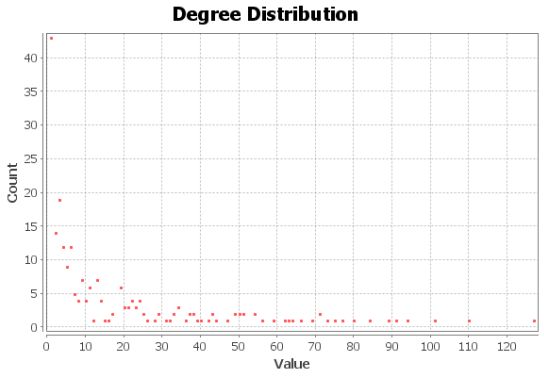
Parameters => Number of iterations: 100, Sum change: 0.007861



Common Hashtags between Celebrities Community Network



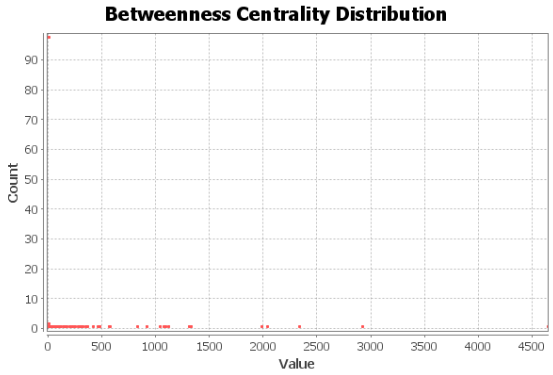
Graph for the Communities Distribution of Common Hashtags Network



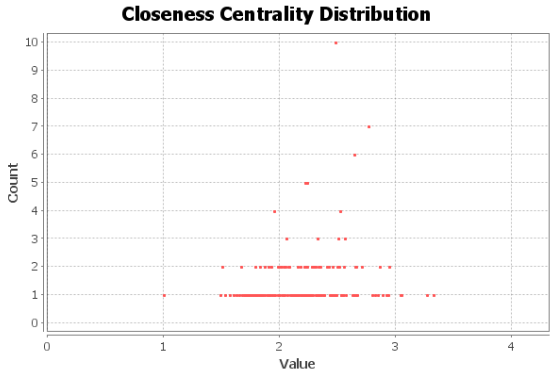
Graph for the Degree Distribution of Common Hashtags Network

Results => Modularity: 0.315, Modularity with resolution: 0.243, Number of Communities: 5

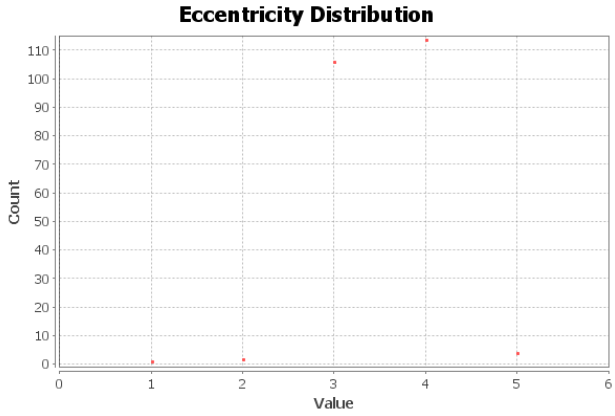
Results => Average Degree: 18.106



Graph for Betweenness Centrality Distribution of Common Hashtags Network

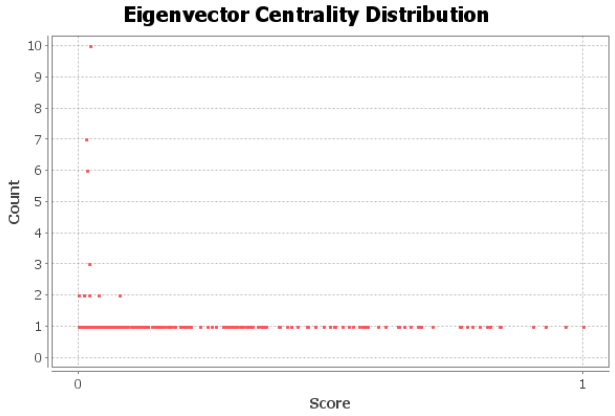


Graph for Closeness Centrality Distribution of Common Hashtags Network



Graph for Eccentricity Distribution of Hashtags Common Network

Results => Diameter: 5, Radius: 1, Average Path length: 2.2723, Number of shortest paths: 49958



Graph for EigenVector Centrality Distribution of Common Hashtags Network

Parameters => Number of iterations: 100, Sum change: 0.004076

Conclusion

After analyzing the three networks we get to know about the strength(count) of the users mention for the celebrities and who are the celebrities which tweeters consider most related to each other and to themselves.

We also get to know the users who tweet most about celebrities and whether or not they are loyal to one celebrity or many.

My study shows that as in accordance to the general perspective the social media also shows the same relationship among the celebrities, making the celebrities from the same community(politics, cinema and cricket) to be related more to each other than with other community and the users being more common between those community celebrities. Surely there are some exceptions but they are overshadowed by the strong relation between the celebrities in single community.

For the common users network there were 8 different communities with modularity of 0.351, in this network two clusters were prominent which constitute mostly actors with their related entertainment groups(brown) and politicians with news channels and Government organizations in other(green), the third celebrity group of cricketers were mixed in both these groups and were less in number as compare to other two groups. Surely some communities such as(violet) has some mix of celebrities such as cricketers and actors which has Indian Super League(@IndSuperLeague)(soccer event) as the central group where these celebrities are team owners and other are participating teams twitter handle. Other smaller isolated groups(green,blue) are also connected to each other with their current involvement specific to each other. Within this network two nodes standout due to their size (@narendramodi and @iamsrk) due to their highest betweenness centrality in the two respective clusters within the network which gives us the conclusion that these two celebrities are most commonly mentioned by the users and all other celebrities withn that cluster have users common with these two.

For the common mention tweets the network is more spatial and diverse with 17 communities and modularity of 0.580, even here there are two major clusters with same two celebrities(@narendramodi and @iamsrk) having main clout due to their high betweenness centrality, but here due to large number of communities there are some other relevant groups like(violet) of @aamaadmiparty and its members which proposes that these members are more mentioned with each other in a tweet than with others. A single community member(@suhelseth)(violet) strikes out prominently due to his common mention with many other celebrities which is a surprise element in this network. Other isolated communities are also present which is due to their common mention with each other prominently.

The third network on the common hashtags between celebrities has 5 communities with modularity 0.315, is more concentrated and could be easily seen to be divided in three major clusters. This network is vastly different from the other two networks as here the three main celebrities due to their betweenness centrality are(@ibnlive, @deepikapadukone and @sonamkapoor) which

is different from the previous one's and are quite surprising which holds out that these celebrities are mostly mentioned with the hashtags concerning with other celebrities as well as most are connected to them. Here a third community(red) is also prominent cluster which contains more of sport celebrities connected with each other due to Indian Super League(@IndSuperLeague) event with stark inclusion of some actors which are not part of that event which makes us believe that these celebrities when mentioned are affiliated with these hashtags.

Some celebrities or organizations have a very large number of mentions than others which may be due to those celebrities(politicians, cricketers and movie stars) being trending in India for some cause (movie release, an event being taking place or due to their certain initiatives) examples of which could be given with @iamsrk, @deepikapadukone movie stars mentioned a lot due to their movie released within this time period or @IndSupLeague an organization which is conducting a football event within India during this time or @narendramodi India Prime Minister for starting a cleanliness drive campaign within the country.

The common mention between some celebrities also has some great numbers like for @iamsrk, @deepikapadukone, @juniorbachchan, @farahkhan are mentioned a great number of times with each other due to a movie release which had all of them working together. The same was for the teams like @FCGoa, @atlidkolkata, @PuneFC and others due to their involvement in a league.

Taking all three networks in consideration finally I could say that one celebrity which stands out in the whole network as a whole would be(@narendramodi) India's Prime Minister as he has been the most prominent player in all three networks which comes out as a great signal for the India country as a whole as most of the celebrities and tweeting community could relate to the supreme power of the country which should bring Governance to a smooth start and portends bright future for this country prospects as people are more united and related with the works of Government.

Finally I could say that if popularity could be gauged by the mention of celebrities it would hold true to what general perspective is. Still as a caveat before considering this study as the real truth among celebrities relationship is that the study is solely based on the social media mention of celebrities and has been done for a short period of time if the research could be extended to a different level it could be validated to some extent with the real world.

Reference

[1] Movies and Actors: Mapping the Internet Movie Database - Bruce W. Herr , Katy Borner

[2] Measurement and Analysis of Online Social Networks - Alan Mislove, Peter Druschel

[3] Fafadia Tech Blog