Social World Connectivity among the Indian Celebrities

Harshit Bhatt

Florida Institute of Technology, Melbourne 32901.
Florida, USA.
{hbhatt2014}@my.fit.edu
http://www.fit.edu/~hbhatt2014

Abstract. This work maps social world(twitter) relationship among the verified Indian celebrities and groups based on their mention in tweets. Extracting the data helped in redefining the Inter-Network of the celebrities and gave various facts regarding users(tweeters) affiliation to the celebrities. 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.

Keywords: Twitter API, Social Media, Relationship between Users and Celebrities

1 Introduction

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

Celebrities in the list are from the field of Indian Cinema, Sports 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 which modeled the various networks 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 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 as it gives a more real dominance of the nodes in the network, the edge strength is the number of relations between the celebrities, more thick edge means more those celebrities are related to each other. Each node is also provided with its Meta Data which include celebrities twitter handle. The network structure is developed through OpenOrd layout of Gephi with node color and community division done in accordance to the modularity class.

With help of Gephi I was able to calculate Degree distribution, Betweenness and Closeness centrality, Eccentricity and Modularity of the network whose result I have shown in the Result section for further detailed study of the networks.

2 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/initial table for the network is named celebrity_tweets which includes the data that is collected directly from the tweets.

- Its 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 four most famous Indian celebrities(based on general prospective) which by their twitter handle are @SrBachchan, @iamsrk, @msdhoni, @pmoindia.
- With the help of Twitter Streaming API I was able to collect all the Verified friends and friends of friends of these four celebrities.
- Friends details of these four celebrities and their friends were saved in their respective named tables(eg. iamsrk_friends) with table having fields as :
  - user_id -> celebrity twitter id
  - user_name -> celebrity twitter handle
- Now a list was made to remove the duplicates from these tables and get a unique list of 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 twitter 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.
- Only those tweets were collected which have either hashtags with celebrities mentioned or only celebrities mentioned or tweets from these celebrities having hashtags.
- The data was collected within a range of 23 days and total tweets collected were 361,296.
- 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.
- There were Six tables deduced 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

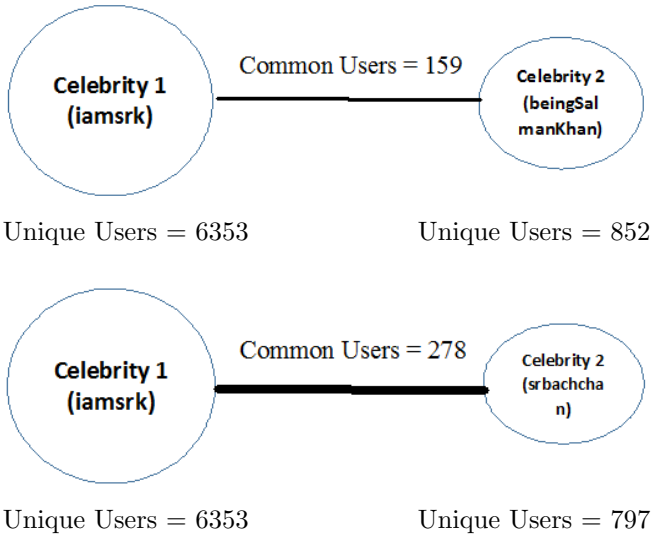
iv. celebrity2_id -> second celebrity twitter id
v. weight -> number of users that are common between them

- I have also saved the users details, it contains information on twitter users who have mentioned the celebrities in another database with tables name as 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 I was 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 -> Betweenness Centrality of the node
iii. Edges -> Common Users between Celebrities
iv. Edge Strength -> Common Users count

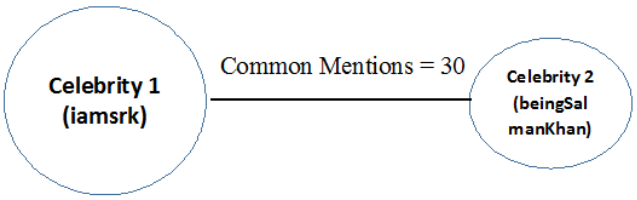
For example =>



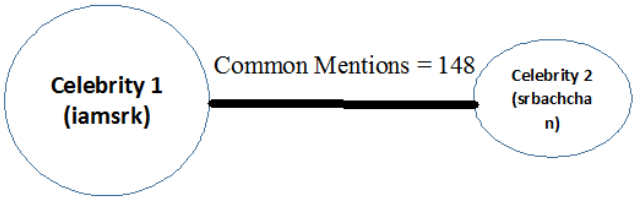
  - b. The second network uses celebrity_mentions_count and celebrity_common_mention_tweets tables to generate a relationship which proposes how commonly two celebrities are mentioned in the same tweets, in this network :

i. Nodes -> Celebrities
ii. Node Size -> Betweenness Centrality of the node
iii. Edges -> Celebrities common in tweets
iv. Edge Strength -> count of Celebrities common mentions

For example =>



Total Mentions = 23443 Total Mentions = 1693



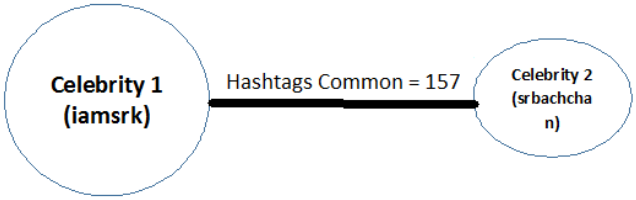
Total Mentions = 23443 Total Mentions = 1087

– 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 -> Betweenness Centrality of the node
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

– For better viewing these networks were exported to GraphML format and were then viewed in Gephi, to divide the network into clusters OpenOrd layout was chosen, to further divide them into communities Modularity algorithm of Gephi was used, the nodes were colored with their respective modularity class color which made it easy to distinguish the nodes.
– The networks were highly cluttered due to large number of edges(2216,21978,22188), hence to make them more viewable I applied Gephi filters of degree range and edge weight range which choose edges above certain weight(20,20,15) and nodes having degree 1 or more, this made some nodes to disappear but still the network gives a generalization of the relationships.

3 Related Work

Many such networks regarding the relationships among celebrities have been developed in the past by researchers. As the celebrities hold special status in general public mind, hence their demeanor, their activities, their words affect people deeply, which made researchers to work on this field as it encloses within it vast opportunities. Hence, there has been many research′s regarding this relationship but I would be here concentrating upon three main research′s 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 actors and their movies and when they have worked with each other, which helps to construct a relationship between the celebrities. The three networks were :

3.1 Fafadia Tech Blog

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 in the year 2013, the data for which they got from the IMDb public database 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.

3.2 Movies and Actors in Hollywood

Another work related to the affiliation between the Movies and Actors in Hollywood was done by Herr, Ke, Hardy, Borner in 2007 [1] where they conducted the same experiment as above for the actors and their affiliation with the movies they have worked on, using the information from IMDb public database. They focused their research on the movies produced between 1890 to 2007 with co-actors network following it. They finally superimposed both these networks on top of each other to get a final network which includes movie name, when it was made and actors who worked in it. The co-actor network includes the relationship among all the actors who have worked with each other in these movies, they nodes are actors with edges showing the strength of their relationship which tells about how many movies they have worked in with each other. The final network had movies distinguished based on their genre with actors receiving the color depending on the genre of movies they have most worked in.

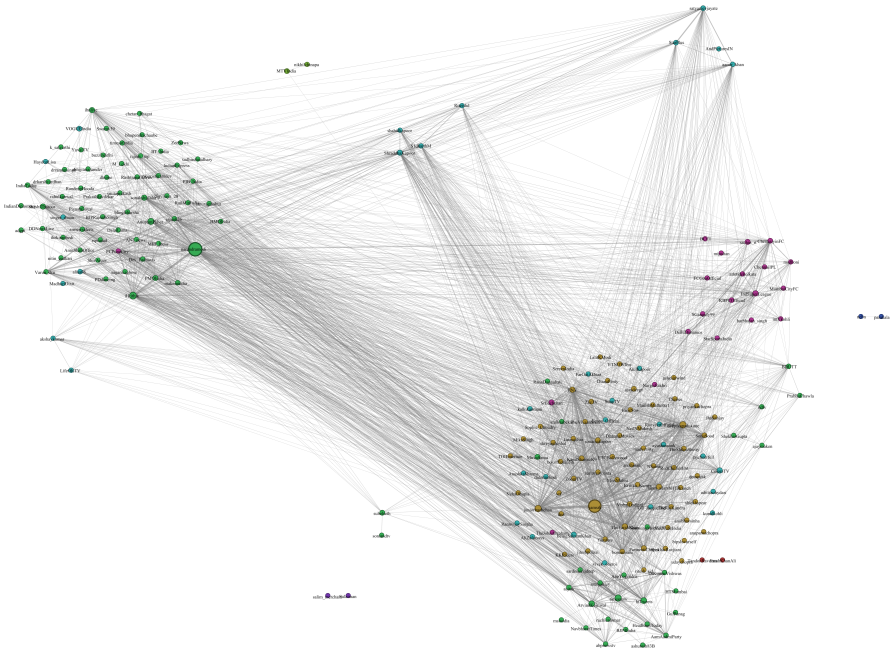
3.3 Analysis and Measurement of Social Media

One other work related to the analysis and measurement of social network [2] also helped me to finalize the basic theme and initial setup of my network where Mislove, Marcon, Gummadi, Druschel, Bhattacharjee 2007 have shown how does the social network has been and could be used as a genuine medium for measuring various real world entities in the near future as more and more people are now concentrating on this medium for communicating and expressing their thoughts and ideas. Their work was done between the year 2005-2007 where they concentrated their research on four of the top social networking sites at that time(Flickr,YouTube,Orkut,Live Journal). Their study was vast with around 11.3 million users having 328 million links. Their research tried to find the similar interest among the users and the resources that have been mostly shared by them. By getting to know about this information they were able to deduce the basic structure of Social Media and how it works. They also predicted social

media functionality in the future, they proposed that as social media continues to grow it would account for the majority of Web traffic and as more people would be joining this network it becomes important to know about their behavior, which in turn could help to provide them a more feasible and related social media structure and this study could also help in improving security of the networks as well as the internet.

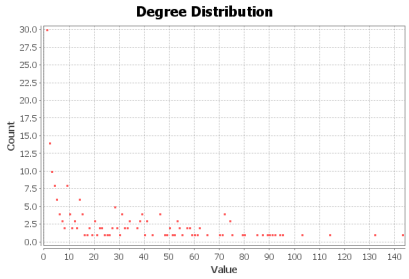
I would like to extend this study to include the sentimental analysis of tweets, which would help to get the idea on how does the general public consider the celebrities. This metric could help in deciding the celebrity status and give us the fact that whether their image is positive or negative in general public mind. Similarly I could also have made a bipartite network between celebrities - twitter users and celebrities - hashtags to get more specific information on this relationship but due to time constraint I have skipped these options for future.

4 Experimental Results



Common Users between Celebrities Community Network

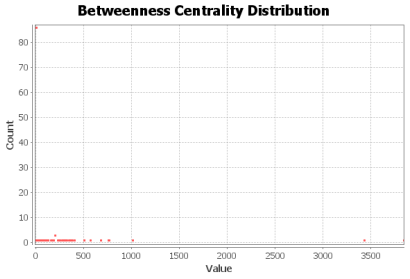
Graph for the Communities Distribution of Common Users Network



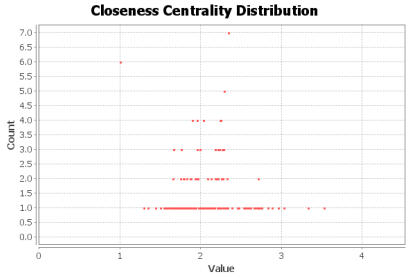
Graph for the Degree Distribution of Common Users Network

Results => Modularity: 0.351, Number of Communities: 8

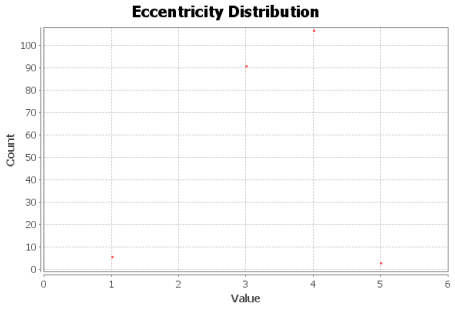
Results => Average Degree: 26.947



Graph for Betweenness Centrality Distribution of Common Users Network

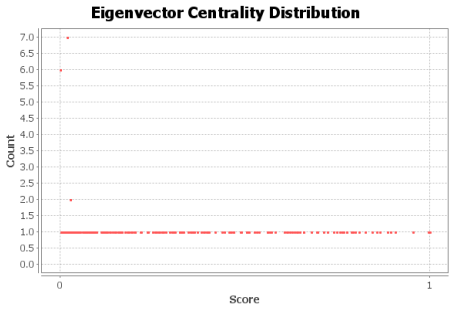


Graph for Closeness Centrality Distribution of Common Users Network



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

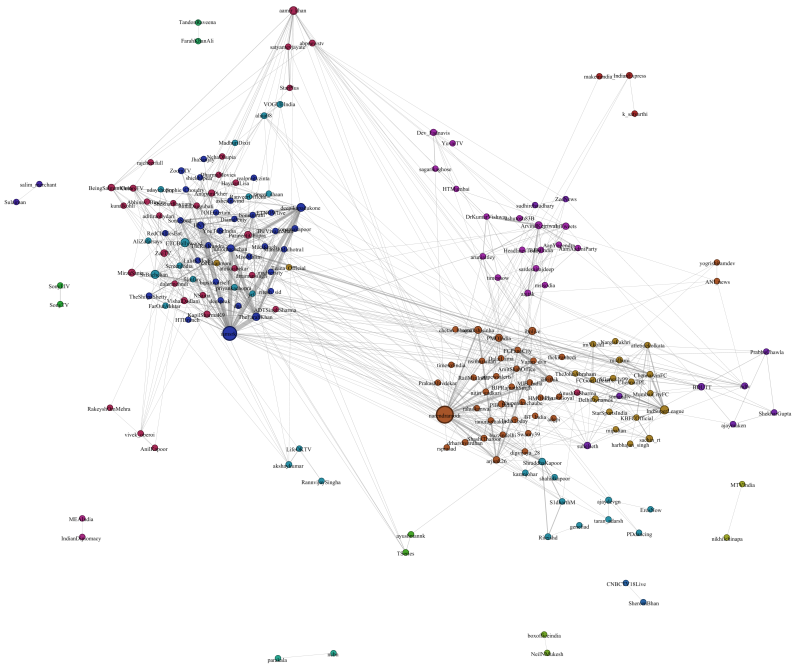


Fig. 1. Celebrities mention in same tweet Community Network

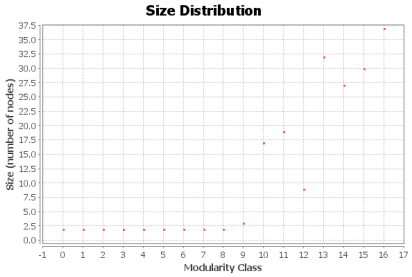


Fig. 2. Graph for the Communities Distribution of Common Mention Network

Results => Modularity: 0.580, Number of Communities: 17

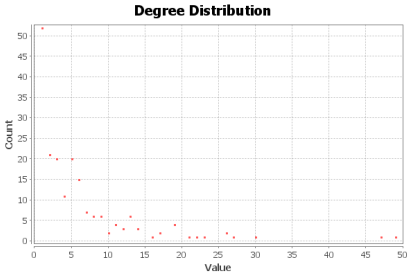


Fig. 3. Graph for the Degree Distribution of Common Mention Network

Results => Average Degree: 6.021

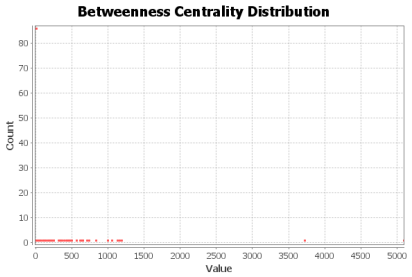


Fig. 4. Graph for Betweenness Centrality Distribution of Common Mention Network

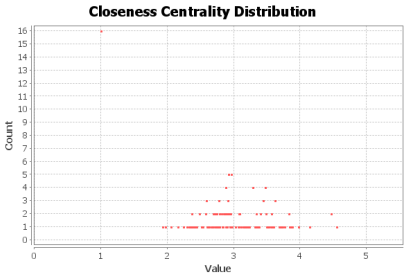


Fig. 5. Graph for Closeness Centrality Distribution of Common Mention Network

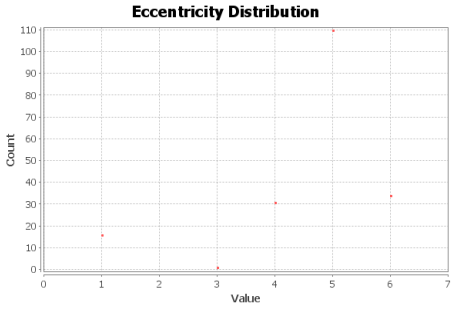


Fig. 6. Graph for Eccentricity Distribution of Common Mention Network

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

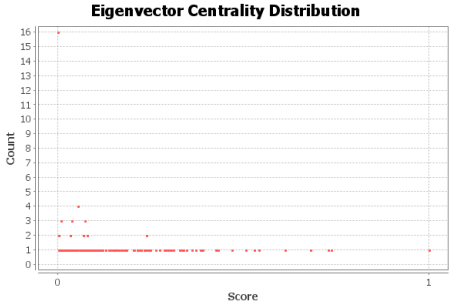


Fig. 7. Graph for EigenVector Centrality Distribution of Common Mention Network

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

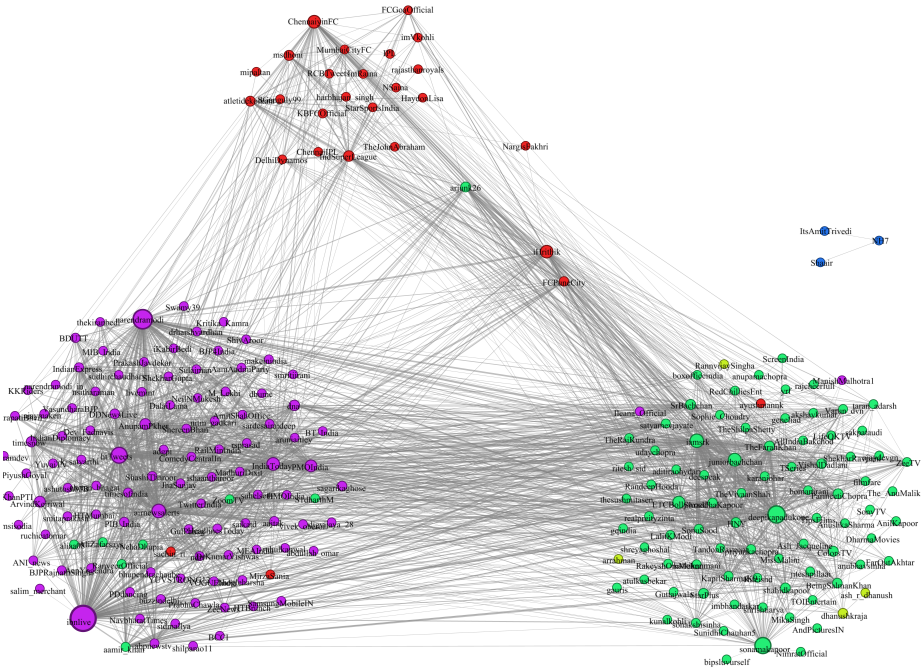


Fig. 8. Common Hashtags between Celebrities Community Network

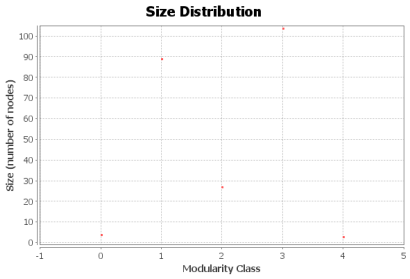


Fig. 9. Graph for the Communities Distribution of Common Hashtags Network

Results => Modularity: 0.315, Number of Communities: 5

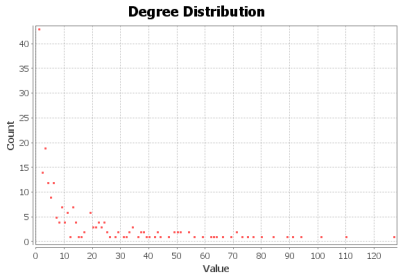


Fig. 10. Graph for the Degree Distribution of Common Hashtags Network

Results => Average Degree: 18.106

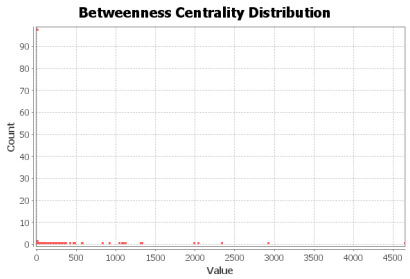


Fig. 11. Graph for Betweenness Centrality Distribution of Common Hashtags Network

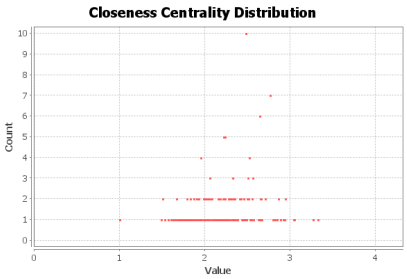


Fig. 12. Graph for Closeness Centrality Distribution of Common Hashtags Network

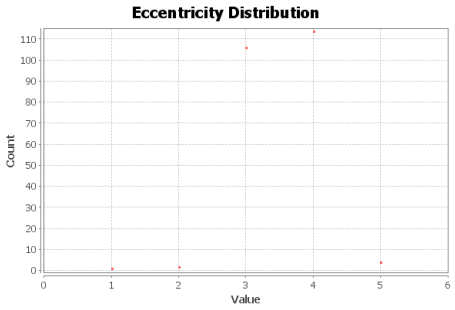


Fig. 13. Graph for Eccentricity Distribution of Hashtags Common Network

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

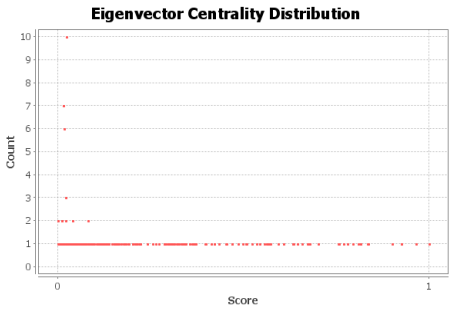


Fig. 14. Graph for EigenVector Centrality Distribution of Common Hashtags Network

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

5 Conclusion

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

We also get to know from the collected data about the users who tweet most on 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 sport) 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 sport stars 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 sport stars and actors, this community has Indian Super League(@IndSuperLeague)(soccer event) as the central node, and the celebrities are team owners with other nodes as participating teams. Other smaller isolated clusters(green,blue) are also connected to each other with their current involvement specific to each other. Within this network two nodes stood out(@narendramodi and @iamsrk) due to their highest Betweenness centrality in the two respective big clusters which gives us the conclusion that these two celebrities are most mentioned by the users and most of the other celebrities within those clusters have users common with these two. There were some movie stars present in the politicians cluster with strong relationship, which proves the goal for designing this network as we could deduce that some of the tweeting population takes interest in both politics and movies simultaneously.

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 suggests 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 the celebrities in them being commonly mention with each other prominently, due to which we could see the large number of communities. This network could be considered as explaining the real nature of associativity between the celebrities. This network largely differs from the one above, and hence we could conclude that although a twitter user tweets mentioning celebrities from different genres but only those celebrities are mentioned with each other in a tweet by users who either work with each other or share some common events. Another striking discovery in this network is that @narendramodi Indian Prime minister is more commonly mentioned with some Movie stars than his colleagues. Another feature of the network is that the connection among the movie stars is far more stronger than the politicians.

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, with most of the other celebrities 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 inclusion of some actors as well which shows that these actors have more hashtags common with the sport stars than their own group of people, which holds the theory true that celebrities are connected to each other not merely on their field of work but due to the events they are involved in.

Some celebrities or organizations have a very large number of mentions than others which may be due to those celebrities(politicians, sport 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 the celebrity which stood out in the whole network is(@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 India as this means most of the celebrities and tweeting community relate to the supreme power of the country which should help in smooth and fair Governance as it has the inclusion of most of the public, it also provide great prospects for India as people are more united and relate to 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.

References

1. Movies and Actors : Mapping the Internet Movie Database - Bruce W. Herr, Katy Borner, Weimao Ke, Elisha Hardy. 2007
2. Measurement and Analysis of Online Social Networks - Alan Mislove, Peter Druschel, Massimiliano Marcon, Krishna P. Gummadi, Bobby Bhattacharjee. 2007
3. Fafadia Tech Blog. 2014