SOCIAL NETWORK ANALYSIS (SNA) TO ANALYSE BOLLYWOOD MOVIES

Sabdick Roy Chowdhury and Kousik Dasgupta[1]

Department of Computer Science and Engineering, Kalyani Government Engineering College,
Kalyani- 741235, India
[1]Email: kousik.dasgupta@gmail.com

Abstract: In today's world Social Network is the mostly used media for communication among the people. The service provided by social network allows people to use and engage with the Internet and with each other. Assessment of these interactions allows critical analysis of trends or opinion. Assessment requires analysis. That is the main reason for the popularity of Social Network Analysis. Social network analysis has undergone a renaissance with the ubiquity and quantity of content from social media, web pages, and sensors. Social network analysis has been a popular research area recently thanks to the availability of many large dataset. The main problem in social network analysis is to recover the underlying network structure given information about how nodes in the network interact with each other in the past. Graph analytics have proven to be valuable tools in Social Network Analysis. This work introduces a novel approach that uses social network analysis for study and analysis of the different stakeholders of Bollywood movie industry. The Indian movie industry is a big business prospect of India. Even the movies released in one year is much higher that any other film industry. Some of the stake holders of the movie industry like stars, directors, music directors are not only renowned in society but also financially prospective. To test this approach, experiments have been conducted on Internet movie database (Imdb) of a particular era. The results obtained from the analysis are encouraging for predicting future trends of the different stake holders of Bollywood movies.

Keywords: Social network analysis; movies; network; cluster coefficient; success of movies.

I. INTRODUCTION

Social Network Analysis provides a mean for interpreting and measuring relationships between a number of social entities, such as people, groups and organisation's. The emphasis on relationships [1] is an important supplement to standard social and behavioral research, which is mainly concerned with attributes of the social entities. In Social Network Analysis the attributes of the individual actors are not essential; the focus is on the structure of their relationships and how the structure of linkages affects individual actors and their relationships. The structure of a network provides insights into network activities and how knowledge is generated and shared within the network. Figure 1 shows the network of Facebook.

An efficient way for scientific researchers to exchange and bring forth knowledge is through collaboration in specialist organisations such as the European Marketing Academy. The structure in networks like this is often hidden because of its informal network characteristics.

This paper introduces a novel approach that uses the basic concept of social network analysis for analysing the effects Bollywood movie industry on the different stake holders. It can be further extended to forecast pre-production trends. Which can be further helpful to upcoming stars of movie industry.

Movie is an integral part of one's daily life. In this nation, it is more than an industry. People are investing money for making a film, which will attract audience. More viewers mean more gross income. So it is important to identify the viewer's choices for a film. Movie Rating Analysis will definitely serve this issue. The aim of this paper

is to analyse the past movies dataset and from that analysis the viability of the upcoming movies through rating system. Various stakeholders such as actors, financiers, directors etc. can use these analyses to make more informed decisions. There are some parameters for social network analysis like Cluster coefficient, Degree distribution. In the latter part of this paper discussion on those topics will be done. There are some other attributes as parameters for analysing. Such as:

- Actor

- Actress

- Director

- Release Time

- Type of movie

- Duration

- Ratings

- Box Office Verdict

By analysing the network through these parameters one will able to find some analytical data regarding this network.

The remainder of this paper is organized as follows. Section 2 surveys current research on movies and the influence of the given parameters on it. Section 3 develops a suggestive proposal of the research. Sections 4 reports finding of the experiments done on movie database. Section 5 gives conclusion and future work.

2. LITERATURE SURVEY

In recent years, there has been an increasing amount of literature on Social Network Analysis. This large and growing body of literature [2-3] has among other things, applied Social Network Analysis as a tool for conducting citation analyses. A large volume of published studies employ Social Network Analysis to reveal patterns and regularities in the way in which scholars work together by focusing on co-authorship in published research papers [4-5].

Social Network Analysis (SNA) has been used for analysis of movies and movie ratings. To achieve this [6] had proposed a Character-net that can represent the relationships between characters using dialogs, and a method that can extract the sequences via clustering communities composed of characters. Weng et al. [7] proposes use of social network analysis to elaborately identify leading roles and the hidden communities. The community identification has been used to perform storyline detection that facilitates more flexible movie browsing and higher-level movie analysis. Doshia et al. [8] proposes the effectiveness of social network analysis and sentiment analysis in predicting trends. It predicts the success of new movies over their first four weeks in the box office after opening. In reference [9], use of web mining approach that combines social network analysis and automatic sentiment analysis has been detailed for predicting trends and real world events in the movie business.
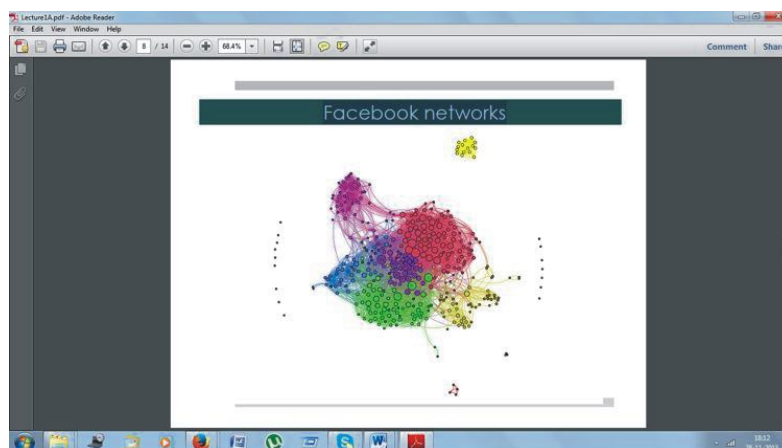
Fig. 1 Facebook Social Network

3. BACKGROUND OF COMPLEX NETWORK

Social Network Analysis pertains to the comprehension of the linkages among social entities and the implications of these linkages. As a tool Social Network Analysis has its origin in the social sciences. In this section a basic overview of network and network analysis pertaining to social network has been given.

A. Network

A networkis a set of nodes connected by a set of ties. The nodes can be anything persons/ individuals, teams, organisations, concepts, patents, etc. In the case of social networks the nodes are individuals. Networks, which are only made of one type of nodes, are homogeneous, they are heterogeneous otherwise. Whereas ties connect pairs of nodes and can be directed (i.e., potentially one-directional, as in giving advice to someone) or undirected (as in being physically proximate) and can be dichotomous (present or absent, as in whether two people are friends or not) or weighted (measured on a scale, as in strength of friendship). It is important to note that, all ties are weighted or have values, even dichotomous relations have binary values (if tie exist a value of 1 is assigned in case of non existence value assigned is 0). However, in SNA dichotomous ties are treated as unweighted ties. When these authors focus their attention on a single node, they call that node the ego and call the set of nodes that ego has ties with alters. When network analysts collect data on ties from a set of nodes, they call it relational data. Relational data can be visualised in matrix form or in graphic form. Table 1 summarizes some terms of network.

B. Network Analysis

The four most important concepts used in network analysis are network density, centrality, betweenness and centralization. These concepts have several measures grouped in it. Additionally, there are four measures of network performance: robustness, efficiency, effectiveness and diversity. Whereas the first set of measures concerns structure, the second set concerns the dynamics and thus depends on a theory explaining why certain agents do certain things (e.g., access to information).

Table 1 Terms used in Social Network

Social Network terms	Definition
Node	The basic element of a network
Tie / Edge	A set of two nodes.
Network	A set of nodes connected by a set of ties
Valued Network	A network whose ties/edges are associated with ameasure of magnitude or strength
Ego	A node which receives particular focus
Alters	The set of nodes that has ties with the ego but notincluding the ego itself

Density is defined as the actual number of ties in a network, expressed as a proportion of the maximum possible number of ties. It is a number that varies between 0 and 1.0. When density is close to 1.0, the network is said to be dense otherwise it is sparse. When dealing with directed ties, the maximum possible number of pairs is used instead.

Centrality: local and global, the concept of centrality encompasses two levels: local and global. Intuitively, a node is central (locally) when it has the higher number of ties with other nodes. Local centrality only considers direct ties (the ties directly connected to that node) whereas global centrality also considers indirect ties (which are not directly connected to that node). For example, in a network with a "star" structure, in which, all nodes have ties with one central node, local centrality of the central node is equal to 1.0.

Betweenness explores further the concept of centrality. Betweenness measures the extent to which a particular node lies "between" the various other nodes in the network: a node with few ties

may play an important intermediary role and so be very central to the network. The betweenness of a node measures the extent to which an agent (represented by a node) can play the part of a broker or gatekeeper with a potential for control over others.

Centralization provides a measure on the extent to which a whole network has a centralized structure. Whereas density describes the general level of connectedness in a network; centralization describes the extent to which this connectedness is organized around particular focal nodes. Centralization and density, therefore, are important complementary measures.

Evolution of network performance can be done as a combination of (1) its robustness to the removal of ties and/or nodes. (2) Its efficiency in terms of the distance to traverse from one node to another and its non-redundant size. (3) Its effectiveness in terms of information benefits allocated to central nodes and (4) its diversity in terms of the history of each of the nodes.

Robustness: Social network analysts have highlighted the importance of network structure in discussion of network's robustness. The robustness can be evaluated by studying how it becomes fragmented as an increasing fraction of nodes is removed. Robustness is measured by an estimate of the tendency of individuals in networks to form local groups or clusters of individuals with whom they share similar characteristics, i.e., clustering. E.g., if individuals A, B, and C are all bioinformatics experts and if A knows B and B knows C, then it is highly likely that A knows C. When the measure of the clustering of individuals is high for a given network, the robustness of that network increase – within a cluster/group where everyone knows everybody it is unlikely that a given person will serve as a lynchpin in the network, potentially destroying connectivity within the network by leaving.

Efficiency: Efficient networks are those in which nodes (individuals or firms) can access instantly a large number of different nodes – sources of knowledge, status, etc., through a relatively small number of ties. Given two networks of equal size, the one with more non-redundant contacts provides more benefits. Social network analysts measure efficiency by the number of non-redundant contacts and the average number of ties an ego has to traverse to reach any alters, this number is referred to as the average path length. The shorter the average path length relative to the size of the network and the lower the number of redundant contacts and the more efficient is the network.

Effectiveness: While efficiency targets the reduction of the time and energy spent on redundant contacts by, decreasing the number of ties with redundant contacts, effectiveness targets the cluster of nodes that can be reached through non-redundant contacts. Each cluster of contacts is an independent source of information. According to [10], one cluster around this non-redundant node, no matter how numerous its members are, is only one source of information, because people connected to one another tend to know about the same things at about the same time. For example, a network is more effective when the information benefit provided by multiple clusters of contacts is broader, providing better assurance that the central node will be informed. Moreover, because non-redundant contacts are only connected through the central node, the central node is assured of being the first to see new opportunities created by needs in one group that could be served by skills in another group.

Diversity: As efficiency is all about reaching a large number of (non-redundant) nodes, node's diversity, not to be confused with network heterogeneity introduced previously, suggests that it is critical from a performance point of view that those nodes are diverse in nature i.e., the history of each individual node within the network is important.

Degree Distribution: In the study of graphs and networks, the degree of a node in a network is the number of connections it has to other nodes and the degree distribution is the probability distribution of these degrees over the whole network.

Cluster Coefficient: The Clustering coefficients of a selected node are defined as the probability that two randomly selected neighbours are connected to each other.

4. PROPOSED FRAMEWORK FOR ANALYSIS OF BOLLYWOD MOVIES

Business houses and producers invest money for making films which, will attract viewers. More viewers mean more gross income from box office. Good amount of revenues are at stake in movie business. Automated analysis of viewer's choice of Bollywood movies can unveil a good insight into the dynamics of business. Thus it is important to identify the viewer's choices. Movie rating analysis will definitely serve this issue. The aim of this work is to analyse the past movies dataset and from that analysis access the viability of the upcoming movies through rating system. Various stakeholders such as actors, financiers, directors etc. can use these analyses to make more informed decisions.

For the experiment dataset comprising bollywood movie information from 2010 to 2015 have been used from Imdb movie database [11]. The dataset contains the following attributes:

- Film Name (As Nodes)
- Actor Name
- Actress Name
- Director Name
- Ratings
- Duration
- Release Time
- Box Office Verdict
- Type of movie
- Review etc.

Some of the attributes had irrelevant or very little information, such attributes were removed as a pre-processing step. A snapshot of the modified dataset has been given in Figure 2.

Social network analysis is an effective means to extract semantic information from movies.Movie analysis through social relationships among different stake holders can support various types of information retrieval better than audio-visual feature analysis, affective content analysis [12-13] or other types of analysis.The analysis on the dataset has been done using Gephi- a social network analysis software [14]. Gephi is open source software for graph and network analysis. It uses a 3D render engine to display large networks in real-time and to speed up the network exploration. Moreover, it provides easy and broad access to network data and allows for filtering, navigating, manipulating and clustering. Two projected networks Actors_Projected_Network and Actress_Projected_Network as given in Figure 3 and Figure 4 respectively. The projected Network has been created considering Film Names as Node. In the networks edges have been created between the nodes with the following rules:

(a) There is an edge between 2 nodes (Film Names) if same actor/ actors work in the both movies. (In case of Actors_ Projected_ Network).

(b) There is an edge between 2 nodes (Film Names) if same actress/actresses work in the both movies. (In case of Actrees_Projected _Network).

(c) All the edges are undirected.

5. ANALYSIS ON THE PROJECTED NETWORK OF IMDB MOVIE DATASET

The analysis on the projected network obtained from Gephi shows some well known network facts like cluster coefficient of different clusters, node degree distribution, finding influential nodes etc. The analysis and finding of the networkare reported in this section.
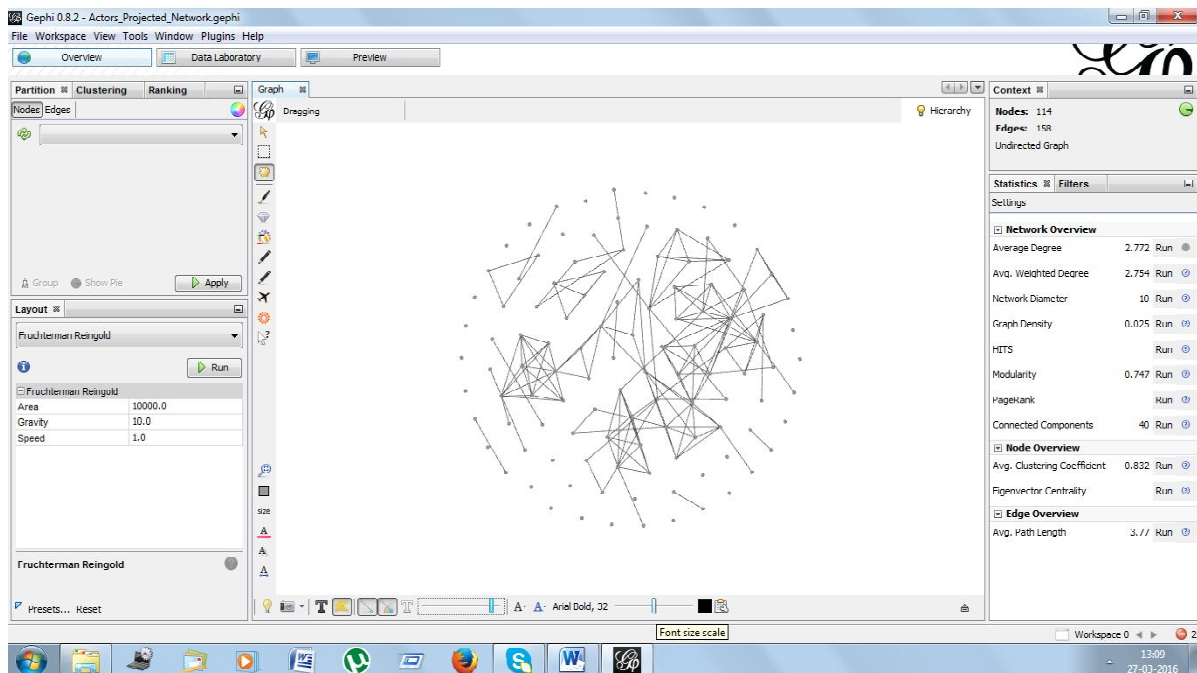
Fig. 2. Bollywood movie dataset

Fig. 3. Actors_project network

1) Edges in Actors_Projected_Network are 158 whereas edges in Actress_Projected_Network are 112- Thus it can be noted that frequency of the actors is higher than the actresses in the Bollywood movies between 2010-2015.

2) The Actors_Projected_Network has:
   - Average Clustering Coefficient: 0.832
   - Total Number of Triangles=148
   - Average Path Length=3.769
   - Number Of Shortest Path=2684
   - Number Of weekly connected components=40
   - Density=0.025
   - Average Weighted Degree=2.772

3) Where as Actress_Projected_Network has:
   - Average Clustering Coefficient: 0.946
   - Total Number of Triangles=145
   - Average Path Length=1.1384
   - Number Of Shortest Path=260
   - Number Of weekly connected components=68
   - Density=0.017
   - Average Weighted Degree=1.965

4) A Node (Film) Raajneeti has the highest degree (13) in the Actors_Projected_Network. That means Actor/Actors of the film Raajneeti is/are mostly connected to the other movies as well as illustrated in Figure 5. In the same lines node (Film) Satyagraha has the 2nd highest degree (10) in the Actors_Projected_Network.Node (Film) Housefull-2 has the 3rd highest degree (9) in the Actors_Projected_Network.

5) Akshay Kumar has the highest degree (6) among the actors in Actors_Projected_Network. Ajay Devgan, Nawazuddin Siddiqui, Amitabh Bachchan, Arjun Rampal has the 2nd highest degree (5) among the actors in Actors_Projected_Network. SRK has the 3rd highest degree (4).

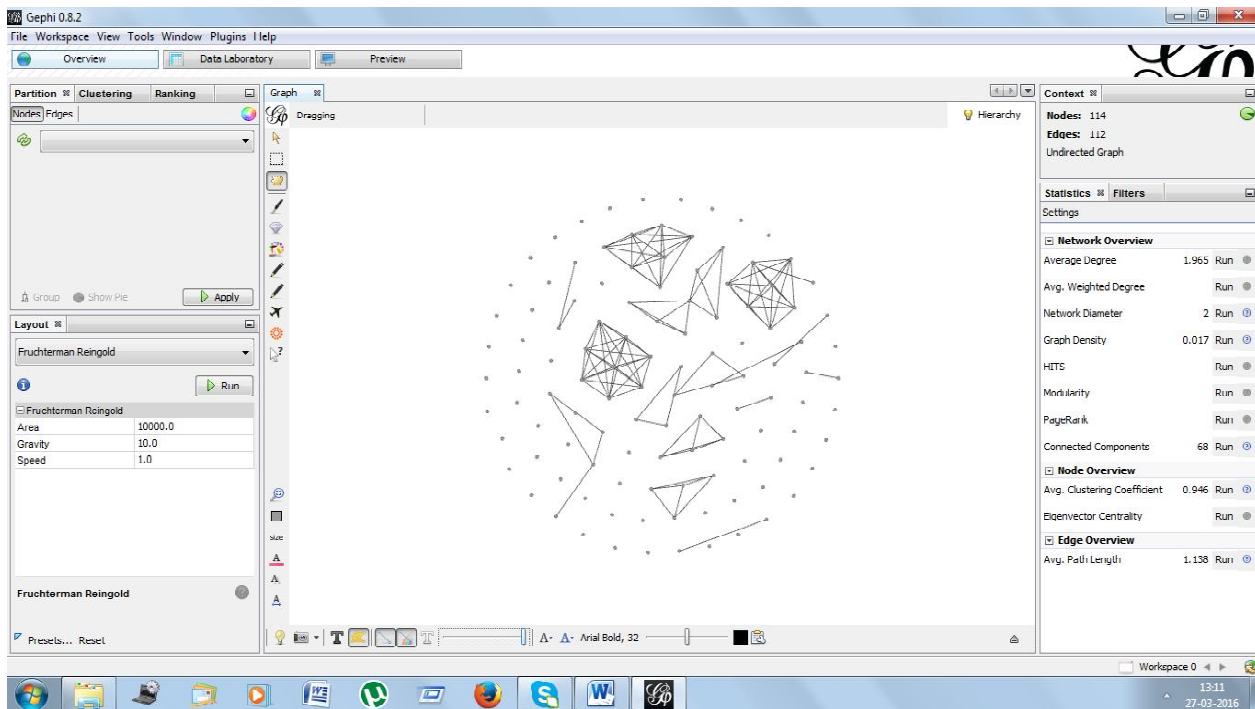After sorting of the dataset (a snapshot is given in Figure 6) its analysis gives:

Fig. 4. Actress_project network

6) Among the top 10 movies (According to ratings) Nawazuddin Siddiqui has the maximum entries (2).

7) Biographical/Action movies got the maximum ratings.

8) Average Movie time of the top 10 rated movies in Bollywood is around 166 minute.

9) 6 of the top 10 rated movies (2010-2015) have degree of below 5 in Actors_Projected_Network. That means most of the actor/actors of the highly rated movies are acted in very few famous movies in bollywood in last 5years.

10) 9 of the top 10 rated movies (2010-2015) released on ordinary day. Thus it may be concluded that there is no relation between release time and ratings. Though there are some exceptions such as- Aamir Khan Movies released around Christmas time. SRK movies mainly released, around Diwali/Eid time (4 out of 5). Salman Khan Movies also released around Eid. Also gross incomes of these movies are 100+ cr. Thus high profile actor based movies has high connection between release date and gross income. The analysis has been corroborated with Box Office India [15] reports.

11) Top 3 rated movies (2010-2015) all are remix of south Indian movies.

12) Deepika Padukone has the highest degree (7) among the actress in Actress_Projected_Network. Kareena Kapoor, Priyanka Chopra has the 2nd highest degree (6) among the actress in Actress_Projected_Network. Other than the movies of these actresses, Jab Tak Hai Jaan movie has the most number of connections in Actress_Projected_Network.

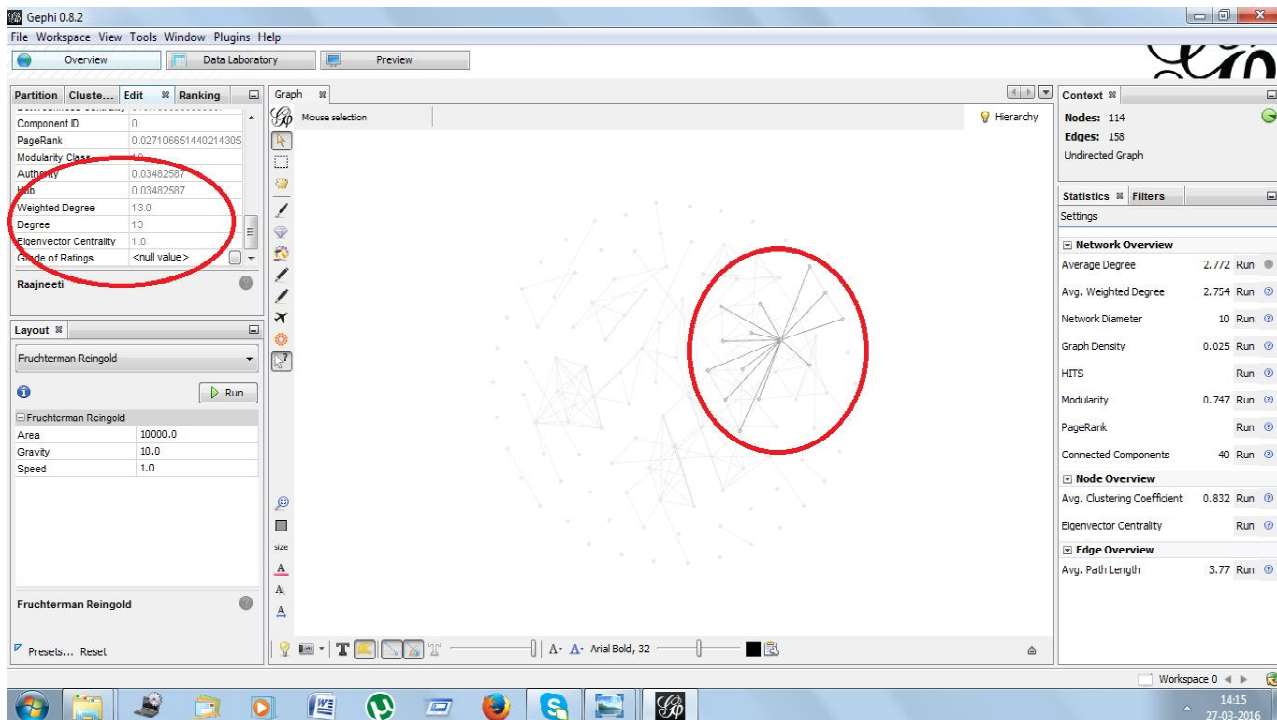13) No actress has double entry in top 10 rated Bollywood movies (2010-2015).

Fig. 5. Connectivity of a movie (Raajneeti)

Fig. 6. Sorted movies on the basis of ratings

9. CONCLUSION AND FUTURE SCOPE

The proposed word is an automated analysis on Bollywood using social network. The findings of the analysis have been reported and are found to be encouraging. From the findings, it can be concluded that average cluster coefficient property of the graphs is a conclusive parameter. Other properties like density, connected components, etc. are highly related with average cluster coefficient. However, it needs to be corroborated with more data by increasing the scope of study to last 15-20 years. Applications of some statistical and soft computing tools on the analysis are also on anvil. More substantial analysis with sentiment in forums can be used for developing automated predicting algorithm, which can be applied for measuring the success of the upcoming movies.

REFERENCES

[1] Wasserman, S., and Faust, K., Social Network Analysis: Methods and Applications, Cambridge: Cambridge University Press, 1994.

[2] Zinkhan, G.M., Roth, M.S. and Saxton, M.J., Knowledge Development and Scientific Status in Consumer-Behavior Research: A Social Exchange Perspective, Journal of Consumer Research, Vol. 19, No.2, pp.282-291, 1992.

[3] Carter, C.R., Leuschner, R. and Rogers, D.S., A Social Network Analysis of the Journal of Supply Chain Management: Knowledge Generation, Knowledge Diffusion and Thought Leadership, Journal of Supply Chain Management, Vol. 43, No.2, pp.15-28, 2007.

[4] Newman, M.E.J., Scientific Collaboration Networks. II. Shortest Paths, Weighted Networks, and Centrality, Phys. Rev. E, Vol. 64, No.1, pp.016132-1-7, 2001.

[5] Barabási, A.L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A. and Vicsek, T., Evolution of the Social Network of Scientific Collaborations, Physica A, Vol. 311, pp. 590-614, 2002.

[6] Park, S., Oh, K. and Jo, G., Social Network Analysis in a Movie Using Character-Net, Multimed Tools Appl., Vol. 59, pp. 601-627, 2013.

[7] Weng, C., Chu, W. and Wu, J., Movie Analysis Based on Roles' Social Network, Proceedings of the IEEE International Conference on Multimedia and Expo, pp.1403-1406, 2007.

[8] Doshia, L., Kraussa, J., Nanna, S., and Gloor, P., Predicting Movie Prices Through Dynamic Social Network Analysis, Proceedings of the Conference on Collaborative Innovations Networks Conference (COINs 2009), pp.6423–6433, 2009.

[9] Krauss, J., Nann, S., Simon, D., Fischbach, K. and Gloor, P., Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis, Proceedings of the ECIS-2008, pp.547-552, 2008.

[10] Burt, R.S., Structural Holes: The Social Structure of Competition, Harvard University Press, 1992.

[11] Top Movies of 2010-2015, http://www.imdb.com/list/ls075761511/, accessed on March 18 2016.

[12] Wang, H.L. and Cheong, L.-F., Affective Understanding in Film, IEEE Trans. on Circuits and Systems for Video Technology, Vol. 16, No.6, pp.689-704, 2006.

[13] Hanjalic, A. and Xu, L.-Q., Affective Video Content Representation and Modeling, IEEE Trans. on Multimedia, Vol. 7, No.1, pp.143-154, 2005.

[14] Bastian, M., Heymann, S. and Jacomy, M., Gephi: an Open Source Software for Exploring and Manipulating Networks, Proceedings of the International AAAI Conference on Weblogs and Social Media, pp.361-362, 2009.

[15] Box Office India, http://www.boxofficeindia.com/Years/years_detail/2015#, accessed on March 18 2016.