An efficient method to detect communities in social networks

MEHJABIN KHATOON (mehjabinkhatoon@gmail.com)
B S Abdur Rahman Crescent Institute of Science & Technology https://orcid.org/0000-0002-4758-2744
W AISHA BANU
B S Abdur Rahman Crescent Institute of Science & Technology

Research

Keywords: Community, Community structure, Social network, Complex network, Community detection

Posted Date: March 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-31561/v2

An efficient method to detect communities in social networks

Mehjabin Khatoon[1] · **W. Aisha Banu**[2]

Abstract
Social networks represent the social structure, which is composed of individuals having social interactions among them. The interactions between the units in a social network represent the relations of the various social contacts and aim at finding different individuals in that network, with similar interests. It is a challenging problem to detect the social interactions between individuals with comparable considerations and desires from a large social network, which can be termed as community detection. Detection of the communities from social networks has been done by other authors previously, and many community identification algorithms were also proposed, but those communities' identification has been achieved on the online available data sets. The proposed algorithm in this paper has been named as Average Degree Newman Girvan (ADNG) algorithm, which can easily identify the communities from the real-time data sets, collected from the social network websites. The approach presented here is based on first determining the average degree of the network graph and then identifying the communities using the Newman Girvan algorithm. The proposed algorithm has been compared with four community detection algorithms, i.e., Leading eigenvector (LEC) algorithm, Fastgreedy (FG) algorithm, Leiden algorithm and Kernighan-Lin (KL) algorithm based on a few metric functions. This algorithm helps to detect communities for different domains, like for any proposed government policy, online shopping products, newly launched products in a market, etc.
Keywords Community · Community structure · Social network · Complex network · Community detection

1 Introduction

For several years, online social media sites have become the most emerging ones which facilitate creating and sharing of information, ideas, individual's interests and other forms of expressions with the aid of networks and communities present in the social networks. Some of the popular social network sites are like Facebook, Instagram, Twitter, WhatsApp, LinkedIn, Pinterest, etc., and these online social media sites contain millions of users. These online social networking sites have tremendous impact in today's context for any individual to express the agreement, disagreement for the events, offers, occasions, politics, etc., which helps an analyzer to determine the required information for the required situation. In these social networks the reviews given by individuals for a particular post or any event, are in hundreds, or in thousands. Those reviews are in the form of likes, shares and comments. Owing to the large number of likes, shares and comments it is usually hard to determine the conclusion regarding - for and against of any particular event. Therefore, community detection methods are one of the convenient ways to analyze the circumstances for those particular events.

Social networks are those complex networks which can be analyzed by forming a network graph of a large number of nodes. Communities which are the sub-graphs of a graph, comprised of nodes, can be explored by using some community detection algorithms. From those formed communities we can do the future analysis. For instance, in the case of a product launch, what changes can be done further in that product, whether to continue that launched product or not, what are the positive and negative reviews for that product, etc. *Social network analysis* can be defined as a measurement of relations, connections between individuals, organizations, groups, computers, URLs, and other different types of informative entities. Social network analysis gives both the visualization and the mathematical analysis of social connectivity (Xu et al. 2013). *Community detection is* a type of analysis process, which is used in the analysis of the social network. Analysis of social network is the technique to analyze the social structures, present in the social networks through the use of the complex network, which can be formed from the social network itself. Communities identification is an ill-defined issue (Fortunato & Hric 2016). There exist no protocols universally for few basic factors like the community definition, algorithms validation and its performances comparison. Community detection is an NP-hard problem which has not been solved yet to a level of satisfaction (Azaouzi et al. 2019). Two major factors obstruct the

Mehjabin Khatoon
mehjabinkhatoon@gmail.com

W. Aisha Banu
aisha@crescent.education

[1] Department of Computer Science and Engineering, B S Abdur Rahman Crescent Institute of Science and Technology, Chennai-48, Tamil Nadu, India.

[1] https://orcid.org/0000-0002-4758-2744

[2] Department of Computer Science and Engineering, B S Abdur Rahman Crescent Institute of science and Technology, Chennai-48, Tamil Nadu, India

computational complexity of this problem. One is the huge size of today's social network websites and another is its dynamic structure evolving over time.

Complex network denotes the presence of structures of community when the grouping of the sets of nodes converts the structure of the network in that form in which internal nodes in the formed group are densely connected while the external nodes are sparsely connected. The power *law* is one of the properties of a complex network and the distribution of the power law is followed by the scale-free networks, which comes under the category of small world networks (Amaral et al. 2000). *Community structure* provides a proper awareness of social relations and also helps in a broad range of applications provided by social networking, as the detection of communities. Among the features of social networks, *community structure* is the most significant feature (Nguyen et al. 2014), which is present in a network if that network nodes can be grouped into some sets of nodes and also internally those networks have a dense connection. The real networks are also called as *complex networks*, whose analysis should be done on the basis of *community structure*. The nodes in the graph of a social network are usually linked according to the relationships based on friendship, the same native, the same workplace, etc. (Choudhury et al. 2013). In networked systems, the problem of determining and characterizing the community *structure* is a major issue. One possible solution is to focus and optimize the function "modularity" which is a constructive approach to determine the quality of the divisions of a network. Newman (2006) showed that the graph modularity expression, represented in the form of a matrix of eigenvectors, also called as modularity matrix, performs better than the other competing methods on the basis of less execution time.

Till date, several algorithms have been developed for the detection of communities which can be applied in numerous types of areas like social science, graph theory, statistics, physics, biology, and linguistics (Vasudevan and Deo 2012). Newman and Girvan (2003) framed an algorithm for determining the structure of communities' present in the networks, that does the division of the nodes in the network into subgroups which are densely connected internally and sparsely connected externally (Vasudevan and Deo 2012). In the perspective of visualization of communities, it helps to visualize the whole network organization and gives more clarity, wherein each community represents a group with common characteristics or it represents individuals with similar interests. In complex networks, identification of communities is an NP-complete problem (Fortunato 2010). Detection of communities helps to determine the network functions and visualize the multiple structures of a network.

Partitioning a graph is an NP-complete problem, which is somewhat difficult for large graphs; but many heuristic algorithms have been developed for partitioning the graphs, like the Kernighan-lin algorithm. Luo et al. (2008) proposed three new algorithms, i.e., Greedy algorithm, Add-all algorithm, and Kernighan-Lin algorithm - all of which starts from a given source vertex to find local optimal community structures in large networks. Lancichinetti et al. (2008) introduced a new benchmark graphs category that determines the difference between the sizes of the communities and the distribution of node degrees. Lancichinetti tested Newman Girvan graphs modular structure, by using *normalized mutual information* which is a metric for calculating the similarity between the graph's divisions.

The research endeavors until now, have encouraged the success of the detected communities in community detection approaches. Nevertheless, it also increases the problem of choosing the appropriate algorithm to apply in different scenarios. The main problem till now is that the proposed approaches have not been compared or analyzed with each other upon unified platforms (Wang et al. 2015). The research work done in this paper is the detection of communities with the help of the proposed ADNG algorithm from the data collected from one of the most famous social networks, i.e., Facebook. The detected communities can predict individuals with positive and negative reviewers. The results of the detected communities have been compared with four community detection algorithms, which are - Fastgreedy, Leading eigenvector, Leiden and Kernighan-Lin algorithms. The researchers nowadays are also working in determining the hidden community structures which are called weak communities (He et al. 2017). Many researchers are still working to determine the large complex network, and obtain the results in a certain specific way which can easily depict the circumstances of any specific interest.

2 Materials and methods

2.1 Materials

The experimental analysis of the ADNG algorithm was executed using the data sets collected from the social network website Facebook. People in the social network upload posts about various things ranging from political, social and economical to private and personal issues dealing with the events of their life, in the form of videos and photographs. The posts in the social network draw responses from many other persons in the form of likes, shares, and comments. The posts on Facebook have a unique id and also every individual user on Facebook have a unique id. A particular post on Facebook receives responses from different individuals in the form of many likes, shares, and comments. So, we have extracted the id of a particular post and the id of individuals who have liked, shared and commented on the posts.

We preferred to extract only real time datasets which have gained maximum common public attention, since in the previous works researchers have considered the already available data sets from UCI machine learning repository. The data sets collected from Facebook are: (i) DATA SET 1: A video post of about a government policy - demonetization, on which many celebrities, political leaders have posted a lot of things in social network, (ii) DATA SET 2: An Amazon post of about a newly launched particular product, i.e., about a T-shirt, (iii) DATA SET 3: A written statement post based on demonetization that has been posted

by Chetan Bhagat (iv) DATA SET 4: A Flipkart post of about #FlipkartBigShoppingDays.

The extracted id of posts and the individual's id for likes, shares, and comments were arranged in Microsoft Excel sheet tables. The tables were then converted into a network graph which will be input for the proposed ADNG algorithm. The implementation was done using R coding. The comments collected from Facebook, corresponding to each post have been segregated into positive and negative comments using the *Sentiment Analysis* (SA) method. The SA method depends upon positive and negative words, for which we have used +1 for a positive word and -1 for a negative word. The SA method has been implemented using Java coding. The final score corresponding to each comment has been used for segregating the comments, which is positive for the positive comment and negative for negative comment. The datasets tables were arranged in two columns - one is Post ID and another is Individual ID. Post ID is same for the likes, shares and comments of a particular post. In order to make a difference in between the vertices of likes, shares, and comments, in the network graph formed from the table of data set of a single post, the Post ID has been initiated with 1, 2 and 3 for the id of shares, positive and negative comments respectively. The Post ID has been kept the same for the likes-id as it can be seen in table 1. We have also included the likes rendered for the sub-comments of a post.

The comments extracted corresponding to each post have been segregated as positive and negative comments by using the Sentiment Analysis (SA) method. A database of positive and negative words was collected from the internet for implementation of the SA method which has been implemented using Java language. The SA method implemented, is depend on the concept of +1 for a positive word and -1 for a negative word. Thus, a score will be calculated based on the SA method, which will be positive for a positive comment and negative for a negative comment.

Table 1 A part of the full DATA SET 2 table, which consists of some likes id, shares id and comments id

Post ID	Individual ID
9465008123_10154703093783124	835874766427515
9465008123_10154703093783124	10152400310023100
1.9465008123_10154703093783124	193973515297
1.9465008123_10154703093783124	1405864056351850
2.9465008123_10154703093783124	700104936753274
2.9465008123_10154703093783124	481707478886524
3.9465008123_10154703093783124	10153110632619000
3.9465008123_10154703093783124	741089122628838
10153110632619000	766281553469225
481707478886524	870715158924

2.2 Proposed Method

The main concept of the proposed algorithm depends on the *average degree (ad)* of the formed graph. The algorithm has been termed as Average Degree Newman Girvan (ADNG) algorithm. The ADNG algorithm has been implemented on the network graphs using the R tool. The input for this algorithm is the network graph which is formed from the data sets arranged in the excel table, that have been collected from Facebook. Initially, the ADNG algorithm will consider a *seed node (s)*, and the same seed node was also considered for all other graphs that have been formed using the scale-free graphs and the Newman Girvan (NG) graphs, which were considered for the comparison of the performance analysis of the ADNG algorithm. The graphs considered for the experimental purpose are the undirected graphs. The seed vertex can be any vertex in the network and can belong to any community. The graphs formed from the collected data are explored at full length. The communities formed from the graph will be having the vertex with comments, likes, and shares.

2.2.1 Steps for identifying communities

1. In the very first step, the seed vertex from the network graph formed from the collected data sets ought to be initialized. The work ought to get initialized from the mentioned number of the seed vertex.

2. In the second step, the degree of the graph is computed. Characterization for the topology of a real network can be determined by degree correlations and clustering hierarchy (Vazquez 2003).

3. In the third step, the vertex with the least degree ought to be deleted. The vertex with least degree refers to those individuals in the whole network whose contribution is less, i.e., whose likes, shares or comments are extremely less in number.

4. In the fourth step, the average degree of the graph is calculated. The repetition from step 1 to step 4 ought to be done in every iteration and it will continue until the average degree value becomes less than the previous iteration.

5. The formed graph after the previous steps ought to be utilized for detecting the communities using Newman Girvan algorithm whose working depends on the factor of "edge betweenness" which determines every shortest path that exists between a pair of nodes.

2.2.2 Average Degree Newman Girvan (ADNG) Algorithm

The proposed algorithm's main concept is based on the average degree of the graph and the Newman Girvan algorithm, as it can be seen in the ADNG algorithm. Metrics for the analysis of graphs like degree which is the number of edges incident to a vertex in a network graph and the clustering coefficient which is the fraction of edges among the neighbors of a vertex have attracted the attention of researchers a lot. However, Barabasi et al. (1999) emphasized that a lot of real networks are also

characterized by power-law degree distributions, and have given an appreciable probability to observe high degree vertices.

Input:
G ←G (V, E) //*input graph with V number of vertices and E number of edges*
s //*seed vertex*
l //*length of the graph to be considered for the application*

Output:
C //*Community formed*

Initialize:
Set seed vertex (*s*)

Procedure:
ad ← Calculate_average_degree(*G*) // *average degree for the input graph*
while(*ad* > *ah*)
{
d ← compute the degree
ad ← calculate average degree
ah ← *ad* //*previous best average degree*
dmin ← nodes with minimum degree
V' ← (V/d > dmin) //*excluding the nodes with a minimum degree*
ad←Find average degree
}
Returns the graph g(V'E')
C ←Forming the community by applying Newman Girvan algorithm.

This research is primarily based on the average degree of the network graph. The fixing of the seed vertex was done to start the whole process from that fixed node. The degree of the graph was determined for the input graph which results in giving the degree of each and every node. We have used the concept of the degree of a node because the degree of each and every node denotes its contribution in that network graph. So, some individuals who had very less contribution in that data set, which was a node in the case of the graph, were deleted according to the proposed ADNG algorithm. The mean or average degree of the graph was then calculated and this process continued for every iteration until the average degree becomes less than the previous iteration. The network graph then formed after this process was partitioned for the formation of communities according to the Newman-Girvan algorithm.

The main purpose of using the concept of degree is that very less valued nodes, like the person whose contribution is very less in the whole set of reviews can be discarded. The collective contribution of all the nodes degree led us to use the concept of average degree, dependent upon makes the deletion of less valuable nodes possible.

The conclusion of the decision regarding the review of a product can be done by considering the precise amount of much valuable nodes. In ADNG algorithm the average degree value increases iteratively but the iteration halts in which the evaluated value is lesser than the previous iteration. Other community detection algorithm which has been considered in this research work does not focus on the concept of the node's contribution, and thus this concept makes the ADNG algorithm a different one from those compared community detection algorithms.

In the graph partitioning method, a good partition involves fewer existence of edges between communities than expectation. In the Newman Girvan algorithm procedure, the communities are detected by the calculation of betweenness scores for all the edges in the network. It then removes the edges from the network - which have the highest scores, and it continues the calculation of edge betweenness for all the remaining edges (Newman 2006). The Newman Girvan (NG) algorithm shares two characteristics; the first one includes the removal of the edges iteratively, so that the network should get split into subgroups or communities, and the removal of edges are done with the help of "betweenness" calculations, and the other characteristic is that the "betweenness" calculations are done in every iteration of node removal. The calculation of "betweenness" measure is done for every edge in the network in $O(mn)$ time whereas, *'m'* represents the number of the edges in the network graph, while *'n'* represents the number of the nodes in the network graph. The complexity of the Newman Girvan algorithm is $O(m^2n)$. The time complexity of the proposed ADNG algorithm is $O(m^2n)$. The space complexity of the proposed ADNG algorithm is $O(m+n)$. Edge-betweenness measures focus on finding edges with the highest betweenness and it is a measure that favors the edges which lie in between the communities and disfavors those edges which lie inside the communities (Newman and Girvan 2003).

3 Results and Discussion

The final graph formed for each post, after the detection of communities is of positive supporters, negative supporters, and sometimes other groups or communities of likes and shares. The posts in a social network can be of any sort. It can be a video, any image or any composed explanation. When an individual shares a specific post in social network sites, generally that individual is by all accounts a constructive supporter for that post, however can likewise give a contrary remark, therefore in the final

formed network graph (which got segregated into communities) that particular individual can be in the negative supporter's community along with the individuals who have given antagonistic remarks. The final network graph will be having lesser number of nodes than the initially formed graph because nodes with very less degree will get expelled, but in the final resultant network graph, it will give estimation whether most of the general public preferred or loathed the items, occasions or any renowned talks.

The initial networks of the data sets have not been shown in this paper since those networks are too much crowded. After applying the ADNG algorithm, the communities formed for the data collected for DATA SET 1 consist of individuals id, which was initially of 1308 nodes but got reduced to less number of nodes and it's shown in Figs. 1 and 2 that are with and without vertex label respectively, formed after the application of ADNG algorithm with 7 groups, in which green color community is of negative supporters and sky color community is of positive supporters. The communities formed are of nodes with positive and negative supporters, likes hit for those supporters, shares done for the post. The communities have been formed according to the *shortest edge betweenness* concept of the Newman & Girvan community detection algorithm. The graph formed after the application of the ADNG algorithm is of the lesser number of nodes than in the initial graph, but the groups of the majority of supporters for positive and negative reviewers can be determined by the final network graph. Similarly, Figs. 3 & 4, Figs. 5 & 6, and Figs. 7 & 8 graphs correspond to DATA SET 2, 3 and 4 respectively. In every network graph, it shows the communities formed, after the application of the ADNG algorithm in which green color community is of negative supporters and sky color community is of positive supporters.

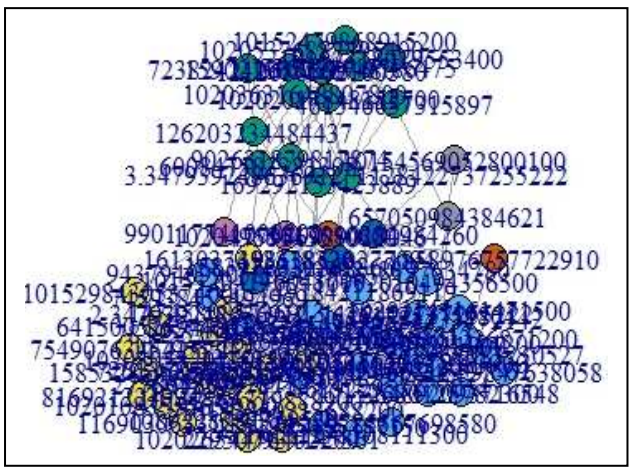

Fig. 1 Graph with vertices label formed for DATA SET 1

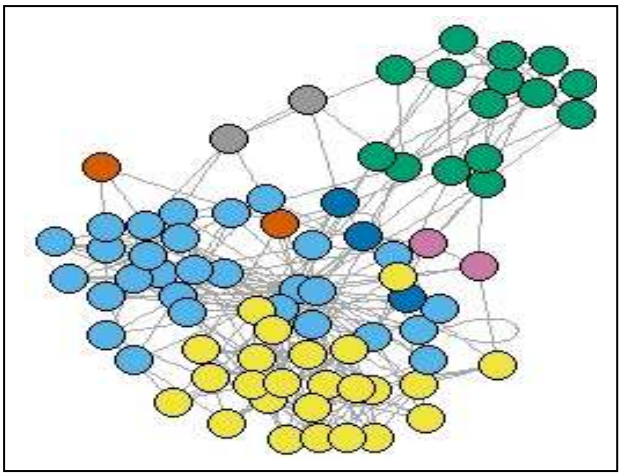

Fig. 2 Graph without vertices label formed for DATA SET 1

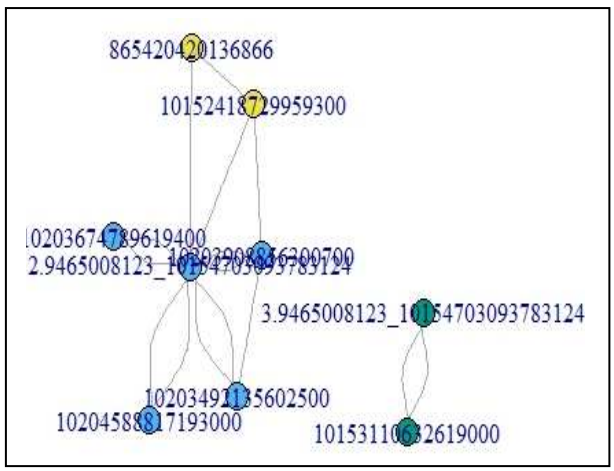

Fig. 3 Graph with vertex label formed for DATA SET 2

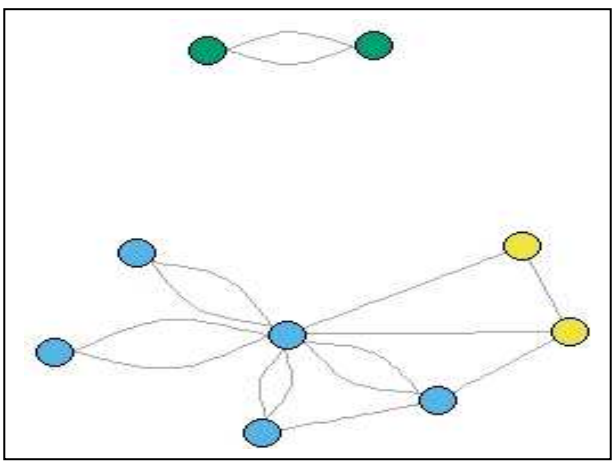

Fig. 4 Graph without vertex label formed for DATA SET 2

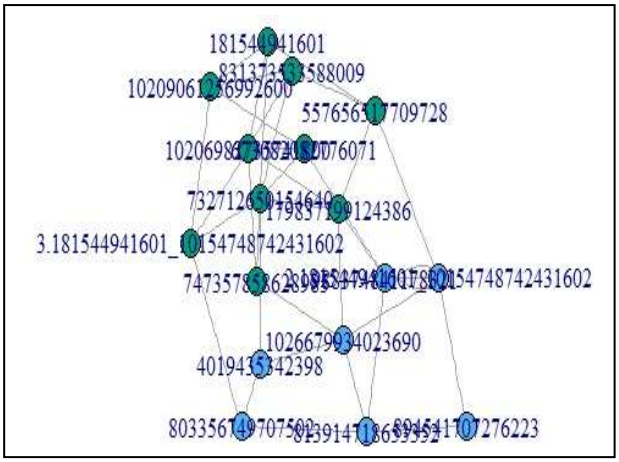

Fig. 5 Graph with vertex label formed for DATA SET 3

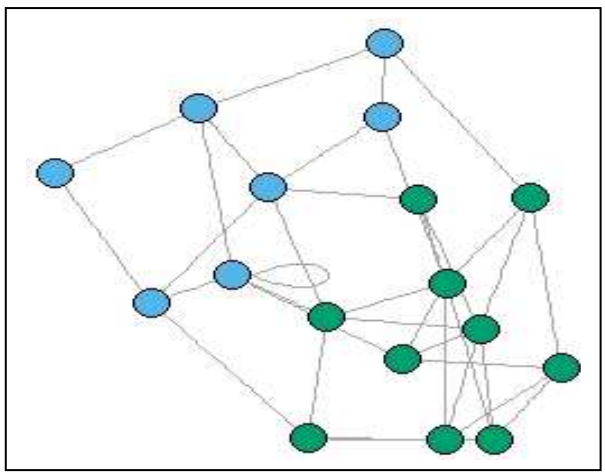

Fig. 6 Graph without vertex label formed for DATA SET 3

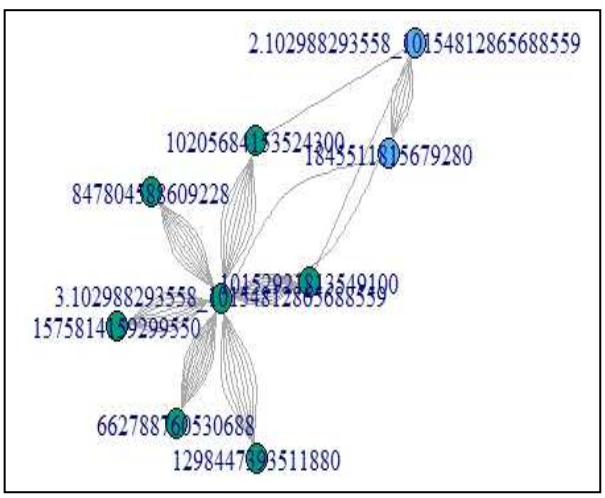

Fig. 7 Graph with vertex label formed for DATA SET 4

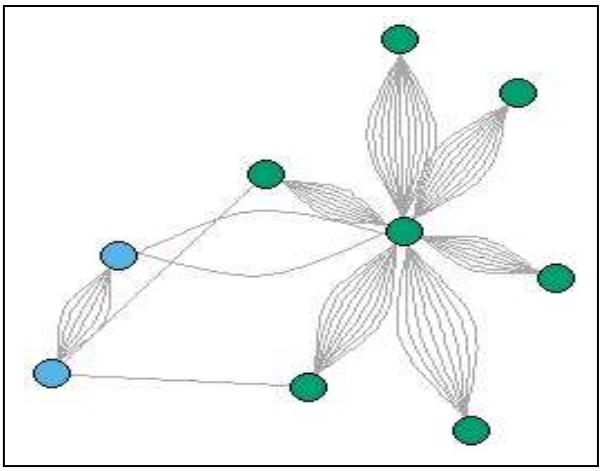

Fig. 8 Graph without vertex label formed for DATA SET 4

3.1 Community significance test

The communities formed according to the ADNG algorithm have been tested by a community significance test, named *Wilcoxon rank-sum test*. The health and social sciences area can be tested by the rank tests like *Wilcoxon rank-sum test* which can be used for exploratory and formal inferences purpose (Lumley et al. 2013). The *Wilcoxon rank-sum test* is the non-parametric version of the two-sample t-test. *Wilcoxon rank-sum test* has been applied on the "internal" and "external" degrees of a community, for quantifying the significance of a community. We can assume the edges which are within a community are internal edges while the edges which connect the vertices of a community with other communities' present in a network graph are external edges. The conditions of differences between the number of internal and external edges which are incident to the vertex of the communities have been considered as the null hypothesis of the test. The value of the test statistic results either in being a community or an anti-community. For community significance, the internal edges are more than the external edges and for anti-community, external edges are more than the internal edges. The *p-value* results by this function will be close to zero in both cases (i.e., for a community and for an anti-community) which informs about the statistical significance of the difference. Table 2 is about the Wilcoxon test results for DATA SET 1, in which total 7 communities formed comprising of 4 communities and 3 anti-communities. Table 3 shows the Wilcoxon test results for DATA SET 2, in which a total of 2 communities and 1 anti-community formed. Table 4 shows the Wilcoxon test results for DATA SET 3, in which a total of 2 communities formed. Table 5 shows the Wilcoxon test results for DATA SET 4 in which a total of 2 communities formed.

Table 2 Wilcoxon test values for DATA SET 1

Groups color	Wilcoxon test statistic "W" value	P-values	Community or anti-community
Green colour	210	0.00004828	Community
Sky colour	621	0.0006025	Community
Yellow colour	492.5	0.00002297	Community
Pink colour	0	0.1939	Anti-community
Grey colour	0	0.2207	Anti-community
Blue colour	0	0.0722	Anti-community
Red colour	2	1	Community

Table 3 Wilcoxon test values for DATA SET 2

Groups color	Wilcoxon test statistic "W" value	P-values	Community or anti-community
Green colour	4	0.1939	Community
Sky colour	24	0.01812	Community
Yellow colour	1	0.6171	Community

Table 4 Wilcoxon test values for DATA SET 3

Groups color	Wilcoxon test statistic "W" value	P-values	Community or anti-community
Green colour	99.5	0.0001634	Community
Sky colour	46	0.005664	Community

Table 5 Wilcoxon test values for DATA SET 4

Groups color	Wilcoxon test statistic "W" value	P-values	Community or anti-community
Green colour	49	0.001796	Community
Sky colour	4	0.1939	Community

3.2 Power law and clustering coefficient

The complex network follows the property of power-law degree distribution and clustering coefficient (Vasudevan and Deo 2012). Social networks are like complex networks which follow the properties of it. A complex network basic property depends upon its edges capacity to form a cluster which can be determined by the clustering coefficient (Yin et al. 2018). Normally, complex networks do not occur in simple networks like lattices or random graphs but usually happen to occur in real networks. In *Power law degree distribution* — the degree of a node or vertex is the number of its neighbors. Li & Chen (2003) analyzed the statistical properties of the linking strengths of "real-world" networks by studying some examples of scientific collaboration networks, and after analyzing, the authors plotted the data in a form which follows the power law distribution, i.e., $P(K)=K^{-\alpha}$ with an exponent $\alpha=2.0\pm0.2$, whereas K is the connection strength.

The characterization of power-law distributions is done by a slower than exponentially decaying probability tail and the loose occurrence denotes the occurrence of the large value with a non-negligible probability (Clegg et al. 2009). In a randomly selected node in an undirected graph G, it has the probability P_k for that node degree to be k. The graph G is a scale free graph if it is having a P_k distribution, as shown in Eq. (1), as heavy tailed, where C > 0 is a constant and $\alpha \in (0, 2)$

$$P_k \sim C_k^{-\alpha} \tag{1}$$

One more property of networked systems like social networks and World Wide Web is *clustering coefficient* or network transitivity, in which probability of two nodes for being neighbors is high if those two nodes are common neighbors to a third node. In a social network perspective, two friends of an individual will know each other more, in comparison to any other random person, due to their regular acquaintanceship with that individual. This effect is quantified by the clustering coefficient that can be defined as in Eq. (2).

$$C = \frac{3\times (number\ of\ triangles\ on\ the\ graph)}{(number\ of\ connected\ triples\ of\ vertices)} \tag{2}$$

This clustering coefficient value is 1 for a fully connected graph, when everyone is familiar with everyone else and has values usually in the range of 0.1 to 0.5 in many real-world networks (Girvan and Newman 2002).

The property of complex network, i.e., power law and clustering coefficient has been tested on the network graphs obtained after the application of ADNG algorithm. Fig. 9 shows the plot for the power law degree distribution for the final graph for **DATA SET 1**, in which value for α =1.007, and the clustering coefficient value is 0.05273672. Fig. 10 shows the plot for the power law degree distribution for the final graph for **DATA SET 2**, in which value for α =1.888889, and the clustering coefficient value is 0.3913043. Fig. 11 shows the plot for the power law degree distribution for the final graph for **DATA SET 3**, in which value for α =1.941176. Fig. 12 shows the plot for the power law degree distribution for the final graph for **DATA SET 4**, in which value for α =1.704744.

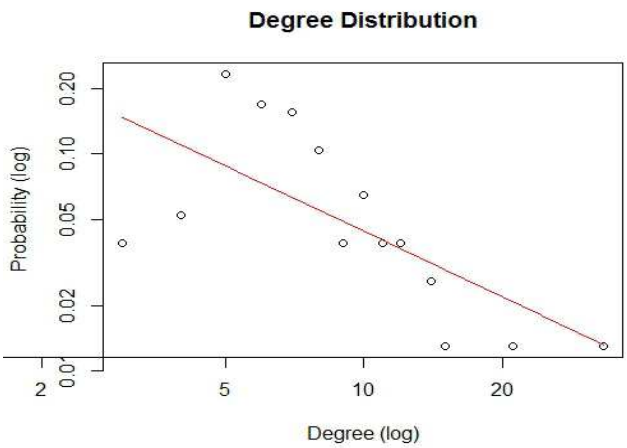

Fig. 9 Plot for power law degree distribution for DATA SET 1

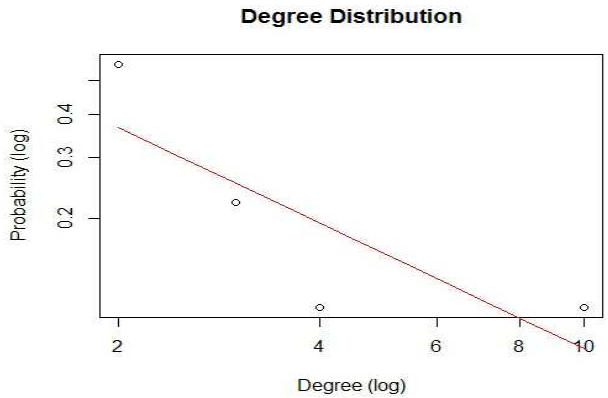

Fig. 10 Plot for power law degree distribution for DATA SET 2

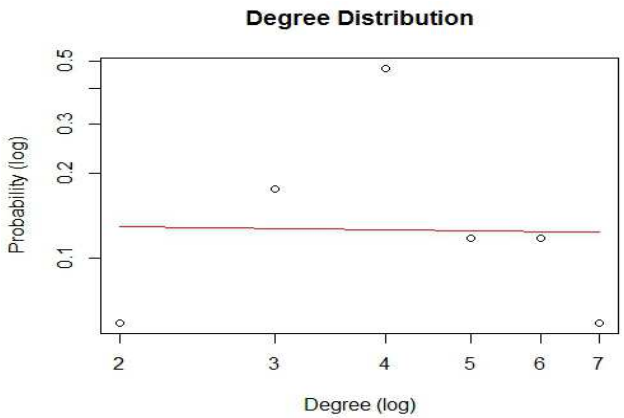

Fig. 11 Plot for power law degree distribution for DATA SET 3

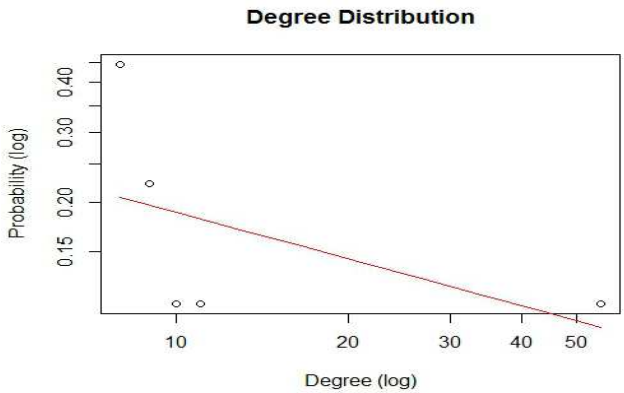

Fig. 12 Plot for power law degree distribution for DATA SET 4

3.3 Other community detection approaches

There are several other approaches for the partition of networks or for detecting communities from the networks. The proposed algorithm has been compared with Leading eigenvector (LEC) algorithm (Porter et al. 2009), Fastgreedy (FG) algorithm (Ciglan and Norvag 2010), Kernighan-Lin (KL) algorithm (Luo et al. 2008) and Leiden (LD) algorithm (Traag et al. 2019).

Leading eigenvector (LEC) algorithm is effective and also the simplest way, for subdividing a network in a recursive way (Porter et al. 2009). Leading eigenvector community detection algorithm follows the top-down hierarchical approach which repeatedly optimizes the modularity function. In this procedure, the graphs are split into two parts with the help of evaluation of leading eigenvector of the modularity matrix, and the separation leads to the increase in the modularity of the graph.

Greedy approaches are also applied for detecting communities, in which assignments of nodes are done in a greedy way and that assignment also contains most of that node's neighbors (Ciglan and Norvag 2010). A node in a greedy approach join a maximum of its neighbors in the same community and at the same time, neighbors of that node would join several communities, then the ties would be broken randomly in a uniform way. **Fastgreedy (FG)** algorithm is a hierarchical, bottom-up approach, in which optimization for the modularity of the graph is done in a greedy manner. The vertexes initially are in separate communities and the communities are merged iteratively to make it locally optimal, and the algorithm runs till there is a possibility to increase the modularity of the graph.

Leiden (LD) algorithm guarantees the communities well connectedness. It is actually an improved version of the previously existing community detection algorithm named Louvain algorithm (Blondel et al. 2008) and was proposed by Traag et al. (2019). It has three steps: (i) nodes local movement (ii) refinement of division and (iii) network aggregation on the basis of refinement of division, by using the non-refined division in the initial division creation process. LD algorithm is much more complex than Louvain algorithm. The community formation in this LD algorithm is not on the basis of searching highest value in terms of

any function but instead it merges with the community which leads to just an increment in the function. The merging process starts with random selection of a community but the preference for merging is given only with those communities which leads to an increment in the value of the function.

The **Kernighan-Lin (KL)** algorithm is designed to find a locally optimal partition of a graph in a heuristic way (Luo et al. 2008). The KL-like algorithm does the movement of nodes even when the modularity gains are temporarily negative because in this situation of nodes movement compensation of the previous losses and the result sub-graph can be formed with much higher modularity. Thus, KL algorithm can also climb out of local optima and get a solution close to the global optima. The KL algorithm results into two partitions, so the functions considered for the comparison between our algorithm and the KL algorithm has separately measured the two different partition of the KL algorithm, i.e., KL algorithm partition 1 and partition 2.

3.4 Scoring functions

The ADNG algorithm is compared with the above-discussed algorithms, NG graphs, and synthetic graphs based on the functions - Jaccard index, Modularity, Conductance, Assortativity degree, Normalized mutual information, and Rand index.

Conductance: Conductance is a function useful in graph partitioning which gives the ratio of the number of cut edges in that graph to the volume of the smallest part. Conductance is a measure which gives the fraction of the total number of edges that point outside the cluster (Yang and Leskovec 2012). It can be calculated as follows as given in Eq. (3):

$$f(S)=c/(2m+c) \quad (3)$$

Whereas S is the set of vertices, m is the total number of edges in S and c is the number of edges which are present in the boundary of S. For a good community structure, number of edges inside a community should be well and uniformly connected, i.e., it should be hard to split that formed community further. So, conductance gives the ratio of the number of edges which point outside the community or the sub-graph to the number of edges which are inside the community or inside the sub-graph.

Modularity: Modularity is a metric which measures the structure of graphs or networks. It is designed to measure the strength of division of a sub-graph formed from a graph. The concept for a true community structure corresponds to the arrangement of edges statistically, which can be quantified by using the modularity metric. Modularity is the subtraction of the number of links or edges inside a group (or a cluster) and the number of edges in an equivalent random network. Newman and Girvan (2003) proposed the *modularity* metric, for determining the quality of a particular division of a network. It derived the grade of the formed subgroup from the graph, and was measured for k communities in this way, as given in Eq. (4).

$$Q= \sum_i^k (e_{ii} - a_i^2) \quad (4)$$

A particular division of a network was considered into *"k"* communities which was defined as $k \times k$ symmetric matrix. The matrix element e_{ij} was the fraction of all the edges in the whole network and which connected the vertices from community i to the vertices in community j. $\sum_i e_{ii}$ was the trace of the matrix and was the fraction of edges which connected the vertices within the same community. The row or column sums $a_i = \sum_j e_{ij}$, represented the fraction of edges that connected to vertices in community i. Modularity values Q (0, 1), represents 0 when the community edges are random and it represents 1 when modularity value reaches in the maximum state, whereas 1 indicates a strong community structure.

Jaccard index: The ADNG algorithm results were compared on the basis of the Jaccard index in which it calculated the percentage of correctly identified nodes by keeping the same seed node in every other algorithm considered for comparison. Alamsyah et al. (2014), mentioned the Jaccard similarity approach as one of the node similarities approaches. Jaccard index can also be called the Jaccard similarity coefficient, which is a statistical measurement used for comparing the similarity and diversity of finite sample sets (Vasudevan and Deo 2012). The value varies between 0 and 1, in which 0 indicates completely dissimilar and 1 indicates completely similar. Jaccard index value can be calculated by Eq. (5).

$$J = (T \cap C) / (T \cup C) \quad (5)$$

Whereas, T represents the set of vertices that are targeted for the resultant community to be formed and C represents the set of vertices actually formed in the resultant community.

Assortativity: Assortativity is an attachment of nodes to other nodes which are similar in some manner. Assortativity actually quantifies the nodes tendency, for the connection between similar nodes in a complex network (Thedchanamoorthy et al. 2014). Newman (2002) first introduced the concept of assortativity. Assortativity is that metric of the graph which represents the extension of associativity between the nodes in a network, in a similar sort or in a dissimilar sort. The most prevalent type of assortativity is degree *assortativity* which is broadly used in network science (Noldus and Mieghem 2015). A network assortativity depends upon the average connectivity between nodes with high degrees, and also on the average connectivity between nodes with

low degrees. Assortativity metric expression, as a scalar value ρ ranges from -1 ≤ ρ ≤ 1.

Normalized mutual information (NMI): Normalized mutual information method is an extensive way to measure the comparison of community detection methods (Amelio and Pizzuti et al. 2015). NMI is an effective measure to do the evaluation of a method, that whether a community detection method finds significant groups of nodes which best fits the underlying community organization. NMI can be calculated as follows as given in Eq. (6).

$$NMI(A,B)=\frac{-2\sum_{i=1}^{R}\sum_{j=1}^{S}C_{ij}\log\left(C_{ijn}/C_i.C_j\right)}{\sum_{i=1}^{R}C_i.\log\left(\frac{C_i}{n}\right)+\sum_{j=1}^{S}C_j.\log\left(\frac{C_j}{n}\right)} \quad (6)$$

Suppose, V is the set of n number of nodes, whereas A and B are two divisions of a network, and the two partitions of V are A = (A_1, . . . ,A_R) and B = (B_1, . . . ,B_S). The C is a contingency table which shows the overlap between A and B, also called confusion matrix and C is of size *R x S*, and *Cij* represents the number of nodes which divisions *Ai* and *Bj* shares. If NMI (A, B) value is 0, that means A is completely dissimilar to B, and if value is 1 that means they are totally similar.

Rand index: Rand index or rand measure is a statistical measurement, in perspective of data clustering, for determining the similarity between two data clusters. Rand index has a value between 0 and 1; in which 0 indicates that there is no similarity in between the two data clusters and 1 indicates that the two data clusters are almost similar. It can be calculated by using Eq. (7), in which TP denotes true positive, TN denotes true negative, FP denotes false positive and FN denotes false negative.

$$Rand\ Index = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (7)$$

3.5 Comparison analysis

In the graphs, ADNG denotes the Average Degree Newman Girvan community identification algorithm, KLAP1 denotes the Kernighan Lin algorithm partition 1, KLAP2 denotes the Kernighan Lin algorithm partition 2, LEC denotes the Leading eigenvector algorithm, FG denotes the Fastgreedy algorithm and LD denotes the Leiden algorithm. The proposed algorithm results have been analyzed for individuals who have supported and not supported the posts of Facebook based on the formed communities.

The proposed ADNG algorithm has been compared with the other algorithms on the basis of the Jaccard index, i.e., the similarity between the Targeted result (T) and the Actual result (C). Fig. 13 shows the performance of our algorithm on the basis of Jaccard index for DATA SET 1, which is better than the other algorithms for both the positive reviewed and negative reviewed communities, except for KLAP2 in the case of positive reviewed community and for LEC in the case of negative reviewed community.

Assortativity degree represents the connectivity of nodes with a similar type of nodes and the values lies between -1 to 1. Fig. 14 represents the performance of our algorithm on the basis of assortativity degree values for DATA SET 1 which shows the negative lowest value, i.e., the highest value is of the ADNG algorithm.

Modularity metric used to measure the structure of graphs and is designed to measure the strength of division of a sub-graph formed from a graph or a network. The valid modularity range lies between 0 and 1, whereas modularity value 1 represents a strong community structure. Fig. 15 shows the modularity performance for our algorithm when compared to the LEC, FG and LD algorithms for DATA SETS 1, 2, 3 and 4. The modularity performance shows better or sometimes same for the ADNG algorithm except in the case of DATA SET 4.

Modularity and Conductance determine the strength of the community structure. For a good community structure, high modularity, and low conductance should be present (Yang and Leskovec 2012). So, the difference between modularity and conductance can be shown as the strength of community structure, i.e., higher the difference, greater is the strength of the community. Fig. 16 shows the performance of the proposed algorithm on the basis of the strength of community structure shows better for DATA SETS 1, 2, 3 and 4.

The communities formed based upon the proposed algorithm has been compared with the communities formed by other algorithms considered for comparison on the basis of similarity measurements, i.e., Normalized mutual information (NMI) measure and the Rand Index measure. Table 6 shows the values that lies between 0 and 1, which proves that none of the communities formed by the proposed algorithm is similar to the communities formed after applying the LEC and FG algorithms on all the four datasets.

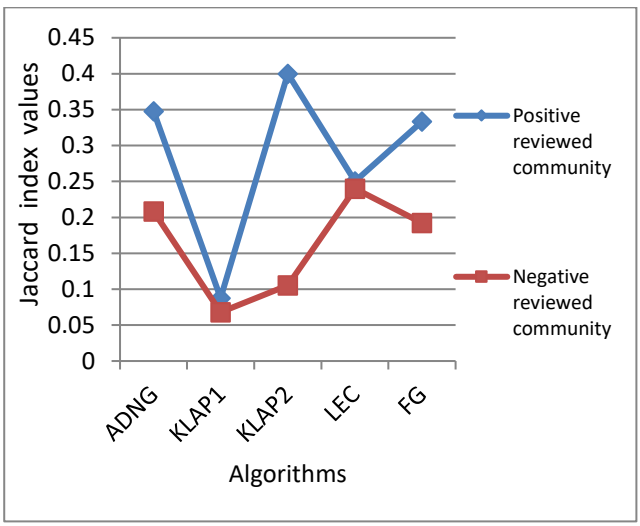

Fig. 13 Jaccard index result for DATA SET 1

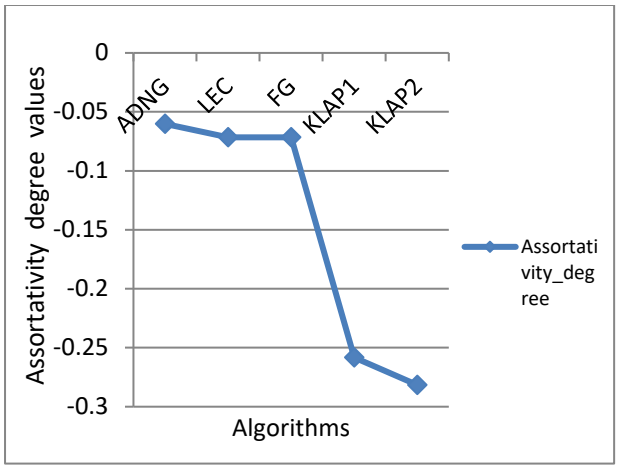

Fig.14 Assortativity degree values for DATA SET 1

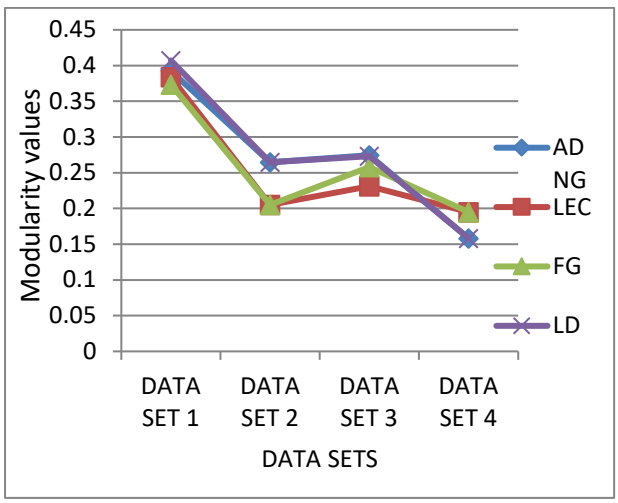

Fig.15 Comparison of modularity metric performance for the ADNG algorithm with LD, LEC and FG algorithms

Fig.16 Strength of the communities formed by the ADNG algorithm

Table 6 Similarity measurements of the proposed algorithm with the LEC and FG algorithms with the help of NMI measure and Rand Index measure

DATA SETS	Compared algorithms	NMI	RAND INDEX
DATA SET 1	LEC	0.5532844	0.8051948
	FG	0.6823823	0.8694463
DATA SET 2	LEC	0.4684876	0.7794118
	FG	0.5178743	0.7794118
DATA SET 3	LEC	0.624236	0.6666667
	FG	0.8418344	0.8333333
DATA SET 4	LEC	0.3643407	0.6111111
	FG	0.3643407	0.6111111

3.6 Comparison with other type of networks

Real world applications are represented by the scale-free networks in a much better way than by the Erdos-Renyi (ER) models network. Erdos-Renyi networks are random networks, in the sense that the edges or the links are created between the nodes by a random variable. Random networks are those networks which consist of connections between the nodes, in a random manner, and most of the nodes consist of links or edges, approximately of the same number. Scale-free networks are those networks in which some nodes will be having a fewer number of links while some other will be having a tremendous number of links and the distribution of nodes links follows a power law, in other sense, that network does not have any "scale". The scale-free networks can be drawn by the Barabasi-Albert model algorithm. The ADNG algorithm's performance has been compared with LEC and FG algorithms by applying those algorithms in scale-free network graphs of 1308 nodes whose average degree is 1.998471, in the synthetic graphs of 100 nodes whose average degree is 20.04, and also in NG (Newman-Girvan) graphs which consist of 100 nodes and its average degree is 19.66. Fig. 17 shows the scale-free network of 1308 nodes, a random network of 100 nodes, and of NG graph with 100 nodes performance compared with other algorithms, i.e., with LEC and FG algorithm, in terms of strength of the formed communities, which was determined by the difference between the modularity and conductance values. Therefore, the graphs and the tables in this discussion section show ADNG algorithm's performance better than the other compared algorithms.

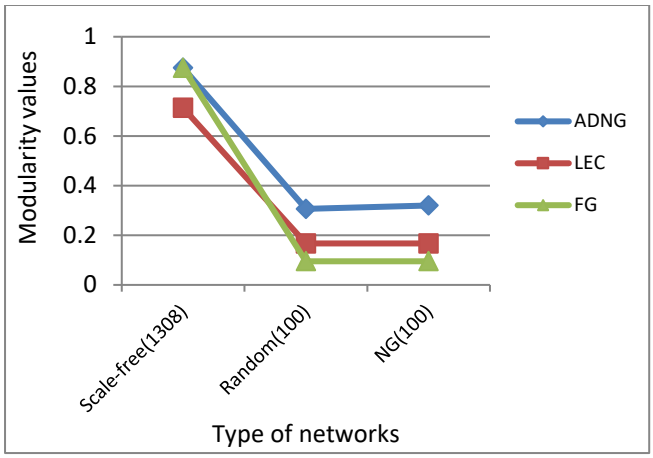

Fig.17 Strength of the communities formed by the ADNG algorithm

4 Conclusion and future work

In today's era, where social networks are considered as the most convenient platform to update people about the practical things, detection of communities from those networks helps to know about the circumstances of any event, individual's reviews about any service, or for any new product launched. The proposed ADNG algorithm detects the majority of positive and negative reviewers for the support of any event, product, etc. The detected communities convey the behavioral information among the entities. The ADNG algorithm has been applied for the real-time data sets of political speech video, for the data sets of a product launched by online shopping websites and it successfully detects the communities for the positive and negative supporters for those data sets. The ADNG algorithm results for the collected datasets have been compared with other community detection approaches by considering the various scoring functions and the results after those comparisons give maximum time positive results. Datasets have been collected from the most popular social network - Facebook, in which a post of any video, any event or for any service which had got tremendous likes, shares, and comments.

Anyways, the proposed ADNG algorithm helps to detect the communities from the social network and can make the work easier for the research communities in the future. For the future research work, community detection can be done by considering a greater number of data sets and by considering data from other social networks also, with some more statistical measurements and getting more relevant results.

Availability of data and materials

The datasets utilized for this research work are all real time data collected from Facebook. These datasets are available from corresponding author only on reasonable request.

Funding

The work was supported by University Grants Commission- Maulana Azad National Fellowship.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MK have conceived the idea, performed the literature survey, analysis and implementation. AB provided guidance, supervised the research work and contributed to the writing of the manuscript. Both authors read and approved the final manuscript.

Acknowledgement

We are highly thankful to University Grants Commission (F1-17.1/2014-15/MANF-2014-15-MUS-WES-38675/ (SAIII/ Website) & February-2015) for providing Maulana Azad National Fellowship, for conducting this research work.

References

Alamsyah A, Rahardjo B, Kuspriyanto (2014) Community Detection Methods in Social Network Analysis. Adv Sci Lett 20: 250-253

Amaral LAN, Scala A, Barthelemy M, Stanley HE (2000) Classes of small-world networks. Proc Natl Acad Sci 97:11149-11152

Amelio A, Pizzuti C (2015) Is Normalized Mutual Information a Fair Measure for Comparing Community Detection Methods? In: IEEE/ACM International Conferences on Advances in Social Network Analysis and Mining, Paris, France, pp 1584-1585

Azaouzi M, Rhouma D, Romdhane LB (2019) Community detection in large-scale social networks: state-of-the-art and future directions. Soc Netw Anal Min 9:23

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks, J Stat Mech Theory Exp. https://10.1088/1742-5468/2008/10/P10008

Barabasi AL, Albert R (1999) Emergence of Scaling in Random Networks. Science 286: 509-512

Choudhury D, Bhattacharjee S, Das A (2013) An Empirical Study of Community and Sub-Community Detection in Social Networks Applying Newman-Girvan Algorithm. In: 1st International Conference on Emerging Trends and Applications in Computer Science (ICETACS), Shillong, India, pp 74-77

Ciglan M, Norvag K (2010) Fast detection of size-constrained communities in large networks. In: International Conference on Web Information Systems Engineering, WISE, Hong Kong, China, pp 91-104

Clegg GR, Cairano-Gilfedder CD, Zhou S (2009) A critical look at power law modeling of the Internet. Comput Commun

33: 259-268
Fortunato S (2010) Community detection in graphs. Phys Rep 486: 75-174
Fortunato S , Hric D (2016) Community detection in networks: A user guide. Phys Rep 659:1-44
Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99:7821-7826
He K, Li Y (2017) Hidden Community Detection in Social Networks. arXiv:1702.07462v1
Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78: 046110
Li C, Chen G (2003) Network connection strengths: Another power-law? arXiv:cond-mat/0311333
Lumley T, Scott A (2013) Rank Tests with Data from a Complex Survey. In: Proceedings of 59th ISI World Statistics Congress, Hong Kong, pp 900-905
Luo F, Wang JZ, Promislow E (2008) Exploring local community structures in large networks. Web Intel and Agent Systs: An Int J 6:387- 400
Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89: 208701(1-4)
Newman MEJ, Girvan M (2003) Finding and evaluating community structure in networks. Phys Rev E 69: 026113(1-16)
Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103: 8577-8582
Nguyen NP, Dinh TN, Shen Y, Thai MT (2014) Dynamic Social Community Detection and Its Applications. PLoS One 9: e91431
Noldus R, Mieghem PV (2015) Assortativity in Complex Networks. J of Complex Networks 3:507-542
Porter MA, Onnela JP, Mucha PJ (2009) Communities in Networks. Not Am Math Soc 56: 1082-1097
Thedchanamoorthy G, Piraveenan M, Kasthuriratna D, Senanayake U (2014) Node assortativity in complex networks: An alternative approach. Procedia Comput Sci 29: 2449-2461
Traag VA, Waltman L, Van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-connected communities, Sci Rep, 9: 5233
Vasudevan M, Deo N (2012) Efficient community identification in complex networks. Soc Netw Anal Min 2:345–359
Vazquez A (2003) Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. Phys Rev E 67:056104(1-15)
Wang M, Wang C, Yu JX, Zhang J (2015) Community Detection in Social Networks: An In-depth Benchmarking Study with a Procedure-Oriented Framework. In: Proceedings of the VLDB Endowment, 41st International Conference on Very Large Data Bases, Kohala Coast, Hawaii, pp 998-1009
Xu B, Deng L, Jia Y, Zhou B, Han Y (2013) Overlapping Community Detection on Dynamic Social Network. In: IEEE Sixth International Symposium on Computational Intelligence and Design, Hangzhou, China, pp 321-326
Yang J, Leskovec J (2012) Dening and Evaluating Network Communities based on Ground-truth. In: Proceedings of 2012 IEEE International Conference on Data Mining (ICDM), Brussels , Belgium, pp 745-754
Yin H, Benson AR, Leskovec J (2018) Higher-order clustering in networks. Phys Rev E 97:052306

Figures

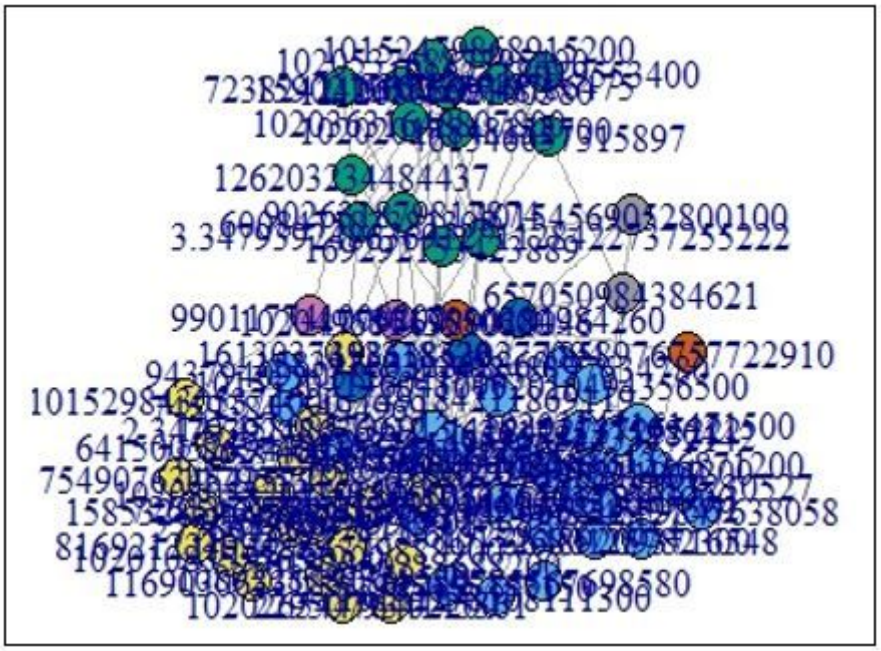

Figure 1

Graph with vertices label formed for DATA SET 1

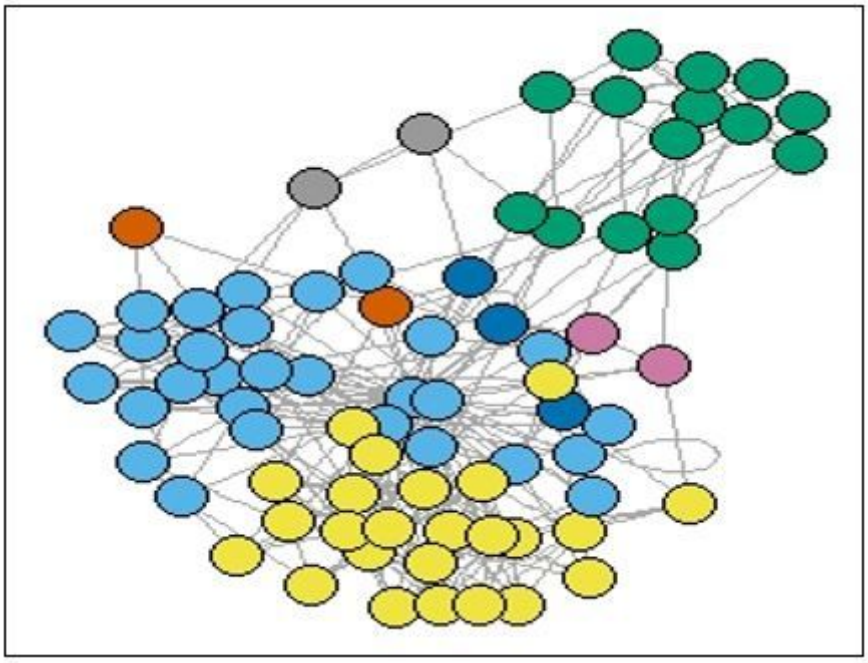

Figure 2

Graph without vertices label formed for DATA SET 1

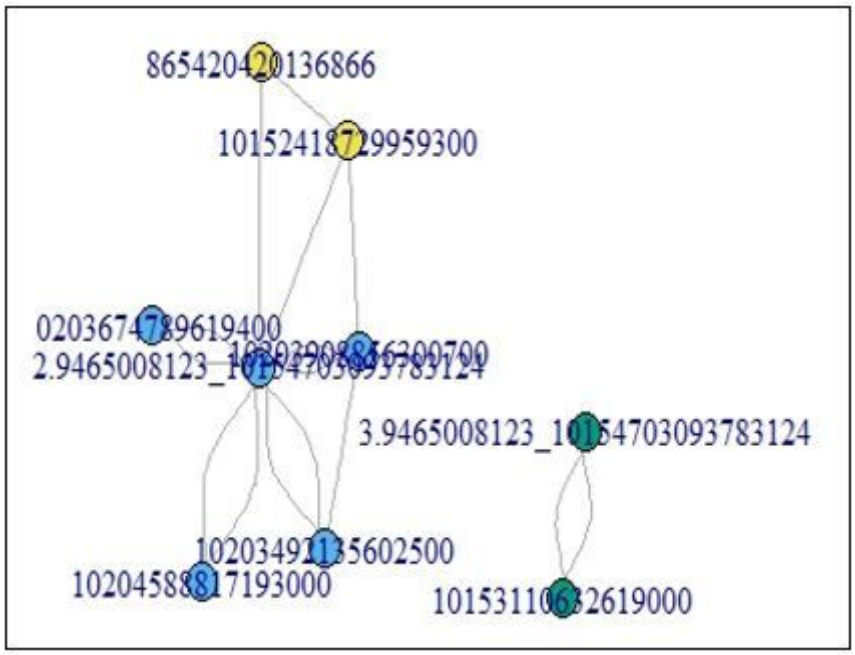

Figure 3

Graph with vertex label formed for DATA SET 2

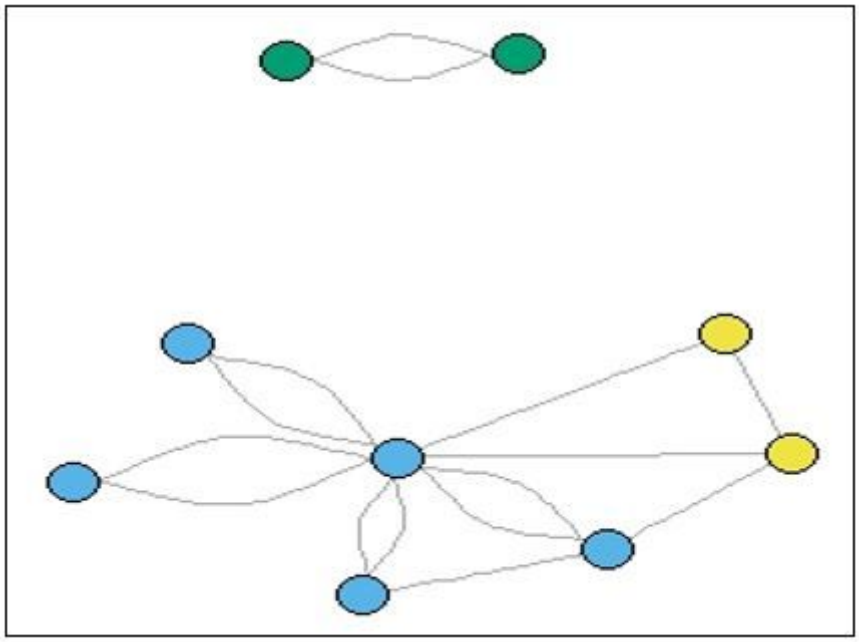

Figure 4

Graph without vertex label formed for DATA SET 2

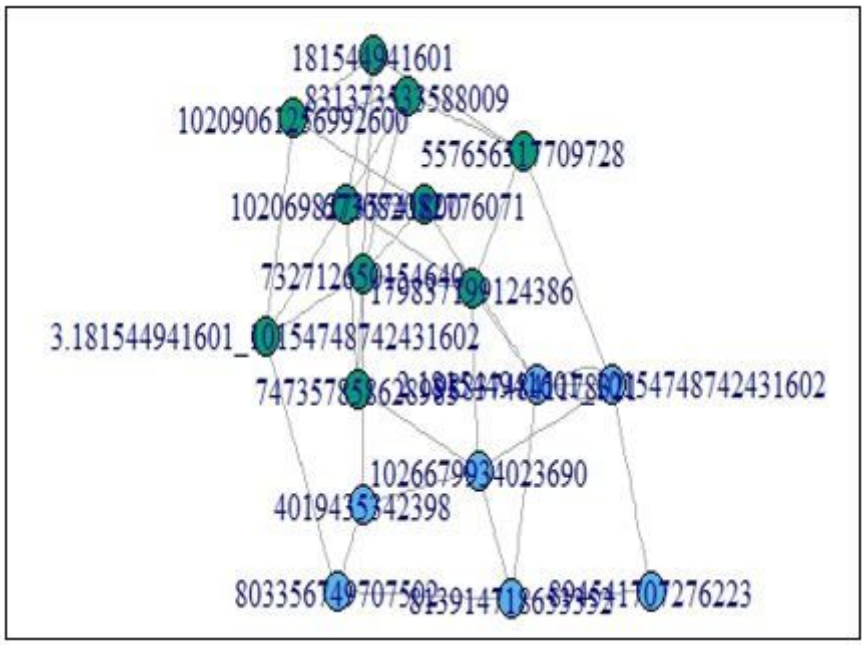

Figure 5

Graph with vertex label formed for DATA SET 3

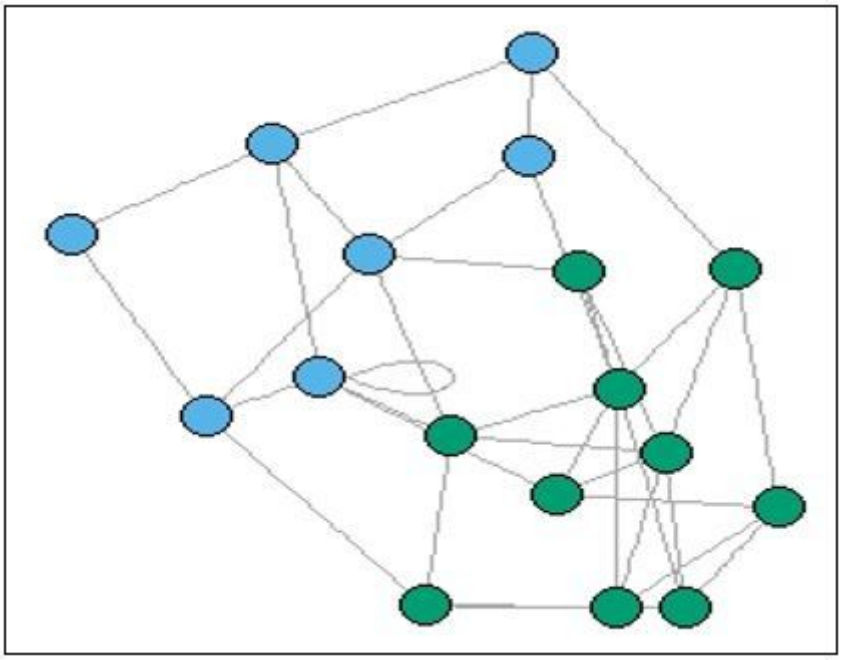

Figure 6

Graph without vertex label formed for DATA SET 3

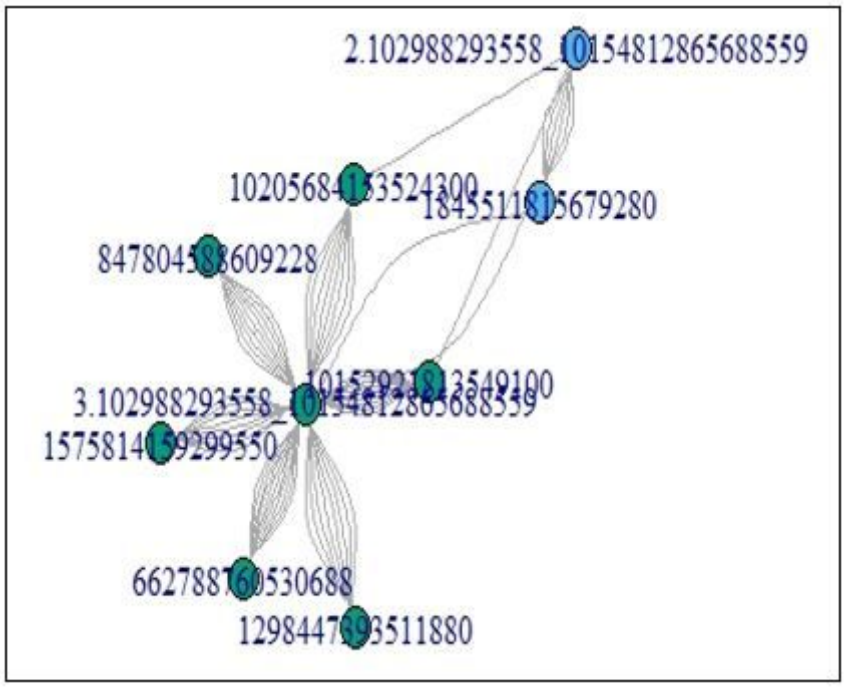

Figure 7

Graph with vertex label formed for DATA SET 4

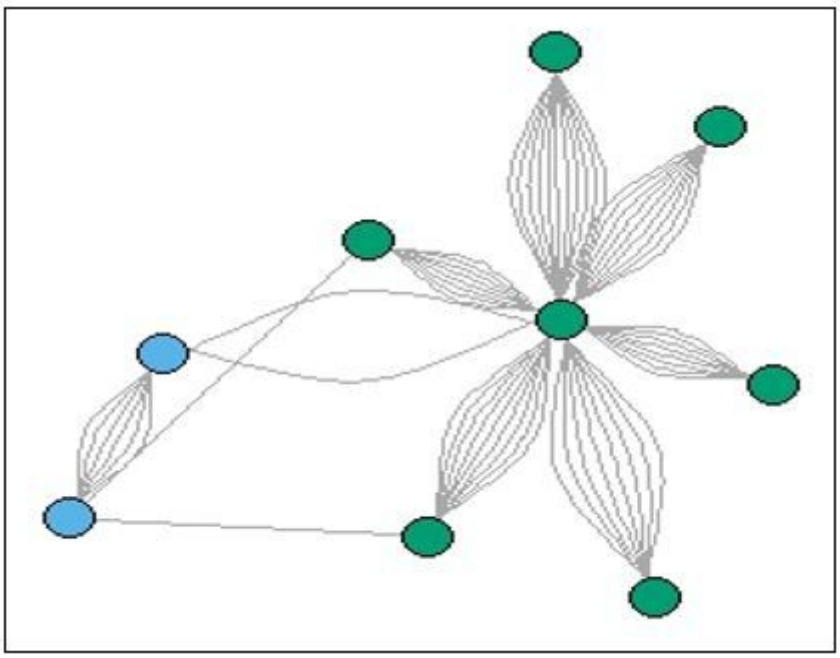

Figure 8

Graph without vertex label formed for DATA SET 4

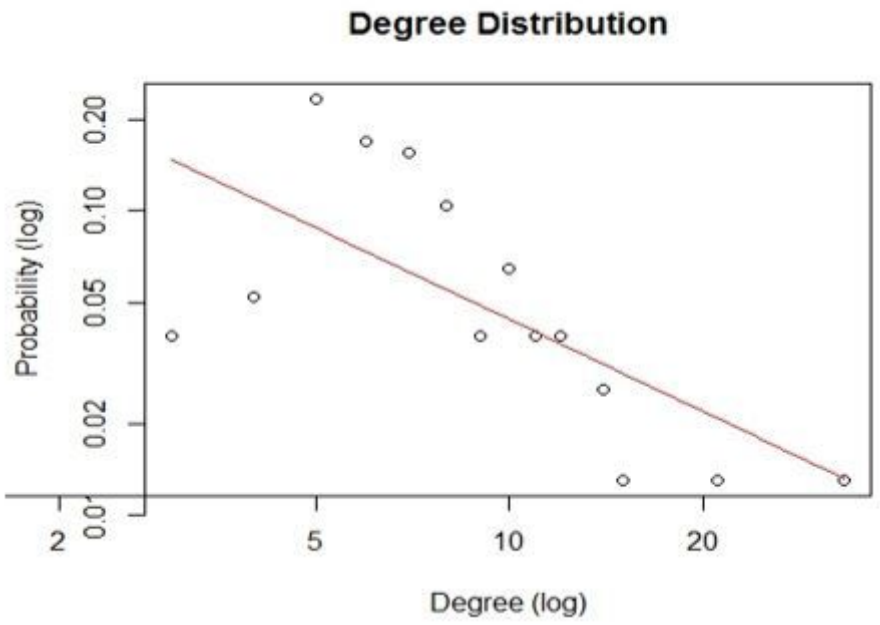

Figure 9

Plot for power law degree distribution for DATA SET 1

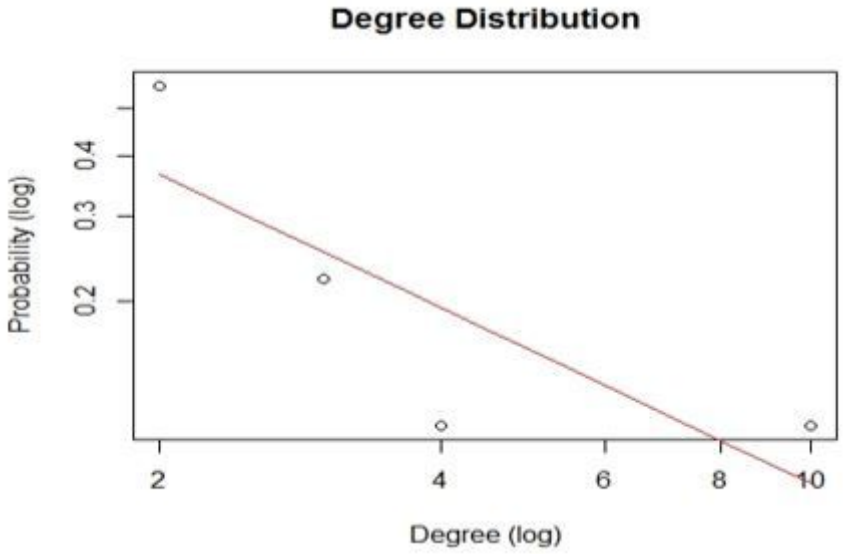

Figure 10

Plot for power law degree distribution for DATA SET 2

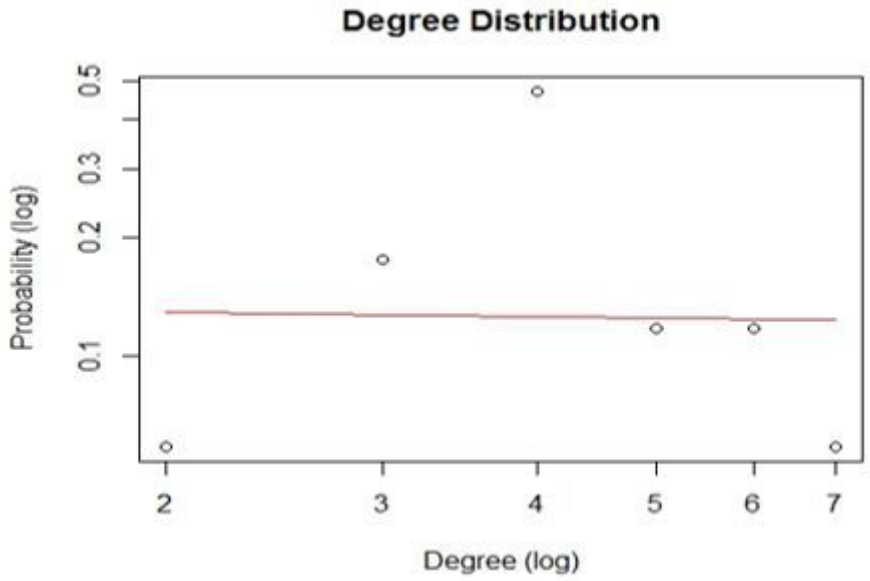

Figure 11

Plot for power law degree distribution for DATA SET 3

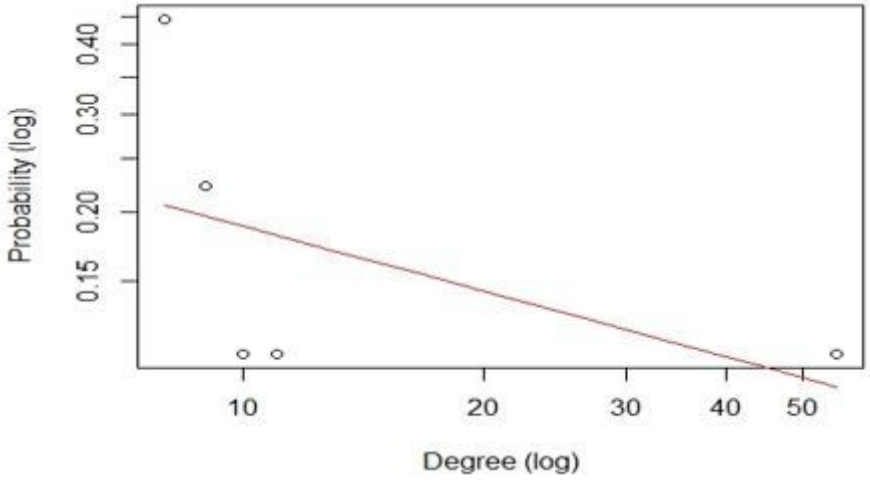

Figure 12

Plot for power law degree distribution for DATA SET 4

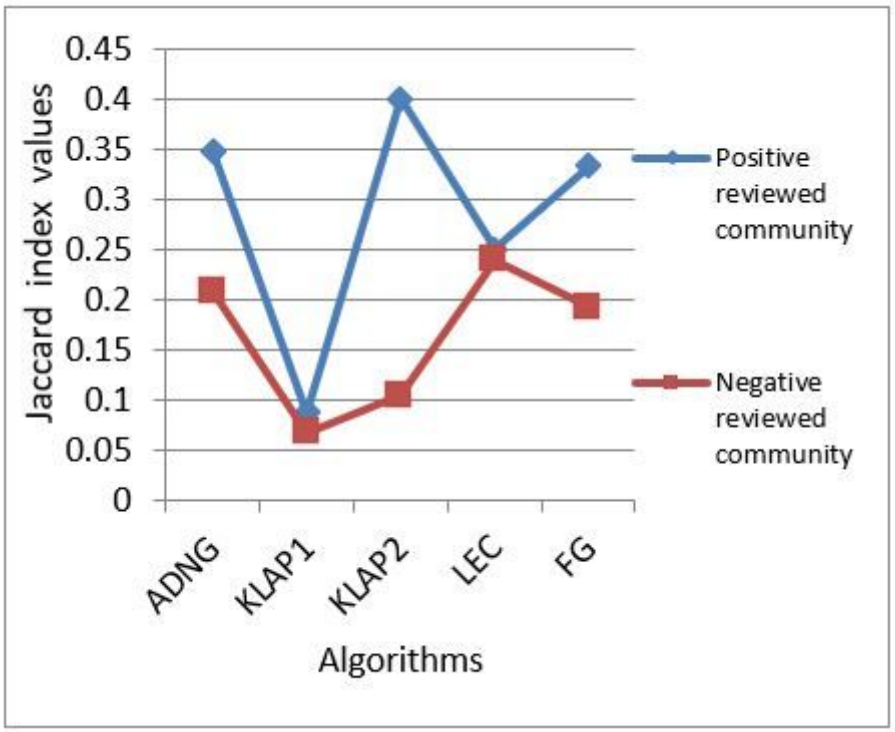

Figure 13

Jaccard index result for DATA SET 1

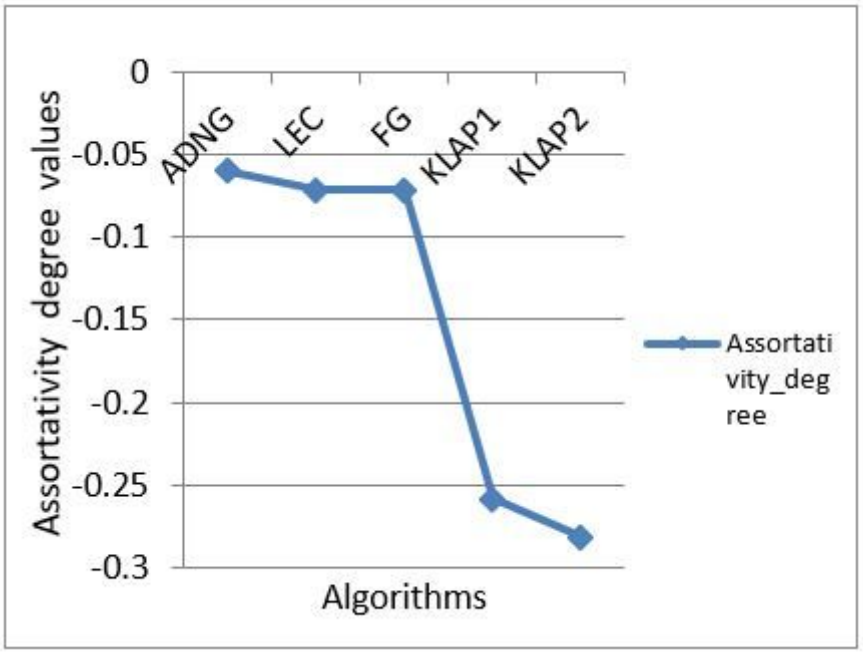

Figure 14

Assortativity degree values for DATA SET 1

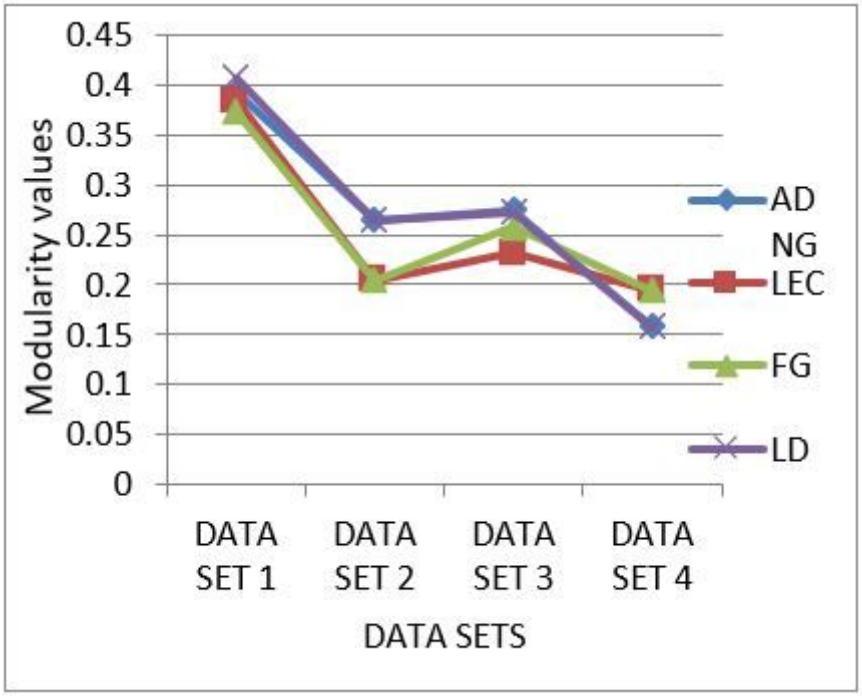

Figure 15

Comparison of modularity metric performance for the ADNG algorithm with LD, LEC and FG algorithms

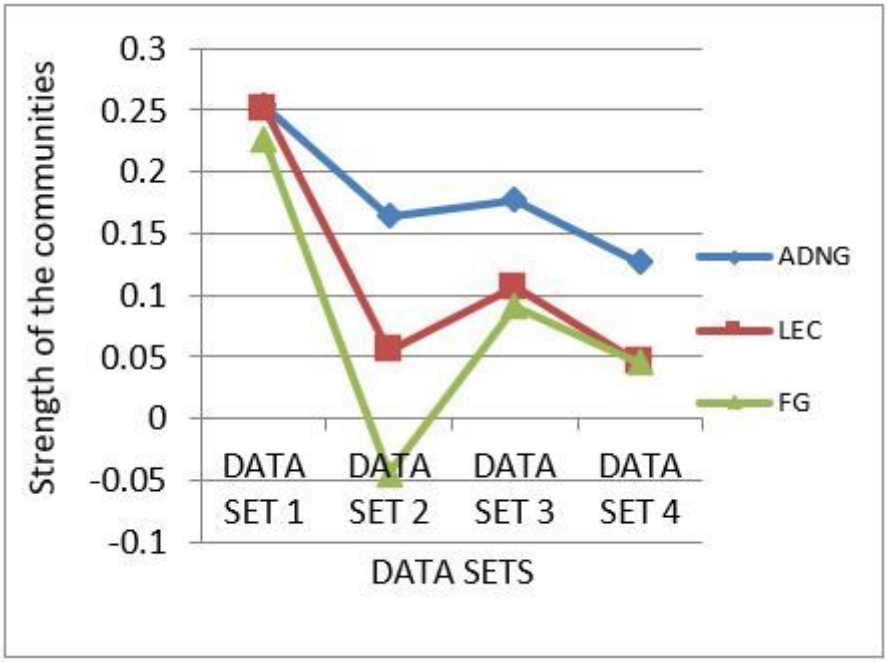

Figure 16

Strength of the communities formed by the ADNG algorithm

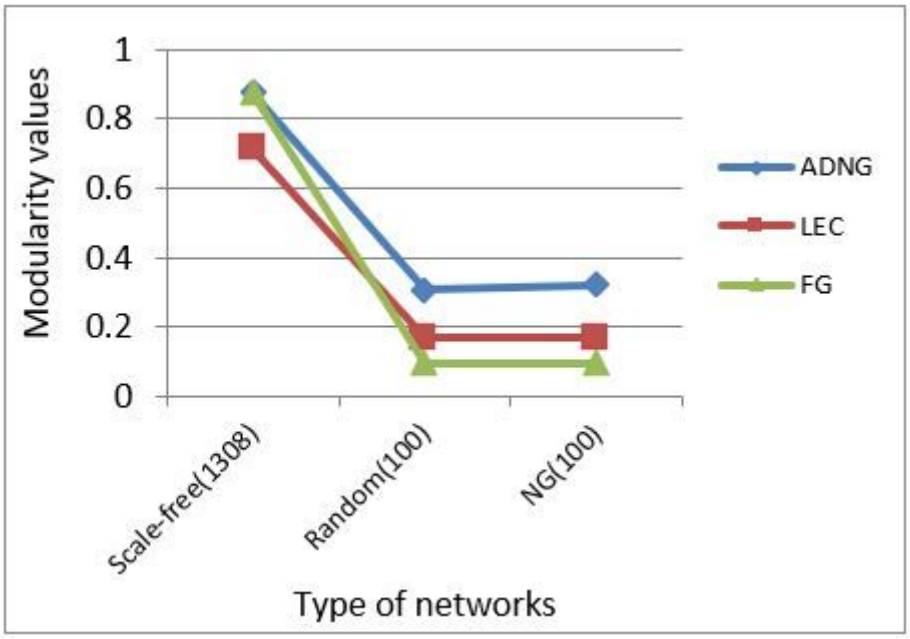

Figure 17

Strength of the communities formed by the ADNG algorithm