Identification of Influential Nodes in Social Network: Graph Approach

Rajnish Kumar[1], Laxmi Ahuja[2], Suman Mann[3], Sanmukh Kaur[4]
[1,2]Amity Institute of Information Technology, Amity University, Uttar Pradesh-201313
[3]Maharaja Surajmal Institute of Technology, New Delhi, India
[4]Amity School of Engineering & Technology, Amity University, Uttar Pradesh-201313
Rajnishghose@gmail.com, lahuja@amity.edu, sumanmann@msit.com, sanmukhkaur@gmail.com

Abstract— The identification of nodes that are actually influencing other nodes is one of the core and fundamental research problems in the social network data sciences. There are multiple proven techniques available globally that depict the influence of the root node on the remaining nodes in the given network. Although methods and methodology is already proven and multiple experiments have been carried out in this area, the current research work is an extension of the influential node identification problem with the perspective of graph traversal. To prove the methodology and new approach a tree and AUC (area under the curve) - based analysis method has been proposed, TArank in which the information collected from the breadth-first search tree is used to identify influential nodes.

From the obtained results and analysis, the effectiveness of TARank in the identification of influential nodes has been validated. The concentration of data at a particular segment of the plot is observed. Additionally, the tree view of the analysis predicts the root and leaf node thus helping in the identification of the most influential nodes in the online social media network.

Keywords: Influential Node, PCC, TARank, Degree Centrality, PCC

1. Introduction

Influential node identification has a significant role in understanding the network structure and functions of online social media like Facebook or Twitter. The BFS tree help in the identification of the root node. The evaluation of each node influence involves designing a Breadth-first search (BFS) tree determining the final target node being termed as a root node [1]. With the help of the BFS tree, a curve may be created with the x-axis and y-axis representing the level number and cumulative scores of all the nodes visited during the traversal process respectively. The area plotted under the curve value is considered the target node's final influence score [2].

The network graph is a popular type of data structure that presents a complete top-down along with bottom-up approaches to make sense of different collaborative systems including social systems, biological systems, traffic methods, and communication systems which are highly impacted due to small section of influential nodes, popularly known as influential spreaders [3]. These generated nodes play a critical role and may considerably improve our knowledge of the given system. For example, the successful and proper detection of influential nodes permits us to control the spread of outbreaks, design a valid marketing plan, and avoid the power system from failing [4].

Since finding the influential node is a common network analysis method, globally a number of measurement techniques are available for evaluating the importance of each node in a network from different perspectives based on some experimental analysis and justification [5]. Methods like H-Index, as well as degree centrality

(DC), are classic neighborhood-based techniques for the evaluation of centrality. Closeness centrality (CC), load centrality (LC), betweenness centrality (BC), and information centrality (IC) are considered to be global methodologies for path-based centrality measurement methods [5]. The iterative alternative method is an example of eigenvector centrality (EC) and PageRank (PR) is a representative method that evaluates the node centrality [2]. Despite multiple research carried out over the period on developing centrality methods for identifying influential nodes, it is yet not determined which method is best among these for centrality measure across different types of networks in various domains for online social networks [6]. There are several nodes on the top stages of the tree. The nodes at the top level of the BFS tree fall over the local area of the root node. The other assessment methods including DC and classical methods, employ the number of nodes in the local area as a typical criterion for the recognition of influential nodes [1, 7].

Graphs have been plotted as per the available first k entries in the score vector, with the x-axis and y-axis denoting the level number, and the corresponding score respectively. The area under the curve (AUC) indicates the overall centrality score [8, 9], and the general approach for the identification of influential nodes is termed as TARank (Tree & AUC-value-based rank). Further, based on the outcome of different tests on the networks around different domains, it has been determined that the proposed method is substantially superior to widely used centrality measures for influential node detection.

2. Original network and details

The network diagram is plotted for the Facebook dataset as shown in Fig: 1. As it is a multi-dimensional dataset, it is not feasible to show every node and datapoint. Thus, the network diagram for the dataset has been plotted as shown in Fig: 1. All 14 nodes are displayed in the network diagram along with the direction of their relationship. For some nodes, the relationship is unidirectional whereas for some nodes the relationship is bidirectional. The basis of the network diagram is degree and based on the degree value, the node relationship is shown.

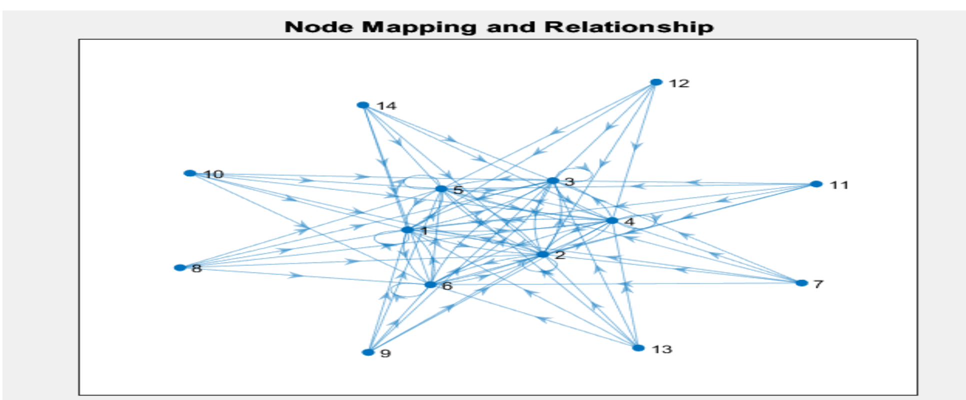


Fig :1 Network Diagram

The tree view and datatype are shown in Fig:2 (a) and (b) have been plotted for the Facebook dataset. The tree view consists of all the available nodes including a root node and leaf nodes with p values assigned to each of the nodes. Since the current work is concentrated on the resolution of identification of the most influential node, hence for the experimental purpose, only an integer-type data set is taken in the scope of activity. The dataset is consisting of 14 variables along with 99002 observations.

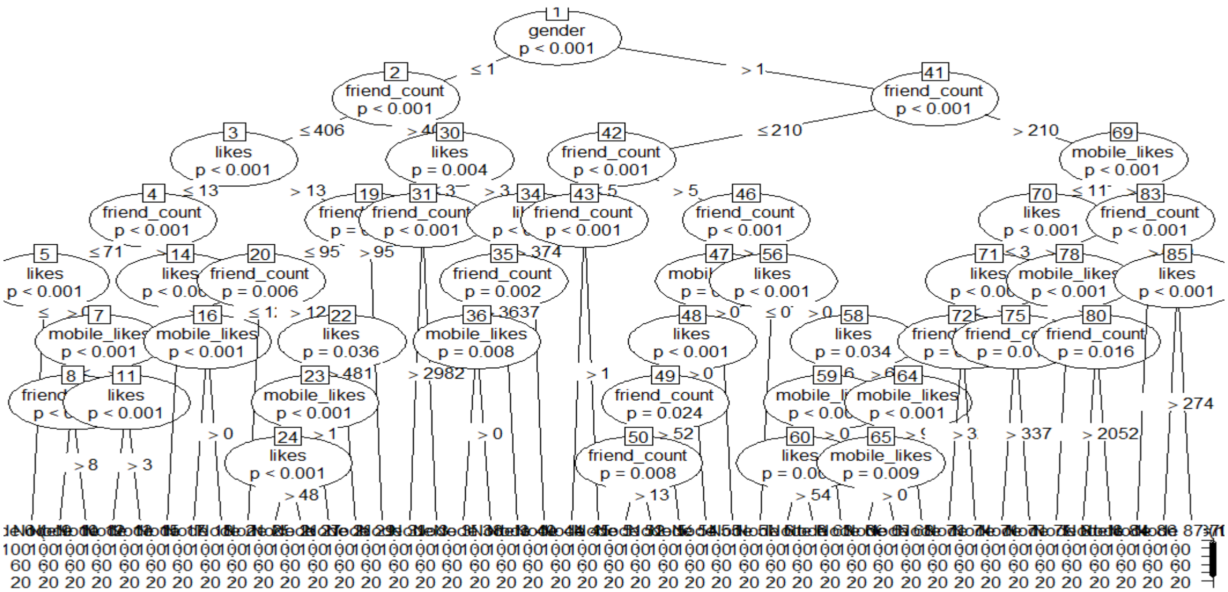


Fig : 2 (a) Tree view of Facebook dataset

```
> str(facebook)
'data.frame':   99002 obs. of  14 variables:
 $ age                  : int  14 14 14 14 14 14 13 13 13 13 ...
 $ dob_day              : int  19 2 16 25 4 1 14 4 1 2 ...
 $ dob_year             : int  1999 1999 1999 1999 1999 1999 2000 2000 2000 2000 ...
 $ dob_month            : int  11 11 11 12 12 12 1 1 1 2 ...
 $ gender               : int  1 2 1 2 1 1 1 2 1 1 ...
 $ tenure               : int  266 6 13 93 82 15 12 0 81 171 ...
 $ friend_count         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ friendships_initiated: int  0 0 0 0 0 0 0 0 0 0 ...
 $ likes                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ likes_received       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ mobile_likes         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ mobile_likes_received: int  0 0 0 0 0 0 0 0 0 0 ...
 $ www_likes            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ www_likes_received   : int  0 0 0 0 0 0 0 0 0 0 ...
```

Fig : 2 (b) Data Type

3. TARank workflow

The traversal follows from the root node to the leaf node. The traversal for the whole branch starts with each node which is considered the root node. A breadth-first search (BFS) tree is designed for the given set of data. The arbitrary centrality measures help in identifying the initial score for each node in the tree within the system [1, 10].Accordingly, a score vector is obtained for each BFS tree. The curve has been plotted by the first k values in the score vector and the AUC value thus obtained is used as the base for centrality score computation in TARank. With this AUC value, the influence of each node is obtained within the tree [10].

3.1 An Overview of the TARank Methodology

a) The raw data is processed and normalized as per the requirement.

b) For the identified response network, each node within the network is considered a root node, and the navigation of each node within the graph in a given level order creates a BFS tree. The centrality measure mechanism like DC helps in identifying the initial significance score.

c) For each of the BFS trees, a score vector known as the aggregate score vector is computed. For the said score vector, h is the biggest level number (the level number of the root node is 1) in the system [11]. The i-th element of the available network within the score vector is determined as the summation of totals of all nodes in the network where the levels are not larger than i. Depending upon the user-specified value, $k(1 \leq k \leq h)$ helps in computing the score vector. The AUC value helps in matching the centrality value for assessing the node influence. Finally, with the help of TARank according to the AUC scores, the most influential nodes are identified [3, 12].

3.2 Performance evaluation mechanism

The experimental activity was performed on real-time data. In order to check and validate the performance of TARank, the correlation coefficient designed by Kendall's tau is applied as the most effective method for the evaluation of performance. Statistical correlation is measured between the node influence obtained SIR (susceptible-infected-recovered) by the well-known spreading model and the AUC score in TARank. In the SIR stretching model, every node affects the existence of another node in the system [13]. The tree view of the Facebook dataset is shown in Figure 2 (b). The basic dataset structure and data type are found in Figure 2(c). To do the evaluation of TARank, Kendall's tau correlation coefficient was employed as an indicator for performance for AUC Score in TARank and SIR model (susceptible-infected-recovered).

There is a total of eight methods currently available that are associated with TARank. These are listed as DC, BC, LC, IC, H-index, PR, EC, and CC. For computing the initial score of each node, Rank is a general framework for any centrality measure. For the experiments, for AUC value calculation, select of following eight methods and denote them as TARank (DC, BC, LC, IC, H-index, PR, EC, and CC) respectively. The plot shows the improvement in the performance of TARank for each participant [1, 14].

4. Data source and processing

The test data was picked from the Kaggle for experimental purposes. Since the focus of the activity was on social media Facebook dataset has been taken as a base for all experimental analysis and approach establishment. The source data can be found at the following URL :

https://www.kaggle.com/datasets/sheenabatra/Facebook-data

It consists of 99903 entries with 14 columns [Fig: 2-b]

As a part of the experimental analysis RStudio (Version 1.4.1106) and Python 3.7 on the Windows/Linux platform is been used. The data processing and experimental analysis have been performed on both RStudio and Python. The steps of data processing are listed below:

Steps Involved :

a) Download the Facebook data from the Kaggle site
b) Copy the data into the respective test environment
c) Import the data in the RStudio and Python for further processing
d) Make the data frame of the raw data for further processing
e) Do the data massaging for the data frame as per requirement using the iloc command in Python
f) Convert the data into experimental data by assigning nodes for each of the column headers.
g) Convert the whole data set into the training and testing dataset for further usage.
h) Verify and validate the finalized data from all experimental analysis

5. Experimental Process

A residual is the measurement of the vertical distance of any point from the regression line. The residual error is the difference between a group of values under observation and their arithmetical mean. The outcome indicates the concentration of data. It is the error between a predicted value and the observed actual value for the given set of data.

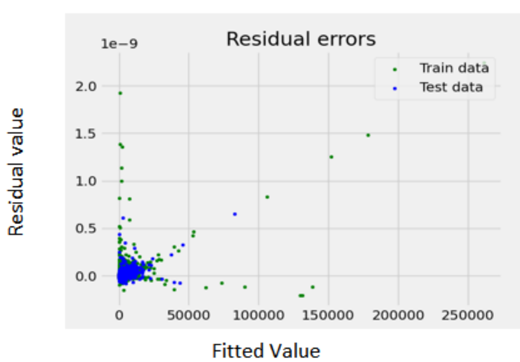


Fig:3 (a)

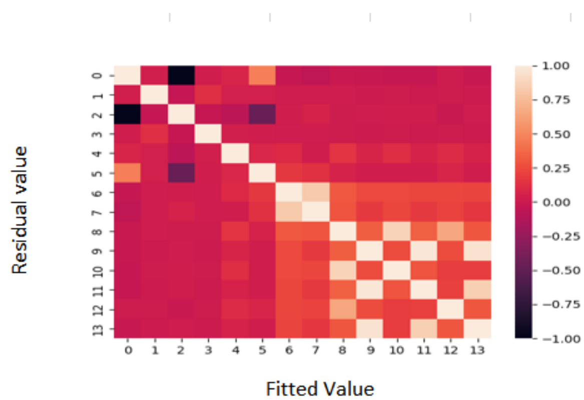


Fig:3(b)

Fig : 3 Residual Error

From the dataset available, the concentration of training data and testing data [Fig:3-a] at a particular point is seen where a few data are seen in the space other than concentration. Additionally, the calculation of the coefficient and variance score of the complete dataset [Fig:3-b] is also done.

Computed Eigen Value and Eigenvector for the given dataset :

```
Eigenvalues is  [ 3.73725422e+00  2.34745235e+00  1.82398341e+00  1.38498874e+00
  1.13284670e+00  6.31399252e-01  9.39679114e-01  8.25616140e-01
  8.70223280e-01  1.68609431e-01  1.38088783e-01 -2.56560512e-16
  1.33068375e-10  1.34800303e-11]
```

(4-a)

```
Eigenvector is  [[ 1.40985569e-02  6.24260930e-01 -9.54658458e-02 -3.67476576e-02
   5.87801788e-02  3.02378864e-01  6.07171031e-02 -3.21359212e-02
   1.28626250e-03 -5.02953547e-03 -1.04037024e-03 -7.07106781e-01
  -3.68580382e-08  1.26829153e-07]
 [-1.24350067e-02  5.50062466e-02  3.74375648e-02  1.47298687e-02
  -6.84014915e-01  2.12678041e-02  1.11984391e-01  6.62346713e-02
  -7.14087256e-01  4.38956387e-03  2.23765594e-03 -1.21980173e-14
  -1.67809647e-08  8.33525047e-09]
 [-1.40985569e-02 -6.24260930e-01  9.54658458e-02  3.67476576e-02
  -5.87801788e-02 -3.02378864e-01 -6.07171031e-02  3.21359212e-02
  -1.28626250e-03  5.02953547e-03  1.04037024e-03 -7.07106781e-01
   9.64760946e-09  1.27689883e-07]
```

(4-b)

Fig : 4 Eigen Computation

Eigen value [Fig : 4-a] and Eigen vector [Fig :4b]- for the given dataset is computed. Following structured method is applied for the calculation of eigenvalue and eigenvector.

Step 1: 3×3 matrix elements in the respective input field are premeditated.

Step 2: “Calculate Eigenvalues” or “Calculate Eigenvectors” to get the results.

Step 3: Finally, plot the eigenvalues or eigenvectors of the matrix

d) PCA calculation: On the given dataset PCA analysis is done as a part of an experiment where training and testing data set are used. The training of our model is done on the training dataset and further tests the model’s suitability on the testing dataset. A confusion matrix is created along with variance. A logistic regression approach has been employed to establish the methodology [15].

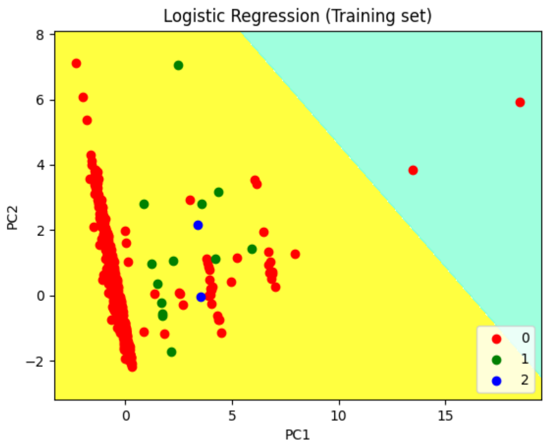


Fig:5(a)

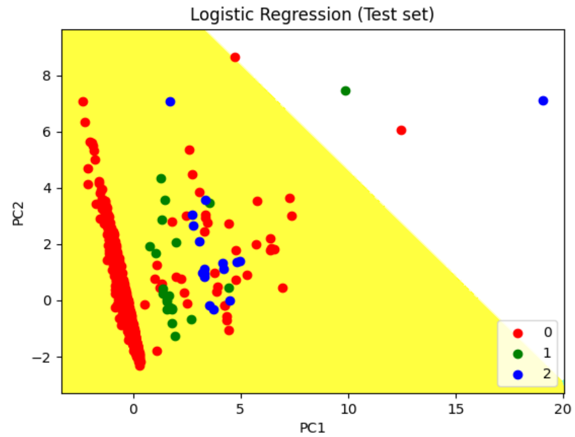


Fig:5(b)

Fig : 5 PCA Calculation

The current experimental data consists of multiple dimension and with the current dataset it is not possible to derive the accuracy of the experimental result [16]. The Facebook dataset is distributed in Training and Testing dataset. Post-distribution of the dataset, Logistic regression for the training dataset is computed[Fig:5(a)]. Once the model is trained then the model is tested on the testing dataset [Fig:5(b)]. Hence, the dimensionality reduction was done, and finally, PC1 (Principal component explained) and PC2 (Principal component remaining) were calculated, PC1 is the linear combination with the largest possible explained variation, and PC2 is the best of what's left.

e) AUC calculation: AUC, (Area under the curve) is the ROC (receiver operating characteristics) curve. It is being used for the verification and validation of model predictions using a probabilistic framework approach [15]. The computation of positivity is done in terms of false positive and true positive and a graph is plotted between the true and false positive rates [5]. AUC calculation is depicted in Fig:6 and Table 1 respectively.

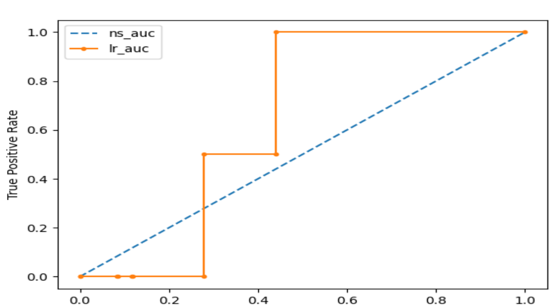


Fig:6(a)

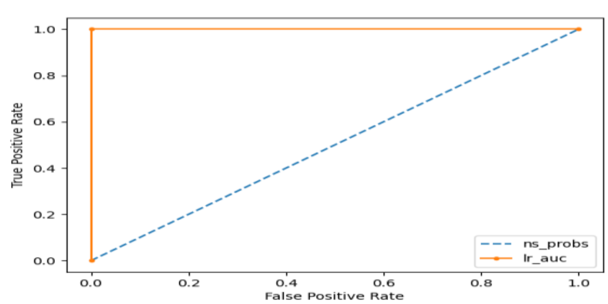


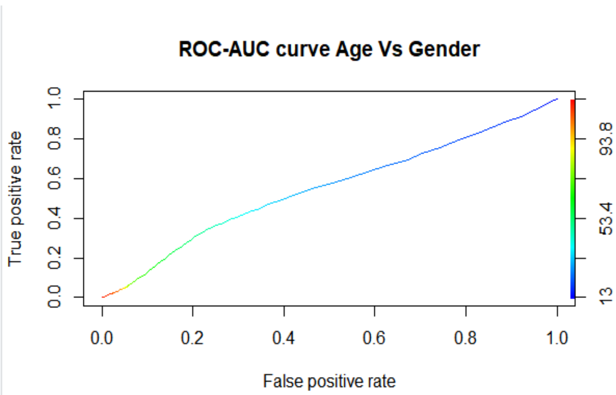
Fig:6(b)

Fig :6 AUC Calculation

The computed values of AUC for both data variance items are as below

Table 1 AUC Calculation

Fig 6-a		Fig 6-b	
Ns_auc: ROC_AUC	0.500	Ns_auc : ROC_AUC	0.512
Lr_auc: ROC_AUC	0.642	Lr_auc : ROC_AUC	0.641



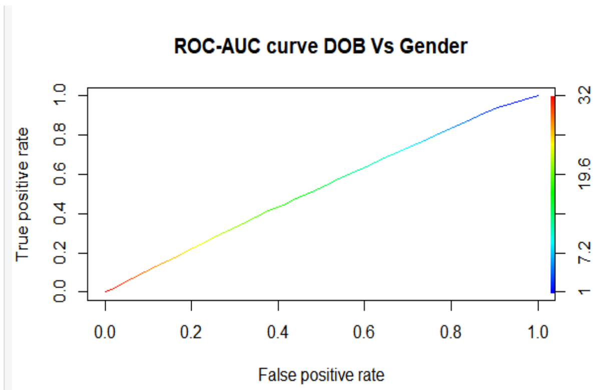
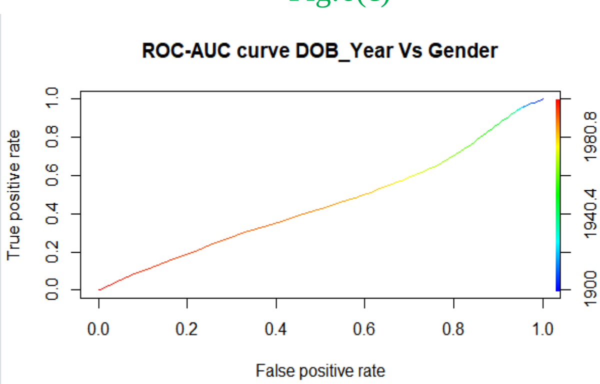
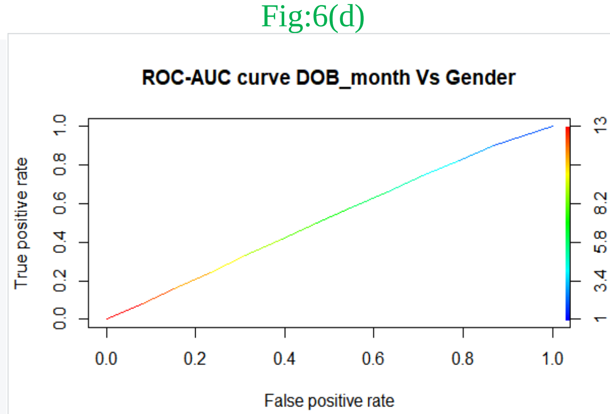


Fig:6(c)

Fig:6(d)





VOLUME: 07 ISSUE: 08 | AUGUST - 2023 SJIF RATING: 8.176 ISSN: 2582-3930

Fig:6(e)

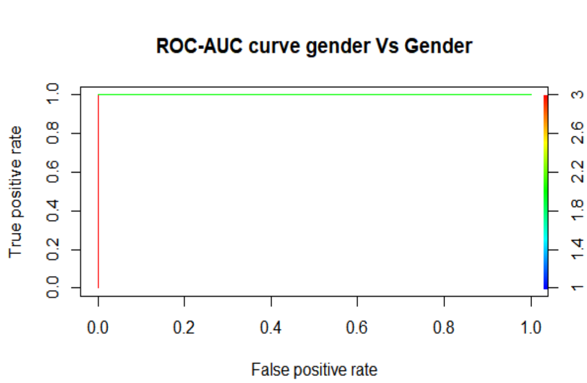


Fig:6(f)

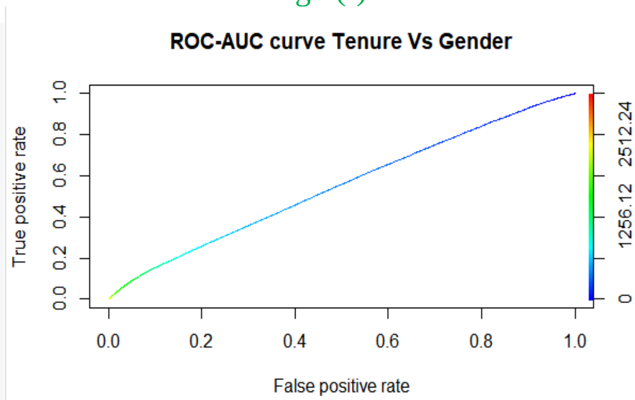


Fig:6(g)

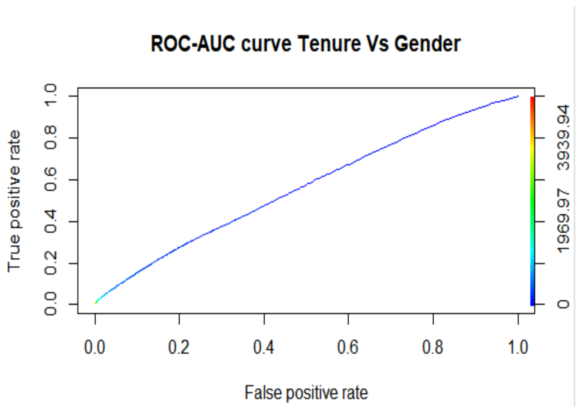


Fig:6(h)

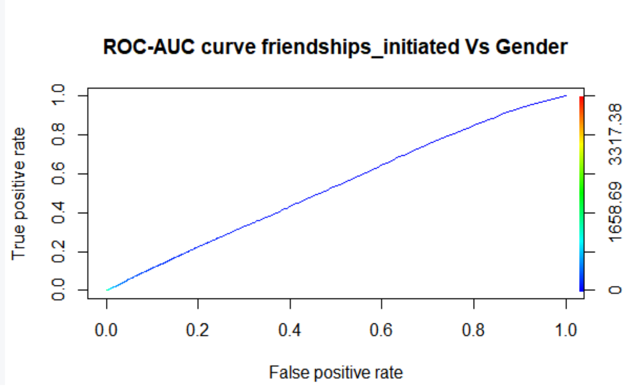


Fig:6(i)

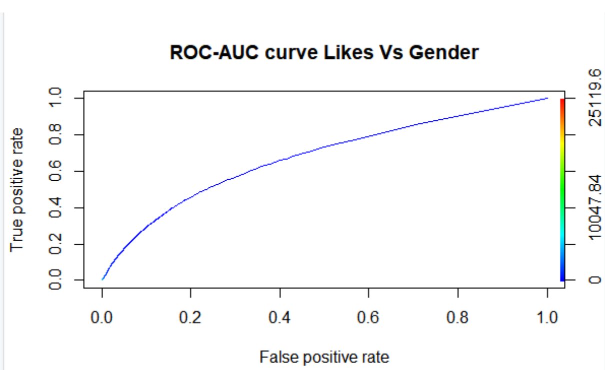


Fig:6(j)

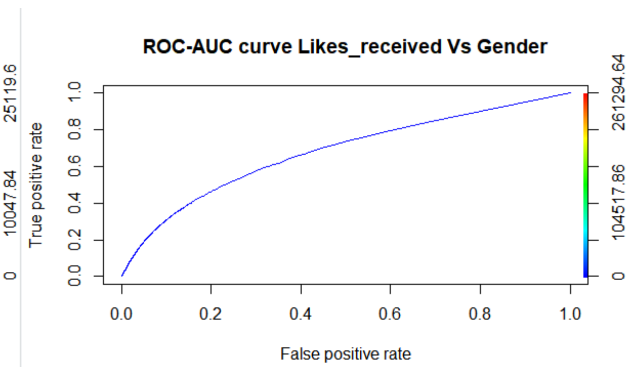
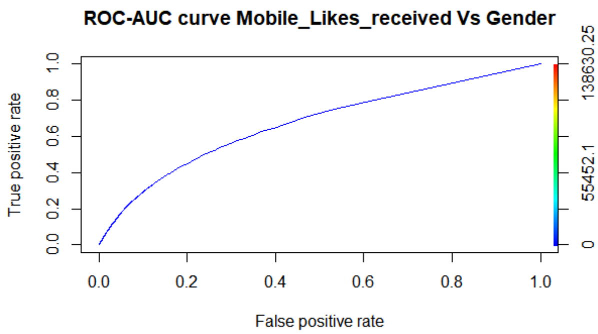


Fig:6(k) Fig:6(l)



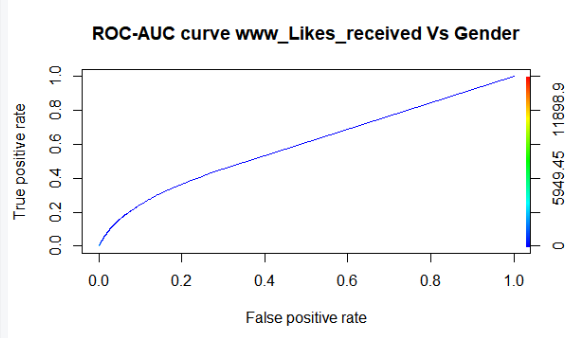


Fig:6(m) Fig:6(n)

The AUC is plotted in fig 6-a and 6-b. It's the plot for the True positive against the false positive and it is plotted for (0,0) and (1,1). Fig 6-a is for the training data where Ns_auc is computed as 0.500 where as Lr_auc is computed as 0.642. Fig 6-b is for the testing dataset where Ns_auc is computed as 0.512 where as the Lr_auc is 0.641. Other plots for AUC is plotted in fig (c) to (n).

f) Precision Versus Recall :

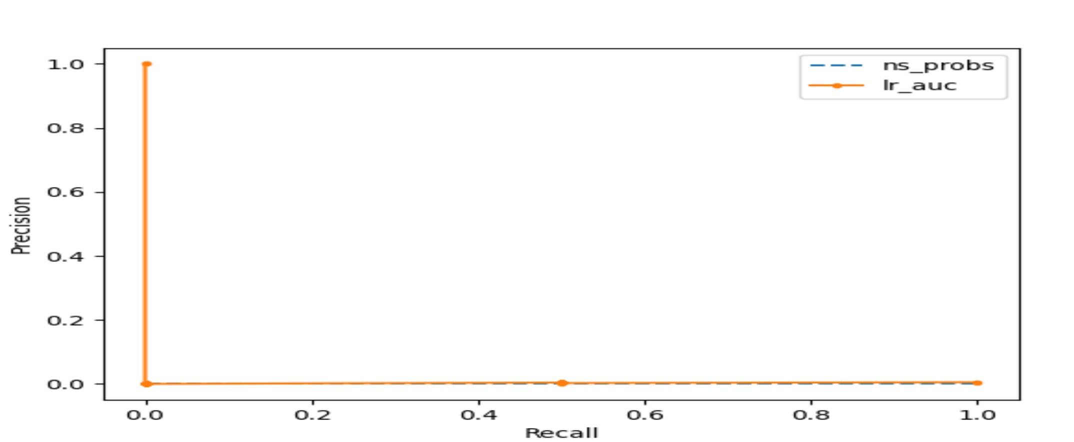


Fig : 7 Precision versus Recall

The plot of the precision versus Recall for the same set of data has been shown in Fig:7 obtained with the Logistic Regression (solver='lbfgs') and the computed score is f1=0.000, AUC =0.003.

ROC Curve :

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 07 ISSUE: 08 | AUGUST - 2023 SJIF RATING: 8.176 ISSN: 2582-3930

The covariance matrix (8- a,b,c,d,e,f) displays the concentration for all the columns available in the experimental dataset.

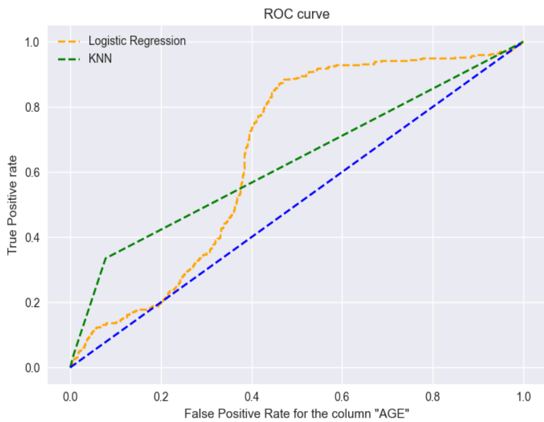


Fig :8(a)

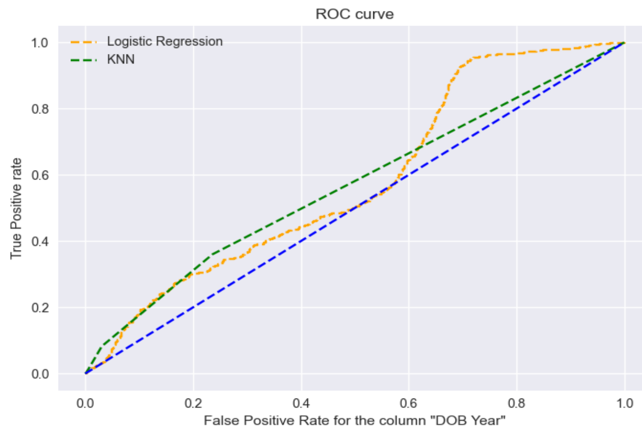


Fig :8 (b)

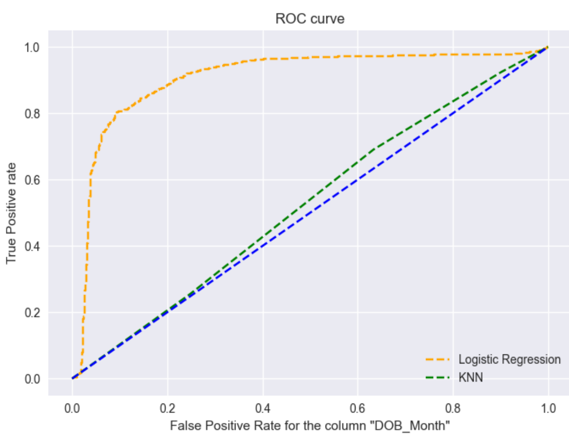


Fig :8 (c)

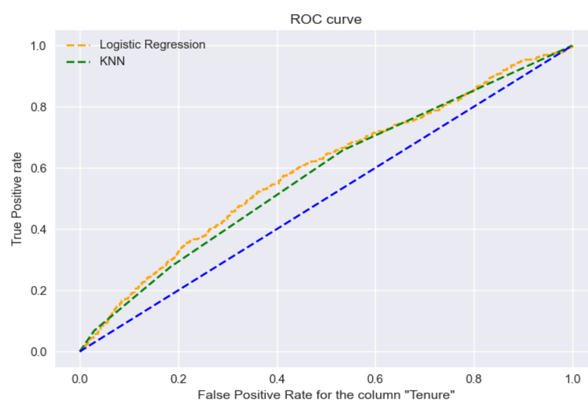


Fig :8 (d)

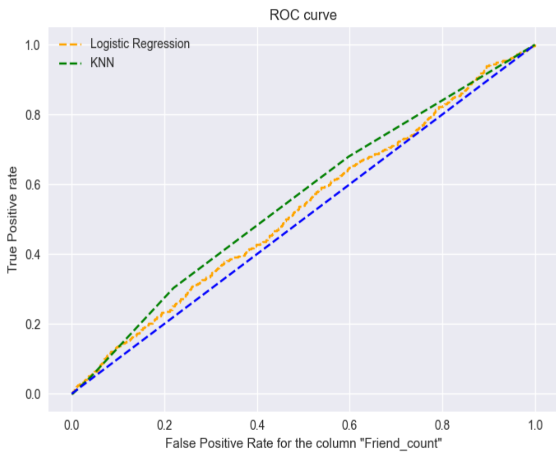


Fig :8(e)

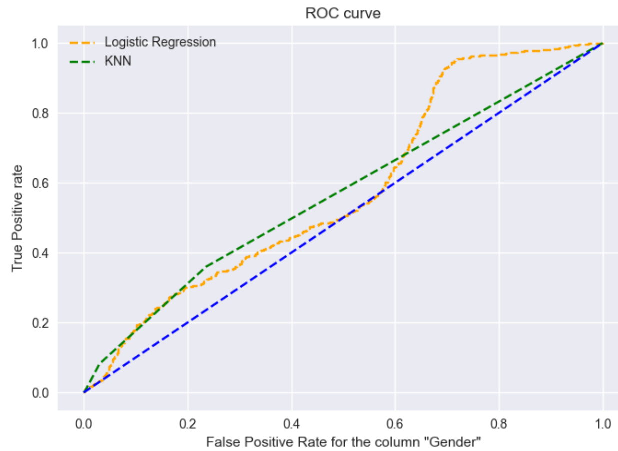


Fig :8(f)

Figure : 8 AUC curve – True versus False

Computed AUC score is 0.7904881915251022 0.6100126521369769 respectively for AUC1 & AUC2 for column AGE [Fig:8-(a)]. Computed AUC score is 0.6515281340156545 & 0. 5854310205588528 respectively for AUC1 & AUC2 for column DOB Year [Fig :8-(b)] . Computed AUC score is 0.6052332894035876 & 0.512090096178257 respectively for AUC1 & AUC2 for column DOB_Month [Fig:8-(c)]. The computed AUC score is 0.6582577882647438 & 0.5883081268144618 respectively for AUC1 & AUC2 for column Tenure [Fig:8-(d)]. The computed AUC score is 0.607350595258181 & 0.5798172319433057 respectively for AUC1 & AUC2 for friend count [Fig:8-(e)]. The computed AUC score is 0.6515281340156545 & 0. 5798172319433057 respectively for AUC1 & AUC2 for gender [Fig:8-(f)].

Degree centrality overall for the Dataset :

Degree centrality shows how many connections a node or person has in the overall network. A node can be very tightly coupled at the center of the network but not so with other nodes on the edge. This is possible across the network.

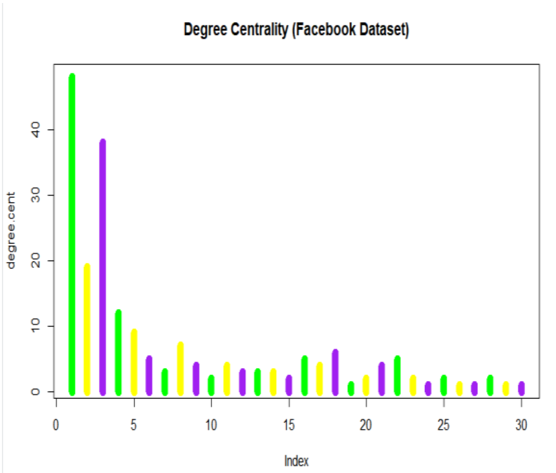


Fig : 9(a)

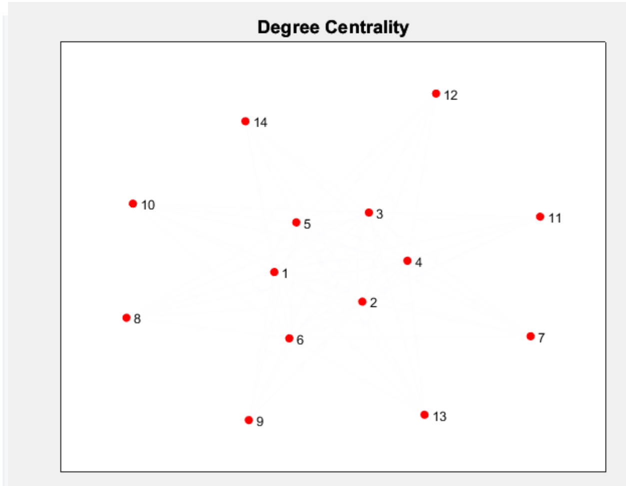


Fig: 9(b)

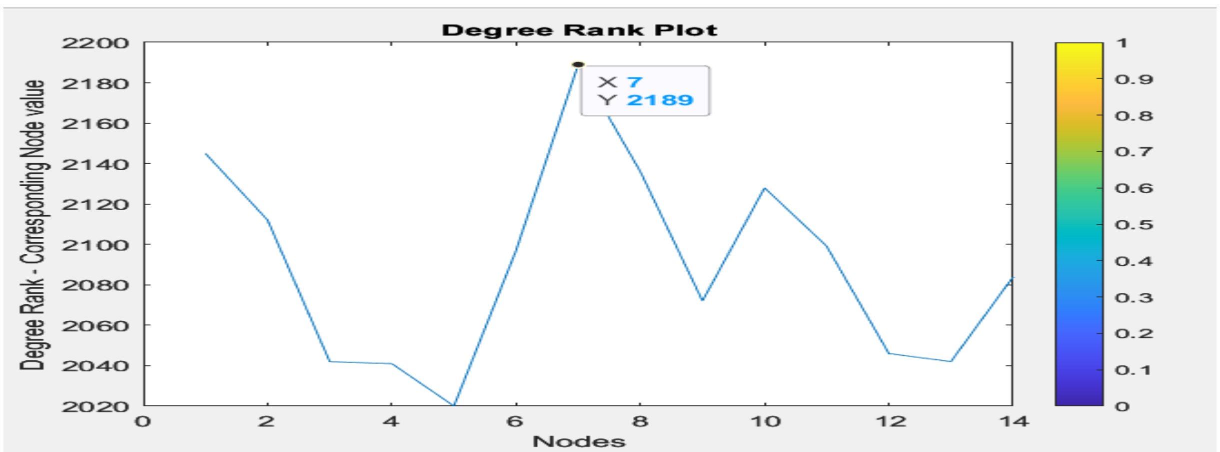


Fig : 9(c)

The fig 9(a), (b) & (c) shows the degree centrality for the Facebook dataset. As the purpose of degree centrality lies to identify the nodes with highest connected network ie most influential node in the network, it is seen that the node 7 has the highest degree rank of 2189. Other nodes in the network have lesser degree rank value than node 7.

H-Index for the Facebook dataset

The h index is a metric for evaluating the cumulative impact of one factor on the overall performance of the whole dataset and measures quantity with quality by comparing sex with other parameters in the dataset. The h index corrects for the disproportionate weight of the high impact of one dataset on overall parameters in the available dataset, the Facebook dataset in the current case.

```
>> hIndex

hIndex =

  Columns 1 through 8

        13          12        1996           3           1         200           0           0

  Columns 9 through 14

         1           5           1           2           0           1
```

Fig : 10(a) Computed H-Index

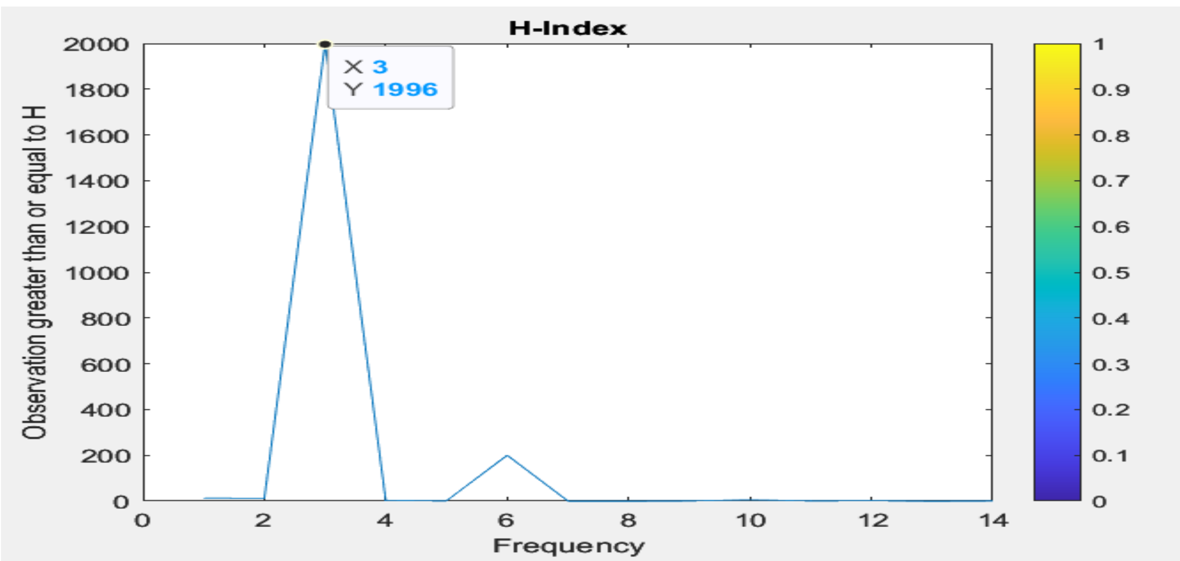


Fig : 10(b) Computed H-Index

From the Fig 10(a) and 10(b), it is very clear that the 3rd column has highest H Index having value as 1996. The third column dob_year has the highest H Index value.

Co-Relation coefficient (CC):

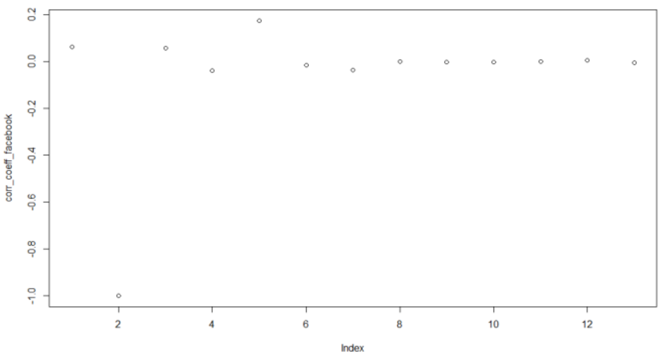


Fig :11(a)

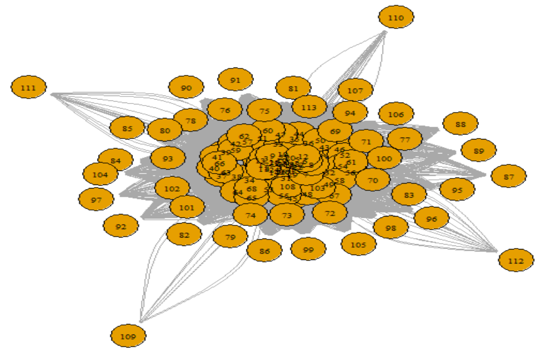


Fig :11(b)

Figure : 13 Correlation coefficient Index versus Coefficient Value

The computed co-relation coefficient which is a plot for Coeff index vs Coeff Value for the given data set is as below [Fig :11(c)].

```
[1]  0.0624920108 -1.0000000000  0.0587919465 -0.0397434580  0.1740534776
 [6] -0.0144935156 -0.0349570412 -0.0003899693 -0.0030127869 -0.0020725352
[11] -0.0008515354  0.0052628614 -0.0048399600
```

Figure : 11 (c)

Where as the centralization for top node is 3.091521 whereas the theoretical_max computed is 19602.

Centrality Plot :

Each node in the network is either important of nonimportant. The nodes which are important in the network are known as the central node. The importance of any node in the network is identified by the centrality plot with the comparison of other nodes which are in the network. The importance of a central node is higher in the context of a noncentral node in the network.

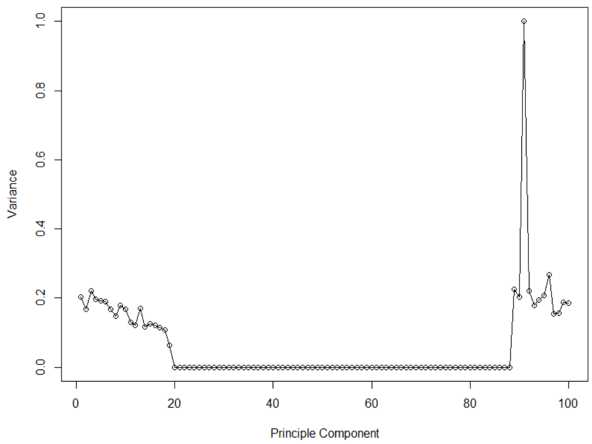


Fig :12 (a)

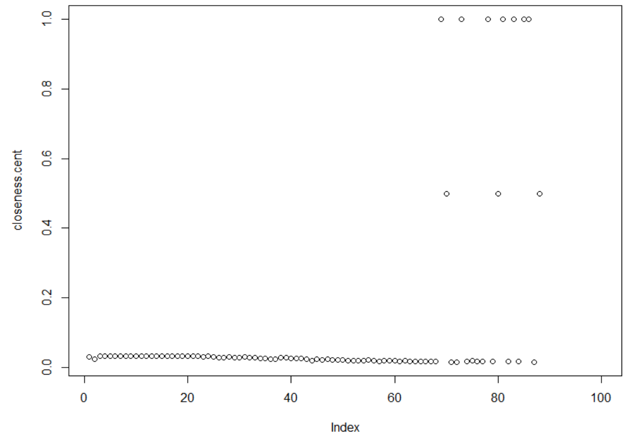


Fig:12 (b)

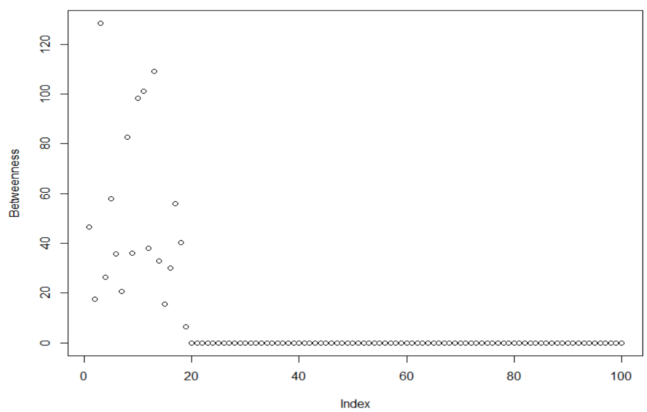


Fig :12 (c)

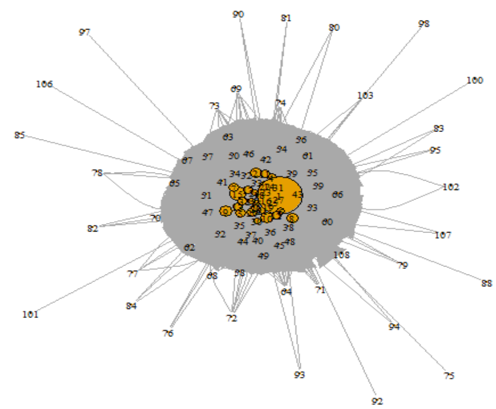


Fig: 12(d)

Fig : 12 Centrality Plot

The experiment was performed on different centrality methodology EC [Fig :12(a)], CC [Fig : 12(b)], Betweenness centrality [Fig :12(c)] and PR[Fig :12(d)] and the plot was done. EC was plotted between Variance and Principle component and the inclination is seen in variance at the end of principal component. CC is plotted on closeness centrality & index and majority of occurrence lies at the lower segment of closeness centrality. Among the defined method H-Index, DC, and EC for TARank, it is observed that DC is better than the other two defined methodologies (H-Index & EC) regarding the time complexity

The inflection probability β is also validated to check how effective is TARank and this is validated with different β values on the designed network. The inflection probability is validated for all eight centrality procedures [17]. For taking the observation eight centrality measures have been taken with the change of inflection probability and see that TARank always performs better in comparison to other measurements. Plot [12] was done to validate all eight significance tests. The TARank variant was validated against the centrality measure and this comparison depicted the p-value in most of the cases as below the significance level which is 0.05. This validation was done for the TARank correlation coefficient and centrality measures.The overall outcome shows that the TARank perform far better than other centrality algorithm evaluation mechanism.

Additionally, validation testing was done for TARank for other spreading models. A similar result was seen for this validation test as well. The performance of TARank still stands better than other mechanisms available globally. The IVI value (Integrated value of influence) for other established mechanisms still remains on the lower side than TARank [5].

6. Experimental Results

The experimental results and plots can be summarized as

a) fig 1 plots the network diagram, whereas 2(a) and 2(b) plots the tree view and data type available in the Facebook dataset.

b) Residual error and EIG in section 3(a) and 3(b) clearly shows the concentration of data at the root. Both the x and y coordinates for the available Facebook dataset lie on the (0,0) axis.

c) H Index for the 3rd node in the Fig 10(a) and 10(b) is highest with the value as 1996.

d) Degree centrality for node 7 is highest in Fig9(a),9(b), and 9(c) having the value as 2189 ie node 7 has got the highest concentration in the overall network plot.

e) Closeness centrality Fig:12(b) and Fig:12(c) and page rank centrality Fig:12(a) show that node 3 has got the highest value in the network.

f) Above said observation in b,c,d,e is confirmed in PCA calculation in Fig 6 (c) to (n). The overall concentration value for PC1 & PC2 is 47.8%.

The experiment was performed on different centrality methodologies EC, CC [Fig: 12(b)], Betweenness centrality [Fig:12(c)], and PR[Fig:12(d)] along with corresponding plots for each of the centrality methodologies. EC was plotted between Variance and Principle components and the inclination is seen in variance at the end of the principal component. CC is plotted on closeness cert & index and the majority of occurrence lies at the lower segment of closeness cert.

Among the defined method H-Index, DC, and EC for TARank, it is observed that DC is better than the other two defined methodologies regarding the time complexity. Hence, from the available information and based on the different experimental results it is clear that out of all available methodologies TARank(DC) is a better approach in comparison to other approaches for the influential node identification for online social media networks.

The inflection probability β is also validated to check how effective is TARank and this is validated with different β values on the designed network. The inflection probability is validated for all eight centrality procedures [18].

A similar result was seen for this validation test as well. The performance of TARank still stands better than other mechanisms available globally. The IVI value (Integrated value of influence) for other established mechanisms still remains on the lower side where weighted networks were used for the identification of influential nodes [19].

Future Aspect of Activity

Further, the activity can be extended to overcome the better performance and accuracy of the outcome for the graph model for identification of the most influential node for the online social network data. Also, further comparison with other established methodologies for the identification of influential networks can be done. Additionally, a similar experiment can be done on a higher volume of data.

6. The TARank framework

For a given network with a set of N nodes

S = {a_1, a_2, a_3, …. .. .. an} and a given set of M edges E = {b1,b2..bM}, primary centrality scores CS = {c1,c2..cN}. For CS, each CS ={c1,c2..cN}, a set of first centrality scores have been produced. For each node's centrality score, BFS tree design has been done [10]. In order to create the BFS Tree T, each node is considered as the root node and the traversal is done in BFS order.

Then, the cum_centrality score of the BFS Tree at the *kth* level is generated as :

Cumulative_score(k) = $\sum_{q=1}^{k} \sum_{V j \in T(q)} CSj,$ (1)

A line chart is drawn for each node having the level number of the BFS tree as the x-axis and cum_centrality score as the y-axis for the dataset provided. Finally, the final centrality score for each node is obtained by calculating the AUC value of the equivalent line chart when the number of levels is no larger than *k is as follows*

AUC(k) = cum_score(1)/2+(cum_score(1)+cum_score(2))/2+….+ (cum_score(k - 1) +cum_score(k))/2 (2)

$= \sum_{q=1} k((k - q + \frac{1}{2}) \sum_{V j \in T(q)} CSj)$ (3)

As per the existing formula, the AUC score is essentially a weighted linear combination of initial node scores in which top nodes are associated with large weight coefficients. Overall, it is concluded that the TARank is overall the best approach for the identification of the most influential node for the online social media network.

8. Conclusion

In the present work, by employing the TARank model, the most influential node out of all influential nodes has been identified. The residual error of the given dataset has been plotted which is the difference between the observed value and the predicted values of data.

This plot was done to certify the quality of the model designed. Additionally, AUC for each node has been calculated and plotted to establish the occupancy of the node in the overall plot. Different graphs including eigenvector, eigenvalue, and degree centrality have been plotted for validation of the results. Closeness Centrality, Between-ness Centrality, and Page Rank Centrality were plotted to see how close a node is to all other nodes in the network. Plotted BFS Tree shows the data flow, the root node, and the leaf node in the tree.

In the future as part of the improvement in the current methodology, more impressive commendation methods can be anticipated to be based on the general framework as discussed earlier. Additionally, online streaming of data implementation is possible using Apache Hadoop and Spark on Facebook or other online social media network like Twitter. Thus it is concluded that the framework defined and stated here may be considered a basic graph theoretic mechanism for identifying the most influential nodes in the area of social network science. Real-time implementation in different fields like medical science, banking, and financial domain and leader identification during the election is possible as per requirement.

9. References

[1]. Frainay, C., and Jourdan, F. (2016). Computational methods to identify metabolic sub-networks based on metabolomic profiles. Brief. Bioinform. 18, 43–56.
[2]. Lordan, O., Sallan, J.M., and Simo, P. (2014). Study of the topology and robustness of airline route networks from the complex network approach: a survey and research agenda. J. Transp. Geogr. 37, 112–120.
[3]. Dorogovtsev, S.N., Goltsev, A.V., and Mendes, J.F.F. (2008). Critical phenomena in complex networks. Rev. Mod. Phys. 80, 1275.
[4]. Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., and Makse, H.A. (2010). Identification of influential spreaders in complex networks. Nat. Phys. 6, 888–893.
[5]. Motter, A.E., and Lai, Y.C. (2002). Cascade-based attacks on complex networks. Phys. Rev. E 66, 065102.
[6]. Freeman, L.C. (1978). Centrality in social networks conceptual clarification. Soc. Netw. 1, 215–239.
[7]. Sabidussi, G. (1966). The centrality index of a graph. Psychometrika 31, 581–603.
[8]. Goh, K.I., Kahng, B., and Kim, D. (2001). Universal behavior of load distribution in scale-free networks. Phys. Rev. Lett. 87, 278701.
[9]. Bavelas, A. (1948). A mathematical model for group structures. Hum. Organ. 7, 16–30.
[10]. Stephenson, K., and Zelen, M. (1989). Rethinking centrality: methods and examples. Soc. Netw. 11, 1–37.
[11]. Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems 30, 107–117.
[12]. Lu€, L., Chen, D., Ren, X., Zhang, Q., Zhang, Y., and Zhou, T. (2016). Vital nodes identification in complex networks. Phys. Rep. 650, 1–63.
[13]. Das, K., Samanta, S., and Pal, M. (2018). Study on centrality measures in social networks: a survey. Soc. Netw. Anal. Min. 8, 13.125.
[14]. Liu, X., Hong, Z., Liu, J., Lin, Y., Rodrı´guez-Pato´ n, A., Zou, Q., and Zeng, X. (2020). Computational methods for identifying the critical nodes in biological networks. Brief. Bioinform. 21, 486–497.
[15]. Salavaty, A., Ramialison, M., and Currie, P.D. (2020). The integrated value of influence: an integrative method for the identification of the most influential nodes within networks. Patterns 1, 100052.
[16]. Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. J. ACM 46, 604–632.
[17]. Kendall, M.G. (1938). A new measure of rank correlation. Biometrika 30, 81–93.

[18]. Teng, B., Zhao, C., Liu, X., and He, Z. (2015). Network inference from APMS data: computational challenges and solutions. Brief. Bioinform. 16, 658–674.
[19]. Lu€, L., Zhou, T., Zhang, Q., and Stanley, E.H. (2016). The h-index of a network node and its relation to degree and coreness. Nat. Commun.7, 10168.