VOL. 14, NO. 5, MARCH 2019 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

KEY NODE SELECTION NETWORK ANALYSIS AND CENTRALITY MEASUREMENTS ON A DATASET OF CANCER DOCUMENTS

V MNSSVKR Gupta and CH.V. Phani Krishna
Department of Computer Science and Engineering, KLEF, Vaddeswaram, Guntur, Andhra Pradesh, India
E-Mail: guptavkrao@gmail.com

ABSTRACT

Network analysis deals with interdisciplinary study of analyzing relationships. Networks possess inherent topological characteristics and integrate multiple sources of high throughput data. A dataset which contains 1000 article titles on cancer disease were considered to evaluate the importance of centrality measurements of nearly 47 cancer genes, and their associated data is also considered to select the most central group of nodes from a network. Centrality analysis revealed that the variables "*cancer*" and "*patient*" were reported to have high values than others which suggest the fact that these two parameters are highly influential in communicating with other nodes. Further, key node selection analysis comprising eight centrality measurements resulted in "*cancer*" as the most central group of nodes from a network.

Keywords: network analysis, cancer, network centrality, key node.

1. INTRODUCTION

Network analysis has witnessed prominent role in the field of computational sciences as it deals with social, interactive data in the context of analyzing relationships among objects in data. Network analysis is popular in various domains like social science, mathematics, computer science and bioinformatics [1] [2] [3] Network modeling is the interdisciplinary study of relationships. The *nodes* or *members* of the network can be groups or organizations. Network structure can be studied at many different levels the dyad, triad, subgroup, or even the entire network [4]. The distance between two nodes in a network can be measured by determining the minimum number of steps between them [5]. Networks possess inherent topological characteristics that impart emergent properties of biological relevance, such as functional robustness [5]. Network analysis integrates multiple sources of high throughput data and link data sets with subsequent modeling efforts, thereby enabling continuous refinement of systems analysis [6]. Network biology is one of the rapidly developing area of research controlled by a complex system-level network of molecular interactions [7]. Network analysis equates the assembly of pairwise connections (edges) between discrete objects (nodes) coalesces to form a network, or graph [8]. Graphs that contain many cohesive subsets as well as short paths, on average, are often termed *small world* networks. Characterizations of the centrality of each data point in the network are based on the degree (*degree* centrality), on the lengths of paths from one data point to all other variables (*closeness* centrality), or on the extent to which the shortest paths between other data points pass through the given data point (*betweenness* centrality). Measures of network *centralization* signify the extent of heterogeneity among data in these different forms of centrality [9]. Also *Centralization* measurements reveal whether a network has a star-like topology or the nodes of the network have on average the same connectivity. *Graph partitioning* or community structure detection algorithms elucidate within- and between-community edges. The present paper describes the centrality measurements of nearly 47 cancer genes and their associated data to select the most central group of nodes from a network.

2. MATERIALS AND METHODS

Networks are a natural way to represent information. Nodes in such networks organize into densely linked groups that are commonly referred as network communities. Nodes in such networks organize into densely linked clusters. To extract communities from a given undirected network, it is advisable to choose a scoring function (e.g., modularity) that quantifies the intuition that communities correspond to densely linked sets of nodes [10]. The degree of a node in an undirected graph is the number of connections or edges the node has to have with other nodes. A dataset of 1000 entries with the term 'cancer' in pubmed literature database was selected to perform network analysis, to analyse the important terms and their distribution in the network. A term document matrix (TDM) was created from the corpus where rows correspond to documents and columns correspond to terms. A sparse adjacency matrix was constructed and nearly 47 terms appeared and the centrality measurements were carried out on these terms using keyplayer package [11].

2.1 Network analysis

Centralization measurements were carried out to assess whether a network has a star-like topology and also to understand whether the nodes of the network have on an average the same connectivity. Network centralities such as Degree centrality [12], Closeness centrality [13] [14], Eigenvector centrality and Betweenness centrality were evaluated.

2.2 Network centralities

In general, central nodes or intermediate nodes affect the topology of the network. Some points are not central to the data, however, they might have crucial role, hence it is important to detect such nodes using Degree centrality, Closeness centrality and Eigenvector Centrality. Centralization measures whether a network has a star-like

topology or not and also measures on an average does the network has the same connectivity or not. If the centralization value is nearer to 1, then the network represents a star-like topology. If the centralization value nearer to 0, it is more likely that the nodes of the network have same connectivity.

Degree Centrality shows that an important node is such a node which is involved in a large number of interactions. For a node *i*, the degree centrality is calculated as:

$C_d(i) = deg(i).$

Nodes with very high degree centrality are called *hubs* since they are connected to many neighbours. Degree centrality is the simplest network centrality measure. It only takes into account the degree of a node, which is equal to the number of nodes that a given node is connected to [15].

Closeness Centrality indicates important nodes that can communicate quickly with other nodes of the network. Let $G = (V, E)$ be an undirected graph. Then, the centrality is defined as :

$$C_{clo}(i) = \frac{1}{\sum_{t \in V}^{|V|} dist(i, j)}$$

Where *dist(i, j)* denotes the distance or the shortest path *p* between the nodes *i* and *j*.

Eigenvector Centrality ranks high values to the nodes that are connected to important neighbours [16].
Betweenness centrality displays high ranks for the nodes which are intermediate between neighbours [17]. These nodes are crucial in communicating between two neighbours. Hence, *betweenness centrality* shows significant nodes on a high fraction of paths between other nodes in the network.

Centralization is the extent to which the centrality of the network is concentrated with only a few nodes. If centrality is more evenly distributed then this number will be low. Any centrality measure can have a centralization score applied to it.

Degrees of centrality with indegree, outdegree and total degreeness are provided here. Degree is a way of measuring node activity. There are a few ways of measuring degree.

- Total-degree is the count of all edges incident (connected to) a node
- In-degree is the count of all edges that point *in* to a node
- Out-degree is the count of all edges that point *out* of a node

2.3 M-reach degree centrality

M-reach degree centrality generalizes the degree centrality by delimiting specific neighbourhoods. It describes the number of nodes that node i can reach in M steps as well as the number of nodes that can reach node i in M steps.

3. RESULTS AND DISCUSSIONS

In this section results and its actual operation is discussed.

3.1 Dataset

A dataset having 1000 article titles with cancer disease, published in various journal resources were considered from PubMed database (www.ncbi.nlm.nih.gov/pubmed). The dataset with titles, author names and year were segregated and used as csv file input. Then Term document matrix (TDM) is constructed. TDM contains the frequencies of terms that occur in a collection of n document. For constructing the term document matrix(TDM), a frequency of terms > 99% are considered , in other words less than 1% sparse was employed on TDM. This was resulted in generating a 47 terms as binary word matrix **.**Also it has to assumed that before the construction of binary matrix, documents which does not contain any terms are excluded and the document for which a term is repeated more than once is counted as 1 entry. Here we have chosen Binary method because it is used to find the relative similarity of those terms that have higher probability of occurring together in a column.

All outdegree and indegree values generated by M-reach degree centrality for 47 terms of 1000 articles with cancer disease are given in Table-1.

Table-1 describes the M-reach degree centralities for 47 terms of 1000 articles with cancer disease. By using Indegree data, different graphs are plotted which are shown in Figure-1, Figure-3, Figure-4. similarly for outdegree data different graphs are drawn which are shown in Figure-2, Figure-5, and Figure-6. We also drawn a plot for Variable and Degree which is shown in Figure-7

Table-1. Outdegree and indegree values of variables in the dataset generated by M-reach degree centrality.

S. No.	Variable	outdegree	indegree	total	S. No.	Variable	outdegree	indegree	total
1	activ	31	31	62	25	inhibitor	26	26	52
2	advanc	27	27	54	26	invas	32	32	64
3	analysi	28	28	56	27	metastasi	27	27	54
4	associ	30	30	60	28	model	22	22	44
5	bladder	27	27	54	29	novel	24	24	48
6	breast	39	39	78	30	outcom	27	27	54
7	cancer	46	46	92	31	pathway	29	29	58
8	carcinoma	35	35	70	32	patient	41	41	82
9	chemotherapi	25	25	50	33	potenti	31	31	62
10	circul	13	13	26	34	predict	23	23	46
11	clinic	29	29	58	35	promot	26	26	52
12	colorect	33	33	66	36	prostat	36	36	72
13	detect	21	21	42	37	protein	28	28	56
14	develop	23	23	46	38	regul	27	27	54
15	diseas	30	30	60	39	review	21	21	42
16	effect	27	27	54	40	signal	30	30	60
17	evalu	22	22	44	41	surviv	28	28	56
18	express	30	30	60	42	target	34	34	68
19	factor	24	24	48	43	therapeut	21	21	42
20	function	28	28	56	44	therapi	27	27	54
21	growth	28	28	56	45	treatment	30	30	60
22	health	13	13	26	46	trial	19	19	38
23	human	31	31	62	47	tumor	38	38	76
24	inhibit	29	29	58					

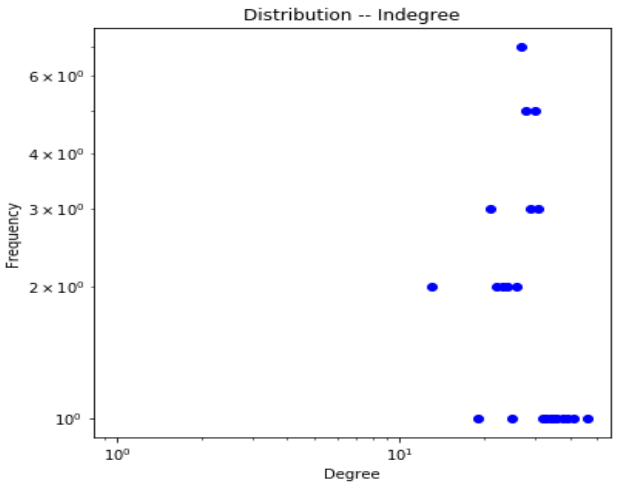

Figure-1. Indegree measure of each node.

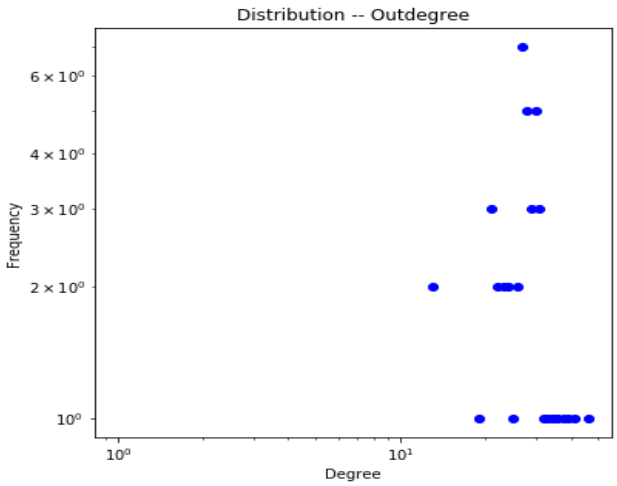

Figure-2. Out degree of node.

Figure-1 shows the indegree measures of each node for 47 terms of 1000 articles with cancer disease using M-reach degree centrality

Figure-2 shows the outdegree measures of each node for 47 terms of 1000 articals with cancer disease using M-reach degree centrality. Cumulative distribution is calculated for indegree values which are depicted in Figure-3. Similarly complementary cumulative distribution is calculated for indegree values which are depicted in Figure-4.

As calculated and depicted for indegree, Cumulative distribution is calculated for outdegree values which is depicted in Figure-5. Similarly complementary cumulative distribution is calculated for outdegree values which are depicted in Figure-6. CDF and PDF comparison for degree vs frequency is shown in Figure-7. M-reach Degree centrality of each node for 47 terms of 1000 articles is shown in Figure-8.

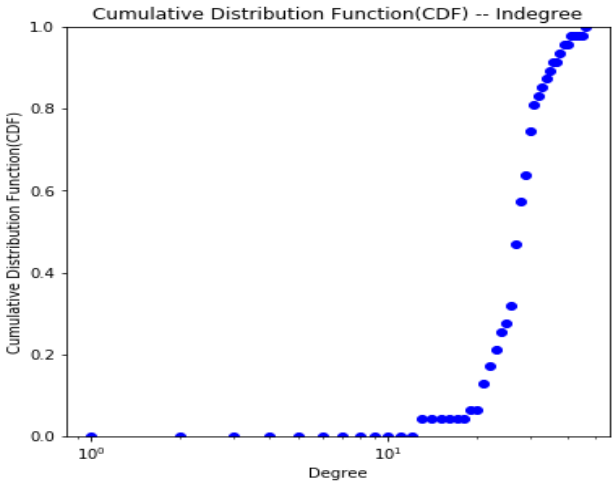

Figure-3. Cumulative distribution function interms of indegree.

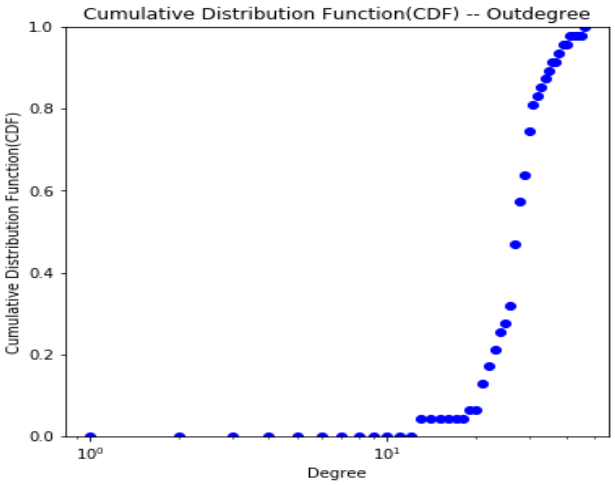

Figure-5. Cumulative distribution function interms of outdegree.

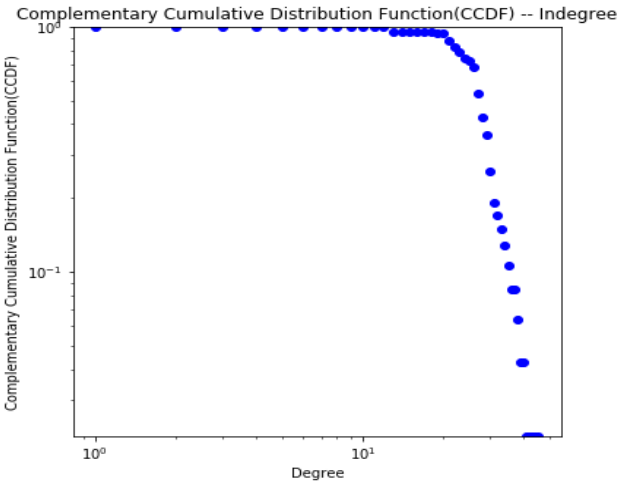

Figure-4. Complementary cumulative distribution function for indegree.

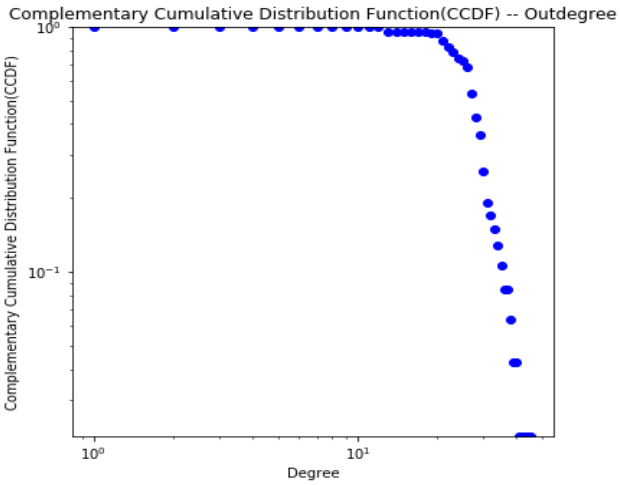

Figure-6. Complementary cumulative distribution function interms of out degree.

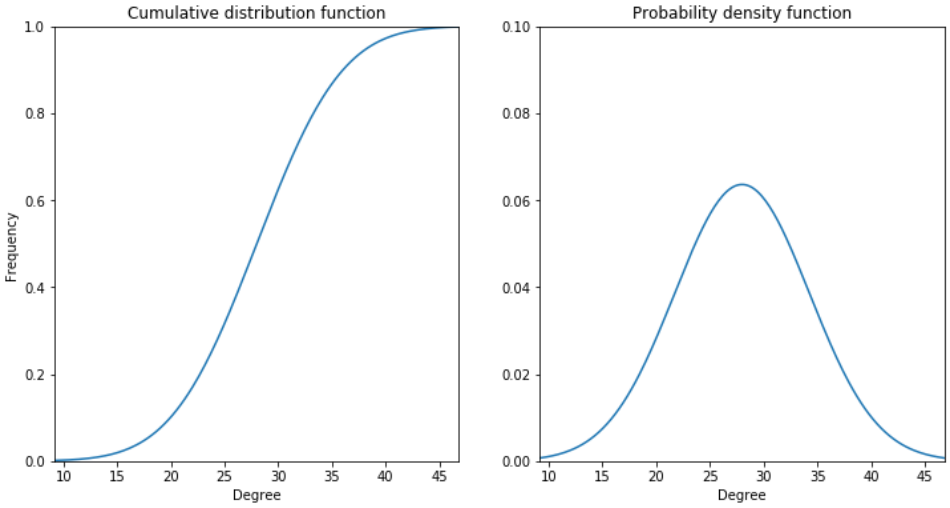

Figure-7. CDF and PDF comparison for degree vs frequency.

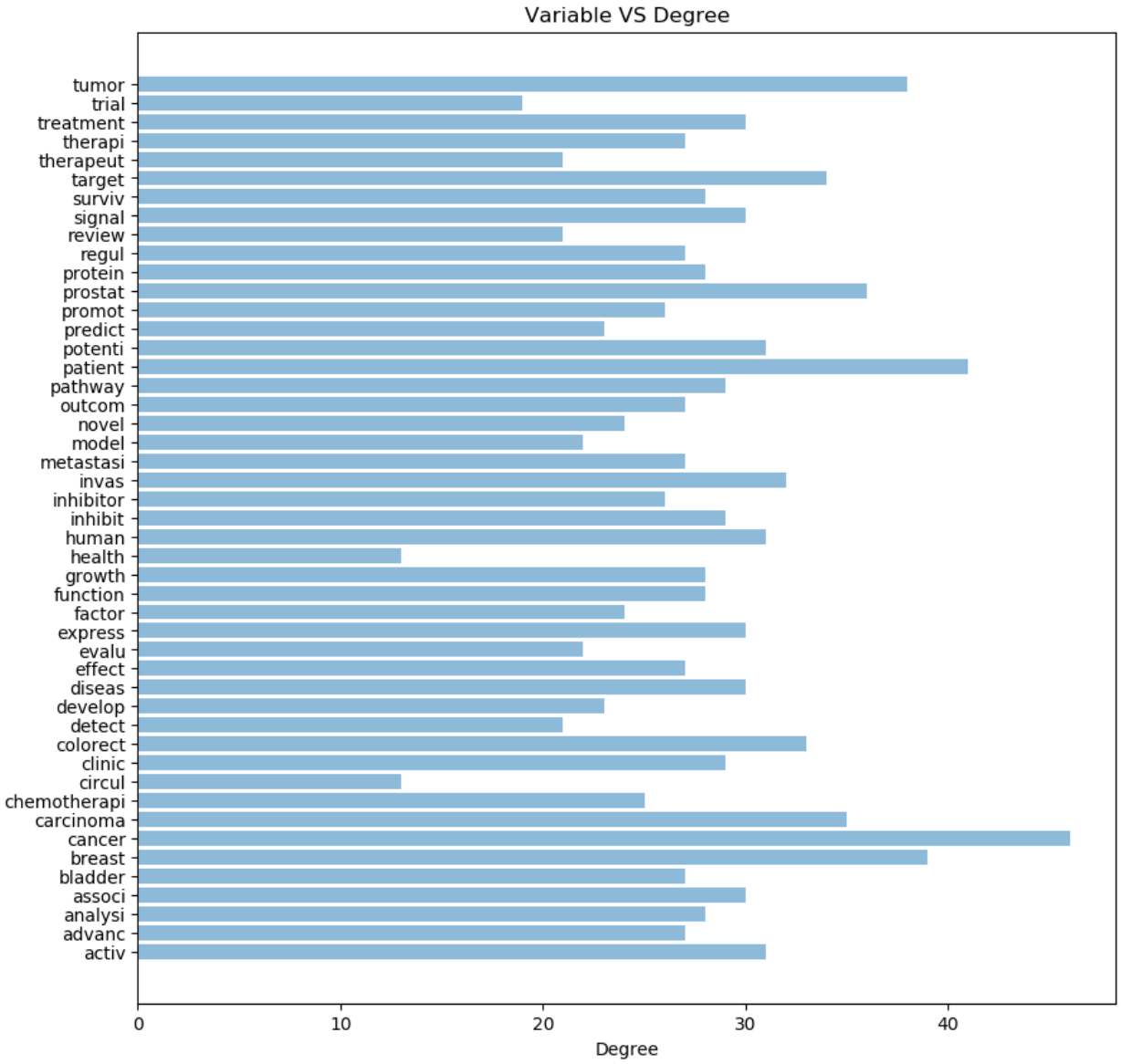

Figure-8. Shows the M-reach Degree centrality of each node for 47 terms of 1000 articles with cancer disease using Table-1.

3.2 M-reach closeness centrality

One way to refine the M-reach degree centrality is to use the inverse of distance to measure the status between nodes (Table-2).

Table-2. outdegree and in-degree values of variables in the dataset generated by M-reach closeness centrality.

S. No.	Variable	outdegree	indegree	total	S. No.	Variable	outdegree	indegree	total
1	activ	0.113332	0.113332	0.226665	25	inhibitor	0.093189	0.093189	0.186378
2	advanc	0.126711	0.126711	0.253422	26	invas	0.11967	0.11967	0.23934
3	analysi	0.118495	0.118495	0.236989	27	metastasi	0.135261	0.135261	0.270522
4	associ	0.138799	0.138799	0.277598	28	model	0.10035	0.10035	0.200701
5	bladder	0.126812	0.126812	0.253624	29	novel	0.09353	0.09353	0.18706
6	breast	0.183691	0.183691	0.367382	30	outcom	0.123858	0.123858	0.247716
7	cancer	0.227744	0.227744	0.455488	31	pathway	0.144796	0.144796	0.289592
8	carcinoma	0.11019	0.11019	0.22038	32	patient	0.193509	0.193509	0.387019
9	chemotherapi	0.134818	0.134818	0.269636	33	potenti	0.115848	0.115848	0.231697
10	circul	0.131508	0.131508	0.263017	34	predict	0.12349	0.12349	0.24698
11	clinic	0.121382	0.121382	0.242763	35	promot	0.130088	0.130088	0.260176
12	colorect	0.150239	0.150239	0.300479	36	prostat	0.155349	0.155349	0.310697
13	detect	0.115702	0.115702	0.231403	37	protein	0.099268	0.099268	0.198536
14	develop	0.099023	0.099023	0.198046	38	regul	0.138994	0.138994	0.277989
15	diseas	0.119762	0.119762	0.239524	39	review	0.112639	0.112639	0.225279
16	effect	0.098804	0.098804	0.197608	40	signal	0.131703	0.131703	0.263406
17	evalu	0.098906	0.098906	0.197813	41	surviv	0.126963	0.126963	0.253927
18	express	0.118714	0.118714	0.237428	42	target	0.154935	0.154935	0.30987
19	factor	0.116536	0.116536	0.233073	43	therapeut	0.127625	0.127625	0.255251
20	function	0.098734	0.098734	0.197469	44	therapi	0.14111	0.14111	0.282221
21	growth	0.142401	0.142401	0.284802	45	treatment	0.126908	0.126908	0.253817
22	health	0.083171	0.083171	0.166341	46	trial	0.117979	0.117979	0.235959
23	human	0.111316	0.111316	0.222632	47	tumor	0.140697	0.140697	0.281394
24	inhibit	0.134638	0.134638	0.269276					

A consensus from table 1 displayed "*cancer*" (7th variable in dataset) as the most central group of nodes from a network followed by "*patient*" (32nd variable). In terms of eigenvector centrality, m-reach degree and m-reach closeness centrality, variable 10 is represented by the term "*circulation*". Hence, one has to choose a particular centrality measure for selecting the most central player depends on the objectives of that work. If the objective is to find the variable that is repeatedly presented in the dataset, then indegree is a suitable measure. However, if the objective is to spread the information most widely, then outdegree or closeness may be a better option. Therefore, for identifying central variable that communicates with other nodes, these centrality measures would help in locating key players.

3.3 Fragmentation centrality

Fragmentation centrality measures the extent to which a network is fragmented after a node is removed from the network [18], data given in Table-3. Fragmentation centrality is the opposite of the cohesion centrality. Table-3 shows the Fragmentation Centrality of 47 variables from 1000 articles with cancer disease.

Table-3. Fragmentation centrality values of variables in the dataset.

S. No.	Variable	Fragment value	S. No.	Variable	Fragment value
1	activ	0.872393	25	inhibitor	0.871493
2	advanc	0.872983	26	invas	0.872671
3	analysi	0.872618	27	metastasi	0.873388
4	associ	0.873633	28	model	0.871811
5	bladder	0.872987	29	novel	0.871508
6	breast	0.875593	30	outcom	0.872856
7	cancer	0.894081	31	pathway	0.874023
8	carcinoma	0.872285	32	patient	0.877248
9	chemotherapi	0.873343	33	potenti	0.872519
10	circul	0.873196	34	predict	0.872845
11	clinic	0.872782	35	promot	0.873511
12	colorect	0.874029	36	prostat	0.874259
13	detect	0.872494	37	protein	0.871763
14	develop	0.871801	38	regul	0.87385
15	diseas	0.872674	39	review	0.872376
16	effect	0.871743	40	signal	0.873226
17	evalu	0.87176	41	surviv	0.873012
18	express	0.872695	42	target	0.875471
19	factor	0.87255	43	therapeut	0.873106
20	function	0.87174	44	therapi	0.873747
21	growth	0.874303	45	treatment	0.872998
22	health	0.871048	46	trial	0.872595
23	human	0.872371	47	tumor	0.874712
24	inhibit	0.873387			

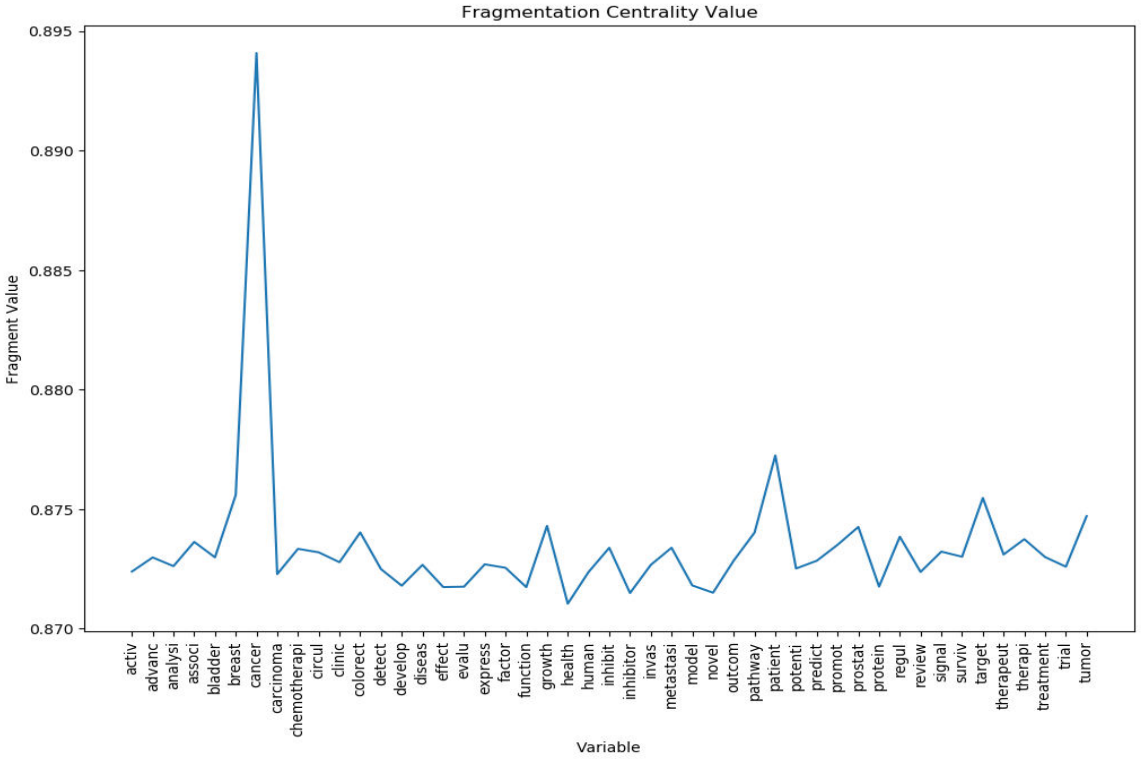

Figure-9. Shows that the fragmentation centrality values for 47 terms of 1000 articles with cancer disease using Table-3 information.

From the Figure-9 of degree centrality, closeness and fragment centrality measures, it was observed that the variable "*cancer*" and "*patient*" were reported to have high values than others which suggest the fact that these two parameters are highly influential in communicating with other nodes.

3.4 Key node selection

The main goal of the work is to select the most central group of nodes from a network based on greedy search algorithm of Borgatti [18]. The algorithm chooses a set of nodes as seeds and then swaps the selected nodes with unselected ones, if the swap increases the group centrality. The greedy algorithm converges fast and to facilitate the search in large networks, parallel computation can be opted. During parallel computation, for each cluster and each iteration the algorithm randomly picks a node from the candidate set and the residual set, respectively, and swaps the two if it improves the centrality score of the candidate set. It repeats this process until exhausting the specified iterations and rounds and then combines the results from the clusters [19]. The analysis resulted in finding top two term variables that are ought to play key role in network dimensionality based community structure detection were given in Table-4 which showed that the most central group is similar for all centrality measures employed.

Table-4. Centrality measures and identified key players.

Method	Keyplayer1	Keypalyer2	Centrality
degree centrality	7	32	45
closeness centrality	7	32	1
betweenness centrality	7	47	260
eigenvector centrality	7	10	0.33
fragment centrality	7	32	0.211
mreach.degree	7	32	45
mreach.closeness	7	10	0.255
parallel computation	7	10	4269

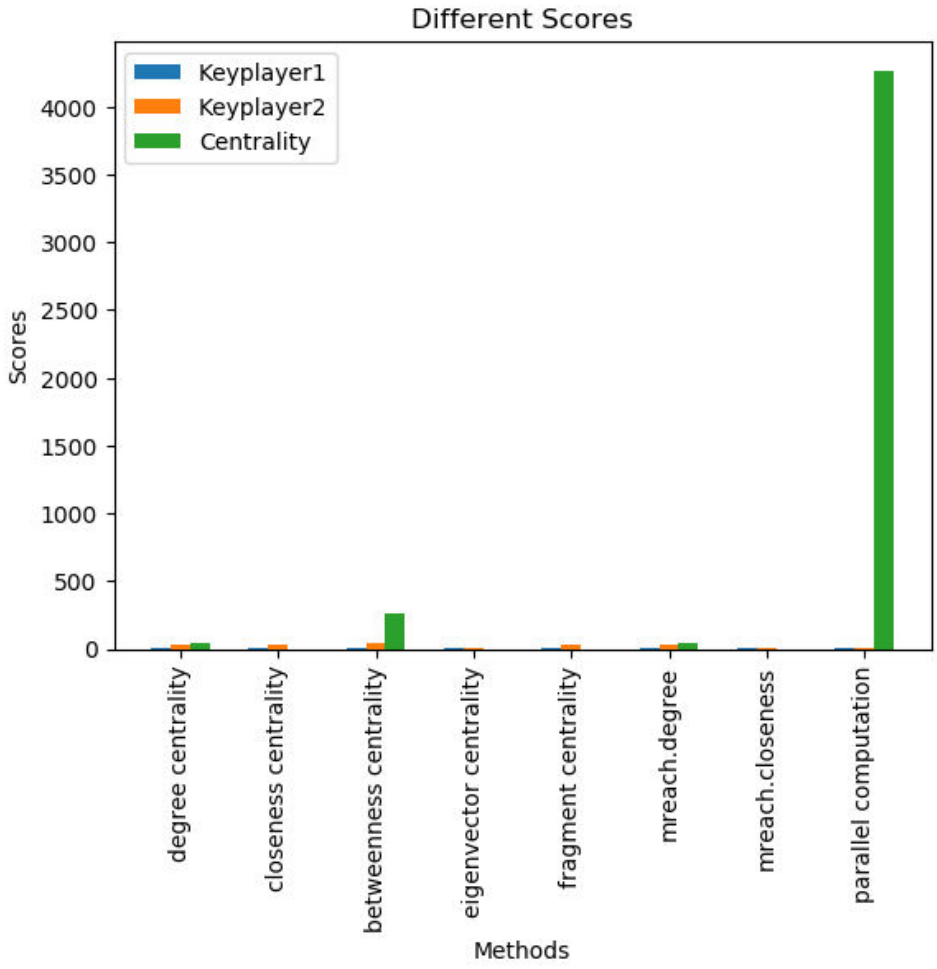

Figure-10. Centrality values of different centrality methods.

Figure-10. Shows the scores calculated by using different centrality methods for 47 variables of 1000 articles with cancer disease.

4. CONCLUSIONS

Network analysis on a dataset of 1000 entries with the term 'cancer' in PubMed literature database was chosen to study the most central group of nodes from a network based on greedy search algorithm. From the degree centrality, closeness and fragment centrality measures, the variable "*cancer*" and "*patient*" were reported as highly influential in communicating with other nodes. A consensus of call centrality measures also resulted in "*cancer*" as the most central group of nodes from a network followed by "*patient*". Finally, locating key players in a dataset, centrality measures are the better option of choice as central nodes or intermediate nodes affect the topology of the network.

REFERENCES

[1] Scott John. 2000. Social Network Analysis. 2nd eition. London: Sage

[2] Carrington P. J., Scott J. & Wasserman S. (Eds.). 2003. Models and methods in social network analysis. New York: Cambridge University Press.

[3] V MNSSVKR GUPTA, CVP Krishna. 2018. K-core analysis and modeling for network centralities. International Journal of Engineering and Technology (UAE) 7(2.7): 168-171.

[4] Carrington PJ, Scott J, Wasserman S. 2005. editors. Models and methods in social network analysis. Cambridge university press.

[5] Barabasi AL, Oltvai ZN. 2004. Network biology: Understanding the cell's functional organization. Nature reviews genetics. 5(2): 101-13.

[6] Sauer U, Heinemann M, Zamboni N. 2007. Getting closer to the whole picture. Science (Washington). 316(5824): 550-1.

[7] Gardy JL, Lynn DJ, Brinkman FS, Hancock RE. 2009. Enabling a systems biology approach to

immunology: focus on innate immunity. Trends in immunology. 30(6): 249-62.

[8] Barabási AL, Albert R. 1999. Emergence of scaling in random networks. Science. 286(5439): 509-12.

[9] Harary F., Norman D. & Cartwright D. 1965. Structural models for directed graphs. New York: Free Press.

[10] Pavlopoulos Georgios A, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, Schneider R, Bagos PG. 2011. Using graph theory to analyze biological networks. BioData mining. 4(1).

[11] An W, Liu YH. 2016. keyplayer: An R Package for Locating Key Players in Social Networks. R Journal. 8(1).

[12] Levy SF, Siegal ML. 2008. Network hubs buffer environmental variation in Saccharomyces cerevisiae. PLoS biology. 6(11): e264.

[13] Ma HW, Zeng AP. 2003. The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics. 19(11): 1423-30.

[14] V MNSSVKR Gupta, CVP Krishna. 2018. A Dataset of 1000 Cancer TitlesUsing Network Analysis and Community. Journal of Advanced Research in Dynamical and Control Systems. pp. 707-712.

[15] Freeman LC. 1978/79. Centrality in social networks: conceptual clarification, Social Networks. 1: 215-239.

[16] Newman ME. 2003. The structure and function of complex networks. SIAM Rev. 45(2): 167-256.

[17] Freeman LC. 1977. A set of measures of centrality based on betweenness. Sociometry. 40: 35-41.

[18] Borgatti SP. 2006. Identifying sets of key players in a Social network. Computational & Mathematical Organizational Theory. 12(1): 21-34.

[19] An Weihua. 2015. Multilevel Meta Network Analysis with Application to Studying Network Dynamics of Network Interventions. Social Networks. 43: 48-56.