International Journal of Engineering & Technology, 7 (2.7) (2018) 168-171

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

K-core analysis and modeling for network centralities

V MNSSVKR Gupta[1]*, Ch.V.Phani Krishna[2]

[1]Research Scholar, Department of CSE,K L Educational Foundation,Vaddeswaram,Guntur, AP-522502, India
[2]Professor, Department of CSE, K L Educational Foundation,Vaddeswara,Guntur, AP-522502, India
**E-mail: guptavkrao@gmail.com*

Abstract

Network modeling is the interdisciplinary study of relationships. Network analysis deals with relational data and Network modeling represents the interdisciplinary study of relationships. Network structure can be studied at many different levels. Around 1000 article titles on cancer, published in journal resources were considered as a dataset. Data exploration was done through displaying nodes and edges in various layouts. With a term frequency limit of 100, nearly 64 terms appeared which less than 1% sparse is. Word cloud data was plotted using word frequencies from term matrix data. An undirected network graph plotted and evaluated density, average path length and modularity, which were found to be within limits. K-cores have also been used to analyze the connectivity of a network. Network centralities such as Degree centrality, Closeness, Eigenvector and between's centrality resulted in node *carcinoma* being more central in the network.

Keywords: *Network analysis, network structure, community structure, network.*

1. Introduction

Cancer has been reported as one of the most serious health problem faced by the human community [1]. Solid tumors are formed due to many cancers. The tumors are formed with consolidation of tissues, blood cancer like leukemia from tumors. Some tumors are cancers and some are not. This male grant natural travel to other parts of the body using blood or lymph system as the medium and so far new tumors. Cancerous tumors are malignant and travel to distant places in the body through the blood or the lymph system and form new tumors. Cancer is a genetic disease and genetic changes may be inherited [2]. Over 100 types of cancers affect humans [3], [4].

In males, lungs prostate, colorectal and stomach are the most predominant organs of cancer. In female's breast, colorectal, lung and cerrical cancers are most common [5]. After lung cancer, breast cancer is the second among universal causes of deaths in women [6].

This work uses network analysis for the study. The network analysis is a combination of activities like hypothesizing, creation of modules, research using empirical evidence and high level data analysis. The aim is to understand the structure of the network. The network structure can be understood at different stages like the clyad, triad, sub group or sometimes the total network. The theories of network can be postulated at multiple levels as a reason of this variety of structural questions can be studied at the same time. For this, methods that go beyond the regular methods, where every individual is treated as a separate analysis unit. The same is also true for the entire network.

2. Materials and methods

2.1 Dataset

A dataset of 1000 article titles associated with the cancer disease published in various journal resources were considered in the study. Data was considered from PubMed database (www.ncbi.nlm.nih.gov/pubmed). Data in the form of article title information was only considered. The dataset with titles, author names and year were segregated and used as csv file input.

2.2 Network analysis

Network analysis is used for identification of drug targets, proteins or gene function identification, formulating effective plans for treatment of different diseases etc in the field of biology and medicine. Networks are created as *nodes* and *edges*. Visual depiction of networks was made to understand the network data [7]. Data exploration was about 9 crore people suffered from cancer in 2015 [8]. Adjacency matrix was used as data structure to store network graph representations. Different network properties such as *graph density* and *sparsity* provided valuable insight into the internal organization of a network. *Clustering Coefficients* calculated as they imply the measurement that shows the tendency of a graph to be divided into clusters. A cluster is a subset of vertices that contains lots of edges connecting these vertices to each other. *Centralization* measurements were carried out to assess whether a network has a star-like topology or whether the nodes of the network have on average the same connectivity.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Network centralities such as Degree centrality [9], Closeness centrality [10], Eigenvector centrality and Betweenness centrality were evaluated.

2.3 K-cores: Coreness of a network

The k-core of a graph is the maximal sub graph in which every vertex has at least degree k. The cores of a graph form layers: the (k+1)-core is always a sub graph of the k-core. This function calculates the coreness for each vertex [11] [12].

3. Results and discussion

The dataset was extracted from PubMed database. It was observed that the term 'cancer' in pubmed literature database resulted in 3528413 entries (as on 31st August 2017), however, only top 1000 entries were selected as final dataset. The dataset exported as excel csv file was visually inspected to identify common words that appear more times in title subjects. Few words such as *'and', 'with', 'for', 'in', 'on', 'of', 'from', 'a', 'to', 'an', 'as', 'do', 'after', 'it', 'study', 'effect', 'test', 'clinic', 'like' 'use', 'role'* were excluded as commonly occurring words. Hyperlinks, if any, punctuations, numbers and extra spaces between words were also excluded from the dataset.
A term document matrix (TDM) was created from the corpus where rows correspond to documents and columns correspond to terms. With a frequency limit of 100, nearly 64 terms appeared and the term frequencies are given in Table-1 and Figure 1.

Table 1: Frequency of all 64 terms that showed against a frequency limit of 100.

word	freq	word	freq	word	freq
cancer	555.2098	factor	158.9051	receptor	120.075
patient	359.84	pathway	158.9051	develop	117.0428
tumor	301.1464	review	155.3592	model	117.0428
breast	288.5218	surviv	155.3592	predict	117.0428
carcinoma	246.1379	outcom	144.436	protein	117.0428
therapi	231.5993	advanc	140.6942	adenocarcinoma	115.9172
target	220.4183	effect	140.6942	enhanc	108.6393
activ	193.5458	inhibitor	140.6942	impact	108.6393
prostat	193.5458	metastasi	140.6942	manag	108.6393
associ	192.0856	novel	140.6942	molecular	108.6393
regul	192.0856	circul	136.899	mutat	108.6393
express	183.9964	evalu	136.899	prognost	108.6393
diseas	177.4468	chemotherapi	132.1525	gastric	105.8098
human	175.9842	invas	130.6133	women	105.8098
potenti	174.114	function	129.1397	screen	104.6933
analysi	167.3273	signal	125.1711	biomark	104.3255
inhibit	165.8629	therapeut	125.1711	malign	104.3255
treatment	165.8629	health	122.6163	prevent	104.3255
clinic	163.8709	bladder	121.1398	squamous	104.3255
colorect	163.8709	growth	121.1398	nanoparticl	101.4183
promot	162.4059	trial	121.1398		
detect	158.9051	ovarian	120.075		

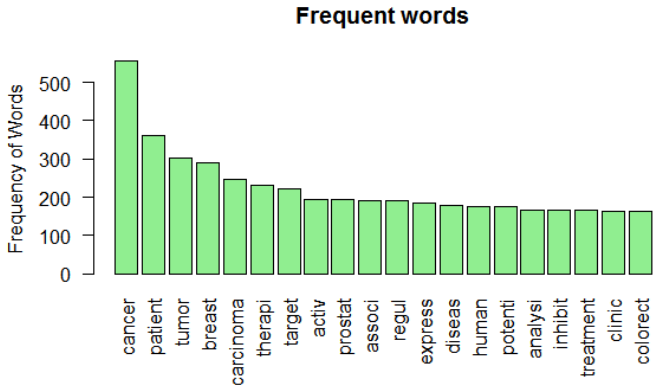

Figure 1: Graphical representation of top 20 terms within a frequency limit of 100.

2.4 K-cores: Coreness of a network

K-cores is used to analyze the connectivity of a network, and implemented in several domains, such as clustering, community structures etc. In order to better understand the concept, coreness structure was visualized for the network (Figure 2) and depicted in table 2.

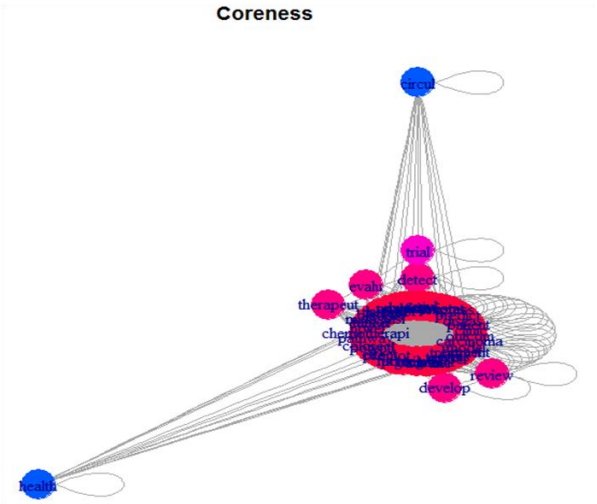

Figure 2: Coreness of the dataset network showing core elements and objects 'health' and 'circul' with low core values.

It can be observed from Figure 2 that the core elements can be differentiated from periphery elements where 'health' and 'circul' represented low core value 15 (blue color) whereas the central objects have core values of 23 and pink color objects represented core value 22.

Table 2 Coreness(g)

activ	advanc	analysi	associ	bladder	breast
23	23	23	23	23	23
cancer	carcinoma	chemotherapi	circul	clinic	colorect
23	23	23	15	23	23
detect	develop	diseas	effect	evalu	Express
22	22	23	23	22	23
factor	function	growth	health	human	Inhibit
23	23	23	15	23	23
inhibitor	invas	metastasi	model	novel	Outcom
23	23	23	23	23	23
pathway	patient	potenti	predict	promot	prostat
23	23	23	23	23	23
protein	regul	review	signal	surviv	Target
23	23	22	23	23	23
therapeut	therapi	treatment	trial	tumor	
22	23	23	21	23	

2.5 Network Centralities

Generally, central nodes or intermediate nodes affect the topology of the network. Some points are not central to the data, however, might have crucial role, hence it is important to detect such nodes using Degree centrality, Closeness centrality and Eigenvector Centrality. The values of centrality measurements are provided in Table 3. The *Betweenness* is shown in Table 4.
Betweenness centrality displays high ranks for nodes which are intermediate between neighbors. The centralities of all terms representing nodes are given below; where it is evidenced that node *carcinoma* is more central

Table 3: Degree centrality, Closeness and Eigenvector centrality measurements of network graph.

Measurement	**Value**
Degree centrality	0.3913
Betweenness centrality	0.0224
Closeness centrality	0.5662
Eigenvector centrality	0.3720

.

Table 4: Betweenness(g)

activ	advanc	analysi	associ	bladder	breast
37.5	3.5	22.0	46.0	1.0	25.0
cancer	carcinoma	chemotherapi	circul	clinic	colorect
1.0	56.0	2.0	7.5	4.5	28.0
detect	develop	diseas	effect	evalu	Express
5.0	0.0	39.0	12.5	5.5	27.0
factor	function	growth	health	human	inhibit
13.5	11.0	3.0	0.0	90.0	7.0
inhibitor	invas	metastasi	model	novel	Outcom
1.0	3.5	4.0	1.0	3.0	16.0
pathway	patient	potenti	predict	promot	prostat
2.0	38.0	39.0	0.0	5.0	49.5
protein	regul	review	signal	surviv	target
3.0	11.0	1.0	4.5	4.0	59.0
therapeut	therapi	treatment	trial	tumor	
3.0	16.0	25.0	0.0	21.0	

4. Conclusions

A term document matrix from the corpus of a dataset of 1000 article titles published in various journal resources on cancer disease resulted in 64 terms against a frequency limit of 100. An agglomerative hierarchical clustering analysis ensued clusters of 27 terms with different types of cancer appearing under different clades. K-core analysis revealed periphery elements such as ‘health’ and ‘circul’ with central elements having core values and it was observed that breast cancer and lung cancer predominated. Betweenness centrality, which displays high ranks for nodes evidenced that node *carcinoma* is more central

References

[1] Parkin DM, Laara E, Muir CS., "Estimates of worldwide frequency of sixteen major cancers in 1980", International Journal of Cancer, vol.41, issue.2, pp.184–197, 1988.

[2] World Health Organization, "Cancer Fact sheet N°297", February 2015.

[3] Anand P, Kunnumakkara AB, Kunnumakara AB, Sundaram C, Harikumar KB, Tharakan ST, Lai OS, Sung B, Aggarwal BB, "Cancer is a preventable disease that requires major lifestyle changes", Pharmaceutical Research, Vol.25, issue.9, pp.2097–2116, 2008

[4] Stewart BW, Wild CP, "World cancer report 2014", Health, 2017.

[5] “How is cancer diagnosed?". American Cancer Society, 2013

[6] Ravikumar S, Fredimoses M, Gnanadesigan M. “Anticancer property of sediment actinomycetes against MCF-7 and MDA-MB-231 cell lines”, Asian Pacific Journal of Tropical Biomedicine, Vol.2,issue.2, pp.92-96, 2012

[7] Linton C. Freeman, “Visualizing Social Networks", Journal of Social Structure, Vol.1, issue.1, 2000.

[8] Fruchterman, Thomas M. J.; Reingold, Edward M, "Graph Drawing by Force-Directed Placement", Software – Practice & Experience, Wiley, Vol.21, issue.11, pp.1129–1164, 1991.

[9] Levy SF, S ML, “Network hubs buffer environmental variation in Saccharomyces cerevisiae”, PLoS Biol, Vol.6, issue.11, 2008.

[10] Ma H-W, Z A-P, “The connectivity structure, giant strong component and centrality of metabolic networks”, Bioinformatics, Vol.19, issue.11, 2003.

[11] Vladimir Batagelj, Matjaz Zaversnik, "An O(m) Algorithm for Cores Decomposition of Networks”, 2002.

[12] Seidman SB., "Network structure and minimum degree", Social networks. Vol.5, issue.3, pp.269-87, 1983.