

Discovering Prominent Key Opinion Leaders: A Corona Pandemic Case

Priya Shelke, PhD

Faculty, Information Technology
Vishwakarma Institute of Information Technology
Pune, India

Manas Kulkarni

Student, Information Technology
Vishwakarma Institute of Information Technology
Pune, India

Suruchi Dedgaonkar

Faculty, Information Technology
Vishwakarma Institute of Information Technology
Pune, India

Sanul Raskar

Student, Information Technology
Vishwakarma Institute of Information Technology
Pune, India

ABSTRACT

A 2019 novel coronavirus disease (COVID- 19) outbreak in Wuhan, China has spread quickly nationwide. This paper seeks to provide a forum for the detection of KOLs (Key Opinion Leader) in pharmaceuticals and used network analysis to study coronavirus. Our system comprises of the following modules-module 1 is the user interface for entering keywords that show a collection of KOLs and knowledge database, module 2 consists of libraries for network forming dependent on keyword entry, module 3 consists of algorithms for analysis of network formed. This paper contains a literature study that was conducted using network research to classify KOLs. The paper contains the following sections I) Literature survey II) Network Analysis Algorithms III) Our Work IV) Result with COVID-19 keyword V) Comparison with Page Rank algorithm.

Keywords

KOLs, covid19, coronavirus, network, analysis, author, publication, influencer, leaders, pharmaceuticals

1. INTRODUCTION

In December 2019, a group of patients with pneumonia of unknown cause was linked to a seafood wholesale market in Wuhan, China. Through the use of unbiased sequencing in samples from patients with pneumonia a previously unknown betacoronavirus was discovered. This novel coronavirus was named 2019 nCoV, which forms a clade within the subgenus sarbecovirus, Ortho Coronaviridae subfamily. Human airway epithelial cells were used to isolate a novel coronavirus. It is different from both MERS-CoV and SARS-CoV, 2019-nCoV is the seventh member of the family of coronaviruses that infect humans [1]. Till 30 May 2020, it has infected 58,03,416 people. Total deaths because of this virus, all over the world are 3,59,791 [2]. Here in this paper a platform is proposed that will facilitate finding the right researcher (KOL) for the job. KOLs are the representatives of the scientific community who, over a long period of time, perform research in a particular domain. They hold in-depth, specialized and current expertise in their profession and therefore have the ability to develop their competence policies and guidance. This paper aims to concentrate on the most prominent KOLs who have been operating on Covid 19. Author target to use European PubMed Central's data for collecting information about such researchers, their

publications, and citations. Using this data, a network of authors will be created and then analyzed against network analysis algorithms.

2. LITERATURE SURVEY

Hengmin Zhou and Daniel Zeng [3] have researched Finding Leaders from Opinion Networks. They have proposed a page rank like algorithm and given name to that algorithm as an Opinion Rank algorithm. They had collected data from e-commerce websites on which visitors could read new and old reviews about a variety of items to help them decide on the purchase. They used this dataset to create social networks. Their work was mainly based on the opinion of people. To evaluate the performance of their proposed Opinion Rank method, they have compared it with two benchmark methods, PageRank and Random Sampling, also they have used activity-based rank and registering time as third and fourth comparing methods. Jiangjiao Duan, Jianping Zeng and BanghuiLuo [4] have worked on Identification of Opinion Leaders Based on User Clustering and Sentiment Analysis. They found out Opinion Leaders based on User Clustering and Sentiment Analysis. Guba forum, a well-known Chinese stock forum They selected it as a message source. They have used k-means clustering for which 5 attributes such as length of message and number of Identification of KOLs in Pharmaceuticals Using Network Analysis replies are selected. Then to find opinion leaders, fuzzy membership function is used which measures the degree of truthfulness. The actual value of stock and the predicted value is compared to find the opinion leaders. Huanhuan Liu, Xiaoqing Yu, Jing Lu [5] have worked on identifying Top-N Opinion Leaders for finding the people who have an important influence on message propagation on local social networks. They apply different centrality measures at once on China's local social network 'sina microblog' which is called a hybrid of Facebook and Twitter. To find the accuracy of their proposed algorithm they compare their results with the results of the Page Rank and the HITS algorithm and got to the conclusion that their proposed algorithm works better than the mentioned algorithms. The accuracy of the synthesis centrality algorithm increases with the number of opinion leaders' Increasing as compared to page rank and HITS algorithm.

3. PROPOSED WORK

Our system is divided into three main modules as seen in Fig. 2.

A. Module 1

ReactJS is used to build our WebApp which allows users to enter the keyword to find Key Opinion Leaders. If the keyword is valid and research papers regarding that keyword are present in the database then those papers are sent to Module 2 for network formation. If the data is not present then the user is prompted to scrape the data. Authors are scraping data from the European PubMed website - an archive of life sciences journal literature. For the scraping process, technologies like Scrapy and Selenium are used. Data is stored in the Elasticsearch database which is a NoSQL database. A record consists of a research paper title, authors, citation, published year, keyword, and paper ID.

B. Module 2

This module consists of Network Formation. When the keyword is sent from Reactjs to the Django server, records are retrieved from the database using the keyword. The data is pre-processed to meet requirements for the NetworkX library's input. In network, every node represents an author while every edge represents at least one publication common between the two nodes (authors). First distinct author nodes are formed. Then according to paper, edges are formed between nodes with edge weight based on the number of citations for that publication. If the authors of the edge have contributed to more than one research paper together then edge weight will be the sum of citations of those papers.

From the below table, three sample records will be given as input to Python's NetworkX library. Author A and B have contributed to Paper 1 and Paper 3 so the weight of edge A to B is 400 i.e. sum of citations of papers 1 and 3[6].

Table 1. Sample data for network formation

Paper	Authors	Citation
Paper 1	A, B,C,D	100
Paper 2	A,L,M	200
Paper 3	A,B,X,Y,Z	300

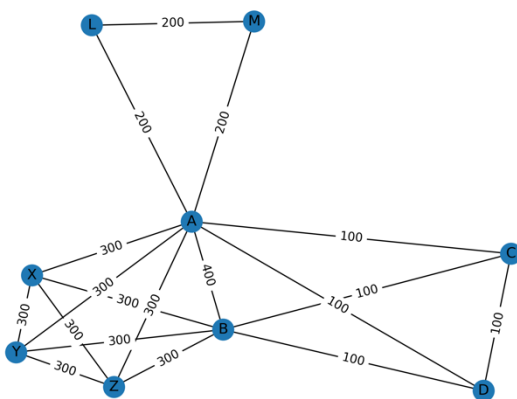


Fig 1: Weighted Network of authors

C. Module 3

This module consists of algorithms used for analysing the network. The algorithms include Degree Centrality, Closeness Centrality, Betweenness Centrality, Eigenvector Centrality[7].

- *Degree centrality*: Degree of a node means number of nodes that a particular node is connected to. Using degree centrality algorithm the degree of each node is calculated.
- *Closeness centrality*: Closeness centrality algorithm is used to calculate the minimum distance of each node from all other nodes in the network and do their sum. This results in greater weight lesser the distance and vice versa.
- *Betweenness centrality*: Betweenness centrality algorithm is used to calculate the sum of minimum distance between two nodes if the path between those two nodes is passing through the node for which we are calculating betweenness centrality and divide it by the total number of shortest paths present between those two nodes.
- *Eigenvector Centrality*: In this algorithm authors first formulate adjacency matrix A (as have given in formula) $n \times n$ matrix A and then consider vector v of having dimension $n \times 1$. Vector v contains values of each node in terms of their influence; greater value node will have greater influence in network. And one damping factor is also used in order to control increasing vector v values after each iteration. Keep on doing iterations till it converges to some value.

The network formed in Module 2 is then given as input to Degree Centrality, Closeness Centrality, Betweenness Centrality, and Eigenvector Centrality Algorithms [8]. The output of these algorithms is sorted in descending order based on the centrality values. Then only the top 10 authors with the highest centrality values from these outputs are considered for further process. Now system have 40 authors and need to find 10 key opinion leaders from them. First, find the frequency of occurrences of distinct authors in the output and if they have occurred in most of the algorithm outputs then consider them as Key Opinion Leaders. If such common authors are less then to find KOLs, take the highest valued authors from centrality algorithms and priority for the algorithm can be given as Eigenvector centrality, Betweenness Centrality, Closeness Centrality, and Degree Centrality. The top 10 influencers identified from these algorithms are then stored in the database. The output from Module 3 is a list of 10 Key Opinion Leaders which is shown on WebApp using D3 visualization of force-directed graph and in the table.

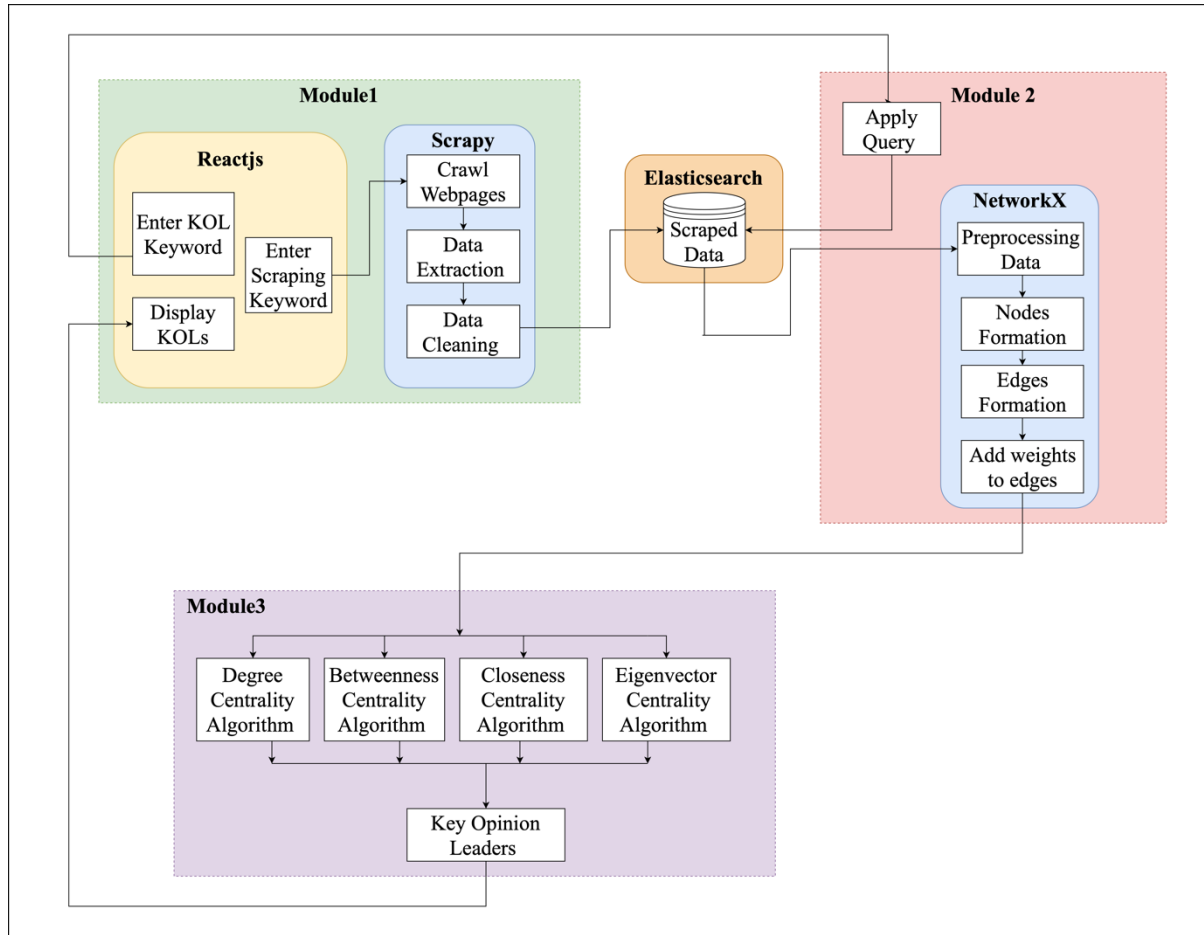


Fig 2: Block Diagram

4. RESULT WITH COVID-19 KEYWORD

Total 1236 coronavirus research papers were scraped from PubMed, out of which 986 are the most cited papers and 250 are the most recent papers published in the year 2020. The list of 10 Key Opinion Leaders is displayed in the table and the network is rendered which consists of red-coloured nodes as KOLs. Authors who have contributed more will have more edges associated with him and can find such nodes in the centre of the network. Authors with fewer contributions will be with fewer edges and their nodes can be found away from the centre. There are 4655 nodes and 25411 edges in this network. Centrality algorithms like Degree Centrality, Closeness Centrality, Betweenness Centrality, and Eigenvector Centrality will take this network as input and find out respective centrality measures for all nodes. The output is sorted in descending order and only the top 10 authors which have the highest centrality measure value are selected from each centrality algorithms. Selected authors from the output of algorithms for the keyword COVID-19 are shown in the tables below.

Table 2. Output of degree centrality algorithm

Author	Degree Centrality
Yuen KY	12.253975075204126
Poon LL	10.987107864202837
Guan Y	10.951869359690589
Drosten C	10.614740008594758
Chan KH	10.367640739149119

Peiris JS	9.890631714654061
Lim W	9.796089385474861
Osterhaus AD	9.132574129780833
Fouchier RA	9.096691018478728
Bellini WJ	8.341856467554791

Table 3. Output of betweenness centrality algorithm

Author	Betweenness Centrality
Hsueh PR	3.493760408993646e+231
Liu Q	1.1499097228367653e+227
Ippolito G	2.409394581467681e+226
Wang Y	1.4736662939897249e+226
Zhang Y	7.946733940474649e+225
Zhang J	9.157158840551274e+223
Wang W	4.64548721039794e+222
Drosten C	1.3530410062465845e+222
Zhang H	4.894722547460989e+221
Zumla A	1.8773661869966268e+220

Table 4. Output of closeness centrality algorithm

Author	Closeness Centrality
Drosten C	0.005687738652300591
Guan Y	0.005687738652300591
Li W	0.005687738652300591
Zhang LJ	0.005687738652300591

Cheng VC	0.005687738652300591
Chan KS	0.005687738652300591
Hung IF	0.005687738652300591
Lau SK	0.005687738652300591
Chan KH	0.005687738652300591
Tsoi HW	0.005687738652300591

Table 5. Output of eigenvector centrality algorithm

Author	Eigenvector Centrality
Yuen KY	0.3581499446892206
Guan Y	0.3281281131086142
Poon LL	0.3205739920215614
Chan KH	0.3159702144073491
Peiris JS	0.31143111961630515
Lim W	0.2318847662998414
Cheng VC	0.21612569407408053
Woo PC	0.15127453879946545
Lau SK	0.14468542677561083
Zheng BJ	0.13189581723722946

From 40 authors, the most influential authors are then selected which are most common among four algorithm outputs and are termed as Key Opinion Leaders for a respective keyword

of pharmaceuticals term. Key Opinion Leaders for our COVID-19 network are shown in table below.

Table 6. Key opinion leaders for covid-19 keyword

Key Opinion Leaders
Drosten C
Chan KH
Guan Y
Poon LL
Peiris JS
Yuen KY
Lim W
Lau SK
Cheng VC
Woo PC

Research Paper Title	Paper Authors	Cited	PMID	Date	Keyword
A novel coronavirus associated with severe acute respiratory syndrome.	Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, Tong S, Urbani C, Comer JA, Lim W, Rollin PE, Dowell SF, Ling AE, Humphrey CD, Shieh WJ, Guarner J, Paddock CD, Rota P, Fields B, DeRisi J, Yang JY, Cox N, Hughes JM, LeDuc JW, Bellini WJ, Anderson LJ, SARS Working Group	1393	PMID:12690092	10 Apr 2003	covid19
Identification of a novel coronavirus in patients with severe acute respiratory syndrome.	Drosten C, Günther S, Preiser W, van der Werf S, Brodt HR, Becker S, Rabenau H, Panning M, Kolesnikova L, Fouchier RA, Berger A, Burguière AM, Cinatl J, Eickmann M, Escrich N, Grywna K, Kramme S, Manuguerra JC, Müller S, Rickerts V, Stürmer M, Vieth S, Klenk HD, Osterhaus AD, Schmitz H, Doerr HW	1330	PMID:12690091	10 Apr 2003	covid19
as a possible cause of severe acute respiratory syndrome. Coronavirus	Peiris JS, Lai ST, Poon LL, Guan Y, Yam LY, Lim W, Nicholls J, Yee WK, Yan	1118	PMID:12711465	01 Apr 2003	covid19

Fig 3: Most cited papers of Coronavirus in database

Emerging novel coronavirus (2019-nCoV)-current scenario, evolutionary perspective based on genome analysis and recent developments.	Malik YS,Sircar S,Bhat S,Sharun K,Dhama K,Dadar M,Tiwari R,Chaicumpa W	0	PMID:32036774	01 Dec 2020	covid19
The different clinical characteristics of corona virus disease cases between children and their families in China - the character of children with COVID-19.	Su L, Ma X,Yu H,Zhang Z,Bian P,Han Y,Sun J,Liu Y,Yang C,Geng J,Zhang Z,Gai Z	0	PMID:32208917	01 Dec 2020	covid19
COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses.	Shereen MA,Khan S,Kazmi A,Bashir N,Siddique R	0	PMID:32257431	16 Mar 2020	covid19
COVID-19: From bench to bed side.	Singh A,Shaikh A,Singh R,Singh AK	0	PMID:32283498	09 Apr 2020	covid19
The 2019 Novel Coronavirus: A Crown Jewel of Pandemics?	Azamfirei R	0	PMID:32104726	31 Jan 2020	covid19
Emergence of Novel Coronavirus and COVID-19: whether to stay or die out?	Biswas A,Bhattacharjee U,Chakrabarti AK,Tewari DN,Banu H,Dutta S	0	PMID:32282268	13 Apr 2020	covid19
Coronaviruses: origin and evolution.	Sen S,Anand KB,Karade S,Gupta RM	0	PMID:32341622	27 Apr 2020	covid19
Host Immune Response and Immunobiology of Human SARS-CoV-2 Infection.	Kumar S,Nyodu R,Maurya VK,Saxena SK	0	PMCID:PMC7189399	30 Apr 2020	covid19
The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak - an update on the status.	Guo YR,Cao QD,Hong ZS,Tan YY,Chen SD,Jin HJ,Tan KS,Wang DY,Yan Y	0	PMID:32169119	13 Mar 2020	covid19

Fig 4: Most recent papers of Coronavirus in database (up to year 2020)

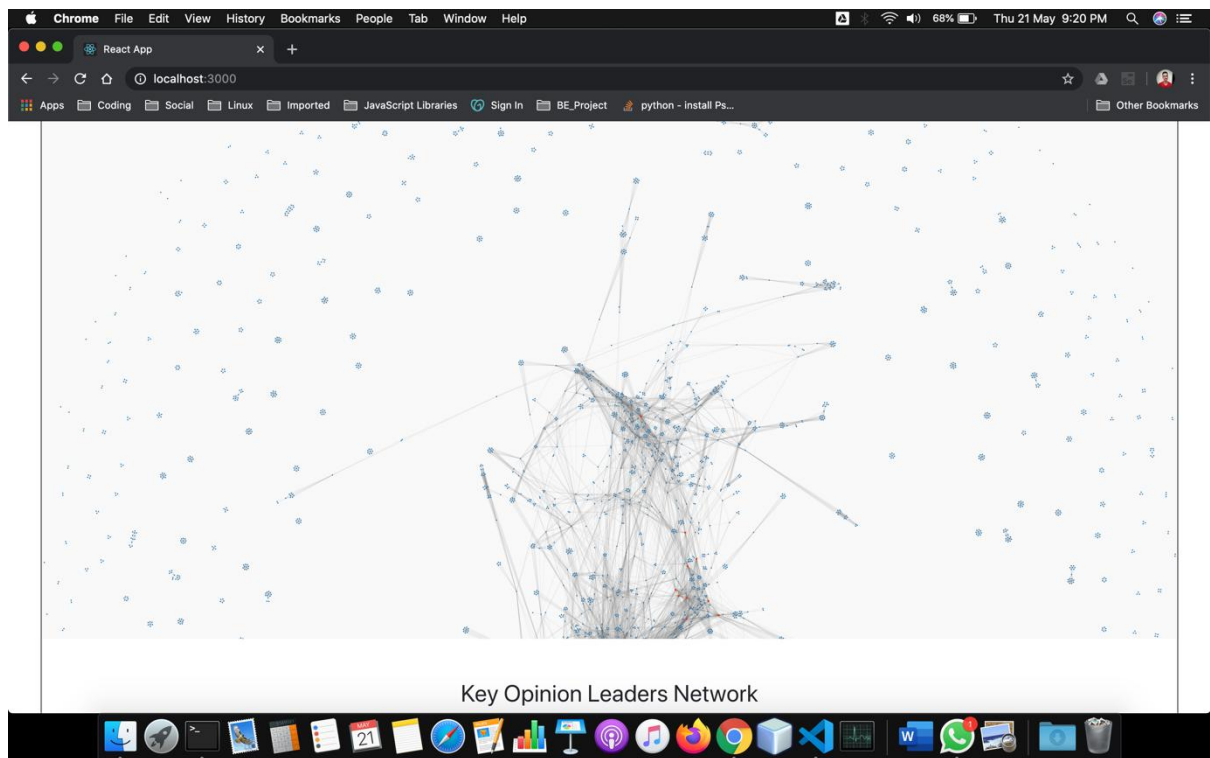


Fig 5: Complete Network

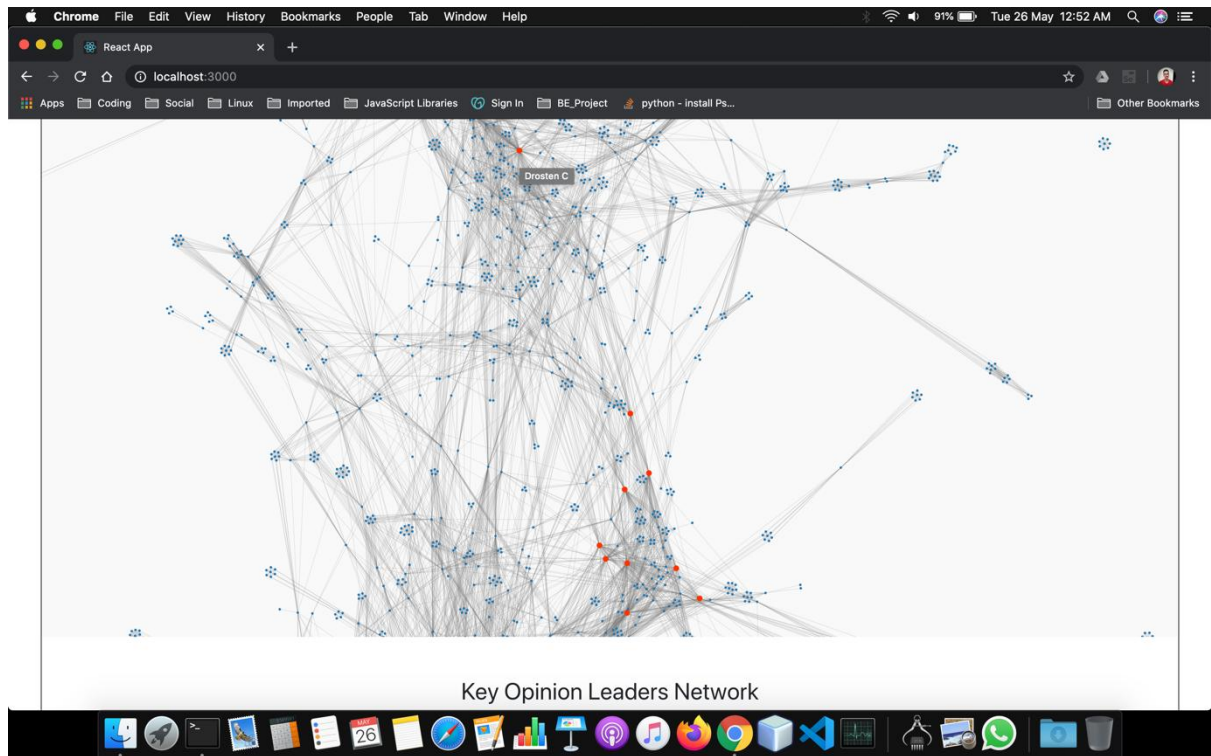


Fig6: KOLs in the Network

5. COMPARISON WITH PAGE RANK ALGORITHM

For comparison, the same network is given as input to both the Centrality algorithm and the Page Rank algorithm. The output of the page rank algorithm was sorted in descending order of values and only the top 10 are considered for comparison. Our system have considered documents by author, their total citations, and h-index from Scopus - a citation database. H-index is author level metric used to measure the impact of their publications. It is the highest number of papers included that have at least the same number of citations[9]. Page Rank Algorithm is used by Google to rank web content using hyperlinks. In this algorithm value for each node is assigned based on incoming links and edge weight depends upon score value of origin node. This algorithm tells that nodes with more incoming links are influential and others that are connected to these nodes are also influencing to some extent [10].

Table 7. Kols using centrality algorithms

Author	Total documents	Total Citations	h-index
Drosten C	388	21002	74
Chan KH	255	18156	68
Guan Y	324	40634	97
Poon LL	225	22710	73
Peiris JS	752	54595	110
Yuen KY	1024	49838	99
Lim W	97	16509	50
Lau SK	434	15065	59
Cheng VC	264	12180	54
Woo PC	551	18780	66

Table 8. Kols using page rank algorithm

Author	Total Documents	Total Citations	h-index
Drosten C	388	21002	74
Yuen KY	1024	49838	99
Poon LL	225	22710	73
Osterhaus AD	1310	71162	114
Guan Y	324	40634	97
Fouchier RA	432	42080	94
Chan KH	255	18156	68
Peiris JS	752	54595	110
Gorbalenya AE	184	16643	69
Memish ZA	643	61772	85

Out of 10 authors 6 authors are common in both algorithms for COVID-19 keyword. Common authors in outputs are highlighted with yellow color.

Our system has tested on different keywords with approximately 1000 papers for each keyword and results are shown in Table IX. The network for all keywords had on average 4,000 nodes and 25,000 edges. Total considered KOLs column show number of authors considered for comparison and Common authors shows authors common between our system and Page Rank algorithm.

Table 9. Comparison for various keywords

Keyword	Total considered KOLS	Common Authors
COVID-19	10	6
Typhoid	10	7
Asthma	10	5
Cancer	10	7
Malaria	10	4

Thus, it can be determined that approximately 60% KOLs of centrality algorithms match with Page Rank algorithm. The non-common authors also have good h-index score which represents that they also have significant contribution towards their research interest. This match can be increased by considering a greater number of KOLs, however system has considered only top 10 KOLs for comparison purpose.

6. CONCLUSION

In this paper, authors have tried to identify KOLs that operate in the Covid19 field. They have developed a weighted network and used network analysis algorithms to classify the influential researchers who can play a critical role in discovering a way to save the planet from this COVID-19 pandemic. We contrasted our algorithm with the Page-rank algorithm, which showed approximately 60 percent matching KOLs.

7. REFERENCES

- [1] Zhu, Na, Dingy Zhang, Weanling Wang, Xingang Li, Bo Yang, Jindong Song, Xiang Zhao et al. "A novel coronavirus from patients with pneumonia in China, 2019." *New England Journal of Medicine* (2020).
- [2] COVID-19 live Map- <https://news.google.com/covid19/map>
- [3] Zhou, H., Zeng, D. and Zhang, C., 2009, June.: Finding leaders from opinion networks In: *2009 IEEE International Conference on Intelligence and Security Informatics* (pp. 266-268). IEEE
- [4] Duan, J., Zeng, J., Luo, B.: Identification of Opinion Leaders Based on User Clustering and Sentiment Analysis. In: *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*
- [5] Liu, H., Yu, X., Lu, J.: Identifying Top-N Opinion Leaders on Local Social Network. In: *Smart and Sustainable City 2013 (ICSSC 2013)*, IET International Conference on Date 19-20 Aug. 2013.
- [6] Identifying Key opinion leaders using social network analysis. Cognizant System Private Limited's White Paper on June 2015.
- [7] Zhou, H., Zeng, D., Zhang, C.: Finding Leaders from Opinion Networks. In: *Intelligence and Security Informatics, 2009. ISI '09. IEEE International Conference on Date 8-11 June 2009.*
- [8] Duan, J., Zeng, J., Luo, B.: Identification of Opinion Leaders Based on User Clustering and Sentiment Analysis. In: *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*
- [9] Scopus - <https://www.scopus.com>
- [10] EigenCentrality & PageRank - <https://cambridge-intelligence.com/eigencentality-pagerank/>