Citation Network Analysis of Scientific Publications for Coronavirus 2019

Gulcin Ozlem Atalay, Nazrin Abbasli, Fatih Enes Usta
Istanbul Sehir University
34865 Dragos, Istanbul, Turkey
{gulcinatalay2019,nazrinabbasli2019,fatihusta}@std.sehir.edu.tr

11 May,2020

Abstract

The aim of citation network analysis is to acknowledge the key publications and dominant researchers in the COVID-19 area. The data of this analysis has been prepared by the leading research groups. The data contains 47,000 scholarly articles about COVID-19, SARS-Cov-2 and other related corona viruses. Citation network graph is a directed graph with nodes as articles and edges are citations. In this study we measured the indegree centrality scores (number of citations), tried to find the most influential node which has highest page rank score, calculated the subcommunities and the most important keywords used in the articles.

1 Introduction

At December 2019, the world came across a new epidemic, while China announced to the world, a new kind of virus, which comes from the coronavirus family was seen in their country, namely in the city of Wuhan. In short time period, it is detected that the virus is highly infectious and quarantine is a requirement to protect healthy people from diseased ones. In time 'quarantine' has become a new way all living all over the world, schools were closed, some employees took compulsory holidays, some forced to leave their job or started to work at home. Transportation is limited and arranged according to decrease the contact of people as much as possible. Addition to this frightening virus, those strict quarantine conditions also affect people economically bad.

While the virus, which is defined as Covid-19 by researchers, has been spreading throughout people in a furious way, lots of researches simultaneously started all over the world to define the symptoms, to protect healthy people, to develop a medical treatment or vaccine or to compensate for the negative economic effects of the virus, on countries, all around the world. All scientific studies focused almost only this new issue and lots of articles related these new virus are published.

1.1 Motivation

Although the scientific studies we are interested in resulted from an emergency situation, a pandemic issue world come across almost every hundred year period, it is a chance to analyse such a big worldwide problem and measure the scientific reflections against this issue.

i. We are interested in the flow of knowledge on pandemic issues.

ii. We constructed our citation structure to make comments on our publication network.

2 Literature Review

Citation is the quotation source from research papers, books and websites in a research work. Scientific research papers (SRP) can contain several hundreds of references. Citation is used to measure the popularity. The history of citation network analysis goes back to the early 1960s to Garfield E, Sher IH, and Torpie RJ (1964), who made a citation network analysis about genetics, namely DNA. Although they mentioned in their paper about Dr. G.Allen as the first one at citational network analysis and the one who prepared a bibliographic citation network diagram for publications about nucleic acids, the scientific studies about citation network analysis started with Garfield et al [1]. The first feasible implementations of the graph theory in citation networks analysis was made by Garner (1965) [2]. In 1973, 'co-citation' term as a new measure of the relationship between two documents was used in a study made by Henry Small [3]. Another very important development about citation network analysis was made by Hummon and Doreian (1989); they introduced us with indices (NPPC, SPLC, SPNP) and main path analysis [4].

Citation network analysis is somewhere in-between social network analysis and network science. Citation graph is a directed graph where nodes stand for publications and citations are stand for edges, directed from cited paper to the citing one [5]. In particular, citation analysis can be used for:

i. Evaluating information resources and scientific contributions,

ii. Following the diffusion of ideas, and flows of information,

iii. Studying uses of scientific research papers.

It has been found that centrality measures defined by Freeman (1977) are related to the number of citations (degree centrality, closeness centrality, betweenness centrality) [6].

- Degree centrality measures the number of papers that the paper directly linked (the capacity to communicate directly with others).

- Closeness centrality measures that "how long it will take information to spread from a given vertex to others in the network" [7].

- Betweenness centrality measures the popularity of a published paper.

3 Proposed Methodology

3.1 Data Preparation and Exploration

At the beginning we have to prepare our data since it is collection of json files including 47304 articles inside. The data used for this study was taken from a Kaggle challenge, COVID-19 Open Research Dataset Challenge (CORD-19) [8]. The data is imported from file to a nested dictionary and a data frame to follow different steps. The program used in this study for calculations and visualisations is Networkx 2.4. First, for preparing our data frame we imported json files and collected them on a list with some cleaning applications. In bibliography words like "proc." or "magazine" were cleaned and the data that including "publisher's note", "world health organization", "fields virology", "united states census", "geneva: world health organization" were totally eliminated. At the end we have collected title of the articles with reference titles in a list and we turned it to a data frame. The dictionary is collected in the same way; it is constructed from imported json files with more data for each article, like title, author, year, DOI etc. The natural language processing is applied to dictionary for calculating the maximum frequent keywords used in the articles and filtering the network using those keywords.

3.2 Network properties and Visualisation

3.2.1 Graph implementation

The nodes and edges are defined by using our data frame in Networkx. The graph is constructed as a directed graph; since it is a requirement of the nature of citation analysis. If an article A makes citation from article B, it is not possible for the cited article B to make citation from A, since it is published before.

Listing 1: Number of nodes in graph

```
>>> number_of_nodes = len(list(G.nodes))
>>> print("Number of nodes:", number_of_nodes)
Number of nodes: 1271588
```

So there are 1271588 nodes and 2264535 edges in the graph.

Listing 2: Number of edges in graph

```
>>> number_of_edges = len(list(G.edges))
>>> print("Number of edges:", number_of_edges)
Number of edges: 2264535
```

3.2.2 Centrality Scores and Page Rank

The maximum in-degree score means the paper which took highest citation from others. The highest cited paper is seen as the most important paper as a reference in our network. Maximum in-degree score calculated for the network is 0.00080765. The maximum cited paper with its title is seen in listing 3 with a citation amount of 1027. This paper is published by New England Journal of Medicine, in 2012.

Listing 3: Maximum in-degree score

```
>>> in_deg_centrality =nx.in_degree_centrality(G)
>>> max(in_deg_centrality.values())
0.0008076521700835256
>>> max(dict(G.in_degree()).items(),
        key=lambda x:x[1])
isolation of a novel coronavirus from a man with
pneumonia in saudi arabia,1027
```

Page Rank computes a ranking of the nodes in the graph G based on the structure of the incoming links and it works on directed graphs. It was originally designed as an algorithm to rank web pages. So the page rank score is also a measurement of the importance of a node in the network. In our network calculated page rank scores are given in table 1, and they are in consistent with the maximum in-degree score. The article which has maximum page rank score also has maximum in-degree score and cited most.

There is also out-degree score calculated for finding the article, which makes the most citation and have the longest reference list. As seen in listing 4 the maximum

	Title	Pageranks
194	isolation of a novel coronavirus from a man wi...	2.223313e-05.
193	identification of a novel coronavirus in patie...	2.161468e-05
192	a novel coronavirus associated with severe acu...	2.043571e-05
...	...	...
0	the rna pseudoknots in foot-and-mouth disease ...	7.671025e-07

Table 1: Page Ranks of Network.

out-degree centrality score is 0.123094212 and the paper makes highest citation with 156525. Since it has no title in data set, the second highest also calculated. It is not surprising with such high citation amounts, to see their being bibliographic study.

Listing 4: Maximum out-degree score

```
>>>out_deg_centrality=nx.out_degree_centrality(G)
>>>max(out_deg_centrality.values())
0.12309421219310988
>>> print("1:",first_highest,":..,
"2:",second_highest,":...)
1:  nan : 156525
2: bibliography of the current world
literature : 8696
```

In figure 1, the ego graph of the article which has maximum second out-degree score was drawn due to get more clear visual sight and to show its centrality.

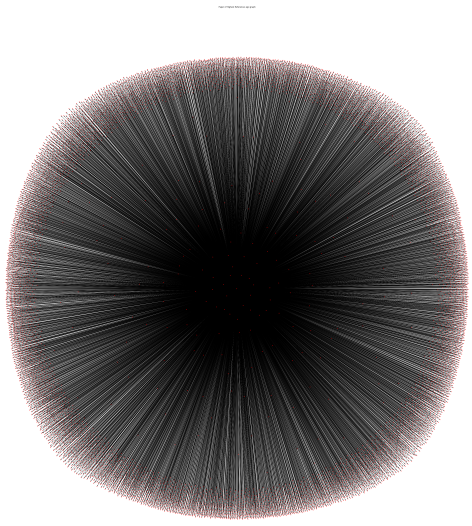

Figure 1: Ego graph of second highest out degree central article

Also another randomly selected paper's ego graph with reference titles is given in figure 2. The paper is cited 0 times. Its annexe contain 34 references.

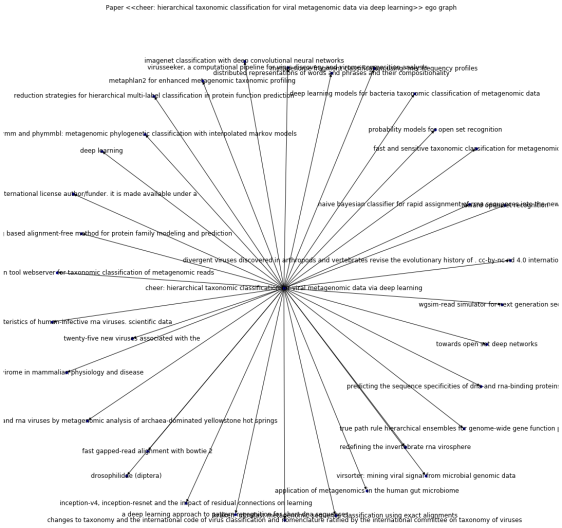

Figure 2: One random paper's ego graph

3.2.3 Keyword Detection and Visualization

After calculating page rank and centrality scores, we chose some keywords to see their visualization in our network and to divide our data with some logic. First, we chose 'corona', 'wuhan', 'sars', 'virus', 'china' etc., which were seen important and applied those words to filter our network. Secondly, we split the words of titles to tokens and cleared from stop words and calculated the most common words in titles, which were used more than 10,000 times. These were seen better to select as keywords for us and we used those keywords and drew our graph accordingly. The top 10 most frequent words used in network are: 'virus', 'respiratory', 'viral', 'human', 'infectious', 'infection', 'protein', 'influenza', 'coronavirus'. 'virus', 'infectious', 'disease' were selected and 3 graphs were drawn accordingly. As seen in figure 3, they look like each other and probably most of the articles were same.

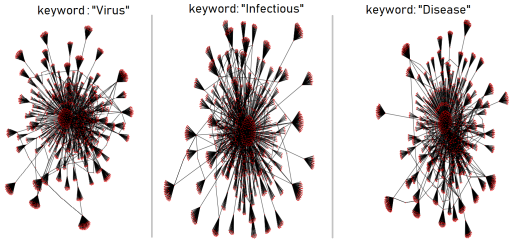

Figure 3: Graph with 3 frequently used words.

3.2.4 Sub-community Search

Since we have a directed, graph to calculate the sub-communities we should convert it to undirected graph. When connected components of sub-communities were calculated it has seen that there are 940 sub-communities in the network but there is one dominated giant component which has size of 1,258,569.00 and the others are mostly below hundred in size. So it can be said that the whole network is almost one community, which is the giant component.

Listing 5: Number of communities

```
>>> print(len(und_G))
1271588
>>> print(len(list(und_G.subgraph(c) for c in
nx.connected_components(und_G))))
940
>>> [len(c) for c in
     nx.connected_components(undirected_G)
     if len(c) > 10]
[1258569,
 35,
 35,
 ..
 ..
 11]
```

4 Conclusions

In conclusion, in this research, citation network analysis of 47,000 scholarly articles about COVID-19, SARS-CoV-2 and other related corona viruses were studied. The most cited paper, its in degree centrality and page rank scores were calculated. The paper which made highest citation with longest reference list and its out degree centrality score was found. The most important keywords were found to apply as a filter to the network and their graphs were drawn. At the end the number of sub-communities and their sizes in the network were calculated and the giant component is defined.

References

[1] E. Garfield, I. Sher, and R. Torpie, *The Use of Citation Data in Writing the History of Science*. Institute for Scientific Information Inc. Philadelphia, Pennsylvania, USA, 1964.

[2] R. A. Garner, L. Lunin, and L. Baker, *Three Drexel information science studies*. Drexel University Press, 1967.

[3] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the American Society for Information Science and Technology*, vol. 24, pp. 265–269, 1973.

[4] N. P. Hummon and P. Doreian, "Connectivity in a citation network: The development of dna theory," *Social Networks*, vol. 11, pp. 39–63, 1989.

[5] L. F. Courtney, E. A. Hobson, T. C. Mendelson, R. L. Rodríguez, R. J. Safran, S. C. Scordato, Elizabeth, M. R. Servedio, C. A. Stern, L. B. Symes, and M. Kopp, "Theory meets empiry: A citation network analysis," *BioScience*, vol. 68, pp. 805–812, 2018.

[6] E. Yan and Y. Ding, "Applying centrality measures to impact analysis: A co-authorship network analysis," *Journal of the American Society for Information Science and Technology*, vol. 60, pp. 2107–2118, 2009.

[7] L. Yin, H. Kretschmer, R. A. Hanneman, and Z. Liu, "Connection and stratification in research collaboration: An analysis of the collnet network," *Information Processing and Management*, vol. 42, pp. 1599–1613, 2016.

[8] Kaggle, "Cord-19-research-challenge," 2020. Data retrieved from Kaggle open research data challenge for Covid-19, https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.