Building Citation Network from dataset: Citation Graph about COVID-19 Vaccination.

Innocente Simone, Shekhar Suman Patel

I. Abstract

***Abstract*—Citation networks are important tools in order to find relevant information and in particular when you need to focus on specific scientific domain.**
By analyzing citation network, you can understand how much a particular work has an impact to other related papers.
Specially, starting from the 2019, a lot of Covid-19 related papers have been published, and, in order to exploit this amount of works, in the current project, we are building a citation network given a dataset of Covid-19's papers.
Our work is focused on the vaccination for Covid-19, by doing that, we will discover the most important articles regarding this topic.
By exploiting the abstract, keywords and titles, we can extract papers from a dataset, and then parsing the document in order to retrieve the references. In that way we are able to build a citation network.

II. Introduction

In December 2019, an outbreak of an unknown pneumonia was reported in the city of Wuhan. Subsequently, this new disease, that was increasingly spreading, was named Covid-19.
On 11 March 2020, the WHO, World Health Organization, after assessing the severity levels and global spread of SARS-CoV-2 infection, declared that the outbreak of Covid-19 can be considered a pandemic.
In the wake of this new disease, numerous scientific papers have been written in order to analyse it and understand its strengths and weaknesses.

The rapidity with which Covid-19 was expanding led to an urgent need for a vaccine. However, a vaccine for an infectious disease had not been produced in less than several years and no vaccine existed to prevent coronavirus infection in humans. The increase in the number of infected people, therefore, led to the use of compressed schedules that shortened the standard vaccine development timeline, in some cases combining clinical trial phases over months, a process typically conducted sequentially over years.
Just as numerous articles have been written about Covid-19, as many if not more, have been written about the vaccine.
All these papers, with the respective references, create a citation network which is an important tool in order to discovering important information concern a specific topic. With it, you are able to understand the papers which have an impact on other related articles.

In this work, we are exploiting *Web Of Science* [9] in order to obtain relevant papers that concern Covid-19 and the vaccine against it.
Together with Web Of Science, we decided to use another freely available dataset for Covid-19 articles, *CORD_19* [6]. However, due to compatibility issues, we decided to use the second dataset only for including papers that have a high number of references. Web of Science and CORD_19 will be better described in the section *Dataset*.

Citation Network and its analysis are methods to map scientific papers of a particular field, in our case COVID-19 and the vaccine.
They are very useful for understanding the history of the papers and the connection between them. In particular is even more relevant in this field and in this period, where the vaccine is a priority.
Some basic property of citation networks are: [8]

- Citation networks are directed. The edges go from one document to the other and this means that the first paper is citing the second one;
- All edges in the citation networks point backwards in time;
- Vertices and edges added to the citation networks are permanent and cannot be removed;

To develop our citation network, we relied on two important dataset which contains a very high number of scientific papers. We used Web Of Science [9] to obtain the most cited articles concerning the Covid-19 vaccine.
The second resource that we utilized, was CORD-19 [6], a open research and freely available dataset. We use the number of reference as heuristic in order to retrieve titles of different papers.
Thus, we get the titles of the papers with a number of references greater than 40 and then we used Web Of Science to retrieve the paper in a format compatible for CitNetExplorer. [7]
The third tool that we used was CitNetExplorer in order to create the citation network from the file that you can download from Web Of Science. However, due to some problem with a specific tag in these file, we had to perform a preprocessing step before using CitNetExplorer.
After that we were able to exploit this tool and create the citation network. We saved the publications and citations files, and afterwards, by using the python library *networkx* [11], we created the citation network, where the nodes represents the papers and the edges the citation.

From the file obtained from Web of Science we feed the CitNetExplorer, which builds the citation network. It provides an interactive network with feature such zoom in and out with attributes of node display if we click the node.
CitNetExplorer comes with own set of algorithm for network

analysis. In addition to this it provides the tow text files called the Publications and Citations i.e the node and edges data files. For visualizing the whole citation network used Gephi which is another open-source visualization and exploration software for networks and graphs.

Finally we wanted to know which papers were most citated and we could obtain information about the clustering of topics on the research works in order to know areas of research.
The CitNetExplorer and Gephi come with most analysis algorithms to be implemented and thus task specific analysis could be done easily.

III. Related Work

In this section we briefly discuss existing methods for building a citation networks.
Our goal is to create a network of the most cited papers which are present in a dataset. Therefore we focused our search about the related works, to the creation of citation networks and extracting information from dratabase.
In *Myopia Control: A Citation Network Study* [1] the dataset was developed by using keyword as ”myopia control” in order to retrieve all papers of that specific field. By using CitNetExplorer, a software tool for visualizing the most cited publications, they extracted the ”citation score” in order to obtain the number of citations in the network.
In the end, for their purpose, subnet analysis was performed for checking the therapies for myopia.

An identical procedure was used in *Current State and Future Trends: A Citation Network Analysis of the Academic Performance Field* [5] where, in the same way as the previous paper, the database used was Web of Science (WOS), and by using the keyword ”Academic performance” they found all the desired articles.
With CitNetExplorer, they analyzed the publications and, with the ”citation score” discovered the most cited publications. In order to find a group for each paper a clustering function was applied.

In the literature you can find paper about generating automatically citation graphs, in particular in *Automatically Generating Citation Graphs (and Variants) for Systematic Reviews* [2] where the main focus was on systematic reviews. In order to extract the reference they used a tool called *Grobid*, which is able to extract metadada from scientific publications. This permits to have the information about publications for the graph. By using Grobid, the references are also extracted, which are the required information for the edges in the citation network. With this knowledge, a data model is developed which is used to create the citation network.

Another automatic method for building citation network is described in the paper *Building direct citation networks* [3], where is proposed an algorithm with computer implementation that prevent cyclic paths.
The main idea of the program is to iteratively check the list of the documents, save the references for each of them, and then check again if one reference is present in the document's list and create an edge from a cited document to the citing one.

In *Extraction and Visualization of Citation Network* [4] they proposed a particular method based on the weighting procedure for finding the relationship between paper and its citation. This paper proposed a system which ranks the most relevant cited papers by using this mechanism of weighting. In particular the procedure assign different importance by considering the keyword, the category and the common author of both paper and cited paper. After assigning these weights to the papers, you can easily find the top related articles.

IV. Dataset

CORD_19 [6] is a powerful resource for getting scientific papers. Specifically, is a Covid-19 Open Research Dataset freely available, in order to discover relevant information more quickly from the literature.
Indeed by using AI-based techniques and Natural Language processing you can extract useful information.
Born from the collaboration between the Allen Institute for AI (AI2), the White House Office of Science and Technology Policy (OSTP), the National Library of Medicine (NLM), the Chan Zuckerburg Initiative (CZI), Microsoft Research, Kaggle and Georgetown Univeristy's Center for Security and Emerging Technology (CSET), CORD_19 was released on the 16 March of 2020 with around 28K papers.
Now, it contains over 500K scholarly articles including 200K with full text, about Covid-19, SARS-CoV-2 and related coranaviruses.
As described in the official paper [6], they performed some preprocessing on the metadata and the full text. Specifically, concerning the metadata, they apply three operations:

1) Clustering duplicate papers by using identifiers;
2) Selecting canonical metadata for each cluster;
3) Filtering clusters to remove non-papers.

Instead, for the preprocessing step of the text, they used a parsing pipeline for the content that come from PDF files.

1) Parse PDFs to TEI XML files using GROBID;
2) Parse TEI XML to JSON;
3) Postprocess to clean up links between citations and bibliography.

The paper's metadata are saved in a .csv file. Instead, the text for each paper is parsed and stored in a collection of JSON files.

CORD_19 contains different attribute in the metadata file, in particular we have:

- **cord_uid**: Id for each paper;
- **title**: the paper title;
- **abstract**: the paper's abstract;
- **publish_time**: the published date of the paper;
- **authors**: the authors of the paper;
- **pdf_json_files**: path for the parsed version of the paper's PDF;
- **url**: all the URLs associated.

One advantage of using CORD_19, is that is freely available to everyone. This leads to several research on application about searching, exploration and visualization.

Another very important resource for our project is **Web Of Science**. Developed by the *Institute for Scientific Information (ISI)* and now owned by *Clarivate*, Web of Science is a website that allow you to access several databases to citation date for many different academic disciplines. You can acquire and analyze the database information in a timely manner.
This powerful tool let us filter the collection and export it in different way, like as plain text file or excel file.
You can choose to download various kind of content, for example title, authors, abstracts, keywords, etc. For our purpose, in particular we need the cited references from the papers and some metadata.
Moreover, the website allow to analyze the result with plots and tables, you can sort the results by relevance, date, citation or usage. In case you can choose to use quick filters like for example the highly cited.

The main difference with respect to CORD_19 is that with Web Of Science, you don't have access to the full paper's text, however, for our goal is not really essential. Indeed we just need metadata and the cited reference in order to develop a citation graph.
Web of Service works particularly well with another tool for visualizing citation network: CitNetExplorer, which will be discussed in the next session.

The code for this project and all the files needed are at this link.

V. Methodology

Firstly, we got papers from Web Of Science by focus our research on articles that could contain the word "Covid-19 Vaccine" since was our field.
Web Of Science allows also to put some quick filters to the results, so we searched the highly cited papers in order to get the articles which are more cited.
This website permits to export at the same time, up to 500 records with the respective content, which in our case include also the cited references.
Web Of Science export the papers in a unique file that we renamed *JOI*. Every line in this file correspond to a specific tag which indicate a particular characteristic of the paper, for example *TI* is the document title, *AU* is the list of the authors or *PT* is the type of publication.

The role that CORD_19 has in our project is to find Covid-19 Vaccine articles with a high number of reference, so we download the last available release from the official website.
After that, we created a python script in order to access this dataset. In particular we worked with paper's metadata, since in the github page of CORD_19 they recommended to use metadata.csv and then augment the data when needed with the full text which is parsed.
We searched the titles with the metadata, and if the title contains the words *Covid-19* and *Vaccine*, we access the parsed document in order to get the number of references for each paper.
We selected articles with a number of reference greater than 40, we used this kind of heuristic since we assumed that it could be a good evaluation for a cited paper. After that we search them with Web Of Science in order to have files which were compatible with CitNetExplorer.

After that, we created the JOI file which is used in order to exploit CitNetExplorer and, to do that, all the files derived from the papers filtered from CORD_19 and obtained from Web Of Science, were concatenated in a unique file and it contains all the information needed for create the citation network.

However, in some cases, in the generated JOI file was not available a particular tag called **PY**, which indicate the publication year. This caused an error with the tool and was not able to create the network.
In order to make the JOI file work with CitNetExplorer, we performed a preprocessing step, where we add the missing tag. For convenience, we decided to add '2019' as publication year for the missing field, since it was the year of the discovery of the virus. In this way we were able to give the data to CitNetExplorer in order to visualize the citation network.

Specifically, CitNetExplorer is a software tool for visualizing and analyzing citation of scientific publications. Using this program, networks could be explored interactively and clusters of closely related publications could also be identified. The tool is freely available for non-commercial research and teaching purposes. and provides, with zooming and scrolling over the citation network, to making indirectly related citations visible.
Also, network graph algorithms are provided in order to give further information such as clustering, shortest paths, core publication etc.
Finally, citation network could be saved into publications and citation text file which will be respectively the nodes and the edges.

In particular, both the file will be used for creating a citation graph by utilizing the library networkx [11]. We performed this operation with a python script that create the directed network from the files and moreover, build the .csv files that *Gephi* [10] needs to have in order to define a better visualization for the graph.

Indeed, our last step include Gephi, a free and open-source software for visualize and explore graphs. In this way we were able to use a better layout and check some metrics like betweenness centrality, closeness, diameter, clustering, coefficient, pageRank.
We can also perform community detection by exploiting the modularity.

VI. Results

The network graph generated with CitNetExplorer and then, created with the python script can be viewed in the figure 2. It is a directed network composed by 1422 nodes, each of them correspond to a publications related to Covid-19 vaccine.
Instead, there are 7763 edges that connect all the nodes. If an edge from a node goes to another one, means that the first

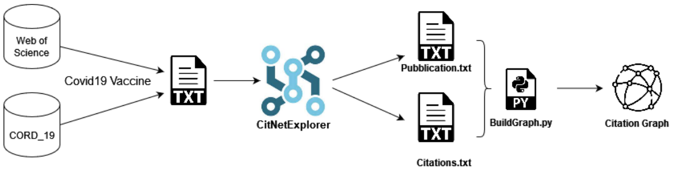

Fig. 1. Workflow of the project

TABLE I
METRICS OF THE CITATION NETWORK

Metric	***Value***
Nodes	1422
Edges	7763
Type	Directed
Average Degree	5,459
Diameter	11
Density	0.004
Avg. Path Length	3.229

paper is citing the second one.
Each nodes has seven attributes:

- **Title**;
- **Publication year**;
- **Authors**;
- **Source Paper**;
- **DOI**;
- **Citation Score**;

In order to visualize the entire network, we used Gephi and we choose the Fruchterman Reingold layout which works well for large network like this one.
In the image 2, color and size are based on the in-degree centrality, in this way we can visualize better, the most cited paper, which is the blue bigger circle.
In particular, the title of the most cited paper is **Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine**, written in 2020 from the New England journal of medicine.
This paper has a citation score of 283, this means that 283 different articles cited this paper.
In the table II we listed all the papers that have the highest citation score, so they are the articles which are more cited.
As we can see, several papers regards the safety and the efficacy of the Covid-19 Vaccine. Instead other papers does not contain in the title the words 'Covid-19' and 'Vaccine', but in a way they are still related to this topic.

We computed several metrics with the help of Gephi. In the table I you can find the number of nodes and edges, the average degree, diameter, density and the average path length.

In the figure 3 there is a visualization of the different communities founded by using Gephi and the modularity. In particular, we have cluster with:

1) 244 elements (16,39%);
2) 204 elements (14,35%);
3) 187 elements (13,15%);
4) 157 elements (11,04%);
5) 150 elements (10,55%);
6) 110 elements (7,74%);
7) 85 elements (5,98%);
8) 57 elements (4,01%);
9) other small communities.

By analyzing the different communities, it turned out that different cluster mainly correspond to different topic regarding Covid-19 and Covid-19 Vaccine.
For example, the largest cluster contains papers that concern the safety of the COVID-19 vaccine like *"Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine"*, *"Safety and immunogenicity of two rna-based covid-19 vaccine candidates"* or *"Evaluation of the bnt162b2 covid-19 vaccine in children 5 to 11 years of age"*.
Another community instead contains a lot of articles about the development of the vaccine like *"Sars-cov-2 vaccines in development"*, *"Development of an inactivated vaccine candidate, bbibp-corv, with potent protection against sars-cov-2"* or *"Coronavirus vaccine development: from sars and mers to covid-19"*.
In one more community instead, you can find paper that are more specific about the sociological vision of the covid-19 vaccine instead about the scientific part of it. For example *"A global survey of potential acceptance of a Covid-19 vaccine"*, *"Acceptability of a covid-19 vaccine among adults in the united states: how many people would get vaccinated?"* or *"Vaccine hesitancy: the next challenge in the fight against covid-19"*

VII. CONCLUSION

In this paper we build up and analyzed the citation network for Covid-19 Vaccination.
It was an important step to build up human immunity to Covid-19 and save thousands or millions of people across the globe.

Such citation networks would enable researchers to filter out the most cited papers regarding their areas of interests and speed up the process.
With Covid-19 spreading and affecting people across the world, a focused and faster access to works within one's interest, would be great as well as collaborating with other groups in order to build up on work for such vaccination development.

In the table II you can see the most cited paper in this network. The top three papers are from the New England Journal of medicine.
The first two papers talk about the safety and efficacy of the vaccine, the third one instead, is a paper which concern the developing of the vaccine.
This also might indicate the earlier development and trials of Covid-19 vaccine roll-out in United Kingdom. From the table I the reader can get the metrics of the citation network. The results showed a varied citation score from the avg of 5.459 to 283 of the highest cited paper which might be because of the timing of the publication and world-wide immediate response needed. similarly as seen from the figure 3 3 we found out that there were 8 clusters of papers which could infer the less differentiated areas of research across the globe.

So this is way not only we provide with a simple approach to building citation network. But also to use a more reliant data set with Web of Science and CORD_19, and a methodology that is quite able to use algorithms involved in graph theory for further analysis.

Both CitNetExplorer and Gephi are open source and are feature rich and similarly with basic data processing one could be able to generate both the nodes and edges file.

In future we would like to try to add some more citation using social media platform and do analysis and that would be using the product's API for data collection. Also *VOSviewer* [9] is another tool which can used for building and analyzing citation networks and has capability of text mining the papers.

VIII. GitHub

In this section you can find the link to the code that we used.
At the following link there is the code available for this project. GitHub repository

We divided the code in different folder. In particular we want point out that the code, available in the folder **Cord19** needs the metadata and the parsed file of the dataset CORD_19, which is available here. We didn't put these files directly into our repository because they the size of them are too big.

As regards Web Of Science, the university of Tartu has free access to it by using the UT username and password. In the UT Website you can find database that are available for the UT students.
Instead at the following one you can find the direct dataset where you can perform search. Web of Science

References

[1] Villa-Collar C, Alvarez-Peregrina C, Sanchez-Tena MA, *Myopia Control: A Citation Network Study.* Med Hypothesis Discov Innov Ophthalmol. 2020; 9(3): 208-214.

[2] Sven Groppe, Lina Hartung, *Automatically Generating Citation Graphs (and Variants) for Systematic Reviews*, 2021.

[3] Bruno Miranda Henrique, Vinicius Amorim Sobreiro,Herbert Kimura, *Building direct citation networks*

[4] Mohsin Naqvi, Ahmad Usman Mailk, Sohail Razzaq, Muhammad Tanvir Afzal, *Extraction and Visualization of Citation Network*

[5] Clara Martinez-Perez, Cristina Alvarez-Peregrina, Cesar Villa-Collar, Miguel Ángel Sánchez-Tena, *Current State and Future Trends: A Citation Network Analysis of the Academic Performance Field*

[6] Wang, Lucy Lu and Lo, Kyle and Chandrasekhar, Yoganand and Reas, Russell and Yang, Jiangjiang and Burdick, Doug and Eide, Darrin and Funk, Kathryn and Katsis, Yannis and Kinney, Rodney Michael and Li, Yunyao and Liu, Ziyang and Merrill, William and Mooney, Paul and Murdick, Dewey A. and Rishi, Devvret and Sheehan, Jerry and Shen, Zhihong and Stilson, Brandon and Wade, Alex D. and Wang, Kuansan and Wang, Nancy Xin Ru and Wilhelm, Christopher and Xie, Boya and Raymond, Douglas M. and Weld, Daniel S. and Etzioni, Oren and Kohlmeier, Sebastian, *CORD-19: The COVID-19 Open Research Dataset*, 2020, https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1

[7] Van Eck, N.J., Waltman, L. *CitNetExplorer: A new software tool for analyzing and visualizing citation networks. Journal of Informetrics, 8(4), 802-823*, 2014. Official Website

[8] Miray Kas,*Structures and Statistics of Citation Networks*, 2011.

[9] Web of Science - Official Website

[10] Gephi Official Website

[11] Networkx Official Website

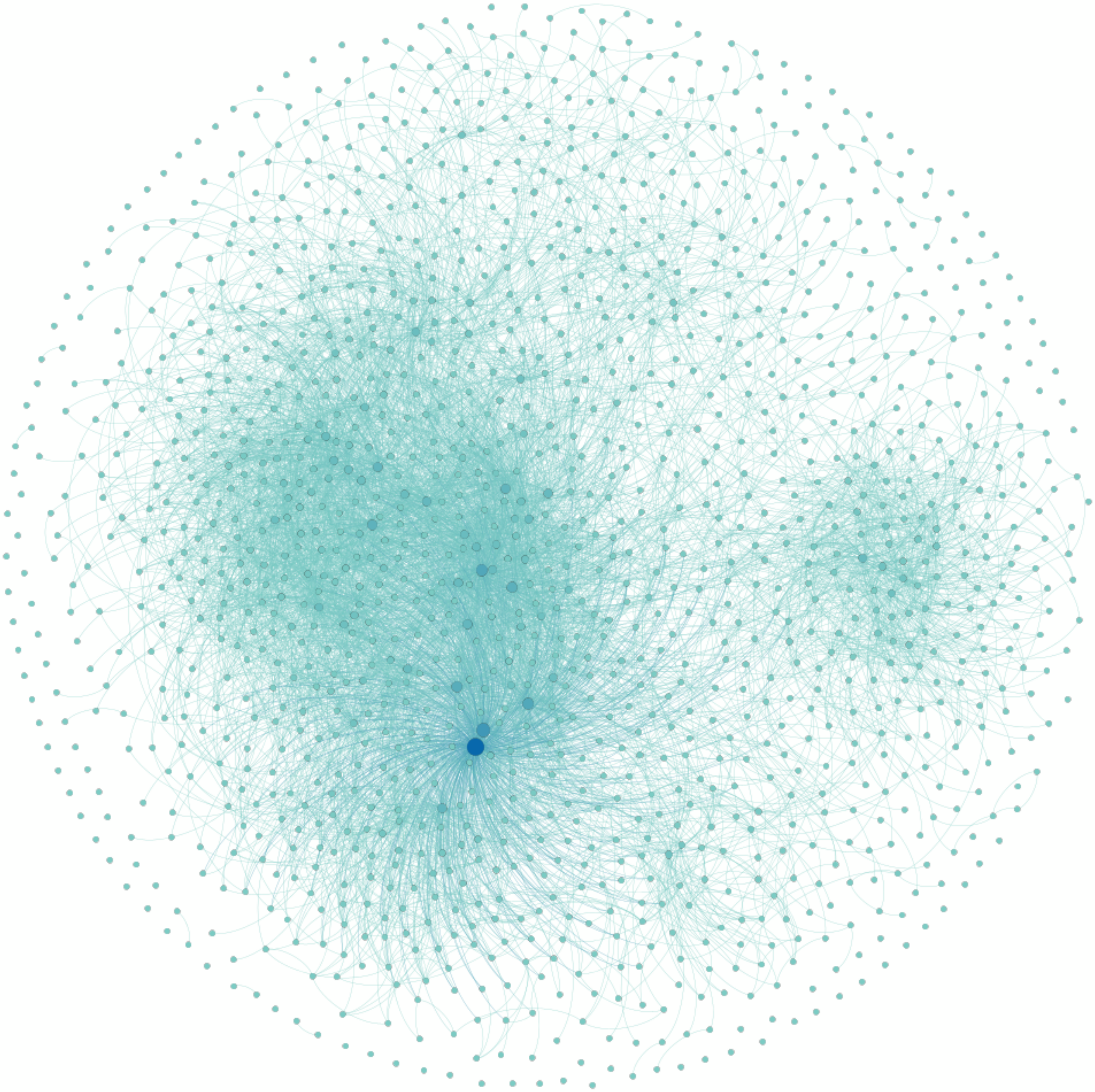

Fig. 2. Visualization of the citation network with Gephi. Color and size of nodes are based on the in-degree centrality.

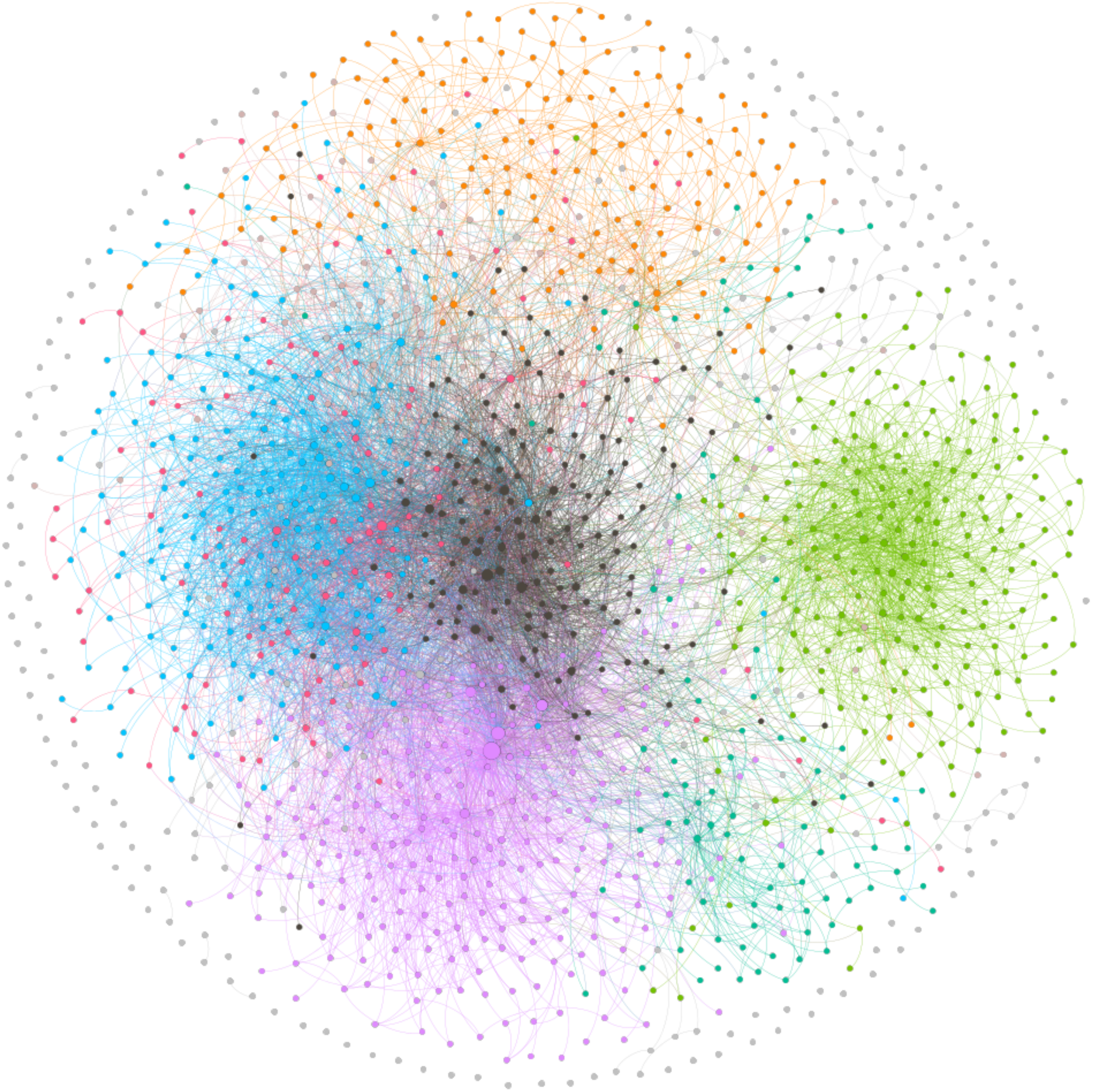

Fig. 3. Citation Network with communities

TABLE II

PAPERS WITH THE HIGHEST CITATION SCORE

Title	***Authors***	***Source***	***Year***	***Cit_Score***
Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine	Fernando P. Polack, ...	New England journal of medicine	2020	283
Efficacy and safety of the mrna-1273 sars-cov-2 vaccine	Lindsey R. Baden, ...	New England journal of medicine	2021	189
An mRNA Vaccine against SARS-CoV-2 — Preliminary Report	Lisa A. Jackson ...	New England journal of medicine	2020	138
Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK	Merryn Voysey DPhil	Lancet	2021	134
Safety and Immunogenicity of Two RNA-Based Covid-19 Vaccine Candidates	Edward E. Walsh, ...	New England journal of medicine	2020	116
Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial	Pedro M Folegatti ...	Lancet	2020	107
Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals	Alba Grifoni, ...	Cell	2020	100
Development of an inactivated vaccine candidate for SARS-CoV-2	Qiang Gao, ...	Science	2020	87
Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus	Better Korber, ...	Cell	2020	84
BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting	Noa Dagan, ...	New England journal of medicine	2021	83