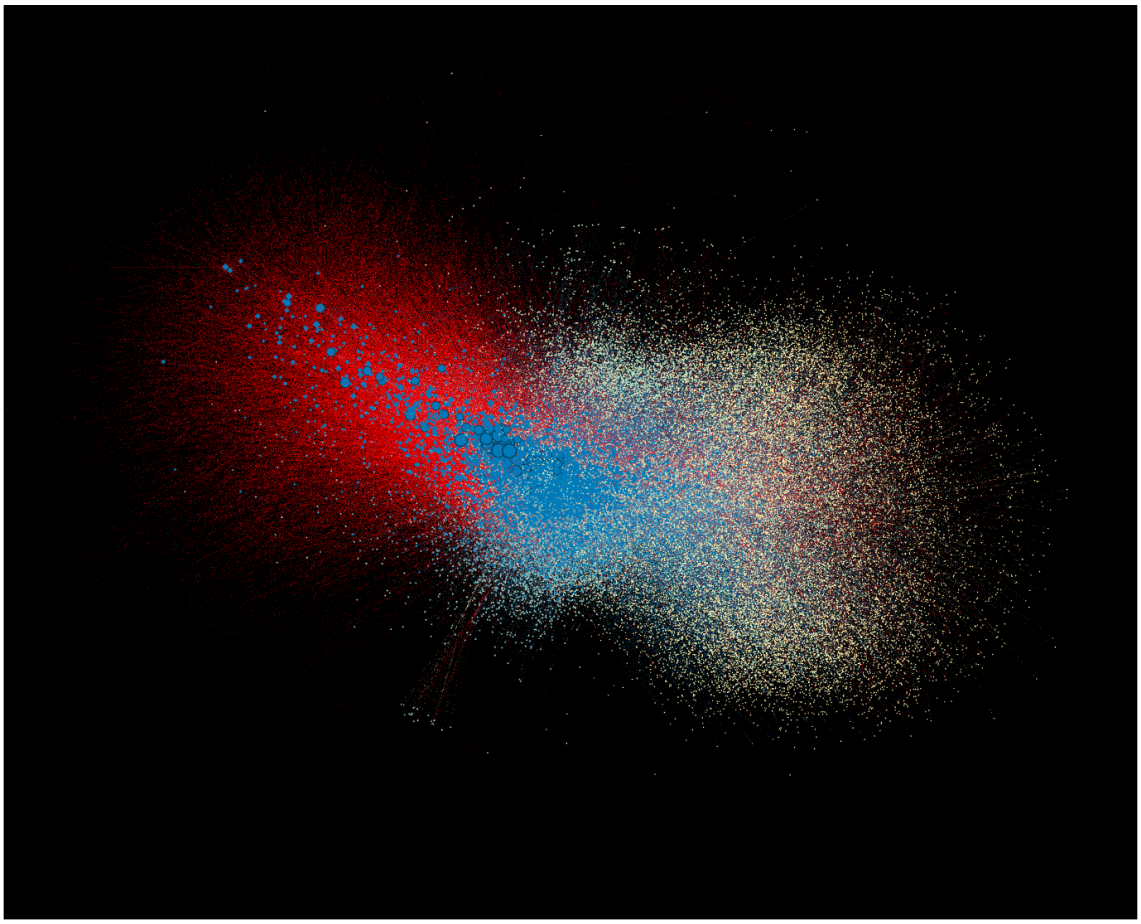

Visualization of the citational landscape in coronaviruses research

Identifying key papers within a citational landscape

The case of Coronavirus' research

Network Analysis - Project's final report
Course: Digital Humanities & Digital Knowledge

Authors:

- Lorenzo Paolini: lorenzo.paolini6@studio.unibo.it | 0001056503
- Tommaso Battisti: tommaso.battisti4@studio.unibo.it | 0001056557

12 gennaio 2023

1. Context

The ongoing COVID-19 pandemic has had a profound effect on the scientific community, with researchers from various disciplines collaborating to gain insights into the virus and develop effective treatments and vaccines. In general, Coronaviruses have been significant research topics for many years, and the current pandemic has highlighted the importance of cross-disciplinary collaboration in addressing global challenges. For many years, experts from various disciplines have investigated the characteristics and the effects of these viruses, resulting in a wide and manifold array of citations in the scientific literature.

In this study, we analyze the citational context of papers related with Coronaviruses, in order to illustrate a particular workflow we devised to analyze *context-specific citational landscapes*. To do that, we relied on the field of *Bibliometrics*, a sub-field of *Library and Information Science* that focuses on the use of quantitative methods to study the production, dissemination, and reception of research publications. In particular, this work uses methodologies from *Graph Theory*, *Network Analysis*, and *Citation Analysis* to retrieve and investigate citational patterns.
The corpus we focused on includes articles about Coronaviruses and their citational panorama, covering a time span that goes from the year 1904 to the 2020. By mapping the connections between them, we hope to gain a better understanding of the evolution of Coronaviruses' research over a long period of time, including the recent emergence of COVID-19.

To this aim, we built a network of citations consisting of nodes representing individual papers, and edges representing citations between them. Moreover, we also investigated the network mapping all the journals containing such articles, from which we gained insights that led us to deepen the analysis. In particular, the network of papers was subjected to a process of "*citational re-weight*", where the strength of the connections between papers get also determined by the relative importance of the journals in which they are published, with respect to the analyzed context. This weighted approach can determine a better understanding of the influence of different research papers and journals, identifying the ones that have been instrumental in shaping the direction of Coronavirus-related research. Overall, the aim is to provide a comprehensive picture of the citational context of Coronaviruses' research by developing a reliable measure for the study of citations within a specific context of inquiry.

In the following sections, we will first discuss the motivations that led us to devise this approach. Then, we will describe our dataset and all the techniques we used to model data and to develop our measures, presenting step by step our specific workflow. We will finally present the results of our analysis and discuss their implications with respect to the structure of the networks under analysis.

2. Problem and Motivation

In studying the development of Coronaviruses' related research, analyzing the papers' citational context can provide valuable insights about the ways in which different research studies are connected and their interdependence. Identifying the most influentials papers and being aware of how they have shaped the direction of research within a specific context could possibly help researchers to understand the current state of knowledge, as well as to identify potential areas for future research.
In such context, it is valuable to have a measure for papers' importance that goes beyond the simple in-degree count, even though it is, as today, still considered a reference. The reason is mainly rooted in the so called *Matthew effect*, pointed out by Merton [1968], and also by de Solla Price [1965] under the name of *cumulative advantage*. These two names are both referring to the same concept: popular and famous articles, which have yet received many citations, will tend to increase the number of citations received during the years. This is a possible source of problems, in particular when dealing with extended periods of time. The solution we devised in order to overcome this problem takes into account different kind of centrality scores. Additionally, we decided to mitigate the connections between papers by means of the relative importance of the journals in which they have been published. From this process, we aim to extract a more realistic view of the importance that the papers under analysis had for the context of inquiry.

The urgency to develop a new way to investigate citational data comes from the amount of money spent each year by governments and institutions to asses the quality of research by means of citation counts. However, according to Martin and Irvine [1983], "*indicators based on citation counts are seen as reflecting the utility, rather than the quality or importance, of the research work*", leading people to think whether all those money are objectively well spent. We completely rely on this perspective, feeling the need to produce a meaningful contribution and hoping to capture the deep patterns arising from context based networks, by taking into account both *quality* and *utility* in order to better analyze citational data. For what concerns *utility*, we relied on citation counts[1] as a starting point. Additionally, we also tried to assess the *quality* of these articles that, as stated by Nieminen et al. [2006], "*is not necessarily associated with the number of citations [received by a paper]*". In our opinion, the *quality* of research could be better assessed by means of *centrality* measures, thanks to which we have been able to analyze network patterns and derive meaningful conclusions.

In conclusion, the problem we want to address, otherwise our research questions, could be defined as follows: *is it possible to define a reliable workflow that determines which articles are the most significant within a circumscribed context, by considering a set of different measures?* More specifically, *how different would be the result if we go beyond the simple in-degree and consider additional characteristics in computing the importance of each node? And what kind of conclusions could we draw by encoding in the process also the relative importance of the journals in which these articles have been published?* And finally, *how well does our algorithm correlates with the reference in-degree measure?*

[1] Considered a reference measure in bibliometrics analyses, as stated by Pride [2020].

3. Dataset

Data have been gathered from *The Coronavirus Open Citations Dataset*, which is an open dataset under the administration of *OpenCitations*, containing information about 189697 citations and 49719 citing or cited articles. The version we have used contains citations between papers published from 1904 to 2020, it can be found on *Zenodo*, stored in JSON format, under a Creative Commons CC0 waiver (see Peroni [2020]).
However, as declared on the website[2], the dataset is still missing many relevant citations coming both from articles whose journal's publishers are not participating in the *Initiative for Open Citations* (*I4OC*) and opening their reference lists at *Crossref*, and from preprints, which also did not deposit their reference lists in *Crossref*. This could be a possible source of problems, but we will deal with it in the following sections.

To build the relevant networks, we have used the following files:

- metadata.json - contains the metadata of all the collected articles (*Article Id*, *Author*, *Publication Year*, *Article Title*, and *Journal Title*);
- citations.json - contains all the information about citing and cited entities (*Citation Id*, *Citation Source Id*, and *Citation Target Id*).

We devised a function to analyze these articles titles in order to retrieve the research area to which they belong to. In order to do that, we used the Academic Vocabulary corpus[3], collected by Mark Davies and Dee Gardener, who analyzed academic texts coming from the Corpus of Contemporary American English (COCA).

Additionally, by crossing data, we built two different networks relying on the Python Package *NetworkX*. Overall, *Python* has been used as the main language for the analysis of the dataset, and the computation of the results. We used *Math, Pandas* and *NLTK* Python libraries for data analysis purposes. *Numpy*, *Random* and *Statistics* Python packages have been used to modify and re-adapt *NetworkX* functions to our aims. Furthermore, we used *Seaborn* to produce some visualizations inside the project's *Jupyter Notebooks*, and *JobLib* to speed up the different operations. Finally, the operations performed in Python have been re-computed inside *Gephi*. This last software has also been used to produce network visualizations. The final outputs, as well as the source code of the project, are publicly available on the dedicated *Github* repository[4].

4. Validity and Reliability

Our objective is to test and evaluate the effectiveness of the workflow we have developed. The dataset we decided to use includes citations from various academic fields (detailed in **Table 1**), giving us the possibility to gain a deeper understanding of the research landscape.

---

2 https://opencitations.github.io/coronavirus/

3 https://www.academicvocabulary.info/x.asp

4 https://github.com/Postitisnt/COVID-19-Citations-Network-Analysis.git

Articles in category	
Medicine	42355
Science	3040
History	37
Law	26
Sociology	52
Humanities	244
Finance	17
Religion	22
Education	1250
Unclassified	2676

Table 1. Overall count of articles for category

By applying the analysis mentioned in the section above we discovered that, while most of our data belong to the medical domain, the other fields are represented by a much smaller number of articles. We believe that this distribution is a quite representative and valid model for analyzing the research's landscape of Coronaviruses. Indeed, it is likely that the vast majority of articles related to this topic would come from the medical field, in particular if we consider the time-span under analysis. While it is unlikely that between 1904 and 2020, academic fields such as religion, education, sociology or finance, dealt with *Coronaviruses*[5].

As pointed out in the previous section, the lack of articles from which suffers the dataset could be critical for the final outcome. In fact, even though the main objective of this work is to illustrate our newly devised workflow, it is unlikely that it will produce reliable results for what concerns the ranking of key-papers in absolute terms[6]. Nonetheless, there will be always room for improving this project, as well as the possibility to integrate new data in the model[7], and the final workflow could be reapplied to an extended dataset, as well as to different kind of networks besides those related with *Coronaviruses*. Thus, the reliability of this project does not merely depend on the context of application (intended as the isolated topic of research), but on the objectivity and the effectiveness of the used measures and workflow. This holds as long as they are applied to a circumscribed context of inquiry and to objective data, which are the basic conditions that make this work reliable and consistent.

5. Measures

5.1. Introduction to the workflow

The workflow we developed is partially rooted in the work by Diallo S.Y. et al. [2016], in which the authors compare different centrality measures to define which one performs better in identifying key papers within a particular journal. The research demonstrated that Eigenvector centrality best correlates with citations counts (at least among the measures taken in consideration), providing information that slightly rearranged the final output. Starting from this assumption, we added some complexity layers and enlarged our view on the overall structure of the networks, in order to build an abstraction useful to retrieve information from different sources of information.

[5] However, it is possible for a dataset accounting for more recent publications to show a more uniform distribution. This is likely due to the increased importance attributed to COVID-19 in recent years, whose effects generated interest from a wider range of disciplines. Nonetheless, we expect that the medical domain will still present a higher number of articles overall.

[6] It is indeed possible that the *real* most important paper is not included in the dataset due to the missing articles.

[7] We would also like to remark that the context of Coronavirus' research is just instrumental to apply our workflow.

Additionally, we conducted a structural analysis of the networks to draw meaningful conclusions related to their shapes.

Finally, to compare the outcomes of our workflow with the ones produced by more standard measures (PageRank, Eigenvector centrality, and In-Degree count), we have scaled all the values in a range [-1,1] in order to obtain a better grasp on our data before computing the *Kendall's Tau-b Correlation coefficient*. The reason behind the choice of *Kendall's Tau-b*, in addition to being non-parametric, lies on the fact that we expect many concordant pairs in our network[8], in particular from nodes with low in-degree, which often result in ties with eigenvector scores or with the outcome produced by our algorithm[9]. These ties would have led to a slightly biased final output if we had decided to use *tau-a*. Additionally, since the inputs for this operation are all distributed along the same scale ($[-1,1]$), *tau-c*'s complexity would have being unnecessary.

5.2. Network analyses - Workflow

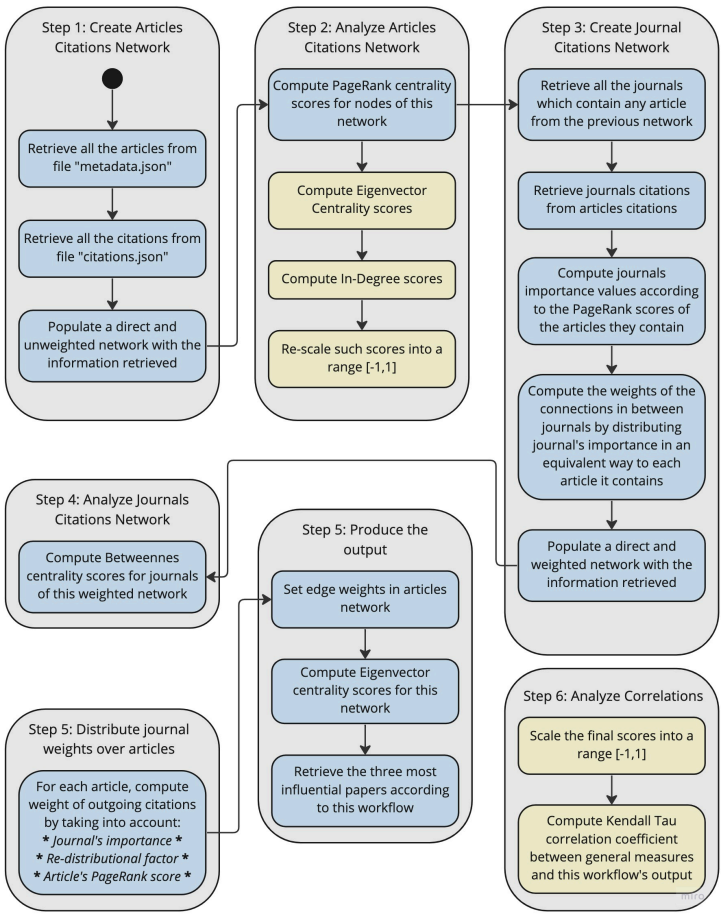

Figure 1. Workflow's overview. Blue blocks represent structural elements; Yellow blocks represent additional checks to compute final results and verify our model's consistency.

[8] It has been checked thanks to the scaling operation.

[9] In particular, this happen with our algorithm only when such nodes have also been published in a low-ranked journal.

As a first step, in order to extract useful information from citation networks, we adopted a slightly different version of the Eigenvector centrality: the *PageRank*, whose effectiveness has been firstly pointed out by Page et al. [2001]. At this step, it seemed somehow necessary to take into account the *in-degree* value of network nodes by means of this algorithm[10]. Thus, we built the directed and unweighted papers' network to which we applied the PageRank algorithm to retrieve nodes' centrality scores. The main reason for the adoption of PageRank, instead of Eigenvector centrality, consists in the need to provide papers without incoming citations with an *a priori* weight, so to avoid a centrality score propagation of 0. The scores obtained have been subsequently used to extract a measure for journals importance. To this aim, the importance of journal A is given as follows:

$$\Phi_A = \sum_{i=0}^{n_A} \tau_i$$

With Φ_A representing the importance of journal A, which contains n_A articles, each with a PageRank centrality value τ.

Once computed the Φ-value for the journals, we built the weighted and directed journals' citation network. The weight of the edge $A \rightarrow B$ (ω_{AB}) has been computed as the reciprocal of the relative importance of the citing journal A in the context (Φ_A), times the number of citations going from journal A to journal B (n_{AB}):

$$\omega_{AB} = \frac{1}{\Phi_A * n_{AB}}$$

Since, in our opinion, the relevance of a journal in a context can be meaningfully given by the junction position it assumes within the network, once all the weight have been obtained and applied to the edges, the journals' Betweenness Centrality has been computed. The reason behind the computation of the reciprocal in the previous step lies in the fact that the Betweenness **C**entrality considers edges weights as distances, leading to a situation in which a higher weight means higher distance and thus lower importance, unless the reciprocal value is computed.

Finally, we built a new network of citations for the papers but, this time, we gave to each citation (edge) a value that is the weight (relative importance) of the journal in which each article is contained. We represent with ω_{a_γ} the strength of each connection γ pointing from $a \in A$ to each target article x ($a \rightarrow x$), while τ_a is the PageRank value of paper a. We also took into account a *re-distributional factor* by counting the overall number of out-going citations from a, represented as c_a.

[10] To recall what we said above, the aim is to extract some key articles from a network by taking a perspective centered on both the *quality* and the *utility* of such works in the panorama under examination.

The resulting weight of the edges has been computed as:

$$\omega_{a_\gamma} = \frac{\Phi_A}{c_a} \cdot \tau_a$$

Finally, the most influential papers have been extracted using the Eigenvector centrality applied to the new weighted graph. We did not use the PageRank since we overcame the problem of nodes without incoming citations in the previous steps.

5.3. Networks' structure[11]

The structures of the networks have been analyzed under three different points of view. We started by computing the *Small World Coefficient* ω, in order to determine the presence of small world properties in our networks[12]. We decided to compute this coefficient, instead of the σ coefficient, for two main reasons:

- As stated by Watts and Strogatz [1998], small-world networks are "highly clustered, like regular lattices, yet have small characteristic path lengths, like random graphs". The σ coefficient, computed by comparing the network's clustering and path length to those of a comparable random network only, is less reliable than ω[13], which can "characterize networks as small-world, random, or lattice, or at least tendencies toward one of these types" (Telesford et al. [2011]).
- The second reason is mainly utilitarian, the time complexity of the algorithms is quite high. Therefore, we chose the coefficient whose computation was more amenable to parallelization, to obtain the output in a more efficient manner.

The ω coefficient is defined as:

$$\omega = \frac{L_r}{L} - \frac{C}{C_l}$$

With C and L representing, respectively, the *average clustering coefficient* and the *average shortest path length* of the network we are studying; L_r representing the *average shortest path length* of an equivalent random graph, and C_l the *average clustering coefficient* of an equivalent lattice graph. The steps involved in obtaining the final coefficient were based on the building of a lattice graph, and of some random graphs to average their clustering coefficient.

---

[11] Note that the measures used to analyze the overall structure of the networks are less likely to be generalizable to networks that are not made up of citations, for which other analyses could be more suitable.

[12] In a small-world network we expect to find *shorter-than-expected* distances between couples of nodes: most nodes are not neighbors of one another, but the neighbors of any node are likely to be neighbors of each other, thus most nodes can be reached from every other node by a small number of steps.

[13] With reliable here we mean the potentiality of the coefficient to capture small-worldness with respect to the definition given by Watts and Strogatz. Being the σ coefficient computed with respect to a random graph, without considering the clustering with respect to a lattice graph, it can define as "small-word" networks with very low clustering, which is indeed what we do not want.

Once obtained the ω coefficient, we moved to the analysis of the possible modular configurations of the networks, which involved the use of the *NetworkX* library, which provides a module to compute the *Modularity Coefficient Q* as follows:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{d_i d_j}{2m}) \delta_{g_i g_j}$$

Where m is the number of edges in the network, A is the adjacency matrix of the network, d_i and d_j are the degrees of nodes i and j, and $\delta_{g_i g_j}$ assumes the value of 1 if i and j belong to the same community and 0 otherwise.

Finally, we decided to compute the *Scaling Coefficient* α to understand whether the nodes' degree of our networks follow a *Power Law Distribution*. The α coefficient is obtained as follows:

$$\alpha = 1 + n(\sum_{i} ln \frac{d_i}{d_{min} - \frac{1}{2}})^{-1}$$

Where n is the number of nodes in our network, d_i is the degree of the node i, and d_{min} is the minimum degree value that can be found in the graph under analysis.

6. Results

The outcomes we obtained are presented in **Table 2**.

Network	***N***	***E***	***C***	***L***	***Q***	α	ω
Journals	5661	39100	0.33	2.99	0.0067	1.5435	0.2003
Papers	49719	150536	0.07	1.63	0.3620	1.7226	-0.9625

Table 2. Values obtained from the analyses. *N*, nodes; *E*, edges; *C*, average clustering coefficient; *L*, average shortest path length; *Q*, modularity coefficient; α, scaling coefficient; ω, small-world coefficient.

The network of journals, with 5661 nodes and 39100 edges, has an average clustering coefficient $C = 0.33$ and an average shortest path length $L = 2.99$. The papers' network has instead 49719 nodes and 150536 edges, and it presents an average clustering coefficient $C = 0.07$ and an average shortest path length $L = 1.63$. Furthermore, the modularity analyses shows that, while the journals' network presents a modularity coefficient $Q = 0.0067$, the network composed by papers presents a higher value $Q = 0.3620$. Additionally, their scaling coefficients are, respectively, $\alpha_p = 1.7226$ and $\alpha_j = 1.5435$. Finally, computing the small-world coefficient, we got $\omega_j = 0.2003$ for the journals' network, and $\omega_p = -0.9675$ for papers' network.

The five most influential papers retrieved by means of the main reference measures and the final outcomes of our workflow are detailed in **Table 3**. Between the first two measures, there are no significant differences other than the order of the papers. Beyond this, the *a priori* Eigenvector Centrality computed on the unweighted network carried to the foreground a majority of papers that have never been included in the first five key papers before.

Measure	*Ranking by internal Id*	*Value*	*Journal*
In-degree	39264	1.0	New England Journal Of Medicine
	25837	.9766	New England Journal Of Medicine
	17440	.9201	New England Journal Of Medicine
	21306	.5478	The Lancet
	12204	.4152	Science
PageRank	39264	1.0	New England Journal Of Medicine
	17440	.7403	New England Journal Of Medicine
	21306	.6090	The Lancet
	25837	.3598	New England Journal Of Medicine
	12204	.1904	Science
Eigenvector Centrality	34948	1.0	Journal Of Infectious Diseases
	26230	.6793	The Lancet
	3907	.5997	Veterinary Pathology
	11156	.5281	Experimental And Molecular Pathology
	23638	.4668	American Journal Of Epidemiology
Workflow's final measure	25837	1.0	New England Journal Of Medicine
	13497	.7437	Plos Pathogens
	17440	.6276	New England Journal Of Medicine
	39264	.5734	New England Journal Of Medicine
	25492	.2358	Journal Of The American Geriatrics Society

Table 3. Key papers found with different measures used during the workflow. Note that values have been rescaled within a range between [-1, 1] and are approximated.

For what concerns the validation of our outcomes, we analyzed the correlation between an article's citation count and its centrality value computed by means of our workflow and by the other measures reported in **Table 3**, using citation count as a proxy for the significance of papers. The main assumption we decided to adopt here is based on the work of Diallo S.Y. et al. [2016], in which the authors tend to consider citation count as a "*good metric that indicates the historical value of a paper*", therefore we rely on the fact that it should correlate well with measures that tries to retrieve the same kind of historical significance of a paper.

Once all the output values have been re-scaled within the range $[-1,1]$, we computed the Kendall's correlation coefficient τ_b. The results obtained show a strong positive correlation between the outcomes produced by our algorithm and citations count ($\tau_b = 0.7386$), as well as between the PageRank algorithm's output and the number of ingoing citations ($\tau_b = 0.7095$). Instead, even thought the Eigenvector centrality measure still positively correlates with the in-degree count, the correlation coefficient's value is in this case lower ($\tau_b = 0.4997$) with respect to the other two.

7. Conclusion

Overall, while the average clustering coefficient of papers' network is lower than the network of journals' one, the opposite is true with respect to the modularity coefficient. Furthermore, as seems to be suggested by the higher clustering coefficient, the journals' network is the one that present a greater small-word effect, since the value of the ω is quite close to 0. This, implies that there is a strong connection between the journals within the network, thus suggesting a lot of collaboration and cross-fertilization of ideas between the research fields involved. Clearly, seen the dataset, this last hypothesis could be mainly referred to medical and scientific fields, while there are not sufficient data to suggest the same for the remaining fields.

In general, we can say that it is not unusual for a network of papers citations to exhibit a higher modularity than a journals' one. In fact, papers belonging to the same research area are more likely to form "communities", thus resulting in a higher modularity coefficient. A visualization of the main communities within the papers' network is provided in **Figure 2**.

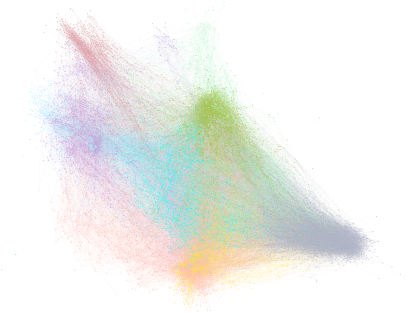

Figure 2. Visualization of Papers' network according to modularity classes extracted with Gephi. ForceAtlas2 layout - Modularity coefficient resolution: 1.0

The contrary holds for journals, since they normally aggregate papers of the same research area. In this sense, the communities could be seen as the journals themselves. However, the papers' value for Q is still closer to 0 than to 1. An intuitive explanation may be that the network covers a wide time-span, resulting in papers that, while

belonging to the same research areas, hold fewer connections between one another due to the characteristics of the research panorama and the evolution of the field over time[14].

A similar explanation may hold for the outcomes of the ω coefficient. The network of journals presents the best balance between high clustering and low path length, meaning that it is the one that most resemble small-world characteristics. On the contrary, the network of papers has a value very close to -1, showing a strong tendency towards a lattice graph. This is confirmed by the fact that the network of papers has a very low average clustering coefficient. Indeed, it is very unlikely for such a sparse network to present dense connections between nodes leading to small-world characteristics.
As explained before, a possible reason for these two opposite values can be the temporal range of publication. In fact, while journals are less, if not at all, influenced by the temporality of the citations (meaning that clusters can be easily formed since they contain articles from different years), papers are. To prove that, it is enough to think that citations cannot point to "future-papers" since they do not exist yet. Thus, the typical triadic form assumed by clusters are difficult, if not even impossible, to be formed[15].

Focusing on the scaling coefficient α, we can conclude that the degree distribution of the papers' network is the one that best approximates a power-law. Since we can empirically talk about a power-law when $2 \leq \alpha \leq 3$, we can say that the distribution of our main network is only near to such distribution. What holds, however, and what makes the value so close to the range, is the fact that only few nodes have a very high degree, as shown in **Figure 3**.

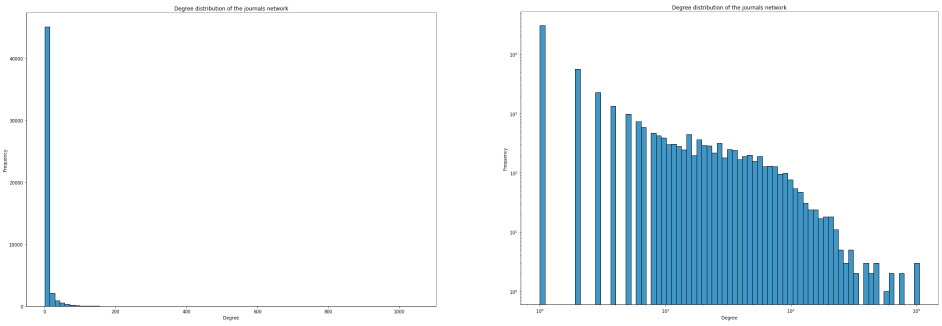

Figure 3. Degree distribution of papers' network. On the left, normal plot; on the right, log-log plot.

Such long tail distribution seems to suggest, seen that the academic areas are mainly distributed along the medical and scientific fields[16], that the papers within the network

[14] Moreover, the overall size of the network could be an extremely important factor. In fact, it may become more difficult for papers to establish connections with one another, as there are more papers to potentially reference. This could lead to an overall sparser structure.

[15] For example, if a paper from 2019 cites a paper published in 2001, which in turn cited a paper from 1999, it is impossible that the last paper has as direct neighbor the first one, if not because it gets cited by it too.

[16] Indeed, another possible interpretation of these phenomena could have been the one that supports the hypothesis that we are dealing with a broad panorama in terms of research fields. We can exclude this hypothesis thanks to the analysis performed on our dataset.

are highly diverse in terms of the topics they cover. This hypothesis could lead us to the idea that there are many subfields - or subtopics - within the main academic fields, each of which has being investigated by a subset of the papers in the network.

The differences in the computation of the five key-papers on the basis of the different measures, show their individual contribution to the final importance value. As an example, the three papers from the *New England Journal Of Medicine* presented in **Table 3** are the most relevant in terms of in-degree and PageRank. However, they do not appear in the key papers computed with the Eigenvector centrality. Nonetheless, the final outcomes of our workflow include them in a different order even though they are precisely computed by the Eigenvector centrality, but this time on the weighted network. Therefore, the differences in the two eigenvector-based measurements have been caused by the mediation of the journals importance values, caused, in turn, by the PageRank centrality scores of the papers, which ultimately depend on the in-degree. The importance of a paper in a specific context is thus the result of multiple components capturing some specific features of the articles' publication and citational context. Our measure, does not completely revolutionize the key papers' list with respect to the in-degree measure, but re-balances it in a meaningful way. This last point could be a sign of the reliability of our workflow.

With respect to networks structures, the computation weighs the edges of a sparse and more modular - than clustered - network (papers' network) by means of the relative importance gained by the nodes of the journals' network, which presents opposite characteristics. The relation between the papers and their publishing "container" has been taken into account, and assumes particular relevance if positioned within a specific and circumscribed context. Additionally, the results of our algorithm highly correlates with citations' count, suggesting a good prediction of the historical significance of papers in this well defined context.

In conclusion, we extended the simple in-degree measure by means of additional steps, and abstracted the context to a higher level, resulting in a better understanding of the patterns delineating the reference network. We took in consideration different aspects and characteristics of the citational network, such as the relative importance of the journals publishing the analyzed papers, and we found a solution to extract key articles within a well defined context. Additionally, we could also state that our work extends the research of Diallo S.Y. et al. [2016] by considering a wider time-span, but also by adding a different measure that the authors did not consider in their research, the PageRank. This additional measure seems to correlate better with citations counts in a temporally extended context with respect to the eigenvector, thus suggesting a meaningful comparison between the Eigenvector centrality and the PageRank even in the more circumscribed context analyzed by the reference authors.

8. Critiques

By analyzing the results, with respect to how well our algorithm correlates with citation count, we can consider our work a good starting point to conduct further researches. In particular, we are aware of the incompleteness of the specific citational context we have investigated. Nonetheless, our aim, before determining which are the most relevant papers in a specific context, was to define an algorithm able do that. Meaning that our dataset yet represents a good testing field for our hypothesis, large enough to give us a grasp on how to handle big networks, but also well defined with respect to the research area to which such articles belong[17].

The temporal extension of the reference context can cause the rise of critiques related to the hypothesis of having articles much more cited than other due to temporal factors. However, the concept of *citations' half-life*, defined by Burton and Kepler [1960], can mitigate such claims. The authors state that papers citations can be metaphorically juxtaposed with radioactive materials, which have a *rate of obsolescence*. In fact, the greatest amount of incoming citations of an article are demonstrated to be received during the first five years after its publication[18]. Nonetheless, in dealing with citations, the imbalance between older and newer works is naturally found.

In conclusion, we can consider our work a good starting point in the direction of defining a new *Altmetric*[19]. We do not pretend to have solved the original problem of finding an alternative and better measure, also because of the lack of tests and case studies to which our new workflow has been applied. Nonetheless, the model we have defined seems to be pretty solid, and could also be improved by abstracting it at higher levels, resulting in the possibility to take into account additional aspects of the networks. One possibility to improve and deepen this analysis could take into account the *incremental ratio* of citations that a journal have received during years, in order to see its temporal trend and develop a better measure to analyze the relative importance of journals. Another possibility, could involve different measures trying to capture different aspects of these network, thus resulting in the definition of additional models that could be meaningfully compared between each other.

[17] To recall what was said above, the belonging of these articles is coherent with the context we are investigating, thus resulting in a good approximation of the entirety of academic literature related to Coronaviruses.

[18] This is just an average, obviously it depends on the specific research area to which the paper belongs.

[19] Namely, *alternative metrics*, a group of measures in bibliometrics' research.

BIBLIOGRAPHY

- Burton, Robert E. and R.W. Kebler. 'The "half-life" of some scientific and technical literatures'. American Documentation, 11(1):18–22, 1960.
- Diallo, Saikou Y., Christopher J. Lynch, Ross Gore, and Jose J. Padilla. 2016. 'Identifying Key Papers within a Journal via Network Centrality Measures'. *Scientometrics* 107 (3): 1005–20. https://doi.org/10.1007/s11192-016-1891-8.
- Martin, Ben R., and John Irvine. 1983. 'Assessing Basic Research'. *Research Policy* 12 (2): 61–90. https://doi.org/10.1016/0048-7333(83)90005-7.
- Merton, Robert K. 1968. 'The Matthew Effect in Science: The Reward and Communication Systems of Science Are Considered.' *Science* 159 (3810): 56–63. https://doi.org/10.1126/science.159.3810.56.
- Neylon, Cameron, and Shirley Wu. 2009. 'Article-Level Metrics and the Evolution of Scientific Impact'. *PLoS Biology* 7 (11): e1000242. https://doi.org/10.1371/journal.pbio.1000242.
- Nieminen, Pentti, James Carpenter, Gerta Rucker, and Martin Schumacher. 2006. 'The Relationship between Quality of Research and Citation Frequency'. *BMC Medical Research Methodology* 6 (1): 42. https://doi.org/10.1186/1471-2288-6-42.
- Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry (1999) The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab. (http://ilpubs.stanford.edu:8090/422/)
- Peroni, SIlvio. 2020. 'Coronavirus Open Citations Dataset'. Zenodo. https://doi.org/10.5281/ZENODO.3756802.
- Price, Derek J. de Solla. 1965. 'Networks of Scientific Papers: The Pattern of Bibliographic References Indicates the Nature of the Scientific Research Front.' *Science* 149 (3683): 510–15. https://doi.org/10.1126/science.149.3683.510.
- Pride, David. 2022. 'Identifying and Capturing the Semantic Aspects of Citations'. https://doi.org/10.21954/OU.RO.000146FF.
- Telesford, Qawi K., Karen E. Joyce, Satoru Hayasaka, Jonathan H. Burdette, and Paul J. Laurienti. 2011. 'The Ubiquity of Small-World Networks'. *Brain Connectivity* 1 (5): 367–75. https://doi.org/10.1089/brain.2011.0038.
- Watts, Duncan J., and Steven H. Strogatz. 1998. 'Collective Dynamics of "Small-World" Networks'. *Nature* 393 (6684): 440–42. https://doi.org/10.1038/30918.

SITOGRAPHY

- Academic Vocabulary Corpus (https://www.academicvocabulary.info/x.asp)
- The Coronavirus Open Citations Dataset (https://opencitations.github.io/coronavirus/)