Content Aware Analysis of Scholarly Networks: A Case Study on CORD19 Dataset

Mehmet Emre Akbulut[*]
Politecnico di Milano
mehmetemre.akbulut@mail.polimi.it

Yusuf Erdem Nacar[*]
Technical University of Munich
yusuf.nacar@boun.edu.tr

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

1 Introduction

The core elements of scientific research include articles, researchers, and institutions. Since scientific research is the cumulative effort of researchers to increase the understanding of the world around us, the relationships between these elements are as important as the scientific results themselves.

Gaining insight into the relationship between the core elements of scientific research can be useful for a variety of purposes, such as guiding scientific effort toward better use of resources, inferring comparative results between fields of research, better representing the importance of certain research fields and research groups.

Current scientific literature continues to grow at a rapid pace every day. Let alone being able to follow the growth of communities in which we are not a part, it has become very difficult to even find conferences, journals or other prominent studies in our field. Naturally, examining the academic community in detail becomes a great burden for most young or experienced researchers, which results in missing out promising researchers and useful works. To overcome this problem, most researchers represent the scientific literature as a wide network consists of different entities such as researchers, institutes, etc. In this paper, we aimed to analyze current literature and demonstrate different approaches to this problem with some practical applications.

The academic writers, their studies and the citation connection between them composes the scientific community, which forms a wide network of **authors** and **articles**. Authors are identified as the entities that creates the knowledge in the community through the articles they have published. The citation network which is derived from the published work is the most common representation of this knowledge, which is very simple yet effective to analyze the communities. **Social Network Analysis** is a way of measuring and mapping various aspects of relationships between different entities such as people, organizations and groups [10]. At first step, we started our analysis from simple representation of the network which is a graph of authors and articles. Then, we focus on the possible interpretations of the centrality metrics, PageRank and its variations in the real world scenarios.

Apart from the approach above, it is clear that the proposed citation network lacks of the semantic meaning of the published works. Also representation of the topics is missing in the constructed graph of authors and their articles. Even though some solutions based on Natural Language Processing are available in the literature, most of these works require to process and analyze the content of the published articles via their open access files or abstracts. So, we propose a new and robust method to represent and retrieve the data in scientific network by considering the topics as an entity in the research graph.

[*]Equal contribution.

The topics are derived from a pipeline based on Named Entity Recognition and Knowledge Base of the relevant graph. Naturally, we have to focus on a specific domain to use the knowledge base effectively. Eventually, we showed the applicability of our method on a chosen domain and dataset, which are COVID-19 and CORD19 Dataset.

Our motivation in this paper is to develop a way to provide the researchers with quantifiable information about the relationships between these elements so that it can be used for such purposes. This quantifiable information includes graph measures of individual elements of a graph as well as the graph measures of the whole graph. Briefly, we propose a pipeline to create, analyze and store the research network which consists of authors, articles, named entities and relationships between them.

2 Related Works

The idea of Social Network Analysis was firstly proposed in 1969 by Philip Mayer and Julia C. Mitchell in their work on Urban situation in Central African Towns. Over time, Social Network Analysis has been used to understand relationships between different groups, organizations, communities, and any other possible network [9]. Social Network analysis has a interdisciplinary nature as being in the intersection of sociology, mathematics, statistics and computer science. In addition to these, the rise of Big Data in recent years lead to analyze communities in more efficient way. For example, it has been a common approach to analyze connected social media data to detect misinformation or influential behaviours [1], [12].

Scientific community is another domain which can be analyzed to derive meaningful information because of its interconnected nature. One of the earliest work in the literature is the *Impact Factor* proposed by Garfield (1972) in order to calculate the effectiveness and impact of the journals [3]. He calculated the *ImpactFactor* with the formula:

$$\mathit{ImpactFactor}(j, i) = A/B$$

where A is the number of times articles published in journal j in years $i-1$ and $i-2$ were cited in indexed journals and B is the number of articles, reviews, proceedings or notes published in journal j years $i-1$ and $i-2$. Following his works many researcher aimed to rank articles, authors and journals based on their impact on the scientific community [2].

Considering the article ranking methods, plenty of earlier works are derived from PageRank algorithm [7]. Mostly this approach cause biased results due to the fact that the older papers ,which have naturally higher citations than the newer ones, are assigned with higher ranking. CiteRank was proposed to remove the bias caused by PageRank through assigning more probability to new articles so that random surfer model can choose these articles [13]. Additionally, FutureRank [8] and P-Rank [16] algorithms were proposed in order to make use of different aspects such as time-indicator, authorship, journal information, etc. P-Rank hifted the focus to the understanding heterogeneous network representation of the entities [17]. The heterogeneous network proposed in the P-Rank algorithm consists of author, article and journal layers which propagates information among themselves to rank the specific article in the main network.

Apart from these works, HITS algorithm was proposed by Kleinberg. HITS algorithm uses authority and hub concepts to exploit local structure of the network [4]. The W-Rank then used the both PageRank and HITS algorithms in order to utilize link weights based on citation and authorship relationships [17].

The existing solutions in the literature generally ignore the importance of the different edges in the heterogeneous network. Some researchers use topics for academic search by using topic modelling and its integration into the random walk framework [11], however most of these methods lack of motivation to use topics in a weighting scheme to understand the nature of the community. Even though the time information is also used to evaluate link importance in citation relations in CiteRank algorithm [13], it suffers from the absence of semantic meaning and heterogeneity. Also we firstly present a way to asses the semantic meaning through approaching to the topics as not only a similarity measure but also an entity in the graph.

In this paper, we introduce the use of the MedCAT Concept Annotation Tool [5] and relevant knowledge base in the coronavirus domain. Based on the heterogeneous network built with academic entities, we have conducted bunch of experiments for the different link weighting schemes, whereas proposing a different approach for ground truth.

3 Methodology

Given the nature of any research field, the network of scientific knowledge and researchers is immensely complex. Therefore, to make sense of how aspects of these graphs relate to each other, one would not only need the quantifying information on the graph but also how this information changes as the graph itself evolves.

There are some assumptions before building the network we proposed:

- Old articles tend to have higher citation, which leads to biased results in ranking algorithms.
- The articles in the prestigious journals tend to have be higher influence on the network without their momentary citation count.
- The prestigious authors tend to publish articles with the bigger influence.

- Important articles are cited by other important articles. The meaning of all the citations are not the same.
- The semantically similar papers tend to cite each others and such a citation has more importance if their topics are dominant in the network.

Considering these assumptions, our network structure mainly based on the citation network which is composed of article nodes. The authors of and journal, if possible, of the article are connected to it with author and journal vertices. These attributes can be though as another layer of the network which is used to propagate information in prior works. The time information is also used to weight citation link between articles. Such network structure is very similar to PageRank + HITS approaches we mentioned above with various weigthing schemes. Differently, we add the topic layer, similar to author and journal layers, to the network by also exploring the influence of topics on citation weights. Eventually, a lightweight semantic network can be a part of the graph, which helps to analyze semantic relationships between articles in the network. We also believe that such as approach enables us to exploit topic based search and understand topic-wise prestige of articles in the network.

3.1 Heterogeneous Network

The heterogeneous approach has become a common approach when investigating the scientific communities. From the formal perspective a heterogeneous graph can be defined as:

$$G(V,E) = (V_{ar} \cup V_{au} \cup V_{ju}, \\ E_{ar-ar} \cup E_{ar-au} \cup E_{ar-ju}) \quad (1)$$

where V_{ar}, V_{au} and V_{ju} are the vertices of article, author and journal networks, whereas E_{ar-ar}, E_{ar-au} and E_{ar-ju} are the edges respectively. Based on this definition, we consider the topics as a part of the graph by adding vertices between articles and topics. Unlike the author and journal layer, the topics have a connection among themselves, which is analyzed later. So eventually, our heterogeneous network has the formula:

$$G(V,E) = (V_{ar} \cup V_{au} \cup V_{ju} \cup V_{tp}, \\ E_{ar-ar} \cup E_{ar-au} \cup E_{ar-ju} \cup E_{ar-tp} \cup E_{tp-tp}) \quad (2)$$

where V_{tp} represents the vertices of the topic network, whereas E_{ar-tp} stands for the edges from articles to topics. E_{tp-tp} is a term added for topic network which has a hierarchy derived from the thesaurus and ontology of biomedical concepts tanks to Unified Medical Language System (UMLS).

3.2 Link Weighting

Depending on the type of vertices, different weighting schemes can be employed. At this step, each relationship type is analyzed separately.

3.2.1 Article-Author

It is quite obvious that authors contribute differently to their published works. Although in some disciplines such as computer science, the ordering of the authors implies the importance and contribution, it is not a standardized approach in scientific community. Additionally, it can lead to underestimating the contribution of the authors, which is a serious issue in academic community. Some research includes $H - Index$ based solutions however this approach can assign lower ranks young researchers which have lower $H - Index$.

3.2.2 Article-Journal

The journals tends to publish similar quality articles in line with their own prestige, so we believe that we reach the articles published with equal probability starting from a specific journal. There are only two possibility in the article-journal networks, which are *publish* or *not publish* [17].

3.2.3 Article-Article

The citation network has more information than other sub-networks, which stems from the complex and versatile nature of the citation relationship. In addition to the graph based approach, many works explore the Natural Language Processing based techniques to understand the importance of the citation. However most of these works require huge amount of effort and data to train and classify the relevant models to understand whether a citation is influential or not. We use a hybrid approach by combining graph attributes and processing the abstract information to find similarities between articles. Using the abstract is more robust, efficient and fast than trying to find citation text in the document. Also it prevents us from being limited to articles that are only open access PDF. Briefly, the citation weight should have two different parts which are semantic-based similarity and network-based similarity [17]. Based on the work proposed by Zhang et al. we improve the network-based similarity by adding parameters related to authorship and journals in which published.

$$S(P_1,P_2) = \alpha \cdot \frac{|(In_{P_1} \bigcup Out_{P_1}) \bigcap (In_{P_2} \bigcup Out_{P_2})|}{\sqrt{|In_{P_1} \bigcup Out_{P_1}| \times |In_{P_2} \bigcup Out_{P_2}|}} \\ + \beta \cdot \frac{|A_{P_1} \bigcup A_{P_2}|}{\sqrt{|A_{P_1}| \times |A_{P_2}|}} \\ + \gamma \cdot J_{P1-P2} \quad (3)$$

where In_P and Out_P are incoming and outgoing links, whereas A_P is the authors of the the article P. J_{P1-P2} can be 1 or 0 depending on whether articles P_1 and P_2 were published in the same journal or not. The coefficients α, β and γ are 0.6, 0.3 and 0.1, respectively.

We analyzed the topic related weighting in next chapter.

3.3 Topic Linking and Semantic Weighting

Having a graph-structured representation of the research world allows the addition of explicit connections to other graph-structured knowledge representations. One prime example of such knowledge representations is ontologies. The possible connections between a selected ontology, Unified Medical Language System (UMLS), have been investigated. The connections between the papers and the UMLS concepts are constructed by passing the abstracts of the papers through a named entity recognizer called MedCAT.

In this paper, we propose a method that is based on Named Entity Recognition and Linking (NER+L) to extract the relevant concepts from the article abstracts based on MedCAT and Unified Medical Language System.

MedCAT [5] is an content annotation tool based on Word2Vec embeddings which can be used to extract information from medical documents to link them to medical ontologies such as UMLS [6].

3.4 Ranking Algorithm

Based on the HITS algorithm [4], we use a weighted iteration and updates of authorities and hubs in the network.

3.4.1 Hub Scores

The hub scores in the scientific network can be interpreted as the quality and impact of the hubs which the relevant entity belongs to. We analyzed the hub scores of articles, authors, journals and topics in this direction. We followed the slight derivation of HITS algorithm by normalizing the hub score with the number of the links [15], because of the risk that authors and journal which publish huge number of articles can dominates the hubs. In addition to work of Wang et al., we present the hub scores of the topics by considering the topics as a part of the heterogeneous network. Apart from these, to understand current state of the network we also followed the time-aware approaches similar to prior works [15]. Time-aware weights for article-author, article-journal and article-article links are needed to score the hubs.

The hub score of an author i:

$$H(A_i) = \frac{\sum_{P_j \in L_i} w_{ar-au}(i,j) \cdot A(P_j)}{|L_i|} \tag{4}$$

where L_i is the articles published by author i, $A(P_j)$ is the authority score of article j, $w_{ar-au}(i,j)$ is the time-aware weight between author i and article j, and $H(A_i)$ is the hub score of the author i. Then all the hub scores of authors are normalized to 1.

The hub score of a journal i:

$$H(J_i) = \frac{\sum_{P_j \in K_i} w_{ar-ju}(i,j) \cdot A(P_j)}{|K_i|} \tag{5}$$

where K_i is the articles published in journal i, $A(P_j)$ is the authority score of article j, $w_{ar-ju}(i,j)$ is the time-aware weight between journal i and article j, and $H(J_i)$ is the hub score of the journal i. Then all the hub scores of journals are normalized to 1.

The hub score of a topic i:

$$H(T_i) = \frac{\sum_{P_j \in M_i} w_{ar-tp}(i,j) \cdot A(P_j)}{|M_i|} \tag{6}$$

where M_i is the articles published related to topic i, $A(P_j)$ is the authority score of article j, $w_{ar-tp}(i,j)$ is the weight between topic i and article j, and $H(T_i)$ is the hub score of the topic i. Then all the hub scores of topics are normalized to 1.

The hub score of an article i:

$$H(P_i) = \frac{\sum_{P_j \in N_i} w_{ar-ar}(i,j) \cdot A(P_j)}{|N_i|} \tag{7}$$

where N_i is the articles cites or cited by the article i, $A(P_j)$ is the authority score of article j, $w_{ar-ar}(i,j)$ is the weight between article i and article j, and $H(P_i)$ is the hub score of the article i. Then all the hub scores of articles are normalized to 1.

3.4.2 Authority Scores

4 Experiments and Results

4.1 Dataset

We use the public dataset CORD19 which is a corpus of academic articles about COVID-19 and coronavirus scientific network published by Semantic Scholar Team at the Allen Institute for AI [14]. The final version was released on June 2, 2022 with 1M+ papers and around 370k full text support. We have performed a pre-processing that eliminates the duplicate articles, articles with no year information or journal information and broken rows. Also the dataset has not any information about the citations among the articles. To create a graph representation of CORD19 Dataset, we have used Semantic Scholar API to fetch the citation information related to each articles. By doing this operation, we also aimed to synchronize dataset with the Semantic Scholar Academic Graph API. Eventually, we have created and pıblished a graph dataset containing 728675 articles, 2210182 authors, and 5875663 citations. The dataset represents the citation network of the CORD19 Dataset on January 2024.

For experiments it would be computationally challenging to use all the graph to find the the related metrics, so we have performed our analysis on a subset of the graph containing 19981 articles, 121431 author, 2925 journals, and 209788 citations.

The semantic network was created using the MedCAT [5] outputs of the article abstracts. Here we follow the Unified Medical Language System [6] and the unique identifiers in the Metathesaurus Corpus. The extracted **Concept Unique Identifiers (CUI)** are used to build the semantic network by creating the links to related article by setting a accuracy threshold, %75 in our pipeline. Some Concept Unique Identifiers are discarded by considering the their type ids, see Appendix.

do not define them as baseline. We investigated the change of

4.1.1 Ground Truth and Evaluation Criteria

Finding a ground truth to analyze the academic networks is a common issue that most researchers tried to answer [9]. In our paper, we do not follow the prior ground truth because it would be contradictory to use metrics and algorithms based on old approaches as ground truth while trying to find a novel metric or trying to show that new parameters change the results. Since there was no human-based annotation on the dataset we worked with, we made a comparative evaluation criterion. In this case, we tried to analyze how different parameters in the algorithm change the result and what different semantic network causes.

Briefly, we investigated the change of most ranked papers and correlation coefficient by changing the parameters in order to see their effect on the main algorithm. The distance and irrelevance of the settings to the basis PageRank dominated algorithm can been as a differentiation from the traditional PageRank solutions.

4.1.2 Experiment Setup

We conducted experiments by changing the parameters in the algorithm to see their behaviours. 6 different kind of metrics:

- Alpha: traditional PageRank scores of the articles in citation network
- Beta: Authority scores propagated from author-article network
- Gamma: Authority scores propagated from journal-article network
- Delta: Authority scores propagated from topic-article network
- Omega: Authority scores propagated from article-article network, namely citation network
- Sigma: Time Factor of the articles

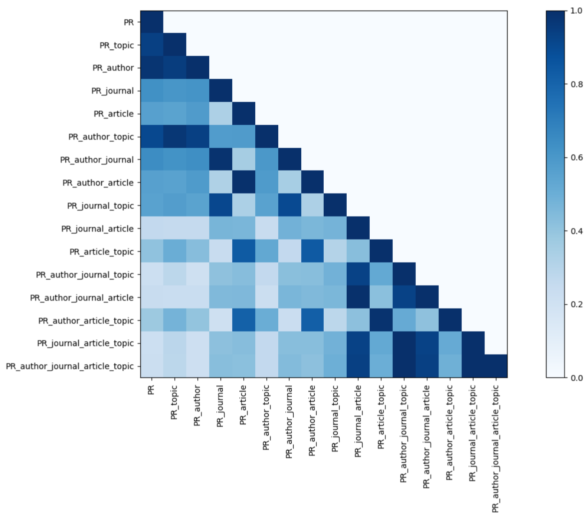
Among our experiments, we have chosen 12 settings to show the relationship between semantic network and the other networks. We ensured that equation $\alpha+\beta+\gamma+\delta+\omega+\sigma=1$ is met in all settings with the 0.1 random jump probability coming from $1-(\alpha+\beta+\gamma+\delta+\omega+\sigma)$. See the appendix to see exact values of parameters for the 12 different settings.

Apart from these, based on the prior works mentioned above, we set the time factor to 0.1 in all the calculations.

We have created various configurations through changing the coefficients above.

- **PageRank**: the PageRank score of the articles in the network. Base metric for the following combinations.
- **Author Information**: Author-article network information is added to network. It represent the authority scores coming from authors to articles.
- **Topic Information**: Topic-article network information is added to network. It represent the authority scores coming from topics to articles.
- **Journal Information**: Journal-article network information is added to network. It represent the authority scores coming from journals to articles.
- **Article Information**: Citation network information is added to network. It represent the authority scores coming from articles to articles.

4.2 Experiments and Result



The traditional PageRank algorithm is used as a baseline for the experiments. Even hough it is not a good choice to be a ground truth, understanding which networks have lower correlation may help us to differentiate the behaviours of the article, author, journal and topic networks.

In Figure 1, the correlation scores decrease as the new metric is added to calculation function. However, adding only topic or author to the calculation does not change results as much as adding journal or article information. Nevertheless, topic weighted calculation causes different results the versions with no topic information considering their similarity to the baseline PageRank only calculation. For instance, $PR\_journal$ setting has 0.62 correlation score with PageRank, which is very low considering that it still has PageRank coefficient $\alpha = 0.5$. Adding the topic value leads to 0.55 correlation score. Similarly, $PR\_article$ has 0.56 correlation, while $PR\_article\_topic$ is much lower with the correlation 0.41. Also consider the settings $PR\_author\_journal$ with 0.64 correlation to PageRank and $PR\_journal\_topic$ with 0.55 correlation. At this step, it is clear that topic network propagates more information than the author network considering its affect on the basic settings. Article information is the strongest among the settings. Naturally, it has very low correlation with the baseline setting because of the highly informative nature of the citation network. On the other hand, we believe that sparsity of the journal network is the main reason of the such affect of journal information.

Apart from the basic settings, the complex settings in which at least 3 network is used also produce similar results. We investigated the correlation of the settings with respect to setting in which every network is used, namely $PR\_author\_journal\_article\_topic$. Its lowest correlation among the complex settings is with the $PR\_author\_article\_topic$, which is similar to the above results indicating importance of journal information. Even though other correlations are similar, the second highest drop in correlation occurs when we discard the topic information from this setting.

4.2.1 Top Articles

Setting	Topic	Author	Journal	Article
Author	84	–	–	–
Journal	77	71	–	–
Article	16	18	15	–
PR	55	58	42	14

Table 1: Number of Common Papers Between Different Settings

We have analyzed the top 10 results with respect to different settings. At this time, we enable a network to dominate the results by assigning it to maximum possible coefficient value and nullifying the others. Note that the PageRank coefficient is not null in the above experiments, however its coefficient is zero in this experiment.

Similar to above experiments, the article network is not correlated to PageRank dominated setting. Actually, article network has lower correlation to all the other settings. Also, the journal dominated setting has more common articles with the topic dominated setting than the author dominated setting. This behaviour demonstrates the journal-topic relationship is stronger than the journal-author relationship, which also can be derived from the correlation scores above. Using topic instead of author together with journal information causes more differentiated results than the PageRank dominated settings.

Table 3 shows the top 10 ranked articles from 5 different settings. In the topic dominated setting column it is easy to spot that most of the articles have broader topics than the best ranked articles of other settings. They mainly include **Severe Acute Respiratory Syndorme (SARS)** text in their title, naturally in their abstracts. Also these articles mostly tried to understand and explain the literature. As we observed, the general terms such as *"respiratory"*, *"acute"* and *"cell"* tend to occur frequently in the abstract of these articles. Based on this phenomena, we can conclude that the topic dominated settings tend to give results favoring the literature reviews, reports and broader articles.

In the table, there are some papers among top papers with very low number of citations, such as *"Ultrastructural analysis of SARS-CoV-2 interactions with the host cell via high resolution scanning electron microscopy1*. Along with the papers in the rank 3 and 4, three of the top 5 articles in journal dominated setting belong to journal *Scientific Reports, Nature*. Considering the academic prestige of the Nature, it is not a surprise to see these papers when the journal information dominate the information flow in the network.

4.2.2 Effect of Citation

See the Appendix A

The effect of number of citations also important to understand the different behaviour of the settings. To achieve this, we calculate the Spearman's rank correlation coefficient between the settings and the citation count based ranking.

Metric	Correlation
Topic vs Citation Count	0.2525
Author vs Citation Count	0.2735
Journal vs Citation Count	0.2789
Article vs Citation Count	0.2575
PR vs Citation Count	0.6151

Table 2: Correlation of different settings with citation count

It is a widely known fact that PageRank based solutions can be biased because highly cited papers can dominates the scores. The purpose of this research is proposing a framework to investigate the network in different perspectives without the biased citation based approaches. At this point of view, the correlation between topic and citation count is lower than the author, journal and article.

Table 3: Table 3: The top 10 ranked articles

Topic	Author	Journal	Article	PageRank
The clinical pathology of severe acute respiratory syndrome (SARS): a report from China	Isolation from Man of ”Avian Infectious Bronchitis Virus-like” Viruses (Coronaviruses*) similar to 229E Virus, with Some Epidemiological Observations	Clinical Characteristics of Coronavirus Disease 2019 in China	Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia	Identification of a novel coronavirus in patients with severe acute respiratory syndrome.
A novel coronavirus associated with severe acute respiratory syndrome.	Antigenic relationships among the coronaviruses of man and between human and animal coronaviruses.	Ultrastructural analysis of SARS-CoV-2 interactions with the host cell via high resolution scanning electron microscopy	A Novel Coronavirus from Patients with Pneumonia in China, 2019	Identification of severe acute respiratory syndrome in Canada.
Identification of severe acute respiratory syndrome in Canada.	SEROEPIDEMIOLOGIC STUDIES OF CORONAVIRUS INFECTION IN ADULTS AND CHILDREN1	Stability of SARS-CoV-2 on critical personal protective equipment	The Human Respiratory System and its Microbiome at a Glimpse	Evidence of human metapneumovirus in Australian children
A cluster of cases of severe acute respiratory syndrome in Hong Kong.	Clinical Characteristics of Coronavirus Disease 2019 in China	Genomic mutations and changes in protein secondary structure and solvent accessibility of SARS-CoV-2 (COVID-19 virus)	Antigenic relationships among the coronaviruses of man and between human and animal coronaviruses.	A Novel Coronavirus from Patients with Pneumonia in China, 2019
Clinical Characteristics of Coronavirus Disease 2019 in China	A Novel Coronavirus from Patients with Pneumonia in China, 2019	Forecasting the spread of COVID-19 under different reopening strategies	First Case of 2019 Novel Coronavirus in the United States	Importation and Human-to-Human Transmission of a Novel Coronavirus in Vietnam
A Novel Coronavirus from Patients with Pneumonia in China, 2019	Visualization by Immune Electron Microscopy of a 27-nm Particle Associated with Acute Infectious Nonbacterial Gastroenteritis	A Novel Coronavirus from Patients with Pneumonia in China, 2019	The Impact of COVID-19 on Italy: A Lesson for the Future	Assessing spread risk of Wuhan novel coronavirus within and beyond China, January-April 2020: a travel network-based modelling study
Evidence of human metapneumovirus in Australian children	Middle East Respiratory Syndrome Coronavirus (MERS-CoV): A Perpetual Challenge	Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia	Prophylactic and therapeutic remdesivir (GS-5734) treatment in the rhesus macaque model of MERS-CoV infection	First Case of 2019 Novel Coronavirus in the United States
Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia	Studies With Human Coronaviruses II. Some Properties of Strains 229E and OC43	Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention.	SEROEPIDEMIOLOGIC STUDIES OF CORONAVIRUS INFECTION IN ADULTS AND CHILDREN1	A novel coronavirus associated with severe acute respiratory syndrome.
First Case of 2019 Novel Coronavirus in the United States	Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia	[Asymptomatic infection of COVID-19 and its challenge to epidemic prevention and control].	Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020	Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia
Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention.	First Case of 2019 Novel Coronavirus in the United States	First Case of 2019 Novel Coronavirus in the United States	Human Coronavirus in Hospitalized Children With Respiratory Tract Infections: A 9-Year Population-Based Study From Norway	A major outbreak of severe acute respiratory syndrome in Hong Kong.

5 Conclusion and Future Work

In this paper, we have explored the possible ways of using the topic related information through propagating it with a hybrid HITS algorithm following the prior works. We have showed that proper parameter selection ,with respect to author, journal, topic or article network, can lead to different ranking for articles. At this point we see that topic network has more power to change the article rankings than author network. Also, adding topic information lead to lower correlation to the number of citations compared to other networks. This phenomena arises from not only propagation from topic network but also semantic weighting of the citation links.

We believe that using a topic network that also has interconnected links between their nodes such as subclass relation, can lead better results with the help of strong representation of the semantic relationships. In this case, using the UMLS Metathesaurus knowledge graph can be a suitable future work beyond our research.

A significant contribution would be the proposing a topic matrix in the calculation to find the most ranked papers for a given topic at any time, which can improve the quality of the semantic ranking in the network.

References

[1] Wasim Ahmed, Josep Vidal-Alaball, Joseph Downing, and Francesc López Seguí. Covid-19 and the 5g conspiracy theory: Social network analysis of twitter data. *Journal of Medical Internet Research*, 22, 2020.

[2] Liwei Cai, J. Tian, Jiaying Liu, Xiaomei Bai, Ivan Lee, Xiangjie Kong, and Feng Xia. Scholarly impact assessment: a survey of citation weighting solutions. *Scientometrics*, 118:453 – 478, 2018.

[3] Eugene Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178 4060:471–9, 1972.

[4] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In *ACM-SIAM Symposium on Discrete Algorithms*, 1999.

[5] Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel M Bean, Aurelie Mascio, Leilei Zhu, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert J Stewart, Anoop Dinesh Shah, Wai Keong Wong, Zina M. Ibrahim, James T. H. Teo, and Richard J. B. Dobson. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083, 2020.

[6] Donald A. B. Lindberg, Betsy L. Humphreys, and Alexa T. McCray. The unified medical language system. *Methods of Information in Medicine*, 32:281 – 291, 1993.

[7] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking : Bringing order to the web. In *The Web Conference*, 1999.

[8] Hassan Sayyadi and Lise Getoor. Futurerank: Ranking scientific articles by predicting their future pagerank. In *SDM*, 2009.

[9] Yu-Sheng Su, Chien-Liang Lin, Shih-Yeh Chen, and Chin-Feng Lai. Bibliometric study of social network analysis literature. *Libr. Hi Tech*, 38:420–433, 2019.

[10] Tracy M. Sweet. Social network analysis. *The Reviewer's Guide to Quantitative Methods in the Social Sciences*, 2018.

[11] Jie Tang, Ruoming Jin, and Jing Zhang. A topic modeling approach and its integration into the random walk framework for academic search. *2008 Eighth IEEE International Conference on Data Mining*, pages 1055–1060, 2008.

[12] Julia Vassey, Tom Valente, Joshua Barker, Cassandra A. Stanton, Dongmei Li, Linnea Irina Laestadius, Tess Boley Cruz, and Jennifer B. Unger. E-cigarette brands and social media influencers on instagram: a social network analysis. *Tobacco Control*, 32:e184 – e191, 2022.

[13] Dylan Walker, Huafeng Xie, Koon-Kiu Yan, and Sergei Maslov. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007:P06010 – P06010, 2006.

[14] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics.

[15] Yujing Wang, Yunhai Tong, and Ming Zeng. Ranking scientific articles by exploiting citations, authors, journals, and time information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2013.

[16] Erjia Yan, Ying Ding, and Cassidy R. Sugimoto. P-rank: An indicator measuring prestige in heterogeneous scholarly networks. *J. Assoc. Inf. Sci. Technol.*, 62:467–477, 2011.

[17] Yu Zhang, Min Wang, Florian Gottwalt, Morteza Saberi, and Elizabeth Chang. Ranking scientific articles based on bibliometric networks with a weighting scheme. *J. Informetrics*, 13:616–634, 2019.

Appendices

A Number of citation of Top Papers

Title	Citations
Identification of a novel coronavirus in patients with severe acute respiratory syndrome.	759
Genomic mutations and changes in protein secondary structure and solvent accessibility of SARS-CoV-2 (COVID-19 virus)	3
Assessing spread risk of Wuhan novel coronavirus within and beyond China, January-April 2020: a travel network-based modelling study	21
Studies With Human Coronaviruses II. Some Properties of Strains 229E and OC43	5
A novel coronavirus associated with severe acute respiratory syndrome.	727
Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia	1104
The clinical pathology of severe acute respiratory syndrome (SARS): a report from China	125
Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72,314 Cases From the Chinese Center for Disease Control and Prevention.	1043
Antigenic relationships among the coronaviruses of man and between human and animal coronaviruses.	24
Prophylactic and therapeutic remdesivir (GS-5734) treatment in the rhesus macaque model of MERS-CoV infection	85
Asymptomatic infection of COVID-19 and its challenge to epidemic prevention and control.	2481
Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020	125
A cluster of cases of severe acute respiratory syndrome in Hong Kong.	195
Ultrastructural analysis of SARS-CoV-2 interactions with the host cell via high resolution scanning electron microscopy	2
The Impact of COVID-19 on Italy: A Lesson for the Future	5
Visualization by Immune Electron Microscopy of a 27-nm Particle Associated with Acute Infectious Non-bacterial Gastroenteritis	3
Importation and Human-to-Human Transmission of a Novel Coronavirus in Vietnam	130
Isolation from Man of "Avian Infectious Bronchitis Virus-like" Viruses (Coronaviruses*) similar to 229E Virus, with Some Epidemiological Observations	18
The Human Respiratory System and its Microbiome at a Glimpse	1
A major outbreak of severe acute respiratory syndrome in Hong Kong.	408
Human Coronavirus in Hospitalized Children With Respiratory Tract Infections: A 9-Year Population-Based Study From Norway	10
SEROEPIDEMIOLOGIC STUDIES OF CORONAVIRUS INFECTION IN ADULTS AND CHILDREN1	45
A Novel Coronavirus from Patients with Pneumonia in China, 2019	1563
Evidence of human metapneumovirus in Australian children	17
Clinical Characteristics of Coronavirus Disease 2019 in China	1575
Forecasting the spread of COVID-19 under different reopening strategies	0
Identification of severe acute respiratory syndrome in Canada.	232
Middle East Respiratory Syndrome Coronavirus (MERS-CoV): A Perpetual Challenge	7
Stability of SARS-CoV-2 on critical personal protective equipment	2
First Case of 2019 Novel Coronavirus in the United States	524

B Parameters and Settings

Metric	α	β	γ	δ	ω	σ
PR	0.8	0.0	0.0	0.0	0.0	0.1
PR_author	0.5	0.3	0.0	0.0	0.0	0.1
PR_journal	0.5	0.0	0.3	0.0	0.0	0.1
PR_article	0.5	0.0	0.0	0.0	0.3	0.1
PR_author_journal	0.4	0.2	0.2	0.0	0.0	0.1
PR_author_article	0.4	0.2	0.0	0.0	0.2	0.1
PR_journal_article	0.4	0.0	0.2	0.0	0.2	0.1
PR_author_journal_article	0.2	0.2	0.2	0.0	0.2	0.1
PR_topic	0.5	0.0	0.0	0.3	0.0	0.1
PR_author_topic	0.3	0.2	0.0	0.3	0.0	0.1
PR_journal_topic	0.3	0.0	0.2	0.3	0.0	0.1
PR_article_topic	0.3	0.0	0.0	0.3	0.2	0.1
PR_author_journal_topic	0.2	0.1	0.1	0.3	0.2	0.1
PR_author_article_topic	0.2	0.1	0.0	0.3	0.2	0.1
PR_journal_article_topic	0.2	0.0	0.1	0.3	0.2	0.1
PR_author_journal_article_topic	0.2	0.133	0.133	0.2	0.133	0.1
Topic Dominated	0.1	0.0	0.0	0.7	0.0	0.1
Author Dominated	0.1	0.7	0.0	0.0	0.0	0.1
Journal Dominated	0.1	0.0	0.7	0.0	0.0	0.1
Article Dominated	0.1	0.0	0.0	0.7	0.0	0.1