# Journal Co-citation Network based on Wikipedia Entries

Stephen Kurniawan (3044444)
stephen.kurniawan.leuphana@gmail.com
Evgeniya Chetneva (3044086)
evgeniya.chetneva@stud.leuphana.de
Bek-Myrza Nurmatov (3044116)
bek-myrza.nurmatov@stud.leuphana.de

August 13, 2022

**Abstract**

The study conducts a thorough comparison of journal co-citation network based on Wikipedia references between the results based on our data set from 2020 and original paper with a data set from 2018. A new data set includes over 29 million citations. After preprocessing and filtering the data related to journals the sample was reduced to 2 million citations, which is as much as twice that in the original paper. Although the newer data set contained more information in regard to Wikipedia entries, references, and articles, we observed similar trends, for example, the most popular journals remain in the top positions over the years. In addition to the general summary statistics, we built a network that included the most significant journals and provided the description of the findings as well as network analysis information, which we also compared with the results from the original paper. Overall, our main goal was to track the changes that happened during these two years, as well as provide our own discoveries. Supplemental materials are available at https://github.com/stepkurniawan/network-analysis-wikipedia-journals.

***Keywords:*** *Wikipedia; Journals; Co-citation; Citation; Network; Graph.*

## 1 Introduction

Our project is based on the paper "Science through Wikipedia: A novel representation of open knowledge through co-citation networks" (Arroyo-Machado, Torres-Salinas, Herrera-Viedma, and Romero-Frías (2020)), to which we will refer as the "original paper". The main goal of the original paper is to present an overview of science from Wikipedia's perspective through creating and analyzing the co-citation networks of journals.

Wikipedia is one of the largest online sources of information and not only it is free, but everyone can contribute by adding and editing entries, therefore allowing for quick accumulation of knowledge. However, to verify the provided information in the entries, Wikipedia strongly recommends including references, with a preference for specialized publications, which refers to academic and peer review publications as well as scholarly monographs and textbooks. Therefore, Wikipedia creates the so-called "social construction of knowledge", which connects information, provided by Wikipedia, and academic research. It offers the opportunity the investigation how Wikipedia and Science interrelate.

According to Bellomi and Bonato (2005), Wikipedia pages are growing exponentially over time. The difference of one year could mean much in the data growth and can lead to different results. By introducing a newer data set, we can compare the similarities and the differences in the analysis.

The main objective of the authors (Arroyo-Machado et al. (2020)) was to discover the different visions offered by Wikipedia by using co-citation networks (Small (1973)) at different levels of aggregation: 1) journal co-citation maps 2) main field co-citation maps 3) field co-citation maps. This objective is strongly grounded in the exploration and analysis of the created networks. Through these maps, the authors intended to obtain a full understanding of how scientific articles, published in journals, are used and consumed, from the perceptive of the Wikipedia entries.

## 2 Network Model

In the original paper, the authors created three co-citations networks, as mentioned before, however in our project, we will focus only on the first one (network of journals), and therefore we will explain the details only for this particular map.

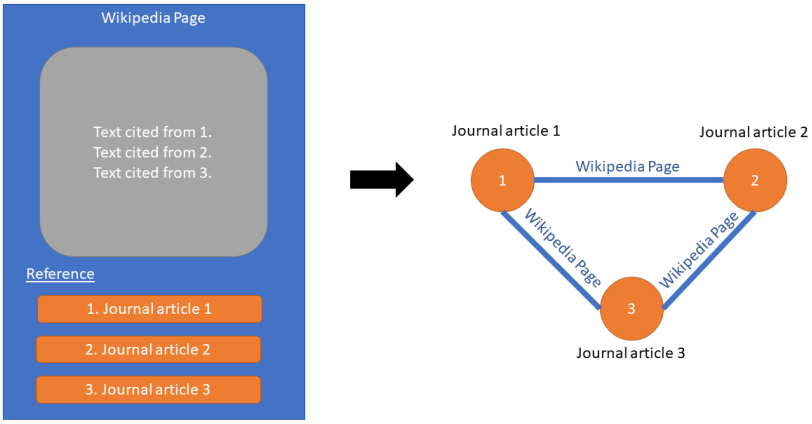


Figure 1: The concept for creation of networks.

Figure 1 presents the overall concept for the creation of the networks, which goes as follows. The main source of information are Wikipedia entries, which contain references to the other sources. From the reference, we can select references to the academic articles and create a network with the co-citation of articles. Every two articles that were mentioned together in one Wikipedia entry has a co-citation connection with the weight of 1. But the goal is to analyze journals, therefore the connection between articles and journals is made, which allowed for the creation of a co-citation network of journals. Hence, in the main network, which is a co-citation network of journals, journals are represented as nodes and edges referring to the co-citation link. The size of the node depends on the number of citations each journal received, while the thickness of the edge - the number of co-citations,

| citations |
| --- |
| {{Cite journal \| doi = 10.1371/journal.pone.0089165\| title = The Origin and Early Evolution of Sauria: Reassessing the Permian Saurian Fossil Record and the Timing of the Crocodile-Lizard Divergence \| journal = PLoS ONE\| volume = 9\| issue = 2\| pages = e89165\| year = 2014\| last1 = Ezcurra \| first1 = M. N. D. \| last2 = Scheyer \| first2 = T. M. \| last3 = Butler \| first3 = R. J. \| pmid=24586565 \| pmc=3937355}} |

Figure 2: The fragment of the column citations.

existing between each two journals. The authors added the information about areas of science in terms of colors to each node. Therefore, the full map contains information about citations and co-citations for each journals, as well as its belongingness to the certain area of science.

# 3 Contribution

## 3.1 Network Data Source

We are using a newer and larger data set from the Singh, West, and Colavizza (2021) paper. It includes 29.3 million citations from 6.1 million English Wikipedia articles as of May 2020, and is classified as being books, journal articles, or Web content. Using the data source (Singh et al. (2021)) we imported the data to a python data frame. In the original paper, they are using data set from April 2018, with 847 thousand citations from 193 thousand Wikipedia articles. A more comprehensive comparison will be discussed in Subsection 4.1.

## 3.2 Data Cleaning

The python data frame contains 29.276.667 rows and 32 columns. The main steps of the data cleaning process include filtering rows by the journal citations, and columns by the name of the journals and Wikipedia page titles.

Filtering the journal citations was made with the column **type_of_citation**, which has an attribute **cite journal**. Afterwards, the columns **citations** and **page_title** were chosen. The column **citations** contains a long list of attributes attribute from which the journal name should be extracted (fig. 2).

After extracting the journal name from the column **citations** it was combined with the column page title, forming a journal data frame that will be used further.

## 3.3 Data Preparation

In the data preparation, we are mainly focusing on two tasks. First, we introduce *SCOPUS* database to enhance our data to identify the subject area of each journal. And finally, we are extracting the information from each stage to compare the centrality and visualize it in the following Subsection 3.4.

Table 1: Example table of descriptive statistics of the main variables.

| Page Title | Journal Name |
|---|---|
| Ballitore | The Economic History Review |
| N-Propyl azide | Bioorganic & Medicinal Chemistry Letters |
| Pristimantis satagius | The IUCN Red List of Threatened Species |
| Long-term impact of alcohol on the brain | The American Journal of Geriatric Psychiatry |
| Sulfamerazine | Journal of the American Chemical Society |
| Photonic molecule | The Journal of Physical Chemistry C |
| Graphene production techniques | The Journal of Physical Chemistry Letters |
| Transition metal phosphido complexes | Organometallics |
| Donald Crothers | Nature |
| Systellommatophora | Zoological Journal of the Linnean Society |
| Christopher Dye | New England Journal of Medicine |
| Percy Williams Bridgman | Physics Today |
| ... | ... |

Table 2: Full node table after getting enhanced by SCOPUS area data set.

| Index | Journal Name | Number of Citations | Area |
|---|---|---|---|
| 0 | nature | 35475 | Multi |
| 1 | journal of biological chemistry | 30200 | Life Sciences |
| 2 | pnas | 29544 | Multi |
| 3 | science | 25924 | Multi |
| 4 | PLOS one | 12602 | Multi |
| ... | ... | ... | ... |
| 176651 | bull inst fr afr noire a | 1 | NaN |
| 176652 | httpwwwluxurytraveladvisorcom | 1 | NaN |
| 176653 | httpwwwluxurytravelmagazinecom | 1 | NaN |
| 176654 | newcastle herald | 1 | NaN |
| 176655 | overland literary journal | 1 | NaN |

#### 3.3.1 SCOPUS Area Data - Node Enhancement

*SCOPUS* is a database that includes journals, books, and other public articles with their metadata, which includes areas and fields (service.elsevier.com). According to SCOPUS, the journals are classified into four main subject areas (”Physical Sciences” (PS), ”Health Sciences” (HS), ”Social Sciences” (SSH), and ”Life Sciences” (LS)) and one transversal area called ”Multidisciplinary”. In the original paper, the authors were using SCOPUS to add colors to the network graph to enhance area information of each journal. We also did the same by merging areas in the journals. The resulting table can be seen in the Table 2.

#### 3.3.2 Edges Pruning and Trimming

In the original paper, they used the PathFinder algorithm (PFnet) from Quesada (2005) to prune the network so that it reduces to a minimum covering the tree. This algorithm keeps only the strongest co-citation links between all pairs of nodes and offers a diaphanous view of large networks (Arroyo-Machado et al. (2020)). For journals specifically, they are trimming the edges that have lower than 50 co-citations.

Due to the exponential growth of Wikipedia pages in the past years, our limited computational power, and the recent algorithm, we decided to trim the data set directly without applying the Pathfinder algorithm. The assumption is a heavy weight on the edges means higher importance because it indicates high co-citing phenomena. Therefore, by trimming the small weighted relationships, we preserve the central information, while reducing the

Table 3: Comparison of the different numbers of data between the original paper and ours.

| | Original paper's data set | Our data set | Changes (%) |
|---|---|---|---|
| **References to scientific articles** | 847.512 | 1.996.721 | 236% |
| **Wikipedia entries** | 193.802 | 486.710 | 251% |
| **Scientific articles published** | 598.746 | 1.535.906 | 257% |
| **Journals** | 14.149 | 176.515 | 1248% |

network size.

We chose to trim the edges with weights less than 1400 using the same reasoning as the original paper, to have a limited number of nodes to fit the display. Before trimming, we have 5 Million edges and 176 thousand nodes, it becomes 641 edges and 237 nodes after the trimming process. The number of edges is very similar to the original paper which has 629 edges after their PathFinder and trimming process. The original network has more nodes, even though having a similar number of edges.

### 3.4 Data Visualization

For the creation of the network, we used *Gephi* (gephi.org), since it provides various tools for network presentation and analysis. In order to create a network we used two data sets: one with the information about nodes, specifically journals names, the number tations for the node sizes, and the area for the assignment of the colors, and another with the information about edges, specifically, two journals and the number of their co-citations for the thickness of the edge. After creating the network, we removed the total of 30 nodes that had no connection with the main network, resulting in 190 nodes and 599 edges. For better visualization, we used *Force Atlas 2* layout and the *Nooverlap* tool for creating the margin between nodes. In contrast to the original paper, we decided to put five largest nodes on the edges of the graphs, since most of the nodes had connections with several of them and therefore better be put in the center. We also supplemented the network with the legend that explains the areas as well as their combination and added labels to the 20 most cited journals.

## 4 Results and Discussion

### 4.1 General Description

In the original paper, they analyzed only 847.512 references to scientific articles across 193.802 Wikipedia entries, with a total of 598.746 scientific articles published across 14.149 journals cited in the year 2018. Our data set is generally more than twice larger over the board with 1.996.721 references to scientific articles across 486.710 Wikipedia entries, having a total of 1.535.906 scientific articles across 176.515 cited journals from the year 2020.

As seen in Table 3, the enormous difference in the number of journals could stem from the quality of our data set. Even though we have done data cleaning algorithms, many human mistakes such as typos and abbreviations can affect the overall quality of the data.

Table 4: The top 10 most cited journals in Wikipedia and its areas. Multidisciplinary means having multiple areas, while empty cell means that the journal name is not found.

| Journal name | Number of Citations | Area |
|---|---|---|
| nature | 35.475 | Multidisciplinary |
| journal of biological chemistry | 30.200 | Life Sciences |
| pnas | 29.544 | Multidisciplinary |
| science | 25.924 | Multidisciplinary |
| PLOS one | 12.602 | Multidisciplinary |
| cell | 9.189 | Life Sciences |
| zootaxa | 8.241 | Life Sciences |
| genome res | 7.275 | Life Sciences & Health Sciences |
| lancet | 6.484 | Health Sciences |
| the astrophysical journal | 6.304 | Physical Sciences |

In the original paper, each Wikipedia entry includes an average of 4,37 ($\pm 8,35$) references to scientific articles, while in our data set, it includes 4,09 ($\pm 8,98$). The difference between them is not significant. It could be because the time difference between two data sets is not large, therefore the style in which people write wiki entries does not change. Another similar feature is the number of articles that receive only one citation. In the original paper, they found 489.235 articles with that property - which corresponds to 57,73% of all the references - in the study sample. In our data set, we found 1.270.531 articles with such a feature, which corresponds to 63,63% of all the references.

In the original paper, the high standard deviations are explained by looking at the top 1% of Wikipedia entries with high references, which have the mean and standard deviation of 60,87 ($\pm 32,75$), representing 13,92% of all references. In our case, we have 56,53 ($\pm 35,53$), representing 14,9% of all references. Similar to the original paper, our top 1% of entries is related to listings, history, genes, common diseases, and paleontology.

Actually, we think that using mean and standard deviations to measure a highly skewed data set is less than optimum, however we did it in order to compare it with the original paper. Based on the data comparison above, we concluded that the data set that we are using is not only similar but enhances the old data set that was used in the original paper.

In total, filtering only on journals gives rise to 1.997.167 cited journals and 22.717.803 journals co-citations through Wikipedia. The most cited journals are Nature, Journal of Biological Chemistry, PNAS, Science, and PLOS ONE, which can be seen in Table 4. Where the most co-cited journals are often self-citing, the first non-self-citing journal is Nature-Science, which co-cited each other for 51.706 times in Wikipedia (Tab. 5). We allow the possibility to co-cite journals multiple times from the same Wikipedia entry since it is possible and likely to cite the same journal but have different articles.

## 4.2 Journals by Areas

### 4.2.1 Full Network

The original paper reports a mean of 42,36 and a standard deviation of 269,22 for the number of articles and a mean of 59.9 with a standard deviation of 458,54 for the number of citations in the journals. While in our data, we observed the mean of 11,28 and a standard

Table 5: The top 10 most co-cited journals relationships in Wikipedia.

| Journal 1 | Journal 2 | Number of Co-citations |
|---|---|---|
| nature | science | 51.706 |
| nature | pnas | 43.631 |
| journal of biological chemistry | pnas | 40.297 |
| pnas | science | 33.636 |
| journal of biological chemistry | nature | 24.171 |
| cretaceous research | zootaxa | 21.597 |
| cell | journal of biological chemistry | 18.217 |
| cell | nature | 16.761 |
| cell | pnas | 14.544 |
| plos one | pnas | 14.383 |

deviation of 174,55 in regard to citations for each journal, so our data set contains more journals with each journal receiving fewer citations on average. As mentioned, it is caused by the low quality of journal names in our data set.

Comparing the areas, the original paper reported that ”Social Sciences and Humanities” contains the largest amount of journals - 3.279 (23,2%). This area also remained in our first position with 5.784 (27,5%) journals. The subsequent areas in the original paper are ”Health Sciences” with 3.077 (21,7%) journals, ”Physical Sciences” with 2.489 (17,6%) journals, ”Life Sciences” with 1.298 (9,2%) journals and only 31 (0,2%) journals belong to the multidisciplinary area. The rest of the journals (3.975, 28,1%) belong to two or three areas simultaneously. In our data, the order of the areas also remained the same with ”Health Sciences” containing 4.449 (21,2%) journals, ”Physical Sciences” - 4.064 (19,3%), ”Life Sciences” - 2.088 (9,9%), 75 (0,4%) journals belonging to the Multidisciplinary, and 4.551 (21,7%) journals belonging to several areas. The provided percentage for our data set are shown based on the number of journal for which we know the area (21.011) and not on total number of journals in the data set.

In the original paper, the three most cited journals are Nature (26.434 citations), PNAS (24.104), and the Journal of Biological Chemistry (21.921), while in our data Nature is still the most cited journal with a total of 35.475 citations, however, the Journal of Biological Chemistry took the second place with 30.200 citations and PNAS only the third with 29.544 citations

The most referenced area is ”Life Sciences” with 300.402 references for the full data set and 153.994 for the trimmed network. However, for the full version, the second place belongs to ”Physical Sciences” with 173.621 references, but for the trimmed version it is Multidisciplinary with 111.226 references, while ”Physical Sciences” is in the third place with 61.035 references.

#### 4.2.2 Trimmed Network

It can be seen from the network of the original paper (Arroyo-Machado et al. (2020)) in Figure 3 that the vast majority of journals fall under the ”Life Sciences” category (36,6%) and the ”Life Sciences & Health Sciences” category (19,2%), the ”Health Sciences” category (14,5%), and the ”Physical Sciences” category (14,5%). Nature, Science, and PNAS play

a major role in the network, and they also demonstrate their multidisciplinary nature by strongly co-citing with journals from other fields. A majority of PNAS connections relate to journals in the "Life Sciences" and "Health Sciences & Life Sciences" areas. Despite being listed under "Health Sciences & Life Sciences", PLOS ONE also has strong co-citation links with journals in other fields.

Figure 4 illustrates the co-citation network of journals from our data set. It shows that most journals also belong to the "Life Sciences" area (74, 31,2%), followed by "Physical Sciences" (47, 19,8%) and the combination of "Life Sciences" & "Health Sciences" (25, 10,5%). Table 6 contains statistics for all areas. Three out of the four largest intermediates (Science, PNAS, and Nature) belong to the Multidisciplinary area and also have strong co-citation links with each other. PLOS ONE also was included in the Multidisciplinary area, even though in the original paper it was described as a combination of "Health Science" and "Life Science".

Most of the other journals have a connection to two or more largest components (also including the Journal of Biological Chemistry). Additionally, most of the "Life Sciences" journals have a connection with the Journal of Biological Chemistry and PNAS, while most of the "Physical Sciences" journals connect with Science, Nature, and PLOS ONE (the fifth largest component).

Overall, our network (Fig.4) has more edges, because we did not use PathFinder algorithm to trim the excessive edges. Since many of the nodes have two (or more) edges, it creates tension and placed them in the middle of those two journals they are connected with. Therefore, we can see in the figure that a large number of nodes are in the middle.

In the trimmed version of the network, the first thing that is mentioned in the original paper is the degree (Barabási and Albert (1999)) of journals, with PNAS (251), Nature (76), and Journal of Biological Chemistry (41) having the highest degree. In our data, Nature has the highest degree (87), although its degree showed a moderate increase compared to the original paper, the degree of PNAS decreased drastically moving to the second position with 78, while Journal of Biological Chemistry (68) remained in the third position with a higher degree. Table 7 presents data about the degrees of the top 10 journals of the original paper and for our data set.

Table 8 presents the comparison of the betweenness (Leydesdorff (2007)) of the top 10 journals from original paper and from our data set. The first key difference is that the betweenness for our data is in general lower, due to the fact that we used another algorithm for the network creation and therefore kept all the connections. Regarding the top three journals, for original paper it was PNAS (0,881), Nature (0,389) and Science (0,220), while for our data it was Nature (0,352), Journal of Biological Chemistry (0,255) and PNAS (0,220).

Table 9 presents the comparison of the closeness (Sabidussi (1966)) of the top 10 journals from original paper and from our data set. Regarding the top three journals, in original paper it was PNAS (0,49), Nature (0,41) and Cold Spring Harbor Symposia on Quantitative Biology (0,38), while for our data it was Nature (0,63), PNAS (0,60) and Science (0,54). The

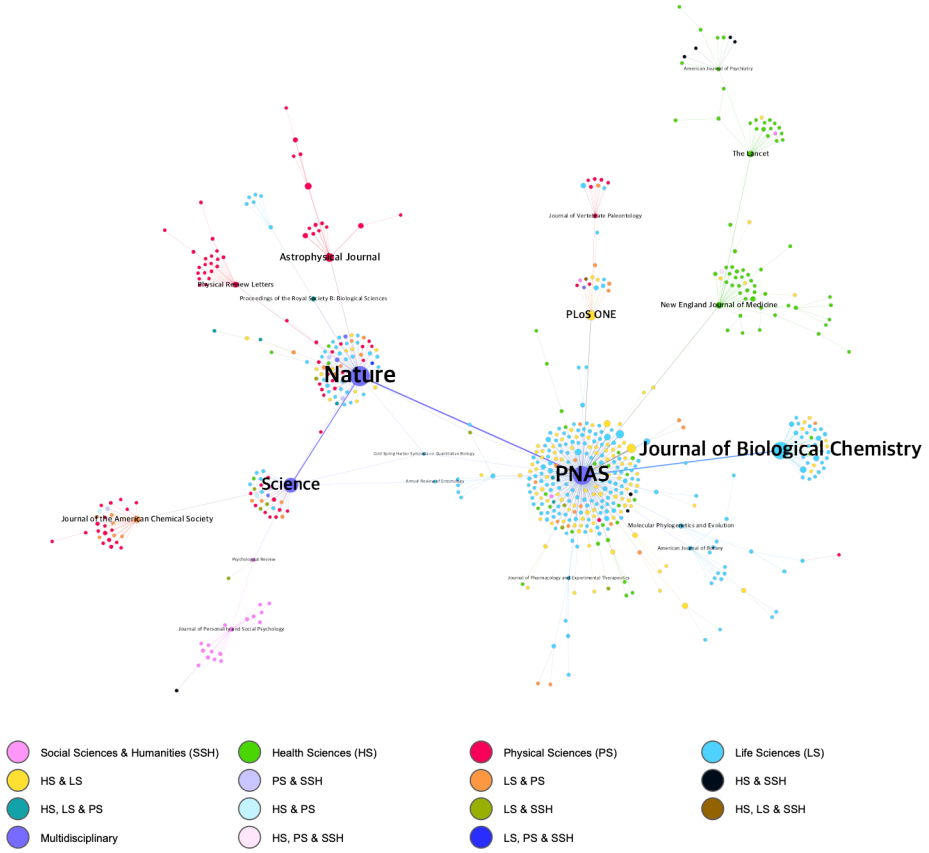


Figure 3: Co-citation network graph of journals from original paper Arroyo-Machado et al. (2020).

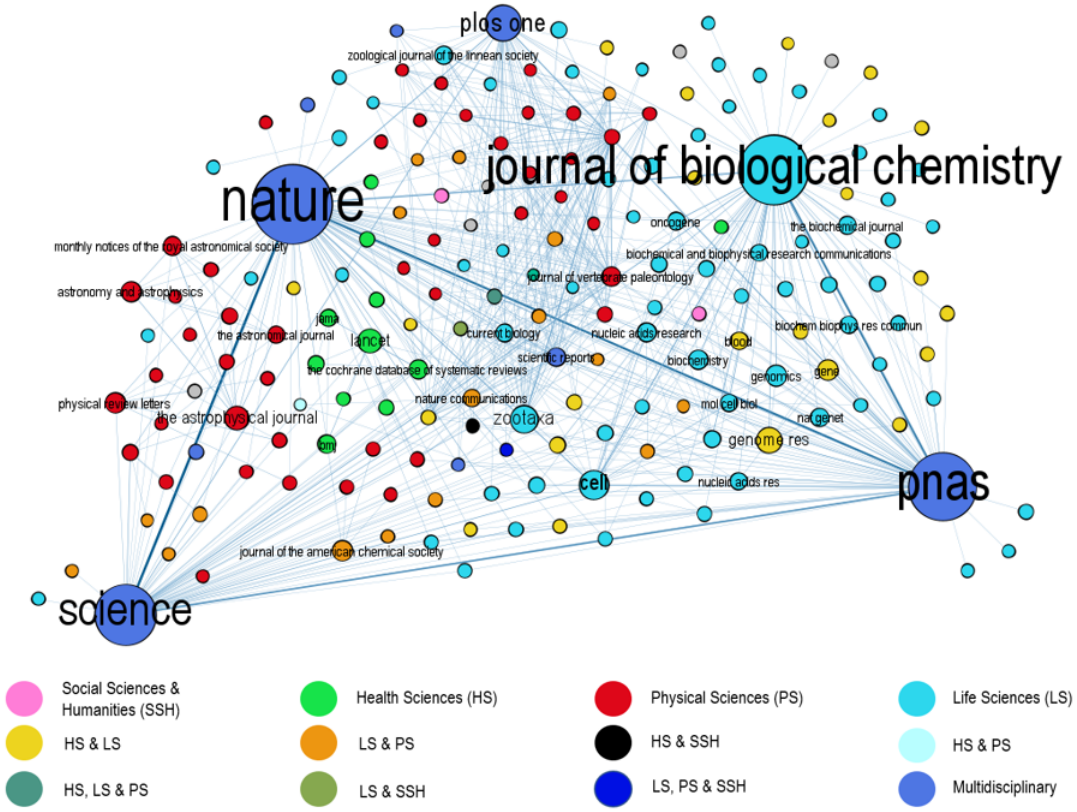


Figure 4: Co-citation network graph of journals from our data set.

reason why for our data set the closeness tend to be higher connects with the fact of the more connections between nodes and therefore the sum of the shortest paths for each node also tends to be higher.

Table 10 presents the comparison of the eigenvectors (Zaki and Meira (2014)) of the top 10 journals from the original paper and from our data set. In the original paper, the first place goes to PNAS (1,0), while the second and third to Nature (0,198) and Journal of Biological Chemistry (0,112). In our dataset the first place with eigenvector 1,0 goes to Nature, followed by PNAS (0,969) and Science (0,751). The difference can be explained by the fact that in the original network PNAS has the largest number of connections, while also directly connected with the nodes with a high number of connections. While in our case, each node including the biggest ones has more connections, therefore we can observe more nodes with a high eigenvector.

### 4.3 Conclusion

After comparing our results with the original paper, we can conclude several things. The first one is that the number of references, Wikipedia entries, and articles cited has grown about 2,5 times from 2018 to 2020. Another interesting fact is that the already famous journals like Nature, Science, and PNAS are still the top journals that are being referenced, and therefore accumulating more articles. It also means that multidisciplinary areas are still in the top-cited and co-cited journals.

Unfortunately, in specific cases like the number of nodes on the trimmed network, the original paper did not provide any information that we can compare. Another surprising change over the years is that PLOS ONE changed its area, coming from "Health Sciences & Life Sciences" to Multidisciplinary.

## 5 Journals per Areas Network Graphs

Since we have analyzed the journals by area intensively. It is also interesting to see how the network graph is formed to see which journals are dominating the center of each area and their relationships with other journals.

Firstly, in the area of Social Sciences and Humanities (Fig. 5), even though the Journal of Personality and Social Psychology is clearly having the largest node - cited a lot in Wikipedia - it does not have much co-citation. Therefore, the co-citation to the other nodes got trimmed in the process. It is also interesting that the many major co-cited journals are related to East European, such as Russian Review, The Slavonic and East European Review, Russian History, etc.

Secondly, in the area of Health Sciences (Fig. 6), the node that received the largest number of citations (Lancet) also has a strong co-citation link to the two next largest nodes, namely BMJ and The Cochrane Database. In addition, nodes with a higher number of citations have a co-citation link with each other, while the journals with the less total number of citations mostly connect only to one of the larger nodes.


Figure 5: The network graph of Social Sciences and Humanities area.

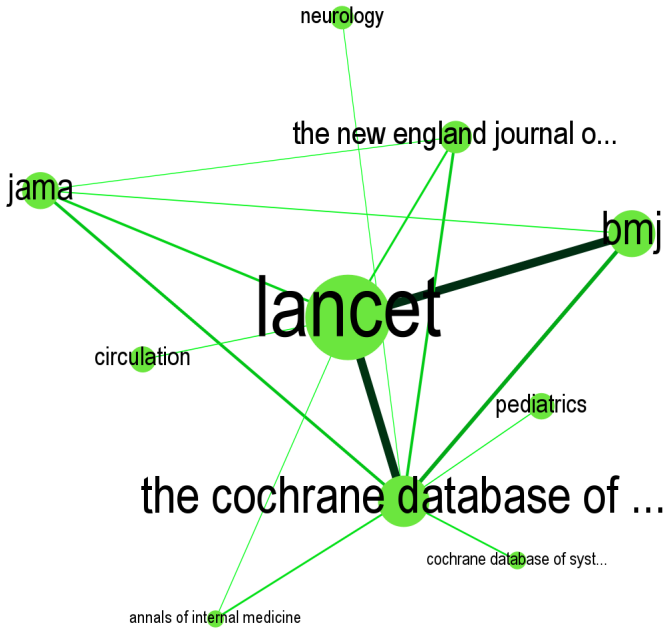


Figure 6: The network graph of Health Sciences area.

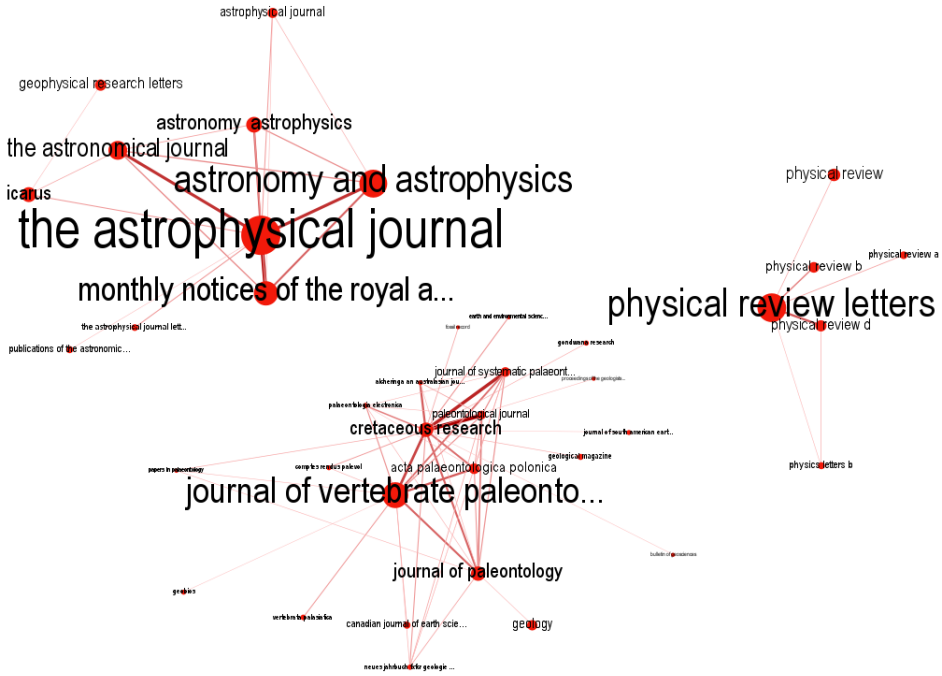


Figure 7: The network graph of Physical Sciences area.

Thirdly, in the area of Physical Sciences (Fig. 7), we can observe three islands of nodes with no connection between them. On the biggest island, the node with the largest number of citations is the Journal of Vertebrate Paleontology, and although it shares the co-citation link with several journals, it is the Cretaceous Research that has the strongest co-citations links (with the journal of Systematic paleontology and Paleontological Journal). Overall, this island mostly contains journals about paleontology and geology, however, it is exactly paleontological journals that contain most citations to co-citations links, creating the basis for this particular network.

The second island contains journals about astronomy and astrophysics, with The Astrophysical Journals being the largest node and having co-citation links with most of the other journals in this network. Another interesting point is as aforementioned, we can see the results of the problems with initial journal naming since we have duplicates, for example, ”Astronomy Astrophysics” and ”Astronomy and Astrophysics”, also ”Astrophysical Journal” and ”The Astrophysical Journal” are referring to one journal but have slightly different names.

The smallest island is a network of Physical Review Letters (being the largest component) and its several versions, which are denoted by letters from A to D. The only exception is the second-largest node ”Physical Review”, which also has a connection with the main component.

Fourthly, the center of the area of Life Sciences (Fig. 8) is the Journal of Biological Chemistry. It is both the most cited journal and the most co-cited journal on Wikipedia. It has high co-citation relationships with Cell, Oncogene, and Molecular and Cellular Biology. The figure also shows that many abbreviated journal names are referring to the

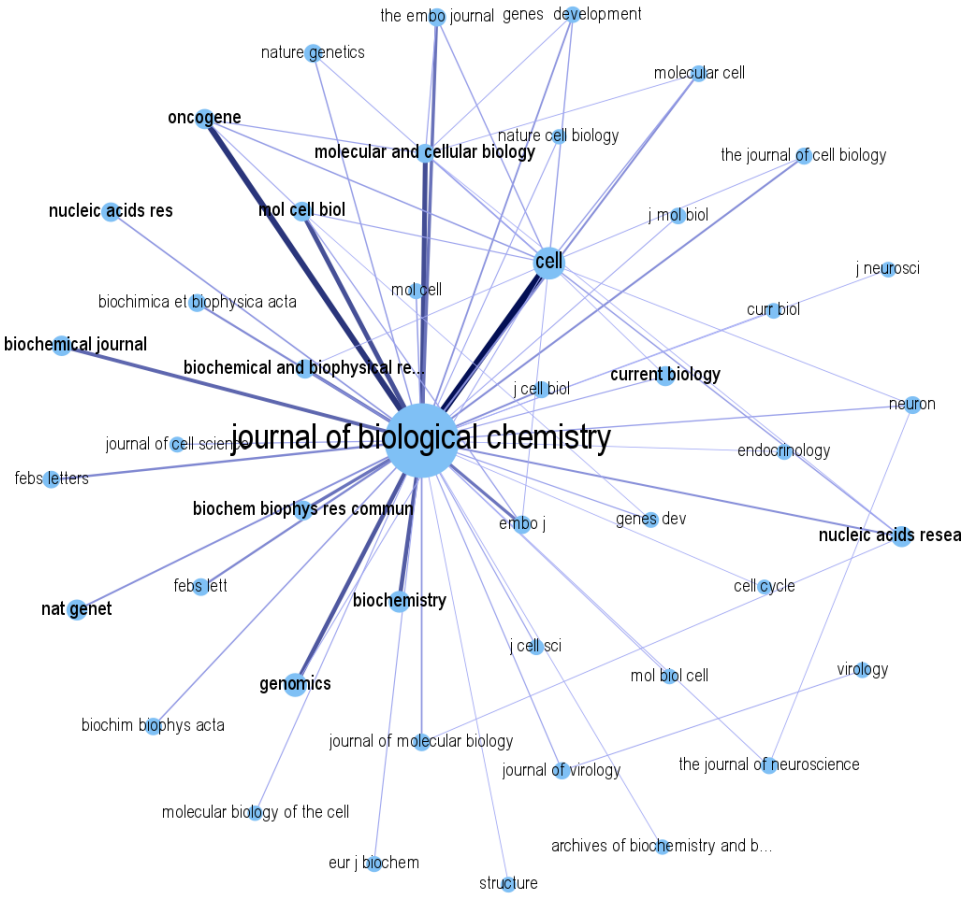


Figure 8: The network graph of Life Sciences area.

same journal. For example, we hypothesize that ”mol cell biol”, ”mol cell”, ”j mol biol”, ”j cell biol”, and ”mol biol cell” are different abbreviations for the journal of Molecular and Cellular Biology.

And last but not least, the Multidisciplinary area (Fig. 9) is made of the main large nodes that we can see in the co-citation network of journals in Figure 4. The sheer size of the number of citation and co-citation in this area is larger compared to the other areas, however, on the other side, the number of multidisciplinary journals are less.

## 6 Limitations

While working on the project, we issues several limitations, which can be divided into three areas:

1. the data set,
2. computational complexity,
3. insufficient information in the original paper.

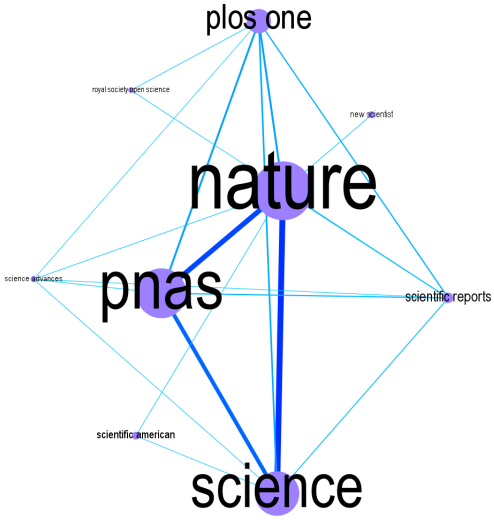


Figure 9: The network graph of Multidisciplinary area.

The major problem that we faced was due to the low quality of citation data representation in the data set that we get from Singh et al. (2021). We already mentioned one of the problems through our project, which is the low quality of journal names. Not only did we have to derive the name from the string that contained other information, but also the names themselves were inconsistent across the rows. For example, sometimes the journal name was given in the full format and sometimes only an abbreviation was presented or some words were shortened. It is as if that part of the data set is written manually by hand. In addition, some journal names include the article, while others did not. In the future, Natural Language Processing (NLP) algorithm could be implemented in the journal names to fix the errors in the data mining process.

Given the huge volume of the data set, it would take an effort to merge the similar strings together. Therefore in data cleaning, we applied only global data cleaning algorithms and did not forget to mention that some information that we present might include minor inaccuracy due to the inconsistent naming.

The second problem was in regard to the assignment of the areas. Since our initial data set did not include the standard journal area code (*ASJC*) information for the journals, we had to retrieve it from another data set and merged it with ours based on the journal names. However, due to the already mentioned inconsistency with the journal name, some journals were not assigned to the area.

The second group of problems is connected with the Pathfinder algorithm. First, the authors of the original paper provided very little information about the Pathfinder algorithm that they used for the network creation, therefore it is a problem to identify the specific algorithm and its parameters. In addition, running the algorithm on our amount of data requires a lot of time and computation power, which in our current situation, we did not have. It resulted in our decision not to use the Pathfinder algorithm for our network at all.

## References


Arroyo-Machado, W., Torres-Salinas, D., Herrera-Viedma, E., & Romero-Frías, E. (2020, 02). Science through wikipedia: A novel representation of open knowledge through co-citation networks. *PLOS ONE*, *15*(2), 1-20. Retrieved from https://doi.org/10.1371/journal.pone.0228713 doi: 10.1371/journal.pone.0228713

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512. Retrieved from https://www.science.org/doi/abs/10.1126/science.286.5439.509 doi: 10.1126/science.286.5439.509

Bellomi, F., & Bonato, R. (2005). Network Analysis for Wikipedia.

(n.d.). Retrieved from https://github.com/stepkurniawan/network-analysis-wikipedia-journals

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology*, *58*(9), 1303-1319. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20614 doi: https://doi.org/10.1002/asi.20614

Quesada, B. (2005). *Visualizacion y analisis de grandes dominios cientıficos mediante redes pathfinder (PFNET)* (Unpublished doctoral dissertation). Universidad de Granada.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*, 581-603.

Singh, H., West, R., & Colavizza, G. (2021). Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia. *MIT Press Direct*, *2*(1), 63–79. Retrieved from https://zenodo.org/record/3940692#.YvOttXZBxPa

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Retrieved from http://doi.wiley.com/10.1002/asi.4630240406

Walker, M. (2020). *Python Data Cleaning Cookbook: Modern techniques and Python tools to detect and remove dirty data and extract key insights*. Packt Publishing.

Zaki, M. J., & Meira, W., Jr. (2014). Data mining and analysis: Fundamental concepts and algorithms. *Cambridge University Press*.

# Appendix

Table 6: Percentage of nodes by area

| Area | Number of nodes | Percentage |
|---|---|---|
| LS | 74 | 31.2% |
| PS | 47 | 19.8% |
| NaN | 28 | 11.8% |
| LS & HS | 25 | 10.5% |
| LS & PS | 18 | 7.6% |
| SSH | 14 | 5.9% |
| HS | 14 | 5.9% |
| Multi | 10 | 4.2% |
| SSH & PS | 2 | 0.8% |
| LS, PS & HS | 1 | 0.4% |
| SSH & HS | 1 | 0.4% |
| LS & SSH | 1 | 0.4% |
| PS & HS | 1 | 0.4% |
| LS, SSH & PS | 1 | 0.4% |
| Total | 237 | 100.0% |

Table 7: The comparison of the degree of the top 10 journals.

| Degree of the journals of the original paper | Degree of the journals for our data set |
|---|---|
| PNAS (251) | Nature (87) |
| Nature (76) | PNAS (78) |
| Journal of Biological Chemistry (41) | Journal of Biological Chemistry (68) |
| Science (33) | Science (61) |
| New England Journal of Medicine (32) | Cretaceous Research (46) |
| Journal of the American Chemical Society (26) | PLOS ONE (39) |
| Physical Review Letters (19) | Journal of Vertebrate Paleontology (33) |
| The Lancet (18) | Scientific Reports (28) |
| Journal of Personality and Social Psychology (17) | Cell (23) |
| PLOS ONE (15) | Journal of Paleontology (23) |

Table 8: The comparison of the betweenness of the top 10 journals.

| **The Betweenness of the journals of the original paper** | **The Betweenness of the journals for our data set** |
|---|---|
| PNAS (0,881) | Nature (0,352) |
| Nature (0,389) | Journal of Biological Chemistry (0,255) |
| Science (0,220) | PNAS (0,220) |
| New England Journal of Medicine (0,214) | Cretaceous Research (0,216) |
| Journal of Biological Chemistry (0,120) | Science (0,155) |
| The Lancet (0,087) | PLOS ONE (0,100) |
| Journal of the American Chemical Society (0,081) | Journal of Vertebrate Paleontology (0,057) |
| PLOS ONE (0,072) | Physical Review Letters (0,052) |
| Physical Review Letters (0,063) | Journal of the American Chemical Society (0,042) |
| Psychological Review (0,059) | The Lancet (0,042) |

Table 9: The comparison of the closeness of the top 10 journals.

| **The Closeness of the journals of the original paper** | **The Closeness of the journals for our data set** |
|---|---|
| PNAS (0,491) | Nature (0,636) |
| Nature (0,413) | PNAS (0,6) |
| Cold Spring Harbor Symposia on Quantitative Biology (0,383) | Science (0,548) |
| Neuroscience and Biobehavioral Reviews (0,367) | PLOS ONE (0,529) |
| Nature Structural and Molecular Biology (0,366) | Journal of Biological Chemistry (0,524) |
| Annual Review of Entomology (0,358) | Cretaceous Research (0,493) |
| New England Journal of Medicine (0,357) | Journal of Vertebrate Paleontology (0,481) |
| Journal of Biological Chemistry (0,344) | Current Biology (0,480) |
| Journal of the National Cancer Institute (0,343) | Scientific Reports (0,475) |
| PLOS ONE (0,338) | Journal of Paleontology (0,469) |

Table 10: The comparison of the eigenvectors of the top 10 journals.

| **The Eigenvector of the journals of the original paper** | **The Eigenvector of the journals for our data set** |
|---|---|
| PNAS (1,0) | Nature (1,0) |
| Nature (0,198) | PNAS (0,969) |
| Journal of Biological Chemistry (0,112) | Science (0,751) |
| New England Journal of Medicine (0,103) | PLOS ONE (0,693) |
| Cold Spring Harbor Symposia on Quantitative Biology (0,083) | Scientific Reports (0,586) |
| Neuroscience and Biobehavioral Reviews (0,081) | Journal of Vertebrate Paleontology (0,582) |
| Nature Structural and Molecular Biology (0,077) | Journal of Biological Chemistry (0,579) |
| PLOS ONE (0,076) | Cretaceous Research (0,564) |
| Molecular and Cellular Neurosciences (0,070) | PeerJ (0,541) |
| Journal of Neuroscience (0,070) | Journal of Paleontology (0,528) |