Citation Network Analysis of Co-Authorship in Google Scholar

A comparative analysis of manually added and automatically generated co-authorship-networks

Project Report
(Type B)

Analyzing Networks
Summer Semester 2022
Leuphana Universität Lüneburg

Submitted by:

Seike

Maria

Evgeniya Zakharova

Submitted to: Prof. Dr. Peter Niemeyer

Lüneburg, 14.08.2022

Table of Contents

List of Figures

List of Tables

Introduction

From funding over recruitment to choosing a thesis supervisor, academic practice regularly gives rise to the question how to measure scholars' scientific impact. In recent years, the analysis of co-authorship networks has gained popularity to address this problem. By placing the formal collaboration between authors at the center of analysis, this perspective accounts for the increasingly interdependent nature of scientific ventures across disciplines and enables new insights into the role that researchers, disciplines, institutions and countries play in the globalizing world of academia (Acedo et al., 2006; Glänzel & Schubert, 2006).

A useful tool for exploring co-authorship networks is Google Scholar, an online search engine for scholarly literature. One of its advantages are the public profiles, where authors can manually add information like publications and co-authors. In contrast to automatically generated lists, which may be subject to the name disambiguation problem[1] or include papers with thousands of "co-authors", the manually added information likely reflects actual cooperation and may, thus, reflect co-authorship relations more accurately (Kalhor et al., 2022).

Kalhor et al. (2022) draw on this functionality in "A new insight to the analysis of co-authorship in Google Scholar" to analyze the Manually Added Co-authorship Network (MACN) regarding community attributes, assortativity coefficients, exponential random graph model coefficients, and structural characteristics. This project seeks to probe their results by applying their measures to an ordinary co-authorship network (OCN). For this purpose, we decided to use the dataset that Chen et al. (2017) generated for "Building and analyzing a global co-authorship network using Google Scholar". Kalhor et al. provided us directly with their data and Chen et al.'s data is publicly available on Github[2]. We jointly implemented all coding from scratch in Python via Google Colab (see Appendix A for links) with the library *NetworkX* for networks.

This report first introduces in more detail the reference paper (section I). Subsequently, we describe the network concepts and data by Kalhor et al. (2022) (section II) and Chen et al. (2017) (section III), as well as how we replicated and visualized their network models. We proceed to report the results from applying Kalhor et al.'s (2022) key measures to our implementations of the MACN and the OCN (section IV). In the discussion, we compare our results and address our project's major limitations (section V). The final section concludes.

[1] Name disambiguation problem applies when searching publications by a specific author by name, which bears the risk of accidentally including articles published by different authors with the same name (Chen at al., 2017).

[2] https://github.com/chenyang03/co-authorship-network

I Story of the Reference Paper

Google Scholar is one of the most convenient search engines for scholarly literature. It encompasses millions of books, publications, and research papers by authors from all over the world. Some users maintain public profiles where they add personal information such as name, age, country, institution and career data like the number of publications, citations and h-index[3]. Special algorithms generate continuous and automatic updates including the suggestion of a list of potential co-authors to users. An author's name will be added to the list once they were mentioned in one of the user's publications. Users may select the co-authors from this list and add them to their profiles manually (Kalhor et al., 2022).

Kalhor et al. (2022) claim that all previous works focused on automatically generated data, where co-authorship is assumed whenever two names appear in the author list of the same paper. A major limitation of this approach is that those authors may never have worked with each other directly and do not even acknowledge each other as co-authors. To address this concern, Kalhor et al. (2022) analyzed the manually added co-authorship network (MACN) instead, which is based on the collaborators that users consciously add to their profile. With this approach, the authors do not only seek a more accurate representation of scientific collaboration but are also able to engage with a number of new intriguing questions:

1. What is the motivation of the manually added co-authorship?
2. Who adds more collaborations?
3. Whom are users willing to include?

To address these questions, Kalhor et al. (2022) calculated the following network metrics for the MACN: edge reciprocity, modularity, average clustering, assortativity and rich-club coefficients, betweenness, closeness, weighted degree, eigenvector centralities, and PageRank. Moreover, they implemented exponential random graph models (ERGM) to include and control for node attributes, structural configurations, and edge attributes.

Kalhor et al. (2022) additionally analyzed the Field of interest network (FIN) and the Affiliated institute network (AIN). However, we decided not to include these network models in our project to be able to engage in depth with the MACN, which is at the core of the paper's analysis. Moreover, the datasets for the FIN and AIN were not available.

[3] H-index refers to an author's number of publications h that have received at least h citations (Hirsch, 2005)

II Network Model of the Reference Paper

2.1 Network Data and Concept

Kalhor et al. (2022) kindly provided us directly with their dataset when we reached out to them via email. They collected the data with a web crawler developed in Python. Up until March 31, 2021, they crawled the Google Scholar profiles of the most cited researchers from universities in different countries, collected information about their co-authors and enriched the dataset with demographic information as predicted by the NameToGAN tool and standard fields of interest. As we received an anonymized version of the data, authors are represented by user IDs and identificatory features such as author URL and institute URL that are mentioned in the paper are not included. The data is organized in three files. The main dataset *Co-authorship* contains the network data for the MACN, where authors correspond to nodes and each manually added co-author relationship is a directed, unweighted link. Accordingly, the dataset maps a source to a target user ID to the extent that the former has manually added the latter as co-author on Google Scholar. The second dataset *Users fields* contains the user IDs and corresponding standard fields of interest. The third and final dataset *Users attributes* matches user IDs with other attributes including university, citation count, h-index, country and gender.

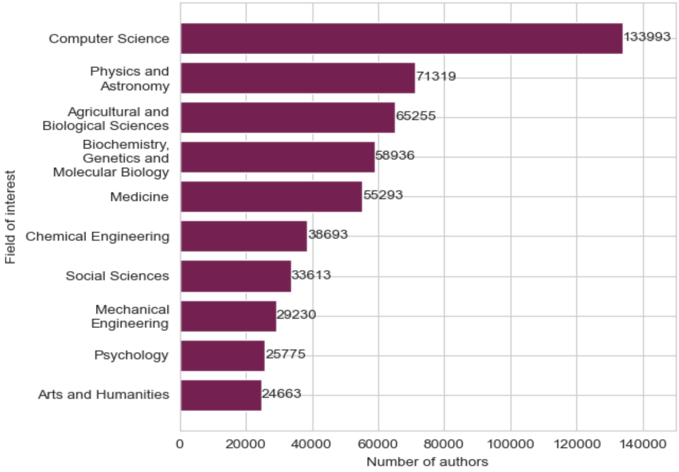
2.2 Exploratory Data Analysis

To gain an initial understanding of the data, this section presents a brief exploratory data analysis. The *Co-authorship* dataset has one column for the source user ID (data type: string/mixed) and one for the target user ID (string/mixed). Each of the 307,364 rows represents one manually added co-authorship relation (i.e., link) from source author to target author. The data is made up of 134,113 unique user IDs (i.e., nodes). On average, each user appears 2.292 times as a source and the three most common sources are repeated 34, 32 and 29 times. They are all male and associated with Computer Science as a standard field of interest but stem from different universities in different countries (Canada, Germany, Turkey). Remarkably, the same three authors are the most frequent targets and upon further exploration of the data, it becomes apparent that each author shows up the same number of times as source and target. This symmetry, suggesting that all source-target relationships are repeated in the other direction, is unexpected given that the network is directed. However, we decided not to modify the dataset to avoid distorting our replication of Kalhor et al.'s (2022) calculations (section IV).

The *Users fields* dataset comprises the user ID (string/mixed) of 850,827 authors and the standard field of interest (string/mixed) for 808,737 of them. Accordingly, 95% of authors have been assigned one out of 40 unique fields of interest. Out of those, about one-third are associated with one of the three most popular fields, namely Computer Science (133,993 users), Physics and Astronomy (71,319 users) and Agricultural and Biological Science (65,255 users) (figure 1). The fields of interest in the dataset match those listed by Kalhor et al. (2022), except for “Engineering (miscellaneous)”, which counts 2,902 authors in the dataset but is not explicitly mentioned in the paper.

Figure 1

Most frequent fields of interst in the MACN

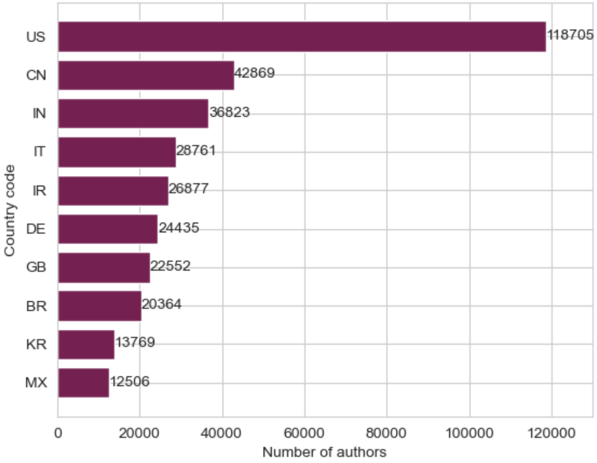


Note. Based on replication of the MACN by Kalhor et al. (2022).

Finally, the *Users attributes* dataset contains the user ID (string/mixed), university ID (unsigned integer), citation count (integer), h-index (integer), country (string/mixed) and gender (string/mixed) for 556,932 authors. While Kalhor et al. (2022) started with a list of 65 universities, the final dataset, for which they additionally considered the co-authors of the researchers from the initial institutions, contains 5,752 unique universities from 226 countries. The United States, China and India top the list with 118,705, 42,869 and 36,823 (figure 2) users respectively, in sum reflecting 36% of all users. On average, the dataset includes 97 authors per university, the most frequently occurring university is associated with 4,998 authors. The citation counts vary between

Figure 3

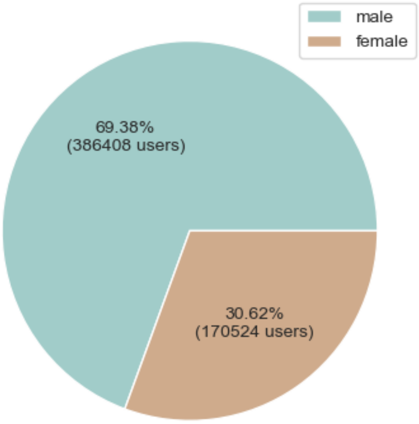
Most frequent countries in the MACN



Note. Based on replication of the MACN by Kalhor et al. (2022).

Figure 2

Gender distribution in the MACN



Note. Based on replication of the MACN by Kalhor et al. (2022).

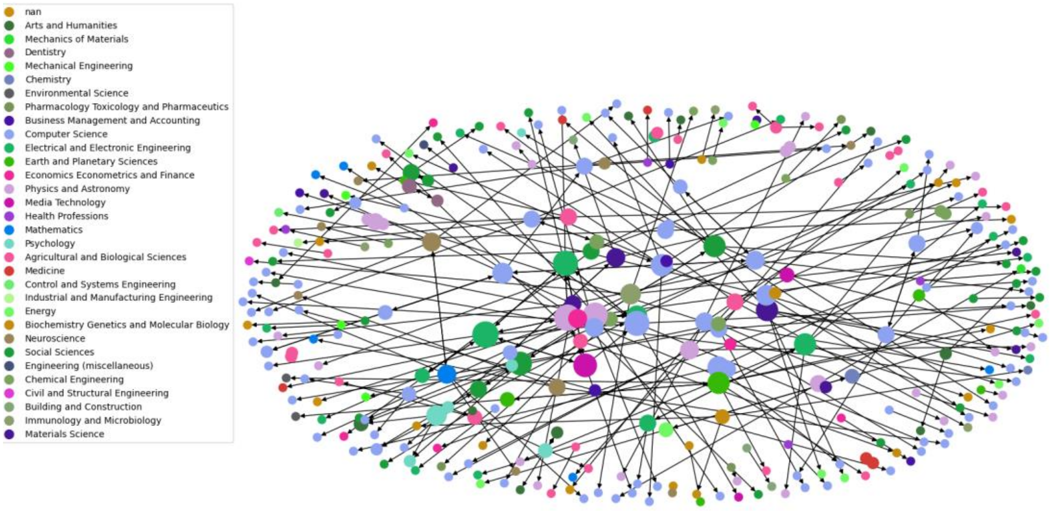
1 and 1,092,925 with a mean of 3,280 and the h-index ranges from 1 to 314 with a mean of 18.913. Moreover, the majority of included authors (69%) are male (figure 3).

2.3 Network Model

We defined the elements of the MACN, where the nodes represent user IDs and links manually added co-authorship relations between users. The MACN is directed from source to target ID and unweighted. Relevant node features include the respective author's field of interest, university, citation count, h-index, country and gender. During our process, we encountered some limitations in the visualization of the network due to a lack of computational resources, for this reason, we decided to reduce the dataset size. The calculations were applied to the entire data (section IV), but for the visualization we sliced the first 300 rows, which is 0.1% of the initial *Co-authorship* dataset and contains 126 unique user IDs. The biggest node degree is 10. A sample graph of the MACN is presented in figure 4. The graph consists of 300 links in order to see the general picture of the MACN.

Figure 4

Subgraph of the MACN


nan
Arts and Humanities
Mechanics of Materials
Dentistry
Mechanical Engineering
Chemistry
Environmental Science
Pharmacology Toxicology and Pharmaceutics
Business Management and Accounting
Computer Science
Electrical and Electronic Engineering
Earth and Planetary Sciences
Economics Econometrics and Finance
Physics and Astronomy
Media Technology
Health Professions
Mathematics
Psychology
Agricultural and Biological Sciences
Medicine
Control and Systems Engineering
Industrial and Manufacturing Engineering
Energy
Biochemistry Genetics and Molecular Biology
Neuroscience
Social Sciences
Engineering (miscellaneous)
Chemical Engineering
Civil and Structural Engineering
Building and Construction
Immunology and Microbiology
Materials Science

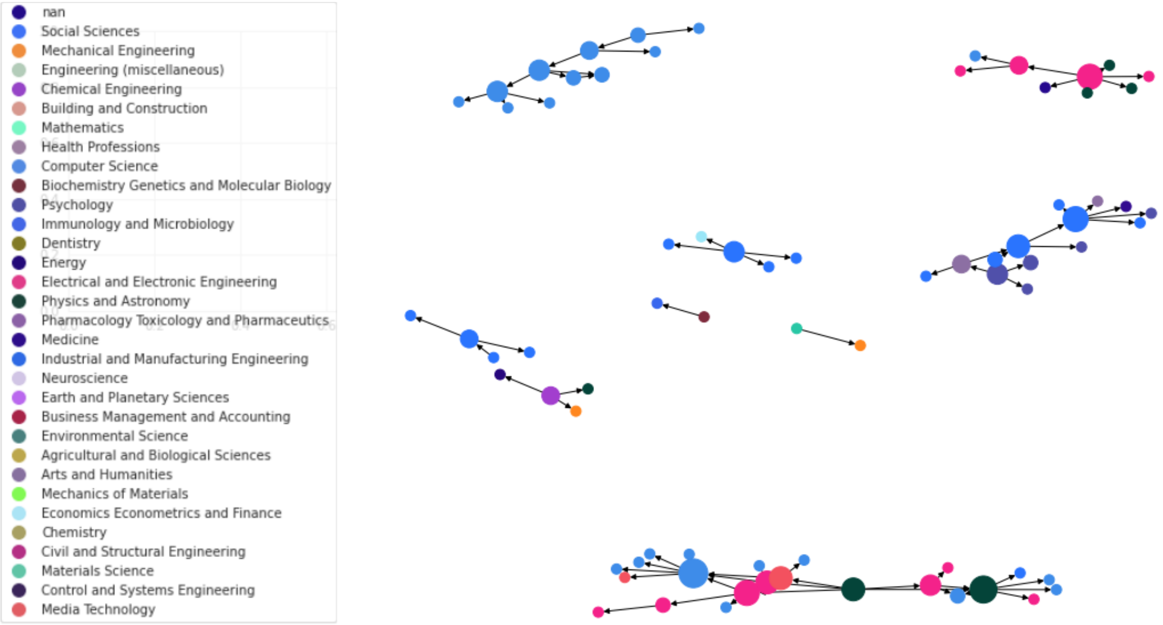
Note. Digraph (directed): the arrows point to the authors who were added. The size of the node is proportional to the node degree. The color of the node reflects the field of study. The default layout was used.

To build this graph the nodes and links were taken from the *Co-authorship* dataset together with the direction of arrows. The users' fields of interest are represented through colors and the size of each node is proportional to its degree. We applied two different layouts (default - figure 4 and spring - figure 5) to understand the behavior of the network.

By considering the number of links, we can see that the dominant field of interest is Computer Science, which is in line with Kalhor et al.'s (2022) observations. In this subgraph of the MACN, there are a lot of connections between only two authors, and only some have a big number of co-authors. We suppose that the graph visualization of the whole dataset will have the same structure.

Figure 5

Graph of the reduced MACN's components


nan
Social Sciences
Mechanical Engineering
Engineering (miscellaneous)
Chemical Engineering
Building and Construction
Mathematics
Health Professions
Computer Science
Biochemistry Genetics and Molecular Biology
Psychology
Immunology and Microbiology
Dentistry
Energy
Electrical and Electronic Engineering
Physics and Astronomy
Pharmacology Toxicology and Pharmaceutics
Medicine
Industrial and Manufacturing Engineering
Neuroscience
Earth and Planetary Sciences
Business Management and Accounting
Environmental Science
Agricultural and Biological Sciences
Arts and Humanities
Mechanics of Materials
Economics Econometrics and Finance
Chemistry
Civil and Structural Engineering
Materials Science
Control and Systems Engineering
Media Technology

Note. Several enlarged directed components. Colors define the field of interest. Spring layout was used.

Additionally, we applied the spring layout in order to change the structure of the visualized graph and elaborate on it with further analysis. It shows more information about the connections between authors regarding the field of interest. It is likely to see a trend to collaborate with authors from the same subject area or close to it. We zoomed in on some components of the MACN to have a detailed representation of them (figure 5). The components contain several nodes from 2 to more than 20.

III Extension

3.1 Motivation

As for the extension, the intuitive way to continue our group's research was to accomplish a comparative analysis of the MACN by Kalhor et al. (2022) and the ordinary co-authorship network (OCN) by Chen et al. (2017), which contains the data of the automatically generated lists of potential co-authors from Google Scholar profiles. Kalhor et al. (2022) refer to Chen et al.'s (2017) paper both as a relevant study on Google Scholar and for results comparison. However, the comparison only considers two metrics (clustering coefficient and average shortest path length) to arrive at the conclusion that the MACN and OCN are structurally different to the extent that the former is less connected than the latter. Thus, we decided to extend the comparison to the other network measures implemented by Kalhor et al. (2022) to explore further how the choice of network model may affect our view on one and the same phenomenon. Moreover, we were interested in investigating how relevant Kalhor et al.'s (2022) new and innovative approach is to the more traditional approaches to co-authorship network analysis.

3.2 Added Dataset

We obtained the data used by Chen et al. (2017) from Github where one of the authors, Yang Chen, has made it publicly available on his profile[4]. The authors collected their data by crawling and analyzing Google Scholar profiles along with filtering out inaccurate ones. The unique user IDs were extracted on May 29, 2015. Later the relevant information was gathered by parsing HTML source code of authors' profiles, including "name, title and affiliation, list of scientific labels, email domain, homepage URL, with photo or not, citation metrics (total citations, h-index, i10-index, and these values in the last five years), number of citations of each year, each article's information with its number of citations, and self-claimed co-author list" (Chen et al., 2017). One of the problems with the automatically generated publications is that Google Scholar only considers the first letter of an author's first name and the last name. Thus, some publications are inaccurately added to the profile of a user because s/he happens to have the same first initial and last name as one of the actual authors. Therefore, Chen et al. (2017) picked only those authors where there was a possibility to verify 95% of their list of publications. To build the network, all the publications of each author were scanned and links were formed if and only if two authors have the same publication.

[4] https://github.com/chenyang03/co-authorship-network

After cleaning and preprocessing the data, Chen et al. (2017) generated two datasets. The dataset *Edges* contains the network data and maps a source to a target author. Authors correspond to the nodes of the OCN and links reflect the presence of a co-authorship relation between the source and the target author, which occurs when the authors share an automatically added publication on their Google Scholar profile. Thus, the links are unweighted and, in contrast to the network concept of Kalhor et al. (2022), undirected. The second dataset *Info* links author ID with total number of citations, h-index, g-index, academic title and academic fields.

The *Edges* dataset is made up of 1,234,019 rows, each of which reflects a unique co-authorship relationship between the authors in the “source” and “target”-column (data type: string/mixed). In total, the dataset contains 227,074 unique authors that appear between 1 and 463 times with a mean of 9 appearances. The *Info* dataset maps author IDs (integer) to total number of citations (integer), h-index (integer), g-index (integer), academic title (integer; 3-professors / 2-postdocs / 1-students / 0-unknown) and three academic fields (integer; 1-author associated with field / 0-author not associated with field), namely computer science, biology, and sociology. One author can belong to one, multiple or no fields. The dataset contains 402,392 unique users and no missing values. The most common field is computer science (82,966 authors / 21%) followed by biology (82,361 authors / 20%) and sociology (1,602 authors / 0.4%). The citation counts vary between 0 and 230,238 with a mean of 868 and the h-index ranges from 0 to 212 with a mean of 8.335.

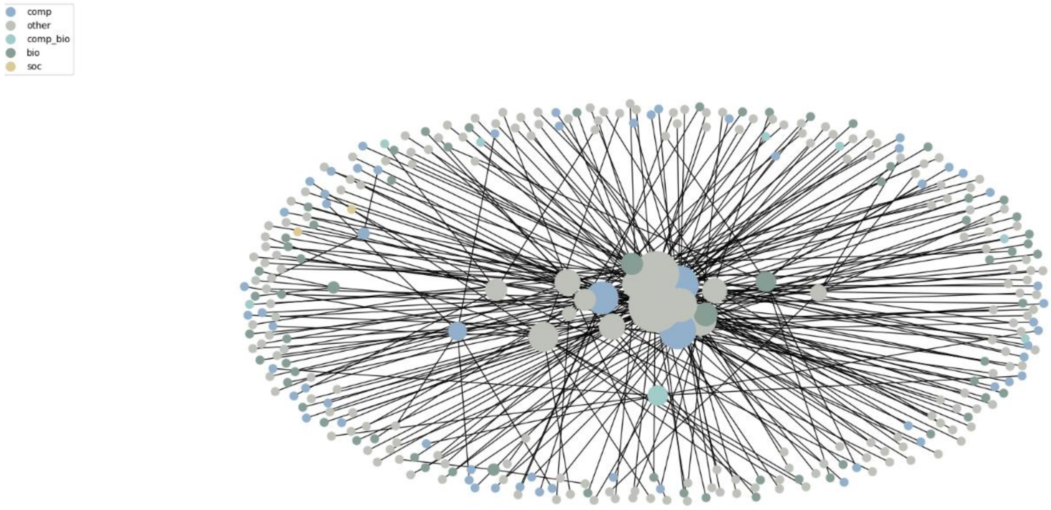
3.4 Network Model

The nodes of the OCN represent authors and the links the co-authorship relations between users. The network is undirected and unweighted. Relevant node features include citation count, h-index, g-index, academic title and academic field.

We decided to reduce the dataset size because a lack of computational resources caused limitations for the visualization of the network. The methods were also calculated only on a subset of the data. In the end, the dataset was sliced to the first 300 rows, which is 0.024 % of the initial *Edges* dataset and contains 30 unique source IDs. The biggest node degree is 43. A sample graph is presented in figure 6.

Figure 6

Subgraph of the OCN



Note. Undirected and unweighted network. The size of the node is proportional to the node degree. The color of the node reflects the field of study. Default layout was used.

To ensure comparability, we chose the same attributes for this graph as we did for the graph of MACN. We modified the dataset for the field of interest so that it presents the same format as Kalhor et al.'s (2022). Therefore, each user has been assigned one field of interest according to the following pattern:

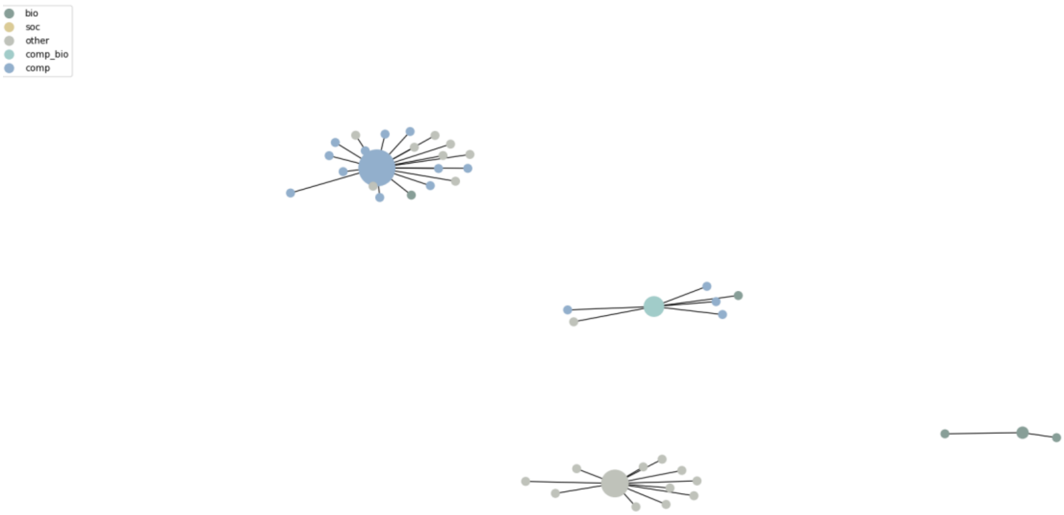
- computer science, biology or sociology if the original dataset identifies a unique field for the author
- any combination of the above disciplines if the original dataset assigned multiple disciplines to one author (e.g., "comp_bio", "comp_soc",...)
- "other" if all fields were "0" in the original dataset

From the graph we can conclude that the combination of zeros is the most frequent field, which is equivalent to the field "other". Some nodes have a large degree. This can be attributed to working with a reduced dataset, which includes only a few authors and their lists of co-authors.

The spring layout was implemented. Figure 7 displays some enlarged components of the OCN. The components demonstrate the connections between authors from the same field of interest with some other areas. Unfortunately, it is not possible to examine the graph accurately because the exact subjects are not mentioned in the dataset, except only three and their combinations.

Figure 7

Graph of the reduced OCN's components



Note. Several enlarged undirected components. Colors define the field of interest. Spring layout was used.

IV Results

In this section, we will present the different metrics calculated in the MACN and the OCN. The following metrics are the same ones included in Kalhor et al.'s (2022) paper, some metrics were excluded due to their complexity.

4.1 Reference Paper (Kalhor et al., 2022)

In addition to plotting a subgraph of the MACN (section 2.3, figure 4), we replicated the measures conducted by Kalhor et al. (2022) on our implementation of the whole network. We used *NetworkX* for building the MACN from the dataset provided by the authors. In total, we had 134,113 nodes and 307,364 directed, unweighted links in the MACN. Then, we added the different attributes included in the datasets to the nodes of the network, namely: field of interest, university, citations count, h-index, country, and gender. Additionally, we calculated the degree of each node and added it as an attribute called "node degree". The highest node degree is 68, which is in line with our exploratory data analysis that found that the most popular author appears 34 times as source and 34 times as target (section 2.2).

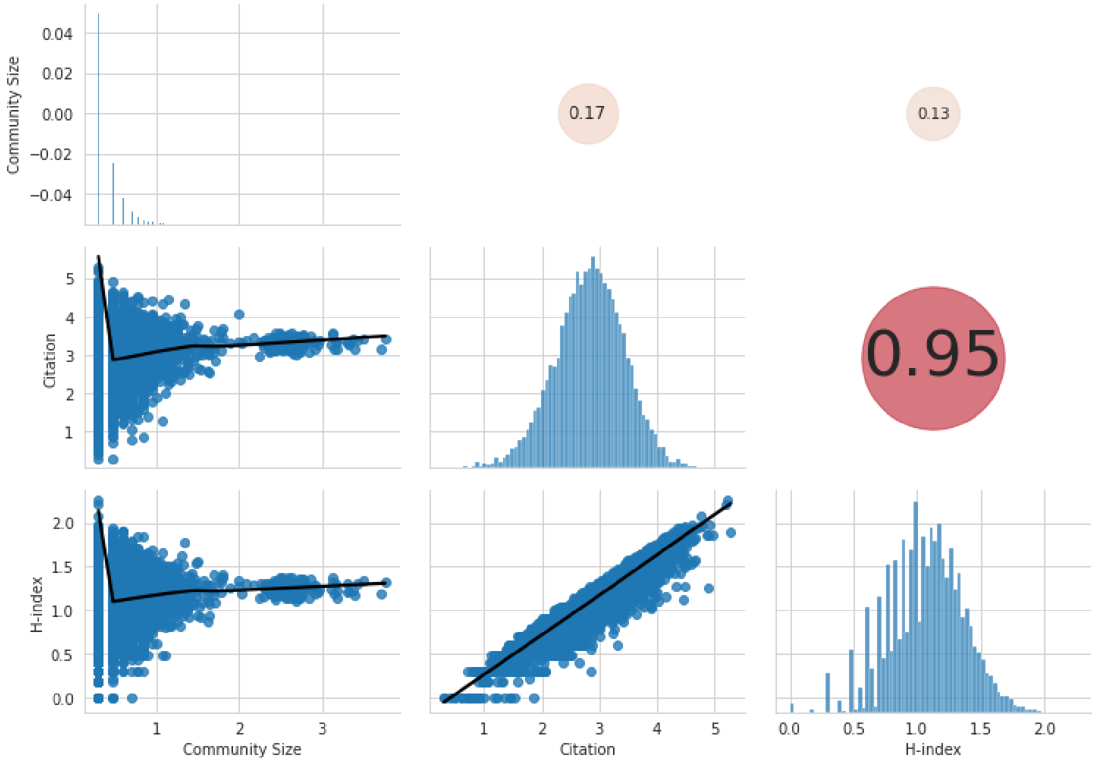
4.1.2 Community Analysis

The community analysis involved finding the different communities in the MACN and doing calculations. The paper originally used the *infomap* algorithm for this purpose. However, as the algorithm is complex and not included in the *NetworkX* package, we decided to use one of the *community algorithms* included in *NetworkX*. More specifically, we applied *greedy_modularity_communities*, which uses Clauset-Newman-Moorse greedy modularity maximization to find the community partition with the largest modularity (Hagberg et al., 2008). The implementation of this method was simpler than *infomap* and still successful in finding the communities of MACN. In total, there were 15,356 communities identified in the MACN.

Turning to the community calculations, the paper analyzed the correlation using Pearson correlation on the attributes of the communities such as size, mean citation count and mean h-index. In our experiment, we also computed the sizes of the communities, the mean citation count and the mean h-index. We plotted the correlations of these attributes in a grid plot as shown on figure 8.

Figure 8

Bivariate relationships between MACN's community attributes



The graph shows the bivariate relationship between communities' attributes similar to the one presented by Kalhor et al. (2022). There is a strong positive linear relationship between mean citation count and mean h-index of communities, which confirms the results in the paper. Equally in line with the authors' findings, the correlation between community size and other variables is weak.

4.1.3 Assortativity Coefficient

According to Kalhor et al. (2022), assortativity is one of the important metrics calculated on social networks. Assortativity measures the similarity of connections in the graph with respect to the given attribute (Hagberg et al., 2008). We calculated the assortativity coefficients for the same attributes as the ones calculated by Kalhor et al. (2022). Our assortativity coefficient results are presented in table 1.

Table 1

Assortativity coefficients for different attributes in the MACN

Attribute	Assortativity coefficient		Difference
	Our experiment	Kalhor et al.	
Field of interest	0.393	0.525	-25%
h-index	0.038	0.224	-83%
Country	0.200	0.220	-9%
Institute	0.184	0.193	-5%
Citation count	0.029	0.162	-82%
Gender	0.104	0.095	9%
Node degree	0.042	0.088	-52%

In general, our results were relatively close to Kalhor et al. (2022), except for the h-index and citation count that are both noticeably lower in our experiment. However, we do not know the reason for this difference. The assortativity coefficients were positive for all attributes, which indicates a tendency between nodes (i.e., authors) with a similar value for that attribute to make a connection (i.e., add each other as a co-author). The highest assortativity coefficient is for the field of interest, which can deduce that authors with the same field of interest tend to collaborate more.

4.1.4 ERGM Coefficients

Kalhor et al. (2022) made an exponential random graph model (ERGM) based on the attributes of the nodes of the MACN to find the features that are statistically significant in forming the network links. In our experiment, we tried coding an ERGM, however it required the implementation of Markov Chain Monte Carlo sampling techniques and probability models, which ultimately turned out to be too complex and time-consuming. For these reasons, we did not calculate the ERGM coefficients.

4.1.5 Structural Metrics of MACN

There are several structural metrics that were performed on the MACN. Like Kalhor et al. (2022), we focused on edge reciprocity, average clustering coefficient and average shortest path length. Firstly, the edge reciprocity of a directed network indicates the ratio of the number of links pointing in both directions to the total number of links in the network (Costa et al., 2007). Kalhor et al.'s (2022) paper indicates that only 31.10% of the links in the MACN are bidirectional. In our experiment, we got 100% reciprocity in our MACN, this result supports our finding in the exploratory data analysis (section 2.2). We cannot fully trust the dataset provided by the authors due to this large difference. Additionally, we calculated the edge reciprocity on the sample subgraph we built for illustrating our MACN, where we got edge reciprocity of 28.67%, which is closer to the paper and confirms Kalhor et al.'s (2022) finding that many authors do not tend to add all of their co-authors to their Google Scholar profile.

Secondly, Kalhor et al. (2022) calculated a clustering coefficient of 0.188 and an average shortest path length of 6.384 for only the largest weakly connected component of the MACN. In our experiment, we attempted to perform these calculations. We created a subgraph with the largest weakly connected component and calculated the clustering coefficient of 0.160, which is close to the paper's result. It seems that there are weak ties in this network, meaning that the probability of the co-authors of a user collaborating is low.

Lastly, for the average shortest path length, the subgraph of the largest weakly connected component was too large with 84,533 nodes to calculate the path in Google Colab. For this reason, we decided to calculate the average shortest path length from a specific source node. We chose the "yAXdVbMAAAAJ" as it is the most frequent author in the dataset according to our exploratory data analysis. We calculated the average of all the paths from the source node to other nodes, which resulted in 11.354 steps from the source node to a random target node.

4.1.6 Centrality Metrics of MACN

Kalhor et al. (2022) did not measure centrality for the MACN but for the Fields of Interest Network (FIN), which is of a smaller scale. However, we do not have the FIN dataset and decided to calculate centrality metrics in the MACN. Our results show the same three authors in the top five for degree centrality and PageRank, which are: “yAXdVbMAAAAJ”, “TxKNCSoAAAAJ”, and “QY-earAAAAAJ”. According to these results, PageRank works almost the same as the degree centrality in this network.

Additionally, we calculated the betweenness centrality and closeness centrality in the subgraph of the MACN which was used for the illustration because these metrics use paths for their calculations. In this case, our results show two other authors in the top five for these centrality metrics, which are: “S8mtOroAAAAJ” and “FOfDj5gAAAAJ”. Given that we only used a subgraph, we can unfortunately not compare these results with Kalhor et al. (2022).

4.2 Extension (Chen et al., 2017)

Chen et al. (2017) built a global co-authorship network to study the collaboration among authors and performed social network analysis, which involves calculating different metrics. For our extension, we plotted a subgraph of the OCN (figure 6, section 3.4). Additionally, we calculated the same metrics as the MACN for Chen et al.’s (2017) dataset. The results are shown below.

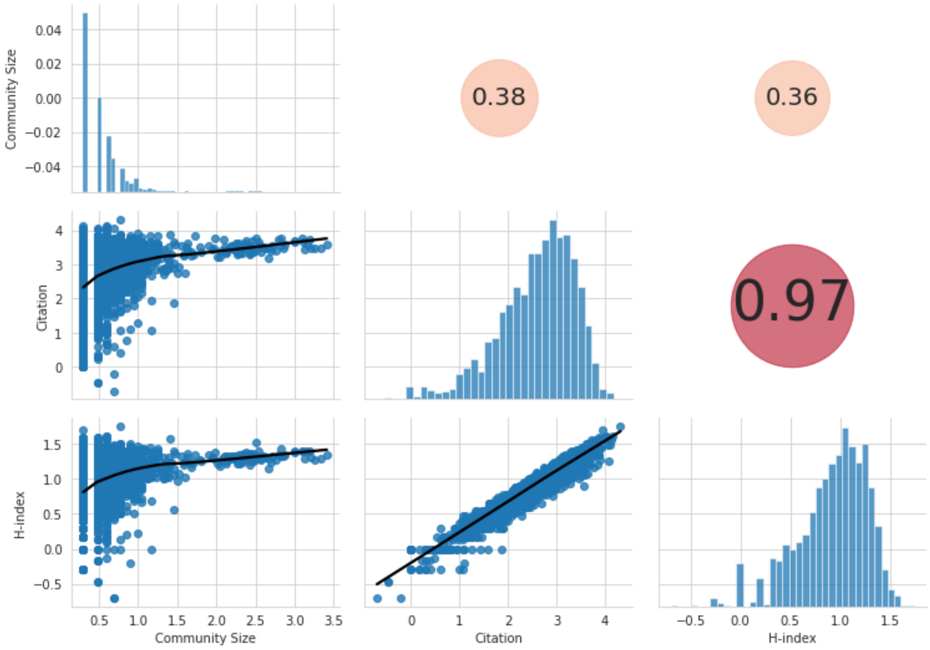
We also used *NetworkX* for building the OCN. In total, the dataset had 277,074 nodes and 1,234,019 undirected and unweighted links. Due to OCN’s large scale, we decided to perform the calculations on only a subgraph of OCN with 44,723 nodes and 50,000 links. Subsequently, we added the different attributes available in the datasets to the nodes of the network. The relevant attributes included for the nodes were: field of interest, citation count, and h-index. Moreover, we calculated the degree of each node and added it as an attribute called “node degree”. The largest node degree for an individual node was 158.

4.2.2 Community Analysis

The community analysis of the subgraph of OCN was performed with the same algorithm as MACN. In total, there were 2,468 communities identified in the subgraph of OCN. For the calculations, following the same correlation method and similar attributes as the MACN, the bivariate correlations of the community size, mean citation count and mean h-index are shown in figure 9.

Figure 9

Bivariate relationships between OCN's community attributes



The graph shows the bivariate relationship between communities' attributes of OCN's subgraph, which is very similar to the one we presented for the MACN. There is a strong positive linear relationship between mean citation count and mean h-index communities. However, the correlation between the community size and the rest of the variables is weak.

4.2.3 Assortativity Coefficient

We calculated the assortativity coefficients for the same attributes as the ones calculated for the MACN. The assortativity coefficient results are on table 2.

All the assortativity coefficients for the OCN's subgraph are very different to the MACN. The assortativity coefficients were positive for all attributes except node degree, which indicates a tendency between nodes (i.e., authors) with a similar value for that attribute to make a connection (i.e., collaborate as co-authors). The highest assortativity coefficient is for the field of interest, which suggests that authors with the same field of interest tend to collaborate more.

Table 2

Assortativity coefficients for different attributes in the OCN

Attribute	Assortativity coefficient (Our calculations)		Difference
	Chen et al. (OCN subgraph)	Kalhor et al. (MACN)	
Field of interest	0.417	0.525	-21%
h-index	0.014	0.224	-94%
Citation count	0.005	0.162	-97%
Node degree	-0.119	0.088	-235%

4.2.4 Structural Metrics of OCN

The main difference between the OCN and the MACN is that the OCN is undirected while the MACN is directed. For this reason, some metric calculations could not be performed. The structural metrics were performed on the largest connected component of the OCN´s subgraph with 35,880 nodes. The result for the clustering coefficient on the largest connected component is 0.038, which is much lower than our result for the MACN with 0.160. Therefore, it seems that there are weak ties in this network also, meaning that the probability of the co-authors of a user collaborating is low.

The other metric is the average shortest path length. Because the largest connected component is too large to calculate the path in Google Colab, we decided to calculate the average shortest path length from a specific node. We chose node "455" as it is the most frequent author in the dataset according to our exploratory data analysis. We calculated the average of all the paths from this node and it resulted in 6.833 steps on average from the node "455" to a random node.

4.2.5 Centrality Metrics of OCN

The centrality metrics of the subgraph of the OCN are not very reliable as only one same author is presented in the top five for degree centrality and PageRank, which is author “5544”. Additionally, we calculated the betweenness centrality and closeness centrality in the subgraph of the OCN which was used for the illustration because these metrics use paths for their calculations. In this case, our results are inconclusive as the betweenness is zero for all the top five authors and the closeness centrality shows the same value for the top five authors.

V Discussion

In this section, we will point out the main differences and similarities encountered in our experiments with the MACN and OCN. This comparison analyzes the visualizations and metrics calculations implemented by us and not the original papers. The main difference between the MACN and OCN is that the MACN is directed while OCN is undirected. Both are unweighted.

The subgraphs of the MACN and OCN with the reduced data show the same structure and visual characteristics. We conclude that the subgraph of the MACN portrays the whole network better than OCN’s subgraph, because the OCN has bigger node degrees than the MACN. Some authors in the OCN have a large number of co-authors. This can be explained by the fact that the authors were added automatically and not manually as in the MACN. One of the similarities we encountered is that the authors from the same or adjacent fields of interest are collaborating more with each other. Also, there are connected and unconnected components in both networks, which means that they are not fully connected.

In comparing the metrics calculated between MACN and OCN, the main similarity was in the correlation plot, which shows a strong positive correlation between mean citation and mean h-index. They also presented several differences. For example, the largest node degree was 68 in the MACN and 158 in the OCN, which indicates that authors tend to add fewer co-authors manually. As shown in the result section, the assortativity coefficients differ greatly in both networks, however, the assortativity coefficient of the field of interest is the largest one in both networks, which can indicate that authors with the same field of interest tend to collaborate more. On the structural metrics, both networks have low clustering coefficients which indicates weak ties in their networks meaning that the probability of the co-authors of a user collaborating is low. Unfortunately, the centrality metrics could not be compared.

The main limitations we encountered in our project was that the *Co-authorship* dataset from Kalhor et al. (2022) duplicated source-target as target-source relationship, which can explain some of the differences found in the metrics compared to the results in the paper. Additionally, we had to use subsets of the whole datasets in most cases for the visualization and metric calculations because the dataset was too large for our platform to handle. For the whole datasets, we encountered RAM overload and very long running times. The fact that we conducted the calculations on differently sized datasets also limits the comparability of results. Lastly, we had to dispense with the implementation of the ERGM because it would have been disproportionately time-consuming for the scope of the present project, which means that we could not investigate further the attributes that are statistically significant in forming the network links.

For future research, it would be interesting to dive deeper into analyzing the fields of interest as an attribute for visualizing graphs. Also, the graphs can be extended by including other attributes from the users' information. Additionally, doing research on new algorithms or methods to calculate path lengths on large datasets would be beneficial to ease the analysis of this and other types of networks.

Conclusion

This project was conducted to analyze and compare two network representations of co-authorship based on data from the platform Google Scholar. On the one hand, we considered the approach by Kalhor et al. (2022) who based their analysis on the collaborations that authors can add manually to their Google Scholar profile. On the other hand, we extended the investigation by applying Kalhor et al.'s (2022) measurements to Chen et al.'s (2017) co-authorship network that automatically infers co-authorship from the author lists in the publications that users have added on their profile.

For each network, we conducted an exploratory analysis of the data and visualized subgraphs to gain an initial understanding of the models. This preliminary analysis revealed that both networks are unweighted, and the data is structured similarly with separate files for the network data, mapping source to target users, and additional node information. However, one key difference is that the MACN is directed whereas the OCN is undirected. Furthermore, we computed several community analysis metrics, assortativity coefficients, structural metrics and centrality measures for both networks. A comparison of the results yielded that the networks are similar in many respects including the fact that the largest assortativity coefficient of field of interest is the largest (i.e., authors are more likely to collaborate with authors from the same or adjacent fields) and the low clustering coefficients (i.e., the probability that the co-authors of a user collaborate is low). Moreover, a strong positive correlation between mean citation count and h-index of communities was present in both networks. However, a comparison of node degrees shows that authors tend to add fewer collaborators manually.

In conclusion, this project has contributed a comparison of the insights we can gain about one and the same phenomena using two distinct network concepts and datasets. In the spirit of Brandes et al. (2013), who argued that “the essential point must be that the abstraction into a network is helpful to scientific inferences”, it has shown that such a comparison may contribute to knowledge development in diverse ways. It may provide additional support for certain observations because the results support each other or, conversely, it may reveal previously invisible aspects of a phenomenon, like the fact that manually generated co-authorship lists may represent only a fraction of an author’s actual formal collaboration.

References

Acedo, F. J., Barroso, C., Casanueva, C., & Galan, J. L. (2006). Co-authorship in management and organizational studies: An empirical and network analysis. *The Journal of Management Studies*, *43*(5), 957–983. https://doi.org/10.1111/j.1467-6486.2006.00625.x

Brandes, U., Robins, G., McCranie, A., & Wasserman, S. (2013). What is network science? *Network Science*, *1*(1), 1–15.https://doi.org/10.1017/nws.2013.2

Chen, Y., Ding, C., Hu, J., Chen, R., Hui, P., & Fu, X. (2017). Building and analyzing a global co-authorship network using Google scholar data. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*.

Costa, L. F., Rodrigues, F. A. , Travieso, G., Villas Boas, P. R. (2007). Characterization of complex networks: a survey of measurements. Adv Phys 56(1):167–242

Glänzel, W., & Schubert, A. (2006). Analyzing scientific networks through co-authorship. In *Handbook of Quantitative Science and Technology Research* (pp. 257–276). Kluwer Academic Publishers.

Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX*. https://networkx.org/documentation/stable/index.html

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46). https://doi.org/10.1073/pnas.0507655102

Kalhor, G., Asadi Sarijalou, A., Sharifi Sadr, N., & Bahrak, B. (2022). A new insight to the analysis of co-authorship in Google Scholar. *Applied Network Science*, *7*(1). https://doi.org/10.1007/s41109-022-00460-4

Appendix: Links to Google Colab Codes

Link to code for the replication of the reference paper (Kalhor et al., 2022):

https://colab.research.google.com/drive/1A9BZ-gblZpQTxz0CV8t348foOK6ssPE2?usp=sharing

Link to code for the replication of the extension (Chen, 2017):

https://colab.research.google.com/drive/1JnrbK6Sl56h1cXpC-Cc8x0BDj468vmWm?usp=sharing