IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 10, No. 4, December 2021, pp. 1103~1114
ISSN: 2252-8938, DOI: 10.11591/ijai.v10.i4.pp1103-1114

Measuring scientific collaboration in co-authorship networks

Karam H. Thanoon[1], Nagham A. Sultan[2], Basim Mahmood[3], Dheyaa S. Kadhim[4]
[1]Department of Software, University of Mosul, Mosul, Iraq
[2,3]Department of Computer Science, University of Mosul, Iraq
[3]BioComplex Laboratory, Exeter, UK
[4]Department of Mechanical Engineering, University of Mosul, Mosul, Iraq

Article Info

Article history:

Received Oct 23, 2020
Revised Aug 25, 2021
Accepted Sep 10, 2021

Keywords:

Academic-performance evaluation
Co-authorship networks
Collaboration networks
Data mining
Intelligent web crawler
Network measurements
Scientific collaboration

ABSTRACT

Scientific research is currently considered one of the key factors in the development of our life. It plays a significant role in managing our business, study, and work more conveniently. One of the important aspects when it comes to scientific research is the level of collaboration among researchers/disciplines. The collaboration between two different disciplines contributes to obtaining more reliable solutions for our everyday issues. Therefore, it is needed to understand the collaboration patterns among researchers and come up with convenient strategies for strengthening this kind of collaboration. In this work, we aim at investigating the patterns of scientific collaboration among researchers across disciplines. To this end, we generate a co-authorship network for several disciplines. The generated network reveals many interesting facts regarding the collaboration patterns among researchers who work in the same/different disciplines. We involve several measurements in this study that evaluate different aspects, which is of interest to the research communities since most of the studies in the literature measure specific aspects. Moreover, we propose a novel metric for measuring scientific collaboration in a research community and use it to benchmark the collaboration among disciplines. Finally, we use the obtained results/facts in providing recommendations for scientific communities.

Corresponding Author:

Basim Mahmood
Department of Computer Science
University of Mosul
University Street, University of Mosul, Mosul, 41002, Iraq
Email: bmahmood@uomosul.edu.iq

1. INTRODUCTION

Data mining and artificial intelligence fields are intertwined and have caused a paradigm shift in the literature of data analysis. It currently plays a crucial role in analyzing and comprehensively understanding our data [1]. The analysis of data can be performed in many different techniques such as traditional statistical approaches. However, with the advent of complex networks, researchers have become able to deeply investigate the relations among data entities. In complex networks, data can be formed as a network structure with nodes and edges such that friendship networks, citation networks, collaboration networks, road networks, gene networks, and co-authorship networks. In the context of this work, a co-authorship network can be represented as a graph, in which the nodes denote network authors and edges among them that are formed when co-authoring articles. Co-authorship networks have been used for investigating the collaboration patterns that might exist among scientific researchers [2]. In a co-authorship network [3], two or more authors are considered to be connected if they have co-authored an article. Figure 1 shows a simple example of how a co-authorship network is generated. Given 6 authors (R1 to R6), and 3 articles (Article 1,

Article 2, and Article 3). As can be seen in the figure, a network of 6 nodes is generated including the edges among them. These edges are generated based on co-authoring in articles. The concept of co-authorship is also used to measure the scientific status of a researcher in a particular research community as well as predicting future potential collaboration [4], [5]. Investigating co-authorship networks is important since it plays a significant role in understanding the dissemination of knowledge within research communities. The analysis approach of this work is based on concepts inspired by the complex networks field [6]. This field has emerged from computer science graph theory, statistics, and sociology. Furthermore, using this field of study enables us to deeply investigate the relations among network actors (authors). This approach is a technique that generates a network and measures its properties. For instance, one can analyze the relationships among research groups or collaborating teams using network measurements at different levels. The characteristics of a network can be described in two levels [7]: at the entire network level and individual level. In the former, we can measure the density, diameter, clusters (research communities) of a network. In the latter, we can analyze the centrality of network nodes such as degree, betweenness, and closeness centralities. Each of these measurements can be used to extract a specific fact on the network/node. However, these measurements, separately, cannot measure the overall performance of a network/node.

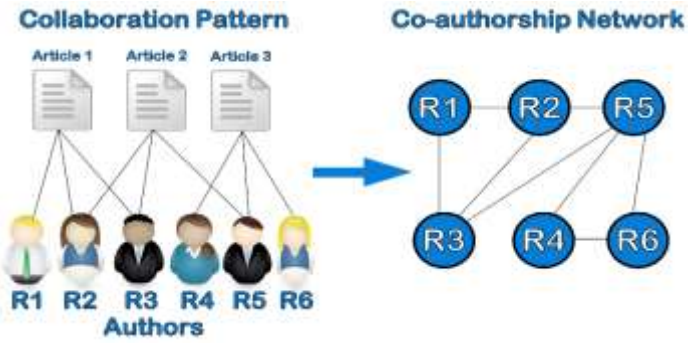

Figure 1. An Example of generating a co-authorship network

Recently, the area of co-authorship networks has attracted research communities due to its role in improving and strengthening the collaboration patterns among researchers. Revealing the patterns of this kind of collaboration has been studied in the literature such as the distinguished studies of Newman [3] and Girvan-Newman [8], they used three bibliographic datasets for three fields of study; biology, mathematics, and physics. The goal of these studies was to investigate the collaboration patterns among the authors who work in the same area of study as well as among the three mentioned areas. They used statistical tools in the analysis approach. One of the interesting results that were obtained, the biological scientists have a strong tendency to co-author papers with authors from the same field, and this tendency is significantly decreased for mathematicians or physicists. These results were also confirmed in the study of Coccia and Bozeman [9]. In another study by Newman [10], they generated three networks for three areas of research; Computer Science, Physics, and Biomedical Research. They investigated and studied these networks and found the best-connected scientists in terms of the strength of collaboration. In the analysis, they used network centrality measurements. In 2013, Divakarmurthy and Menezes [11] investigated the collaboration patterns and citation patterns among the authors of the association for computing and machinery (ACM). They generated two collaboration networks; the first one was based on the citation of articles and the second was based on publications only without considering articles' citations. They used several network measurements such as degree centrality, betweenness centrality, closeness centrality, and the characteristics of community structure in evaluating authors. They also compared the results of the two generated networks to rank authors. Furthermore, the characteristics of Social Networks can also be considered as a useful tool in understanding the collaboration patterns among researchers as presented in the distinguished work of Barabasi *et al.* [12]. They considered the evolution of the social network of scientific collaboration in deeply understanding what is driving the collaboration patterns that exist among researchers.

The main problem in the literature is the lack of providing a deep evaluation of the performance of research communities/researchers in terms of collaboration. Most of the works perform the evaluation based on fixed factors (e.g., the number of published articles and citations) or the evaluation is performed using a single network measurement that reflects one fact on a network/author. Furthermore, the majority of works in the literature were performed using specific repositories. This case might force the evaluation to be biased to

the themes (e.g., area of research) of the repository considered in the analysis, which leads to unreliable results.

According to the aforementioned description, we have distinguished two main gaps and can be summarized as:

- Most of the works in the literature considered few fields of research from a particular repository (e.g., ACM, IEEE). In fact, investigating the relations/collaboration among disciplines need to include as many disciplines as possible and varying the repositories aiming at enriching the dataset with fruitful items and eventually provides more reliable results.
- Most of the proposed approaches use a particular measurement that evaluates and reflects one fact on a discipline/author. In this case, the evaluation process looks at the problem from one angle, which is not sufficient when seeking a comprehensive analysis.

In this article, we try to fill the aforementioned gaps and come up with results and recommendations for research communities. Hence, the contributions of this work can be summarized as:

- Generate a co-authorship network that includes mainly 6 major disciplines (and subdisciplines). Many repositories are used and the data collection is performed on Google Scholar. This makes the dataset colorful of items and the analysis to be more comprehensive and reliable.
- Propose a novel metric that can evaluate the strength of collaboration in a research community (discipline). This metric is based on several network measurements, which gives a view from different angles (utilizing different features) and obtain a more accurate evaluation.

We believe that filling the aforementioned gaps will make the difference between our work and the works in the literature. There are many web applications (e.g., Publons, Scopus, Google Scholar) that evaluate the scientific status of an institution/researcher using well-known indicators such as h-index, Cite-Score in Scopus, or Impact Factor (SCIE and SSCI) in Clarivate. However, our approach evaluates the scientific collaboration based on the relations among disciplines/researchers and this feature is not available in current web applications. The main advantage of our work is that; it can be used (or integrated) to implement a web application that provides recommendations for a research community in real-time, which is a new service that can be shared.

This article is organized as; the next section includes our research methodology including the dataset collection and network measurements. Section 3 contains the results obtained and a discussion on the generated co-authorship network and the proposed metric. Finally, we conclude our article in section 4.

2. RESEARCH METHOD

2.1. Data collection

The dataset of this work was collected from worldwide authors. The collection process was performed on Google Scholar. We designed a special-purpose crawler for retrieving the data using the R language. The collected dataset included information on researchers and the articles they have authored/co-authored. We generated a co-authorship network using the collected dataset that includes researcher name, discipline, affiliation, number of articles published, articles' titles, number of co-authors in each article and co-authors names, journal/conferences name for each article, and publishing year. The total number of authors in our dataset was about 3444 authors after removing the noisy data. The strategy that was used in dealing with the co-authorship network was based on classifying authors into groups of disciplines. The fields of chemistry, physics, and biology were formed as a science group (SC). In the same way, engineering group (EN) including architectural, civil, mechanical, and electrical fields; computing group (CO) including computer science, mathematics, statistics, and operations research; education group (EDU); agriculture group (AG); and business and administration group (BA) including marketing, finance, accounting, and business administration. The dataset contains 6 main disciplines, which are our targeted disciplines in this work. It should be mentioned that the number of articles used in this work was 12,532. Dividing this number on the number of authors taken in this work (3444) leads to having an average of 3.6 papers per author in our co-authorship network, taking into considerations there are single-author papers with no collaborators. The data collection had the issue of author name disambiguation. To solve this issue, we used a particular strategy, which states that distinguishing authors can be performed through their ORCID numbers. However, in case of this identifier is not available for a particular author, we used his/her name along with his/her affiliation as the primary key in the database. Furthermore, we took into considerations that this issue cannot exist with the authors from the same institution. As mentioned, our co-authorship network contains international authors from worldwide institutions. During the data collection, our crawler went through two depths. More precisely, in depth-1 the crawler started randomly with a publication and took its main author and co-authors (if available). Then, the crawler looked for those co-authors if they appear as the main authors/co-authors in other publications and extract their information, which is depth-2. We strongly believed that this strategy

further enriched our dataset with colorful patterns of collaboration, which is our purpose in this work. The generated network included 3444 nodes and 4240 edges.

2.2. Network measurements

The analysis of this work was based on many network measurements, each of which has the ability to reveal a particular fact on the co-authorship network. The main reason behind using these indicators was to enable readers to understand each measurement and what can evaluate in a co-authorship network. These measurements can be either used at the node (author) level or network (community or discipline) level. Now, we consider our network graph G=(N, E), where *N* represents network nodes and *E* represents network edges. The measurements we have used in this work are:

– Clustering Coefficient (*C*): it reflects the tendency of network nodes to cluster together [13]. The value of *C* depends on the number of triangles that are formed by a particular node (3 nodes connected to each other). In a co-authorship network, *C* measures the tendency of authors in co-authoring articles and can be local clustering coefficient (C_i) or global clustering coefficient (C_G). The former can be defined for each author as:

$$C_i = \frac{2|\{l_{jk}: n_j, n_k \in N_i, l_{jk} \in E\}|}{k_i(k_i - 1)} \quad (1)$$

where l_{jk} is an article between the authors n_j and n_k. N_i is the total network authors and k_i is the neighbors' authors in the network. On the other hand, the average clustering coefficient (global) (C) of a network G can be defined as:

$$C_G = \frac{\sum_{i=1}^{n} C_i}{N} \quad (2)$$

where C_i is defined in (1) and N is the number of network authors.

– Average Path Length (l): for all possible pairs of authors in the network, it is defined as the average number of paths (steps) for all the shortest paths among the pairs of authors [13]. It shows the average shortest distance among authors and can be defined as:

$$l = \frac{1}{n(n-1)} \sum_{i \neq j} \quad (3)$$

where d_{ij} is the length between author i and author j.

– Diameter (O): for a network, it is the longest path among all the shortest paths [14]. In our work, it calculates the distance between the farthest authors in the network.

– Density (D): it is the proportion of the number of network edges to the number of potential (possible) edges in that network, which means how close a network to be a fully-connected [14]. It shows the collaboration density among authors as well as the potential collaboration among authors and can be defined as:

$$D_G = \frac{2\big(E(G)\big)}{N(N-1)} \quad (4)$$

– Communities (cu): refers to the groups of nodes in a network that are densely connected with each other. In co-authorship networks, it reveals the research groups that have articles in common (collaborative groups). In our work, we used the Girvan-Newman Clustering algorithm [8] to find the number of research communities in the network and for each discipline. This algorithm detects the links (edges) that connect network communities then removes these links and leaves only the communities themselves. This technique uses a centrality measurement called betweenness.

– Betweenness Centrality (C_b): it shows how many times a node appears in the shortest path of network pairs [14]. It reveals the importance of a particular node in the flow of information within a network. In other words, it represents the importance of an author in a research community. In this work, C_b shows how influential an author in a research community and within a discipline. C_b of the author j can be defined:

$$C_b(j) = \sum_{i \neq j \neq k} \frac{\sigma_{ik}(j)}{\sigma_{ik}} \quad (5)$$

where σ_{ik} is the shortest path between the authors i and k. $\sigma(j)$ is the number of paths that pass through author j.

- Degree Centrality (C_d): it reflects the number of connections that a particular node has in a network [14]. In the context of this work, it reflects the actual number of papers an author has published.
- Closeness Centrality (C_c): it represents the reciprocal of the sum of all the shortest paths of a node to other network nodes [14]. It shows how close an author to other authors in a research community and can be described as:

$$C_c(i) = \frac{N-1}{\sum_j d(ji)} \quad (6)$$

where $d(ij)$ is the distance between the authors i and j. This measurement will be further used in the proposed metric.

3. RESULTS AND DISCUSSION

3.1. Co-authorship network

As mentioned, our co-authorship network (CAN) was generated based on a particular strategy. It states that when two authors have participated in co-authoring an article, a link is established between them. This strategy was also followed in [8], [10], [11], [15]. CAN network consisted of 3444 authors and the number of edges was 4240 that connecting the authors. The visualization of the network is shown in Figure 2, which shows the dense level of co-authoring articles among authors from different disciplines. In the figure, the size of nodes reflects the frequency of collaboration of a particular author in the CAN network.

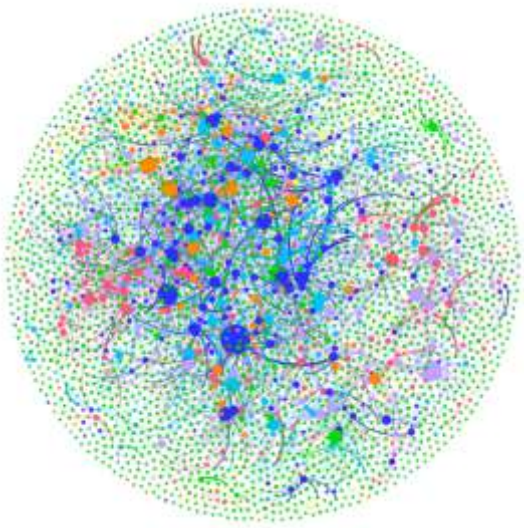

Figure 2. Visualization of CAN network, different colors reflect different disciplines

Moreover, according to the Girvan-Newman Clustering algorithm [8], CAN include 486 potential research communities with a strong modularity level of 0.875 as shown in Figure 3. Also, all the potential communities include authors from different disciplines, which is a positive indicator of the future of scientific collaboration among disciplines. It should be mentioned that the number of communities is dynamic and changed over time.

Figure 4 depicts the degree distribution of the CAN network, which followed a power-law. According to [3], [12], co-authorship networks follow this kind of distribution since there exist few authors with a high degree (authored/co-authored large number of articles), while a large number of authors with low

degree. This phenomenon reflects one of the most important features in co-authorship networks. Based on the distinguished work of Barabasi and Bonabeau [16], when the degree distribution of a network follows a power-law, the network is considered to be Scale-Free Network. Therefore, the nodes of CAN network evolution are governed by the preferential attachment feature according to [12], [17]. This feature reflects important facts on a network. One of these facts is that the newly connected authors prefer to attach to the highly connected authors within the network. It can be inferred, fresh researchers should try to consider senior researchers for their future collaborations because this leads to improve their positions in the research community and eventually improve the quality of their researches.

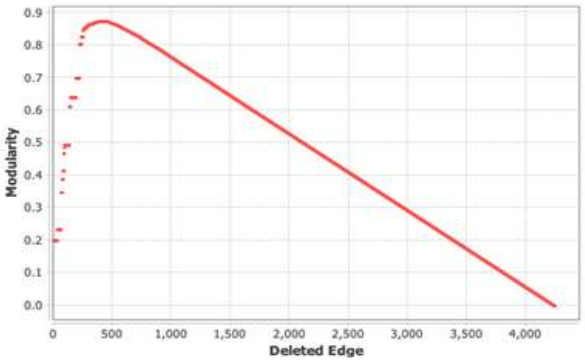

Figure 3. Modularity level; measures the strength of splitting a network into groups or communities. High values of modularity reflect dense connections among the extracted communities (disciplines)

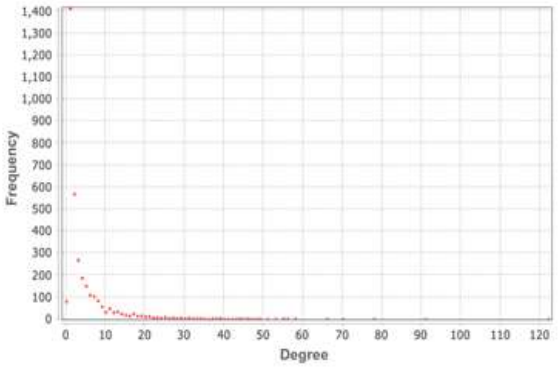

Figure 4. Degree distribution of CAN network (a power-law distribution)

Furthermore, we benchmarked the CAN network with other co-authorship networks in the literature. The goal of this comparison is to see how our network performs compared to other networks. Table 1 presents a comparison between CAN and 4 other co-authorship networks, namely, ACM, Biology, and Physics networks [11], and Engineering network [18]. According to the aforementioned table, the C value of CAN reflected a weak tendency of authors to collaborate. Also, the tendency (C) of authors from the same discipline is stronger than in authors from different disciplines. The l value was higher compared to the benchmarking networks. This means the shortest distance among CAN authors is longer than what was obtained from other networks. Moreover, CAN is less dense compared to the benchmarking (Density (D)=0.001), this is due to the number of disciplines (and subdisciplines) compared to the other networks. Finally, the Diameter (O) is 19 in CAN, which is higher than the other networks.

The above-mentioned results confirm our first claim in this work. The performance of CAN in terms of network measurements (C, l, D, O) underperformed the benchmarking networks. This means evaluating a coauthorship network needs to have a dataset that is colorful in disciplines and repositories.

Table 1. Comparison between the can network and other similar co-authorship networks

Network	C	l	D	O
CAN (multidisciplinary from several repositories)	0.016	7.613	0.001	19
ACM (multidisciplinary from ACM repository)	0.060	4.990	0.01	15
Biology (uni-disciplinary from ACM repository)	0.880	4.920	0.01	6
Physics uni-disciplinary from ACM repository	0.450	6.190	0.01	9
Engineering (uni-disciplinary from CNPq repository)	0.293	8.464	0.015	13

3.2. Disciplines network

In this section, we extracted 6 sub-networks from CAN, each of which represents a particular discipline group. Practically, it is needed to perform a comparison among these disciplines in terms of the measurements mentioned in section 2.2. Table 2 presents the disciplines with the corresponding values of measurements. It can be observed that the SC discipline group has the highest number of communities, which reflects the highest level of collaboration among the other disciplines. In fact, a high number of communities in a network reflects the strong tendency of their authors to collaborate.

Furthermore, when observing the number of communities in a network, the number of authors in that network should also be observed and taken into considerations. Therefore, we see that the performance of a network in terms of forming research communities should be measured according to (as we propose) the ratio (r) of the number of scientific communities (or research groups) (cu) to the actual number of authors (number of nodes) as:

$$r(dis) = \frac{1}{N} cu(dis) \tag{7}$$

where dis denotes a discipline. According to r, most of the groups have almost close levels of forming communities as shown in Table 2. This leads to a better evaluation when investigating the scientific collaboration among authors in a co-authorship network. Furthermore, in the CAN network, the largest number of scientific communities has existed in the SC group. It also has the second-highest tendency of authors to cluster and collaborate with each other. It is inferred that increasing the number of research groups plays a significant role in increasing the research productivity of that group. Therefore, we found that about *35%* of the research articles in our network belong to the SC group. This result was also verified in other works in the literature such as [3], [10], [12], [19], [20]. These works found that the authors in the field of Science have strong tendencies to collaborate. On the other hand, although its small number of communities, the engineering discipline (EN) reflects the strong relations (*C=0.347*) of its authors to collaborate and co-author articles with a high dense level of collaboration (*D=0.011*) compared to the other disciplines. Yet, there is a difference between the number of communities and the strengths of these communities (the strength of authors' connections in a community). Furthermore, when comparing our results with the results of [11], the worldwide researchers of the Engineering discipline reflect approximately the same pattern of C in terms of their tendency to collaborate. As can be seen, each network measurement can be considered as an indicator to reflect a specific fact on network communities and the collaboration patterns in disciplines.

Table 2. In comparisons of the disciplines, the values are ordered based on the number of communities that each group has in its network (from high to low)

Discipline network	# of cu	# of authors	r	C	l	D	O
SC	118	624	0.189	0.319	5.330	0.005	15
EDU	89	549	0.162	0.309	2.575	0.006	7
CO	81	398	0.203	0.234	4.018	0.007	9
AG	67	462	0.145	0.28	2.439	0.008	6
EN	66	620	0.106	0.347	4.056	0.011	10
BA	65	791	0.082	0.232	4.806	0.01	10

3.3. CAN best connected authors

As mentioned, network measurements can be used in two levels (network and node levels). In this section, the CAN network is analyzed based on node-level measurements. We aim at using centrality measurements for evaluating authors from different disciplines. The goal of this analysis is to reveal the disciplines that have the best-connected authors based on authors' relations. The other reason for this analysis is to have an indicator for our further discussions. Table 3 ranks CAN authors based on the values of betweenness centrality (C_b) measurement and the frequency of collaboration with other authors. In the context of co-authorship networks, C_b reveals how influential an author in a research community. It shows the number of times an author appears in the shortest paths of network pairs. The results show that the science discipline (SC) has 4 authors out of the top 10 best-connected authors. Furthermore, in addition to the frequency of collaborations, the position of an author in a research community is another important factor that can be used in assessing an author. For instance, the CAN network has authors with more than 122 published articles but their positions in the network do not make them influential. This means the positions of authors in a research community play a crucial role in the level of collaboration in the whole community. In section 3.1, we showed that the CAN network is scale-free and the concept of preferential attachment [21] is applicable. Therefore, increasing the level of collaboration among CAN authors can also be obtained when the authors tend to be connected and attached (collaborate) to best-connected authors within the network. This specific case leads to an increase in the number of triangles in the network (3 authors are connected and collaborated), which eventually increases the global clustering coefficient of the network. The concept of preferential attachment is based on the concepts of clustering coefficient and degree centrality. This means when an author has a strong tendency to collaborate and he/she has frequently collaborated with other authors, the probability of preferential attachment is also increased. The results show that the probability of

preferential attachment in Science and Engineering disciplines is the highest compared to other disciplines (0.758 and 0.701 respectively), while all the other disciplines have less than 0.3.

3.4. CAN best connected authors

This section shows the current trends in collaboration among disciplines. The analysis of this section is important insofar as it investigates the integration of the disciplines that are considered in this work. Figure 5 depicts how CAN disciplines are connected and collaborated with each other in terms of co-authoring articles. In this figure, each discipline group is represented as a node, and the edge between the two disciplines is formed if there was a collaboration between them. Usually, the collaboration among disciplines is almost performed when inspiring theories from a discipline and involving them in another one. The figure also depicts the level of collaboration between each pair of disciplines, which is represented by the weight of edges The size of nodes reflects the actual size of the disciplines in terms of the number of co-authored articles. It is clear that some pairs of disciplines reflect a high level of collaboration such as the pairs (CO-EDU, BA-CO, EN-CO, and SC-AG). The integration among disciplines leads the value of C to be 0.583, which is acceptable, and l among CAN pairs equals 1.58, which is long compared to the size of this graph. Moreover, the highest value of C_b was gained by the engineering discipline (EN), which reflects the strong tendency of the authors in this discipline to collaborate and contribute more to other disciplines. Practically, the integration of a particular field of research with other fields opens the horizon to the authors of both fields to come up with new contributions that will significantly improve the quality of research and share the knowledge.

Table 3. Top 10 best-connected authors according to their disciplines. The items in this table are ascendingly ordered based on the values of C_b. The frequency of collaborations is also listed in the table, which expresses the number of co-authored articles of an author

Discipline Group	Field of Study	C_b	Collaboration Frequency
SC	Physics	476227.25	122
SC	Chemistry	393431.53	56
EDU	Computer Education	365928.79	34
AG	Animal Production	325392.82	33
EN	Architectural	285042.29	31
SC	Chemistry	245861.18	38
BA	BA	242574.54	41
BA	Marketing	209667.36	44
SC	Chemistry	203790.99	46
BA	BA	195877.80	35

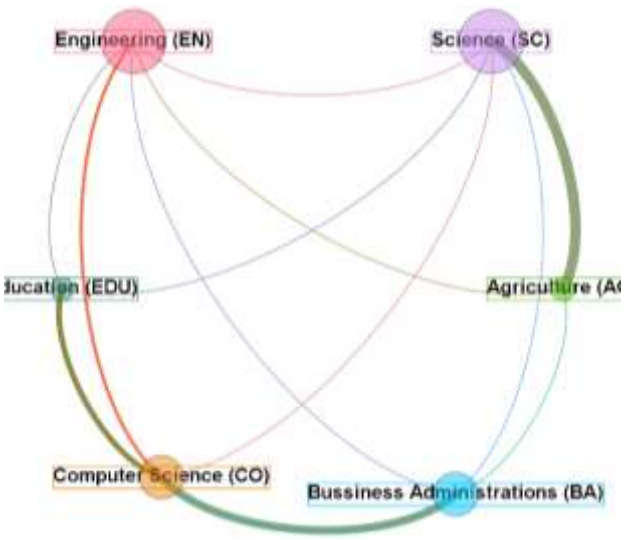

Figure 5. The collaboration among the disciplines in CAN network. The weight of the edges reflects the strength of the collaboration for each pair of disciplines

3.5. The proposed approach

In the previous sections, we measured the level of collaboration among authors/disciplines. The way we followed was based on standard network measurements. In this section, we describe the proposed metric and how it is used for measuring scientific collaboration in disciplines.

In developing the proposed metric, we were inspired by the concept of social capital in sociology [20]-[22]. In social communities, social capital can be defined as the collectively shared values, cooperation, and reciprocity among individuals [20]. Social capital is not a measurement for network nodes, instead; it evaluates the relations among nodes [23]. Therefore, we propose to incorporate centrality measurements in developing a novel metric for measuring the global level of collaboration in academic communities. Furthermore, clustering coefficient measurement is an expression of the authors' tendency to collaborate with others. Therefore, we propose to incorporate it into the proposed metric. Since we are investigating scientific collaboration among academic communities, we also propose to use the term collaboration capital (CC). This term is used in business and economic contexts [24] referring to the collaborative processes that improve the outcomes. The value of CC for a particular author represents the collective centralities and the clustering values of that author. On the other hand, the value of CC for a community (discipline) represents the average values of all authors in a discipline (means CC). The selected measurements give our proposed metric the ability to deeply investigate the relations among authors as well as their positions in CAN structure. We believe this way is efficient in evaluating the collaboration capital in a research community. Moreover, our metric looks at the network from many different points of view because it combines the characteristics of many measurements (features) in one strong metric.

The value of CC for a particular author i in a scientific community (cu) can be defined as:

$$CC_{cu}(i) = \sum_{i=1}^{n} \big(CE(i) + c(i)\big) \tag{8}$$

where c is the clustering coefficient of author i in a community cu. CE represents the collective value of the centrality measurements of a particular author and can be formalized as:

$$CE(i) = (C_b(i) + C_c(i))^{C_d(i)} \tag{9}$$

where C_b, C_c, and C_d are betweenness as shown in (5), closeness as shown in (6), and degree centralities for author i respectively. The power (C_d) in (9) reflects the frequency of collaborations of an author. This will contribute to increasing CC because the frequency of collaboration is considered as an important factor that should be given its actual role when measuring scientific collaboration.

According to the aforementioned description, the CC of a particular community or discipline represents the average collaboration capital values of the authors in that discipline as:

$$CC_{cu} = \frac{\sum_{i=1}^{N} CC_{cu}(i)}{N} \tag{10}$$

In the experimental results, the value of CC of a discipline follows a power-law distribution (long-tail distribution). All the values were normalized to be in the range of 0 and 1. The CC of each discipline is shown in Table 4. It can be observed that the CO and EN disciplines outperform the other disciplines in terms of collaboration capital as well as the strong ability of their authors to collaborate. On the other hand, SC discipline does not reflect a high performance of CC. Although four of the top 10 best-connected authors in the CAN network are from the SC discipline, the results show that its collaboration capital underperformed the aforementioned two disciplines. We believe this is due to First**:** the high level of variations in the clustering coefficient of authors, which is reasonable because the degree distribution of the CAN network follows a power-law. Second**,** the creation of links among network authors is controlled by the concept of preferential attachment as stated in [25].

According to the performance of the proposed metric (collaboration capital), we see that the results show a different perspective from what we have analyzed in the previous sections. For instance, the performance of the Science discipline is changed when it comes to collaboration capital. Therefore, using one measurement in evaluating scientific collaboration is insufficient and not reliable. This result confirms our second claim in this work. The evaluation should be performed using a metric that looks at the problem from many different angles aiming at having a clear view and coming up with an accurate and reliable assessment.

Table 4. Mean collaboration capital for CAN disciplines

Discipline	# of Authors	Mean CC
CO	398	0.145
EN	620	0.126
SC	624	0.08
AG	462	0.075
BA	791	0.07
EDU	549	0.043

3.6. Recommendations

Based on the obtained results, we summarize our recommendations by the following:

- The concepts of complex networks can be considered as powerful tools in understanding the collaboration patterns in co-authorship networks and this is due to the ability of each concept in analyzing network relations from a different perspective. Therefore, developers can benefit from these concepts when designing academic assessment tools.
- The preferential attachment feature in coauthorship networks plays a significant role in improving research quality. Therefore, it is of benefit for the institutions to encourage this kind of feature within their academic settings.
- The process of measuring the level of collaboration does not depend on a specific metric, it depends on the aspect we are investigating. For instance, one can measure the tendency of a researcher/community to collaborate with others; the clustering coefficient can be involved in this case. In the same context, when exploring the most influential authors in a community; betweenness centrality measurement can be used.
- The number of authors in a discipline is not an important factor for strengthening the collaboration level. Instead, the number of research groups (communities) is an effective indicator that contributes to increasing the level of collaboration among authors. This feature also enriches the quality of research with colorful experiences and provides more trusted and flexible solutions for our life issues.
- Improving the collaboration level is not only about the quantity of the co-authored articles, it is also about with whom the authors are connected (collaborated).
- The concept of social capital in sociology is applicable in co-authorship networks and can be in the form of CC, which is very useful when it comes to scientific collaboration assessment.
- According to the obtained results, there are some facts on the scientific collaboration in disciplines such that; the authors in Science and Agriculture disciplines reflected high performance in terms of co-authoring articles, the Engineering discipline authors have strong tendencies to collaborate, and Computer Science discipline outperforms the other disciplines in terms of collaboration capital.

4. CONCLUSION

In this work, we investigated and analyzed the collaboration patterns among authors from different disciplines considering 6 of them, namely; science, engineering, computing, education, administration and economics, and agriculture. We generated a co-authorship network called CAN network containing all the aforementioned disciplines (and subdisciplines). The dataset in this work was collected from Google Scholar including 3444 authors and 12,532 articles retrieved from many different repositories. We also proposed a novel metric to measure what we called CC. The analysis of this work showed that; accurate evaluating of coauthorship networks can be obtained when including more disciplines and involve different repositories in the dataset. The results also showed that measuring the scientific collaboration in a research community needs to adopt a metric that can capture most of the possible features in that community aiming at producing a more precise evaluation. In future work, we plan to investigate our coauthorship network using more measurements (e.g., Eigencentrality and bridging centrality). We also plan to develop a social-driven approach using some social theories (e.g., assortativity) in exploring research communities and predict future collaboration among disciplines.

ACKNOWLEDGEMENTS

We would like to thank the Computer Science Department at the University of Mosul for all the provided support in making this work achieved.

REFERENCES

[1] X. Wu, X. Zhu, G. -Q. Wu and W. Ding, "Data mining with big data," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014, doi: 10.1109/TKDE.2013.109.

[2] S. Kumar, "Co-authorship networks: a review of the literature," *Aslib Journal of Information Management*, vol. 67, no. 1, pp. 55-73, 2015, doi: 10.1108/AJIM-09-2014-0116.

[3] M. E. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. suppl 1, pp. 5200-5205, 2004, doi: 10.1073/pnas.0307545100.

[4] F. Narin, K. Stevens, and E. S. Whitlow, "Scientific co-operation in Europe and the citation of multinationally authored papers," *Scientometrics*, vol. 21, no. 3, pp. 313-323, 1991, doi: 10.1007/BF02093973.

[5] A. K. Tiwari, G. Ramakrishna, L. K. Sharma, and S. K. Kashyap, "Academic performance prediction algorithm based on fuzzy data mining," *IAES International Journal of Artificial Intelligence*, vol. 8, no. 1, pp. 26-32, 2019, doi: 10.11591/ijai.v8.i1.pp26-32.

[6] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268-276, 2001, doi: 10.1038/35065725.

[7] E. J. Cornblath, D. M. Lydon-Staley, and D. S. Bassett, Harnessing networks and machine learning in neuropsychiatric care," *Current Opinion in Neurobiology*, vol. 55, pp. 32–39, 2019, doi: 10.1016/j.conb.2018.12.010.

[8] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821-7826, 2002, doi: 10.1073/pnas.122653799.

[9] M. Coccia and B. Bozeman, "Allometric models to measure and analyze the evolution of international research collaboration," *Scientometrics*, vol. 108, no. 3, pp. 1065-1084, 2016, doi: 10.1007/s11192-016-2027-x.

[10] M. E. Newman, "Who is the best connected scientist? a study of scientific coauthorship networks," in *Complex Networks, Springer*, vol. 650, pp. 337-370, 2004, doi: 10.1007/978-3-540-44485-5_16.

[11] P. Divakarmurthy and R. Menezes, "The effect of citations to collaboration networks," in *Complex Networks. Springer*, vol. 424, pp. 177-185, 2013, doi: 10.1007/978-3-642-30287-9_19.

[12] A.-L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A: Statistical mechanics and its applications*, vol. 311, no. 3-4, pp. 590-614, 2002, doi: 10.1016/S0378-4371(02)00736-7.

[13] D. J. Watts and S. H. Strogatz, "Collective dynamics of small world networks," *Nature*, vol. 393, no. 6684, p. 440, 1998, doi: 10.1038/30918.

[14] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47-97, 2002, doi: 10.1103/RevModPhys.74.47.

[15] B. M. Mahmood, N. A. Sultan, K. H. Thanoon, and D. S. Khadhim, "Collaboration networks: university of mosul case study," *AL-Rafidain Journal of Computer Sciences and Mathematics*, vol. 14, no. 1, pp. 117-133, 2020, doi: 10.33899/csmj.2020.164679.

[16] A.-L. Barabasi and E. Bonabeau, "Scale-free networks," *Scientific American*, vol. 288, no. 5, pp. 60-69, 2003, doi: 10.1038/scientificamerican0503-60.

[17] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási, "Quantifying the evolution of individual scientific impact," *American Association for the Advancement of Science, Science*, vol. 354, no. 6312, 2016, doi: 10.1126/science.aaf5239.

[18] R. L. D. Andrade and L. C. Rêgo, "Exploring the co-authorship network among cnpq™s productivity fellows in the area of industrial engineering," *Pesquisa Operacional, SciELO Brasil*, vol. 37, no. 2, pp. 277-310, 2017, doi: 10.1590/0101-7438.2017.037.02.0277.

[19] A. Gazni, C. R. Sugimoto, and F. Didegah, "Mapping world scientific collaboration: Authors, institutions, and countries," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 2, pp. 323-335, 2012, doi: 10.1002/asi.21688.

[20] J. S. Coleman, "Social capital in the creation of human capital," *American Journal of Sociology*, vol. 94, pp. 95-120, 1988, doi: 10.1086/228943.

[21] R. D. Putnam, "Bowling alone: America™s declining social capital," in *Culture and Politics*, pp. 223-234, 2000, doi: 10.1007/978-1-349-62965-7_12.

[22] B. Mahmood, M. Tomasini and R. Menezes, "Estimating memory requirements in wireless sensor networks using social tie strengths," *2015 IEEE 40th Local Computer Networks Conference Workshops (LCN Workshops)*, 2015, pp. 695-698, doi: 10.1109/LCNW.2015.7365916.

[23] B. Mahmood and R. Menezes, "The role of human relations and interactions in designing memory-related models for sensor networks," *Sensors and Transducers*, vol. 199, no. 4, pp. 42-51, 2016, doi: 10.5220/0005672600250034.

[24] I. Boughzala, "Collaboration 2.0 through the New Organization (2.0) Transformation," in *Knowledge Management 2.0: Organizational Models and Enterprise Strategies*, *IGI Global*, pp. 1-16, 2012, doi: 10.4018/978-1-61350-195-5.ch001.

[25] A. Vázquez, "Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations," *Physical Review*, vol. 67, no. 5, 2003, doi: 10.1103/PhysRevE.67.056104.

BIOGRAPHIES OF AUTHORS

Basim Mahmood, obtained his Ph.D. degree in Computer Science from Florida Institute of Technology, Melbourne, USA in 2015. He is currently a computer science assistant professor at the University of Mosul, Iraq. He is also a permanent member of the BioComplex laboratory, Exeter, UK. His main fields of interest are complex networks, data mining, and big data analysis.

Nagham A. Sultan, obtained her M.Sc. degree in Computer Science from the University of Mosul, Mosul, Iraq in 2017. Her main field of interest is big data analysis and the Internet of Things (IoT).

Karam H. Thanoon, obtained his Ph.D. degree in Computer Science from the University of Mosul, Mosul, Iraq in 2013. His main fields of interest are big data analysis, software engineering, and information security.

Dheya S. Kadhim, obtained his B.Sc. degree in Information Management, University of Mosul, Mosul, Iraq in 2003. He works as a researcher in the field of complex networks. His main area of interest is complex networks.